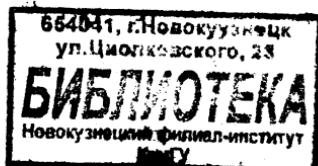


24016

ББК 65в673
И 97



Печатается по решению
редакционно-издательского
совета МОУ ДПО ИПК

Ишкова Л.В. Эконометрика для начинающих: теория и практика: Учебное пособие для студентов, аспирантов, преподавателей экономики и эконометрики (в двух частях). Ч.2 (введение в практическую эконометрику). – Новокузнецк: ИПК, 2003. – 118 с.

ISBN 5-7291-0292-5

Рецензенты: доктор физ.-мат. наук, профессор, зав.кафедрой высшей математики СибГИУ В.И.Базайкин;
доктор экон. наук, профессор, декан экономического факультета НФИ КемГУ И.Г.Степанов

Часть вторая данного учебного пособия является логическим продолжением части первой, написанной Л.В.Ишковой и изданной в 2002 г.

Автором пособия поставлена задача формирования у студентов, аспирантов, слушателей факультетов повышения квалификации, преподавателей экономики и эконометрики практических навыков реализации эконометрических методов в процессе анализа конкретных экономических процессов и проблем.

Часть 2 пособия содержит анализ типовых эконометрических задач, в основе которых встречаются, как правило, два типа выборочных данных:

- пространственные данные (*cross-sectional data*);
- временные ряды (*time-series data*).

Раздел 3 части второй содержит наряду с другими развернутое решение одной задачи эконометрического анализа конкретного временного ряда (от вычисления простейших статистических характеристик временного ряда до составления долгосрочного прогноза развития экономического процесса).

ББК 65в673

И 4306010000
7С2(03) – 2003

ISBN 5-7291-0292-5

© МОУ ДПО ИПК, 2003
© Ишкова Л.В., 2003

Раздел 1.

ПРИМЕНЕНИЕ РАЗЛИЧНЫХ СПОСОБОВ ПРЕДСТАВЛЕНИЯ И ОБРАБОТКИ СТАТИСТИЧЕСКИХ ДАННЫХ

Задачей статистического описания выборки является получение такого ее представления, которое позволит наглядно выявить вероятностные характеристики. Применяются различные формы упорядочения данных в выборке: по возрастанию, по совпадающим значениям, по интервалам и т.п. При анализе какого-то конкретного показателя X в фиксированный момент времени (либо без учета фактора времени) наблюдаемые значения обычно располагают по *неубыванию*: $x_1 \leq x_2 \leq \dots \leq x_n$. Разность между максимальным и минимальным значениями СВ X называется *размахом выборки*: $d = x_{\max} - x_{\min}$.

Пусть количество различных значений в выборке равно k ($k \leq n$). Значения x_i , где $i = 1, 2, \dots, k$, называются *вариантами*. Если значение x_i встретилось в выборке n_i раз, то число n_i называют *частотой* значения x_i , а величину $\omega_i = n_i / n$ – *относительной частотой* значения x_i . Тогда данные выборки можно сгруппировать в *обычный статистический ряд* (таблица 1.1):

Таблица 1.1

X	x_1	x_2	x_3	x_k	$\sum n_i = n$
n_i	n_1	n_2	n_3	n_k	
$\omega_i = n_i / n$	n_1/n	n_2/n	n_3/n	n_k/n	$\sum n_i/n = 1$

По статистическому ряду можно построить эмпириическую функцию распределения: $F^*(x) = n_x / n$, где n_x – число значений случайной величины X , меньших, чем рассматриваемое x_i (лежащих левее на числовой оси); n – объем выборки.

По определению $F^*(x)$ обладает следующими свойствами:

1. $0 \leq F^*(x) \leq 1$.
2. Для любых $x_1 \leq x_2$ $F^*(x_1) \leq F^*(x_2)$.
3. $F^*(x) = 0$ при $x \leq x_1$; $F^*(x) = 1$ при $x \geq x_k$.

При большом объеме выборки ее элементы могут быть сгруппированы в *интервальный статистический ряд*. В общем случае он имеет следующий вид (таблица 1.2).

Таблица 1.2

$[x_{i-1}, x_i)$	$[x_0, x_1)$	$[x_1, x_2)$	$[x_{k-1}, x_k)$
n_i	n_1	n_2	n_k
n_i / n	n_1/n	n_2/n	n_k/n

В этой таблице k непересекающихся интервалов равной длины h (h – шаг разбиения); n_i – количество наблюдаемых значений СВ X , попадающих в i -й интервал; $\omega_i = n_i / n$ – относительная частота попадания СВ X в i -й интервал.

Интервальный статистический ряд наглядно может быть представлен в виде *гистограммы* – графика, в котором по оси абсцисс откладываются подынтервалы, на i -ом из которых строится прямоугольник высотой n_i / nh .

Справка для вычисления выборочных характеристик

Для любой СВ X , кроме определения ее функции распределения, желательно указать числовые характеристики, важнейшими из которых являются *средняя величина (математическое ожидание)*, *дисперсия*, *среднее квадратическое отклонение*.

Рассмотрим числовые характеристики выборки (n – объем выборки).

Выборочное среднее – это среднее арифметическое наблюдаемых значений выборки:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad i = 1, 2, \dots, n.$$

Выборочное среднее при задании выборки в виде статистического ряда записывается так:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i, \quad k \text{ – число групп в выборке.}$$

Выборочная дисперсия:

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

Во многих случаях удобно пользоваться формулой:

$$D = \frac{1}{n} \sum_{i=1}^n (x_i^2 - 2x_i \cdot \bar{x} + (\bar{x})^2) = \bar{x}^2 - \bar{x}^2.$$

При задании выборки в виде статистического ряда выборочная дисперсия находится по формуле:

$$D_s = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2.$$

Выборочное среднее квадратическое отклонение:

$$\sigma = \sqrt{D} = \sqrt{\frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2} = \sqrt{\bar{x}^2 - \bar{x}^2}.$$

При задании выборки в виде *интервального* статистического ряда в формулах выборочного среднего, выборочной дисперсии, выборочного среднего квадратического отклонения вместо x_i рассматривается среднее

значение i -го подынтервала: $\bar{x}_i = \frac{x_{i-1} + x_i}{2}$, а n_i есть частота вариантов в интервале.

Выборочный коэффициент вариации V определяется отношением выборочного среднего квадратического отклонения к выборочной средней, выраженным в процентах: $V = \frac{\sigma}{\bar{x}} \cdot 100\%$. Коэффициент вариации – безразмерная величина, удобная для сравнения величин рассеяния двух выборок, имеющих различные размерности. Совокупность считается количественно однородной, если коэффициент вариации не превышает 33%.

Примеры решения избранных задач

Задача 1.1. Найти математическое ожидание и дисперсию случайной величины $A = 4B - 3C + 2$, если даны $M(B) = 3$, $M(C) = 4$, $D(B) = 2$, $D(C) = 1,5$ и известно, что B и C – независимые переменные.

Решение

Используем свойство математического ожидания: $M(X \pm Y) = M(X) \pm M(Y)$. Получаем:

$$M(A) = 4M(B) - 3M(C) + 2 = 4 \cdot 3 - 3 \cdot 4 + 2 = 2.$$

Используем свойства дисперсии: 1) $D(kX) = k^2 D(X)$; 2) $D(C) = 0$. Получаем:

$$D(A) = 4^2 \cdot D(B) + (-3)^2 \cdot D(C) + D(2) = 16 \cdot 2 + 9 \cdot 4 + 0 = 68.$$

Задача 1.2. Функция распределения случайной величины X имеет вид:

$$F^*(x) = \begin{cases} 1 & \text{при } x \leq 0; \\ \frac{x}{3} & \text{при } 0 < x \leq 2; \\ 2x & \text{при } x > 2. \end{cases}$$

Найти вероятность того, что случайная величина X примет значение в интервале $[1; 4]$. Найти плотность вероятности случайной величины X . Вычислить квантиль $x_{0,5}$ и 40%-ную точку случайной величины X .

Решение

Воспользуемся свойством функции распределения: вероятность попадания случайной величины X в интервал $[x_1; x_2]$ равна приращению ее функции распределения на этом интервале:

$$P(x_1 \leq X < x_2) = F^*(x_2) - F^*(x_1).$$

Получаем:

$$P(1 \leq X < 4) = F^*(4) - F^*(1) = 2 \cdot 4 - \frac{1}{3} = \frac{23}{3} = 7\frac{2}{3}.$$

Для нахождения плотности вероятности СВ X $\varphi(x)$ воспользуемся свойством плотности вероятности непрерывной СВ: функция распределения непрерывной случайной величины может быть выражена через плотность вероятности по формуле:

$$F^*(x) = \int_{-\infty}^x \varphi(x) dx \Rightarrow \varphi(x) = (F^*)'(x).$$

Получаем:

$$\varphi(x) = \begin{cases} 0 & \text{при } x \leq 0, \\ \frac{1}{3} & \text{при } 0 < x \leq 2, \\ 2 & \text{при } x > 2. \end{cases}$$

Квантилем уровня q (q -квантилем) называется такое значение x_q случайной величины, при котором функция ее распределения принимает значение, равное q , то есть:

$$F^*(x_q) = P(X < x_q) = q.$$

На основании определения приходим к выводу:

$$F^*(x_{0,5}) = 0,5, \text{ то есть } \frac{x_{0,5}}{3} = 0,5 \Rightarrow x_{0,5} = 1,5.$$

40%-ная точка случайной величины X или квантиль $x_{1-0,4} = x_{0,6}$ находится аналогично из уравнения $\frac{x_{0,6}}{3} = 0,6 \Rightarrow x_{0,6} = 1,8$.

Задача 1.3. Анализируется прибыль X (%) предприятий отрасли. Обследованы $n = 100$ предприятий, данные по которым занесены в следующий статистический ряд (таблица 1.3):

Таблица 1.3

X	5	10	15	20	25
n_i	5	20	40	25	10
n_i/n	0,05	0,2	0,4	0,25	0,1

Задание

Построить эмпирическую функцию распределения $F^*(x)$.

Решение

$$F^*(x) = \begin{cases} 0, & x \leq 5; \\ 0,05, & 5 < x \leq 10; \\ 0,25, & 10 < x \leq 15; \\ 0,65, & 15 < x \leq 20; \\ 0,90, & 20 < x \leq 25; \\ 1 & x > 25. \end{cases}$$

Задача 1.4. Анализируется доход населения, для чего извлечена выборка объема $n = 300$. По уровню дохода население подразделяется на число групп $k = 6$ при шаге разбиения $h = 20$. Полученные по выборке данные сгруппированы в следующий интервальный статистический ряд (таблица 1.4).

Таблица 1.4

$[x_{i-1}, x_i)$ у.е.	$[0, 20)$	$[20, 40)$	$[40, 60)$	$[60, 80)$	$[80, 100)$	$[100, 120)$
№ интервала	1	2	3	4	5	6
n_i	10	50	80	100	40	20
$\omega_i = n_i / nh$	1/600	5/600	8/600	10/600	4/600	2/600

Задание

Построить гистограмму (график функции $\omega_i = n_i / nh$) и по ее виду выдвинуть предположение о виде закона распределения случайной величины X – дохода населения. Учесть, что в последнюю группу могут быть включены все субъекты, чей доход превышает 100 у.е.

Решение

По оси абсцисс отложим номера интервалов, по оси ординат – функцию распределения n_i / nh . (М: 1: 1/600) (рис. 1.1).

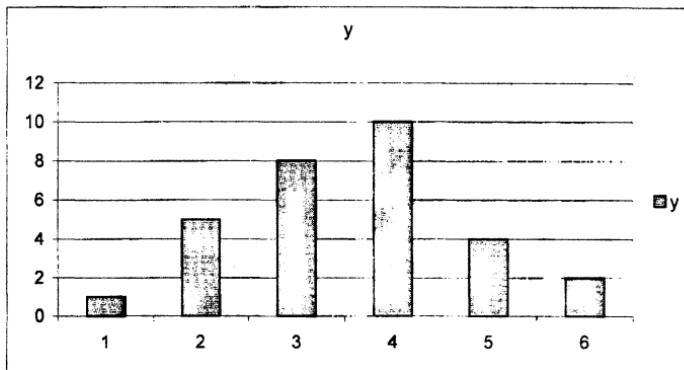


Рис. 1.1. Гистограмма к анализу интервального ряда к задаче 1.4

Форма гистограммы в наибольшей степени соответствует нормальному закону распределения. Поэтому естественным является предположение о нормальном распределении СВ X .

Задача 1.5. Анализируются объемы ежедневных продаж некоторого товара за 60 дней. Получены следующие данные:

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

Задание

- 1) Построить корреляционное поле выборки; сделать вывод об однородности выборки.
- 2) Построить статистический ряд.
- 3) Определить размах выборки.
- 4) Построить эмпирическую функцию распределения.
- 5) Построить полигон частот.
- 6) Определить математическое ожидание, дисперсию, среднее квадратическое отклонение, коэффициент вариации наблюдаемой случайной величины.

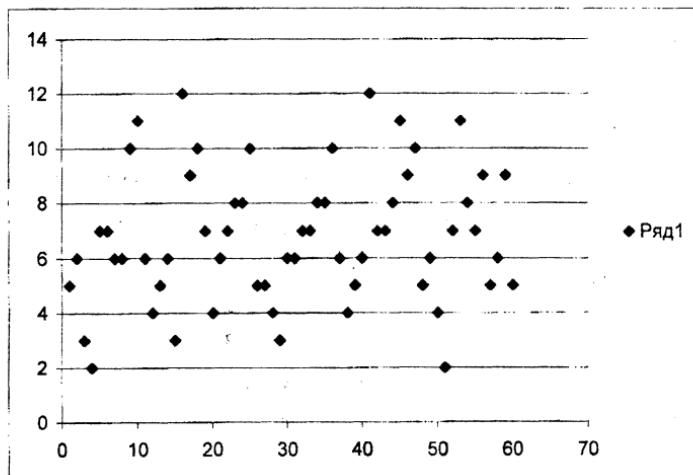


Рис. 1.2. Корреляционное поле статистической выборки к задаче 1.5

Решение

1) Корреляционное поле изображаем в осях: абсцисса – номер элемента в полной выборке, ордината – числовое значение элемента (рис.1.2).

Вид построенного поля корреляции СВ говорит об однородности статистической выборки, то есть о ее достаточно высоком качестве.

2) Строим статистический ряд.

Пусть переменная X – изменяющийся объем ежедневных продаж некоторого товара; n – объем выборки ($n = 60$); k – количество различных значений в выборке ($k = 11$, $k < n$); x_i – варианты значений ($i = 1, 2, \dots, k = 11$); n_i – частота значения x_i (число, показывающее, сколько раз это значение встречается в выборке); $\omega_i = \frac{n_i}{n}$ – относительная частота значения x_i .

Расположив варианты по возрастанию, сгруппируем имеющуюся информацию в статистический ряд, представленный в таблице 1.5.

Таблица 1.5

x_i	2	3	4	5	6	7	8	9	10	11	12	$k = 11$
n_i	2	3	5	8	12	10	6	4	5	3	2	$\sum_{i=1}^k n_i = n$
$\omega_i = \frac{n_i}{n}$	0,03	0,05	0,08	0,13	0,2	0,17	0,1	0,07	0,08	0,05	0,03	$\sum_{i=1}^k \frac{n_i}{n} = 1$

В построенном статистическом ряду вследствие приближенных вычислений (округлений) $\sum_{i=1}^k \frac{n_i}{n} \neq 1$.

3) Размахом выборки d называется разность между максимальным и минимальным значениями рассматриваемой случайной величины X :

$$d = x_{\max} - x_{\min} = 12 - 2 = 10.$$

4) Эмпирическую функцию распределения F^* можно построить по статистическому ряду: $F^* = \frac{n_x}{n}$, где n_x – число значений случайной величины X , меньших, чем x .

$$F^*(x) = \begin{cases} 0, & x \leq 2, \\ 0,03, & 2 < x \leq 3, \\ 0,08, & 3 < x \leq 4, \\ 0,17, & 4 < x \leq 5, \\ 0,3, & 5 < x \leq 6, \\ 0,5, & 6 < x \leq 7, \\ 0,67, & 7 < x \leq 8, \\ 0,77, & 8 < x \leq 9, \\ 0,83, & 9 < x \leq 10, \\ 0,92, & 10 < x \leq 11, \\ 0,97, & 11 < x \leq 12, \\ 1, & x > 12. \end{cases}$$

5) Полигон частот строится в осях: ордината n_i (2,3,5,8,12,10,6,4,5,3,2) и абсцисса x_i (2,3,4,5,6,7,8,9,10,11,12) (рис. 1.3).

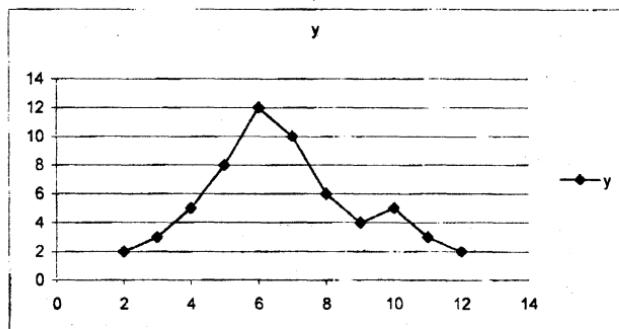


Рис. 1.3. Полигон частот к задаче 1.5

6) Найдем выборочные характеристики (n – объем выборки).

Средняя величина (математическое ожидание) выборки при задании выборки в виде статистического ряда записывается так:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i, k - \text{число групп в выборке.}$$

Получим:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{1}{60} \sum_{i=1}^{11} n_i x_i = \frac{1}{60} (4 + 9 + 20 + 40 + 72 + 70 + 48 + 36 + 50 + 33 + 24) = 6,77.$$

Дисперсия выборки в случае задания ее в виде статистического ряда находится по формуле:

$$D = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2.$$

Для нахождения дисперсии составим таблицу 1.6.

Таблица 1.6

n_i	x_i	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
2	2	7	-5	25	50
3	3		-4	16	48
5	4		-3	9	45
8	5		-2	4	32
12	6		-1	1	12
10	7		0	0	0
6	8		1	1	6
4	9		2	4	16
5	10		3	9	45
3	11		4	16	48
2	12		5	25	50
					$\sum_{i=1}^{11} n_i (x_i - \bar{x})^2 = 352$

$$D = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2 = \frac{1}{60} \sum_{i=1}^{11} n_i (x_i - \bar{x})^2 = \frac{1}{60} \cdot 352 = 5,87.$$

Выборочное среднее квадратическое отклонение σ :

$$\sigma = \sqrt{D} = 2,42.$$

Все задания выполнены.

Задача 1.6. Анализируется продолжительность телефонных разговоров с клиентами некоторой справочной телефонной службы. Случайным образом отобраны 60 телефонных разговоров и зафиксированы их длительности (в секундах): 39, 60, 40, 52, 32, 68, 77, 61, 68, 60, 47, 49, 70, 55, 66, 80, 35, 67, 70, 55, 42, 52, 60, 82, 70, 55, 47, 39, 50, 58, 45, 50, 53, 33, 49, 54, 55, 70, 62, 60, 60, 40, 59, 64, 70, 55, 54, 35, 48, 52, 57, 55, 82, 70, 51, 35, 49, 60, 55, 47.

Задание

- 1) Сгруппировать данные в обычный статистический ряд и вычислить выборочные характеристики (среднее значение, дисперсию и среднее квадратическое отклонение рассматриваемой величины).
- 2) Построить интервальный статистический ряд, включающий 5 подинтервалов (при этом выбрать самостоятельно величину шага h и объяснить выбор).
- 3) Вычислить выборочные числовые характеристики рассматриваемой величины на основании построенного интервального статистического ряда с числом подинтервалов $k = 5$.
- 4) Построить интервальный статистический ряд, включающий 7 подинтервалов, и вычислить на его основании выборочные числовые характеристики рассматриваемой величины.
- 5) Сравнить результаты вычислений в пунктах 1, 3, 4; сделать вывод.

Решение

- 1) Сгруппируем данные выборки в обычный статистический ряд.

Пусть переменная X – продолжительность телефонных разговоров; n – объем выборки ($n = 60$); k – количество различных значений в выборке ($k = 30$, $k < n$); x_i – варианты значений ($i = 1, 2, \dots, k = 30$); n_i – частота значения x_i (число, показывающее, сколько раз это значение встречается в выборке); $\omega_i = \frac{n_i}{n}$ – относительная частота значения x_i . Расположив варианты по возрастанию, сгруппируем имеющуюся информацию в статистический ряд, представленный в таблице 1.7 (поскольку ряд длинный, запишем варианты, их частоты и относительные частоты в три строчки).

Таблица 1.7

x	32	33	35	39	40	42	45	47	48	49	$k = 30$
n_i	1	1	3	2	2	1	1	3	1	3	
$\omega_i = \frac{n_i}{n}$	0,02	0,02	0,05	0,03	0,03	0,02	0,02	0,05	0,02	0,05	$\sum_{i=1}^{30} n_i = 60$
x	50	51	52	53	54	55	57	58	59	60	
n_i	2	1	3	1	2	7	1	1	1	6	$\sum_{i=1}^{30} \omega_i = 1$
$\omega_i = \frac{n_i}{n}$	0,03	0,02	0,05	0,02	0,03	0,12	0,02	0,02	0,02	0,1	
x	61	62	64	66	67	68	70	77	80	82	
n_i	1	1	1	1	1	2	6	1	1	2	
$\omega_i = \frac{n_i}{n}$	0,02	0,02	0,02	0,02	0,02	0,03	0,1	0,02	0,02	0,03	

В построенном статистическом ряду вследствие приближенных вычислений (округлений) $\sum_{i=1}^k \frac{n_i}{n} \neq 1$.

Найдем выборочные характеристики ($n = 60$ – объем выборки).

Средняя величина (математическое ожидание) выборки при задании выборки в виде статистического ряда записывается так:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i, \quad k \text{ – число групп в выборке.}$$

Получим:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i x_i = \frac{1}{60} \sum_{i=1}^{30} n_i x_i = \frac{1}{60} (32 + 33 + 105 + 78 + 80 + 42 + 45 + 141 + 48 + 147 + 100 + 51 + 156 + \\ + 53 + 108 + 385 + 57 + 58 + 59 + 360 + 61 + 62 + 64 + 66 + 67 + 136 + 420 + 77 + 80 + 164) = 55,58.$$

Дисперсия выборки в случае задания ее в виде статистического ряда находится по формуле:

$$D = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2.$$

Для нахождения дисперсии составим таблицу 1.8.

Таблица 1.8

n_i	x_i	\bar{x}	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$n_i (x_i - \bar{x})^2$
1	32	55,58	-23,58	556,02	556,02
1	33		-22,58	509,86	509,86
3	35		-20,58	423,54	1270,61
2	39		-16,58	274,90	549,79
2	40		-15,58	242,74	485,47
1	42		-13,58	184,42	184,42
1	45		-10,58	111,94	111,94
3	47		-8,58	73,62	220,85
1	48		-7,58	57,46	57,46
3	49		-6,58	43,30	129,89
2	50		-5,58	31,14	62,28
1	51		-4,58	20,98	20,98
3	52		-3,58	12,82	38,45
1	53		-2,58	6,66	6,66
2	54		-1,58	2,50	5,00
7	55		-0,58	0,34	2,35
1	57		1,42	2,02	2,02
1	58		2,42	5,86	5,86
1	59		3,42	11,70	11,70
6	60		4,42	19,54	117,22
1	61		5,42	29,38	29,38
1	62		6,42	41,22	41,22
1	64		8,42	70,90	70,90
1	66		10,42	108,58	108,58
1	67		11,42	130,42	130,42
2	68		12,42	154,26	308,52
6	70		14,42	207,94	1247,62

1	77		21,42	458,82	458,82
1	80		24,42	596,34	596,34
2	82		26,42	698,02	1396,03
					$\sum_{i=1}^k n_i (x_i - \bar{x})^2 = 8736,58$

$$D = \frac{1}{n} \sum_{i=1}^k n_i (x_i - \bar{x})^2 = \frac{1}{60} \sum_{i=1}^{30} n_i (x_i - \bar{x})^2 = \frac{1}{60} \cdot 8736,58 = 145,6.$$

Выборочное среднее квадратическое отклонение σ :

$$\sigma = \sqrt{D} = \sqrt{145,6} = 12,07.$$

2) Составим интервальный ряд, включающий 5 подинтервалов.

Выборка содержит $n = 60$ значений. Размах выборки $d = 50$. При пяти подинтервалах шаг разбиения $h = 10$.

Учтем, что левый край подинтервала принадлежит ему, а правый – открыт. Итак, вся выборка из 60 значений может быть разделена на следующие 5 равных по длине подинтервала:

$$32 \leq x < 42,$$

$$42 \leq x < 52,$$

$$52 \leq x < 62,$$

$$62 \leq x < 72,$$

$$72 \leq x < 82.$$

Расположив подинтервалы по возрастанию, сгруппируем имеющуюся информацию в интервальный статистический ряд, представленный в таблице 1.9.

Таблица 1.9

$[x_{i-1}, x_i)$	$[32,42)$	$[42,52)$	$[52,62)$	$[62,72)$	$[72,82)$
n_i	9	11	24	12	4
$\frac{n_i}{n}$	0,15	0,18	0,4	0,2	0,07

3) Вычислим выборочные числовые характеристики рассматриваемой величины X на основании построенного интервального статистического ряда с числом подинтервалов $k = 5$.

Следует учесть, что при задании выборки в виде интервального статистического ряда (в отличие от обычного статистического ряда) в формулах для вычисления среднего значения, дисперсии и среднего квадратиче-

скогого отклонения вместо x_i рассматривается среднее значение i -го подынтервала: $\bar{x}_i = \frac{x_{i-1} + x_i}{2}$.

Среднее значение выборки будет равно:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \bar{x}_i = \frac{1}{n} \sum_{i=1}^k n_i \frac{x_{i-1} + x_i}{2} = \frac{1}{60} \sum_{i=1}^5 n_i \frac{x_{i-1} + x_i}{2} = \frac{1}{60} \left(9 \cdot \frac{32+42}{2} + 11 \cdot \frac{42+52}{2} + 24 \cdot \frac{52+62}{2} + 12 \cdot \frac{62+72}{2} + 4 \cdot \frac{72+82}{2} \right) = \frac{1}{60} (333 + 517 + 1368 + 804 + 308) = 55,5.$$

Дисперсия выборки в случае задания ее в виде интервального ряда находится по формуле:

$$D = \frac{1}{n} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 = D = \frac{1}{n} \sum_{i=1}^k n_i \left(\frac{x_{i-1} + x_i}{2} - \bar{x} \right)^2.$$

Для нахождения дисперсии ($k = 5$) составим таблицу 1.10.

Таблица 1.10

№ подын-терва-ла	n_i	$\frac{x_{i-1} + x_i}{2}$	\bar{x}	$\frac{x_{i-1} + x_i}{2} - \bar{x}$	$(\frac{x_{i-1} + x_i}{2} - \bar{x})^2$	$n_i (\frac{x_{i-1} + x_i}{2} - \bar{x})^2$
1	9	37	55,5	-18,5	342,25	3080,25
2	11	47		-8,5	72,25	794,75
3	24	57		1,5	2,25	54
4	12	67		11,5	132,25	1587
5	4	77		21,5	462,25	1849
						$\sum_{i=1}^5 n_i (\frac{x_{i-1} + x_i}{2} - \bar{x})^2 = 7365$

$$D = \frac{1}{n} \sum_{i=1}^k n_i \left(\frac{x_{i-1} + x_i}{2} - \bar{x} \right)^2 = \frac{1}{60} \cdot \sum_{i=1}^5 n_i \left(\frac{x_{i-1} + x_i}{2} - 55,5 \right)^2 = \frac{1}{60} \cdot 7365 = 122,75.$$

Выборочное среднее квадратическое отклонение σ :

$$\sigma = \sqrt{D} = 11,08.$$

4) Построим интервальный статистический ряд, включающий 7 подынтервалов, вычислим на его основании выборочные числовые характеристики рассматриваемой величины X .

Выборка содержит 60 значений. Вспомним, что размах выборки $d = 50$. Это значит, что при 7 подынтервалах шаг разбиения $h = 7$. Итак, вся выборка из 60 данных может быть разделена на следующие 7 равных по длине подынтервалов:

$$32 \leq x < 39,$$

$$39 \leq x < 46,$$

$$46 \leq x < 53,$$

$$53 \leq x < 60,$$

$$60 \leq x < 67,$$

$$67 \leq x < 74,$$

$$74 \leq x < 81.$$

Расположив подынтервалы по возрастанию, сгруппируем имеющуюся информацию в интервальный статистический ряд ($k = 7$), представленный в таблице 1.11.

Таблица 1.11

$[x_{i-1}, x_i)$	[32,39)	[39,46)	[46,53)	[53,60)	[60,67)	[67,74)	[74,81)
n_i	5	6	12	14	10	9	4
$\frac{n_i}{n}$	0,08	0,1	0,2	0,23	0,17	0,15	0,07

Найдем числовые характеристики СВ X в условиях задания выборки в виде интервального ряда с числом подынтервалов $k = 7$ (формулы будут те же, что и при обработке интервального статистического ряда с числом подынтервалов $k = 5$).

Среднее значение выборки будет равно:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \bar{x}_i = \frac{1}{n} \sum_{i=1}^k n_i \frac{x_{i-1} + x_i}{2} = \frac{1}{60} \sum_{i=1}^7 n_i \frac{x_{i-1} + x_i}{2} = \frac{1}{60} \left(5 \cdot \frac{32+39}{2} + 6 \cdot \frac{39+46}{2} + 12 \cdot \frac{46+53}{2} + 14 \cdot \frac{53+60}{2} + 10 \cdot \frac{60+67}{2} + 9 \cdot \frac{67+74}{2} + 4 \cdot \frac{74+81}{2} \right) = \frac{1}{60} (177,5 + 255 + 594 + 791 + 635 + 6345 + 310) = 56,62$$

Дисперсия выборки в случае задания ее в виде интервального ряда находится по формуле:

$$D = \frac{1}{n} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2 = D = \frac{1}{n} \sum_{i=1}^k n_i \left(\frac{x_{i-1} + x_i}{2} - \bar{x} \right)^2$$

Для нахождения дисперсии ($k = 7$) составим таблицу 1.12.

Таблица 1.12

№ подын- терва- ла	n_i	$\frac{x_{i-1} + x_i}{2}$	\bar{x}	$\frac{x_{i-1} + x_i}{2} - \bar{x}$	$(\frac{x_{i-1} + x_i}{2} - \bar{x})^2$	$n_i (\frac{x_{i-1} + x_i}{2} - \bar{x})^2$
1	5	35,5	56,62	-21,12	446,05	2230,25
2	6	42,5		-14,12	199,37	1196,22
3	12	49,5		-7,12	50,69	608,28
4	14	56,5		-0,12	0,01	0,14
5	10	63,5		6,88	47,33	473,3

6	9	70,5		13,88	192,65	1733,85
7	4	77,5		20,88	435,97	1743,88
						$\sum_{i=1}^5 n_i \left(\frac{x_{i-1} + x_i}{2} - \bar{x} \right)^2 = 7985,92.$

$$D = \frac{1}{n} \sum_{i=1}^k n_i \left(\frac{x_{i-1} + x_i}{2} - \bar{x} \right)^2 = \frac{1}{60} \sum_{i=1}^7 n_i \left(\frac{x_{i-1} + x_i}{2} - 56,62 \right)^2 = \frac{1}{60} \cdot 7985,92 = 133,1.$$

Выборочное среднее квадратическое отклонение σ :

$$\sigma = \sqrt{D} = 11,54.$$

5) Систематизируем числовые характеристики случайной величины X, полученные для обычного статистического ряда (вариант I), для интервально-го статистического ряда с числом подынтервалов k = 5 (вариант II) и для ин-тервального статистического ряда с числом подынтервалов k = 7 (вариант III) (таблица 1.13).

Таблица 1.13

Сравнение числовых характеристик СВ X по трем вариантам вычислений

Вариант	Числовые характеристики		
	\bar{x}	D	σ
I	55,68	144,48	12,02
II	55,5	120,01	10,95
III	56,62	133,1	11,54

Вывод: значения числовых характеристик случайной величины зависят от способа представления выборки.

Задачи для самостоятельного решения

Задача 1-А. В Кузбассе в 2002 г. изучалась стоимость основных производственных фондов (ОПФ) малых предприятий. Результаты приведены в виде интервального ряда (таблица 1.14).

Таблица 1.14

Стоимость ОПФ, млн.руб.	14 - 16	16 - 18	18 - 20	20 - 22	22 - 24
Число пред-приятий	2	6	10	4	3

Задание

- Перейти от интервального ряда к дискретному.
- Вычислить среднюю стоимость ОПФ малых предприятий на момент рассмотрения.
- Найти дисперсию стоимости ОПФ.
- Найти среднее квадратическое отклонение изучаемой величины.
- Вычислить коэффициент вариации изучаемой величины.

Задача 1-В. На предприятии работают 100 рабочих. Их сменная выработка некоторых изделий позволяет выделить следующие интервалы и группы: 170 – 190 (10 рабочих); 190 – 210 (20 рабочих); 210 – 230 (50 рабочих); 230 – 250 (20 рабочих).

Задание

1. Представить информацию в виде интервального статистического ряда.
2. Вычислить среднесменную выработку рабочих.
3. Найти дисперсию, среднее квадратическое отклонение и коэффициент вариации сменной выработки рабочих.

Задача 1-С. Дан ряд распределения случайной величины X (таблица 1.15).

Таблица 1.15

x_i	1	4	5
p_i	0,4	0,1	0,5

Задание

1. Найти математическое ожидание $M(x)$, дисперсию $D(x)$ и среднее квадратическое (стандартное) отклонение σ случайной величины X.
2. Определить функцию распределения $F^*(x)$.

Задача 1-Д. Данна функция распределения случайной величины X:

$$F^*(x) = \begin{cases} 0 & \text{при } x < 1; \\ 0,3 & \text{при } 1 < x \leq 2; \\ 0,7 & \text{при } 2 < x \leq 3; \\ 1 & \text{при } x > 3. \end{cases}$$

Задание

1. Найти ряд распределения СВ X.
2. Найти математическое ожидание $M(x)$ и дисперсию $D(x)$.

Задача 1-Е. Случайная величина X, сосредоточенная на интервале $[-1;3]$, задана функцией распределения $F^*(x) = \frac{1}{4}x + \frac{1}{4}$.

Задание

Найти вероятность попадания СВ X в интервал $[0;2]$.

Раздел 2.

ПРОСТРАНСТВЕННЫЕ СТАТИСТИЧЕСКИЕ ВЫБОРКИ. ПАРНАЯ РЕГРЕССИЯ И КОРРЕЛЯЦИЯ

Пространственные (или перекрестные) статистические выборки получают при фиксации значений некоторых переменных величин в один и тот же момент времени на разных объектах пространства. Если при этом переменных величин две (одна из них – объясняемая или результат, другая – объясняющая или фактор), речь идет о парной регрессии.

Парная регрессия - уравнение связи двух переменных y и x :

$$y = \hat{f}(x),$$

где y — зависимая или объясняемая переменная (результативный признак);

x – независимая или объясняющая переменная (признак-фактор).

Различают линейные и нелинейные регрессии.

Линейная регрессия: $y = a + b \cdot x + \varepsilon$.

Нелинейные регрессии делятся на два класса: 1) регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам; 2) регрессии, нелинейные по оцениваемым параметрам.

Регрессии, нелинейные по объясняющим переменным:

- полиномы разных степеней, например, $y = a + b_1 \cdot x + b_2 \cdot x^2 + b_3 \cdot x^3 + \varepsilon$;
- равносторонняя гипербола.

Регрессии, нелинейные по оцениваемым параметрам:

- степенная $y = a \cdot x^b \cdot \varepsilon$,
- показательная $y = a \cdot b^x \cdot \varepsilon$,
- экспоненциальная $y = e^{a+b \cdot x} \cdot \varepsilon$.

Построение уравнения регрессии сводится к оценке ее параметров. Для оценки параметров регрессий, линейных по параметрам, используют метод наименьших квадратов (МНК).

Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b :

$$\begin{cases} na + b \sum x = \sum y, \\ a \sum x + b \sum x^2 = \sum yx. \end{cases}$$

Можно воспользоваться готовыми формулами, которые вытекают из этой системы:

$$a = \bar{y} - b \cdot \bar{x}, \quad b = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}.$$

Тесноту связи изучаемых явлений оценивает:

а) для линейной регрессии линейный коэффициент парной корреляции r_{xy} ($-1 \leq r_{xy} \leq 1$):

$$r_{xy} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{(\overline{y^2} - \bar{y}^2)(\overline{x^2} - \bar{x}^2)};$$

в) для нелинейной регрессии индекс корреляции ρ_{xy} ($0 \leq \rho_{xy} \leq 1$):

$$\rho_{xy} = \sqrt{1 - \frac{\sigma_{ocm}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sum (y - \hat{y}_x)^2}{\sum (y - \bar{y})^2}}.$$

Оценку качества построенной модели даст коэффициент (индекс) детерминации, а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации - среднее отклонение расчетных значений от фактических:

$$\bar{A} = \frac{1}{n} \sum \left| \frac{y - \hat{y}}{y} \right| \cdot 100\%.$$

Допустимый предел значений A - не более 8 – 10 %.

Средний коэффициент эластичности $\bar{\varepsilon}$ показывает, насколько процентов в среднем по совокупности изменится результат y от своей средней величины при изменении фактора x на 1% от своего среднего значения:

$$\bar{\varepsilon} = f'(x) \frac{\bar{x}}{\bar{y}}.$$

Задача дисперсионного анализа заключается в анализе дисперсии зависимой переменной:

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

где $\sum (y - \bar{y})^2$ - общая сумма квадратов отклонений;

$\sum (\hat{y}_x - \bar{y})^2$ - сумма квадратов отклонений, обусловленная регрессией (“объясненная” или “факторная”);

$\sum (y - \hat{y}_x)^2$ - остаточная сумма квадратов отклонений.

Долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака характеризует коэффициент (индекс) детерминации R^2 :

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

Коэффициент детерминации - квадрат коэффициента или индекса корреляции.

F-тест - оценивание качества уравнения регрессии - состоит в проверке гипотезы H_0 о статистической незначимости уравнения и показателя тесноты связи. Для этого выполняется сравнение фактического $F_{\text{факт}}$ и критического (табличного) $F_{\text{табл}}$ значений *F-критерия Фишера*. $F_{\text{факт}}$ определяется из соотношения значений факторной и остаточной дисперсий, рассчитанных на одну степень свободы:

$$F_{\text{факт}} = \frac{\sum (\hat{y} - \bar{y})^2 / m}{\sum (y - \hat{y})^2 / (n - m - 1)} = \frac{r_{xy}^2}{1 - r_{xy}^2} (n - 2),$$

где n - число единиц совокупности;

m - число параметров при переменных x .

$F_{\text{табл}}$ - это максимально возможное значение критерия под влиянием случайных факторов при данных степенях свободы и уровне значимости α . Уровень значимости α - вероятность отвергнуть правильную гипотезу при условии, что она верна. Обычно α принимается равной 0,05 или 0,01.

Если $F_{\text{табл}} < F_{\text{факт}}$, то H_0 - гипотеза о случайной природе оцениваемых характеристик отклоняется и признается их статистическая значимость и надежность. Если $F_{\text{табл}} > F_{\text{факт}}$, то гипотеза H_0 принимается и признается статистическая незначимость, ненадежность уравнения регрессии.

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются *t-критерий Стьюдента* и *доверительные интервалы* для каждого из показателей. Выдвигается гипотеза H_0 о случайной природе показателей, то есть о незначимом их отличии от нуля. Оценка значимости коэффициентов регрессии и корреляции с помощью *t-критерия Стьюдента* проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_b = \frac{b}{m_b}; \quad t_a = \frac{a}{m_a}; \quad t_r = \frac{r}{m_r}.$$

Случайные ошибки параметров линейной регрессии и коэффициента корреляции определяются по формулам:

$$m_a = \sqrt{\frac{\sum (y - \hat{y}_x)^2 / (n-2)}{\sum (x - \bar{x})^2}} = \sqrt{\frac{S_{\text{окн}}^2}{\sum (x - \bar{x})^2}} = \frac{S_{\text{окн}}}{\sigma_x \sqrt{n}},$$

$$m_b = \sqrt{\frac{\sum (y - \hat{y}_x)^2}{(n-2)} \cdot \frac{\sum x^2}{n \sum (x - \bar{x})^2}} = \sqrt{\frac{S_{\text{окн}}^2}{n-2} \frac{\sum x^2}{n \sigma_x^2}} = S_{\text{окн}} \frac{\sqrt{\sum x^2}}{n \sigma_x},$$

$$m_{r_{xy}} = \sqrt{\frac{1 - r_{xy}^2}{n-2}}.$$

Сравнивая фактическое и критическое (табличное) значения t -статистики, принимаем или отвергаем гипотезу H_0 .

Связь между F -критерием Фишера и t -статистикой Стьюдента выражается равенством:

$$t_r^2 = t_h^2 = \sqrt{F}.$$

Если $t_{\text{табл}} < t_{\text{факт}}$, то H_0 отклоняется, т.е. a , b и r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x . Если $t_{\text{табл}} > t_{\text{факт}}$, то гипотеза H_0 не отклоняется и признается случайная природы формирования a , b или r_{xy} .

Для расчета доверительного интервала определяем предельную ошибку Δ для каждого показателя:

$$\Delta_a = t_{\text{табл}} m_a; \quad \Delta_b = t_{\text{табл}} m_b.$$

Формулы для расчета доверительных интервалов имеют следующий вид:

$$\gamma_a = a \pm \Delta_a; \quad \gamma_{a_{\min}} = a - \Delta_a; \quad \gamma_{a_{\max}} = a + \Delta_a;$$

$$\gamma_b = b \pm \Delta_b; \quad \gamma_{b_{\min}} = b - \Delta_b; \quad \gamma_{b_{\max}} = b + \Delta_b.$$

Если в границы доверительного интервала попадает ноль, то есть нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

Прогнозное значение y_p определяется путем подстановки в уравнение регрессии $\hat{y}_x = a + b \cdot x$ соответствующего (прогнозного) значения x_p . Вычисляется средняя стандартная ошибка прогноза $m_{\hat{y}_p}$:

$$m_{\hat{y}_p} = \sigma_{\text{окн}} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x - \bar{x})^2}},$$

где $\sigma_{\text{окн}} = \sqrt{\frac{\sum (y - \hat{y})^2}{n-m-1}}$. Затем строится доверительный интервал прогноза:

$$\hat{y}_p = \hat{y}_p \pm \Delta_{\hat{y}_p}; \quad \hat{y}_{\hat{y}_{p_{\min}}} = \hat{y}_p - \Delta_{\hat{y}_p}; \quad \hat{y}_{\hat{y}_{p_{\max}}} = \hat{y}_p + \Delta_{\hat{y}_p},$$

где $\Delta_{\hat{y}_p} = t_{\text{мод}} \cdot m_{\hat{y}_p}$.

Примеры решения избранных задач

Задача 2.1. В городах Кузбасса в 2002 г. изучалась зависимость между среднедушевым прожиточным минимумом в день одного трудоспособного (Х, руб.) и среднедневной заработной платой (Y, руб.). Для обработки были выбраны данные по наиболее крупным населенным пунктам (таблица 2.1).

Таблица 2.1

Город	Y	X
Кемерово	112	148
Новокузнецк	115	165
Междуреченск	87	134
Осинники	73	152
Прокопьевск	89	132
Мыски	67	139
Белово	76	121

Задание

- Составить статистический ряд на основании предложенной выборки данных.
- Вычислить числовые характеристики статистического ряда.
- Построить линейное уравнение парной регрессии у от x.
- Рассчитать линейный коэффициент парной корреляции.
- Проверить гипотезу о статистической значимости коэффициента корреляции.
- Оценить статистическую значимость параметров регрессии.
- Оценить с помощью коэффициента детерминации R^2 качество регрессионного уравнения в целом.
- Сделать вывод о возможности применения построенного уравнения регрессии для прогнозирования ежедневной заработной платы при известном прогнозном значении среднедушевого ежедневного прожиточного минимума.

Решение

- Составляя статистический ряд, расположим данные по неубыва-нию объясняющей переменной X (таблица 2.2).

Таблица 2.2

№	1	2	3	4	5	6	7
Название города	Белово	Прокопьевск	Осинники	Мыски	Кемерово	Междуреченск	Новокузнецк
x	121	132	134	139	148	152	165
y	76	89	87	67	112	73	115

2. Найдем числовые характеристики выборки.

Средняя величина (математическое ожидание) величины X находится по всей совокупности значений этой переменной величины (повторяющиеся значения отсутствуют, оснований для выделения групп нет):

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \text{ где объем совокупности } n = 7 \text{ (i = 1, 2, 3, 4, 5, 6, 7).}$$

Получаем:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{7} \sum_{i=1}^7 x_i = \frac{1}{7} (121 + 132 + 134 + 139 + 148 + 152 + 165) = \frac{991}{7} = 141,57.$$

Средняя величина (математическое ожидание) переменной величины Y находится также по всей совокупности значений переменной величины Y:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{7} (76 + 89 + 87 + 67 + 112 + 73 + 115) = \frac{619}{7} = 88,43.$$

Для вычисления дисперсий переменных величин X и Y составим таблицу 2.3.

Таблица 2.3

x _i	y _i	\bar{x}	\bar{y}	(x _i - \bar{x})	(y _i - \bar{y})	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$
121	76	141,57	88,43	-20,57	-12,43	423,12	154,50
132	89			-9,57	0,57	91,58	0,32
134	87			-7,57	-1,43	57,30	2,04
139	67			-2,57	-21,43	6,60	459,24
148	112			6,43	23,57	41,34	555,54
152	73			10,43	-15,43	108,78	238,08
165	115			23,43	26,57	548,96	706
				$\sum_{i=1}^7 (x_i - \bar{x})^2 = 1277,68$		$\sum_{i=1}^7 (y_i - \bar{y})^2 = 2115,72$	

$$\text{Дисперсия } D_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = D_x = \frac{1}{7} \sum_{i=1}^7 (x_i - 141,57)^2 = \frac{1}{7} \cdot 1277,68 = 182,52.$$

$$\text{Дисперсия } D_y = \frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = D_y = \frac{1}{7} \sum_{i=1}^7 (y_i - 88,43)^2 = \frac{1}{7} \cdot 2115,72 = 302,24.$$

Выборочное среднее квадратическое отклонение $\sigma = \sqrt{D}$.

Для переменной величины X:

$$\sigma_x = \sqrt{182,52} = 13,51.$$

Для переменной величины Y :

$$\sigma_y = \sqrt{302,24} = 17,38.$$

3. Построим уравнение парной линейной регрессии y от x . Для этого рассчитаем его параметры b и a : $b = \frac{\bar{y} \cdot \bar{x} - \bar{y} \cdot \bar{x}}{\bar{x}^2 - \bar{x}^2}$ и $a = \bar{y} - b \cdot \bar{x}$. Предварительно необходимо вычислить $\bar{x} \cdot y$, $\bar{y} \cdot \bar{x}$, x_i^2 , \bar{x}^2 .

Строим расчетную таблицу 2.4 (первые четыре колонки скопируем из таблицы 2.3).

Таблица 2.4

x_i	y_i	\bar{x}	\bar{y}	$x_i \cdot y_i$	$\bar{x} \cdot y_i$	x_i^2	\bar{x}^2	\bar{x}^2
121	76	141,57	88,43	9196	12651,71	14641	20225	20042,06
132	89			11748		17424		
134	87			11658		17956		
139	67			9313		19321		
148	112			16576		21904		
152	73			11096		23104		
165	115			18975		27225		

Получаем:

$$b = \frac{\bar{y} \cdot \bar{x} - \bar{y} \cdot \bar{x}}{\bar{x}^2 - \bar{x}^2} = \frac{12651,71 - 141,57 \cdot 88,43}{20225 - 20042,06} = \frac{12651,71 - 12519,03}{182,94} = \frac{132,68}{182,94} = 0,72.$$

$$a = \bar{y} - b \bar{x} = 88,43 - 0,72 \cdot 141,57 = 88,43 - 101,93 = -13,5.$$

Искомое уравнение регрессии принимает вид:

$$\hat{y} = -13,5 + 0,72 \cdot x,$$

где \hat{y} - модельные значения объясняемой переменной.

Вывод: найденное значение параметра b означает, что с увеличением среднедушевого прожиточного минимума на 1 руб. среднедневная заработная плата увеличивается на 0,72 руб.

4. Линейный коэффициент парной корреляции рассчитывается по формуле:

$$r_{xy} = \frac{\bar{y} \cdot \bar{x} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y} = \frac{12651,71 - 12519,03}{13,51 \cdot 17,38} = \frac{132,68}{234,8} = 0,56.$$

Вывод: поскольку для идеальной линейной зависимости $|r_{xy}| = 1$, а в нашем случае $|r_{xy}| = 0,56$, то это означает: если линейная зависимость между рассматриваемыми переменными существует, то выражена она крайне слабо. Необходима дополнительная проверка построенного уравнения регрессии.

5. Осуществим статистическую проверку значимости найденного коэффициента корреляции в соответствии со следующей гипотезой:

$$H_0: r_{xy} = 0,$$

$$H_1: r_{xy} \neq 0.$$

Наблюдаемая статистика Стьюдента для r_{xy} находится по формуле:

$$T_{\text{набл}} = \frac{r_{xy} \cdot \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}.$$

$$\text{В нашем случае } T_{\text{набл}} = \frac{0,56 \cdot \sqrt{(7-2)}}{\sqrt{(1-0,56^2)}} = \frac{0,56 \cdot 2,24}{0,83} = \frac{1,25}{0,83} = 1,5.$$

Для уровня значимости 0,05 и числа степеней свободы $n-2 = 5$ по таблице Стьюдента находим критическое значение статистики:

$$T_{\text{крит}} = 2,571.$$

Поскольку $T_{\text{набл}} < T_{\text{крит}}$, нулевая гипотеза принимается и коэффициент корреляции необходимо считать незначимым.

6. Осуществим статистическую проверку значимости найденных параметров регрессии.

Для параметра b :

$$H_0: b = 0,$$

$$H_1: b \neq 0.$$

Гипотеза в такой постановке обычно называется гипотезой о статистической значимости вычисленного параметра.

Если H_0 принимается, то говорят, что параметр b статистически незначим. При отклонении H_0 параметр b считается статистически значимым, что указывает на наличие определенной линейной зависимости между объясняющей и объясняемой переменными.

Поскольку предполагается, что $\beta = 0$ (из того, что полагается $b = 0$), то значимость параметра регрессии b проверяется по формуле:

$$T = \frac{|b - \beta|}{S_b} = \frac{|b|}{S_b} = \frac{|b|}{\sqrt{S_b^2}}.$$

то есть с помощью анализа отношения величины параметра b к его стандартной ошибке (корню квадратному из его необъясненной дисперсии).

Необъясненная дисперсия параметра b в общем виде записывается так:

$$S_b^2 = \frac{\sum e_i^2}{n(n-2)(x^2 - \bar{x}^2)} = \frac{\sum (y_i - a - bx_i)^2}{n(n-2)(x^2 - \bar{x}^2)}.$$

По аналогичной схеме на основе T -статистики проверяется гипотеза о статистической значимости параметра a :

$$T = \frac{|a - \bar{a}|}{S_a} = \frac{|a|}{\sqrt{S_a^2}},$$

где $S_a^2 = \bar{x}^2 S_b^2$.

Отметим, что более важным является анализ статистической значимости параметра b , так как именно в нем скрыто влияние объясняющей переменной на зависимую переменную.

Многие необходимые промежуточные вычисления для нахождения стандартной ошибки параметра b находятся в таблице 2.4. Для получения недостающих промежуточных вычислений подготовим таблицу 2.5.

Таблица 2.5

x_i	y_i	$\hat{y}_i = a + b x$	$e_i = y_i - \hat{y}_i$	$e_i^2 = (y_i - \hat{y}_i)^2$
121	76	73,62	2,38	5,66
132	89	81,54	7,46	55,65
134	87	82,98	4,02	16,16
139	67	86,58	-19,58	383,38
148	112	93,06	18,94	358,72
152	73	95,94	-22,94	526,24
165	115	105,3	9,7	94,09
				$\sum_{i=1}^7 e_i^2 = 1439,9$

Необъясненная дисперсия коэффициента b будет равна:

$$S_b^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n(n-2)(\bar{x}^2 - \bar{x}^2)} = \frac{\sum_{i=1}^7 (y_i - \hat{y}_i)^2}{7(7-2)(20225 - 20042,06)} = \frac{1439,9}{35 \cdot 182,94} = \frac{1439,9}{6402,9} = 0,22.$$

Стандартная ошибка параметра b будет равна:

$$S_b = \sqrt{S_b^2} = \sqrt{0,22} = 0,47.$$

Тогда наблюдаемая статистика для b будет равна:

$$T_{\text{набл}} = \frac{b}{S_b} = \frac{0,72}{0,47} = 1,53.$$

По таблице распределения Стьюдента при уровне значимости $\alpha = 0,05$ и числе степеней свободы $v = n - 2 = 5$ находим соответствующее критическое значение T -статистики для параметра b :

$$T_{kp} = T_{\frac{\alpha}{2}, n-2} = T_{0,025, 5} = 2,571.$$

Видим, что для параметра b : $T_{\text{набл}} = 1,53 < T_{kp} = 2,571$.

Вывод: гипотеза H_0 принимается и параметр b статистически незначим; между переменными Y и X скорее всего существует зависимость, от-

личная от линейной. В подобных случаях проверять значимость параметра a необязательно.

7. Для парной регрессии коэффициент детерминации $R^2 = r_{xy}^2$. В нашем случае $R^2 = (0,56)^2 = 0,31$.

Известно, что чем теснее линейная связь между переменными величинами, тем ближе коэффициент детерминации к единице. При оценивании регрессий по временным рядам объемных показателей величина R^2 действительно может быть весьма близкой к единице.

Если же уравнение регрессии строится по перекрестным данным, а не по временным рядам (в решаемой задаче именно такой является выборка переменных величин), то коэффициент детерминации обычно не превышает $0,6 - 0,7$. В нашем случае тот факт, что он равен $0,31$, означает, что только 31% вариации заработной платы Y объясняется вариацией фактора X , что явно недостаточно и не позволяет принять построенное уравнение регрессии.

8. Поскольку предложенная спецификация регрессионной модели оказалась ошибочной, использовать ее для прогнозирования не имеет смысла.

Задача 2.2. По семи территориям Кемеровской области за 2002 г. известны значения двух признаков (таблица 2.6).

Таблица 2.6

Территория (номер по списку)	y_i	x_i
1	71,9	50,2
2	64,3	64,1
3	63,0	62,3
4	59,8	65,9
5	58,1	63,9
6	57,4	52,3
7	52,4	60,3

Задание

1. Для характеристики зависимости Y от X рассчитать параметры следующих функций:

- а) линейной;
- б) степенной;
- в) показательной;
- г) равносторонней гиперболы.

2. Вычислить для каждой модели коэффициенты корреляции и детерминации; оценить каждую модель через среднюю ошибку аппроксимации и F-критерий Фишера.

Решение

1-a) Линейная функция. Для расчета параметров линейной регрессии $\hat{y} = a + bx$ по выборке объемом n необходимо решить систему линейных уравнений относительно a и b :

$$\begin{cases} n \cdot a + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x^2 = \sum_{i=1}^n xy. \end{cases}$$

В результате получаем уравнения:

$$b = \frac{\bar{y} \cdot \bar{x} - \bar{y} \cdot \bar{x}}{\bar{x}^2 - \bar{x}^2} \quad \text{и} \quad a = \bar{y} - b \cdot \bar{x}.$$

Выполним промежуточные вычисления, необходимые для нахождения параметров регрессии и дальнейшего решения задачи (вычисления коэффициентов корреляции, детерминации, средней ошибки аппроксимации) (таблица 2.7).

Таблица 2.7

	y_i	x_i	$y_i \cdot x_i$	x_i^2	y_i^2	\bar{y}	$y_i - \bar{y}$	$A_i = \left \frac{y_i - \bar{y}}{y_i} \right \cdot 100\%$
1	71,9	50,2	3609,38	2520,04	5169,61	57,50	14,4	20,03
2	64,3	64,1	4121,63	4108,81	4134,49	62,51	1,79	2,78
3	63,0	62,3	3924,90	3881,29	3969	59,37	3,63	5,76
4	59,8	65,9	3940,82	4342,81	3576,04	63,15	-3,35	5,60
5	58,1	63,9	3712,59	4083,21	3375,61	62,43	-4,33	7,45
6	57,4	52,3	3002,02	2735,29	3294,76	58,26	-0,86	1,50
7	52,4	60,3	3159,72	3636,09	2745,76	61,14	-8,74	16,68
Σ	426,9	419	25471,06	25307,54	26265,27			59,8
Средние	$\bar{y}_i = 60,98$	$\bar{x}_i = 59,86$	$\bar{y}_i \cdot \bar{x}_i = 3638,72$	$\bar{x}_i^2 = 3615,36$	$\bar{y}_i^2 = 3752,18$			$\bar{A} = 8,54\%$

В результате получаем: 21147,356

$$b = \frac{y_i \cdot x_i - \bar{y}_i \cdot \bar{x}_i}{x_i^2 - \bar{x}^2} = \frac{3638,72 - 60,98 \cdot 59,86}{3615,36 - 59,86^2} = \frac{-11,54}{-32,14} = 0,36.$$

$$a = \bar{y} - b \cdot \bar{x} = 60,98 - 0,36 \cdot 59,86 = 39,43.$$

Уравнение регрессии имеет вид:

$$\hat{y} = 39,43 + 0,36 \cdot x.$$

Вывод: с увеличением среднедневной заработной платы на 1 руб. доля расходов на покупку продовольственных товаров увеличивается в среднем на 0,36%-ных пункта.

Рассчитаем линейный коэффициент парной регрессии по формуле:

$$r_{xy} = \frac{\bar{y} \cdot \bar{x} - \bar{y} \cdot \bar{x}}{\sigma_x \cdot \sigma_y} \Rightarrow r_{xy} = b \frac{\sigma_x}{\sigma_y},$$

где $\sigma_x = \sqrt{\bar{x}^2 - \bar{x}^2}$, $\sigma_y = \sqrt{\bar{y}^2 - \bar{y}^2}$.

Воспользуемся данными из таблицы 2.7 и получим:

$$\sigma_x = \sqrt{\bar{x}^2 - \bar{x}^2} = \sqrt{3615,36 - 3583,22} = \sqrt{32,14} = 5,67,$$

$$\sigma_y = \sqrt{\bar{y}^2 - \bar{y}^2} = \sqrt{3752,18 - 60,98^2} = \sqrt{3752,18 - 3718,56} = \sqrt{33,62} = 5,8.$$

$$\text{Следовательно, } r_{xy} = b \frac{\sigma_x}{\sigma_y} = \frac{0,36 \cdot 5,67}{5,8} = 0,35.$$

Коэффициент детерминации R^2 для парной регрессии равен квадрату коэффициента корреляции:

$$R^2 = r_{xy}^2 = (0,35)^2 = 0,12.$$

Поскольку полученное значение коэффициента детерминации близко к нулю, линейная связь между переменными величинами очень слабая. Только 12% вариаций результата объясняется вариацией фактора x .

Ошибка аппроксимации в каждом отдельном случае определяется модулем отношения отклонения расчетного значения от фактического к фактическому значению и измеряется в процентах, то есть:

$$A_i = \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%.$$

Средняя ошибка аппроксимации находится по формуле:

$$\bar{A} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \cdot 100\%.$$

В таблице 2.7 получено значение средней ошибки аппроксимации $\bar{A} = 8,54\%$ (допустимый предел составляет 8 – 10 %), то есть в среднем расчетные значения отклоняются от фактических на 8,54%.

И, наконец, осуществим анализ построенной линейной модели с помощью F-статистики (критерия Фишера). Задача эта связана с анализом дисперсии переменной величины Y :

$$\sum (y_i - \bar{y})^2 = \sum k_i^2 + \sum e_i^2.$$

Очевидно, что $\sum (y_i - \bar{y})^2$ – общая (полная) сумма квадратов отклонений наблюдаемых точек от среднего значения переменной величины (она может интерпретироваться как мера общего разброса (рассеивания) переменной Y относительно \bar{y}); $\sum k_i^2 = \sum (\hat{y}_i - \bar{y})^2$ – объясненная сумма квадра-

тов, интерпретируемая как мера разброса, объяснимого с помощью регрессии; $\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$ - остаточная (необъясненная) сумма квадратов, являющаяся мерой остаточного, не объясненного уравнением регрессии разброса (разброса точек вокруг линии регрессии).

Долю дисперсии, объясняемую регрессией, в общей дисперсии результативного признака Y характеризует коэффициент детерминации R^2 . Таким образом, для оценки качества построенной линейной регрессии осуществим анализ статистической значимости найденного выше коэффициента детерминации. Поскольку всегда $0 \leq R^2 \leq 1$, проверим гипотезу:

$$H_0: R^2 = 0,$$

$$H_1: R^2 > 0.$$

Для проверки используем F-статистику Фишера:

$$F = \frac{\sum k_i^2/m}{\sum e_i^2/(n-m-1)} = \frac{\sum (\hat{y}_i - \bar{y})^2/m}{\sum (y_i - \hat{y}_i)^2/(n-m-1)},$$

где m – число объясняющих переменных (в нашей задаче $m = 1$).

В данной задаче уравнение F-статистики легко приводится к виду:

$$F = \frac{R^2}{1-R^2} (n-2).$$

Получаем значение $F_{\text{набл}}$:

$$F_{\text{набл}} = \frac{0,12}{1-0,12} \cdot 5 = 0,68.$$

Соответствующее критическое значение статистики Фишера $F_{kp} = F_{\alpha; m; n-m-1}$ (в некоторых справочных материалах $m = k_1$ или λ_1 ; $m-n-1 = k_2$ или λ_2). Находим по таблице:

$$F_{kp} = F_{0,05; 1; 5} = 6,61.$$

Так как $F_{\text{набл}} = 0,68 < F_{kp} = 6,61$, то при уровне значимости $\alpha = 0,05$ нулевая гипотеза принимается и делается вывод: параметры построенной линейной регрессии статистически незначимы. Причина этого, скорее всего, в невысокой тесноте выявленной зависимости и небольшой по объему выборке данных.

1-б) Степенная функция. Степенную модель будем строить в виде $y = a \cdot x^b$.

Осуществим линеаризацию функции путем логарифмирования обеих частей уравнения и дальнейшей замены переменных:

$$\lg y = \lg a + b \cdot \lg x,$$

$$Y = C + b \cdot X,$$

где $Y = \lg y$, $X = \lg x$, $C = \lg a$.

Чтобы вычислить параметры полученного в результате замены переменных линейного уравнения, выполним промежуточные вычисления (таблица 2.8).

Таблица 2.8

	y_i	x_i	Y_i	X_i	$Y_i X_i$	Y_i^2	X_i^2
1	71,9	50,2	1,86	1,70	3,16	3,46	2,89
2	64,3	64,1	1,81	1,81	3,28	3,28	3,28
3	63,0	62,3	1,80	1,79	3,22	3,24	3,20
4	59,8	65,9	1,78	1,82	3,24	3,17	3,31
5	58,1	63,9	1,76	1,80	3,17	3,10	3,24
6	57,4	52,3	1,76	1,72	3,03	3,10	2,96
7	52,4	60,3	1,72	1,78	3,06	2,96	3,17
Σ	426,9	419	12,49	12,42	22,16	22,31	22,05
Средние	$\bar{y}_i = 60,98$	$\bar{x}_i = 59,86$	$\bar{Y}_i = 1,78$	$\bar{X}_i = 1,77$	$\bar{Y}_i \bar{X}_i = 3,16$	$\bar{Y}_i^2 = 3,19$	$\bar{X}_i^2 = 3,15$

Вычислим параметры b и C по формулам:

$$b = \frac{\bar{Y}_i \cdot \bar{X}_i - \bar{Y}_i \cdot \bar{X}_i}{\bar{X}_i^2 - \bar{X}_i^2} \quad \text{и} \quad C = \bar{Y}_i - b \cdot \bar{X}_i.$$

Получаем:

$$b = \frac{3,16 - 1,78 \cdot 1,77}{3,15 - 1,77^2} = \frac{0,0094}{0,017} = 0,55; \quad C = 1,78 - 0,55 \cdot 1,77 = 0,81.$$

В результате получаем линеаризованное уравнение:

$$\hat{Y} = 0,81 + 0,55X.$$

Выполним его потенцирование, то есть вернемся к реальным переменным величинам, и получим степенную модель:

$$\hat{y} = 10^{0,81} \cdot x^{0,55} = 6,46 \cdot x^{0,55}.$$

Подготовим промежуточные вычисления (таблица 2.9).

Таблица 2.9

y_i	$y_i - \bar{y}_i$	$(y_i - \bar{y}_i)^2$	x_i	$x_i^{0,55}$	$\hat{y}_i = 6,46 \cdot x_i^{0,55}$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$A_i = \left \frac{y_i - \hat{y}_i}{y_i} \right \cdot 100\%$
71,9	10,92	119,25	50,2	8,62	55,68	16,22	263,09	23,59
64,3	3,32	11,02	64,1	9,86	63,69	0,61	0,37	3,33
63,0	2,02	4,08	62,3	9,70	62,66	0,34	0,11	2,81
59,8	-1,18	1,39	65,9	10,01	64,66	-4,86	23,62	5,37
58,1	-2,88	8,29	63,9	9,84	63,57	-5,47	29,92	6,71
57,4	-3,58	12,82	52,3	8,81	56,91	0,49	0,22	2,26
52,4	-8,58	73,62	60,3	9,53	61,56	-9,16	83,90	14,92
426,9		230,47	419				401,23	58,99
$\bar{y}_i = 60,98$			$\bar{x}_i = 59,86$					$\bar{A} = 8,43$

Для оценки тесноты связи изучаемых явлений в случае нелинейной

$$\text{регрессии вычисляется индекс корреляции } \rho_{xy} = \sqrt{1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}.$$

$$\text{В нашем случае имеем: } \rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^7 (y_i - \bar{y})^2}{\sum_{i=1}^7 (y_i - \bar{y})^2}} = \sqrt{1 - \frac{401,23}{230,47}}.$$

Учитывая, что выражение под корнем отрицательное, делаем вывод: оценить тесноту связи между переменными величинами не удается.

Средняя ошибка аппроксимации, как видно из таблицы 4.9, равна 8,43%.

1-в) *Показательная функция.* Показательную модель получим в виде $y = a \cdot b^x$.

Точно так же, как и в предыдущем случае со степенной моделью, проведем процедуру линеаризации функции путем ее логарифмирования и дальнейшей замены переменных:

$$\lg y = \lg a + \lg b \cdot x \Rightarrow Y = C + B \cdot x.$$

Для дальнейших расчетов подготовим таблицу 2.10.

Таблица 2.10

	y_i	Y_i	x_i	$Y_i \cdot x_i$	Y_i^2	x_i^2
1	71,9	1,86	50,2	93,37	3,46	2520,04
2	64,3	1,81	64,1	116,02	3,28	4108,81
3	63,0	1,80	62,3	112,14	3,24	3881,29
4	59,8	1,78	65,9	117,30	3,17	4342,81
5	58,1	1,76	63,9	112,46	3,10	4083,21
6	57,4	1,76	52,3	92,05	3,10	2735,29
7	52,4	1,72	60,3	103,72	2,96	3636,09
Σ	426,9	12,49	419	747,06	22,31	25307,5
Сред ние	$\bar{y}_i =$ 60,98	$\bar{Y}_i =$ 1,78	$\bar{x}_i =$ 59,86	$\bar{Y}_i \cdot \bar{x}_i =$ 106,72	$\bar{Y}_i^2 =$ 3,19	$\bar{x}_i^2 =$ 3615,36

Вычислим параметры В и С по формулам:

$$B = \frac{\bar{Y}_i \cdot \bar{x}_i - \bar{Y}_i \cdot \bar{x}_i}{\bar{x}_i^2 - \bar{x}_i^2} \quad \text{и} \quad C = \bar{Y}_i - B \cdot \bar{x}_i.$$

Получаем:

$$B = \frac{106,72 - 1,78 \cdot 59,86}{3615,36 - 59,86^2} = \frac{0,17}{32,14} = 0,005; \quad C = 1,78 - 0,005 \cdot 59,86 = 1,48.$$

В результате получаем линеаризованное уравнение:

$$\hat{Y} = 1,48 + 0,005 X.$$

Осуществим потенцирование полученного уравнения и запишем его в обычном виде:

$$\hat{y} = 10^{1,48} \cdot 10^{0,005x} = 30,20 \cdot 1,01^x.$$

Подготовим промежуточные вычисления для оценки тесноты связи рассматриваемых переменных величин и показательной модели регрессии в целом (таблица 2.11).

Таблица 2.11

	y_i	x_i	$1,01^{x_i}$	$\hat{y}_i = 30,20 \cdot 1,01^{x_i}$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$A_i = \left \frac{y_i - \hat{y}_i}{y_i} \right \cdot 100\%$
1	71,9	50,2	1,65	49,83	22,07	487,08	30,70
2	64,3	64,1	1,89	57,38	6,92	47,89	10,76
3	63,0	62,3	1,86	56,17	6,83	46,65	10,84
4	59,8	65,9	1,93	58,29	1,51	2,28	2,52
5	58,1	63,9	1,89	57,08	1,02	1,04	1,75
6	57,4	52,3	1,68	50,74	6,66	44,35	11,60
7	52,4	60,3	1,82	54,96	-2,56	6,55	4,88
\sum	426,9	419				635,84	73,05
Средние	$\bar{y}_i = 60,98$	$\bar{x}_i = 59,86$					$\bar{A} = 10,43$

Для оценки тесноты связи изучаемых явлений в случае нелинейной

регрессии вычисляется индекс корреляции $\rho_{xy} = \sqrt{1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}$.

В нашем случае имеем: $\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^7 (y_i - \hat{y}_i)^2}{\sum_{i=1}^7 (y_i - \bar{y})^2}} = \sqrt{1 - \frac{635,84}{230,47}}.$

Выражение под корнем отрицательное, что делает невозможным оценить количественно связь между изучаемыми переменными величинами в контексте показательной модели (ясно, что она практически отсутствует).

Средняя ошибка аппроксимации, как видно из таблицы 4.11, равна 10,43% (при допустимом пределе 8 – 10%).

Сравнивая полученные характеристики показательной модели с характеристиками линейной и степенной моделей, видим, что она еще хуже описывает взаимосвязь рассматриваемых переменных величин.

1-г) Гипербола. Уравнение равносторонней гиперболы будем получать в виде

$$y = a + b \cdot \frac{1}{x}.$$

Линеаризацию проведем путем замены переменной $z = \frac{1}{x}$, после чего получим уравнение:

$$y = a + b \cdot z.$$

Для расчетов параметров линеаризованного уравнения заполним таблицу 2.12.

Таблица 2.12

	y_i	x_i	z_i	$y_i z_i$	z_i^2	y_i^2
1	71,9	50,2	0,0199	1,4308	0,0004	5169,61
2	64,3	64,1	0,0156	1,0031	0,0002	4134,49
3	63,0	62,3	0,0160	1,008	0,0002	3969
4	59,8	65,9	0,0152	0,9090	0,0002	3576,04
5	58,1	63,9	0,0156	0,9064	0,0002	3375,61
6	57,4	52,3	0,0191	1,0963	0,0004	3294,76
7	52,4	60,3	0,0166	0,8698	0,0003	2745,76
Σ	426,9	419	0,0168	7,2234	0,0019	26265,27
Средние	$\bar{y}_i = 60,98$	$\bar{x}_i = 59,86$	$\bar{z}_i = 0,0168$	$\bar{y}_i z_i = 1,0319$	$\bar{z}_i^2 = 0,0003$	$\bar{y}_i^2 = 3752,18$

Значения параметров регрессии a и b найдем по формулам:

$$b = \frac{\sum y_i z_i - \bar{y}_i \cdot \bar{z}_i}{\sum z_i^2 - \bar{z}_i^2} \quad \text{и} \quad a = \bar{y}_i - b \cdot \bar{z}_i.$$

Получаем:

$$b = \frac{1,0319 - 60,98 \cdot 0,0168}{0,0003 - 0,0168^2} = \frac{0,0074}{0,000018} = 411; \quad a = 60,98 - 411 \cdot 0,0168 = 54,07.$$

В результате получаем уравнение регрессии:

$$\hat{y} = 54,07 + 411 \cdot \frac{1}{x}.$$

Подготовим промежуточные вычисления (таблица 2.13).

Таблица 2.13

y_i	$y_i - \bar{y}_i$	$(y_i - \bar{y}_i)^2$	x_i	$\frac{1}{x_i}$	$\hat{y}_i = 54,07 + 411 \cdot \frac{1}{x_i}$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$A_i = \left \frac{y_i - \hat{y}_i}{y_i} \right \cdot 100\%$
71,9	10,92	119,25	50,2	0,0199	62,25	9,65	93,12	13,42
64,3	3,32	11,02	64,1	0,0156	60,48	3,82	14,59	5,94
63,0	2,02	4,08	62,3	0,0160	60,65	2,35	5,52	3,73
59,8	-1,18	1,39	65,9	0,0152	60,32	-0,52	0,27	0,87
58,1	-2,88	8,29	63,9	0,0156	60,48	-2,38	5,66	4,10

57,4	-3,58	12,82	52,3	0,0191	61,92	-4,52	20,43	7,87
52,4	-8,58	73,62	60,3	0,0166	60,89	-8,49	72,08	16,14
426,9		230,47	419	0,118			211,67	52,07
$\bar{y}_i =$ 60,98		$\bar{x}_i =$ 59,86	$\bar{z}_i =$ 0,0168					$A = 7,44$

Для оценки тесноты связи изучаемых явлений в случае нелинейной

$$\text{регрессии вычисляется индекс корреляции } \rho_{xy} = \sqrt{1 - \frac{\sigma_{\text{ост}}^2}{\sigma_y^2}} = \sqrt{1 - \frac{\sum_{i=1}^n (y_i - \bar{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

$$\text{В нашем случае имеем: } \rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^7 (y_i - \bar{y})^2}{\sum_{i=1}^7 (y_i - \bar{y})^2}} = \sqrt{1 - \frac{211,67}{230,47}} = \sqrt{0,08} = 0,28.$$

Полученное значение индекса корреляции говорит о том, что связь между переменными величинами Y и X умеренная, прямая (знак "+").

Средняя ошибка аппроксимации, как видно из таблицы 4.13, равна 7,4%.

Для проверки качества построенной регрессии используем F-статистику Фишера:

$$F = \frac{\sum_{i=1}^m (\hat{y}_i - \bar{y})^2 / m}{\sum_{i=1}^{m-1} (\hat{y}_i - \bar{y}_i)^2 / (n-m-1)} = \frac{\sum_{i=1}^7 (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^7 (\hat{y}_i - \bar{y}_i)^2} \cdot (n-2),$$

где m – число объясняющих переменных (в нашей задаче m = 1), n – объем выборки.

Для вычисления F-статистики необходимо найти $\sum_{i=1}^7 (\hat{y}_i - \bar{y})^2$ – сумму квадратов отклонений, обусловленных регрессией. Составим таблицу 2.14.

Таблица 2.14

Промежуточные вычисления для отыскания F-статистики к задаче 4.2-2

	\hat{y}_i	\bar{y}_i	$(\hat{y}_i - \bar{y}_i)$	$(\hat{y}_i - \bar{y}_i)^2$
1	62,25	60,98	1,27	1,61
2	60,48		-0,5	0,25
3	60,65		-0,33	0,11
4	60,32		-0,66	0,43
5	60,48		-0,5	0,25
6	61,92		0,94	0,88
7	60,89		-0,09	0,01
Σ				5,51

Получаем значение $F_{\text{набл.}}$:

$$F_{\text{набл.}} = \frac{5,51}{1 - 211,67} \cdot 5 = -0,13.$$

Соответствующее критическое значение статистики Фишера $F_{kp} = F_{\alpha; m; n - n - 1}$ (в некоторых справочных материалах $m = k_1$ или λ_1 ; $m - n - 1 = k_2$ или λ_2). Находим по таблице:

$$F_{kp} = F_{0,05; 1; 5} = 6,61.$$

Так как $F_{\text{набл.}} = -0,13 < F_{kp} = 6,61$, то на уровне значимости $\alpha = 0,05$ нулевая гипотеза принимается и делается вывод: параметры построенной линейной регрессии статистически незначимы. Причина этого, скорее всего, в невысокой тесноте выявленной зависимости и небольшой по объему выборке данных.

2. Сравнивая полученные характеристики параболической модели с характеристиками линейной, степенной и показательной моделей, видим, что параболическая модель несколько лучше других описывает взаимосвязь рассматриваемых переменных величин в условиях данной задачи. Однако и она не вполне адекватна эмпирическим данным. Исследование необходимо продолжить.

Задачи для самостоятельного решения

Задача 2-А. Зависимость среднемесячной производительности труда от возраста рабочих характеризуется моделью: $y = a + bx + cx^2$. Ее использование привело к результатам, представленным в таблице 2.15.

Таблица 2.15

№ п/п	Производительность труда рабочих, тыс. руб., у		№ п/п	Производительность труда рабочих, тыс. руб., у	
	фактическая	расчетная		фактическая	расчетная
1	12	10	6	11	12
2	8	10	7	12	13
3	13	13	8	9	10
4	15	14	9	11	10
5	16	15	10	9	9

Оцените качество модели, определив ошибку аппроксимации, индекс корреляции и F-критерий Фишера.

Задача 2-В. По совокупности 30 предприятий торговли изучается зависимость между признаками: x - цена на товар А, тыс. руб.; y - прибыль торгового предприятия, млн. руб.

При оценке регрессионной модели были получены следующие промежуточные результаты:

$$\sum (y_j - \hat{y}_x)^2 = 39000;$$
$$\sum (y_j - \bar{y})^2 = 120000.$$

1. Поясните, какой показатель корреляции можно определить по этим данным.
2. Постройте таблицу дисперсионного анализа для расчета значения F-критерия Фишера.
3. Сравните фактическое значение F-критерия с табличным. Сделайте выводы.

Задача 2-С. Изучается зависимость потребления материалов u от объема производства продукции x . По 20 наблюдениям были получены следующие варианты уравнения регрессии:

$$1. y = 3 + 2x + \epsilon.$$

(6,48)

$$2. \ln y = 2,5 + 0,2 \ln x + \epsilon, \quad r^2 = 0,68.$$

(6,19)

$$3. y = 3 + 1,5x + 0,1x^2, \quad r^2 = 0,701.$$

(3,0) (2,65)

В скобках указаны фактические значения t-критерия.

1. Определите коэффициент детерминации для 1-го уравнения.
2. Запишите функции, характеризующие зависимость u от x во 2-м и 3-м уравнениях.
3. Определите коэффициенты эластичности для каждого из уравнений.
4. Выберите наилучший вариант уравнения регрессии.

Задача 2-Д. Пусть имеется следующая модель регрессии, характеризующая зависимость u от x :

$$y = 8 - 7x + \epsilon.$$

Известно также, что $r_{xy} = -0,5$; $n = 20$.

1. Постройте доверительный интервал для коэффициента регрессии в этой модели: а) с вероятностью 90%; б) с вероятностью 99%.
2. Проанализируйте результаты, полученные в п.1, и поясните причины их различий.

Задача 2-Е. Для трех видов продукции А, В и С модели зависимости удельных постоянных расходов от объема выпускаемой продукции выглядят следующим образом:

$$\begin{aligned}y_A &= 600, \\y_B &= 80 + 0,7x, \\y_C &= 40x^{0,5}.\end{aligned}$$

1. Определите коэффициенты эластичности по каждому виду продукции и поясните их смысл.

2. Сравните при $x = 1000$ эластичность затрат для продукции В и С.

3. Определите, каким должен быть объем выпускаемой продукции, чтобы коэффициенты эластичности для продукции В и С были равны.

Задача 2-Ф. Исследуя спрос на телевизоры марки N, аналитический отдел компании ABC по данным, собранным с 19 торговых точек компании, выявил следующую зависимость:

$$\ln y = 10,5 - 0,8 \ln x + \varepsilon,$$

(2,5) (-4,0)

где y – объем продаж телевизоров марки N в отдельной торговой точке;

x - средняя цена телевизора в данной торговой точке.

В скобках приведены фактические значения t -критерия Стьюдента для параметров уравнения регрессии.

До проведения этого исследования администрация компании предполагала, что эластичность спроса по цене для телевизоров марки N составляет -0,9. Подтвердилось ли предположение администрации результатами исследования?

Задача 2-Г. Получены функции:

$$\begin{array}{ll}1. y = a + bx^3 + \varepsilon, & 5. y = b + cx^2 + \varepsilon, \\2. y = a + b \ln x + \varepsilon, & 6. y = 1 + a(1 - x^b) + \varepsilon, \\3. \ln y + a + b \ln x + \varepsilon, & 7. y = a + b \frac{x}{10} + \varepsilon. \\4. y = a + bx^2 + \varepsilon,\end{array}$$

Определите, какие из представленных выше функций линейны по переменным, линейны по параметрам, нелинейны ни по переменным, ни по параметрам.

Задача 2-Н. По территории Центрального района известны данные за 1995 г. (таблица 2.16).

Таблица 2.16

Район	Доля денежных доходов, направленных на прирост сбережений во вкладах, займах, сертификатах и на покупку валюты, в общей сумме среднедушевого денежного дохода, %, у	Среднемесячная начисленная заработка платы, тыс. руб., х
Брянская обл.	6,9	289
Владимирская обл.	8,7	334
Ивановская обл.	6,4	300
Калужская обл.	8,4	343
Костромская обл.	6,1	356
Орловская обл.	9,4	289
Рязанская обл.	11,0	341
Смоленская обл.	6,4	327
Тверская обл.	9,3	357
Тульская обл.	8,2	352
Ярославская обл.	8,6	381

Задание

- Постройте поле корреляции и сформулируйте гипотезу о форме связи.
- Рассчитайте параметры уравнений линейной, степенной, экспоненциальной, полулогарифмической, обратной, гиперболической парной регрессии.
- Оцените тесноту связи с помощью показателей корреляции и детерминации.
- Дайте с помощью среднего (общего) коэффициента эластичности сравнительную оценку силы связи фактора с результатом.
- Оцените с помощью средней ошибки аппроксимации качество уравнений.
- Оцените с помощью F-критерия Фишера статистическую надежность результатов регрессионного моделирования. По значениям характеристик, рассчитанным в пунктах 4, 5 и данном пункте, выберите лучшее уравнение регрессии и дайте его обоснование.
- Рассчитайте прогнозное значение результата, если прогнозное значение фактора увеличится на 10% от его среднего уровня. Определите доверительный интервал прогноза для уровня значимости $\alpha = 0,05$.

Раздел 3.

ВРЕМЕННЫЕ РЯДЫ. ПОСТРОЕНИЕ ТРЕНДА

Модели, построенные по данным, характеризующим один и тот же объект за ряд последовательных моментов (периодов), называются *моделями временных рядов*.

Временной ряд - это совокупность значений какого-либо показателя за несколько последовательных моментов или периодов.

Каждый уровень временного ряда формируется из *трендовой* (T), *циклической* (S) и *случайной* (ε) компонент.

Модели, в которых временной ряд представлен как сумма перечисленных компонент, - *аддитивные модели*, как произведение - *мультипликативные модели временного ряда*.

Аддитивная модель имеет вид: $Y = T + S + \varepsilon$,

мультипликативная модель: $Y = T \cdot S \cdot \varepsilon$.

Построение аддитивной и мультипликативной моделей сводится к расчету значений T , S и ε для каждого уровня ряда.

В процессе построения модели необходимо учесть следующее.

Среднее значение объясняемой переменной \bar{y}_t вычисляется по формуле:

$$\bar{y}_t = \frac{\sum_{t=1}^n y_t}{n}.$$

Исправленная дисперсия объясняемой переменной вычисляется по формуле:

$$S^2 = \frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y}_t)^2.$$

Исправленное среднее квадратическое отклонение наблюдаемой переменной (или эмпирический стандарт) S найдется по формуле:

$$S = \sqrt{S^2}.$$

Для устранения случайной составляющей выполняется сглаживание временного ряда. Если сглаживание временного ряда осуществляется, например, по $m = 5$ точкам, то для $p = 1$ (степень полинома), т.е. в случае линейного сглаживания новое значение временного ряда вычисляется как среднее арифметическое пяти заданных значений ряда. Таким образом, новое, третье значение вычисляется по пяти первым заданным, четвертое - по пяти заданным, начиная со второго и т.д., то есть:

$$\tilde{y}_t = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5},$$

где y_t – заданное значение элемента временного ряда;

\tilde{y}_t – сглаженное значение элемента временного ряда ($t = 3, \dots, n - 3$).

Для вычисления сглаженных первых и последних $\frac{m-1}{2}$ значений ряда (то есть при $m = 5$ вычисляют два первых и два последних ряда) при линейном сглаживании используют формулу:

$$\tilde{y}_1 = \frac{1}{5}(3y_1 + 2y_2 + y_3 - y_5),$$

$$\tilde{y}_2 = \frac{1}{10}(4y_1 + 3y_2 + 2y_3 + y_4),$$

$$\tilde{y}_{n-1} = \frac{1}{10}(4y_n + 3y_{n-1} + 2y_{n-2} + y_{n-3}),$$

$$\tilde{y}_n = \frac{1}{5}(3y_n + 2y_{n-1} + y_{n-2} - y_{n-4}).$$

Начинается вычисление с первых двух членов сглаженного ряда. Затем вычисляются два последних члена сглаженного ряда. И, наконец, последовательно вычисляются члены сглаженного ряда, начиная с третьего (как средние арифметические пяти заданных членов, каждый раз отсчитывая эту пятерку, отступая еще на один член от начала ряда). По полученным результатам строится график сглаженного ряда.

Задачей автокорреляционного анализа временного ряда является установление степени и временного интервала зависимости последующих членов ряда от предыдущих. Наличие корреляционной связи между последующими и предыдущими членами ряда также служит информативным признаком временного ряда.

Коэффициенты автокорреляции рассчитываются по формуле:

$$r_k = \frac{\overline{(y_t \cdot y_{t+k})} - \bar{y}_t \cdot \bar{y}_{t+k}}{S_1 \cdot S_2},$$

где $\bar{y}_t, \bar{y}_{t+k}, \overline{y_t \cdot y_{t+k}}$ – средние значения рядов y_t и y_{t+k} ; S_1 и S_2 – исправленные средние квадратические отклонения рядов y_t и y_{t+k} ; $k = 1, 2, \dots$.

Средние значения величин $\bar{y}_t, \bar{y}_{t+k}, \overline{y_t \cdot y_{t+k}}$ вычисляются как средние арифметические этих значений:

$$\overline{y_t \cdot y_{t+k}} = \sum_{t=1}^{n-k} \frac{y_t \cdot y_{t+k}}{n-k}; \quad \bar{y}_{t+k} = \sum_{t=1}^{n-k} \frac{y_{t+k}}{n-k}; \quad \bar{y}_t = \sum_{t=1}^{n-k} \frac{y_t}{n-k}.$$

Исправленные средние квадратические отклонения также вычисляются известным путем:

$$S_1 = \sqrt{S_1^2} = \sqrt{\sum_{t=1}^{n-k} \frac{(y_t - \bar{y}_t)^2}{n-1}}; \quad S_2 = \sqrt{S_2^2} = \sqrt{\sum_{t=1}^{n-k} \frac{(y_{t+k} - \bar{y}_{t+k})^2}{n-k-1}}.$$

Затем находятся значения коэффициентов автокорреляции для различных значений k ($k=1, 2, 3, \dots$).

Последовательность полученных коэффициентов автокорреляции ряда называется *автокорреляционной функцией* временного ряда.

Коррелограммой обычно называют график зависимости значений автокорреляционной функции от величины *лага* (порядка коэффициента автокорреляции). Строится коррелограмма в осях k – величина лага, r – значение автокорреляционной функции (величина коэффициента автокорреляции).

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, а следовательно, и лаг, при котором связь между текущим и предыдущими уровнями ряда наиболее тесная, то есть при помощи анализа автокорреляционной функции и коррелограммы можно выявить структуру ряда.

Если в рассматриваемой задаче наиболее высоким оказался коэффициент автокорреляции *первого порядка*, это означает, что исследуемый ряд содержит *только тенденцию*. Если наиболее высоким оказывается коэффициент автокорреляции порядка τ , то тогда ряд содержит циклические колебания с периодичностью в τ моментов времени. Если же ни один из коэффициентов автокорреляции не является значимым, это означает одно из двух: либо ряд не содержит тенденции и циклических колебаний, либо ряд содержит сильную нелинейную тенденцию и необходим дополнительный анализ.

Степень p полиномиального тренда устанавливается чаще всего методом переменных разностей. Этот метод заключается в вычислении переменных разностей и проверке гипотезы о равенстве дисперсий предыдущих и последующих разностей.

Сначала вычисляют *первые разности*:

$$\Delta^1 y_t = y_{t+1} - y_t,$$

где $t = 1, \dots, n - 1$.

Затем по первым разностям вычисляют *вторые разности*:

$$\Delta^2 y_t = \Delta^1 y_{t+1} - \Delta^1 y_t,$$

где $t = 1, \dots, n - 2$.

Далее последовательно вычисляются *разности 3-го, 4-го и т.д. порядков* (до n -го порядка):

$$\Delta^m y_t = \Delta^{m-1} y_{t+1} - \Delta^{m-1} y_t,$$

где $t = 1, \dots, n - m$.

На каждом шаге, начиная с $m = 0$, вычисляют:

а) дисперсии разностей m -го порядка по формуле:

$$S_m^2 = \frac{\sum_{i=1}^{n-m} (\Delta^m y_i - \bar{\Delta^m y}_i)^2}{(n-m-1) \cdot (2m)!} \cdot (m!)^2;$$

б) для каждого двух (предыдущей и последующей) дисперсий проверяют гипотезу о равенстве дисперсий по критерию Фишера:

$$F_m = \begin{cases} S_{m-1}^2 / S_m^2 & \text{при } S_{m-1}^2 > S_m^2 \\ S_m^2 / S_{m-1}^2 & \text{при } S_m^2 > S_{m-1}^2 \end{cases};$$

Проверка заключается в сравнении вычисленной статистики Фишера F_m с ее критическим значением $F_{kp} = F(\alpha, k_1, k_2)$, где α - принятый уровень значимости; $k_1 = n - m$, $k_2 = n - m - 1$ (степени свободы).

Для 5% уровня значимости критические значения распределения Фишера приведены в таблице 3.1.

Таблица 3.1

Степени свободы	Критические значения распределения Фишера						
	k_2 / k_1	5	10	15	20	25	30
5	5,0	4,7	4,6	4,6	4,5	4,5	
10	3,3	3,0	2,8	2,8	2,7	2,7	
15	2,9	2,5	2,4	2,3	2,3	2,2	
20	2,7	2,3	2,2	2,1	2,1	2,0	
25	2,6	2,2	2,0	1,9	1,9	1,8	
30	2,5	2,2	2,0	1,9	1,9	1,8	

Последовательность дисперсий $S_m^2 = \frac{\sum_{i=1}^{n-m} (\Delta^m y_i - \bar{\Delta^m y}_i)^2}{(n-m-1) \cdot (2m)!} \cdot (m!)^2$ убывает с

ростом m , и при некотором значении $p = m - 1$ выполняется неравенство $F_m < F_{kp}$ (это означает, что сравниваемые дисперсии отличаются незначимо). В противном случае процедура вычислений разности и их дисперсий продолжается. Полученное значение p и является степенью полиномиального тренда. Если $p = 1$, искомое уравнение тренда в общем виде выглядит следующим образом: $\hat{y}_t = b t + a$.

Коэффициенты полиномиального тренда вычисляются путем решения соответствующей системы уравнений (используя МНК).

Для $p = 1$ коэффициенты тренда оцениваются по решению системы:

$$\begin{cases} b \sum_{t=1}^{10} t^2 + a \sum_{t=1}^{10} t = \sum_{t=1}^{10} t \cdot y_t, \\ b \sum_{t=1}^{10} t + a n = \sum_{t=1}^{10} y_t. \end{cases}$$

Сравнение эмпирических коэффициентов b и a с некоторыми теоретически ожидаемыми значениями β и α этих коэффициентов может быть осуществлено по схеме статистической проверки гипотез.

Для проверки гипотез:

$$\begin{array}{ll} H_0 : b = \beta, & H_0 : a = \alpha, \\ H_1 : b \neq \beta & H_1 : a \neq \alpha \end{array}$$

используется статистика Стьюдента: $T = \frac{(b - \beta)}{S_b}$ и $T = \frac{(a - \alpha)}{S_a}$ соответственно.

Функции S_b и S_a представляют собой стандартные ошибки эмпирических коэффициентов b и a . Находятся эти ошибки как корни квадратные из соответствующих необъясненных дисперсий:

$$S_b = \sqrt{S_b^2} \quad \text{и} \quad S_a = \sqrt{S_a^2}.$$

Необъясненные дисперсии коэффициентов в общем виде записываются так:

$$S_b^2 = \frac{\sum e_i^2}{n(n-2)(\bar{x}^2 - \bar{t}^2)} = \frac{\sum (y_i - a - bt)^2}{n(n-2)(\bar{x}^2 - \bar{t}^2)},$$

$$S_a^2 = \bar{t}^2 S_b^2.$$

Учитывая обозначения переменных в нашей задаче, выражения для необъясненных дисперсий коэффициентов можно переписать в таком виде:

$$S_b^2 = \frac{\sum (y_i - a - bt)^2}{n(n-2)(\bar{t}^2 - \bar{t}^2)},$$

$$S_a^2 = \bar{t}^2 S_b^2.$$

При справедливости H_0 Т-статистика имеет распределение Стьюдента с числом степеней свободы $v = n - 2$, где n – объем выборки.

Предположение $H_0: b = \beta$ или $a = \alpha$ отклоняется на основании данного критерия, если:

$$|T_{\text{набл}}| = \left| \frac{b - \beta}{S_b} \right| \geq T_{\frac{\chi}{2}, n-2}, \quad \text{или} \quad |T_{\text{набл}}| = \left| \frac{a - \alpha}{S_a} \right| \geq T_{\frac{\chi}{2}, n-2},$$

где $T_{\frac{\chi}{2}, n-2}$ – критическая статистика $T_{\text{кр}}$, χ – требуемый уровень значимости (в нашей задаче $\chi = 0,05$).

При невыполнении вышезаписанного условия считается, что нет оснований для отклонения гипотезы H_0 .

На начальном этапе статистического анализа построенной модели наиболее важной является задача проверки справедливости установленной линейной зависимости между объясняющей и объясняемой переменными.

Эта проблема может быть решена по рассмотренной выше схеме, но проверяемые гипотезы выглядят несколько иначе.

Для коэффициента b :

$$H_0: b = 0,$$

$$H_1: b \neq 0.$$

Гипотеза в такой постановке обычно называется *гипотезой о статистической значимости вычисленного коэффициента*.

Если H_0 принимается, то говорят, что коэффициент b статистически незначим. При отклонении H_0 коэффициент b считается статистически значимым, что указывает на наличие определенной линейной зависимости между объясняющей и объясняемой переменными.

Поскольку предполагается, что $\beta = 0$ (из того, что полагается $b = 0$), то значимость коэффициента регрессии b проверяется по формуле:

$$T = \frac{|b - \beta|}{S_b} = \frac{|b|}{S_b} = \frac{|b|}{\sqrt{S_b^2}}$$

то есть с помощью анализа отношения величины коэффициента b к его стандартной ошибке.

По аналогичной схеме на основе Т-статистики проверяется гипотеза о статистической значимости коэффициента a :

$$T = \frac{|a - \alpha|}{S_a} = \frac{|a|}{S_a}$$

Отметим, что более важным является анализ статистической значимости коэффициента b , так как именно в нем скрыто влияние объясняющей переменной на зависимую переменную.

Для оценивания качества трендовой модели в целом необходимо рассмотреть ряд остатков – разностей значений ряда и значений тренда

$$e_t = y_t - \hat{y}_t.$$

Затем проверяют следующие гипотезы:

а) о случайности ряда остатков методом поворотных точек (поворотная точка – точка экстремума, то есть точка, в которой значение величины больше или меньше, чем значения в соседних точках); число таких точек d можно найти по графику ряда остатков. Среднее число точек поворота для случайного ряда \bar{d} и их дисперсия S_d^2 равны:

$$\bar{d} = \frac{(2n-4)}{3}, \quad S_d^2 = \frac{(16n-29)}{90}.$$

Затем вычисляют статистику $Z = \frac{|\bar{d} - \bar{\bar{d}}|}{S_d}$, и если $Z < 1,96$, то гипотеза о

случайности ряда остатков принимается и на уровне значимости 5% можно сделать вывод о том, что тренд существует;

б) о равенстве математического ожидания ряда остатков нулю по статистике

$$T = \frac{\bar{e}_t \sqrt{n}}{S_e},$$

где \bar{e}_t - среднее значение ряда остатков, S_e - среднее квадратическое отклонение ряда остатков; на 5% уровне значимости вычисленное значение Т сравнивается с критическим значением T_{kp} , взятым из таблицы 3.2 при ($n - 1$) степенях свободы:

Таблица 3.2

Критические значения распределения Стьюдента

k	3	5	7	10	13	16	20	30	∞
T_{kp}	3,18	2,57	2,45	2,23	2,16	2,12	2,09	2,04	1,96

Если вычисленное значение статистики окажется меньше критического ее значения ($T < T_{kp}$), то гипотеза о равенстве математического ожидания ряда остатков нулю принимается и модель на уровне значимости 5% считается адекватной;

в) об отсутствии автокорреляции ряда остатков; при этом используется критерий Дарбина-Уотсона со статистикой DW:

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}.$$

Если $DW \in [2;4]$, следует использовать вспомогательную статистику $DW' = 4 - DW$.

Расчетное значение DW (или DW') сравнивается с верхним D_2 и нижним D_1 критическими значениями статистики DW, представленными в таблице 3.3, для различной длины ряда n и числа определяемых параметров k на уровне значимости 5% (0,05).

Таблица 3.3

Критические значения статистики DW для различной длины ряда n и числа определяемых параметров k на уровне значимости 5%

n	K=1		K=2		K=3	
	D_1	D_2	D_1	D_2	D_1	D_2
10	0,98	1,34	0,94	1,52	0,67	1,72
15	1,08	1,36	0,95	1,54	0,82	1,75
20	1,35	1,49	1,28	1,57	1,21	1,65

Если расчетное значение критерия DW больше верхнего табличного значения D_2 , то гипотеза о независимости уровней остаточной последовательности, то есть об отсутствии в ней автокорреляции, принимается.

Если значение DW меньше нижнего табличного значения D_1 , то эта гипотеза отвергается и модель считается неадекватной.

Если же значение D находится между значениями D_2 и D_1 , включая сами эти значения, то считается, что нет достаточных оснований сделать тот или иной вывод и необходимы дальнейшие исследования, например, по большей выборке данных.

г) о возможности осуществления оценки соответствия тренда статистическим данным с помощью коэффициента детерминации; известно, что *коэффициент детерминации R^2 является суммарной мерой общего качества уравнения регрессии (его соответствия статистическим данным)*.

В общем случае коэффициент детерминации рассчитывается по формуле:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y}_i)^2}.$$

Для этого коэффициента в общем случае справедливо соотношение $0 \leq R^2 \leq 1$. Чем слабее линейная связь между X и Y , тем R^2 ближе к нулю, и чем эта связь значительнее, тем ближе R^2 к единице.

Коэффициент детерминации является мерой, позволяющей определить, в какой степени найденная прямая регрессии дает лучший результат для объяснения поведения зависимой переменной Y , чем горизонтальная прямая $Y = \bar{y}$.

Трендовая модель считается адекватной, если подтверждены все четыре гипотезы (а,б,в,г).

Осуществим выбор модели, имеющей наибольший процент достоверных прогнозов для краткосрочного прогнозирования (на один шаг вперед).

Для краткосрочного прогнозирования используют следующие модели:

- 1) прогноз по одному последнему значению

$$y_{n+1}^{(1)} = y_n;$$

- 2) прогноз по двум последним значениям

$$y_{n+1}^{(2)} = 2y_n - y_{n-1};$$

- 3) прогноз по трем последним значениям

$$y_{n+1}^{(3)} = \frac{4y_n + y_{n-1} - 2y_{n-2}}{3};$$

- 4) прогноз по четырем последним значениям

$$y_{n+1}^{(4)} = \frac{2y_n + y_{n-1} - y_{n-3}}{2};$$

5) прогноз по пяти последним значениям

$$y_{n+1}^{(5)} = \frac{8y_n + 5y_{n-1} + 2y_{n-3} - y_{n-5} - 4y_{n-4}}{10}.$$

Для выбора модели необходимо по каждому варианту вычислить абсолютные погрешности прогнозных значений членов ряда по формуле:

$$\Delta_k = |y_{n+1}^{(k)} - y_{n+1}|.$$

Затем найденные абсолютные погрешности необходимо сравнить с определенным самим исследователем критическим значением погрешности (по условию решаемой эконометрической задачи). Найденные разности фиксируются как K^+ или K^- (в зависимости от того, какой знак имеет эта разность).

Процент достоверных прогнозов по каждой модели определяется по формуле $\delta_1 = \frac{\sum K^+}{\sum K} \cdot 100\%$.

Долговременный прогноз на число шагов k вперед производят на основе уравнений тренда:

- $\tilde{y}_{n+k} = a(n+k) + b$ (если тренд линейный);
- $\tilde{y}_{n+k} = a(n+k)^2 + b(n+k) + c$ (если тренд параболический).

Примеры решения избранных задач

Задача 3.1. По данным за 18 месяцев построено уравнение регрессии зависимости прибыли предприятия y (млн. руб.) от цен на сырье x_1 (тыс. руб. за 1 т) и производительности труда x_2 (ед. продукции на 1 работника):

$$\hat{y} = 200 - 1.5 \cdot x_1 + 4.0 \cdot x_2.$$

При анализе остаточных величин были использованы значения, приведенные в таблице 3.4.

Таблица 3.4

№	y	x_1	x_2
1	210	800	300
2	720	1000	500
3	300	1500	600
...

$$\sum \varepsilon_i^2 = 10500, \quad \sum (\varepsilon_i - \varepsilon_{i-1})^2 = 40000.$$

Задание

1. По трем позициям рассчитать \hat{y}_i , ε_i , ε_{i-1} , ε_i^2 , $(\varepsilon_i - \varepsilon_{i-1})^2$;

2. Рассчитать критерий Дарбина-Уотсона;

3. Оценить полученный результат при 5%-ном уровне значимости;

4. Указать, пригодно ли уравнение для прогноза.

Решение

1. Регрессионное значение объясняемой переменной \hat{y} , определяется путем подстановки фактических значений x_1 и x_2 в уравнение регрессии:

$$\hat{y}_1 = 200 - 1,5 \cdot 800 + 4,0 \cdot 300 = 200,$$

$$\hat{y}_2 = 200 - 1,5 \cdot 1000 + 4,0 \cdot 500 = 700,$$

$$\hat{y}_3 = 200 - 1,5 \cdot 1500 + 4,0 \cdot 600 = 350.$$

Остатки ε_i рассчитываются по формуле:

$$\varepsilon_i = y_i - \hat{y}_i.$$

Следовательно:

$$\varepsilon_1 = 210 - 200 = 10, \quad \varepsilon_2 = 720 - 700 = 20, \quad \varepsilon_3 = 300 - 350 = -50, \quad \varepsilon_1^2 = 100, \quad \varepsilon_2^2 = 400, \quad \varepsilon_3^2 = 2500.$$

Результаты вычислений представим в виде таблицы 3.5.

Таблица 3.5

№	\hat{y}_i	ε_i	ε_{i-1}	$(\varepsilon_i - \varepsilon_{i-1})$	$(\varepsilon_i - \varepsilon_{i-1})^2$	ε_i^2
1	200	10	-	-	-	100
2	700	20	10	10	100	400
3	350	-50	20	-70	4900	2500
...
Σ					40000	10500

2. Критерий Дарбина-Уотсона рассчитывается по формуле:

$$d = \frac{\sum (\varepsilon_i - \varepsilon_{i-1})^2}{\sum \varepsilon_i^2} = \frac{40000}{10500} = 3,81.$$

3. Фактическое значение d сравниваем с табличными значениями при 5%-ном уровне значимости. При $n = 18$ месяцев и $m = 2$ (число факторов) нижнее значение d' равно 1,05, а верхнее – 1,53. Так как фактическое значение d близко к 4, можно считать, что автокорреляция в остатках характеризуется отрицательной величиной. Чтобы проверить значимость отрицательного коэффициента автокорреляции, найдем величину:

$$4 - d = 4 - 3,81 = 0,19,$$

что значительно меньше, чем d' . Это означает наличие в остатках автокорреляции.

4. Уравнение регрессии не может быть использовано для прогноза, так как в нем не устранена автокорреляция в остатках, которая может

иметь разные причины. Автокорреляция в остатках может означать, что в уравнение не включен какой-либо существенный фактор. Возможно также, что форма связи неточна, а может быть, в рядах динамики имеется общая тенденция.

Задача 3.2. Имеются данные о величине дохода на одного члена семьи и расхода на товар A , приведенные в таблице 3.6.

Таблица 3.6

Показатель	1997	1998	1999	2000	2001	2002
Расходы на товар A , руб.	30	35	39	44	50	53
Доход на одного члена семьи, % к 1997 г.	100	103	105	109	115	118

Задание

1. Определить ежегодные абсолютные приrostы доходов и расходов и сделать выводы о тенденции развития каждого ряда.
2. Перечислить основные пути устранения тенденции для построения модели спроса на товар A в зависимости от дохода.
3. Построить линейную модель спроса, используя первые разности уровней исходных динамических рядов.
4. Пояснить экономический смысл коэффициента регрессии.
5. Построить линейную модель спроса на товар A , включив в нее фактор времени; интерпретировать полученные параметры.

Решение

1. Обозначим расходы на товар A через y , а доходы одного члена семьи – через x . Ежегодные абсолютные приросты определяются по формулам:

$$\Delta y_t = y_t - y_{t-1}, \quad \Delta x_t = x_t - x_{t-1}.$$

Соответствующие расчеты можно оформить в виде таблицы 3.7.

Таблица 3.7

y_t	Δy_t	x_t	Δx_t
30	-	100	-
35	5	103	3
39	4	105	2
44	5	109	4
50	6	115	6
53	3	118	3

Значения Δy не имеют четко выраженной тенденции, они варьируют вокруг среднего уровня, что означает наличие в ряде динамики линейного тренда (линейной тенденции). Аналогичный вывод можно сделать и по ряду x : абсолютные приросты не имеют систематической направленности, они примерно стабильны, а следовательно, ряд характеризуется линейной тенденцией.

2. Так как ряды динамики имеют общую тенденцию к росту, то для построения регрессионной модели спроса на товар А в зависимости от дохода необходимо устранить тенденцию. С этой целью модель может строиться по первым разностям, то есть $\Delta y = f(\Delta x)$, если ряды динамики характеризуются линейной тенденцией.

Другой возможный путь учета тенденции при построении моделей – найти по каждому ряду уравнение тренда:

$$\hat{y}_t = f(t) \quad u \quad \hat{x}_t = f(t)$$

и отклонения от него:

$$dy = y_t - \hat{y}_t, \quad dx = x_t - \hat{x}_t.$$

Далее модель строится по отклонениям от тренда:

$$dy = f(dx).$$

При построении эконометрических моделей чаще используется другой путь учета тенденции – включение в модель фактора времени. Иными словами, модель строится по исходным данным, но в нее в качестве самостоятельного фактора включается время, то есть $\bar{y}_t = f(x, t)$.

3. Модель имеет вид:

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

Для определения параметров a и b применяется МНК. Система нормальных уравнений следующая:

$$\begin{cases} \sum \Delta y = n \cdot a + b \cdot \sum \Delta x, \\ \sum \Delta y \Delta x = a \cdot \sum \Delta x + b \cdot \sum \Delta^2 x. \end{cases}$$

Применительно к нашим данным имеем:

$$\begin{cases} 23 = 5 \cdot a + 18 \cdot b, \\ 88 = 18 \cdot a + 74 \cdot b. \end{cases}$$

Решая эту систему, получим:

$$a = 2,565 \quad u \quad b = 0,565,$$

откуда модель имеет вид:

$$\Delta \hat{y} = 2,565 + 0,565 \cdot \Delta x.$$

4. Коэффициент регрессии $b = 0,565$ руб. Он означает, что с увеличением прироста душевого дохода на 1% расходы на товар А увеличиваются со средним ускорением, равным 0,565 руб.

5. Модель имеет вид:

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

Применяя МНК, получим систему нормальных уравнений:

$$\begin{cases} \sum y = n \cdot a + b \cdot \sum x + c \cdot \sum t, \\ \sum yx = a \cdot \sum x + b \cdot \sum x^2 + c \cdot \sum x \cdot t, \\ \sum yt = a \cdot \sum t + b \cdot \sum x \cdot t + c \cdot \sum t^2. \end{cases}$$

Расчеты приведены в таблице 3.8.

Таблица 3.8

t	y	x	yx	y_t	x_t	x_t^2	t^2
1	30	100	3000	30	100	10000	1
2	35	103	3605	70	206	10609	4
3	39	105	4095	117	315	11025	9
4	44	109	4796	176	436	11881	16
5	50	115	5750	250	575	13225	25
6	53	118	6254	318	708	13924	36
21	251	650	27500	961	2340	70664	91

Система уравнений примет вид:

$$\begin{cases} 251 = 6a + 650b + 21c, \\ 27500 = 650a + 70664b + 2340c, \\ 961 = 21a + 2340b + 91c. \end{cases}$$

Решая ее, получим:

$$a = -5,42; \quad b = 0,322; \quad c = 3,516.$$

Уравнение регрессии имеет вид:

$$y = -5,42 + 0,322 \cdot x + 3,516 \cdot t.$$

Параметр $b = 0,322$ фиксирует силу связи y и x . Его величина означает, что с ростом дохода на одного члена семьи на 1% при условии неизменной тенденции расходы на товар A возрастают в среднем на 0,322 руб.

Параметр $c = 3,516$ характеризует среднегодовой абсолютный прирост расходов на товар A под воздействием прочих факторов при условии неизменного дохода.

Задача 3.3. Эмпирический временной ряд представляет объем производства продукции предприятия (по месяцам) y_t за 10 месяцев 2002 года в сопоставимых ценах, млн.руб. (информация представлена в таблице 3.9).

Таблица 3.9

t	1	2	3	4	5	6	7	8	9	10
Месяц	Январь	Февраль	Март	Апрель	Май	Июнь	Июль	Август	Сентябрь	Октябрь
y_t	51	55	62	70	81	75	116	115	125	120

Задание

1. Построить график эмпирического временного ряда.
2. Вычислить среднее значение объясняемой переменной \bar{y}_t , ее исправленную дисперсию S^2 и исправленное среднее квадратическое отклонение s .
3. Осуществить линейное сглаживание эмпирического временного ряда по $m = 5$ точкам выборки. Построить график сглаженного ряда. Сделать вывод о монотонности ряда.
4. Провести автокорреляционный анализ эмпирического временного ряда (рассчитать последовательно коэффициенты автокорреляции между членами ряда, проверить их значимость путем сравнения с критическими значениями при 5% уровне значимости, построить график критического уровня коэффициентов автокорреляции и коррелограмму – график автокорреляционной функции).
5. Определить степень полиномиального тренда методом переменных разностей с использованием F-статистики (статистики Фишера).
6. Вычислить коэффициенты полиномиального тренда путем решения соответствующей системы уравнений (используя МНК).
7. Оценить статистическую значимость найденных коэффициентов.
8. Записать уравнение тренда; восстановить график эмпирического временного ряда и в тех же осях построить график тренда.
9. Провести оценку качества трендовой модели.
10. Осуществить кратковременный (на один шаг вперед) и долгосрочный (на три шага вперед) прогнозы временного ряда.
11. Сделать вывод по результатам решения задачи в целом.

Решение к заданию 1

График эмпирического временного ряда объема производства продукции строим в осях t – время и y_t – наблюдаемая величина, т.е. объем производства продукции (рис. 3.1).

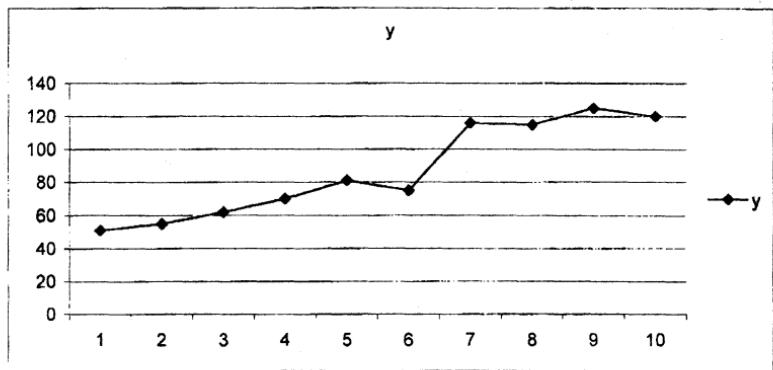


Рис. 3.1. График эмпирического временного ряда (изменение во времени объема производства продукции)

Решение к заданию 2

Среднее значение объема производства продукции \bar{y}_t вычислим по формуле:

$$\bar{y}_t = \frac{\sum_{t=1}^n y_t}{n} = \frac{\sum_{t=1}^{10} y_t}{10} = \frac{51 + 55 + 62 + 70 + 116 + 115 + 125 + 120}{10} = \frac{870}{10} = 87.$$

Исправленную дисперсию значения производства вычислим по формуле:

$$S^2 = \frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y}_t)^2.$$

Выполним промежуточные вычисления (таблица 3.10).

Таблица 3.10

y_t	\bar{y}_t	$y_t - \bar{y}_t$	$(y_t - \bar{y}_t)^2$
51	87	-36	1296
55		-32	1024
62		-25	625
70		-17	289
81		-6	36
75		-12	144
116		29	841
115		28	784
125		38	1444
120		33	1089
			$\sum_{t=1}^{10} (y_t - \bar{y}_t)^2 = 7572$

Следовательно, величина исправленной дисперсии наблюдаемой переменной Y_t будет равна:

$$S^2 = \frac{1}{n-1} \sum_{t=1}^n (y_t - \bar{y}_t)^2 = \frac{1}{9} \cdot 7572 = 757,2.$$

Исправленное среднее квадратическое отклонение наблюдаемой переменной (или эмпирический стандарт) S найдем по формуле:

$$S = \sqrt{S^2} = \sqrt{757,2} = 27,52.$$

Решение к заданию 3

Сглаживание временного ряда выполняется для устранения случайной составляющей. Если сглаживание временного ряда осуществляется по $m = 5$ точкам, то для $p = 1$ (степень полинома), т.е. в случае линейного сглаживания новое значение временного ряда вычисляется как среднее арифметическое пяти заданных значений ряда. Таким образом, новое, третье значение вычисляется по пяти первым заданным, четвертое – по пяти заданным, начиная со второго и т.д., то есть

$$\tilde{y}_t = \frac{y_{t-2} + y_{t-1} + y_t + y_{t+1} + y_{t+2}}{5},$$

где y_t – заданное значение элемента временного ряда;

\tilde{y}_t – сглаженное значение элемента временного ряда ($t = 3, \dots, n-3$).

Для вычисления сглаженных первых и последних $\frac{m-1}{2}$ значений ряда (то есть при $m = 5$ вычисляют два первых и два последних ряда) при линейном сглаживании используют формулы:

$$\tilde{y}_1 = \frac{1}{5}(3y_1 + 2y_2 + y_3 - y_5),$$

$$\tilde{y}_2 = \frac{1}{10}(4y_1 + 3y_2 + 2y_3 + y_4),$$

$$\tilde{y}_{n-1} = \frac{1}{10}(4y_n + 3y_{n-1} + 2y_{n-2} + y_{n-3}),$$

$$\tilde{y}_n = \frac{1}{5}(3y_n + 2y_{n-1} + y_{n-2} - y_{n-4}).$$

Начнем вычисления с первых двух членов сглаженного ряда:

$$\tilde{y}_1 = \frac{1}{5}(3 \cdot 51 + 2 \cdot 55 + 62 - 81) = \frac{244}{5} = 48,8,$$

$$\tilde{y}_2 = \frac{1}{10}(4 \cdot 51 + 3 \cdot 55 + 2 \cdot 62 + 70) = \frac{563}{10} = 56,3.$$

Вычислим два последних члена сглаженного ряда (9-й и 10-й):

$$\tilde{y}_9 = \frac{1}{10}(4 \cdot y_{10} + 3 \cdot y_9 + 2 \cdot y_8 + y_7) = \frac{1}{10}(4 \cdot 120 + 3 \cdot 125 + 2 \cdot 115 + 116) = \frac{1201}{10} = 120,1;$$

$$\tilde{y}_{10} = \frac{1}{5}(3 \cdot y_{10} + 2 \cdot y_9 + y_8 - y_6) = \frac{1}{5}(3 \cdot 120 + 2 \cdot 125 + 115 - 75) = \frac{650}{5} = 130.$$

А теперь последовательно вычислим члены сглаженного ряда, начиная с третьего (как средние арифметические пяти заданных членов, каждый раз отсчитывая эту пятерку, отступая еще на один член от начала ряда):

$$\tilde{y}_3 = \frac{y_1 + y_2 + y_3 + y_4 + y_5}{5} = \frac{51 + 55 + 62 + 70 + 81}{5} = \frac{319}{5} = 63,8,$$

$$\tilde{y}_4 = \frac{y_2 + y_3 + y_4 + y_5 + y_6}{5} = \frac{55 + 62 + 70 + 81 + 75}{5} = \frac{343}{5} = 68,6,$$

$$\tilde{y}_5 = \frac{y_3 + y_4 + y_5 + y_6 + y_7}{5} = \frac{62 + 70 + 81 + 75 + 116}{5} = \frac{404}{5} = 80,8,$$

$$\tilde{y}_6 = \frac{y_4 + y_5 + y_6 + y_7 + y_8}{5} = \frac{70 + 81 + 75 + 116 + 115}{5} = \frac{457}{5} = 91,4,$$

$$\tilde{y}_7 = \frac{y_5 + y_6 + y_7 + y_8 + y_9}{5} = \frac{81 + 75 + 116 + 115 + 125}{5} = \frac{512}{5} = 102,4,$$

$$\tilde{y}_8 = \frac{y_6 + y_7 + y_8 + y_9 + y_{10}}{5} = \frac{75 + 116 + 115 + 125 + 120}{5} = \frac{551}{5} = 110,2.$$

Повторим график эмпирического ряда и в тех же осях построим график сглаженного ряда (рис. 3.2).

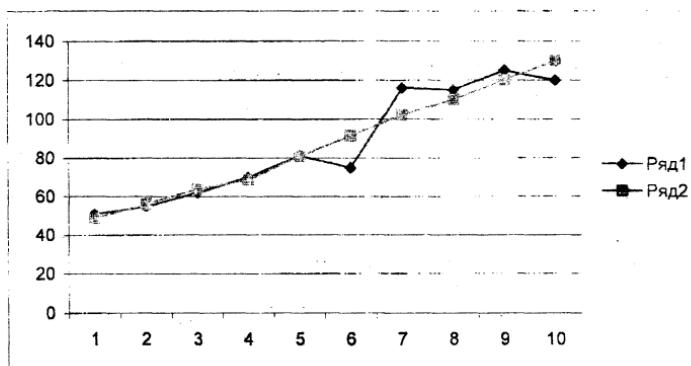


Рис. 3.2. Эмпирический и сглаженный ряды

График сглаженного ряда показывает монотонное возрастание значений ряда во времени.

Решение к заданию 4

Задачей автокорреляционного анализа временного ряда является установление степени и временного интервала зависимости последующих членов ряда от предыдущих. Наличие корреляционной связи между последующими и предыдущими членами ряда также служит информативным признаком временного ряда.

Рассчитаем коэффициенты автокорреляции по формуле:

$$r_k = \frac{\bar{y}_t \cdot \bar{y}_{t+k} - \bar{y}_t^2 \cdot \bar{y}_{t+k}^2}{S_1 \cdot S_2}$$

где $\bar{y}_t, \bar{y}_{t+k}, \bar{y}_t \cdot \bar{y}_{t+k}$ – средние значения рядов y_t и y_{t+k} ; S_1 и S_2 – исправленные средние квадратические отклонения рядов y_t и y_{t+k} ; $k = 1, 2, \dots$.

Средние значения величин $\bar{y}_t, \bar{y}_{t+k}, \bar{y}_t \cdot \bar{y}_{t+k}$ вычисляются как средние арифметические этих значений:

$$\bar{y}_t \cdot \bar{y}_{t+k} = \frac{\sum_{t=1}^{n-k} y_t \cdot y_{t+k}}{n-k}; \quad \bar{y}_{t+k} = \frac{\sum_{t=1}^{n-k} y_{t+k}}{n-k}; \quad \bar{y}_t = \frac{\sum_{t=1}^{n-k} y_t}{n-k}.$$

Исправленные средние квадратические отклонения также вычисляются известным путем:

$$S_1 = \sqrt{S_1^2} = \sqrt{\frac{\sum_{t=1}^{n-k} (y_t - \bar{y}_t)^2}{n-1}}; \quad S_2 = \sqrt{S_2^2} = \sqrt{\frac{\sum_{t=1}^{n-k} (y_{t+k} - \bar{y}_{t+k})^2}{n-k-1}}.$$

Найдем значения коэффициента автокорреляции для различных значений k .

1) $k = 1$.

Составим таблицу промежуточных вычислений для определения средних $\bar{y}_t, \bar{y}_{t+1}, \bar{y}_t \cdot \bar{y}_{t+1}$ (таблица 3.11).

Таблица 3.11

Таблица промежуточных вычислений для определения числителя коэффициента автокорреляции r_1 при $k = 1$

										Средняя
y_t	51	55	62	70	81	75	116	115	125	83,3
y_{t+1}	55	62	70	81	75	116	115	125	120	91
$y_t \cdot y_{t+1}$	2805	3410	4340	5670	6075	8700	13340	14375	15000	8190,6

На основе таблицы 3.11 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_1^2 величины y_t при $k = 1$ (таблица 3.12).

Таблица 3.12

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_t при $k = 1$

y_t	\bar{y}_t	$(y_t - \bar{y}_t)$	$(y_t - \bar{y}_t)^2$
51	83,3	-32,3	1043,29
55		-28,3	800,89
62		-21,3	453,69
70		-13,3	176,89
81		-2,3	5,29
75		-8,3	68,89
116		32,7	1069,29
115		31,7	1004,89
125		41,7	1738,89
			$\sum = 6362,01$

Следовательно, исправленная дисперсия величины y_t при $k = 1$ будет равна: $S_1^2 = \frac{1}{8} \sum_{t=1}^9 (y_t - 83,3)^2 = \frac{6362,01}{8} = 795,251$.

Исправленное среднее квадратическое отклонение величины y_t при $k = 1$ будет равно: $S_1 = \sqrt{S_1^2} = \sqrt{795,251} = 28,2$.

На основе таблицы 3.12 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_2^2 величины y_{t+1} (таблица 3.13).

Таблица 3.13

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_{t+1} при $k = 1$

y_{t+1}	\bar{y}_{t+1}	$(y_{t+1} - \bar{y}_{t+1})$	$(y_{t+1} - \bar{y}_{t+1})^2$
55	91	-36	1296
62		-29	841
70		-21	441
81		-10	100
75		-16	256
116		25	625
115		24	576
125		34	1156
120		29	841
			$\sum = 6132$

Следовательно, исправленная дисперсия величины y_{t+1} при $k = 1$ будет равна $S_2^2 = \frac{1}{8} \sum_{t=1}^9 (y_{t+1} - 91)^2 = \frac{6132}{8} = 766,5$.

Исправленное среднее квадратическое отклонение величины y_{t+1} при $k = 1$ будет равно $S_2 = \sqrt{S_2^2} = \sqrt{766,5} = 27,686$.

По результатам промежуточных вычислений, для $k = 1$ получаем значение коэффициента автокорреляции r_1 :

$$r_1 = \frac{(\bar{y}_t \cdot \bar{y}_{t+1} - \bar{y}_t \cdot \bar{y}_{t+1})}{S_1 \cdot S_2} = \frac{8190,6 - 83,3 \cdot 91}{28,2 \cdot 27,686} = \frac{610,3}{780,745} = 0,782.$$

2) $k = 2$.

Составим таблицу промежуточных вычислений для определения средних $\bar{y}_t, \bar{y}_{t+2}, \bar{y}_t \cdot \bar{y}_{t+2}$ (таблица 3.14).

Таблица 3.14

Таблица промежуточных вычислений для определения чисителя коэффициента автокорреляции r_2 при $k = 2$

									Средняя
y_t	51	55	62	70	81	75	116	115	78,125
y_{t+2}	62	70	81	75	116	115	125	120	95,5
$y_t \cdot y_{t+2}$	3162	3850	5022	5250	9396	8625	14500	13800	7950,625

На основе таблицы 3.14 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_1^2 величины y_t при $k = 2$ (таблица 3.15).

Таблица 3.15

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_t при $k = 2$

y_t	\bar{y}_t	$(y_t - \bar{y}_t)$	$(y_t - \bar{y}_t)^2$
51	78,125	-27,125	735,766
55		-23,125	534,766
62		-16,125	260,016
70		-8,125	66,015
81		2,875	8,266
75		-3,125	9,766
116		37,875	1434,516
115		36,875	1359,766
			$\sum = 4408,875$

Следовательно, исправленная дисперсия величины y_t при $k = 2$ будет равна: $S_1^2 = \frac{1}{7} \sum_{t=1}^8 (y_t - 78,125)^2 = \frac{4408,875}{7} = 629,71$.

Исправленное среднее квадратическое отклонение величины y_t при $k = 2$ будет равно: $S_1 = \sqrt{S_1^2} = \sqrt{629,71} = 25,1$.

На основе таблицы 3.15 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_2^2 величины y_{t+2} (таблица 3.16).

Таблица 3.16

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_{t+2} при $k = 2$

y_{t+2}	\bar{y}_{t+2}	$(y_{t+2} - \bar{y}_{t+2})$	$(y_{t+2} - \bar{y}_{t+2})^2$
62	95,5	-33,5	1122,25
70		-25,5	650,25
81		-14,5	210,25
75		-20,5	420,25
116		20,5	420,25
115		19,5	380,25
125		29,5	870,25
120		24,5	600,25
			$\sum = 4674$

Следовательно, исправленная дисперсия величины y_{t+2} при $k = 2$ будет равна $S_2^2 = \frac{1}{7} \sum_{i=1}^8 (y_{t+2} - 95,5)^2 = \frac{4674}{7} = 667,714$.

Исправленное среднее квадратическое отклонение величины y_{t+2} при $k = 2$ будет равно $S_2 = \sqrt{S_2^2} = \sqrt{667,714} = 25,84$.

По результатам промежуточных вычислений, для $k = 2$ получаем значение коэффициента автокорреляции r_2 :

$$r_2 = \frac{(y_t \cdot y_{t+2} - \bar{y}_t \cdot \bar{y}_{t+2})}{S_1 \cdot S_2} = \frac{7950,625 - 78,125 \cdot 95,5}{25,834 \cdot 25,84} = \frac{489,688}{667,55} = 0,733.$$

3) $k = 3$.

Составим таблицу промежуточных вычислений для определения средних $\bar{y}_t, \bar{y}_{t+3}, y_t \cdot y_{t+3}$ (таблица 3.17).

Таблица 3.17

Таблица промежуточных вычислений для определения числителя коэффициента автокорреляции r_3 при $k = 3$

y_t	51	55	62	70	81	75	116	$\bar{y}_t = 72,857$
y_{t+3}	70	81	75	116	115	125	120	$\bar{y}_{t+3} = 100,286$
$y_t \cdot y_{t+3}$	3570	4455	4650	8120	9315	9375	13920	$y_t \cdot y_{t+3} = 7629,286$

На основе таблицы 3.17 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_3^2 величины y_t при $k = 3$ (таблица 3.18):

Таблица 3.18

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_t при $k = 3$

y_t	\bar{y}_t	$(y_t - \bar{y}_t)$	$(y_t - \bar{y}_t)^2$
51	72,857	-21,857	477,728
55		-17,857	318,872
62		-10,857	117,874
70		-2,85	8,1225
81		8,143	66,308
75		2,143	4,592
116		43,143	1861,318
			$\sum = 2854,814$

Следовательно, исправленная дисперсия величины y_t при $k = 3$ будет равна: $S_1^2 = \frac{1}{6} \sum (y_t - 72,857)^2 = \frac{2854,814}{6} = 475,802$.

Исправленное среднее квадратическое отклонение величины y_t при $k = 3$ будет равно: $S_1 = \sqrt{S_1^2} = \sqrt{475,802} = 21,813$.

На основе таблицы 3.18 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_2^2 величины y_{t+3} (таблица 3.19).

Таблица 3.19

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_{t+3} при $k = 3$

y_{t+3}	\bar{y}_{t+3}	$(y_{t+3} - \bar{y}_{t+3})$	$(y_{t+3} - \bar{y}_{t+3})^2$
70	100,286	-30,286	917,242
81		-19,286	371,95
75		-25,286	639,382
116		15,714	246,93
115		14,714	216,502
125		24,714	610,782
120		19,714	388,642
			$\sum = 3391,43$

Следовательно, исправленная дисперсия величины y_{t+3} при $k = 3$ будет равна $S_2^2 = \frac{1}{6} \sum (y_{t+3} - 100,286)^2 = \frac{3391,43}{6} = 565,238$.

Исправленное среднее квадратическое отклонение величины y_{t+3} при $k = 3$ будет равно $S_2 = \sqrt{S_2^2} = \sqrt{565,238} = 23,775$.

По результатам промежуточных вычислений, для $k = 3$ получаем значение коэффициента автокорреляции r_3 :

$$r_3 = \frac{(y_t \cdot y_{t+3} - \bar{y}_t \cdot \bar{y}_{t+3})}{S_1 \cdot S_2} = \frac{7629,286 - 72,857 \cdot 100,286}{21,813 \cdot 23,775} = \frac{322,749}{518,604} = 0,622.$$

4) $k = 4$.

Составим таблицу промежуточных вычислений для определения средних \bar{y}_t , \bar{y}_{t+4} , $\bar{y}_t \cdot y_{t+4}$ (таблица 3.20).

Таблица 3.20

Таблица промежуточных вычислений для определения числителя коэффициента автокорреляции r_4 при $k = 4$

y_t	51	55	62	70	81	75	$\bar{y}_t = 65,7$
y_{t+4}	81	75	116	115	125	120	$\bar{y}_{t+4} = 105,3$
$y_t \cdot y_{t+4}$	4131	4125	7192	8050	10125	9000	$\bar{y}_t \cdot y_{t+4} = 7103,83$

На основе таблицы 3.20 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_1^2 величины y_t при $k = 4$ (таблица 3.21):

Таблица 3.21

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_t при $k = 4$

y_t	\bar{y}_t	$(y_t - \bar{y}_t)$	$(y_t - \bar{y}_t)^2$
51	65,7	-14,7	216,09
55		-10,7	114,49
62		-3,7	13,69
70		4,3	18,49
81		15,3	234,09
75		9,3	86,49
			$\sum = 683,34$

Следовательно, исправленная дисперсия величины y_t при $k = 4$ будет равна: $S_1^2 = \frac{1}{5} \sum (y_t - 65,7)^2 = \frac{683,34}{5} = 136,668$.

Исправленное среднее квадратическое отклонение величины y_t при $k = 4$ будет равно: $S_1 = \sqrt{S_1^2} = \sqrt{136,668} = 11,69$.

На основе таблицы 3.21 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_2^2 величины y_{t+4} (таблица 3.22):

Таблица 3.22

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_{t+4} при $k = 4$

y_{t+4}	\bar{y}_{t+4}	$(y_{t+4} - \bar{y}_{t+4})$	$(y_{t+4} - \bar{y}_{t+4})^2$
81	105,3	-24,3	590,49
75		-30,3	918,09
116		10,7	114,49
115		9,7	94,09
125		19,7	388,09

120		14.7	216.09
			$\sum = 2321,34$

Следовательно, исправленная дисперсия величины y_{t+4} при $k = 4$ будет равна $S_2^2 = \frac{1}{5} \sum_{t=1}^6 (y_{t+4} - 105,3)^2 = \frac{2321,34}{5} = 464,268$.

Исправленное среднее квадратическое отклонение величины y_{t+4} при $k = 4$ будет равно $S_2 = \sqrt{S_2^2} = \sqrt{464,268} = 21,547$.

По результатам промежуточных вычислений, для $k = 4$ получаем значение коэффициента автокорреляции r_4 :

$$r_4 = \frac{(\overline{y_t \cdot y_{t+4}} - \bar{y}_t \cdot \bar{y}_{t+4})}{S_1 \cdot S_2} = \frac{7103,83 - 65,7 \cdot 105,3}{11,69 \cdot 21,547} = \frac{185,62}{251,884} = 0,737.$$

5) $k = 5$.

Составим таблицу промежуточных вычислений для определения средних $\bar{y}_t, \bar{y}_{t+5}, y_t \cdot y_{t+5}$ (таблица 3.23).

Таблица 3.23

Таблица промежуточных вычислений для определения числового значения коэффициента автокорреляции r_5 при $k = 5$

y_t	51	55	62	70	81	$\bar{y}_t = 63,8$
y_{t+5}	75	116	115	125	120	$\bar{y}_{t+5} = 110,2$
$y_t \cdot y_{t+5}$	3825	6380	7130	8750	9720	$y_t \cdot y_{t+5} = 7161$

На основе таблицы 3.23 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_i^2 величины y_t при $k = 5$ (таблица 3.24).

Таблица 3.24

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_t при $k = 5$

y_t	\bar{y}_t	$(y_t - \bar{y}_t)$	$(y_t - \bar{y}_t)^2$
51	63,8	-12,8	163,84
55		-8,8	77,44
62		-1,8	3,24
70		6,2	38,44
81		17,2	295,84
			$\sum = 578,8$

Следовательно, исправленная дисперсия величины y_t при $k = 5$ будет равна: $S_i^2 = \frac{1}{4} \sum_{t=1}^5 (y_t - 63,8)^2 = \frac{578,8}{4} = 144,7$.

Исправленное среднее квадратическое отклонение величины y_t при $k = 5$ будет равно: $S_1 = \sqrt{S_1^2} = \sqrt{144,7} = 12,029$.

На основе таблицы 3.24 составим таблицу промежуточных вычислений для определения исправленной дисперсии S_2^2 величины y_{t+5} (таблица 3.25).

Таблица 3.25

Таблица промежуточных вычислений для определения исправленной дисперсии величины y_{t+5} при $k = 5$

y_{t+5}	\bar{y}_{t+5}	$(y_{t+5} - \bar{y}_{t+5})$	$(y_{t+5} - \bar{y}_{t+5})^2$
75	110,2	-35,2	1239,04
116		5,8	33,64
115		4,8	23,04
125		14,8	219,04
120		9,8	96,04
			$\sum = 1610,8$

Следовательно, исправленная дисперсия величины y_{t+5} при $k = 5$ будет равна $S_2^2 = \frac{1}{4} \sum_{i=1}^5 (y_{t+5} - 110,2)^2 = \frac{1610,8}{4} = 402,7$.

Исправленное среднее квадратическое отклонение величины y_{t+5} при $k = 5$ будет равно $S_2 = \sqrt{S_2^2} = \sqrt{402,7} = 20,067$.

По результатам промежуточных вычислений, для $k = 5$ получаем значение коэффициента автокорреляции r_5 :

$$r_5 = \frac{(y_t \cdot y_{t+5} - \bar{y}_t \cdot \bar{y}_{t+5})}{S_1 \cdot S_2} = \frac{7161 - 63,8 \cdot 110,2}{12,029 \cdot 20,067} = \frac{130,24}{241,386} = 0,539.$$

Получены следующие значения коэффициентов автокорреляции:

$$r_1 = 0,782; r_2 = 0,733; r_3 = 0,622; r_4 = 0,737; r_5 = 0,539.$$

Эта последовательность полученных коэффициентов автокорреляции ряда данных называется *автокорреляционной функцией* временного ряда.

Коррелограммой обычно называют график зависимости значений автокорреляционной функции от величины *лага* (порядка коэффициента автокорреляции). Строится коррелограмма в осях k – величина лага, r – значение автокорреляционной функции (величина коэффициента автокорреляции) (приложение, рис. 3).

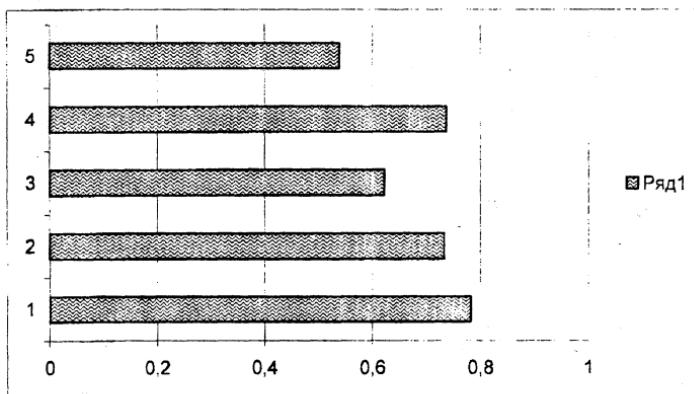


Рис. 3.3. График автокорреляционной функции (коррелограмма)

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, а следовательно, и лаг, при котором связь между текущим и предыдущим уровнями ряда наиболее тесная, то есть при помощи анализа автокорреляционной функции и коррелограммы можно выявить структуру ряда.

В рассматриваемой задаче наиболее высоким оказался коэффициент автокорреляции *первого порядка*. Это означает, что исследуемый ряд содержит *только тенденцию*. Если бы наиболее высоким оказался коэффициент автокорреляции порядка τ , то тогда ряд содержал бы циклические колебания с периодичностью в τ моментов времени. Если же ни один из коэффициентов автокорреляции не является значимым, это означает одно из двух: либо ряд не содержит тенденции и циклических колебаний, либо ряд содержит сильную нелинейную тенденцию и необходим дополнительный анализ.

Решение к заданию 5

Установим степень p полиномиального тренда методом переменных разностей. Этот метод заключается в вычислении переменных разностей и проверке гипотезы о равенстве дисперсий предыдущих и последующих разностей.

Сначала вычисляют первые разности:

$$\Delta^1 y_t = y_{t+1} - y_t,$$

где $t = 1, \dots, n - 1$.

Затем по первым разностям вычисляют вторые разности:

$$\Delta^2 y_t = \Delta^1 y_{t+1} - \Delta^1 y_t,$$

где $t = 1, \dots, n - 2$.

Далее последовательно вычисляются разности 3-го, 4-го и т.д. порядков (до m -го порядка):

$$\Delta^m y_t = \Delta^{m-1} y_{t+1} - \Delta^{m-1} y_t,$$

где $t = 1, \dots, n - m$.

На каждом шаге, начиная с $m = 0$, вычисляют:

а) дисперсии разностей m -го порядка по формуле:

$$S_m^2 = \frac{\sum_{t=1}^{n-m} (\Delta^m y_t - \bar{\Delta}^m y_t)^2}{(n-m-1) \cdot (2m)!} \cdot (m!)^2;$$

б) для каждого двух (предыдущей и последующей) дисперсий

роверяют гипотезу о равенстве дисперсий по критерию Фишера:

$$F_m = \begin{cases} \frac{S_{m-1}^2}{S_m^2} & \text{при } S_{m-1}^2 > S_m^2 \\ \frac{S_m^2}{S_{m-1}^2} & \text{при } S_m^2 > S_{m-1}^2. \end{cases}$$

Проверка заключается в сравнении вычисленной статистики Фишера F_m с ее критическим значением $F_{kp} = F(\alpha, k_1, k_2)$, где α - принятый уровень значимости; $k_1 = n - m$, $k_2 = n - m - 1$ (степени свободы).

Для 5% уровня значимости критические значения распределения Фишера приведены в таблице 3.26.

Таблица 3.26

Степени свободы k_2 / k_1	Критические значения распределения Фишера					
	5	10	15	20	25	30
5	5,0	4,7	4,6	4,6	4,5	4,5
10	3,3	3,0	2,8	2,8	2,7	2,7
15	2,9	2,5	2,4	2,3	2,3	2,2
20	2,7	2,3	2,2	2,1	2,1	2,0
25	2,6	2,2	2,0	1,9	1,9	1,8
30	2,5	2,2	2,0	1,9	1,9	1,8

$$\sum_{t=1}^{n-m} (\Delta^m y_t - \bar{\Delta}^m y_t)^2$$

Последовательность дисперсий $S_m^2 = \frac{\sum_{t=1}^{n-m} (\Delta^m y_t - \bar{\Delta}^m y_t)^2}{(n-m-1) \cdot (2m)!} \cdot (m!)^2$ убывает с

ростом m , и при некотором значении $p = m - 1$ выполняется неравенство $F_m < F_{kp}$ (это означает, что сравниваемые дисперсии отличаются незначительно). В противном случае процедура вычислений разности и их дисперсий продолжается. Полученное значение p является степенью полиномиального тренда.

Составим таблицу 3.27.

Таблица 3.27

Таблица переменных разностей

t	1	2	3	4	5	6	7	8	9	10	$\Delta^m y_t$
$\Delta^0 y_t$	51	55	62	70	81	75	116	115	125	120	87
$\Delta^1 y_t$	4	7	8	11	-6	41	-1	10	-5		7,667
$\Delta^2 y_t$	3	1	3	-17	47	-42	11	-15			-1,125
$\Delta^3 y_t$	-2	2	-20	64	-89	53	-26				-2,571
$\Delta^4 y_t$	4	-22	84	-153	142	-79					-4
$\Delta^5 y_t$	-26	106	-237	295	-221						-16,6
$\Delta^6 y_t$	132	-343	532	-516							-48,75
$\Delta^7 y_t$	-475	875	-1048								-216
$\Delta^8 y_t$	1350	-1923									-286,5
$\Delta^9 y_t$	-3273										-3273

Составив таблицу переменных разностей, приступаем к вычислению дисперсий разностей.

Дисперсия разностей нулевого порядка ($m = 0$) совпадает с дисперсией эмпирического ряда:

$$S_0^2 = S^2 = 925,778.$$

Дисперсию разностей первого порядка ($m = 1$) вычислим по формуле:

$$S_m^2 = \frac{\sum_{t=1}^{n-m} (\Delta^m y_t - \bar{\Delta^m y}_t)^2}{(n-m-1) \cdot (2m)!} \cdot (m!)^2 \Rightarrow S_1^2 = \frac{\sum_{t=1}^9 (\Delta^1 y_t - \bar{\Delta^1 y}_t)^2}{(10-1-1) \cdot 2}.$$

Составим таблицу промежуточных вычислений (таблица 3.28).

Таблица 3.28

Таблица промежуточных вычислений для отыскания дисперсии разностей первого порядка

$\Delta^1 y_t$	$\bar{\Delta^1 y}_t$	$(\Delta^1 y_t - \bar{\Delta^1 y}_t)$	$(\Delta^1 y_t - \bar{\Delta^1 y}_t)^2$
4	7,667	-3,667	13,447
7		-0,667	0,445
8		0,333	0,111
11		3,333	11,109
-6		-13,667	186,787
41		33,333	1111,089
-1		-8,667	75,117
10		2,333	5,443
-5		-12,667	160,453
			$\sum = 1564,001$

$$\text{Получаем: } S_1^2 = \frac{\sum_{i=1}^9 (\Delta^1 y_i - \bar{\Delta^1 y}_i)^2}{(10-1-1) \cdot 2} = \frac{1564,001}{16} = 97,75.$$

Проведем сравнение дисперсий разностей нулевого и первого порядков ($m = 1$; $m - 1 = 0$).

Из того, что $S_0^2 = 925,778$ и $S_1^2 = 97,75$, следует: $S_{m-1}^2 > S_m^2$, и статистику Фишера F_1 вычислим по формуле:

$$F_1 = \frac{S_0^2}{S_1^2} = \frac{925,778}{97,75} = 9,471.$$

При этом $k_1 = n - m = 10 - 1 = 9$; $k_2 = n - m - 1 = 8$. Соответствующее значение $F_{kp} \approx 3$. Получилось, что $F_1 = 9,471 > F_{kp} = 3$. Это означает, что сравнение дисперсий разностей необходимо продолжить. Для этого нужно вычислить очередную дисперсию S_2^2 .

Дисперсию разностей второго порядка ($m = 2$) найдем по формуле:

$$S_m^2 = \frac{\sum_{i=1}^{n-m} (\Delta^m y_i - \bar{\Delta^m y}_i)^2}{(n-m-1) \cdot (2m)!} \cdot (m!)^2 \Rightarrow S_2^2 = \frac{\sum_{i=1}^8 (\Delta^2 y_i - \bar{\Delta^2 y}_i)^2}{(10-2-1) \cdot (2 \cdot 2)!} \cdot (2!)^2$$

Составим таблицу промежуточных вычислений (таблица 3.29).

Таблица 3.29

Таблица промежуточных вычислений для отыскания дисперсии разностей второго порядка

$\Delta^2 y_i$	$\bar{\Delta^2 y}_i$	$(\Delta^2 y_i - \bar{\Delta^2 y}_i)$	$(\Delta^2 y_i - \bar{\Delta^2 y}_i)^2$
3	-1,125	4,125	17,016
1		2,125	4,516
3		4,125	17,016
-17		-15,875	252,016
47		48,125	2316,016
-42		-40,875	1670,766
11		12,125	147,016
-15		-13,875	192,516
			$\sum = 4616,878$

Следовательно,

$$S_2^2 = \frac{\sum_{i=1}^8 (\Delta^2 y_i - \bar{\Delta^2 y}_i)^2}{(10-2-1) \cdot (2 \cdot 2)!} \cdot (2!)^2 = \frac{4616,878}{7 \cdot 1 \cdot 2 \cdot 3 \cdot 4} \cdot (1 \cdot 2)^2 = \frac{4616,878}{42} = 109,926.$$

Сравним дисперсии первого и второго порядков ($m = 2$; $m - 1 = 1$). Из того, что $S_1^2 = 97,75$ и $S_2^2 = 109,926$ следует $S_{m-1}^2 < S_m^2$, и статистику Фишера F_2 вычислим по формуле:

$$F_2 = \frac{S_2^2}{S_1^2} = \frac{109,926}{97,75} = 1,124.$$

При этом $k_1 = n - m = 10 - 2 = 8$; $k_2 = n - m - 1 = 7$. Соответствующее значение $F_{kp} \approx 3,3$. Получилось, что $F_2 = 1,124 < F_{kp} = 3,3$, то есть S_2^2 незначительно отличается от S_1^2 . Это означает, что сравнение дисперсий разностей можно прекратить и степень полиномиального тренда $p = m - 1 = 2 - 1 = 1$, то есть $p = 1$.

Следовательно, искомое уравнение тренда в общем виде выглядит следующим образом: $\hat{y}_t = b t + a$.

Решение к заданию б

Вычислим коэффициенты полиномиального тренда путем решения соответствующей системы уравнений (используя МНК).

Так как $p = 1$, коэффициенты тренда оценим по решению системы:

$$\begin{cases} b \sum_{t=1}^{10} t^2 + a \sum_{t=1}^{10} t = \sum_{t=1}^{10} t \cdot y_t, \\ b \sum_{t=1}^{10} t + a n = \sum_{t=1}^{10} y_t. \end{cases}$$

Учитывая, что:

$$\sum_{t=1}^{10} t = 1 + 2 + 3 + 4 + 5 + 6 + 7 + 8 + 9 + 10 = 55,$$

$$\sum_{t=1}^{10} t^2 = 1 + 4 + 9 + 16 + 25 + 36 + 49 + 64 + 81 + 100 = 385,$$

$$\sum_{t=1}^{10} y_t = 51 + 55 + 62 + 70 + 81 + 75 + 116 + 115 + 125 + 120 = 870,$$

$$\sum_{t=1}^{10} t \cdot y_t = 1 \cdot 51 + 2 \cdot 55 + 3 \cdot 62 + 4 \cdot 70 + 5 \cdot 81 + 6 \cdot 75 + 7 \cdot 116 + 8 \cdot 115 + 9 \cdot 125 + 10 \cdot 120 = 5539,$$

получаем: $\begin{cases} 385b + 55a = 5539, \\ 55b + 10a = 870. \end{cases}$

Решим систему методом Крамера:

$$\Delta = \begin{vmatrix} 385 & 55 \\ 55 & 10 \end{vmatrix} = 3850 - 3025 = 825; \quad \Delta_b = \begin{vmatrix} 5539 & 55 \\ 870 & 10 \end{vmatrix} = 55390 - 47850 = 7540;$$

$$\Delta_a = \begin{vmatrix} 385 & 5539 \\ 55 & 870 \end{vmatrix} = 334950 - 304645 = 30305.$$

$$b = \frac{\Delta_b}{\Delta} = \frac{7540}{825} = 9,139; \quad a = \frac{\Delta_a}{\Delta} = \frac{30305}{825} = 36,733.$$

Решение к заданию 7

Сравнение эмпирических коэффициентов b и a с некоторыми теоретически ожидаемыми значениями β и α этих коэффициентов может быть осуществлено по схеме статистической проверки гипотез.

Для проверки гипотез:

$$\begin{array}{ll} H_0 : b = \beta, & H_0 : a = \alpha, \\ H_1 : b \neq \beta & H_1 : a \neq \alpha \end{array}$$

используется статистика Стьюдента: $T = \frac{(b - \beta)}{S_b}$ и $T = \frac{(a - \alpha)}{S_a}$ соответственно.

Функции S_b и S_a представляют собой стандартные ошибки эмпирических коэффициентов b и a . Находят эти ошибки как корни квадратные из соответствующих необъясненных дисперсий:

$$S_b = \sqrt{S_b^2} \quad \text{и} \quad S_a = \sqrt{S_a^2}.$$

Необъясненные дисперсии коэффициентов в общем виде записываются так:

$$S_b^2 = \frac{\sum e_i^2}{n(n-2)(\bar{x}^2 - \bar{x}^2)} = \frac{\sum (y_i - a - bx_i)^2}{n(n-2)(\bar{x}^2 - \bar{x}^2)},$$

$$S_a^2 = \bar{x}^2 S_b^2.$$

Учитывая обозначения переменных в нашей задаче, выражения для необъясненных дисперсий коэффициентов можно переписать в таком виде:

$$S_b^2 = \frac{\sum (y_i - a - bt)^2}{n(n-2)(t^2 - t^2)},$$

$$S_a^2 = t^2 S_b^2.$$

При справедливости H_0 Т-статистика имеет распределение Стьюдента с числом степеней свободы $v = n - 2$, где n – объем выборки.

Предположение $H_0: b = \beta$ или $a = \alpha$ отклоняется на основании данного критерия, если:

$$|T_{\text{набл}}| = \left| \frac{b - \beta}{S_b} \right| \geq T_{\frac{\chi}{2}, n-2}, \quad \text{или} \quad \left| T_{\text{набл}} \right| = \left| \frac{a - \alpha}{S_a} \right| \geq T_{\frac{\chi}{2}, n-2},$$

где $T_{\frac{\chi}{2}, n-2}$ – критическая статистика $T_{\text{кр}}$, χ – требуемый уровень значимости

(в нашей задаче $\chi = 0,05$).

При невыполнении вышезаписанного условия считается, что нет оснований для отклонения гипотезы H_0 .

На начальном этапе статистического анализа построенной модели наиболее важной является задача проверки справедливости установленной

линейной зависимости между объясняющей и объясняемой переменными.

Эта проблема может быть решена по рассмотренной выше схеме, но проверяемые гипотезы выглядят несколько иначе.

1) Для коэффициента b :

$$H_0: b = 0,$$

$$H_1: b \neq 0.$$

Гипотеза в такой постановке обычно называется гипотезой о статистической значимости вычисленного коэффициента.

Если H_0 принимается, то говорят, что коэффициент b статистически незначим. При отклонении H_0 коэффициент b считается статистически значимым, что указывает на наличие определенной линейной зависимости между объясняющей и объясняемой переменными.

Поскольку предполагается, что $\beta = 0$ (из того, что полагается $b = 0$), то значимость коэффициента регрессии b проверяется по формуле:

$$T = \frac{|b - \beta|}{S_b} = \frac{|b|}{S_b} = \frac{|b|}{\sqrt{S_b^2}}.$$

то есть с помощью анализа отношения величины коэффициента b к его стандартной ошибке.

По аналогичной схеме на основе Т-статистики проверяется гипотеза о статистической значимости коэффициента a :

$$T = \frac{|a - \alpha|}{S_a} = \frac{|a|}{S_a} = \frac{|a|}{\sqrt{S_a^2}}.$$

Отметим, что более важным является анализ статистической значимости коэффициента b , так как именно в нем скрыто влияние объясняющей переменной на зависимую переменную.

Подготовим таблицу промежуточных вычислений для нахождения стандартной ошибки коэффициента b (таблица 3.30).

Таблица 3.30

Промежуточные вычисления для нахождения стандартной ошибки коэффициента b

t	\bar{t}	\bar{t}^2	t^2	\bar{t}^2	$t^2 - \bar{t}^2$	y_t	$\bar{y}_t = a + bt$	$e_t = y_t - \bar{y}_t$	$e_t^2 = (y_t - \bar{y}_t)^2$
1	5,5	30,25	1	38,5	8,25	51	45,872	5,128	26,296
2				4		55	55,011	-0,011	0,0001
3				9		62	64,15	-2,15	4,622
4				16		70	73,289	-3,289	10,817
5				25		81	82,428	-1,428	2,039
6				36		75	91,567	-16,567	274,465
7				49		116	100,706	15,294	233,906
8				64		115	109,845	5,155	26,574
9				81		125	118,984	6,016	36,192
10				100		120	128,123	-8,123	65,983
									$\sum e_t^2 = 680,894$

Необъясненная дисперсия коэффициента b будет равна:

$$S_b^2 = \frac{\sum (y_i - a - bt)^2}{n(n-2)(t^2 - t_{\text{набл}}^2)} = \frac{680,894}{10 \cdot 8 \cdot 8,25} = \frac{680,894}{660} = 1,032.$$

Стандартная ошибка коэффициента b будет равна:

$$S_b = \sqrt{S_b^2} = \sqrt{1,032} = 1,016.$$

Тогда наблюдаемая статистика для b будет равна:

$$T_{\text{набл}} = \frac{|b|}{S_b} = \frac{|9,139|}{1,016} = 8,995.$$

По таблице распределения Стьюдента при уровне значимости $\alpha = 0,05$ и числе степеней свободы $v = n - 2$ находим соответствующее критическое значение Т-статистики для коэффициента b :

$$T_{\text{кр}} = T_{\frac{\chi^2}{2}, n-2} = T_{0,025; 8} = 2,306.$$

Видим, что для коэффициента b $T_{\text{набл}} = 8,995 > T_{\text{кр}} = 2,306$.

Вывод: гипотеза H_0 отклоняется, коэффициент b статистически значим; между переменными y_i и t скорее всего существует линейная зависимость и степень полиномиального тренда ($p = 1$) мы нашли верно.

2) Для коэффициента a :

$$H_0: a = 0,$$

$$H_1: a \neq 0.$$

Поскольку предполагается, что $\alpha = 0$ (из того, что полагается $a = 0$), то значимость коэффициента регрессии a проверяется по формуле:

$$T = \frac{|a - a_{\text{набл}}|}{S_a} = \frac{|a|}{S_a} = \frac{|a|}{\sqrt{S_a^2}},$$

то есть с помощью анализа отношения величины коэффициента a к его стандартной ошибке. Промежуточных вычислений в этом случае делать не нужно, так как S_b^2 и t^2 мы уже знаем.

Найдем необъясненную дисперсию коэффициента a :

$$S_a^2 = t^2 S_b^2 = 38,5 \cdot 1,032 = 39,732.$$

Стандартная ошибка коэффициента a будет равна:

$$S_a = \sqrt{S_a^2} = \sqrt{39,732} = 6,303.$$

Тогда наблюдаемая статистика для a будет равна:

$$T_{\text{набл}} = \frac{|a|}{S_a} = \frac{36,733}{6,303} = 5,828.$$

Критическое значение Т-статистики для коэффициента a будет таким же, как для коэффициента b (те же число степеней свободы и уровень значимости), то есть 2,306.

Видим, что и для коэффициента a $T_{\text{набл}} = 5,828 > T_{\text{кр}} = 2,306$.

Вывод: и в этом случае гипотеза H_0 отклоняется, коэффициент a статистически значим, то есть отбрасывать свободный член в уравнении тренда не стоит, с ним необходимо считаться.

Решение к заданию 8

По вычисленным коэффициентам записываем линейное (поскольку $p = 1$) уравнение тренда: $\hat{y}_t = 9,139t + 36,733$.

По полученному уравнению для построения графика тренда вычислим трендовые значения объясняемой переменной (таблица 3.31).

Таблица 3.31.

Трендовые значения объясняемой переменной

t	1	2	3	4	5	6	7	8	9	10
\hat{y}_t	45,872	55,011	64,15	73,289	82,428	91,567	100,706	109,845	118,984	128,123

Восстанавливаем график эмпирического временного ряда и в тех же осях t , y_t (\hat{y}_t) строим график тренда (рис. 3.4).

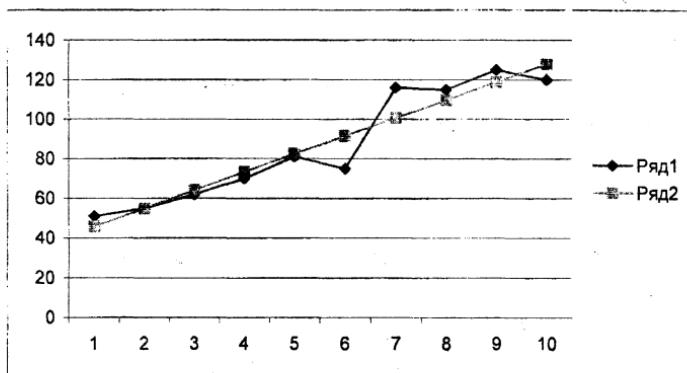


Рис. 3.4. График эмпирического временного ряда и его тенденции (тренда)

Решение к заданию 9

Проведем оценку качества трендовой модели в целом. Для этого необходимо рассмотреть ряд остатков – разностей значений ряда и значений тренда

$$e_t = y_t - \hat{y}_t.$$

Впервые ряд остатков составлен в таблице 3.30, предназначеннной для вычисления стандартной ошибки коэффициента b .

Для оценки качества трендовой модели проверяют следующие гипотезы:

a) о случайности ряда остатков методом поворотных точек (поворотная точка – точка экстремума, то есть точка, в которой значение величины больше или меньше, чем значения в соседних точках); число таких точек d можно найти по графику ряда остатков. Среднее число точек поворота для случайного ряда \bar{d} и их дисперсия S_d^2 равны:

$$\bar{d} = \frac{(2n - 4)}{3}, \quad S_d^2 = \frac{(16n - 29)}{90}.$$

Затем вычисляют статистику $Z = \frac{|d - \bar{d}|}{S_d}$, и если $Z < 1,96$, то гипотеза о

случайности ряда остатков принимается и на уровне значимости 5% можно сделать вывод о том, что тренд существует;

б) о равенстве математического ожидания ряда остатков нулю по статистике

$$T = \frac{\bar{e}_t \sqrt{n}}{S_e},$$

где \bar{e}_t – среднее значение ряда остатков, S_e – среднее квадратическое отклонение ряда остатков; на 5% уровне значимости вычисленное значение Т сравнивается с критическим значением T_{kp} , взятым из таблицы 3.32 с $(n - 1)$ степенями свободы.

Таблица 3.32

Критические значения распределения Стьюдента

k	3	5	7	10	13	16	20	30	∞
T_{kp}	3,18	2,57	2,45	2,23	2,16	2,12	2,09	2,04	1,96

Если вычисленное значение статистики окажется меньше критического ее значения ($T < T_{kp}$), то гипотеза о равенстве математического ожидания ряда остатков нулю принимается и модель на уровне значимости 5% считается адекватной;

в) об отсутствии автокорреляции ряда остатков; при этом используется критерий Дарбина-Уотсона со статистикой DW:

$$DW = \frac{\sum_{i=1}^n (e_i - e_{i-1})^2}{\sum_{i=1}^n e_i^2}.$$

Если $DW \in [2; 4]$, следует использовать вспомогательную статистику $DW' = 4 - DW$.

Расчетное значение DW (или DW') сравнивается с верхним D₂ и нижним D₁ критическими значениями статистики DW, представленными в таблице 3.33, для различной длины ряда n и числа определяемых параметров k на уровне значимости 5% (0,05).

Таблица 3.33

Критические значения статистики DW для различной длины ряда n и числа определяемых параметров k на уровне значимости 5%

n	K=1		K=2		K=3	
	D ₁	D ₂	D ₁	D ₂	D ₁	D ₂
10	0,98	1,34	0,94	1,52	0,67	1,72
15	1,08	1,36	0,95	1,54	0,82	1,75
20	1,35	1,49	1,28	1,57	1,21	1,65

Если расчетное значение критерия DW больше верхнего табличного значения D₂, то гипотеза о независимости уровней остаточной последовательности, то есть об отсутствии в ней автокорреляции, принимается.

Если значение DW меньше нижнего табличного значения D₁, то эта гипотеза отвергается и модель считается неадекватной.

Если же значение D находится между значениями D₂ и D₁, включая сами эти значения, то считается, что нет достаточных оснований сделать тот или иной вывод и необходимы дальнейшие исследования, например, по большей выборке данных.

г) о возможности осуществления оценки соответствия тренда статистическим данным с помощью коэффициента детерминации; известно, что *коэффициент детерминации R²* является суммарной мерой общего качества уравнения регрессии (его соответствия статистическим данным).

В общем случае коэффициент детерминации рассчитывается по формуле:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y}_i)^2}.$$

Для этого коэффициента в общем случае справедливо соотношение $0 \leq R^2 \leq 1$. Чем слабее линейная связь между X и Y, тем R² ближе к нулю, и чем эта связь значительнее, тем ближе R² к единице.

Коэффициент детерминации является мерой, позволяющей определить, в какой степени найденная прямая регрессии дает лучший результат для объяснения поведения зависимой переменной Y, чем горизонтальная прямая $Y = \bar{y}$.

Трендовая модель считается адекватной, если подтверждены все четыре гипотезы (а,б,в,г).

Решение рассматриваемого задания начнем с составления таблицы 3.34 (в этой таблице первые пять столбцов перенесены из таблицы 3.33).

Таблица 3.34

Таблица трендовых значений \hat{y}_t и значений остатков ряда e_t

t	y _t	$\hat{y}_t = 9,139t + 36,733$	$e_t = y_t - \hat{y}_t$	e_t^2	$\sum_{t=1}^{10} e_t$	\bar{e}_t	$e_t - \bar{e}_t$	$(e_t - \bar{e}_t)^2$
1	51	45,872	5,128	26,296	0,025	0,0025	5,125	26,266
2	55	55,011	-0,011	0,00012			-0,013	0,00017
3	62	64,15	-2,15	4,622			-2,152	4,631
4	70	73,289	-3,289	10,817			-3,291	10,83
5	81	82,428	-1,428	2,039			-1,430	2,045
6	75	91,567	-16,567	274,465			-16,570	274,565
7	116	100,706	15,294	233,906			15,291	233,815
8	115	109,845	5,155	26,574			5,152	26,543
9	125	118,984	6,016	36,192			6,013	36,156
10	120	128,123	-8,123	65,983			-8,125	66,016
$\Sigma = 680,894$							$\Sigma = 680,867$	

а) Проверяем гипотезу о случайности ряда остатков:

$$\bar{d} = \frac{2n-4}{3} = \frac{20-4}{3} = 5,33; S_d^2 = \frac{16n-29}{90} = \frac{160-29}{90} = 1,45.$$

По графику ряда остатков можно убедиться, что реальное число точек поворота $d = 3$.

$$\text{Статистика } Z = \frac{|d - \bar{d}|}{S_d} = \frac{|3 - 5,33|}{\sqrt{S_d^2}} = \frac{2,33}{1,204} = 1,935.$$

Так как $Z < 1,96$, то гипотеза о случайности ряда остатков принимается.

б) Проверяем гипотезу о равенстве нулю математического ожидания остатков ряда по статистике $T = \frac{\bar{e}_t \sqrt{n}}{S_e}$.

Функция S_e - среднее квадратическое отклонение в ряду остатков. Оно равно корню квадратному из дисперсии ряда остатков S_e^2 :

$$S_e = \sqrt{\frac{1}{n-1} \sum_{t=1}^n (e_t - \bar{e})^2} = \sqrt{\frac{1}{9} \sum_{t=1}^{10} (e_t - \bar{e}_t)^2} = \frac{1}{3} \sqrt{680,867} = \frac{26,093}{3} = 8,698.$$

Тогда

$$T = \frac{\bar{e}_t \sqrt{n}}{S_e} = \frac{0,0025 \cdot \sqrt{10}}{8,698} = \frac{0,0025 \cdot 3,162}{8,698} = \frac{0,0079}{8,698} = 0,001.$$

Для числа степеней свободы $v = n - 1 = 9$ получаем по таблице 3.32 критическое значение статистики: $T_{kp} \approx 2,23$. Так как найденное значение статистики $T = 0,001 < T_{kp} = 2,23$, приходим к выводу о принятии гипотезы о равенстве нулю математического ожидания ряда остатков.

в) Проверим гипотезу об отсутствии автокорреляции ряда остатков. Воспользуемся статистикой Дарбина-Уотсона (DW):

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

Для ее нахождения составим таблицу 3.35 (в этой таблице первые два столбца составлены на основе четвертого столбца таблицы 3.34).

Таблица 3.35

Таблица промежуточных вычислений для нахождения статистики DW

e_t	e_{t-1}	$(e_t - e_{t-1})$	$(e_t - e_{t-1})^2$
-0,011	5,128	-5,139	26,409
-2,15	-0,011	-2,139	4,575
-3,289	-2,15	-1,139	1,297
-1,428	-3,289	1,861	3,463
-16,567	-1,428	-15,139	229,189
15,294	-16,567	31,861	1015,123
5,155	15,294	-10,139	102,799
6,016	5,155	0,861	0,741
-8,123	6,016	-15,139	229,189
			$\sum = 1612,785$

Получаем:

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2} = \frac{\sum_{t=2}^{10} (e_t - e_{t-1})^2}{\sum_{t=1}^{10} e_t^2} = \frac{1612,785}{680,894} = 2,369.$$

Так как найденное значение статистики $DW \in [2:4]$, будем использовать вспомогательную статистику $DW' = 4 - DW = 4 - 2,369 = 1,631$.

По таблице 3.33 находим верхнее $D_2 = 1,52$ и нижнее $D_1 = 0,94$ табличные значения критической статистики DW_{kp} для длины ряда $n = 10$, числа определяемых параметров $K = 2$ на уровне значимости 5% (0,05).

Так как $DW' = 1,631 > D_2 = 1,52$, то есть вспомогательная статистика больше верхнего значения критической статистики DW'_{kp} , то гипотеза о независимости уровней остаточной последовательности, то есть об отсутствии в ней автокорреляции, принимается.

г) Для расчета коэффициента детерминации нам необходимы значения $\sum_{t=1}^n (y_t - \bar{y}_t)^2$ и $\sum_{t=1}^n e_t^2$. Эти данные есть в ранее составленных таблицах. В результате получаем:

$$R^2 = 1 - \frac{\sum_{t=1}^{10} e_t^2}{\sum_{t=1}^{10} (y_t - \bar{y}_t)^2} = 1 - \frac{680,894}{4671,378} = 1 - 0,146 = 0,854.$$

Столь высокое значение R^2 свидетельствует о соответствии уравнения тренда статистическим данным.

Все четыре гипотезы при проверке качества трендовой модели подтвердились, что говорит об адекватности построенного уравнения.

Решение к заданию 10

Осуществим выбор модели, имеющей наибольший процент достоверных прогнозов для краткосрочного прогнозирования (на один шаг вперед).

Для краткосрочного прогнозирования используют следующие модели:

2) прогноз по одному последнему значению

$$y_{n+1}^{(1)} = y_n;$$

3) прогноз по двум последним значениям

$$y_{n+1}^{(2)} = 2y_n - y_{n-1};$$

4) прогноз по трем последним значениям

$$y_{n+1}^{(3)} = \frac{4y_n + y_{n-1} - 2y_{n-2}}{3};$$

5) прогноз по четырем последним значениям

$$y_{n+1}^{(4)} = \frac{2y_n + y_{n-1} - y_{n-3}}{2};$$

6) прогноз по пяти последним значениям

$$y_{n+1}^{(5)} = \frac{8y_n + 5y_{n-1} + 2y_{n-3} - y_{n-5} - 4y_{n-4}}{10}.$$

Для выбора модели необходимо по каждому варианту вычислить абсолютные погрешности прогнозных значений членов ряда по формуле:

$$\Delta_k = |y_{n+1}^{(k)} - y_{n+1}|.$$

Оценим надежность первой прогнозной модели. Для этого составим таблицу 3.36.

Таблица 3.36

Оценка качества первой прогнозной модели ($\Delta_{kp} = 5$)

y_n	y_{n+1}	$y_{n+1}^{(1)}$	$\Delta_1 = y_{n+1}^{(1)} - y_{n+1} $	K
51	55	51	4	K ⁺
55	62	55	7	K ⁻
62	70	62	8	K ⁻
70	81	70	11	K ⁻
81	75	81	6	K ⁻
75	116	75	41	K ⁻
116	115	116	1	K ⁺
115	125	115	10	K ⁻

125	120	125	5	K
120				
				$\delta_1 = \frac{\sum K^+}{\sum K} \cdot 100\% = \frac{2}{9} \cdot 100\% = 20\%.$

Оценим надежность второй прогнозной модели. Для этого составим таблицу 3.37.

Таблица 3.37

Оценка качества второй прогнозной модели

y _n	y _{n+1}	y _{n+1} ⁽²⁾ = 2y _n - y _{n-1}	Δ ₂ = y _{n+1} ⁽²⁾ - y _{n+1}	K
51	55			
55	62	59	3	K ⁺
62	70	69	1	K [*]
70	81	78	3	K [*]
81	75	92	17	K [*]
75	116	69	47	K [*]
116	115	157	42	K [*]
115	125	114	11	K [*]
125	120	135	15	K [*]
120				
				$\delta_2 = \frac{\sum K^+}{\sum K} \cdot 100\% = \frac{3}{8} \cdot 100\% = 37,5\%.$

Оценим надежность третьей прогнозной модели. Для этого составим таблицу 3.38.

Таблица 3.38

Оценка качества третьей прогнозной модели

y _n	y _{n+1}	y _{n+1} ⁽³⁾ = (4y _n + y _{n-1} - 2y _{n-2}) / 3	Δ ₃ = y _{n+1} ⁽³⁾ - y _{n+1}	K
51	55			
55	62			
62	70	67	3	K ⁺
70	81	77,3	3,7	K [*]
81	75	90	15	K [*]
75	116	80,3	35,7	K [*]
116	115	125,7	10,7	K [*]
115	125	142	17	K [*]
125	120	127,7	7,7	K [*]
120				
				$\delta_3 = \frac{\sum K^+}{\sum K} \cdot 100\% = \frac{3}{7} \cdot 100\% = 43\%.$

Оценим надежность четвертой прогнозной модели. Для этого составим таблицу 3.39.

Таблица 3.39

Оценка качества четвертой прогнозной модели

y_n	y_{n+1}	$y_{n+1}^{(4)} = (2y_n + y_{n-1} - y_{n-3}) / 2$	$\Delta_4 = y_{n+1}^{(4)} - y_{n+1} $	K
51	55			
55	62			
62	70			
70	81	73,5	7,5	K ⁺
81	75	85	10	K ⁻
75	116	80,5	35,5	K ⁻
116	115	113	2	K ⁺
115	125	135,5	10,5	K ⁻
125	120	124,5	4,5	K ⁺
120				
				$\delta_4 = \frac{\sum K^+}{\sum K} \cdot 100\% = \frac{3}{6} \cdot 100\% = 50\%$

Оценим надежность пятой прогнозной модели. Для этого составим таблицу 3.40.

Таблица 3.40

y_n	y_{n+1}	$y_{n+1}^{(5)} = (8y_n + 5y_{n-1} + 2y_{n-2} - y_{n-3} - 4y_{n-4}) / 10$	$\Delta_4 = y_{n+1}^{(4)} - y_{n+1} $	K
51	55			
55	62			
62	70			
70	81			
81	75	86,3	11,3	K ⁻
75	116	87,9	28,1	K ⁻
116	115	114,7	0,3	K ⁺
115	125	128,9	3,9	K ⁺
125	120	140,8	20,8	K ⁻
120				
				$\delta_5 = \frac{\sum K^+}{\sum K} \cdot 100\% = \frac{2}{5} \cdot 100\% = 40\%$

Лучшей оказалась четвертая модель. Для этой модели при $\Delta_{kp} = 5$ достоверность прогноза составляет $\delta = 50\%$. Именно эту модель мы будем использовать для краткосрочного прогноза.

Осуществим краткосрочный прогноз (на один шаг-месяц вперед), то есть вычислим объем производства продукции в ноябре 2002 (11-й месяц):

$$y_{n+1}^{(4)} = \frac{2y_n + y_{n-1} - y_{n-3}}{2} \Rightarrow y_{11} = \frac{2y_{10} + y_9 - y_7}{2} = \frac{240 + 125 - 116}{2} = 124,5 \text{ (млн.руб.)}$$

Долговременный прогноз на число шагов k вперед производят на основе уравнений тренда:

- $\tilde{y}_{n+k} = a(n+k) + b$ (если тренд линейный);
- $\tilde{y}_{n+k} = a(n+k)^2 + b(n+k) + c$ (если тренд параболический).

В нашем случае тренд линейный. Осуществим долговременный прогноз на три месяца вперед ($k = 3$), то есть вычислим объем производства продукции в феврале месяце 2003 года:

$$\hat{y}_{n+k} = b(n+k) + a = 9,139(10+3) + 36,733 = 155,54 \text{ (млн.руб.)}.$$

Решение к заданию 11

Проделанная работа позволила получить закон изменения объема продукции во времени и достаточно уверенно прогнозировать это изменение.

Задачи для самостоятельного решения

Задача 3-А. В табл. 3.41 приводятся данные об уровне дивидендов, выплачиваемых по обыкновенным акциям ($Y, \%$), и среднегодовой стоимости основных фондов компании (X , млн руб.) в сопоставимых ценах за последние девять лет.

Таблица 3.41

Показатель	1	2	3	4	5	6	7	8	9
X	72	75	77	77	79	80	78	79	80
y	4,2	3,0	2,4	2,0	1,9	1,7	1,8	1,6	1,7

Задание

1. Определите параметры уравнения регрессии по первым разностям и дайте их интерпретацию. В качестве зависимой переменной используйте показатель дивидендов по обыкновенным акциям.

2. В чем состоит причина построения уравнения регрессии по первым разностям, а не по исходным уровням рядов?

Задача 3-В. Изучается зависимость объема продаж бензина (y_i) от динамики потребительских цен (x_i). Полученные за последние 6 кварталов данные представлены в табл. 3.42.

Таблица 3.42

Показатель	1 кв.	2 кв.	3 кв.	4 кв.	5 кв.	6 кв.
Индекс потребительских цен, % к кварталу 1	100	104	112	117	121	126
Средний за день объем продаж бензина в течение квартала, тыс. л	89	83	80	77	75	72

Известно также, что $\sum x_i = 680$, $\sum y_i = 476$, $\sum x_i y_i = 53648$. $\sum x_i^2 = 77566$.

Задание

- Постройте модель зависимости объема продаж бензина от индекса потребительских цен с включением фактора времени.
- Дайте интерпретацию параметров полученной вами модели.

Задача 3-С. Имеются поквартальные данные по розничному товарообороту России в 1995 - 1999 гг. (табл. 3.43).

Таблица 3.43

Номер квартала	Товарооборот, % к предыдущему периоду	Номер квартала	Товарооборот, % к предыдущему периоду
1	100,0	11	98,8
2	93,9	12	101,9
3	96,5	13	113,1
4	101,8	14	98,4
5	107,8	15	97,3
6	96,3	16	102,1
7	95,7	17	97,6
8	98,2	18	83,7
9	104,0	19	84,3
10	99,0	20	88,4

Задание

- Постройте график временного ряда.
- Постройте мультипликативную модель временного ряда.
- Оцените качество модели через показатели средней абсолютной ошибки и среднего относительного отклонения.

Задача 3-Д. Пусть имеется следующий временной ряд (таблица 3.44).

Таблица 3.44

i	1	2	3	4	5	6	7	8
x _i	20	10

Известно, что $\sum x_i = 150$, $\sum x_i^2 = 8100$, $\sum_{i=2}^n x_i x_{i-1} = 7350$.

Задание:

- Определите коэффициент автокорреляции уровней этого ряда первого порядка.
- Установите, включает ли исследуемый временной ряд тенденцию.

Задача 3-Е. Данные об урожайности зерновых в хозяйствах некоторой области представлены в таблице 3.45.

Таблица 3.45

Год	Урожайность, ц/га
1995	10,2
1996	10,7
1997	11,7
1998	13,1
1999	14,9
2000	17,2
2001	20,0
2002	23,2

Задание

- Обоснуйте выбор типа уравнения тренда.
- Рассчитайте параметры уравнения тренда.
- Дайте прогноз урожайности зерновых на следующий год.

Задача 3-Ф. Имеются следующие данные (таблица 3.46) об уровне безработицы y_t (%) за 8 месяцев.

Таблица 3.46

Месяц	1	2	3	4	5	6	7	8
Уровень безработицы y_t	8,8	8,6	8,4	8,1	7,9	7,6	7,4	7,0

Задание

- Определите коэффициенты автокорреляции уровней этого ряда первого и второго порядка.
- Обоснуйте выбор уравнения тренда и определите его параметры.
- Интерпретируйте полученные результаты.

Задача 3-Г. Администрация банка изучает динамику депозитов физических лиц за ряд лет (млн. долл. в сопоставимых ценах). Исходные данные представлены в таблице 3.47.

Таблица 3.47

Время t	1	2	3	4	5	6	7	Сумма
Депозиты x	2	6	7	3	10	12	13	53

Учесть, что $\sum x^2 = 511$.

Задание

- Постройте уравнение линейного тренда и дайте интерпретацию его параметров.
- Определите коэффициент детерминации для линейного тренда.

3. Администрация банка предполагает, что среднегодовой абсолютный прирост депозитов физических лиц составляет не менее 2,5 млн. долл. Подтверждается ли это предположение результатами, которые вы получили?

Задача 3-Н. Изучается динамика потребления мяса в регионе. Для этого были собраны данные об объемах среднедушевого потребления мяса y_i (кг) за 7 месяцев. Предварительная обработка данных путем логарифмирования привела к получению результатов, представленных в таблице 3.48.

Таблица 3.48

Месяц	1	2	3	4	5	6	7
$\ln y_i$	2,10	2,11	2,13	2,17	2,22	2,28	2,31

Задание

1. Постройте уравнение экспоненциального тренда.
2. Дайте интерпретацию его параметров.

Раздел 4.

Множественная регрессия и корреляция

Множественная регрессия - уравнение связи с несколькими независимыми переменными:

$$y = f(x_1, x_2, x_3, \dots, x_p),$$

где y – зависимая переменная (результативный признак);

$x_1, x_2, x_3, \dots, x_p$ – независимые переменные (факторы).

Для построения уравнения множественной регрессии чаще используются следующие функции:

- линейная $y = a + b_1 x_1 + b_2 x_2 + \dots + b_p x_p + \varepsilon;$
- степенная $y = a \cdot x_1^{h_1} \cdot x_2^{h_2} \cdot x_3^{h_3} \cdot \dots \cdot x_p^{h_p} \cdot \varepsilon;$
- экспонента $y = e^{a + h_1 x_1 + h_2 x_2 + \dots + h_p x_p + \varepsilon};$
- гипербола $y = \frac{1}{a + b_1 \cdot x_1 + b_2 \cdot x_2 + \dots + b_p \cdot x_p + \varepsilon}.$

Можно использовать и другие функции, приводимые к линейному виду.

Для оценки параметров уравнения множественной регрессии применяют *метод наименьших квадратов* (МНК). Для линейных уравнений и нелинейных уравнений, приводимых к линейным, строится следующая система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии:

$$\begin{aligned} \sum y &= na + b_1 \sum x_1 + b_2 \sum x_2 + \dots + b_p \sum x_p, \\ \sum yx_1 &= a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \dots + b_p \sum x_p x_1, \\ \dots & \\ \sum yx_p &= a \sum x_p + b_1 \sum x_1 x_p + b_2 \sum x_2 x_p + \dots + b_p \sum x_p^2. \end{aligned}$$

Для ее решения может быть применен метод определителей.

Другой вид уравнения множественной регрессии – уравнение в стандартизованном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + \dots + \beta_p t_{x_p},$$

где $t_y = \frac{y - \bar{y}}{\sigma_y}, \quad t_{x_i} = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$ – стандартизованные переменные;

$\beta_i = \frac{x_i - \bar{x}_i}{\sigma_{x_i}}$ – стандартизованные коэффициенты регрессии.

К уравнению множественной регрессии в стандартизованном масштабе применим МНК. Стандартизованные коэффициенты регрессии (β -

коэффициенты) определяются из следующей системы уравнений:

$$t_{b_i} = \frac{b_i}{m_{b_i}} = \sqrt{F_{x_i}}.$$

Связь коэффициентов множественной регрессии b_i со стандартизованными коэффициентами β_i описывается соотношением:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}.$$

Параметр a определяется как $a = \bar{y} - b_1 \bar{x}_1 - b_2 \bar{x}_2 - \dots - b_p \bar{x}_p$.

Средние коэффициенты эластичности для линейной регрессии рассчитываются по формуле:

$$\bar{\varepsilon}_{yx_j} = b_j \frac{\bar{x}_j}{\bar{y}}.$$

Для расчета частных коэффициентов эластичности применяется следующая формула:

$$\varepsilon_{yx_i} = b_i \frac{x_i}{\bar{y}_{x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_p}}.$$

Тесноту совместного влияния факторов на результат оценивает индекс множественной корреляции:

$$R_{yx_1, x_2, \dots, x_p} = \sqrt{1 - \frac{\sigma_{y_{\text{нест}}}}{\sigma_y^2}}.$$

Значение индекса множественной корреляции лежит в пределах от 0 до 1 и должно быть больше или равно максимальному парному индексу корреляции:

$$R_{yx_1, x_2, \dots, x_p} \geq r_{yx_i} \quad (i = 1, p).$$

Индекс множественной корреляции для уравнения в стандартизованном масштабе можно записать в виде:

$$R_{yx_1, x_2, \dots, x_p} = \sqrt{\sum \beta_{yx_i} r_{yx_i}}.$$

При линейной зависимости коэффициент множественной корреляции можно определить через матрицу парных коэффициентов корреляции:

$$R_{yx_1, x_2, \dots, x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{(1)}}},$$

где

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & \dots & r_{yx_p} \\ r_{yx_1} & 1 & r_{x_1 x_2} & \dots & r_{x_1 x_p} \\ r_{yx_2} & r_{x_2 x_1} & 1 & \dots & r_{x_2 x_p} \\ \dots & \dots & \dots & \dots & \dots \\ r_{yx_p} & r_{x_p x_1} & r_{x_p x_2} & \dots & 1 \end{vmatrix} - \text{определитель матрицы парных коэффициентов корреляции};$$

$$\Delta r_{ii} = \begin{vmatrix} 1 & r_{x_1 x_2} & \dots & r_{x_i x_p} \\ r_{x_2 x_1} & 1 & \dots & r_{x_2 x_p} \\ \dots & \dots & \dots & \dots \\ r_{x_p x_1} & r_{x_p x_2} & \dots & 1 \end{vmatrix}.$$

определитель матрицы межфакторной корреляции.

Частные коэффициенты (или индексы) корреляции, измеряющие влияние на y фактора x_i при неизменном уровне других факторов, можно определить по формуле:

$$r_{yx_1 \cdot x_2 \dots x_{i-1} x_{i+1} \dots x_p} = \sqrt{\frac{1 - R^2_{yx_1 x_2 \dots x_{i-1} x_i \dots x_p}}{1 - R^2_{yx_1 x_2 \dots x_{i-1} x_{i+1} \dots x_p}}}.$$

или по рекуррентной формуле:

$$r_{yx_1 \cdot x_2 \dots x_p} = \frac{r_{yx_1 \cdot x_2 \dots x_{p-1}} - r_{yx_p \cdot x_1 x_2 \dots x_{p-1}} r_{x_p x_1 \cdot x_2 \dots x_{p-1}}}{\sqrt{(1 - r^2_{yx_p \cdot x_1 x_2 \dots x_{p-1}})(1 - r^2_{x_p x_1 \cdot x_2 \dots x_{p-1}})}}.$$

Частные коэффициенты корреляции изменяются в пределах от -1 до 1 .

Качество построенной модели в целом оценивает коэффициент (или индекс) детерминации. *Коэффициент множественной детерминации* рассчитывается как квадрат индекса множественной корреляции ($R^2_{x_1 x_2 \dots x_p}$).

Скорректированный индекс множественной детерминации содержит поправку на число степеней свободы и рассчитывается по формуле:

$$\hat{R}^2 = 1 - (1 - R^2) \frac{(n-1)}{(n-m-1)},$$

где n – число наблюдений;

m – число факторов.

Значимость уравнения множественной регрессии в целом оценивается с помощью F -критерия Фишера:

$$F = \frac{R^2}{1-R} \cdot \frac{n-m-1}{m}.$$

Частный F-критерий оценивает статистическую значимость присутствия каждого из факторов в уравнении. В общем виде для фактора x_i частный F-критерий определяется как:

$$F_{част_{x_i}} = \frac{R^2_{yx_1 \dots x_{i-1} x_{i+1} \dots x_p} - R^2_{yx_1 \dots x_{i-1} x_{i+1} \dots x_p}}{1 - R^2_{yx_1 \dots x_{i-1} x_{i+1} \dots x_p}} \cdot \frac{n-m-1}{1}.$$

Оценка значимости коэффициентов чистой регрессии с помощью t-критерия Стьюдента сводится к вычислению значения:

$$t_{b_i} = \frac{b_i}{m_{b_i}} = \sqrt{F_{x_i}},$$

где m_{b_i} - средняя квадратическая ошибка коэффициента регрессии b_i , она может быть определена по следующей формуле:

$$m_{b_i} = \frac{\sigma_y \cdot \sqrt{1 - R_{x_1 \dots x_p}^2}}{\sigma_{x_i} \cdot \sqrt{1 - R_{x_i, x_1 \dots x_{p-1}}^2}} \cdot \frac{1}{\sqrt{n - m - 1}}.$$

При построении уравнения множественной регрессии может возникнуть проблема мультиколлинеарности факторов, их тесной линейной связи. Считается, что две переменные явно коллинеарны, то есть находятся между собой в линейной зависимости, если $r_{x_i x_j} \geq 0,7$.

По величине парных коэффициентов корреляции обнаруживается лишь явная коллинеарность факторов. Наибольшие трудности в использовании аппарата множественной регрессии возникают при наличии мультиколлинеарности факторов. Чем сильнее мультиколлинеарность факторов, тем менее надежна оценка распределения суммы объясненной вариации по отдельным факторам с помощью метода наименьших квадратов.

Для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами.

Если бы факторы не коррелировали между собой, то матрица парных коэффициентов корреляции между факторами была бы единичной матрицей, поскольку все недиагональные элементы $r_{x_i x_j}$ ($x_i \neq x_j$) были бы равны нулю. Так, для включающего три объясняющих переменных уравнения

$$y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$$

матрица коэффициентов корреляции между факторами имела бы определитель, равный 1:

$$\text{Det } |R| = \begin{vmatrix} r_{x_1 x_1} & r_{x_1 x_2} & r_{x_1 x_3} \\ r_{x_2 x_1} & r_{x_2 x_2} & r_{x_2 x_3} \\ r_{x_3 x_1} & r_{x_3 x_2} & r_{x_3 x_3} \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1,$$

так как $r_{x_1 x_1} = r_{x_2 x_2} = r_{x_3 x_3} = 1$ и $r_{x_1 x_2} = r_{x_1 x_3} = r_{x_2 x_3} = 0$.

Если же, наоборот, между факторами существует полная линейная зависимость и все коэффициенты корреляции равны 1, то определитель такой матрицы равен 0:

$$\text{Det } |R| = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0.$$

Чем ближе к 0 определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И, наоборот, чем ближе к 1 определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

Проверка мультиколлинеарности факторов может быть проведена методом испытания гипотезы о независимости переменных $H_0: \text{Det } |R| = 1$.

Доказано, что величина $\left[n - 1 - \frac{1}{6} (2 \cdot m + 5) \lg \text{Det } R \right]$ имеет приближенное распределение χ^2 с $\left(\frac{1}{2} \cdot n \cdot (n-1) \right)$ степенями свободы. Если фактическое значение χ^2 превосходит табличное (критическое), то есть $\chi^2_{\text{факт}} > \chi^2_{\text{табл. (diff., \alpha)}}$, то гипотеза H_0 отклоняется. Это означает, что $\text{Det } |R| \neq 1$. недиагональные ненулевые коэффициенты корреляции указывают на коллинеарность факторов. Мультиколлинеарность считается доказанной.

Для применения МНК требуется, чтобы дисперсия остатков была *гомоскедастичной*. Это означает, что для каждого значения фактора x_j остатки ε_i имеют одинаковую дисперсию. Если это условие не соблюдается, то имеет место *гетероскедастичность*.

При нарушении гомоскедастичности мы имеем неравенства:

$$\sigma_{\varepsilon_i}^2 \neq \sigma_{\varepsilon_j}^2 \neq \sigma^2, \quad j \neq i.$$

При малом объеме выборки для оценки гетероскедастичности может использоваться метод Готфельда-Квандта. Основная его идея состоит в следующем:

- 1) упорядочение n наблюдений по мере возрастания переменной x ;
- 2) исключение из рассмотрения C центральных наблюдений; при этом $(n - C) / 2 > p$, где p – число оцениваемых параметров;
- 3) разделение совокупности из $(n - C)$ наблюдений на две группы (соответственно с малыми и большими значениями фактора x) и определение по каждой из групп уравнений регрессии;
- 4) определение остаточной суммы квадратов для первой (S_1) и второй (S_2) групп и нахождение их отношения $R = S_1 / S_2$.

При выполнении нулевой гипотезы о гомоскедастичности отношение R будет удовлетворять F -критерию со степенями свободы $((n - C - 2p) / 2)$ для каждой остаточной суммы квадратов. Чем больше величина R превышает табличное значение F -критерия, тем более нарушена предпосылка о равенстве дисперсий остаточных величин.

Уравнения множественной регрессии могут включать в качестве независимых переменных качественные признаки (например, профессия, пол, образование, климатические условия, отдельные регионы и т.д.). Чтобы ввести такие переменные в регрессионную модель, их необходимо упорядочить и присвоить им те или иные значения, то есть качественные переменные преобразовать в количественные.

Такого вида сконструированные переменные принято называть *фиксивными переменными*. Например, включать в модель фактор “пол” в виде

фиктивной переменной можно в следующем виде:

$$Z = \begin{cases} 1 - \text{мужской пол}, \\ 0 - \text{женский пол}. \end{cases}$$

Коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение зависимой переменной при переходе от одной категории (женский пол) к другой (мужской пол) при неизменных значениях остальных параметров. На основе t-критерия Стьюдента делается вывод о значимости влияния фиктивной переменной, существенности расхождения между категориями.

Примеры решения избранных задач

Задача 4.1. Имеются данные о сменной добыче угля на одного рабочего $Y(t)$, мощности пласта X_1 (м) и уровне механизации работ X_2 (%), характеризующие процесс добычи угля в 10 шахтах (таблица 4.1).

Таблица 4.1

i	x_{1i}	x_{2i}	y_i
1	8	5	5
2	11	8	10
3	12	8	10
4	9	5	7
5	8	7	5
6	8	8	6
7	9	6	6
8	9	4	5
9	8	5	6
10	12	7	8

Задание

Предполагая, что между переменными Y, X_1, X_2 существует линейная корреляционная зависимость, найти ее аналитическое выражение (уравнение регрессии Y по X_1 и X_2).

Решение

Включение в регрессионную модель нескольких (в данном случае двух) объясняющих переменных усложняет (по сравнению с парной регрессией) получаемые формулы и вычисления. Матричное описание регрессии облегчает как теоретические концепции анализа, так и необходимые расчетные процедуры. Однако для удобства вычислений составим таблицу 4.2.

Таблица 4.2

i	x_{i1}	x_{i2}	y_i	x_{i1}^2	x_{i2}^2	y_i^2	$x_{i1} x_{i2}$	$y_i x_{i1}$	$y_i x_{i2}$	\hat{y}_i	$e_i^2 = (\hat{y}_i - y_i)^2$
1	8	5	5	64	25	25	40	40	25	5,13	0,016
2	11	8	10	121	64	100	88	110	80	8,79	1,464
3	12	8	10	144	64	100	96	120	80	9,64	1,127
4	9	5	7	81	25	49	45	63	35	5,98	1,038
5	8	7	5	64	49	25	56	40	35	5,86	0,741
6	8	8	6	64	64	36	64	48	48	6,23	0,052
7	9	6	6	81	36	36	54	54	36	6,35	0,121
8	9	4	5	81	16	25	36	45	20	5,61	0,377
9	8	5	6	64	25	36	40	48	30	5,13	0,762
10	12	7	8	144	49	64	84	96	56	9,28	1,631
Σ	94	63	68	908	417	496	603	664	445	-	6,329

Теперь введем матричные обозначения (учтем, что в матрицу плана X вводится дополнительный столбец чисел, состоящий из единиц):

$$Y = \begin{pmatrix} 5 \\ 10 \\ \dots \\ 8 \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 8 & 5 \\ 1 & 11 & 8 \\ \dots & \dots & \dots \\ 1 & 12 & 7 \end{pmatrix}.$$

Далее:

$$XX = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 8 & 11 & \dots & 1 \\ 5 & 8 & \dots & 7 \end{pmatrix} \begin{pmatrix} 1 & 8 & 5 \\ 1 & 11 & 8 \\ \dots & \dots & \dots \\ 1 & 12 & 7 \end{pmatrix} = \begin{pmatrix} 10 & 94 & 63 \\ 94 & 908 & 603 \\ 63 & 603 & 417 \end{pmatrix},$$

(суммы можно увидеть в итоговой строке таблицы 4.2);

$$XY = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 8 & 11 & \dots & 12 \\ 5 & 8 & \dots & 7 \end{pmatrix} \begin{pmatrix} 5 \\ 10 \\ \dots \\ 8 \end{pmatrix} = \begin{pmatrix} 68 \\ 664 \\ 445 \end{pmatrix}.$$

Матрицу $A^{-1} = (X'X)^{-1}$ определим по формуле $X'X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 8 & 11 & \dots & 1 \\ 5 & 8 & \dots & 7 \end{pmatrix}$ опре-

делятель матрицы $X'X$; \bar{A} - матрица, присоединенная к матрице $X'X$. Получим:

$$A^{-1} = \frac{1}{3738} \begin{pmatrix} 15027 & -1209 & -522 \\ -1209 & 201 & -108 \\ -522 & -108 & 244 \end{pmatrix} \text{ (убедитесь в этом сами).}$$

Чтобы получить b , умножим A^{-1} на вектор XY и получим:

$$b = \frac{1}{3738} \begin{pmatrix} -13230 \\ 3192 \\ 1372 \end{pmatrix} = \begin{pmatrix} -3,5393 \\ 0,8539 \\ 0,3670 \end{pmatrix}.$$

Делаем вывод, что уравнение множественной регрессии имеет вид:

$$\hat{y} = -3,54 + 0,854x_1 + 0,367x_2.$$

Оно показывает, что при увеличении только мощности пласта X_1 (при неизменном X_2) на 1 м добыча угля на одного рабочего Y увеличивается в среднем на 0,854 т, а при увеличении только уровня механизации работ X_2 (при неизменном X_1) – в среднем на 0,367 т.

Задача 4.2. По данным предыдущей задачи сравнить раздельное влияние на сменную добычу угля двух факторов – мощности пласта и уровня механизации работ.

Решение

Для сравнения влияния каждой из объясняющих переменных по формуле $b'_j = b_j \frac{s_{x_j}}{s_y}$ вычислим стандартизованные коэффициенты регрессии:

$$b'_1 = 0,8539 \cdot \frac{1,56}{1,83} = 0,728; \quad b'_2 = 0,3670 \cdot \frac{1,42}{1,83} = 0,285,$$

а по формуле $E_j = b_j \frac{\bar{x}_j}{\bar{y}}$ вычислим коэффициенты эластичности (опустим расчет характеристик) $\bar{x}_1 = 9,4; \bar{x}_2 = 6,3; \bar{y} = 6,8; s_{x_1} = 1,56; s_{x_2} = 1,42; s_y = 1,83$:

$$E_1 = 0,8539 \cdot \frac{9,4}{6,8} = 1,180; \quad E_2 = 0,3670 \cdot \frac{6,3}{6,8} = 0,340.$$

Таким образом, увеличение мощности пласта и уровня механизации работ только на одно s_{x_1} или на одно s_{x_2} увеличивает в среднем сменную добычу угля на одного рабочего соответственно на 0,728 s_y или на 0,285 s_y , а увеличение этих переменных на 1% (от своих средних значений) приводит в среднем к росту добычи угля соответственно на 1,18% и 0,34%. По обоим показателям на сменную добычу угля большее влияние оказывает фактор “мощность пласта” по сравнению с фактором “уровень механизации работ”.

Задача 4.3. По данным предыдущих двух задач, оценить сменную добычу угля на одного рабочего для шахт с мощностью пласта 8 м и уровнем механизации 6%; найти 95%-ные доверительные интервалы для индивидуального и среднего значений сменной добычи угля на одного рабочего для таких же шахт. Проверить значимость коэффициентов регрессии и по-

строить для них 95%-ные доверительные интервалы. Найти интервальную оценку для дисперсии σ^2 .

Решение

1. В предыдущих задачах уравнение регрессии было получено в виде:

$$\hat{y} = -3,54 + 0,854x_1 + 0,367x_2.$$

По условию необходимо оценить $M_x(Y)$, где $X'_0 = (1 \ 8 \ 6)$. Выборочной оценкой $M_x(Y)$ является групповая средняя, которую найдем по уравнению регрессии:

$$\hat{y} = -3,54 + 0,854 \cdot 8 + 0,367 \cdot 6 = 5,49(\tau).$$

Для построения доверительного интервала для $M_x(Y)$ необходимо знать дисперсию его оценки - $s_{\hat{y}}^2$. Для ее вычисления обратимся к таблице 4.2 (точнее, к ее двум последним столбцам, при составлении которых учтено, что групповые средние определяются по полученному уравнению регрессии) и формуле несмещенной оценки (или выборочной остаточной дисперсии) s^2 :

$$s^2 = \frac{e'e}{n-p-1} = \frac{\sum_{i=1}^n e_i^2}{n-p-1}.$$

Получаем:

$$s^2 = \frac{6,329}{10-2-1} = 0,904 \quad u \quad s = \sqrt{0,904} = 0,951(\tau).$$

Стандартную ошибку групповой средней \hat{y} найдем по формуле:

$$s_{\hat{y}} = s \sqrt{X'_0 (XX)^{-1} X_0}.$$

Вначале найдем:

$$X'_0 (XX)^{-1} X_0 = (1 \ 8 \ 6) \frac{1}{3738} \begin{pmatrix} 15027 & -1209 & -522 \\ -1209 & 201 & -108 \\ -522 & -108 & 244 \end{pmatrix} \begin{pmatrix} 1 \\ 8 \\ 6 \end{pmatrix} =$$

$$\frac{1}{3738} (2223 \ -249 \ 78) \begin{pmatrix} 1 \\ 8 \\ 6 \end{pmatrix} = \frac{1}{3738} (699) = 0,1870.$$

Следовательно, $s_{\hat{y}} = 0,951 \sqrt{0,1870} = 0,411(\tau)$.

Затем для числа степеней свободы $k = 10 - 2 - 1 = 7$ находим значение t -критерия по таблице Стьюдента: $t_{0,95; 7} = 2,36$.

2. Учитывая, что на случай множественной регрессии доверительный интервал для $M_x(Y)$ записывается по формуле $\hat{y} - t_{1-\alpha; k} s_{\hat{y}} \leq M(Y) \leq \hat{y} + t_{1-\alpha; k} s_{\hat{y}}$ (формула стандартной ошибки $s_{\hat{y}}$ записана выше), для рассматриваемой

задачи доверительный интервал для $M_x(Y)$ равен:

$$5,49 - 2,36 \cdot 0,411 \leq M(Y) \leq 5,49 + 2,36 \cdot 0,411,$$

или

$$4,52 \leq M(Y) \leq 6,46 \text{ (т).}$$

С надежностью 0,95 средняя сменная добыча угля на одного рабочего для шахт с мощностью пласта 8 м и уровнем механизации работ 6% находится в пределах от 4,52 до 6,46 т.

3. Найдем доверительный интервал для индивидуального значения y_0^* при $X'_0 = (1 \ 8 \ 6)$ по формуле:

$$\hat{y}_0 - t_{1-\alpha; n-p-1} s_{\hat{y}_0} \leq y_0^* \leq \hat{y}_0 + t_{1-\alpha; n-p-1} s_{\hat{y}_0},$$

где

$$s_{\hat{y}_0} = s \sqrt{1 + X'_0 (X'X)^{-1} X_0},$$

Получаем:

$$s_{\hat{y}_0} = 0,951 \sqrt{1 + 0,1870} = 1,036 \text{ (т).}$$

$$5,49 - 2,36 \cdot 1,036 \leq y_0^* \leq 5,49 + 2,36 \cdot 1,036, \text{ то есть } 3,05 \leq y_0^* \leq 7,93 \text{ (т),}$$

Итак, с надежностью 0,95 индивидуальное значение сменной добычи угля в шахтах с мощностью пласта 8 м и уровнем механизации работ 6% находится в пределах от 3,05 до 7,93 (т).

4. Проверим значимость коэффициентов регрессии b_1 и b_2 . Ранее были получены их значения: $b_1 = 0,854$ и $b_2 = 0,367$. Стандартная ошибка коэффициента регрессии b_1 находится по формуле:

$$s_{b_1} = s \sqrt{\left[(X'X)^{-1} \right]_{11}},$$

Следовательно, для b_1 имеем:

$$s_{b_1} = 0,951 \sqrt{\frac{1}{3738} \cdot 201} = 0,221.$$

Так как наблюдаемая статистика Стьюдента для коэффициента b_1 больше критического ее значения при 5%-ом уровне значимости и числе степеней свободы 7 ($t_{\text{крит}, 1} = \frac{b_1}{s_{b_1}} = \frac{0,854}{0,221} = 3,81 > t_{0,95; 7} = 2,36$), то коэффициент b_1 статистически значим.

Аналогично действуем в отношении коэффициента b_2 :

$$s_{b_2} = 0,951 \sqrt{\frac{1}{3738} \cdot 244} = 0,243, \quad t_{\text{крит}, 2} = \frac{0,367}{0,243} = 1,51 < t_{0,95; 7} = 2,36,$$

что говорит о статистической незначимости коэффициента b_2 .

Доверительный интервал имеет смысл построить только для значимого коэффициента b_1 по формуле:

$$b_1 - t_{1-\alpha; n-p-1} s_{b_1} \leq b_1 \leq b_1 + t_{1-\alpha; n-p-1} s_{b_1},$$

то есть:

$$0,854 - 2,36 \cdot 0,221 \leq \beta_j \leq 0,854 + 2,36 \cdot 0,221 \quad \text{или} \quad 0,332 \leq \beta_j \leq 1,376.$$

Итак, с надежностью 0,95 за счет изменения на 1 м мощности пласта X_1 (при неизменном X_2) сменная добыча угля на одного рабочего Y будет изменяться в пределах от 0,332 до 1,376 (т).

5. Найдем 95%-ный доверительный интервал для параметра σ^2 . В множественной регрессии он строится аналогично парной модели по формуле $\frac{ns^2}{\chi_{\alpha/2; n-2}^2} \leq \sigma^2 \leq \frac{ns^2}{\chi_{1-\alpha/2; n-2}^2}$ с соответствующим изменением числа степеней свободы критерия Пирсона χ^2 , то есть $\frac{ns^2}{\chi_{\alpha/2; n-p-1}^2} \leq \sigma^2 \leq \frac{ns^2}{\chi_{1-\alpha/2; n-p-1}^2}$.

Для уровня значимости 0,05 и числа степеней свободы $k = n - p - 1 = n - 2 - 1 = n - 3$. По таблице для этого критерия находим:

$$\chi_{\alpha/2; n-3-1}^2 = \chi_{0,025; 7}^2 = 16,01;$$

$$\chi_{1-\alpha/2; n-p-1}^2 = \chi_{0,975; 7}^2 = 1,69.$$

Окончательно получаем:

$$\frac{10 \cdot 0,904}{16,01} \leq \sigma^2 \leq \frac{10 \cdot 0,904}{1,69},$$

$$\text{или} \quad 0,565 \leq \sigma^2 \leq 5,349 \quad \text{и} \quad 0,751 \leq \sigma \leq 2,313.$$

Таким образом, с надежностью 0,95 дисперсия возмущений заключена в пределах от 0,565 до 5,349, а их стандартное отклонение – от 0,751 до 2,313 (т).

Задача 4.4. По 30 территориям России имеются данные, представленные в таблице 4.3.

Таблица 4.3.

Признак	Среднее значение	Среднее квадратическое отклонение	Линейный коэффициент парной корреляции
Среднедневной душевой доход x_1 (руб.)	86,8	11,44	
Среднедневная зарплата одного работающего x_2 (руб.)	54,9	5,86	$r_{x_1 x_2} = 0,8405$
Средний возраст безработного x_3 (лет)	33,5	0,58	$r_{x_1 x_3} = -0,2101$ $r_{x_2 x_3} = -0,1160$

Задание

- Построить уравнение множественной регрессии в стандартизованной и естественной формах.
- Рассчитать частные коэффициенты эластичности, сравнить их с β_1 и β_2 , пояснить различия между ними.
- Рассчитать линейные коэффициенты частной корреляции и коэффициент множественной корреляции; сравнить их с линейными коэффициентами парной корреляции и пояснить различия между ними.
- Рассчитать общий и частный критерий Фишера.

Решение

1) Линейное уравнение множественной регрессии y от x_1 и x_2 имеет вид: $y = a + b_1 \cdot x_1 + b_2 \cdot x_2$. Для расчета его параметров применим метод стандартизации переменных и построим искомое уравнение в стандартизованном масштабе: $t_y = \beta_1 \cdot t_{x_1} + \beta_2 \cdot t_{x_2}$. Расчет β -коэффициентов выполним по формулам:

$$\beta_1 = \frac{r_{yx_1} - r_{yx_2} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2} = \frac{0,8405 - 0,2101 \cdot 0,116}{1 - 0,116^2} = \frac{0,8161}{0,9865} = 0,8273;$$

$$\beta_2 = \frac{r_{yx_2} - r_{yx_1} r_{x_1 x_2}}{1 - r_{x_1 x_2}^2} = \frac{-0,2101 + 0,8405 \cdot 0,116}{1 - 0,116^2} = \frac{-0,1126}{0,9865} = -0,1141.$$

Получим уравнение: $t_y = 0,8273 t_{x_1} - 0,1141 t_{x_2}$.

Для построения уравнения в естественной форме рассчитаем b_1 и b_2 , используя формулы для перехода от β_i к b_i .

Получаем:

$$b_1 = \frac{0,8273 \cdot 11,44}{5,86} = 1,6151; \quad b_2 = \frac{-0,1141 \cdot 11,44}{0,58} = -2,2505.$$

Значение параметра a получим из соотношения:

$$a = \bar{y} - b_1 \cdot \bar{x}_1 - b_2 \cdot \bar{x}_2 = 86,8 - 1,6151 \cdot 54,9 + 2,2505 \cdot 33,5 = -73,52276.$$

$$\hat{y}_{x_1 x_2} = -73,52 + 1,62 \cdot x_1 - 2,25 \cdot x_2.$$

2. Для характеристики относительной силы влияния x_1 и x_2 на y рассчитаем средние коэффициенты эластичности:

$$\overline{\mathcal{E}}_{yx_1} = b_1 \frac{\bar{x}_1}{\bar{y}} \Rightarrow \overline{\mathcal{E}}_{yx_1} = \frac{1,62 \cdot 54,9}{86,8} = 1,0246\%; \quad \overline{\mathcal{E}}_{yx_2} = \frac{-2,25 \cdot 33,5}{86,8} = -0,8684\%.$$

Делаем вывод: с увеличением средней заработной платы x_1 на 1% от ее среднего уровня средний душевой доход y возрастает на 1,02% от своего среднего уровня; при повышении среднего возраста безработного x_2 на 1% среднедушевой доход y снижается на 0,87% от своего среднего уровня. Очевидно, что сила влияния средней заработной платы x_1 на средний ду-

шевой доход у оказалась большей, чем сила влияния среднего возраста безработного x_2 . К аналогичным выводам о силе связи приходим при сравнении модулей значений β_1 и β_2 :

$$|\beta_1| = |0,8273| > |\beta_2| = |-0,1141|.$$

Различия в силе влияния фактора на результат, полученные при сравнении $\bar{\sigma}_{x_1}$ и β_j , объясняются тем, что коэффициент эластичности исходит из соотношения средних: $\bar{\sigma}_{x_1} = b_j \frac{\bar{x}_j}{\bar{y}}$, а β - коэффициент – из соотношения средних квадратических отклонений: $\beta_j = b_j \frac{\sigma_{x_j}}{\sigma_y}$.

3. Линейные коэффициенты частной корреляции здесь рассчитываются по рекуррентной формуле:

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1 x_2}^2)}} = \frac{0,8405 - 0,2101 \cdot 0,116}{\sqrt{(1 - 0,2101^2)(1 - 0,116^2)}} = 0,8404;$$

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}} = \frac{-0,2101 + 0,8405 \cdot 0,116}{\sqrt{(1 - 0,8405^2)(1 - 0,116^2)}} = -0,2092;$$

$$r_{x_1 x_2 \cdot y} = \frac{r_{x_1 x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1 - r_{x_1 x_2}^2)(1 - r_{yx_1}^2)}} = \frac{-0,116 + 0,8405 \cdot 0,2101}{\sqrt{(1 - 0,8405^2)(1 - 0,2101^2)}} = 0,1144.$$

Если сравнить значения коэффициентов парной и частной корреляции, то приходим к выводу, что из-за слабой межфакторной связи ($r_{x_1 x_2} = -0,116$) коэффициенты парной и частной корреляции отличаются незначительно; выводы о тесноте и направлении связи на основе коэффициентов парной и частной корреляции совпадают:

$$r_{yx_1} = 0,8405; \quad r_{yx_2} = -0,2101; \quad r_{x_1 x_2} = -0,1160;$$

$$r_{yx_1 \cdot x_2} = 0,8404; \quad r_{yx_2 \cdot x_1} = -0,2092; \quad r_{x_1 x_2 \cdot y} = 0,1144.$$

Расчет линейного коэффициента множественной корреляции выполним с использованием коэффициентов r_{yx_j} и β_j :

$$R_{yx_1 x_2} = \sqrt{r_{yx_1} \cdot \beta_1 + r_{yx_2} \cdot \beta_2} = \sqrt{0,8405 \cdot 0,8273 + 0,2101 \cdot 0,1141} = \sqrt{0,7193} = 0,8481.$$

Зависимость y от x_1 и x_2 характеризуется как тесная, в которой 72% вариации среднего душевого дохода определяются вариацией учтенных в модели факторов: средней заработной платы и среднего возраста безработного. Прочие факторы, не включенные в модель, составляют соответственно 28% от общей вариации y .

4. Общий F-критерий проверяет гипотезу H_0 о статистической значимости уравнения регрессии и показателя тесноты связи ($R^2 = 0$):

$$F_{\text{набл}} = \frac{R_{yx_1x_2}^2}{1 - R_{yx_1x_2}^2} : \frac{m}{n - m - 1} = \frac{R_{yx_1x_2}^2}{1 - R_{yx_1x_2}^2} \cdot \frac{n - m - 1}{m} = \frac{0,7193^2}{0,2807} \cdot \frac{27}{2} = 34,5.$$

Для уровня значимости $\alpha = 0,05$ соответствующее значение

$$F_{\text{крит}} = 3,4.$$

Сравнивая $F_{\text{крит}}$ и $F_{\text{набл}}$, приходим к выводу о необходимости отклонить гипотезу H_0 , так как $F_{\text{крит}} = 3,4 < F_{\text{набл}} = 34,6$. С вероятностью $1 - \alpha = 0,05$ делаем заключение о статистической значимости уравнения в целом и показателя тесноты связи $R_{yx_1x_2}$, которые сформировались под неслучайным воздействием факторов x_1 и x_2 .

Частные F-критерии (F_{x_1} и F_{x_2}) оценивают статистическую значимость присутствия факторов x_1 и x_2 в уравнении множественной регрессии, оценивают целесообразность включения в уравнение одного фактора после другого фактора, то есть F_{x_1} оценивает целесообразность включения в уравнение фактора x_1 после того, как в него включен фактор x_2 . Соответственно F_{x_2} указывает на целесообразность включения в модель фактора x_2 после фактора x_1 :

$$F_{x_1\text{набл}} = \frac{R_{yx_1x_2}^2 - r_{yx_2}^2}{1 - R_{yx_1x_2}^2} \cdot \frac{n - m - 1}{1} = \frac{0,8481^2 - 0,2101^2}{1 - 0,8481^2} \cdot \frac{27}{1} = 64,9.$$

Для уровня значимости $\alpha = 0,05$ соответствующее значение

$$F_{\text{крит}} = 4,21.$$

Сравнивая $F_{\text{крит}}$ и $F_{x_1\text{набл}}$, приходим к выводу о целесообразности включения в модель фактора x_1 после фактора x_2 , так как $F_{x_1\text{набл}} = 64,9 > F_{\text{крит}} = 4,21$. Нулевую гипотезу о несущественности прироста R^2 за счет включения дополнительного фактора x_1 отклоняем и приходим к выводу о статистически подтвержденной целесообразности включения фактора x_1 после фактора x_2 .

Целесообразность включения в модель фактора x_2 после фактора x_1 проверяет F_{x_2} :

$$F_{x_2\text{набл}} = \frac{R_{yx_1x_2}^2 - r_{yx_1}^2}{1 - R_{yx_1x_2}^2} \cdot \frac{n - m - 1}{1} = \frac{0,8481^2 - 0,8405^2}{1 - 0,8481^2} \cdot \frac{27}{1} = 1,234.$$

Низкое значение $F_{x_2\text{набл}} = 1,234$ по сравнению с $F_{\text{крит}} = 4,21$ позволяет принять нулевую гипотезу и сделать вывод о статистической незначимости прироста R^2 за счет включения в модель фактора x_2 после фактора x_1 , то есть о нецелесообразности включения в модель фактора x_2 (среднего возраста безработного). Это означает, что парная регрессионная модель зави-

симости среднего дохода от средней заработной платы является достаточно значимой, надежной и что нет необходимости улучшать ее, включая дополнительный фактор x_2 .

Задачи для самостоятельного решения

Задача 4-А. По 20 предприятиям отрасли были получены результаты регрессионного анализа зависимости объема выпуска продукции y (млн. руб.) от численности занятых на предприятии x_1 (чел.) и среднегодовой стоимости основных фондов x_2 (млн. руб.). Информация представлена в таблице 4.4.

Таблица 4.4

Коэффициент детерминации	0,81		
Множественный коэффициент корреляции	???		
Уравнение регрессии	$\ln y = ??? + 0,48 \ln x_1 + 0,62 \ln x_2$		
Стандартные ошибки параметров	2	0,06	???
t-критерий для параметров	1,5	???	5

Задание

- Напишите уравнение регрессии, характеризующее зависимость y от x_1 и x_2 .
- Восстановите пропущенные характеристики.
- С вероятностью 0,95 постройте доверительные интервалы для коэффициентов регрессии.
- Проанализируйте результаты регрессионного анализа.

Задача 4-В. По 30 наблюдениям, матрица парных коэффициентов корреляции оказалась следующей (таблица 4.5):

Таблица 4.5

	y	x_1	x_2	x_3
y	1,00			
x_1	0,30	1,00		
x_2	0,60	0,10	1,00	
x_3	0,40	0,15	0,80	1,00

Задание

- Постройте уравнение регрессии в стандартизованном виде и сделайте выводы.
- Определите показатель множественной корреляции (некорректированный и скорректированный).
- Оцените целесообразность включения переменной x_1 в модель после введения в нее переменных x_2 и x_3 .

Задача 4-С. По 25 территориям страны изучается влияние климатических условий на урожайность зерновых у (ц/га). Для этого были отобраны две объясняющие переменные:

x_1 – количество осадков в период вегетации (мм);

x_2 – средняя температура воздуха ($^{\circ}\text{C}$).

Матрица парных коэффициентов корреляции этих показателей имеет вид (таблица 4.6):

Таблица 4.6

	y	x_1	x_2
y	1,0		
x_1	0,6	1,0	
x_2	-0,5	-0,9	1,0

Задание

1. Определите частные коэффициенты корреляции результата с каждым из факторов. Прокомментируйте различие полученных парных и частных коэффициентов корреляции результатов.

2. Исследователь, анализирующий данную зависимость, намерен определить на основе приведенной выше матрицы, какое уравнение регрессии лучше строить: а) парную регрессию y на x_1 ; б) парную линейную регрессию y на x_2 ; в) множественную линейную регрессию. Как бы вы ответили на эти вопросы?

3. Постройте уравнение регрессии в стандартизованном масштабе и сделайте выводы.

Задача 4-Д. Анализируется зависимость объема производства продукции предприятиями отрасли черной металлургии от затрат труда и расхода чугуна. Для этого по 20 предприятиям собраны следующие данные:

y – объем продукции предприятия в среднем за год (млн. руб.);

x_1 – среднегодовая списочная численность рабочих предприятия (чел.);

x_2 – средние затраты чугуна за год (млн. т).

Ниже представлены результаты корреляционного анализа этого массива данных.

Матрица парных коэффициентов корреляции для исходных переменных (таблица 4.7):

Таблица 4.7

	y	x_1	x_2
y	1,00		
x_1	0,78	1,00	
x_2	0,86	0,96	1,00

Матрица парных коэффициентов корреляции для натуральных логарифмов исходных данных (таблица 4.8):

Таблица 4.8

	$\ln y$	$\ln x_1$	$\ln x_2$
$\ln y$	1,00		
$\ln x_1$	0,86	1,00	
$\ln x_2$	0,90	0,69	1,00

Задание

- Поясните смысл приведенных выше коэффициентов.
- Используя эту информацию, опишите ваши предположения относительно: а) знаков коэффициентов регрессии в уравнениях парной линейной регрессии y по x_1 ($y = a + bx_1$) и y по x_2 ($y = a + bx_2$); б) статистической значимости коэффициентов регрессии при переменных x_1 и x_2 в линейном уравнении множественной регрессии и в уравнении множественной регрессии в форме функции Кобба-Дугласа.
- Определите значения коэффициентов детерминации в уравнениях парной линейной регрессии $y = a + bx_1$ и $y = a + bx_2$. Какое из этих уравнений лучше?
- Определите частные коэффициенты корреляции для линейного уравнения множественной регрессии.

Задача 4-Е. По 20 предприятиям легкой промышленности получена информация (таблица 4.9), характеризующая зависимость объема выпуска продукции y (млн. руб.) от количества отработанных за год человекочасов x_1 (чел.ч) и среднегодовой стоимости производственного оборудования x_2 (млн. руб.).

Таблица 4.9

Уравнение регрессии	$y = 35 + 0,06x_1 + 2,5x_2$
Множественный коэффициент корреляции	0,9
Сумма квадратов отклонений расчетных значений результата от фактического	3000

Задание

- Определите коэффициент детерминации в этой модели.
- Составьте таблицу результатов дисперсионного анализа.
- Проанализируйте полученные результаты регрессионного анализа.

Задача 4-Ф. В результате исследования факторов, определяющих экономический рост, по 73 странам получено следующее уравнение регрессии:

$$\hat{G} = 1,4 - 0,52P + 0,17S + 11,16I - 0,38D - 4,75 \ln, \quad R^2 = 0,60,$$

(-5,9) (4,34) (3,91) (-0,79) (-2,7)

где \hat{G} - темпы экономического роста (темперы роста среднедушевого ВВП в % к базисному периоду);

P – реальный среднедушевой ВВП, %;

S – бюджетный дефицит, % к ВВП;

I – объем инвестиций, % к ВВП;

D – внешний долг, % к ВВП;

In – уровень инфляции, %.

В скобках указаны фактические значения t-критерия для коэффициентов множественной регрессии.

Задание

1. Проверьте гипотезу о достоверности полученной модели в целом.

2. До получения результатов этого исследования ваш однокурсник заключил с вами пари, что эмпирические результаты по данной модели доказут наличие обратной связи между темпами экономического роста и объемом внешнего долга страны (% к ВВП). Выиграл ли это пари ваш однокурсник?

Задача 4-G. Для изучения рынка жилья в городе по данным о 46 коттеджах было построено уравнение множественной регрессии:

$$y = 21,1 - 6,2x_1 + 0,95x_2 + 3,57x_3; \quad R^2 = 0,7,$$

(1,8) (0,54) (0,83)

где y – цена объекта, тыс. долл.;

x_1 – расстояние до центра города, км;

x_2 – полезная площадь объекта, кв.м;

x_3 – число этажей в доме, ед.;

R^2 – коэффициент множественной детерминации.

В скобках указаны значения стандартных ошибок для коэффициентов множественной регрессии.

Задание

1. Проверьте гипотезу о том, что коэффициент регрессии b_1 в генеральной совокупности равен нулю.

2. Проверьте гипотезу о том, что коэффициент регрессии b_2 в генеральной совокупности равен нулю.

3. Проверьте гипотезу о том, что коэффициент регрессии b_3 в генеральной совокупности равен нулю.

4. Проверьте гипотезу о том, что коэффициенты регрессии b_1 , b_2 , b_3 в генеральной совокупности одновременно равны нулю (или что коэффициент детерминации равен нулю).

5. Поясните причины расхождения результатов, полученных в п.п. 1, 2 и 3, с результатами, полученными в п. 4.

Задача 4-Н. По 19 предприятиям оптовой торговли изучается зависимость объема реализации (y) от размера торговой площади (x_1) и товарных запасов (x_2). Были получены следующие варианты уравнений регрессии:

$$1. y = 25 + 15x_1 \quad r^2 = 0,90;$$

$$2. y = 42 + 27x_2 \quad r^2 = 0,84;$$

$$3. y = 30 + 10x_1 + 8x_2 \quad R^2 = 0,92; \\ (2,5) \quad (4,0)$$

$$4. y = 21 + 14x_1 + 20x_2 + 0,6x_2^2 \quad R^2 = 0,95. \\ (5,0) \quad (12,0) \quad (0,2)$$

В скобках указаны значения стандартных ошибок для коэффициентов регрессии.

Задание

1. Проанализируйте тесноту связи результата с каждым из факторов.
2. Выберите наилучшее уравнение регрессии, обоснуйте принятное решение.

Раздел 5.

Система эконометрических уравнений

Сложные экономические процессы описывают с помощью *системы взаимосвязанных (одновременных) уравнений*.

Различают несколько видов систем уравнений:

- *система независимых уравнений* - когда каждая зависимая переменная y рассматривается как функция одного и того же набора факторов x :

$$\left\{ \begin{array}{l} y_1 = a_{11} \cdot x_1 + a_{12} \cdot x_2 + \dots + a_{1m} \cdot x_m + \varepsilon_1, \\ y_2 = a_{21} \cdot x_1 + a_{22} \cdot x_2 + \dots + a_{2m} \cdot x_m + \varepsilon_2, \\ \dots \\ y_n = a_{n1} \cdot x_1 + a_{n2} \cdot x_2 + \dots + a_{nm} \cdot x_m + \varepsilon_n. \end{array} \right.$$

(для решения этой системы и нахождения ее параметров используется метод наименьших квадратов);

- *система рекурсивных уравнений* - когда зависимая переменная y одного уравнения выступает в виде фактора x в другом уравнении:

$$\left\{ \begin{array}{l} y_1 = a_{11} \cdot x_1 + a_{12} \cdot x_2 + \dots + a_{1m} \cdot x_m + \varepsilon_1, \\ y_2 = b_{21} \cdot y_1 + a_{21} \cdot x_1 + a_{22} \cdot x_2 + \dots + a_{2m} \cdot x_m + \varepsilon_2, \\ y_3 = b_{31} \cdot y_1 + b_{32} \cdot y_2 + a_{31} \cdot x_1 + a_{32} \cdot x_2 + \dots + a_{3m} \cdot x_m + \varepsilon_3, \\ \dots \\ y_n = b_{n1} \cdot y_1 + b_{n2} \cdot y_2 + \dots + b_{n-1} \cdot y_{n-1} + a_{n1} \cdot x_1 + a_{n2} \cdot x_2 + \dots + a_{nm} \cdot x_m + \varepsilon_n. \end{array} \right.$$

(для решения этой системы и нахождения ее параметров используется метод наименьших квадратов);

- *система взаимосвязанных (совместных) уравнений* - когда одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других - в правую:

$$\left\{ \begin{array}{l} y_1 = b_{12} \cdot y_{21} + b_{13} \cdot y_3 + \dots + b_{1n} \cdot y_n + a_{11} \cdot x_1 + a_{12} \cdot x_2 + \dots + a_{1m} \cdot x_m + \varepsilon_1, \\ y_2 = b_{21} \cdot y_1 + a_{23} \cdot y_{32} + \dots + b_{2n} \cdot y_n + a_{21} \cdot x_1 + a_{22} \cdot x_2 + \dots + a_{2m} \cdot x_m + \varepsilon_2, \\ \dots \\ y_n = b_{n1} \cdot y_1 + b_{n2} \cdot y_2 + \dots + b_{n-1} \cdot y_{n-1} + a_{n1} \cdot x_1 + a_{n2} \cdot x_2 + \dots + a_{nm} \cdot x_m + \varepsilon_n. \end{array} \right.$$

Такая система уравнений называется *структурной формой модели*.

Эндогенные переменные - взаимозависимые переменные, которые определяются внутри модели (системы) y .

Экзогенные переменные - независимые переменные, которые определяются вне системы x .

Предопределенные переменные - экзогенные и лаговые (за предыдущие моменты времени) эндогенные переменные системы.

Коэффициенты a и b при переменных - структурные коэффициенты модели.

Система линейных функций эндогенных переменных от всех предопределенных переменных системы - приведенная форма модели:

$$\left\{ \begin{array}{l} \hat{y}_1 = \delta_{11} \cdot x_1 + \delta_{12} \cdot x_2 + \dots + \delta_{1m} \cdot x_m, \\ \hat{y}_2 = \delta_{21} \cdot x_1 + \delta_{22} \cdot x_2 + \dots + \delta_{2m} \cdot x_m, \\ \dots \\ \hat{y}_n = \delta_{n1} \cdot x_1 + \delta_{n2} \cdot x_2 + \dots + \delta_{nm} \cdot x_m, \end{array} \right.$$

где δ - коэффициенты приведенной формы модели.

Необходимое условие идентификации - выполнение счетного правила:

$D + 1 = H$ - уравнение идентифицируемо;

$D + 1 < H$ - уравнение неидентифицируемо;

$D + 1 > H$ - уравнение сверхидентифицируемо,

где H - число эндогенных переменных в уравнении,

D - число предопределенных переменных, отсутствующих в уравнении, но присутствующих в системе.

Достаточное условие идентификации: определитель матрицы, составленной из коэффициентов при переменных, отсутствующих в исследуемом уравнении, не равен нулю, и ранг этой матрицы не менее числа эндогенных переменных системы без единицы.

Для решения идентифицируемого уравнения применяется косвенный метод наименьших квадратов, для решения сверхидентифицированных – двухшаговый метод наименьших квадратов.

Косвенный МНК состоит в следующем:

- составляют приведенную форму модели и определяют численные значения параметров каждого ее уравнения обычным МНК;

- путем алгебраических преобразований переходят от приведенной формы к уравнениям структурной формы модели, получая тем самым численные оценки структурных параметров.

Двухшаговый МНК заключается в следующем:

- составляют приведенную форму модели и определяют численные значения параметров каждого ее уравнения обычным МНК;

- выявляют эндогенные переменные, находящиеся в правой части структурного уравнения, параметры которого определяют двухшаговым МНК, и находят расчетные значения по соответствующим уравнениям приведенной формы модели;

- обычным МНК определяют параметры структурного уравнения, используя в качестве исходных данных фактические значения предопределенных переменных и расчетные значения эндогенных переменных, стоящих в правой части данного структурного уравнения.

Примеры решения избранных задач

Задача 5.1.

1. Оценить следующую структурную форму модели (СФМ) на идентификацию:

$$\begin{cases} y_1 = b_{13} \cdot y_3 + a_{11} \cdot x_1 + a_{13} \cdot x_3, \\ y_2 = b_{21} \cdot y_1 + b_{23} \cdot y_3 + a_{22} \cdot x_2, \\ y_3 = b_{32} \cdot y_2 + a_{31} \cdot x_1 + a_{33} \cdot x_3. \end{cases}$$

2. Исходя из приведенной формы модели (ПФМ) уравнений

$$\begin{cases} y_1 = 2 \cdot x_1 + 4 \cdot x_2 + 10 \cdot x_3, \\ y_2 = 3 \cdot x_1 - 6 \cdot x_2 + 2 \cdot x_3, \\ y_3 = -5 \cdot x_1 + 8 \cdot x_2 + 5 \cdot x_3, \end{cases}$$

найти структурные коэффициенты модели.

Решение

1. Модель имеет три эндогенные (y_1, y_2, y_3) и три экзогенные (x_1, x_2, x_3) переменные. Проверим каждое уравнение системы на необходимое (Н) и достаточное (Д) условия идентификации.

Первое уравнение.

Н: эндогенных переменных – 2 (y_1, y_3); отсутствующих экзогенных – 1 (x_2). Выполняется необходимое равенство: $2 = 1 + 1$, следовательно, уравнение точно идентифицируемо.

Д: в первом уравнении отсутствуют y_2 и x_2 . Построим матрицу из коэффициентов при них в других уравнениях системы:

Уравнение	Отсутствующие переменные	
	y_2	x_2
Второе	-1	a_{22}
Третье	b_{32}	0

$$Det A = -1 \cdot 0 - b_{32} \cdot a_{22} \neq 0.$$

Определитель матрицы не равен 0, ранг матрицы равен 2; следовательно, выполняется достаточное условие идентификации, и первое уравнение точно идентифицируемо.

Второе уравнение.

Н: эндогенных переменных – 3 (y_1, y_2, y_3); отсутствующих экзогенных – 2 (x_1, x_3).

Выполняется необходимое равенство: $3 = 2 + 1$, следовательно, уравнение точно идентифицируемо.

Д: во втором уравнении отсутствуют x_1 и x_3 . Построим матрицу из коэффициентов при них в других уравнениях системы:

Уравнение	Отсутствующие переменные	
	y_1	x_2
Второе	-1	0
Третье	b_{21}	a_{22}

$$\text{Det } A = a_{11} \cdot a_{33} - a_{31} \cdot a_{13} \neq 0.$$

Определитель матрицы не равен 0, ранг матрицы равен 2, следовательно, выполняется достаточное условие идентификации, и второе уравнение точно идентифицируемо.

Третье уравнение.

Н: эндогенных переменных – 2 (y_2, y_3); отсутствующих экзогенных – 1 (x_2).

Выполняется необходимое равенство: $2 = 1 + 1$, следовательно, уравнение точно идентифицируемо.

Д: в третьем уравнении отсутствуют y_1 и x_2 . Построим матрицу из коэффициентов при них в других уравнениях системы:

Уравнение	Отсутствующие переменные	
	y_2	x_2
Второе	-1	0
Третье	b_{21}	a_{22}

$$\text{Det } A = -1 \cdot a_{22} - b_{21} \cdot 0 \neq 0.$$

Определитель матрицы не равен 0, ранг матрицы равен 2, следовательно, выполняется достаточное условие идентификации, и третье уравнение точно идентифицируемо.

Делаем вывод: исследуемая система точно идентифицируема и может быть решена косвенным методом наименьших квадратов.

2. Вычислим структурные коэффициенты модели.

а) Из третьего уравнения ПФМ выразим x_2 (так как его нет в первом уравнении структурной формы):

$$x_2 = \frac{y_3 + 5 \cdot x_1 - 5 \cdot x_3}{8}.$$

Данное выражение содержит переменные y_3, x_1 и x_3 , которые нужны для первого уравнения СФМ. Подставим полученное выражение x_2 в первое уравнение ПФМ:

$$y_1 = 2 \cdot x_1 + 4 \cdot \frac{y_3 + 5 \cdot x_1 - 5 \cdot x_3}{8} + 10 \cdot x_1 \Rightarrow$$

$$\Rightarrow y_1 = 0,5 \cdot y_3 + 4,5 \cdot x_1 + 7,5 \cdot x_3 \quad (\text{первое уравнение СФМ}).$$

б) Во втором уравнении СФМ нет переменных x_1 и x_3 . Структурные параметры второго уравнения СФМ можно будет определить в два этапа.

Первый этап: выразим x_1 из первого или третьего уравнения ПФМ (например, из первого уравнения):

$$x_1 = \frac{y_1 - 4 \cdot x_2 - 10 \cdot x_3}{2} = 0,5 \cdot y_1 - 2 \cdot x_2 - 5 \cdot x_3.$$

Подстановка данного выражения во второе уравнение ПФМ не решило бы задачу до конца, так как в выражении присутствует x_3 , которого нет в СФМ.

Выразим x_3 из третьего уравнения ПМФ:

$$x_3 = \frac{y_3 + 5 \cdot x_1 - 8 \cdot x_2}{5}.$$

Подставим его в выражение x_1 :

$$\begin{aligned} x_1 &= 0,5 \cdot y_1 - 2 \cdot x_2 - 5 \cdot \left(\frac{y_3 + 5 \cdot x_1 - 8 \cdot x_2}{5} \right) = 0,5 \cdot y_1 - y_3 + 6 \cdot x_2 - 5 \cdot x_1; \\ x_1 &= \frac{0,5 \cdot y_1 - y_3 + 6 \cdot x_2}{6}. \end{aligned}$$

Второй этап: аналогично, чтобы выразить x_3 через искомые y_1 , y_3 и x_2 , заменим в выражении x_3 значение x_1 на полученное из первого уравнения ПФМ:

$$x_3 = \frac{y_3 + 5 \cdot (0,5 \cdot y_1 - 2 \cdot x_2 - 5 \cdot x_3) - 8 \cdot x_2}{5} = 0,2 \cdot y_3 + 0,5 \cdot y_1 - 3,6 \cdot x_2 - 5 \cdot x_3.$$

Следовательно,

$$x_3 = 0,033 \cdot y_3 + 0,083 \cdot y_1 - 0,6 \cdot x_2.$$

Подставим полученные значения x_1 и x_3 во второе уравнение ПФМ:

$$y_2 = 3 \cdot \frac{0,5 \cdot y_1 - y_3 + 6 \cdot x_2}{6} - 6 \cdot x_2 + 2 \cdot (0,033 \cdot y_3 + 0,083 \cdot y_1 - 0,6 \cdot x_2) \Rightarrow$$

$$\Rightarrow y_2 = 0,416 \cdot y_1 - 0,434 \cdot y_3 - 4,2 \cdot x_2 \quad (\text{второе уравнение СФМ}).$$

Это уравнение можно было бы получить из ПФМ и другим путем. Суммируя все уравнения, получим:

$$y_1 + y_2 + y_3 = 6 \cdot x_2 + 17 \cdot x_3.$$

Далее из первого и второго уравнений ПФМ исключим x_1 ; умножив первое уравнение на 3, а второе – на (-2), просуммируем их. В результате получим:

$$3 \cdot y_1 - 2 \cdot y_2 = 24 \cdot x_2 + 26 \cdot x_3.$$

Затем аналогичным путем из полученных уравнений исключаем x_3 , а именно:

$$\begin{cases} y_1 + y_2 + y_3 = 6 \cdot x_2 + 17 \cdot x_3 & | -26 \\ 3 \cdot y_1 - 2 \cdot y_2 = 24 \cdot x_2 + 26 \cdot x_3 & | 17, \end{cases}$$

После умножения уравнений на коэффициенты -26 и 17 сложим их и получим:

$$\begin{aligned} 60 \cdot y_2 &= 25 \cdot y_1 - 26 \cdot y_3 - 252 \cdot x_2 \Rightarrow \\ \Rightarrow y_2 &= 0,416 \cdot y_1 - 0,433 \cdot y_3 - 4,2 \cdot x_2. \end{aligned}$$

с) Из второго уравнения ПФМ выразим x_2 , так как его нет в третьем уравнении СФМ:

$$\begin{aligned} y_3 &= -5 \cdot x_1 + 8 \cdot (-0,167 \cdot y_2 + 0,5 \cdot x_1 + 0,333 \cdot x_3) + 5 \cdot x_3 \Rightarrow \\ \Rightarrow y_3 &= -1,336 \cdot y_2 - x_1 + 7,664 \cdot x_3 \quad (\text{третье уравнение СФМ}). \end{aligned}$$

Таким образом, СФМ примет вид:

$$\begin{cases} y_1 = 0,5 \cdot y_3 + 4,5 \cdot x_1 + 7,5 \cdot x_3, \\ y_2 = 0,416 \cdot y_1 - 0,434 \cdot y_3 - 4,2 \cdot x_2, \\ y_3 = -1,336 \cdot y_2 - x_1 + 7,664 \cdot x_3. \end{cases}$$

Задача 5.2

Изучается модель вида:

$$\begin{cases} y = a_1 + b_1(C + D) + \varepsilon_1, \\ C = a_2 + b_2 \cdot y + b_3 \cdot y_{-1} + \varepsilon_2, \end{cases}$$

где y – валовой национальный доход;

y_{-1} – валовой национальный доход предшествующего года;

C – личное потребление;

D – конечный спрос (помимо личного потребления);

ε_1 и ε_2 – случайные составляющие.

Информация за девять лет о приростах всех показателей дана в таблице 5.1.

Таблица 5.1

Год	D	y_{-1}	y	C
1	-6,8	46,7	3,1	7,4
2	22,4	3,1	22,8	30,4
3	-17,3	22,8	7,8	1,3
4	12,0	7,8	21,4	8,7
5	5,9	21,4	17,8	25,8
6	44,7	17,8	37,2	8,6
7	23,1	37,2	35,7	30,0
8	51,2	35,7	46,6	31,4
9	32,3	46,6	56,0	39,1
Σ	167,5	239,1	248,4	182,7

Для данной модели была получена система приведенных уравнений:

$$\begin{cases} y = 8,219 + 0,6688 \cdot D + 0,2610 \cdot y_{-1}, \\ C = 8,636 + 0,3384 \cdot D + 0,2020 \cdot y_{-1}. \end{cases}$$

Требуется:

1. Провести идентификацию модели.
2. Рассчитать параметры первого уравнения структурной модели.

Решение

1. В данной модели две эндогенные переменные (y и C) и две экзогенные переменные (D и y_{-1}). Второе уравнение точно идентифицировано, так как содержит две эндогенные переменные и не содержит одну экзогенную переменную из системы. Иными словами, для второго уравнения имеем по счетному правилу идентификации равенство: $2 = 1 + 1$.

Первое уравнение сверхидентифицировано, так как в нем на параметры при C и D наложено ограничение: они должны быть равны. В этом уравнении содержится одна эндогенная переменная y . Переменная C в данном уравнении не рассматривается как эндогенная, так как она участвует в уравнении не самостоятельно, а вместе с переменной D . В данном уравнении отсутствует одна экзогенная переменная, имеющаяся в системе. По счетному правилу идентификации получаем: $1 + 1 = 2$; $D + 1 > N$. Это больше, чем число эндогенных переменных в данном уравнении, следовательно, система сверхидентифицирована.

2. Для определения параметров сверхидентифицированной модели используется двухшаговый метод наименьших квадратов.

Шаг первый. На основе системы приведенных уравнений по точно идентифицированному второму уравнению определим теоретическое значение эндогенной переменной C . Для этого в приведенное уравнение

$$C = 8,636 + 0,3384 \cdot D + 0,2020 \cdot y_{-1}$$

подставим значения D и y_{-1} , имеющиеся в условии задачи. Получим:

$$\hat{C}_1 = 15,8; \hat{C}_2 = 16,8; \hat{C}_3 = 7,4; \hat{C}_4 = 14,3; \hat{C}_5 = 15,0; \hat{C}_6 = 27,4; \hat{C}_7 = 24,0; \hat{C}_8 = 33,2; \hat{C}_9 = 29,0.$$

Шаг второй. По сверхидентифицированному уравнению структурной формы модели заменим фактические значения C на теоретические \hat{C} и рассчитаем новую переменную $\hat{C} + D$ (таблица 5.2).

Таблица 5.2

Год	D	\hat{C}	$\hat{C} + D$
1	-6,8	15,8	9,0
2	22,4	16,8	39,2
3	-17,3	7,4	-9,9
4	12,0	14,3	26,3
5	5,9	15,0	20,9

6	44,7	27,4	72,1
7	23,1	24,0	47,1
8	51,2	33,2	84,4
9	32,3	29,0	61,3
Σ	167,5	182,9	350,4

Далее к сверхидентифицированному уравнению применяется метод наименьших квадратов. Обозначим новую переменную $\tilde{C} + D$ через Z . Решаем уравнение:

$$y = a_1 + b_1 \cdot Z.$$

Соответствующая система нормальных уравнений будет выглядеть так:

$$\begin{cases} \sum y = n \cdot a_1 + b_1 \cdot \sum Z, \\ \sum y \cdot Z = a_1 \cdot \sum Z + b_1 \cdot \sum Z^2, \end{cases}$$

$$\begin{cases} 248,4 = 9 \cdot a_1 + 350,4 \cdot b_1, \\ 13508,71 = 350,4 \cdot a_1 + 21142,02 \cdot b_1, \end{cases}$$

$$a_1 = 7,678; \quad b_1 = 0,512.$$

Итак, первое уравнение структурной модели будет таким:

$$y = 7,678 + 0,512 \cdot (C + D).$$

Задача 5.3

По одному из штатов США имеются данные за 1990 – 1994 гг. (таблица 5.3).

Таблица 5.3

Год	Годовое потребление свинины на душу населения, y_1 (фунты)	Оптовая цена за фунт, y_2 (долл.)	Доход на душу населения, долл., x_1	Расходы по обработке мяса, % к цене, x_2
1998	60	5,0	1300	60
1999	62	4,0	1300	56
2000	65	4,2	1500	56
2001	62	5,0	1600	63
2002	66	3,8	1800	50

Построить модель вида:

$$\begin{cases} y_1 = f(y_2, x_1), \\ y_2 = f(y_1, x_2). \end{cases}$$

рассчитав соответствующие коэффициенты.

Решение

Система одновременных уравнений с двумя эндогенными и двумя экзогенными переменными имеет вид:

$$\begin{cases} y_1 = b_{12} \cdot y_2 + a_{11} \cdot x_1 + \varepsilon_1, \\ y_2 = b_{21} \cdot y_1 + a_{22} \cdot x_2 + \varepsilon_2. \end{cases}$$

В каждом уравнении две эндогенные и одна отсутствующая экзогенная переменные из имеющихся в системе. Для каждого уравнения данной системы действует счетное правило $2 = 1 + 1$. Это означает, что каждое уравнение и система в целом идентифицированы.

Для определения параметров такой системы применяется косвенный метод наименьших квадратов.

С этой целью структурная форма модели преобразуется в приведенную форму:

$$\begin{cases} y_1 = \delta_{11} \cdot x_1 + \delta_{12} \cdot x_2, \\ y_2 = \delta_{21} \cdot x_1 + \delta_{22} \cdot x_2 + \varepsilon_2, \end{cases}$$

в которой коэффициенты при x определяются методом наименьших квадратов.

Для нахождения значений δ_{11} и δ_{12} запишем систему нормальных уравнений:

$$\begin{cases} \sum y_1 x_1 = \delta_{11} \cdot \sum x_1^2 + \delta_{12} \cdot \sum x_1 x_2, \\ \sum y_1 x_2 = \delta_{11} \cdot \sum x_1 x_2 + \delta_{12} \cdot \sum x_2^2. \end{cases}$$

При ее решении предполагается, что x и y выражены через отклонения от средних уровней, то есть матрица исходных данных составит (таблица 5.4):

Таблица 5.4

	y_1	y_2	x_1	x_2
	-3	0,6	-200	3
	-1	-0,4	-200	-1
	2	-0,2	0	-1
	-1	0,6	100	6
	3	-0,6	300	7
Σ	0	0,0	0	0

Применимельно к ней необходимые суммы оказываются следующими:

$$\sum y_1 x_1 = 1600; \quad \sum y_1 x_2 = -37; \quad \sum x_1^2 = 180000; \quad \sum x_1 x_2 = -1900; \quad \sum x_2^2 = 96.$$

Соответствующая система нормальных уравнений:

$$\begin{cases} 1600 = 180000 \cdot \delta_{11} - 1900 \cdot \delta_{12}, \\ -37 = -1900 \cdot \delta_{11} + 96 \cdot \delta_{12}. \end{cases}$$

Решая ее, получим:

$$\delta_{11} = 0,00609; \quad \delta_{12} = -0,26481.$$

Итак, имеем $y_1 = 0,00609 \cdot x_1 - 0,26481 \cdot x_2$.

Аналогично строим систему нормальных уравнений для определения коэффициентов δ_{21} и δ_{22} :

$$\begin{cases} \sum y_2 x_1 = \delta_{21} \cdot \sum x_1^2 + \delta_{22} \cdot \sum x_1 x_2, \\ \sum y_2 x_2 = \delta_{21} \cdot \sum x_1 x_2 + \delta_{22} \cdot \sum x_2^2. \end{cases}$$

$$\sum y_2 x_1 = -160; \quad \sum y_2 x_2 = 10,2.$$

$$\begin{cases} -160 = 180000 \cdot \delta_{21} - 1900 \cdot \delta_{22}, \\ 10,2 = -1900 \cdot \delta_{21} + 96 \cdot \delta_{22}. \end{cases}$$

Следовательно,

$$\delta_{21} = 0,00029; \quad \delta_{22} = 0,11207,$$

тогда второе уравнение примет вид:

$$y_2 = 0,00029 \cdot x_1 + 0,11207 \cdot x_2.$$

Приведенная форма модели имеет вид:

$$\begin{cases} y_1 = 0,00609 \cdot x_1 - 0,26481 \cdot x_2, \\ y_2 = 0,00029 \cdot x_1 + 0,11207 \cdot x_2. \end{cases}$$

Из приведенной формы модели определяем коэффициенты структурной модели:

$$\begin{cases} y_1 = 0,00609 \cdot x_1 - 0,26481 \cdot x_2, \\ x_2 = \frac{y_2 - 0,00029 \cdot x_1}{0,11207}, \end{cases} \Rightarrow$$

$$\Rightarrow y_1 = 0,00609 \cdot x_1 - 0,26481 \cdot \frac{y_2 - 0,00029 \cdot x_1}{0,11207} = -2,36290 \cdot y_2 + 0,00678 \cdot x_1,$$

$$\begin{cases} y_2 = 0,00029 \cdot x_1 + 0,11207 \cdot x_2, \\ x_1 = \frac{y_1 + 0,26481 \cdot x_2}{0,00609}, \end{cases} \Rightarrow$$

$$\Rightarrow y_2 = 0,00029 \cdot \frac{y_1 + 0,26481 \cdot x_2}{0,00609} + 0,11207 \cdot x_1 = 0,04762 \cdot y_1 + 0,12468 \cdot x_2,$$

Итак, структурная форма модели имеет вид

$$\begin{cases} y_1 = -2,36290 \cdot y_2 + 0,00678 \cdot x_1 + \varepsilon_1, \\ y_2 = 0,04762 \cdot y_1 + 0,12468 \cdot x_2 + \varepsilon_2. \end{cases}$$

Задачи для самостоятельного решения

Задача 5-А. Имеется следующая гипотетическая структурная модель:

$$Y_1 = b_{12}Y_{21} + a_{11}X_1 + a_{12}X_2,$$

$$Y_2 = b_{21}Y_1 + b_{23}Y_3 + a_{22}X_2,$$

$$Y_3 = b_{32}Y_{21} + a_{31}X_1 + a_{33}X_3.$$

Приведенная форма исходной модели имеет вид:

$$Y_1 = 3X_1 - 6X_2 + 2X_3,$$

$$Y_2 = 2X_1 + 4X_2 + 10X_3,$$

$$Y_3 = -5X_1 + 6X_2 + 5X_3.$$

Задание

1. Проверьте структурную форму модели на идентификацию.

2. Определите структурные коэффициенты модели.

Задача 5-В. Строится модель вида:

$$Y_1 = a_1 + b_1Y_2 + c_1X_1 + \varepsilon_1,$$

$$Y_2 = a_2 + b_2Y_1 + c_2X_2 + \varepsilon_2.$$

Задание

Определите структурные коэффициенты, учитывая, что:

$$\sum Y_1 X_1 = 2600; \sum Y_1 X_2 = 4350; \sum Y_2 = 350; \sum X_1 = 750;$$

$$\sum X_2 = 350; \sum X_1^2 = 1200; \sum X_2^2 = 1800; n = 30; \sum X_1 X_2 = 1500; Y_2 = 2X_1 + 3X_2.$$

Задача 5-С. Эконометрическая модель содержит четыре уравнения, четыре эндогенные переменные (y) и три экзогенные переменные (x). Ниже представлена матрица коэффициентов при переменных в структурной форме этой модели (таблица 5.5).

Таблица 5.5

Уравнение	y_1	y_2	y_3	y_4	x_1	x_2	x_3
I	-1	0	b_{13}	b_{14}	C_{11}	0	0
II	0	-1	b_{23}	0	C_{21}	0	0
III	0	b_{32}	-1	0	C_{31}	0	C_{33}
IV	b_{41}	b_{42}	b_{43}	-1	0	C_{42}	C_{43}

Задание

Применив необходимое и достаточное условие идентификации, определите, идентифицируемо ли каждое уравнение модели.

Задача 5-Д. Ниже приводятся результаты расчета параметров некоторой эконометрической модели.

Структурная форма модели:

$$Y_1 = -4 + ???Y_2 - 9.4X_3 + \varepsilon_1,$$

$$Y_2 = 12.83 - 2.67Y_1 + ???X_1 + \varepsilon_2,$$

$$Y_3 = 1.36 - 1.76Y_1 + 0.828Y_2 + \varepsilon_3.$$

Приведенная форма модели:

$$Y_1 = 2 + 4X_1 - 3X_2 + \nu_1,$$

$$Y_2 = 7.5 + 5X_1 + 8X_2 + \nu_2,$$

$$Y_3 = 4 - ???X_1 + ???X_2 + \nu_3.$$

Задание

1. Какими методами получены параметры структурной и приведенной форм модели? Обоснуйте возможность применения косвенного МНК для расчета структурных параметров модели.

2. Восстановите пропущенные характеристики.

Задача 5-Е. Рассматривается следующая модель:

$$S_t = a_1 + b_{11}D_t + b_{12}M_t + b_{13}Un_t + \varepsilon_1,$$

$$C_t = a_2 + b_{21}D_t + b_{22}S_t + b_{23}Un_{t-1} + \varepsilon_2,$$

$$D_t = a_3 + b_{31}S_t + b_{32}C_{t-1} + b_{33}I_t + \varepsilon_3,$$

где S_t - заработка в период t ;

D_t - чистый национальный доход в период t ;

M_t - денежная масса в период t ;

C_t - расходы на потребление в период t ;

C_{t-1} - расходы на потребление в период $t-1$;

Un_t - уровень безработицы в период t ;

Un_{t-1} - уровень безработицы в предыдущий период;

I_t - инвестиции в период t .

Задание

1. Каким методом вы будете оценивать структурные параметры этой модели?

2. Выпишите приведенную форму модели.

3. Кратко охарактеризуйте методику расчета параметров первого и второго структурного уравнения модели.

Задание к задачам F, G, H

1. Применив необходимое и достаточное условие идентификации, определите, идентифицировано ли каждое из уравнений модели.

2. Определите метод оценки параметров модели.

3. Запишите приведенную форму модели.

Задача 5-Ф. Одна из версий модифицированной модели Кейнса имеет вид:

$$C_t = a_1 + b_{11}Y_t + b_{12}Y_{t-1} + \varepsilon_1,$$

$$I_t = a_2 + b_{21}Y_t + b_{22}Y_{t-1} + \varepsilon_2,$$

$$Y_t = C_t + I_t + G_t,$$

где С – расходы на потребление;

Y – доход;

I – инвестиции;

G – государственные расходы;

t – текущий период;

t-1 – предыдущий период.

Задача 5-Г. Модель Менгеса:

$$Y_t = a_1 + b_{11}Y_{t-1} + b_{12}I_t + \varepsilon_1,$$

$$I_t = a_2 + b_{21}Y_t + b_{22}Q_t + \varepsilon_2,$$

$$C_t = a_3 + b_{31}Y + b_{32}C_{t-1} + b_{33}P_t + \varepsilon_3,$$

$$Q_t = a_4 + b_{41}Q_{t-1} + b_{42}R_t + \varepsilon_4,$$

где Y – национальный доход;

C – расходы на личное потребление;

I – чистые инвестиции;

Q – валовая прибыль экономики;

P – индекс стоимости жизни;

R – объем продукции промышленности;

t – текущий период;

t-1 – предыдущий период.

Задача 5-Н. Модель денежного рынка:

$$R_t = a_1 + b_{11} \cdot M_t + b_{12}Y_t + \varepsilon_1,$$

$$Y_t = a_2 + b_{21} \cdot R_t + b_{22}I_t + \varepsilon_2,$$

где R – процентная ставка;

Y – ВВП;

M – денежная масса;

I – внутренние инвестиции;

T – текущий период.