Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Приморская государственная сельскохозяйственная академия»
Институт земледелия и природообустройства

Савельева Е.В., Островская И.Э.

МОДЕЛИРОВАНИЕ И СТАТИСТИЧЕСКАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ НАУЧНЫХ ИССЛЕДОВАНИЙ

Учебное пособие

для обучающихся по направлениям подготовки:

35.06.01 Сельское хозяйство; 35.06.02 Лесное хозяйство;
35.06.04 Технологии, средства механизации и энергетическое оборудование в сельском, лесном и рыбном хозяйстве; 36.06.01 Ветеринария и зоотехния;
38.06.01 Экономика
ФГБОУ ВПО Приморская ГСХА

Уссурийск 2014

УДК 519

ББК 22.1

М 74

Рецензент: Лосев А.С., к.ф.-м. наук, доцент Федерального государственного бюджетного учреждения науки Института прикладной математики Дальневосточного отделения РАН

Сидорова Г.М, к. с.-х. н., доцент, доцент кафедры землеустройства

Моделирование и статистическая обработка результатов научных исследований: учебное пособие для обучающихся по направлениям подготовки 35.06.01 Сельское хозяйство; 35.06.02 Лесное хозяйство; 35.06.04 Технологии, средства механизации и энергетическое оборудование в сельском, лесном и рыбном хозяйстве; 36.06.01 Ветеринария и зоотехния; 38.06.01 Экономика ФГБОУ ВПО Приморская ГСХА / ФГБОУ ВПО Приморская ГСХА; сост. Е.В. Савельева, И.Э. Островская – Уссурийск, 2014. – 80 с.

Учебное пособие «Моделирование и статистическая обработка результатов научных исследований» представляет собой краткое изложение методов обработки и анализа результатов экспериментов и наблюдений с применением информационных технологий для использования их в научно-исследовательской деятельности. Основная цель учебного пособия состоит в том, чтобы оказать помощь аспиранту в изучении вопросов дисциплины в соответствии с программой.

Издается по решению методического совета ФГБОУ ВПО Приморская ГСХА

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Учебное пособие для самостоятельной работы составлено в соответствии с программой курса «Моделирование и статистическая обработка результатов научных исследований», которая занимает важное место в системе прикладного математического образования. Она нацелена на формирование мировоззрения научного исследования.

Содержание дисциплины охватывает круг вопросов, связанных с моделированием и статистической обработкой результатов научных исследований.

В результате изучения дисциплины обучающийся должен:

Знать: методы и технологий обработки экспериментальных данных.

***Уметь*:**

- планировать и организовывать научные эксперименты;

-применять методы статистической обработки данных к исследуемой области;

-строить математические модели исследуемых процессов и явлений;

-анализировать и интерпретировать полученные результаты.

Владеть: навыками статистической обработки экспериментальных данных полученных результатов с помощью компьютерных программ и технологий, построения математических моделей процессов, явлений и объектов, относящихся к исследуемой области.

Материал курса основан на знаниях, полученных обучающимися в результате изучения курсов теория вероятностей, математическая статистика и математическая теория эксперимента.

Целью данного пособия является ознакомить аспирантов с методами обработки и анализа результатов экспериментов и наблюдений с применением информационных технологий для использования их в научно-исследовательской деятельности.

Пособие состоит из 8 разделов. Материал расположен в учебном

пособии по принципу нарастающей сложности рассматриваемых тем и задач.

В целях более эффективного усвоения учебного материала каждая тема содержит краткое теоретическое введение, методические указания с описанием решения конкретных задач, варианты задач для самостоятельного решения и контрольных вопросов.

Краткое изложение теоретического материала не позволяет осветить все вопросы курса статистики и представить формулы по всем описанным показателям. Поэтому при подготовке к практическим и семинарским занятиям рекомендуется более глубокое изучение тем в основной и дополнительной литературе.

1. ПЕРВИЧНАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

1.1 Выборка и ее представление. Числовые характеристики

Математической статистикой называется наука, занимающаяся разработкой методов получения, описания и обработки опытных данных с целью изучения закономерностей случайных массовых явлений.

Определение 1. Генеральной совокупностью называется совокупность всех возможных однородных предметов или явлений, над которыми проводятся наблюдения, или совокупность всех возможных наблюдений, проводимых над некоторой случайной величиной в одинаковых условиях.

Определение 2. Выборочной совокупностью (выборкой) называется совокупность предметов или явлений, отобранная из соответствующей генеральной совокупности.

Определение 3. Объемом совокупности (генеральной или выборочной) называется общее число ее элементов.

• *Выборочный метод* заключается в том, что из генеральной совокупности берется выборка значительно меньшего объема и определяются характеристики выборки, которые принимаются в качестве приближенных значений соответствующих характеристик генеральной совокупности.

Выборки бывают:

а) ***повторные*** – если отобранный объект (перед отбором следующего) возвращается в генеральную совокупность;

б) ***бесповторные*** – если отобранный объект не возвращается в генеральную совокупность.

Пусть из генеральной совокупности извлечена выборка X объемом n. Случайный *выбор элемента* рассматривается как независимое наблюдение над величиной ξ, имеющей некоторое распределение вероятностей. Если те значения, которые приняла случайная величина ξ в n наблюдениях, записать

не в порядке получения, а в порядке возрастания (то есть ***ранжируя***), то получим упорядоченную выборку x_1, x_2, K, x_n ,, называемую ***вариационным рядом***.

- Выборка и вариационный ряд несут одну и ту же информацию, но с вариационным рядом работать легче в силу его упорядоченности.

- Чтобы по данным выборки можно было достаточно точно судить об исследуемом признаке генеральной совокупности, требуется, чтобы выборка правильно представляла пропорции генеральной совокупности, то есть была ***репрезентативной (представительной).***

Определение 4. Вариантой называется значение x_i случайной величины, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных.

Определение 5. Размахом R вариационного ряда называется разность между его наибольшей $x_{\max}$ и наименьшей $x_{\min}$ вариантами: $R = x_{\max} - x_{\min}$.

Определение 6. Частотой (весом) варианты называется численность отдельной группы сгруппированного ряда наблюдаемых данных.

Определение 7. Относительной частотой варианты x_i называется отношение m_i числа повторения x_i к объему выборки n:

$$\tilde{p}_i = \frac{m_i}{n}. \qquad (1)$$

Очевидно, что $n = \sum_{i=1}^{k} m_i$.

По таблице, изображающей вариационный ряд, построим таблицу из двух строк, в верхней строке которой указаны в порядке возрастания варианты x_i , а в нижней – соответствующие им относительные частоты $\tilde{p}_i$ (табл. 1). Такая таблица называется *таблицей статистического распределения.*

Таблица 1.

Значения x_i	x_1	x_2	…	x_k
Относительные частоты $\tilde{p}_i$	$\tilde{p}_1 = \frac{m_1}{n}$	$\tilde{p}_2 = \frac{m_2}{n}$	…	$\tilde{p}_k = \frac{m_k}{n}$

Определение 8. Дискретным вариационным рядом распределения (***распределением частот*** или ***относительных частот***) называется ранжированная совокупность вариант x_i с соответствующими им частотами или относительными частотами.

Алгоритм составления дискретного вариационного ряда:

– найти минимальное ($x_{\min}$) и максимальное ($x_{\max}$) значения выборки;

– в первый столбец таблицы записать варианты значений случайной величины (генеральной совокупности), начиная с $x_{\min}$ и заканчивая $x_{\max}$;

– просмотреть по одному все элементы выборки в протоколе наблюдений, и подсчитать количество значений, соответствующих каждой варианте (то есть m_i);

– подсчитать количество элементов выборки (ее объем).

Определение 9. Интервальным вариационным рядом (интервальным распределением частот или ***относительных частот)*** называется упорядоченная последовательность интервалов варьирования случайной величины с соответствующими частотами или относительными частотами попаданий в каждый из них значений случайной величины.

Алгоритм составления интервального вариационного ряда:

– найти минимальное ($x_{\min}$) и максимальное ($x_{\max}$) значения выборки;

– найти объем n выборки;

– определить оптимальное число интервалов по *формуле Стерджеса*:

$$k = 1 + 3{,}322 \cdot \lg n . \qquad (2)$$

– найти длину интервала по формуле:

$$h = \frac{x_{\max} - x_{\min}}{k} . \qquad (3)$$

– заполнить первый столбец таблицы интервалами исследуемой совокупности. За начало первого интервала рекомендуется брать величину

$$x_{нач} = x_{\min} - 0{,}5 \cdot h. \qquad (4)$$

– просмотреть по одному все элементы выборки в протоколе наблюдений, и подсчитать количество значений, соответствующих каждому интервалу (то есть m_i).

Определение 10. Накопленной частотой $m_x^{нак}$ в точке x называют суммарную частоту членов статистической совокупности со значениями признака, меньшими, чем x.

Определение 11. Накопленной относительной частотой (частостью) $p_x^{нак}$ называется отношение накопленной частоты $m_x^{нак}$ к объему выборки n: $p_x^{нак} = \dfrac{m_x^{нак}}{n}$. (5)

Определение 12. Эмпирической функцией распределения (выборочной или ***функцией распределения выборки)*** называется функция $F^*(x)$, определяющая для каждого значения x относительную частоту события $X<x$:

$$F^*(x) = \tilde{P}(X < x) = p_x^{нак}. \qquad (6)$$

Графическое представление выборки.

Полигон

1. ***Полигон частот*** – это ломаная, отрезки которой соединяют точки $(x_1; m_1), (x_2; m_2), \text{K}, (x_k; m_k)$. (см. табл. 1). Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им частоты m_i. Точки $(x_i; m_i)$ соединяют отрезками прямых и получают полигон частот.

2. ***Полигон относительных частот*** – это ломаная, отрезки которой соединяют точки $(x_1; \tilde{p}_1), (x_2; \tilde{p}_2), \text{K}, (x_k; \tilde{p}_k)$. (см. табл. 1). Для построения полигона относительных частот на оси абсцисс откладывают варианты x_i, а

на оси ординат – соответствующие им относительные частоты $\tilde{p}_i$. Точки $(x_i; \tilde{p}_i)$ соединяют отрезками прямых и получают полигон частот.

Гистограмма.

Гистограмма строится только для интервального вариационного ряда (группированной выборки).

1. ***Гистограмма частот*** – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению m_i/h(плотность частоты).

- Площадь гистограммы частот равна сумме всех частот, то есть объему выборки.

2. ***Гистограмма относительных частот*** – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\tilde{p}_i/h$ (плотность относительной частоты).

Кумулята

Кумулята (кумулятивная кривая) – график накопленных частот, сглаженное графическое изображение эмпирической функции распределения. При построении кумуляты в точке, соответствующей принимаемому значению, для дискретного ряда и в правом конце интервала для интервального ряда строится перпендикуляр, высота которого пропорциональна накопленной частоте, а затем верхние концы перпендикуляров соединяются между собой с помощью отрезков прямых.

Числовые характеристики вариационного ряда (выборки).

Характеристики уровня вариационного ряда

1. Выборочное среднее (среднее арифметическое):

а) ***выборочное среднее (среднее арифметическое) простое***:

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}(x_1 + x_2 + x_3 + \text{K} + x_n), \qquad (1)$$

где $x_i \ (i=\overline{1,n})$ - варианты признака (элементы выборки), n – объем выборки.

- Если частоты вариантов признака различны (т.е. имеют различный удельный вес во всем объеме совокупности), то вычисляется среднее арифметическое, взвешенное по частотам.

б) выборочное среднее (среднее арифметическое) взвешенное:

$$\bar{x}_в = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i \cdot m_i = \frac{1}{n}(x_1 m_1 + x_2 m_2 + \text{K} + x_k m_k), \qquad (2)$$

где x_i - варианты случайной величины (признака), $m_i \ (i=\overline{1,k})$ – соответствующие частоты, k – количество вариантов, n – объем выборки.

2. Мода

а) Для дискретного вариационного ряда ***модой*** M_0 выборки является значение, имеющее максимальную частоту.

б) Для интервального вариационного ряда ***мода*** M_0 вычисляется по приближенной формуле:

$$M_0 \approx x_0 + h \cdot \frac{m_i - m_{i-1}}{(m_i - m_{i-1}) + (m_i - m_{i+1})}, \qquad (3)$$

где: x_0 – начало модального интервала, т.е. интервала, имеющего максимальную частоту, h – длина модального интервала, m_i - частота модального интервала, m_{i-1} и m_{i+1} – частоты соответственно предшествующего и последующего за модальным интервалов.

3. Медиана

Медиана выборки – это значение серединного элемента вариационного ряда.

а) Для дискретного вариационного ряда ***медиана*** M_e находится по формуле:

$$M_e = \begin{cases} \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), & \text{если } n - \text{четное;} \\ x_{\frac{n+1}{2}}, & \text{если } n - \text{нечетное.} \end{cases} \qquad (4)$$

б) Для интервального вариационного ряда ***медиана*** M_e вычисляется по формуле:

$$M_e = x_0 + h \cdot \frac{\frac{n}{2} - T_{i-1}}{m_i}, \qquad (5)$$

где: x_0 - начало медианного интервала, (т.е. интервала, в котором находится серединный элемент); h – длина медианного интервала; n – объем выборки; T_{i-1} – сумма частот интервалов, предшествующих медианному; m_i - частота медианного интервала.

Показатели колеблемости вариационных рядов

1. ***Размах вариации:*** $R = x_{\max} - x_{\min}$, (6)

где $x_{\max}$ – максимальное, $x_{\min}$ – минимальное значения выборки.

2. ***Среднее линейное отклонение*** – это среднее арифметическое, исчисленное из абсолютных отклонений отдельных вариант (x_i) от их выборочной средней ($\bar{x}_в$). Различают:

а) *среднее линейное отклонение (простое):*

$$d = \frac{1}{n} \cdot \sum_{i=1}^{k} |x_i - \bar{x}_в|; \qquad (7)$$

б) *среднее линейное отклонение (взвешенное):*

$$d = \frac{1}{n} \cdot \sum_{i=1}^{k} |x_i - \bar{x}| \cdot m_i. \qquad (8)$$

3. ***Выборочная дисперсия*** $D_в$ – это среднее арифметическое квадратов отклонений отдельных вариант (x_i) этой выборочной средней ($\bar{x}_в$). Бывает:

а) *выборочная дисперсия (простая):*

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x}_в)^2 \qquad (9)$$

или

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i^2 - \bar{x}_в^{\,2}; \qquad (10)$$

б) *выборочная взвешенная дисперсия*:

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} (x_i - \bar{x}_в)^2 \cdot m_i \qquad (11)$$

или

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i^2 \cdot m_i - \bar{x}_в^{\,2}. \qquad (12)$$

в) *исправленная дисперсия*:

$$S^2 = \frac{n}{n-1} \cdot D_в. \qquad (13)$$

4. *Выборочное среднее квадратическое отклонение* $\sigma_в$ – квадратный корень из дисперсии.

Различают:

а) *выборочное среднее квадратическое отклонение простое*:

$$\sigma_в = \sqrt{\frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x}_в)^2} \qquad (14)$$

или

$$\sigma_в = \sqrt{\frac{1}{n} \sum_{i=1}^{k} x_i^2 - \bar{x}_в^{\,2}}; \qquad (15)$$

б) *выборочное среднее квадратическое отклонение взвешенное*:

$$\sigma_в = \sqrt{\frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x}_в)^2 \cdot m_i} \qquad (16)$$

или

$$\sigma_в = \sqrt{\frac{1}{n} \sum_{i=1}^{k} x_i^2 \cdot m_i - \bar{x}_в^{\,2}}. \qquad (17)$$

5. ***Выборочный коэффициент вариации*** – это отклонение выборочного среднего квадратического отклонения к выборочному среднему, выраженное в процентах:

$$V_в = \frac{\sigma_в}{\bar{x}_в} \cdot 100\% \qquad (18)$$

Пример.

Даны результаты измерения роста (с точностью до см) 60 наудачу отобранных студентов:

178, 160, 154, 183, 155, 153, 167, 186, 163, 155, 157, 175, 170, 166, 159, 173, 182, 167, 171, 169, 179, 165, 156, 179, 158, 171, 175. 173, 164, 172, 178, 160, 154, 183, 155, 153, 167, 186, 163, 155, 157, 175, 170, 166, 159, 173, 182, 167, 171, 169, 179, 165, 156, 179, 158, 171, 175. 173, 164, 172

На основе совокупности опытных данных выполнить следующие задания:

Задание 1. Построить интервальный вариационный ряд распределения.

Задание 2. Построить гистограмму частот интервального вариационного ряда.

Задание 3. Составить эмпирическую функцию распределения и построить график.

Задание 4. Рассчитать основные числовые характеристики вариационного ряда:

а) моду и медиану;

б) условные начальные моменты;

в) выборочную среднюю;

г) выборочную дисперсию, исправленную дисперсию генеральной совокупности, исправленное среднее квадратичное отклонение;

д) коэффициент вариации;

е) асимметрию;

ж) эксцесс;

Задание 5. Определить границы истинных значений числовых характеристик, изучаемой случайной величины с заданной надёжностью.

Задание 6. Содержательная интерпретация результатов первичной обработки по условию задачи.

Решение.

Задание 1. Построить интервальный вариационный ряд распределения

n=60;: $x_{\max} = 186, \quad x_{\min} = 153.$

Длина частичного интервала:

$$h = \frac{x_{\max} - x_{\min}}{1 + \log_2 n} = \frac{186 - 153}{1 + \log_2 60} = \frac{33}{1 + 3{,}322 \lg 60} \approx \frac{33}{5{,}9} \approx 5{,}59$$

Примем h = 6. Начало первого интервала $x_{нач} = x_{\min} - \frac{h}{2} = 153 - 3 = 150$

Исходные данные разбиваем на следующие равные интервалы:

(150,156], (156,162], (162,168], (168,174], (174, 180], (180, 186].

Подсчитаем n_i - число студентов попавших в каждый из полученных промежутков.

Таблица 2

X	[150,156]	(156,162]	(162,168]	(168,174]	(174, 180]	(180, 186]
n_i (частота)	8	10	12	14	10	6

Задание 2. Построить гистограмму частот интервального вариационного ряда

Длина интервала h = 6. Найдем плотность частоты n_i / h.

Таблица 3

X	[150,156)	[156,162)	[162,168)	[168,174)	[174, 180)	[180, 186]
n_i (частота)	8	10	12	14	10	6
n_i / h	8/6 ≈ 1,4	10/6 ≈ 1,65	12/6 = 2,0	14/6 ≈ 2,3	10/6 ≈1,65	6/6 = 1

.

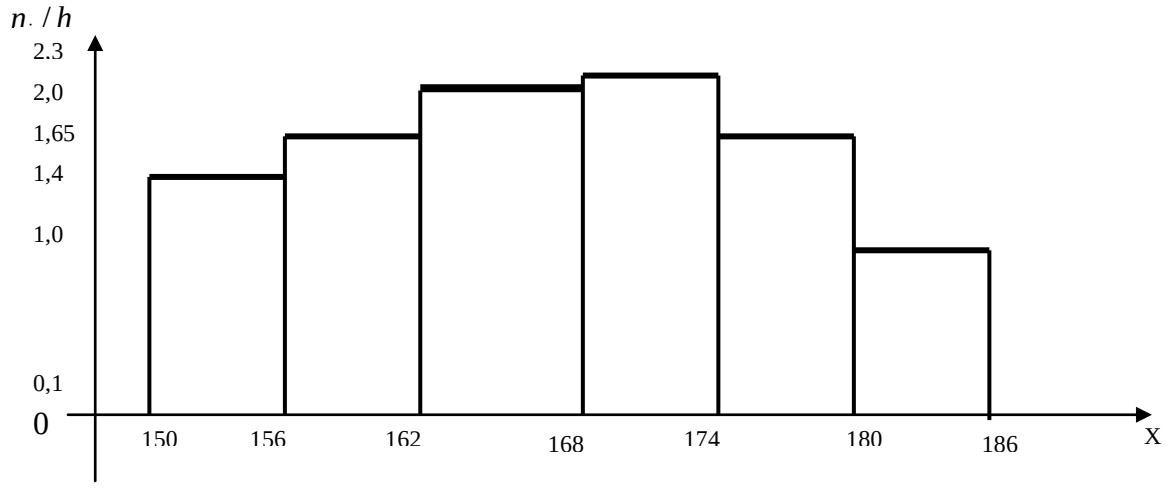


Рисунок 1 - Гистограмма частот вариационного ряда

Задание 3. Записать эмпирическую функцию распределения и построить ее график

Вычислим:

- середину каждого интервала x_i и запишем эти значения в первую строку таблицы;

- относительные частоты $w_i = n_i / n$ и запишем в третью строку таблицы;

- накопительные частоты $\sum w_i$ и запишем в четвертую строку таблицы.

Таблица 4

Значение признака x_i (середина интервала)	153	159	165	171	177	183
n_i (частота)	8	10	12	14	10	6
$w_i = n_i / n$ (частость)	8/60 ≈ 0,133	10/60 ≈ 0,167	12/60 = 0.2	14/60 ≈ 0,233	10/60 ≈ 0,167	6/60 = 0,1
Накопительные относительные частоты $\sum w_i$	0,133	0,133+0,167= 0,3	0,133+0,167+0,2= 0,5	0,733	0,9	1

$$F^*(x) = \begin{cases} 0 & при\ x \le 153 \\ 0{,}133 & при\ 153 < x \le 159 \\ 0{,}3 & при\ 159 < x \le 165 \\ 0{,}5 & при\ 165 < x \le 171 \\ 0{,}733 & при\ 171 < x \le 177 \\ 0{,}9 & при\ 177 < x \le 183 \\ 1 & при\ x > 183 \end{cases}$$

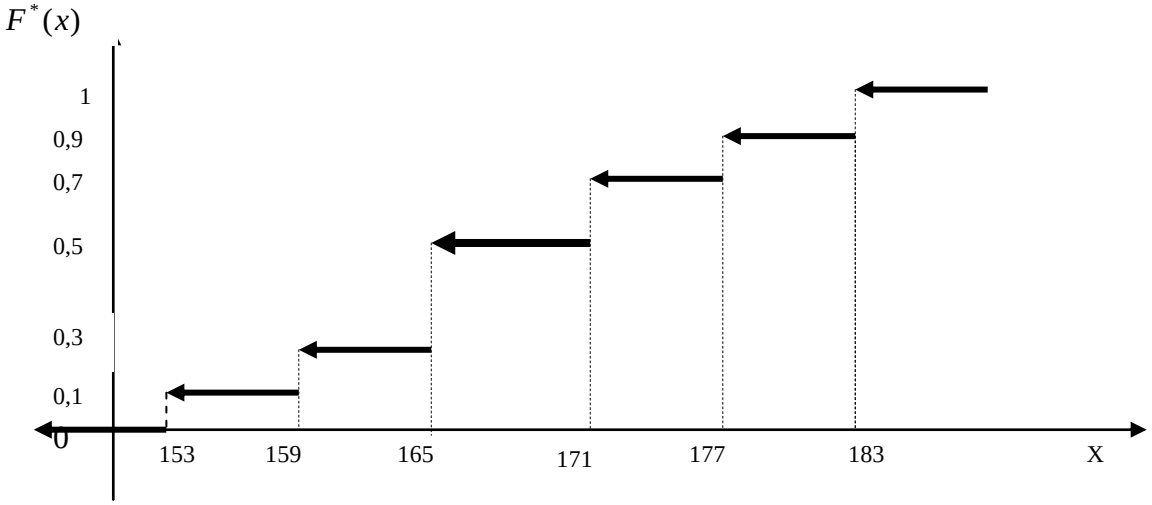


Рисунок 2 - График эмпирической функции распределения

Задание 4. Рассчитать основные числовые характеристики вариационного ряда

а) Мода – это варианта, имеющая наибольшую частоту.

По данным Таблицы 4: $M_0(X)$ = 171.

В нашем примере число интервалов четное

По данным Таблицы 4: (153+159+165+171+177+183)/6=168, $M_e(X)=168$

Для определения остальных числовых характеристик воспользуемся методом произведений.

Введем условные варианты. $u_i = \frac{x_i - C}{h}$, где C – «ложный нуль».

Чаще всего в качестве ложного нуля принимается либо варианта, находящаяся в середине вариационного ряда, либо мода M_0 (варианта x_i, имеющая наибольшую частоту), либо любое другое число, упрощающее расчеты.

Если за C принять какое - либо значение x_i, то соответствующая ему условная варианта u_i будет равна нулю, а слева и справа от нуля будут располагаться соответственно значения $\pm 1, \pm 2, \pm 3, \pm 4$ и т.д.

Примем C = 171, h = 6. Составим расчетную таблицу:

Таблица 5

x_i	n_i	u_i	$n_i \cdot u_i$	$n_i \cdot u_i^2$	$n_i \cdot u_i^3$	$n_i \cdot u_i^4$	$n_i(u_i+1)^4$
153	8	-3	-24	72	-216	648	128
159	10	-2	-20	40	-80	160	10
165	12	-1	-12	12	-12	12	0
171	14	0	0	0	0	0	14
177	10	1	10	10	10	10	160
183	6	2	12	24	48	96	486
$\sum$	$\sum n_i$ = 60		$\sum n_i u_i$ = -34	$\sum n_i u_i^2$ = 158	$\sum n_i u_i^3$ = -250	$\sum n_i u_i^4$ = 926	$\sum$= **798**

Выполним проверку. $\sum n_i u_i^4 + 4\sum n_i u_i^3 + 6\sum n_i u_i^2 + 4\sum n_i u_i + n = \sum n_i (u_i+1)^4$

926+4*(-250)+6*158+4*(-34)+60= -74+948+26=798

б) Условные начальные моменты найдем по формулам:

$M_1 = \frac{\sum n_i u_i}{n}$ $M_1 = \frac{-34}{60} \approx -0{,}56$

$M_2 = \frac{\sum n_i u_i^2}{n}$ $M_2 = \frac{158}{60} \approx 2{,}63$

$M_3 = \frac{\sum n_i u_i^3}{n}$ $M_3 = \frac{-250}{60} = -4{,}2$

$M_4 = \frac{\sum n_i u_i^4}{n}$ $M_4 = \frac{926}{60} \approx 15{,}4$

в) Выборочная средняя $\bar{x}_в = M_1 \cdot h + C \quad \Rightarrow \quad \bar{x}_в = -0{,}56 \cdot 6 + 171 = 167{,}6$

г) Дисперсия (рассеивание) – характеристика рассеяния значений случайной величины X около ее математического ожидания.

Выборочная дисперсия

$D_в = [M_2 - M_1^2] \cdot h^2 \quad D_в = [2{,}63 - (-0{,}56)^2] \cdot 6^2 = [2{,}63 - 0{,}336] \cdot 36 \approx 82{,}4$

Среднее квадратичное отклонение также служит для оценки рассеяния случайной величины X вокруг ее среднего значения $\sigma_в = \sqrt{D_в} \quad \sigma_в = \sqrt{82{,}4} \approx 9{,}1$

Исправленная дисперсия генеральной совокупности

$s^2 = \frac{n}{n-1} D_в = 60/59 * 82{,}4 \approx 83{,}7966$

Исправленное среднее квадратичное отклонение $s = \sqrt{s^2} \approx 9{,}15$

д) Коэффициентом вариации V

$V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\%; \quad V = \frac{9{,}1}{167{,}6} \cdot 100\% \approx 5{,}4\%$

е) Асимметрия (коэффициент асимметрии)

Для вычисления асимметрии и эксцесса найдем центральные моменты:

$\mu_3 = [M_3 - 3M_1M_2 + 2M_1^3] \cdot h^3 \qquad \mu_3 = [(-4{,}2) - 3(-0{,}56) \cdot 2{,}63 + 2(-0{,}56)^3] \cdot 6^3 \approx$ -34,56

$\mu_4 = [M_4 - 4M_1M_3 + 6M_1^2M_2 - 3M_1^4] \cdot h^4$

$\mu_4 = [15{,}4 - 4(-0{,}56)(-4{,}2) + 6(-0{,}56)^2 \cdot 2{,}63 - 3(-0{,}56)^4] \cdot 6^4 \approx 13478{,}4$

Асимметрия $A_s = \frac{\mu_3}{\sigma_в^3} \qquad A_s = \frac{-34{,}56}{(9{,}1)^3} = \frac{-34{,}56}{753{,}6} \approx 0{,}05$

и) Эксцесс вычисляем по формуле:

$$E_x = \frac{\mu_4}{\sigma_в^4} - 3; \qquad E_x = \frac{13478{,}4}{(9{,}1)^4} - 3 = \frac{13478{,}4}{6857{,}76} - 3 \approx 1{,}965 - 3 \approx -1{,}03$$

Задание 5. Определение границ истинных значений числовых характеристик изучаемой случайной величины с заданной надёжностью

Найдем доверительные интервалы, покрывающие неизвестный параметр с надежностью γ=0,95.

а) Доверительный интервал для оценки математического ожидания нормального распределения при неизвестном среднем квадратичном отклонении генеральной совокупности с заданной надежностью найдем по формуле: $\bar{x}_в - t_\gamma \cdot \frac{s}{\sqrt{n}} < \alpha < \bar{x}_в + t_\gamma \cdot \frac{s}{\sqrt{n}}$.

По условию: $\bar{x}_в = 167{,}6$, $s = 9{,}15$, $\gamma = 0{,}95, n = 60.$

По таблице $t_\gamma = t(\gamma; n) = t(0{,}95; 60) = 2{,}001$

Подставим все значения в формулу: $167{,}6 - 2{,}001 \cdot \frac{9{,}15}{\sqrt{60}} < \alpha < 167{,}6 + 2{,}001 \cdot \frac{9{,}15}{\sqrt{60}}$

$165{,}24 < \alpha < 169{,}96$**,** округлим до см. $165 < \alpha < 170$

б) Доверительный интервал для оценки среднего квадратичного отклонения нормального распределения, покрывающий σ с заданной надежностью γ=0,95 найдем по формуле: $s(1-q) < \sigma_Г < s(1+q)$.

По условию $s = 9{,}15$, γ=0,95, n=60. По таблице q=q(γ;n)=0,188

Следовательно: $9{,}15 \cdot (1 - 0{,}188) < \sigma_Г < 9{,}15 \cdot (1 + 0{,}188) \Rightarrow 7{,}43 < \sigma_Г < 10{,}87$

Округлим до см. $7 < \sigma_Г < 11$

Задание 6. Содержательная интерпретация результатов первичной обработки по условию задачи

1. *Интерпретация* $\bar{x}_в$.

Рост студента – величина случайная, ее выборочные значения изменяются в частности от 153 см. до 186 см, однако среднее значение равно 168 см.

2. *Интерпретация доверительного интервала* $\bar{x}_в - \frac{t \cdot_\gamma s}{\sqrt{n}} < a < \bar{x}_в + \frac{t_\gamma \cdot s}{\sqrt{n}}$.

Так как $165 < \alpha < 170$при заданной надежности γ=0,95, то можно утверждать с вероятностью 0,95, что средний рост студента будет колебаться в пределах от 165 до 170 см, то есть из 100 студентов примерно 95 будут иметь средний рост в указанных пределах, в интервале (165; 170) будут находиться основные значения роста студентов.

3. Интерпретация среднего квадратичного отклонения $\sigma_в$.

Так как $\sigma_в$= 9,1 см., то можно сделать вывод, что отклонение роста отдельно взятого студента в среднем составляет 9,1 см.

4. Интерпретация доверительного интервала $s \cdot (-q) < \sigma_Г < s \cdot (+q)$.

По расчетам получили, что $7 < \sigma_Г < 11$, следовательно, можно сделать вывод, что возможные отклонения роста студента с вероятностью 0,95 будут составлять значения, заключенные в промежутке (7; 11) см.

5. Интерпретация коэффициента асимметрии A_s.

Значение $A_S \neq 0$, это говорит о том, что изменения роста студентов в сторону увеличения или уменьшения по отношению к среднему значению происходит неодинаково, так как A_s=0,05>0, то наблюдается левосторонний скос, т.е. рост выше среднего событие более достоверное.

6. Интерпретация полученного значения эксцесса E_x.

Значение E_x= - 1,03, т.к. $E_x \neq 0$, то отклонение от среднего значения $\bar{x}_в$ происходит не плавно и наблюдается отклонение от нормы, а так как E_x=-1,03<0, E_x то отклонения большие, больше нормы, то есть среднее значение резко отличается от других возможных значений случайной величины.

7. Интерпретация значения коэффициента вариации *V.*

Значение V=5,4% - это меньше 10%, следовательно, изменчивость незначительная.

1.2 Задания для самостоятельной работы по теме: «Первичная обработка экспериментальных данных»

1. Дана величина заработной платы специалистов компании (в сотнях руб.).

120	480	175	490	410	425	430	385	335	315	545	265	390	395	360
475	480	225	445	255	425	375	325	320	160	245	395	390	370	380
255	265	445	425	265	410	305	330	335	455	220	370	340	340	
265	275	435	275	415	310	340	245	215	215	275	325	330	340	
355	415	285	265	315	345	345	230	225	285	365	315	255	275	
325	335	295	275	330	315	340	345	215	295	375	345	325	325	
225	255	370	385	355	360	285	345	305	310	235	325	320	335	

2. Имеются данные социологического опроса студентов о влиянии престижа вуза на выбор профессии (в %):

85	76	80	84	88	89	91	88	75	85	79	82	87	90	83
76	82	86	89	88	84	90	89	85	91	89	95	89	79	86
87	81	78	85	88	91	89	87	74	81	85	91	97	90	
87	90	88	86	76	84	88	77	88	82	95	91	84	91	
85	84	74	80	84	91	99	90	87	77	88	85	81	88	
83	89	91	92	88	94	90	88	81	83	90	98	92	86	
89	94	96	89	88	95	99	90	86	78	87	90	93	90	

3. Опрашивалось 100 человек о доверии руководству фирмы. Получены данные (в %).

78	90	90	86	81	77	83	85	92	86	84	84	96	78	84
73	75	83	78	73	84	85	83	88	76	89	78	92	85	83
87	85	87	89	83	76	77	84	83	89	79	81	76	86	
87	76	82	89	74	89	82	87	71	78	75	80	92	74	
85	84	81	83	88	81	83	80	79	82	83	88	81	82	
86	74	91	78	93	84	81	76	74	81	82	86	76	90	
93	83	92	91	83	79	84	90	80	84	81	78	88	85	

1.3 Вопросы для самопроверки

1. Какая совокупность называется генеральной (выборочной)?
2. Какая выборка называется репрезентативной?
3. Какие способы формирования выборки вы знаете?
4. Когда выборка называется повторной (бесповторной)?
5. Какие вариационные ряды вы знаете?

6. Можно ли от дискретного ряда перейти к интервальному и наоборот?

7. От чего зависит число интервалов группировки?

8. Как от простой статистической таблицы данных перейти к вариационному ряду?

9. Как графически изобразить дискретный (непрерывный) вариационный ряд?

10. Каким свойством обладает выборочное среднее?

11. Как вычислить моду и медиану дискретного (интервального) вариационного ряда?

12. Как вычислить дисперсию (среднее квадратичное отклонение) выборочной совокупности?

13. Какими свойствами обладает дисперсия?

14. Что характеризует коэффициент асимметрии?

15. Что характеризует эксцесс?

16. Какое число принимается в качество ложного нуля?

17. По каким формулам от числовых характеристик, вычисленных в условных вариантах производится переход к числовым характеристикам в первоначальных вариантах?

2. ПАРНАЯ РЕГРЕССИЯ. МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (МНК)

2.1 Теоретические основы

Рассмотрим простейшую модель $y = \alpha + \beta x + \varepsilon$. Величина y рассматривается как зависимая переменная, состоящая из двух частей: неслучайной составляющей $\alpha + \beta x$, где x – объясняющая переменная, α и β - параметры, ε - случайный член. Имеется несколько причин включения случайного члена.

1. Невключение объясняющих переменных. Соотношение между x и y является упрощением, и существуют другие факторы, влияющие на y. Или переменные, которые мы хотели бы включить, не можем измерить их, например, психологический фактор. Или мы просто не знаем пока какие ещё переменные влияют на y.

4. Агрегирование переменных. Во многих случаях рассматриваемая зависимость – это попытка объединить вместе некоторое число микроэкономических соотношений. Например, функция суммарного потребления, т.е. объединение решений многих индивидов. Наблюдаемое расхождение объясняет случайный член.

5. Неправильное описание структуры. Структура модели неправильна или не вполне правильна. Например, y зависит не от фактического x, а от y_{t-1} – предыдущего значения, при этом может казаться, что между x и y существует связь. Расхождения при этом описываются ε .

6. Неправильная функциональная спецификация. Математически зависимость x и y описывается не так. Например, зависимость не является линейной.

7. Ошибки измерения. Неизбежны.

Таким образом, ε является суммарным проявлением всех этих причин.

Линейная регрессия сводится к нахождению уравнения вида:

$$\tilde{y}_x = a + b * x + \varepsilon$$

Уравнение такого вида позволяет по заданным значениям фактора *x* иметь теоретические значения результативного признака подстановкой в него фактических значений фактора x.

Построение линейной регрессии сводится к оценке ее параметров — *a* и *b*. Оценки параметров линейной регрессии могут быть найдены разными методами. Можно обратиться к полю корреляции и, выбрав на графике две точки, провести через них прямую линию, затем по графику найти значения параметров. Параметр *a* определим как точку пересечения линии регрессии с

осью *oy,* а параметр *b* оценим исходя из угла наклона линии регрессии как *dy/dx*, где *dy* — приращение результата *y,* а *dx* — приращение фактора *x.*

Классический подход к оцениванию параметров линейной регрессии основан на *методе наименьших квадратов* (МНК).

Можно воспользоваться следующими формулами для определения параметров значений *a* и *b:*

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2}$$

$$b = \frac{\overline{yx} - \bar{y} * \bar{x}}{x^2 - \bar{x}^2}$$

Параметр ***b*** называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу. Знак при коэффициенте регрессии ***b*** показывает направление связи: при $b > 0$ — связь прямая, а при $b < 0$ — связь обратная.

Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии достаточно распространенным в эконометрических исследованиях.

Формально *a* — значение *y* при $x = 0$. Если признак-фактор *x* не имеет и не может иметь нулевого значения, то трактовка свободного члена *a* не имеет смысла. Параметр *a* может не иметь экономического содержания. Попытки экономически интерпретировать параметр *a* могут привести к абсурду, особенно при a<0.

Интерпретировать можно лишь знак при параметре ***a.*** Если $a > 0$, то относительное изменение результата происходит медленнее, чем изменение фактора. Иными словами, вариация результата меньше вариации фактора — коэффициент вариации по фактору *x* выше коэффициента вариации для результата *y: Vx> Vy.*

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя

выступает **линейный коэффициент корреляции** r_{xy}. Имеются разные модификации формулы линейного коэффициента корреляции, например:

$$r_{xy} = \frac{\sum((x-\bar{x})(y-\bar{y}))}{\sqrt{\sum(y-\bar{y})^2 * \sum(x-\bar{x})^2}} \qquad \text{или} \qquad r_{xy} = b\frac{\sigma_x}{\sigma_y}$$

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} \qquad \sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

Как известно, линейный коэффициент корреляции находится в границах - 1 < r_{xy} < 1.

Если коэффициент регрессии $b > 0$, то $0 < r_{xy} < 1$, и, наоборот, при $b<0$ - $1<r_{xy}<0$.

Следует иметь в виду, что величина линейного коэффициента корреляции оценивает тесноту связи рассматриваемых признаков в ее линейной форме. Поэтому близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствия связи между признаками. При иной спецификации модели связь между признаками может оказаться достаточно тесной.

И, наконец, коэффициенты регрессии могут быть определены с помощью ППП Excel, Statgraphic.

Пример 1. По 10 сельскохозяйственным предприятиям имеются данные о себестоимости молока и средней продуктивности молока (табл. 6).

Таблица 6

Себестоимость молока, руб./л	7,5	6,0	5,2	8,3	5,8	6,9	7,8	7,0	5,9	8,0
Средняя продуктивность молока, кг	197	168	135	259	151	186	233	213	147	234

Требуется: Рассчитать параметры уравнения парной линейной регрессии зависимости себестоимости молока от средней продуктивности.

Решение.

Уравнение парной линейной регрессии имеет вид: $\hat{y}_x = a + bx$,

где $\hat{y}_x$ – себестоимость молока, руб./л; x – средняя продуктивность молока, кг; *a, b* – параметры уравнения.

Для определения параметров уравнения *a* и *b* составим систему нормальных уравнений. Исходное уравнение последовательно умножим на коэффициенты при неизвестных *a* и *b*, и затем каждое уравнение просуммируем:

$$\begin{cases} \Sigma y = an + b\Sigma x \\ \Sigma yx = a\Sigma x + b\Sigma x^2 \end{cases},$$

где n – число единиц совокупности.

Для расчетов построим вспомогательную таблицу (табл. 7).

Таблица 7

Сельскохозяйственное предприятие	Себестоимость молока, руб./л	Средняя продуктивность молока, кг	y^2	x^2	xy	$\hat{y}_x$	$y-\hat{y}_x$
	y	x					
1	7,5	197	56,25	38809	1477,5	6,96	0,54
2	6,0	168	36,00	28224	1008,0	6,25	-0,25
3	5,2	135	27,04	18225	702,0	5,44	-0,24
4	8,3	259	68,89	67081	2149,7	8,48	-0,18
5	5,8	151	33,64	22801	875,8	5,83	-0,03
6	6,9	186	47,61	34596	1283,4	6,69	0,21
7	7,8	233	60,84	54289	1817,4	7,84	-0,04
8	7,0	213	49,00	45369	1491,0	7,35	-0,35
9	5,9	147	34,81	21609	867,3	5,73	0,17
10	8,0	234	64,00	54756	1872,0	7,86	0,14
Сумма	68,4	1923	478,09	385759	13544,1	68,43	-0,03

Подставим полученные данные в систему уравнений:

$$\begin{cases} 68{,}4 = 10a + 1923b \\ 13544{,}1 = 1923a + 385759b \end{cases}$$

Разделим каждый член уравнений на коэффициенты при a (в первом уравнении на 10, во втором на 1925):

$$\begin{cases} 6{,}84 = a + 192{,}3b \\ 7{,}043 = a + 200{,}6b \end{cases}$$

Вычтем из второго уравнения первое и найдем параметр *b*:

$$0{,}203 = 8{,}3b;\ \ b = 0{,}0245.$$

Подставив значение b в первое уравнение, найдем значение a:

$$a = 6{,}84 - 192{,}3 \cdot 0{,}0245 = 2{,}13.$$

Параметры уравнения регрессии можно определить и по другим формулам, которые вытекают из системы нормальных уравнений:

$$a = \frac{\Sigma y \Sigma x^2 - \Sigma xy \Sigma x}{n\Sigma x^2 - (\Sigma x)^2} = \frac{68{,}4 \cdot 385759 - 13544{,}1 \cdot 1923}{10 \cdot 385759 - 1923^2} = 2{,}13$$

$$b = \frac{n\Sigma xy - \Sigma y \Sigma x}{n\Sigma x^2 - (\Sigma x)^2} = \frac{10 \cdot 13544{,}1 - 68{,}4 \cdot 1923}{10 \cdot 385759 - 1923^2} = 0{,}0245$$

Уравнение регрессии имеет вид: $\hat{y}_x = 2{,}13 + 0{,}0245x$.

Пример 2. По данным проведенного опроса восьми групп семей известны данные связи расходов населения на продукты питания с уровнем доходов семьи.

Таблица 8

Расходы на продукты питания, y, тыс. руб.	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
Доходы семьи, x, тыс. руб.	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Решение.

Предположим, что связь между доходами семьи и расходами на продукты питания линейная. Для подтверждения нашего предположения построим поле корреляции.

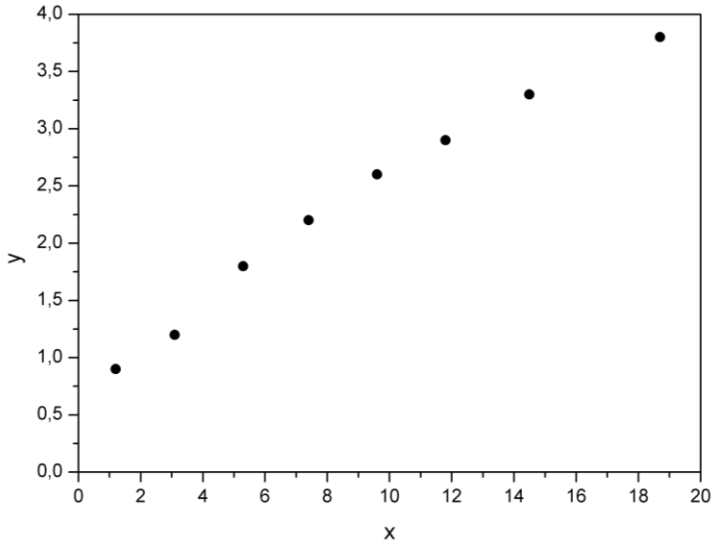


Рисунок 3 - Поле корреляции

По графику видно, что точки выстраиваются в некоторую прямую линию.

Для удобства дальнейших вычислений составим таблицу.

Таблица 9

	x	y	xy	x^2	y^2	$\hat{y}_x$	$y-\hat{y}_x$
1	1,2	0,9	1,08	1,44	0,81	1,038	–0,138
2	3,1	1,2	3,72	9,61	1,44	1,357	–0,157
3	5,3	1,8	9,54	28,09	3,24	1,726	0,074
4	7,4	2,2	16,28	54,76	4,84	2,079	0,121
5	9,6	2,6	24,96	92,16	6,76	2,449	0,151
6	11,8	2,9	34,22	139,24	8,41	2,818	0,082
7	14,5	3,3	47,85	210,25	10,89	3,272	0,028
8	18,7	3,8	71,06	349,69	14,44	3,978	–0,178
Итого	71,6	18,7	208,71	885,24	50,83	18,717	–0,017
Среднее значение	8,95	2,34	26,09	110,66	6,35	2,34	–
σ	5,53	0,935	–	–	–	–	–
σ^2	30,56	0,874	–	–	–	–	–

Рассчитаем параметры линейного уравнения парной регрессии $\hat{y}_x = a + b \cdot x$. Для этого воспользуемся формулами:

$$b = \frac{\operatorname{cov}\ x,\ y}{\sigma_x^2} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2} = \frac{26,09 - 8,95 \cdot 2,34}{30,56} = 0,168$$

$$a = \bar{y} - b \cdot \bar{x} = 2,34 - 0,168 \cdot 8,95 = 0,836.$$

Получили уравнение: $\hat{y}_x = 0,836 + 0,168 \cdot x$. Т.е. с увеличением дохода семьи на 1000 руб. расходы на питание увеличиваются на 168 руб.

2.2 Задания для самостоятельной работы по теме: «Парная регрессия. Метод наименьших квадратов (МНК)»

При изучении химического состава и реологических характеристик желе из ягод смородины было обследовано 10 образцов и получены следующие результаты о содержании сухих веществ X (%) и прочности студня Y (кг), варианты заданий представлены в таблице 9.

В соответствии с выбранным вариантом произвести статистическую обработку данных:

- построить диаграмму рассеяния;
- полагая, что между признаками X и Y имеет место линейная корреляционная вязь, найти выборочное уравнение линейной регрессии. Используя полученное уравнение регрессии, оценить ожидаемое среднее значение признака Y, когда признак X принимает значение, равное a %.
- построить линию регрессии.

Таблица 10

Варианты заданий											
№1		№2		№3		№4		№5		№6	
Прочность студня, кг	Содержание сухих веществ в студне, %	Прочность студня, кг	Содержание сухих веществ в студне, %	Прочность студня, кг	Содержание сухих веществ в студне, %	Прочность студня, кг	Содержание сухих веществ в студне, %	Прочность студня, кг	Содержание сухих веществ в студне, %	Прочность студня, кг	Содержание сухих веществ в студне, %
0,111	59,0	0,278	64,0	0,050	55,0	0,323	66,0	0,222	63,0	0,301	66,0
0,125	59,5	0,284	65,0	0,084	56,0	0,369	67,0	0,237	64,0	0,321	67,0
0,126	60,0	0,291	66,0	0,091	57,0	0,395	68,0	0,265	65,0	0,358	68,0
0,165	61,2	0,302	67,0	0,101	58,0	0,401	69,0	0,275	66,0	0,379	69,0
0,201	62,0	0,323	68,0	0,125	59,0	0,423	70,0	0,303	67,0	0,402	70,0
0,210	63,0	0,375	69,0	0,147	60,0	0,437	71,0	0,345	68,0	0,423	71,0
0,246	64,0	0,401	70,0	0,175	61,0	0,479	72,0	0,378	69,0	0,435	72,0
0,302	65,0	0,421	71,0	0,195	62,0	0,492	73,0	0,401	70,0	0,469	73,0
0,333	66,0	0,520	72,0	0,201	63,0	0,502	74,0	0,419	71,0	0,501	74,0
0,352	67,0	0,530	73,0	0,222	64,0	0,509	75,0	0,425	72,0	0,509	75,0
	a=61,0		a=65,5		a=56,5		a=67,5		a=65,5		a=68,5

2.3 Вопросы для самопроверки

1. Как рассчитываются параметры парной линейной регрессии?

2. Как провести оценку статистической значимости параметров уравнения парной регрессии?

3. Поясните смысл коэффициента корреляции, как оценить его значимость?

4. Как определяется число степеней свободы для факторной и остаточной сумм квадратов?

5. Дайте определение бета–коэффициента. Поясните его смысл.

6. Дайте определение коэффициента эластичности. Поясните его смысл. Как определяется коэффициент эластичности по разным видам регрессионных моделей?

3. ПРОВЕРКА КАЧЕСТВА РЕГРЕССИИ

3.1 Основные понятия

Для практического использования моделей регрессии большое значение имеет их адекватность, т.е. соответствие фактическим статистическим данным. При анализе адекватности уравнения регрессии (модели) исследуемому процессу, возможны следующие варианты:

1. Построенная модель на основе **F-критерия Фишера** в целом адекватна и все коэффициенты регрессии значимы. Такая модель может быть использована для принятия решений и осуществления прогнозов.

2. Модель по F-критерию Фишера адекватна, но часть коэффициентов не значима. Модель пригодна для принятия некоторых решений, но не для прогнозов.

3. Модель по F-критерию адекватна, но все коэффициенты регрессии не значимы. Модель полностью считается неадекватной. На ее основе не принимаются решения и не осуществляются прогнозы.

Корреляционный и регрессионный анализ, как правило, проводится для ограниченной по объёму совокупности. Поэтому показатели регрессии и корреляции – параметры уравнения регрессии, коэффициент корреляции и коэффициент детерминации могут быть искажены действием случайных факторов. Чтобы проверить, на сколько эти показатели характерны для всей генеральной совокупности, не являются ли они результатом стечения случайных обстоятельств, необходимо проверить адекватность построенных статистических моделей.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной. Чтобы иметь общее суждение о качестве модели, из относительных отклонений по каждому наблюдению определяют среднюю ошибку аппроксимации. Проверка адекватности уравнения регрессии (модели) осуществляется с помощью средней ошибки аппроксимации, величина которой не должна превышать 1**0-12%** (рекомендовано).

$$A = \frac{1}{n}\sum\left|\frac{y_i - \hat{y}_x}{y_i}\right| \cdot 100$$

Оценка значимости уравнения регрессии в целом производится на основе **F-критерия Фишера**, которому предшествует **дисперсионный анализ**. В математической статистике **дисперсионный анализ** рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели. Согласно основной идее **дисперсионного анализа, общая сумма квадратов отклонений** переменной (y) от среднего значения ($y_{ср.}$) раскладывается на две части – **«объясненную»** и **«необъясненную»**:

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

Схема дисперсионного анализа имеет следующий вид (n –число наблюдений, m–число параметров при переменной x**):**

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

Определение **дисперсии на одну степень свободы** приводит **дисперсии к сравнимому виду**. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину **F-критерия Фишера**. Фактическое значение F-критерия Фишера сравнивается с табличным значением $F_{табл.}$ (α, k_1, k_2) при заданном уровне значимости α и степенях свободы k_1= m и k_2=n-m-1.При этом, если фактическое значение F-критерия больше табличного $\mathbf{F_{факт} > F_{теор}}$, то признается **статистическая значимость уравнения в целом**. Для парной линейной регрессии m=1 , поэтому:

$$F = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}} = \frac{\sum\left(\hat{y}_x - \bar{y}\right)^2}{\sum\left(y - \hat{y}_x\right)^2} \cdot (n-2)$$

Эта формула в общем виде может выглядеть так:

$$F = \frac{\sigma^2_{факт}(n-m)}{\sigma^2_{ост}(m-1)} \Rightarrow сравнить\ с\ F_{табл}$$

Отношение **объясненной** части дисперсии переменной (y) к **общей дисперсии** называют **коэффициентом детерминации** и используют для характеристики **качества уравнения регрессии** или соответствующей модели связи. Соотношение между **объясненной** и **необъясненной** частями **общей дисперсии** можно представить в альтернативном варианте:

$$R^2 = \frac{\sum(\tilde{y}_x - \bar{y})^2}{\sum(y_i - \bar{y})^2} = 1 - \frac{\sum(y_i - \tilde{y}_x)^2}{\sum(y_i - \bar{y})^2}$$

Коэффициент детерминации $\mathbf{R^2}$ принимает значения в диапазоне от нуля до единицы $0 \le R^2 \le 1$. **Коэффициент детерминации $\mathbf{R^2}$** показывает, какая часть дисперсии результативного признака (y) объяснена уравнением регрессии. Чем больше R^2, тем большая часть дисперсии результативного признака (y) объясняется уравнением регрессии и тем лучше уравнение регрессии описывает исходные данные. При отсутствии зависимости между (y) и (x) коэффициент детерминации R^2 будет близок к нулю. Таким образом, **коэффициент детерминации $\mathbf{R^2}$** может применяться для **оценки качества**

(точности) **уравнения регрессии**. Возникает вопрос, при каких значениях R^2 уравнение регрессии следует считать статистически незначимым, что делает необоснованным его использование в анализе? Ответ на этот вопрос дает **F-критерий Фишера $F_{факт}>F_{теор}$** - делаем вывод о статистической значимости уравнения регрессии. Величина F–критерия связана с коэффициентом детерминации $R^2_{xy}(r^2_{xy})$, и ее можно рассчитать по следующей формуле:

$$F = \frac{r_{xy}^2}{1-r_{xy}^2} \cdot (n-2)$$

Либо при оценке значимости **индекса** (аналог **коэффициента**) **детерминации**:

$$i^2(\eta^2) = 1 - \frac{\sum (y_i - \tilde{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

Использование **коэффициента множественной детерминации** R^2 для оценки качества модели, обладает тем недостатком, что включение в модель нового фактора (даже несущественного) автоматически увеличивает величину R^2. Поэтому, при большом количестве факторов, предпочтительнее использовать, так называемый, улучшенный, скорректированный коэффициент множественной детерминации R^2, определяемый соотношением:

$$\overline{R}^2 = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 : (n-p-1)}{\sum_{i=1}^{n} (y_i - \bar{y})^2 : (n-1)} = 1 - \frac{n-1}{n-p-1}(1-R^2)$$

где p – число факторов в уравнении регрессии, n – число наблюдений. Чем больше величина p, тем сильнее различия между множественным коэффициентом детерминации R^2 и скорректированным R^2. При использовании скорректированного R^2, для оценки целесообразности включения фактора в уравнение регрессии, следует учитывать, что увеличение его величины (значения), при включении нового фактора, не обязательно свидетельствует о его значимости, так как значение увеличивается всегда, когда t-статистика больше единицы (|t|>1). При

заданном объеме наблюдений и при прочих равных условиях, с увеличением числа независимых переменных (параметров), скорректированный коэффициент множественной детерминации убывает. При небольшом числе наблюдений, скорректированная величина коэффициента множественной детерминации R^2 имеет тенденцию переоценивать долю вариации результативного признака, связанную с влиянием факторов, включенных в регрессионную модель. Низкое значение коэффициента множественной корреляции и коэффициента множественной детерминации R^2 может быть обусловлено следующими причинами: в регрессионную модель не включены существенные факторы; неверно выбрана форма аналитической зависимости, не реально отражающая соотношения между переменными, включенными в модель.

Для оценки значимости парного коэффициента корреляции (корень квадратный из коэффициента детерминации)**,** при условии линейной формы связи между факторами, можно использовать **t-критерий Стьюдента**:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = r\sqrt{\frac{n-2}{1-r^2}}$$

При численности объектов анализа до 30 единиц возникает необходимость проверки значимости (существенности) каждого коэффициента регрессии. При этом выясняют насколько вычисленные параметры характерны для отображения комплекса условий: не являются ли полученные значения параметров результатами действия случайных причин. Значимость коэффициентов простой линейной регрессии (применительно к совокупностям, у которых n<30) осуществляют с помощью **t-критерия Стьюдента**. При этом вычисляют расчетные (фактические) значения t-критерия для параметров a_0, a_1:

$$t_{a_0} = \frac{a_0\sqrt{n-2}}{\sigma_\varepsilon} \qquad t_{a_1} = \frac{a_1\sqrt{n-2}}{\sigma_\varepsilon}\sigma_x$$

$$\sigma_\varepsilon = \sqrt{\frac{\Sigma(Y_i - \tilde{Y})^2}{n-m}} \qquad \sigma_x = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-m}}$$

n-число наблюдений, m-число параметров уравнения регрессии, σ_ε - (остаточное) среднее квадратическое отклонение результативного признака от выровненных значений ŷ; σ_x - среднее квадратическое отклонение факторного признака от общей средней.

Вычисленные, по вышеприведенным формулам, значения сравнивают с критическими t, которые определяют по таблице значений Стьюдента с учетом принятого уровня значимости α и числа степеней свободы вариации $k(v)=n-2$. В социально-экономических исследованиях уровень значимости $\boldsymbol{\alpha}$ обычно принимают равным 0,05. Параметр признаётся значимым (существенным) при условии, если $t_{расч.} > t_{табл.}$ В этом случае, практически невероятно, что найденные значения параметров обусловлены только случайными совпадениями.

Критерий Дарбина – Уотсона (Durbin - Watson)

Оценивая качество уравнения регрессии, мы предполагаем, что реальная взаимосвязь переменных линейна. Отклонения от регрессионной прямой являются случайными, независимыми друг от друга величинами с нулевым математическим ожиданием и постоянной дисперсией. Если эти предположения не выполняются, то оценки коэффициентов регрессии не обладают свойствами *несмещенности, эффективности* и *состоятельности*. В этом случае анализ значимости полученных оценок будет неточным.

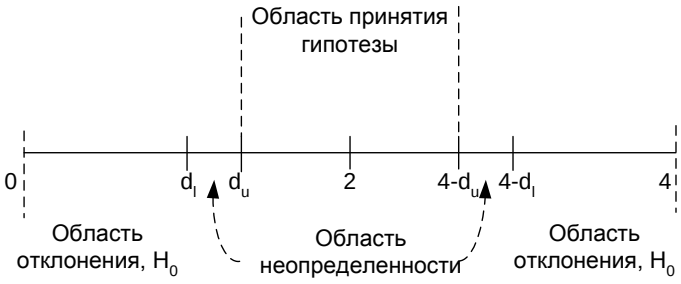
Статистика **Дарбина—Уотсона** используется для проверки гипотезы о том, что остатки построенной регрессионной модели некоррелированны (корреляции равны нулю), против альтернативы: остатки связаны авторегрессионной зависимостью (первого порядка) вида:

$$\varepsilon_i = \rho\, \varepsilon_{i-1} + \delta_i$$

На практике для анализа коррелированности отклонений вместо коэффициента корреляции используют тесно с ним связанную статистику Дарбина—Уотсона, рассчитываемую по формуле

$$d = \frac{\Sigma(e_i - e_{i-1})^2}{\Sigma e_i^2}.$$

Критические точки статистики Дарбина—Уотсона табулированы для различных α, m, n. При проверке гипотезы об отсутствии автокорреляции остатков используется числовой отрезок, на котором отложены $d_l -$ нижняя граница статистики и $d_u -$ верхняя граница:



Статистика Дарбина—Уотсона

Проверка гипотезы проводится по схеме:

- Если $d < d_l$, то гипотеза H_0 отклоняется, принимается H_1 - значительная положительная автокорреляция остатков;
- Если $d > 4 - d_l$, то гипотеза H_0 отклоняется, принимается H_1 - значительная отрицательная автокорреляция остатков;
- Если $d_u < d < 4 - d_u$, то гипотеза H_0 об отсутствии автокорреляции остатков принимается;
- Если $d_l < d < d_u$, или $4 - d_u < d < 4 - d_l$, то гипотеза об отсутствии автокорреляции не может быть ни принята, ни отклонена.

Не обращаясь к таблице критических точек Дарбина—Уотсона можно воспользоваться «грубым» правилом и считать, что автокорреляция остатков отсутствует, если $1{,}5 < d < 2{,}5$. Для более надежных выводов необходимо воспользоваться статистическими таблицами.

Пример. Рассчитать параметры уравнения парной линейной регрессии и проверить качество.

Таблица 11

x	23	29	33	35	38	40	43	45
y	88	83	73	78	68	53	58	48

Решение.

Построим поле корреляции. Как видно из рисунка 4 данные имеют линейную зависимость.

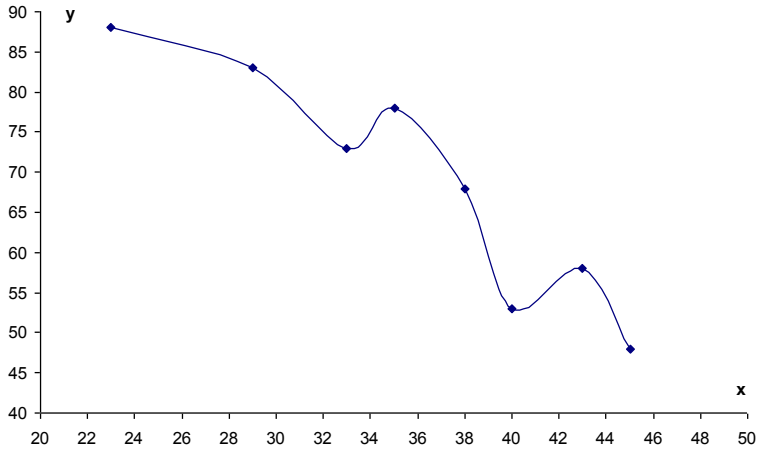


Рисунок 4 - Поле корреляции

Уравнение линейной регрессии имеет вид: $y = a_0 + a_1 \cdot x$.

Составим расчетную таблицу.

Таблица 12

x	y	xy	x^2	y^2	$\tilde{y}$	$y-\tilde{y}$	$(y-\tilde{y})^2$
23	88	2024	529	7744	92,225	-4,225	17,850625
29	83	2407	841	6889	81,119	1,881	3,538161
33	73	2409	1089	5329	73,715	-0,715	0,511225
35	78	2730	1225	6084	70,013	7,987	63,792169
38	68	2584	1444	4624	64,46	3,54	12,531600
40	53	2120	1600	2809	60,758	-7,758	60,186564
43	58	2494	1849	3364	55,205	2,795	7,812025
45	48	2160	2025	2304	51,503	-3,503	12,271009
286	**549**	**18928**	**10602**	**39147**	**548,998**	**0,002**	**178,493378**

Получим:

$$\overline{xy} = \frac{18928}{8} = 2366 \qquad \bar{y} = \frac{549}{8} = 68{,}625 \qquad \bar{x} = \frac{286}{8} = 35{,}75$$

$$\overline{x^2} = \frac{10602}{8} = 1325{,}25 \qquad \overline{y^2} = \frac{39147}{8} = 4893{,}375$$

Следовательно:

$$a_1 = \frac{2366 - 35{,}75 \cdot 68{,}625}{1325{,}25 - (35{,}75)^2} = \frac{-87{,}34375}{47{,}1875} \approx -1{,}851$$

$$a_0 = 68{,}625 + 1{,}851 \cdot 35{,}75 \approx 134{,}798$$

Уравнение примет вид: $\tilde{y} = 134{,}798 - 1{,}851 \cdot x$.

Рассчитаем коэффициент парной корреляции:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}, \quad \text{где } \sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2} = \sqrt{1325{,}25 - (35{,}75)^2} \approx 6{,}869$$

$$\sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2} = \sqrt{4893{,}375 - (68{,}625)^2} = \sqrt{183{,}984375} \approx 13{,}564$$

Тогда $r_{xy} = \frac{2366 - 35{,}75 \cdot 68{,}625}{6{,}869 \cdot 13{,}564} \approx -0{,}937$

Связь между факторами обратная и сильная.

Коэффициент детерминации $r_{xy}^2 \approx 0{,}878$, т.е. 87,8 % вариации изменения y объясняется вариацией изменения x.

Ошибка модели рассчитывается следующим образом:

$$\Delta = \frac{\sqrt{\frac{\sum (y_t - \tilde{y}_t)^2}{n - m - 1}}}{\bar{y}_t} \cdot 100 = \frac{\sqrt{\frac{178{,}493378}{8 - 1 - 1}}}{68{,}625} \cdot 100 \approx 7{,}9\%$$

Модель хорошего качества.

Оценим качество модели с помощью F-критерия Фишера.

$$F_{факт} = \frac{r_{xy}^2}{1 - r_{xy}^2}(n - 2) \Rightarrow F_{факт} = \frac{0{,}878}{1 - 0{,}878}(8 - 2) = 43{,}18$$

При уровне значимости $\alpha = 0{,}05$ (вероятность 95 %) табличное значение равно 5,99. $F_{факт} > F_{табл}$ т.е., гипотеза о случайной природе оцениваемых характеристик не принимается, признается их значимость. Следовательно, модель надежна.

Вычислим фактическое значение критерия Дарбина-Уотсона для этой модели (таблица 13):

Таблица 13

Расчет критерия Дарбина-Уотсона

ε_t	ε_{t-1}	$\varepsilon_t - \varepsilon_{t-1}$	$(\varepsilon_t - \varepsilon_{t-1})^2$	ε_t^2
-4,225	-	-	-	17,850625
1,881	-4,225	6,106	37,283236	3,538161
-0,715	1,881	-2,596	6,739216	0,511225
7,987	-0,715	8,702	75,724804	63,792169
3,54	7,987	-4,447	19,775809	12,5316
-7,758	3,54	-11,298	127,644804	60,186564
2,795	-7,758	10,553	111,365809	7,812025

-3,503	2,795	-6,298	39,664804	12,271009
0,002	**3,505**	**0,722**	**418,198482**	**178,493378**

$$d_{расч} = \frac{\sum_{t=2}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^{\,2}} = \frac{418{,}198482}{178{,}493378} = 2{,}34$$

Зададим уровень значимости $\alpha = 0{,}05$. По таблицам значений критерия Дарбина-Уотсона определим для числа наблюдений n = 8 и числа независимых переменных модели k = 1 критические значения $d_1 = 1{,}08$ и $d_2 = 1{,}36$. $d_{расч} > d_2$, следовательно, форма модели выбрана верно.

3.2 Задания для самостоятельной работы по теме: «Проверка качества регрессии»

Проверить качество моделей, рассчитанных в предыдущем разделе.

3.3 Вопросы для самопроверки

1. Перечислите три группы оценок качества моделирования.

2. Перечислите показатели близости и адекватности.

3. Что такое поле корреляции.

4. Что такое коэффициент детерминации? Что он показывает?

5. Как определяется число степеней свободы для факторной и остаточной сумм квадратов?

6. Как используется F – критерий Фишера для оценки статистической надежности результатов регрессионного моделирования?

9. Дайте определение коэффициента эластичности. Поясните его смысл.

4. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

4.1 Основные понятия

В линейной множественной регрессии $\tilde{y}_x = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \text{K} + b_p \cdot x_p$, параметры при x называются коэффициентами «чистой» регрессии. Они характеризуют среднее изменение результата с изменением соответствующего параметра на единицу при неизменном значении других факторов, закрепленных на среднем уровне.

Классический подход к оцениванию параметров линейной модели основан на **методе наименьших квадратов (МНК).**

Этот метод позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака (y) от расчетных (теоретических) $\tilde{y}_x$ минимальна: $\sum_i (y_i - \tilde{y}_{x_i})^2 \to \min$.

Чтобы найти минимум функции, надо вычислить производные по каждому из параметров и приравнять их к нулю, т.к. равенство нулю производной – необходимое условие экстремума. В результате получается система уравнений, решение которой и позволяет получить оценки параметров регрессии.

Система нормальных уравнений имеет вид:

$$\begin{cases} \sum y = a \cdot n + b_1 \cdot \sum x_1 + b_2 \sum x_2 + \text{K}\ b_p \cdot \sum x_p \\ \sum y \cdot x_1 = a \cdot \sum x_1 + b_1 \cdot \sum x_1^2 + b_2 \sum x_1 \cdot x_2 + \text{K}\ b_p \cdot \sum x_1 \cdot x_p \\ \text{K K K K K K K K K K K K K K K K K K K K K K K} \\ \sum y \cdot x_p = a \cdot \sum x_p + b_1 \cdot \sum x_p \cdot x_1 + b_2 \sum x_p \cdot x_2 + \text{K}\ b_p \cdot \sum x_p^2 \end{cases}$$

Решение системы может быть осуществлено по одному из известных способов: Метод Гаусса, метод Крамера и т.д.

Пример 1. По четырем предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_2 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_1 (%). Требуется написать уравнение множественной регрессии.

Номер предприятия	1	2	3	4

x_1, (%)	1	2	3	5
x_2, (%)	0	1	3	4
y, (тыс. руб.)	6	11	19	28

Решение.

Предположим, что зависимость выработки продукции на одного работника характеризуется следующим уравнением: $\tilde{y}_x = a + b_1 \cdot x_1 + b_2 \cdot x_2$.

На основании исходных данных составляем систему уравнений для определения коэффициентов a, b_1 и b_2.

$$\sum y = 6+11+19+28 = 64;$$

$$\sum x_1 = 1+2+3+5 = 11; \quad \sum x_2 = 0+1+3+4 = 8;$$

$$\sum y \cdot x_1 = 6 \cdot 1 + 11 \cdot 2 + 19 \cdot 3 + 28 \cdot 5 = 225;$$

$$\sum y \cdot x_2 = 6 \cdot 0 + 11 \cdot 1 + 19 \cdot 3 + 28 \cdot 4 = 180;$$

$$\sum x_1^2 = 1^2 + 2^2 + 3^2 + 5^2 = 39; \quad \sum x_2^2 = 0^2 + 1^2 + 3^2 + 4^2 = 26;$$

$$\sum x_1 \cdot x_2 = 1 \cdot 0 + 2 \cdot 1 + 3 \cdot 3 + 5 \cdot 4 = 31.$$

$$\begin{cases} 64 = 4 \cdot a + 11 \cdot b_1 + 8 \cdot b_2 \\ 225 = 11 \cdot a + 39 \cdot b_1 + 31 \cdot b_2 \\ 180 = 8 \cdot a + 31 \cdot b_1 + 26 \cdot b_2 \end{cases}$$

Решим эту систему по методу Крамера. Вычисляем определитель системы:

$$\Delta = \begin{vmatrix} 4 & 11 & 8 \\ 11 & 39 & 31 \\ 8 & 31 & 26 \end{vmatrix} = 4 \cdot 39 \cdot 26 + 11 \cdot 31 \cdot 8 + 11 \cdot 31 \cdot 8 - 8 \cdot 39 \cdot 8 - 31 \cdot 31 \cdot 4 - 11 \cdot 11 \cdot 26 = 26.$$

Аналогично вычисляем частные определители, заменяя соответствующий столбец столбцом свободных членов:

$$\Delta_1 = \begin{vmatrix} 64 & 11 & 8 \\ 225 & 39 & 31 \\ 180 & 31 & 26 \end{vmatrix} = 62; \quad \Delta_2 = \begin{vmatrix} 4 & 64 & 8 \\ 11 & 225 & 31 \\ 8 & 180 & 26 \end{vmatrix} = 88; \quad \Delta_3 = \begin{vmatrix} 4 & 11 & 64 \\ 11 & 39 & 225 \\ 8 & 31 & 180 \end{vmatrix} = 56.$$

Коэффициенты уравнения определяются по формулам:

$$a = \frac{\Delta_1}{\Delta} = \frac{62}{26} \approx 2{,}4; \quad b_1 = \frac{\Delta_2}{\Delta} = \frac{88}{26} \approx 3{,}4; \quad b_2 = \frac{\Delta_3}{\Delta} = \frac{56}{26} \approx 2{,}2.$$

Таким образом, уравнение имеет вид:

$$\tilde{y}_x = 2{,}4 + 3{,}4 \cdot x_1 + 2{,}2 \cdot x_2 .$$

Пример 2. В таблице 10 содержатся данные по пяти предприятиям о прибыли y (млн. руб.), выработке продукции на одного работника x_1 (единиц) и доле продукции, производимой на экспорт x_2 (%). Построить уравнение y по x_1, x_2.

Таблица 14

x_1	11	10	12	18	15	66
x_2	3	2	4	10	11	30
y	2	1	3	8	7	21

Решение.

Уравнение с двумя факторами имеет вид: $\hat{y} = a + b_1 x_1 + b_2 x_2$.

Воспользуемся для вычислений программой Microsoft Excel (Пакет Анализ данных).

корреляция	*Столбец 1*	*Столбец 2*	*Столбец 3*			
Столбец 1	1					
Столбец 2	0,91348012	1				
Столбец 3	0,97666433	0,978600184	1			
ВЫВОД ИТОГОВ						
Регрессионная статистика						
Множественный R	0,99950104					
R-квадрат	0,99900233					
Нормированный R-квадрат	0,99800466					
Стандартная ошибка	0,13912167					
Наблюдения	5					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	2	38,76129032	19,38064516	1001,333333	0,000997672	
Остаток	2	0,038709677	0,019354839			
Итого	4	38,8				
	Коэффициент	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	-4,4129032	0,480504418	-9,183897305	0,011649439	-6,48034687	-2,345459582
x1	0,47580645	0,052264038	9,103897648	0,011851434	0,250932446	0,700680457
x2	0,38870968	0,040867289	9,511511307	0,01087357	0,212871927	0,564547428

Рисунок 5 - Вывод итогов двухфакторной линейной регрессии.

Естественная двухфакторная регрессия принимает форму

$$\hat{y} = -4{,}413 + 0{,}476 x_1 + 0{,}389 x_2 .$$

Оценку математической погрешности приближения реальных значений y_j представляет средняя ошибка $\overline{A}\%$, определяемая на основании расчетов в таблице 15.

Таблица 15

n	x_1	x_2	y	$\hat{y}$	$(y-\overline{y})^2$	$(\hat{y}-\overline{y})^2$	$(y-\hat{y})^2$	$\left\|\frac{y_j-\hat{y}_j}{y_j}\right\|\cdot 100$
1	11	3	2	1,988	4,84	4,893	0,0001	0,60
2	10	2	1	1,124	10,24	9,462	0,0154	12,40
3	12	4	3	2,852	1,44	1,817	0,0219	4,93
4	18	10	8	8,036	14,44	14,715	0,0013	0,45
5	15	11	7	7,000	7,84	7,874	0,0000	0,00
Σ	**66**	**30**	**21**	**21**	38,8	38,761	0,0387	18,38

Итак, $\overline{A}=\frac{100}{5}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_i}\right|=\frac{18{,}38}{5}\cong 3{,}68\,\%$;

Величина приемлемой ошибки колеблется в пределах 5÷10%.

Индекс (линейный коэффициент) множественной корреляции вычисляется по одной из формул:

$$R_{yx_1x_2}=\sqrt{1-\frac{S_{ост}}{S}}$$

$$R_{yx_1x_2}=\sqrt{1-\frac{0{,}0387}{38{,}8}}=\sqrt{0{,}500\cdot 0{,}977+0{,}523\cdot 0{,}979}\cong 0{,}9995.$$

Направленность связи результата с факторами прямая.

Коэффициент детерминации $R^2=(0{,}9995)^2=0{,}9990$ означает долю объясненной дисперсии (вариации) результата по факторам в общей дисперсии результата.

Доля вариации исследуемого признака *y*, объясненная уравнением регрессии, составляет 99,9 %.

Доля вариации *y*, объясняемая неучтенными признаками, составляет 0,1%.

Надежность уравнения в целом по общему критерию Фишера может быть проверена с помощью следующих рассуждений.

Наблюдаемое значение общего критерия Фишера вычисляется по формуле

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

F-критерий принимает значение:

$$F_{набл} = \frac{38{,}7613}{2\cdot\frac{0{,}0387}{5-3}} \cong 1001{,}333.$$

Критическое значение $F_{крит}$ находят в таблице Фишера-Снедекора.

По заданному уровню значимости α (например, α=0,05) гипотезы $H_0 : R^2_{yx_1x_2} = 0$ определяется критическая точка из условия для вероятности

$$P(F_{набл} > F_{кр}(\alpha, n_1, n_2)) = \alpha, \quad F_{кр}(0{,}05{,}2{,}2) = 19{,}00;$$

$F_{набл} = 1001{,}333 > F_{кр}(0{,}05, n_1 = 2, n_2 = 2) = 19$. Гипотезу H_0 следует отвергнуть и принять $H_1 : R_{yx1x2}{}^2 \neq 0$.

Таким образом, уравнение статистически значимо в целом для опытных данных.

Сравнение наблюдаемых значений $T_{I\ набл}$ и $T_{кр}$ показывает статистическую значимость полученных оценок при выдвинутой гипотезах

$$H_0 : \hat{\lambda} = 0, \ \alpha = 0{,}05, \ \hat{\lambda} = \hat{b}_1, \ \hat{\lambda} = \hat{b}_2.$$

Случайные ошибки в оценке b_1, b_2,

$$m_{\hat{b}1} = \frac{0{,}476}{9{,}165} = 0{,}0519, \ m_{\hat{b}2} = \frac{0{,}389}{9{,}487} = 0{,}0410.$$

Точечный прогноз неизвестных параметров b_1, b_2 представляется в виде b_1=0,476±0,052, b_2=0,389±0,041 .

4.2 Задания для самостоятельной работы по теме: «Множественная линейная регрессия»

1. По данным 30 наблюдений постройте модель множественной регрессии удовлетворительного качества (табл. 16). Рассчитайте прогноз

результата, если прогнозные значения независимых факторов будут составлять 112% от их среднего уровня.

Таблица 16

№	Валовой продукт, млн. руб.	Балансовая стоимость оборудования, млн. руб.	Объем промышленного производства, млн. руб.	Количество занятых, тыс.чел.
i	*y*	*x*1	*x*2	*x*3
1	1365	3938	625	161
2	1398	3002	475	186
3	1969	6269	528	267
4	1000	2270	242	129
5	761	1879	197	100
6	1253	4709	407	149
7	1590	3976	702	144
8	1425	2946	521	153
9	1127	3151	369	128
10	1595	3424	447	178
11	1636	4110	645	207
12	2110	3910	717	171
13	1131	3045	299	99
14	2005	4575	464	166
15	768	2126	169	105
16	1682	4692	579	167
17	2146	5873	468	255
18	2865	6906	850	299
19	3133	8678	924	458
20	1706	3988	507	182
21	1456	3840	475	179
22	2616	6368	801	244
23	2657	6396	604	304
24	1538	3632	592	170
25	1249	2681	220	111
26	2960	6675	672	306
27	2255	5298	712	241
28	1423	3185	214	133
29	2488	7364	602	245
30	1426	3812	225	116

2. Исследуется влияние некоторых показателей социально-экономического положения субъектов Центрального федерального округа России на региональный коэффициент смертности. В таблице приводятся официальные статистические данные по субъектам Центрального федерального округа, где:

- *Y* — коэффициент смертности в 2006 году (выражается в промилле «%» и представляет собой число умерших за год на 1000 человек населения);

• X_1 — индекс (темп роста) промышленного производства, в % к 2004 году;

• X_2 — индекс производства продукции сельского хозяйства, в % к 2004 году (для г. Москвы условно принято 100 %);

• X_3 — численность работников малых предприятий, ‰ (чел. на 1000 чел. населения);

• X_4 — среднемесячная номинальная начисленная заработная плата по региону, тыс. руб.;

• X_5 — численность населения на 1 января 2005 года, тыс. чел.

Требуется:

1. Построить матрицу парных коэффициентов линейной корреляции и выявить коллинеарные факторы.

2. Построить линейную регрессионную модель коэффициента смертности, обосновав отбор факторов. Если из-за коллинеарности факторов невозможно построить уравнение регрессии с полным перечнем факторов, то построить несколько моделей.

3. Оценить качество построенных моделей.

4. Дать экономическую интерпретацию параметров лучшего уравнения регрессии и оценить вклад каждого из факторов в вариацию коэффициента смертности с помощью дельта - коэффициентов.

5. Построить три однофакторные нелинейные регрессионные модели зависимой переменной с наиболее подходящим фактором: степенную, гиперболическую и показательную. Сравнить качество моделей. Выбрать лучшую модель.

Примечание. При проверке статистических гипотез уровень значимости α принять равным 0,05.

Таблица 17

Область	Y	X_1	X_2	X_3	X_4	X_5
Белгородская	16,0	108,8	115,8	35,4	6,86	1512
Брянская	19,8	116,0	95,7	25,0	5,24	1346

Владимирская	20,3	100,2	113,3	43,1	6,07	1487
Воронежская	18,8	109,6	102,1	53,3	5,60	2334
Ивановская	22,0	107,6	96,8	36,5	5,37	1115
Калужская	19,2	105,0	94,7	58,4	6,98	1022
Костромская	21,0	108,4	100,3	30,1	5,84	717
Курская	19,7	104,0	101,1	29,8	5,65	1199
Липецкая	17,9	102,5	108,2	33,6	7,19	1190
Московская	17,5	129,6	101,2	61,5	9,51	6630
Орловская	18,5	110,3	101,7	28,4	5,46	842
Рязанская	20,3	106,2	100,9	49,4	6,22	1195
Смоленская	21,5	104,3	92,3	26,3	6,30	1019
Тамбовская	19,3	102,5	110,0	25,6	5,08	1145
Тверская	23,1	104,4	93,0	34,5	6,64	1425
Тульская	22,0	105,0	102,7	36,4	6,34	1622
Ярославская	19,9	104,5	105,9	43,3	7,39	1339
г. Москва	12,4	122,4	100,0	168,9	13,74	10407

4.3 Вопросы для самопроверки

1. Сформулируйте требования, предъявляемые к факторам для включения их в модель множественной регрессии.

2. К каким трудностям приводит мультиколлинеарность факторов и как они могут быть преодолены?

3. Что означает взаимодействие факторов и как оно может быть выражено графически?

4. При каких условиях строится уравнение множественной регрессии с фиктивными переменными?

5. В чем смысл коэффициента детерминации как он соотносится с коэффициентом множественной корреляции?

6. Отличие множественной регрессии от парной регрессии?

7. Какой метод применяют при оценке параметров множественной регрессии?

5. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

5.1 Основные понятия

При нелинейной зависимости признаков, приводимой к линейному виду, параметры множественной регрессии также определяются по МНК с

той лишь разницей, что он используется не к исходной информации, а к преобразованным данным. Так, рассматривая степенную функцию

$$\tilde{y} = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot \text{K} \cdot x_p^{b_p},$$

мы преобразовываем ее в линейный вид:

$$\lg \tilde{y} = \lg a + b_1 \cdot \lg x_1 + b_2 \cdot \lg x_2 + \text{K} + b_p \cdot \lg x_p,$$

где переменные выражены в логарифмах.

Далее обработка МНК та же: строится система нормальных уравнений и определяются неизвестные параметры. Потенцируя значение $\lg a$, находим параметр a и соответственно общий вид уравнения степенной функции.

Вообще говоря, нелинейная регрессия по включенным переменным не таит каких-либо сложностей в оценке ее параметров. Эта оценка определяется, как и в линейной регрессии, МНК. Так, в двухфакторном уравнении нелинейной регрессии

$$\tilde{y} = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_1^2 + a_4 \cdot x_2^2$$

может быть проведена линеаризация, введением в него новых переменных $x_3 = x_1^2$, $x_4 = x_2^2$. В результате получается четырехфакторное уравнение линейной регрессии

$$\tilde{y} = a_0 + a_1 \cdot x_1 + a_2 \cdot x_2 + a_3 \cdot x_3 + a_4 \cdot x_4.$$

Пример. Исследуется зависимость потребительских расходов на душу населения y (тыс.руб.) и средней заработной платой (и выплатами социального характера) x (тыс.руб.) по данным за 2005-2006 г.

Таблица 18

i	Средняя заработная плата x (тыс.руб.)	Расходы на душу населения y (тыс.руб.)
1	10,4	9,0
2	10,6	8,5
3	12,3	10,7
4	12,5	10,1
5	12,5	12,0
6	12,5	11,5
7	12,8	11,8

8	12,8	12,0
9	13,0	11,1
10	13,2	10,0
11	13,5	12,7
12	13,6	13,1
13	14,4	13,9
14	14,6	13,5
15	14,9	13,5
16	14,9	14,6
17	15,4	14,5
18	15,6	15,1
18	15,9	15,6
20	15,9	15,0
21	16,2	15,5
22	16,5	15,9
23	16,7	15,5
24	17,0	16,0

Решение.

Уравнение показательной кривой: $y = b_0 \cdot b_1^{\,x}$.

Приведем к линейному виду: $\ln y = \ln(b_0 \cdot b_1^{\,x})$

$$\ln y = \ln b_0 + \ln b_1^{\,x}$$

$$\ln y = \ln b_0 + x \cdot \ln b_1$$

Пусть $Y = \ln y$, $A = \ln b_0$, $B = \ln b_1$, тогда уравнение имеет вид: $Y = A + B \cdot x$.

Составим расчетную таблицу 19.

Получим:

$$\overline{xY} = \frac{867{,}41765}{24} \approx 36{,}142 \quad \overline{Y} = \frac{61{,}1071163}{24} \approx 2{,}546 \quad \overline{x} = \frac{337{,}7}{24} \approx 14{,}071 \quad \overline{x^2} = \frac{4831{,}95}{24} \approx 201{,}331$$

Следовательно:

$$B = \frac{36{,}142 - 14{,}071 \cdot 2{,}546}{201{,}331 - (14{,}071)^2} = \frac{0{,}317234}{3{,}337959} \approx 0{,}095$$

$$A = 2{,}546 - 0{,}095 \cdot 14{,}071 \approx 1{,}209$$

Тогда $b_0 = e^{1{,}209} \approx 3{,}35$, $b_1 = e^{0{,}095} \approx 1{,}1$

Уравнение примет вид: $\tilde{y} = 3{,}35 \cdot 1{,}1^{x}$.

Вычислим теоретические значения $\tilde{y}$. Рассчитаем тесноту связи – индекс корреляции:

$$\rho_{xy} = \sqrt{1 - \frac{\sum (y - \tilde{y})^2}{\sum (y - \bar{y})^2}} = \sqrt{1 - \frac{10{,}92427539}{119{,}516256}} = \sqrt{0{,}908595903} \approx 0{,}953,$$

связь между факторами очень сильная.

Индекс детерминации $\rho_{xy}{}^2 \approx 0{,}9082$, т.е. 90,82 % вариации изменения y объясняется вариацией изменения x.Средний коэффициент эластичности

$$\overline{Э} = f' \cdot \frac{\bar{x}}{\bar{y}} = \bar{x} \cdot \ln b_1 = 14{,}071 \cdot \ln 1{,}1 \approx 1{,}341$$

При увеличении средней заработной платы на 1% расходы на душу населения увеличатся на 1,341 %.

Оценим статистическую значимость параметров регрессии.

Выдвигаем гипотезу H_0 о статистически незначимом отличии показателей от нуля: $b_0 = b_1 = r_{xy} = 0$.

Таблица 19

n	y	x	Y	xY	x^2	$\tilde{y}$	$y-\tilde{y}$	$(y-\tilde{y})^2$	$y-y_{ср}$	$(y-y_{ср})^2$	A_i
1	9	10,4	2,197224577	22,8511356	108,16	9,026694256	-0,026694256	0,000712583	-3,963	15,705369	0,296602845
2	8,5	10,6	2,140066163	22,68470133	112,36	9,200411872	-0,700411872	0,490576791	-4,463	19,918369	8,240139672
3	10,7	12,3	2,370243741	29,15399802	151,29	10,81869393	-0,118693932	0,014088249	-2,263	5,121169	1,109289083
4	10,1	12,5	2,312535424	28,9066928	156,25	11,02689836	-0,92689836	0,85914057	-2,863	8,196769	9,177211489
5	12	12,5	2,48490665	31,06133312	156,25	11,02689836	0,97310164	0,946926801	-0,963	0,927369	8,10918033
6	11,5	12,5	2,442347035	30,52933794	156,25	11,02689836	0,47310164	0,223825161	-1,463	2,140369	4,113927301
7	11,8	12,8	2,468099531	31,591674	163,84	11,34674192	0,453258079	0,205442886	-1,163	1,352569	3,841170158
8	12	12,8	2,48490665	31,80680512	163,84	11,34674192	0,653258079	0,426746117	-0,963	0,927369	5,443817322
9	11,1	13	2,406945108	31,29028641	169	11,56510857	-0,465108568	0,21632598	-1,863	3,470769	4,190167281
10	10	13,2	2,302585093	30,39412323	174,24	11,78767765	-1,787677654	3,195791395	-2,963	8,779369	17,87677654
11	12,7	13,5	2,541601993	34,31162691	182,25	12,1295882	0,570411804	0,325369626	-0,263	0,069169	4,491431524
12	13,1	13,6	2,57261223	34,98752633	184,96	12,2457482	0,854251798	0,729746135	0,137	0,018769	6,521006093
13	13,9	14,4	2,63188884	37,8991993	207,36	13,21598306	0,68401694	0,467879174	0,937	0,877969	4,920985178
14	13,5	14,6	2,602689685	37,99926941	213,16	13,47032302	0,029676978	0,000880723	0,537	0,288369	0,219829467
15	13,5	14,9	2,602689685	38,78007631	222,01	13,86103997	-0,361039971	0,130349861	0,537	0,288369	2,674370155
16	14,6	14,9	2,681021529	39,94722078	222,01	13,86103997	0,738960029	0,546061925	1,637	2,679769	5,061370062
17	14,5	15,4	2,674148649	41,1818892	237,16	14,53758137	-0,037581366	0,001412359	1,537	2,362369	0,259181837
18	15,1	15,6	2,714694744	42,349238	243,36	14,81735532	0,282644676	0,079888013	2,137	4,566769	1,871819045
19	15,6	15,9	2,747270914	43,68160754	252,81	15,24714397	0,352856032	0,124507379	2,637	6,953769	2,261897641
20	15	15,9	2,708050201	43,0579982	252,81	15,24714397	-0,247143968	0,061080141	2,037	4,149369	1,647626453
21	15,5	16,2	2,740840024	44,40160839	262,44	15,68939896	-0,189398958	0,035871965	2,537	6,436369	1,22192876
22	15,9	16,5	2,766319109	45,6442653	272,25	16,14448189	-0,244481889	0,059771394	2,937	8,625969	1,537621946
23	15,5	16,7	2,740840024	45,7720284	278,89	16,45518045	-0,955180449	0,912369691	2,537	6,436369	6,162454512
24	16	17	2,772588722	47,13400828	289	16,93247545	-0,932475455	0,869510474	3,037	9,223369	5,827971592
сумма	**311,1**	**337,7**	**61,10711633**	**867,41765**	**4831,95**	**312,027249**	**-0,927249006**	**10,92427539**	**-0,012**	**119,516256**	**107,0777763**

Коэффициент Стьюдента для числа степеней свободы $df = n-2 = 24-2 = 22$ и $\alpha = 0{,}05$ составит $2{,}07$.

Определим случайные ошибки:

$$m_{b_0} = \sqrt{\frac{\sum (y-\tilde{y})^2}{n-2} \cdot \frac{\sum x^2}{n\sum (x-\bar{x})^2}} = \sqrt{\frac{10{,}92427539}{24-2} \cdot \frac{483{,}95}{24 \cdot 80{,}229584}} \approx 1{,}116$$

$$m_{b_1} = \sqrt{\frac{\sum (y-\tilde{y})^2 /(n-2)}{\sum (x-\bar{x})^2}} = \sqrt{\frac{10{,}92427539/22}{80{,}229584}} \approx 0{,}079.$$

$$m_{r_{xy}} = \sqrt{\frac{1-r_{xy}^2}{n-2}} = \sqrt{\frac{1-0{,}9082}{24-2}} \approx 0{,}065$$

тогда $t_{b_0} = \frac{b_0}{m_{b_0}} = \frac{3{,}35}{1{,}116} \approx 3{,}002$

$$t_{b_1} = \frac{b_1}{m_{b_1}} = \frac{1{,}1}{0{,}079} \approx 13{,}924$$

$$t_{r_{xy}} = \frac{r_{xy}}{m_{r_{xy}}} = \frac{0{,}953}{0{,}065} = 14{,}662$$

Получим:

$$|t_{b_0}| = 3{,}002 > t_{табл} = 2{,}07$$

$$|t_{b_1}| = 13{,}924 > t_{табл} = 2{,}07$$

$$|t_r| = 14{,}662 > t_{табл} = 2{,}07$$

Т.к. фактические значения t-статистики больше табличного значения, то гипотеза отклоняется, т.е. b_0, b_1, r_{xy} не случайно отличны от нуля и являются статистически значимыми.

Построим доверительные интервалы для коэффициентов регрессии.

$\gamma_{b_0} = b_0 \pm \Delta_{b_0} = b_0 \pm t_{табл} \cdot m_{b_0} = 3{,}35 \pm 2{,}07 \cdot 1{,}116 = 3{,}35 \pm 2{,}31$, т.е $b_0 \in$ [1,04;5,66]

$\gamma_{b_1} = b_1 \pm \Delta_{b_1} = b_1 \pm t_{табл} \cdot m_{b_1} = 1{,}1 \pm 2{,}07 \cdot 0{,}079 = 1{,}1 \pm 0{,}164$, т.е $b_1 \in$ [0,936;1,264]

Оценим качество модели с помощью F-критерия Фишера.

$$F_{факт} = \frac{\rho_{xy}^2}{1-\rho_{xy}^2}(n-2)$$

$$F_{факт} = \frac{0{,}9082}{1-0{,}9082}(24-2) = 217{,}651$$

При уровне значимости $\alpha = 0{,}05$ (вероятность 95 %) табличное значение равно 4,3. $F_{факт} > F_{табл}$, следовательно, гипотеза о случайной природе оцениваемых характеристик не принимается, признается их значимость, следовательно модель надежна.

Средняя ошибка аппроксимации составит:

$$\overline{A} = \frac{1}{n} \cdot A_i = \frac{1}{n} \sum \left| \frac{y - \tilde{y}}{y} \right| \cdot 100\%$$

$$\overline{A} = \frac{1}{24} \cdot 107{,}0777763 \approx 4{,}462\%$$

Качество модели хорошее, т.к. не превышает значение 8 – 10 %.

5.2 Задания для самостоятельной работы по теме: «Нелинейная регрессия»

1. По территории Северного, Северо-Западного и Центрального районов известны данные:

Таблица 20

Район	Потребительские расходы на душу населения, тыс. руб., *y*	Денежные доходы на душу населения, тыс. руб., *x*
Республика Карелия	596-N	913-N
Республика Коми	417+N	1095-N
Архангельская область	354+N	606+N
Вологодская область	526+N	876+N
Мурманская область	934-N	1314-N
Ленинградская область	412+N	593+N
Новгородская область	525-N	754-N
Псковская область	367+N	528+N
Брянская область	364+N	520+N
Владимирская область	336+N	539+N
Ивановская область	409-N	540+N
Калужская область	452-N	682+N
Костромская область	367+N	537+N
Московская область	328+N	589+N
Орловская область	460-N	626-N
Рязанская область	380+N	521+N
Смоленская область	439-N	626-N
Тверская область	344+N	521+N
Тульская область	401-N	658-N

Ярославская область	514-N	746-N

Требуется:

1. Рассчитайте параметры уравнений степенной, показательной, гиперболической парной регрессии.

2. Оцените тесноту связи каждого уравнения с помощью показателей корреляции и детерминации.

3. С помощью среднего коэффициента эластичности дайте сравнительную оценку силы связи фактора с результатом (для каждого уравнения).

4. Оцените с помощью средней ошибки аппроксимации качество уравнений.

5. С помощью F-критерия Фишера оцените статистическую надежность результатов регрессионного моделирования.

6. По значениям характеристик, рассчитанных в пп. 2–5 и данном пункте, выберите лучшее уравнение регрессии и дайте его обоснование.

7. Рассчитайте прогнозное значение результата по линейному уравнению регрессии, если прогнозируется увеличение значения фактора на 10% от среднего уровня. Определите доверительный интервал прогноза для уровня значимости $\alpha = 0,05$.

2. По семи территориям Уральского района известны значения двух признаков (табл. 21.).

Таблица 21.

Район	Расходы на покупку продовольственных товаров в общих расходах, %, y	Среднедневная заработная плата одного работающего, руб., x
Удмуртская респ.	68,8 + N/2	45,1 – K/2
Свердловская обл.	61,2 + M/2	59,0 – N/2
Башкортостан	59,9 + K/2	57,2 – M/2
Челябинская обл.	56,7 + N/2	61,8 – K/2
Пермская обл.	55,0 + K/2	58,8 – N/2
Курганская обл.	54,3 + M/2	47,2 – K/2
Оренбургская обл.	49,3 + K/2	55,2 – M/2

Требуется:

1. Для характеристики зависимости y от x рассчитать параметры следующих функций: степенной, показательной, равносторонней гиперболы.

2. Оценить каждую модель через среднюю ошибку аппроксимации $\overline{A}$ и F-критерий Фишера.

5.3 Вопросы для самопроверки

1. Опишите виды моделей нелинейной регрессии.
2. Приведите примеры использования логарифмических регрессионных моделей.
3. Каков смысл коэффициентов регрессии в логарифмических регрессионных моделях?
4. Приведите примеры использования обратных и степенных моделей.
5. Изменятся ли свойства случайного отклонения при преобразовании уравнения регрессии?

6. ГЕТЕРОСКЕДАСТИЧНОСТЬ И АВТОКОРРЕЛЯЦИЯ

6.1 Основные понятия

В результате построения с помощью МНК уравнения регрессии получается не точное значение, а отличающееся от точного на некоторую величину ε:

$$y = \tilde{y}_x + \varepsilon = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \text{K} + b_p \cdot x_p + \varepsilon .$$

После того как проведена оценка параметров модели, рассчитывая разности фактических и теоретических значений $y - \tilde{y}_x$ можно получить оценки случайной составляющей ε. В задачу регрессионного анализа входит не только построение самой модели, но и исследование остаточных величин.

Необходимость этого объясняется тем, что при использовании МНК предполагалось, что остатки представляют собой независимые случайные

величины и их среднее значение равно 0; они имеют одинаковую (постоянную) дисперсию.

Таким образом, исследование остатков предполагают проверку наличия следующих предпосылок МНК

Случайных характер остатков

Для проверки строится график зависимости остатков ε_i от теоретических значений результативного признака. Если на графике получена горизонтальная полоса, то остатки ε_i представляют собой случайные величины и МНК оправдан, а теоретические значения $\tilde{y}_x$ хорошо аппроксимируют фактические значения *y*. Пример случайности остатков приведен на рисунке:

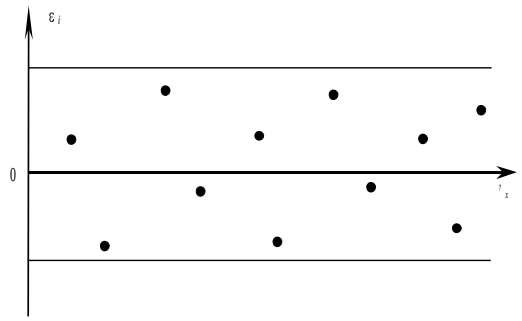
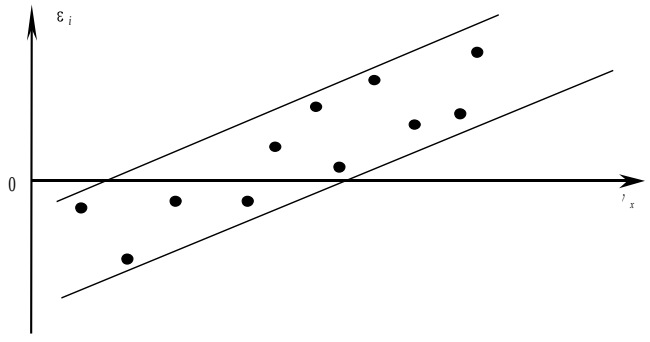


Рисунок 6

Возможны различные случаи зависимости остатков от теоретических значений $\tilde{y}_x$. Приведем примеры



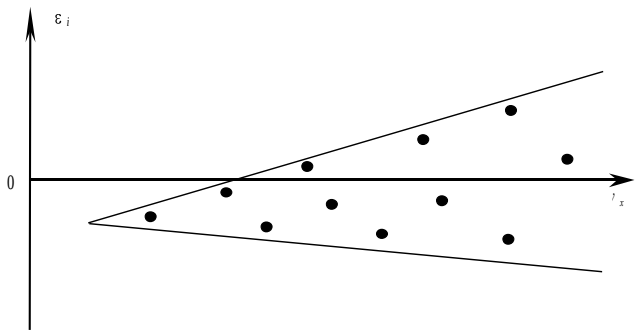


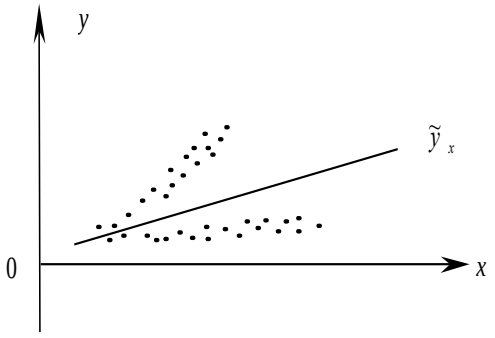
Рисунок 7

Нулевая средняя величина остатков, не зависящая от x_j

Эта предпосылка означает, что $\sum_{k=1}^{n}(y-\tilde{y}_x)=0$. Это условие выполнимо для линейных моделей. Для определения независимость величины остатков от x_j, как и в случае определения независимости ε_i от $\tilde{y}_x$, строится график ε_i от x_j. Если остатки на графике расположены в виде горизонтальной полосы, то они независимы от значений x_j. Если же зависимость присутствует, то модель является неадекватной.

Гомоскедастичность

Гомоскедастичность остатков означает, что дисперсия каждого отклонения одинакова для всех значений *x*. Если это условие не соблюдается, то имеет место **гетероскедастичность**. Наличие гетероскедастичности можно наглядно видеть из поля корреляции (смотри рисунок).



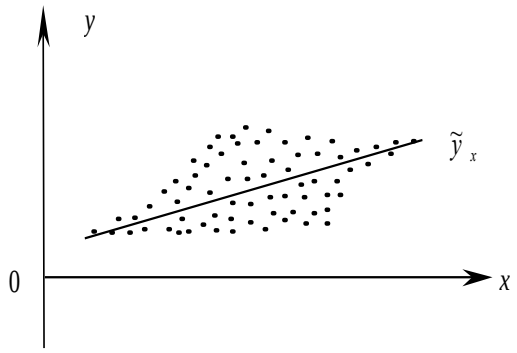


Рисунок 8

Т.к. дисперсия характеризует отклонение то из рисунков видно, что в первом случае дисперсия остатков растет по мере увеличения *x*, а во втором – дисперсия остатков достигает максимальной величины при средних значениях величины *x* и уменьшается при минимальных и максимальных значениях *x*. Наличие гетероскедастичности будет сказываться на уменьшении эффективности оценок параметров уравнения регрессии. Наличие гомоскедастичности или гетероскедастичности можно определять также по графику зависимости остатков от теоретических значений $\tilde{y}_x$.

Отсутствие автокорреляции остатков

Под **автокорреляцией** остатков понимают зависимость распределения значений остатков ε_i друг от друга. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений. Оценить эту зависимость можно вычислив коэффициент корреляции между этими остатками по формуле

$$r_{\varepsilon_i \varepsilon_j} = \frac{\overline{\varepsilon_i \varepsilon_j} - \bar{\varepsilon}_i \cdot \bar{\varepsilon}_j}{\sigma_{\varepsilon_i} \sigma_{\varepsilon_j}}.$$

Если этот коэффициент окажется существенно отличным от нуля, то остатки автокоррелированны.

Пример. Проверить для уравнения регрессии, полученного ранее, выполнение предпосылок МНК.

Вычисляем теоретические значения по уравнению регрессии полученному ранее, а остатки по формуле $\varepsilon = y - \tilde{y}_x$ и записываем в таблицу

Номер предприятия	1	2	3	4
x_1, (%)	1	2	3	5
x_2, (%)	0	1	3	4
y, (тыс. руб.)	6	11	19	28
$\tilde{y}_x$, (тыс. руб.)	5,79	11,31	19,07	27,87
ε, (тыс. руб.)	0,21	-0,31	-0,07	0,13

Теперь для проверки случайного характера остатков построим график их зависимости от теоретических значений $\tilde{y}_x$.

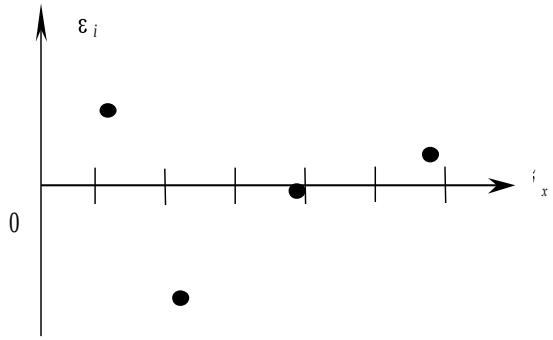


Рисунок 9

Хотя по четырем точкам судить трудно, но в целом можно сделать вывод, что остатки распределены случайно. Из этого же рисунка можно сделать вывод о гомоскедастичности остатков, т. к. дисперсия каждого отклонения одинакова для всех значений *x*.

Вычислим теперь величину суммарного отклонения:

$$\sum_{k=1}^{n}(y-\tilde{y}_x)=0{,}21-0{,}31-0{,}07+0{,}13=-0{,}04.$$

По малости этой величины можно сделать вывод о практически нулевой средней величине остатков.

Коэффициент автокорреляции остатков находим по следующим рядам данных:

ε_i, (тыс. руб.)	-0,31	-0,07	0,13
ε_{i-1}, (тыс. руб.)	0,21	-0,31	-0,07

$$\bar{\varepsilon}_i=\frac{-0{,}31-0{,}07+0{,}13}{3}=-0{,}083;$$

$$\bar{\varepsilon}_{i-1} = \frac{0{,}21 - 0{,}31 - 0{,}07}{3} = -0{,}057;$$

$$\overline{\varepsilon_i \varepsilon_{i-1}} = \frac{(-0{,}31)\cdot 0{,}21 + (-0{,}07)\cdot(-0{,}31) + 0{,}13\cdot(-0{,}07)}{3} = -0{,}018;$$

$$\sigma_{\varepsilon_i} = \sqrt{\frac{\sum_{k=1}^{3} (\varepsilon_{ik} - \bar{\varepsilon}_i)^2}{3}} = \sqrt{\frac{(-0{,}31 + 0{,}083)^2 + (-0{,}07 + 0{,}083)^2 + (0{,}13 + 0{,}083)^2}{3}} = 0{,}18$$

$$\sigma_{\varepsilon_{i-1}} = \sqrt{\frac{\sum_{k=1}^{3} (\varepsilon_{i-1k} - \bar{\varepsilon}_{i-1})^2}{3}} = \sqrt{\frac{(0{,}21 + 0{,}057)^2 + (-0{,}31 + 0{,}057)^2 + (-0{,}07 + 0{,}057)^2}{3}} = 0{,}212$$

Отсюда находим $r_{\varepsilon_i \varepsilon_{i-1}} = \dfrac{\overline{\varepsilon_i \varepsilon_{i-1}} - \bar{\varepsilon}_i \cdot \bar{\varepsilon}_{i-1}}{\sigma_{\varepsilon_i} \sigma_{\varepsilon_{i-1}}} = \dfrac{-0{,}018 - (-0{,}083)\cdot(-0{,}057)}{0{,}18 \cdot 0{,}212} = -0{,}348$

Коэффициент корреляции не так велик, и его можно считать приемлемым. Таким образом, мы установили, что у нас были все предпосылки к тому, чтобы применять МНК и линейное уравнение регрессии к исходным данным.

6.2 Вопросы для самопроверки

1. Что такое гетероскедастичность? Когда она возникает?
2. Из-за чего может возникнуть гетероскедастичность в модели?
3. Перечислите последствия наличия гетероскедастичности в модели.
4. Какие вы знаете еще тесты для обнаружения гетероскедастичности?
5. Как обнаружить гетероскедастичность графически?
6. Как можно скорректировать модель при наличии гетероскедастичности?

7. ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

7.1 Основные понятия

До сих пор в качестве факторов рассматривались экономические переменные, принимающие количественные значения в некотором интервале. Вместе с тем может оказаться необходимым включить в модель

факторы, которые представляют собой различные атрибутивные признаки. Такими признаками, например, являются профессия, пол, образование, климатические условия и т.п. Чтобы ввести такие переменные в регрессионную модель, им должны быть присвоены те или иные цифровые метки, т.е. качественные переменные преобразовать в количественные. Такого вида сконструированные переменные в эконометрике принято называть *фиктивными переменными*.

Рассмотрим применение фиктивных переменных для функции спроса. Предположим, что по группе лиц мужского и женского пола изучается линейная зависимость потребления кофе от цены. В общем виде для совокупности обследуемых уравнение регрессии имеет вид: $y = a + b \cdot x + \varepsilon$, где y – количество потребляемого кофе; x – цена кофе.

Аналогичные уравнения могут быть найдены отдельно для лиц мужского пола: $y_1 = a_1 + b_1 \cdot x_1 + \varepsilon_1$ и женского пола: $y_2 = a_2 + b_2 \cdot x_2 + \varepsilon_2$. Если сила влияния цены на количество потребления кофе одинакова как для мужчин, так и для женщин ($b_1 \approx b_2 \approx b$), то становится возможным построение общего уравнения регрессии с включением в него фактора «пол» в виде фиктивной переменной. Это уравнение может быть записано в виде: $y = a_1 \cdot z_1 + a_2 \cdot z_2 + b \cdot x + \varepsilon$, где z_1, z_2 - фиктивные переменные, принимающие значения:

$$z_1 = \begin{cases} 1 - \textit{мужской пол} \\ 0 - \textit{женский пол} \end{cases}; \quad z_2 = \begin{cases} 0 - \textit{мужской пол} \\ 1 - \textit{женский пол} \end{cases}.$$

Следует отметить, что применение МНК для оценивания параметров a_1 и a_2 приводит к вырожденной матрице исходных данных, а следовательно, и к невозможности получения их оценок.

Выходом из создавшегося положения может явиться переход к уравнению $y = A + A_1 \cdot z_1 + b \cdot x + \varepsilon$, т.е. уравнению, включающему только одну фиктивную переменную. Предположим, что МНК были получены оценки параметров этого уравнения, тогда теоретические значения размера потребления кофе для мужчин будут получены из уравнения $\tilde{y} = A + A_1 + b \cdot x$.

Для женщин соответствующие значения получим из уравнения $\tilde{y} = A + b \cdot x$.

Пример 1. Согласно условию задания 1, дана зависимость объема выпускаемой продукции (шт.) на участке механической обработки от времени (часы), которое включает в себя ремонт, техническое обслуживание оборудования. Исходя из этого, примем объем продукции за результативную переменную, а время простоев оборудования за объясняющую переменную. Выдвинем гипотезу о том, что соответствие сроков выполнения технического обслуживания повлияет на объем выпуска продукции. Тогда введем фиктивную переменную d, которая будет характеризовать качественный признак – соответствие сроков выполнения технического обслуживания. При выполнении сроков обслуживания d=1; при невыполнении d=0.

Таблица 22

y (шт)	x (час) простои	d
3520	10	1
3460	19	1
3100	16	1
3320	26	1
3540	4	1
3310	14	1
3360	21	1
3350	10	1
3150	22	1
3440	8	1
3110	29	1
3290	15	1
3190	3	0
3060	12	0
3270	17	0
3370	14	0
3230	18	0
3300	11	0
3200	14	0
3460	9	0

Решение.

Абстрагируясь от качественного признака во время построения уравнения регрессии, получим следующие данные:

Таблица 23

Результаты моделирования зависимости выпускаемого объема физической продукции от времени простоев по техническим причинам без учета своевременности выполнения технического обслуживания оборудования.

Регрессионная статистика		Дисперсионный анализ		Коэффициенты		P-Значение
Множественный R	0,3817	Значимость F	0,0968	Y-пересечение	3417,6797	1,357E-20
				x (час)	-7,958	0,04841
R-квадрат	0,1457	F	3,0691			

Задействуем фиктивную переменную и сравним полученные результаты:

Таблица 24

Результаты моделирования зависимости выпускаемого объема физической продукции от времени простоев по техническим причинам с учетом своевременности выполнения технического обслуживания оборудования.

Регрессионная статистика		Дисперсионный анализ		Коэффициенты		P-Значение
Множественный R	0,5363	Значимость F	0,055995	Y-пересечение	3386,984	6,3635E-20
				x (час)	-10,366	0,0165
Нормированный R-квадрат	0,2038	F	3,4315	d	109,767	0,0416

Сравнивая результаты моделирования до и после использования фиктивной переменной, можно отметить, что во втором случае наблюдается более высокая точность описания регрессионной моделью эмпирических данных, о чем свидетельствует сопоставление коэффициентов детерминации: без использования фиктивной переменной R^2=14,57%, после использования фиктивной переменной скорректированный R^2=20,38%. Что касается адекватности регрессионной модели, то в первом случае при уровне значимости $\alpha_{кр} = 10\%$ модель является адекватной, т.к. $\alpha_{кр} > \alpha_{факт}$ (0,1>0,96). Во втором случае модель является адекватной уже при $\alpha_{кр} = 6\%$, поскольку 0,060>0,055. Если рассматривать значимость коэффициентов, то и в первом, и во втором случае они окажутся отличными от нуля при $\alpha_{кр} = 5\%$, т.к. 0,05>6.36E-20; 0,05>0,0165; 0,05>0,042. Тем самым, подтверждается

значимость и фиктивной переменной, которая отражает качественный признак. Уравнение множественной регрессии с использованием фиктивной переменной будет иметь вид: $Y = 3386{,}98 - 10{,}37 \times X + 109{,}77 \times d$

На основе проведенного анализа можно сделать вывод о том, что своевременное техническое обслуживание оборудования влияет на ход производственного процесса, а именно на объем выпускаемой продукции: при своевременном техническом обслуживании наблюдается рост объема производства на 109,77 шт. Также важно отметить, что фиктивная переменная является переменной сдвига, поскольку при d=1 происходит увеличение объема производства на 109,77 шт., что наглядно продемонстрировано на графике подбора.

Пример 2. Задана зависимость доходности портфеля ценных бумаг от доходности рынка в целом, которая отражается формулой: $R_p = A + B \times (R_m - R_f)$, где R_p - доходность портфеля ценных бумаг; R_m - доходность рынка; R_f - доходность базисного актива.

Решение.

Нам необходимо определить, насколько качественно управляющий портфелем осуществляет свою деятельность. Высокий уровень управления портфелем характеризуется превышением доходности портфеля над доходностью рынка и изменением стратегии при переходе от положительной доходности рынка к отрицательной. Для этого введем фиктивную переменную d, которая будет отражать качественный признак – качество управления портфелем ценных бумаг. При положительной доходности на рынке d=0; при отрицательной доходности на рынке d=1. Тогда уравнение зависимости портфеля ценных бумаг от доходности рынка в целом будет иметь вид: $R_p = A + B \times (R_m - R_f) + C \times (R_m - R_f) \times d$, где d – фиктивная переменная.

Таблица 23

Rp,%	Rm,%	Rf,%	d	Rm-Rf	d*(Rm-Rf)
4	7	4	0	3	0
16	12	4	0	8	0
28	23	4	0	19	0

44	32	4	0	28	0
27	21	4	0	17	0
12	9	4	0	5	0
-3	2	4	0	-2	0
10	-9	4	1	-13	-13
-3	-23	4	1	-27	-27
2	-12	4	1	-16	-16
1	-5	4	1	-9	-9
5	0	4	0	-4	0
10	12	4	0	8	0

Построим регрессионные модели для первого и второго случая и проведем их сравнительный анализ.

Таблица 24

Результаты моделирования зависимости доходности портфеля ценных бумаг от рынка в целом без учета качества управления портфелем.

Регрессионная статистика		Дисперсионный анализ		Коэффициенты		P-Значение
Множественный R	0,8598	Значимость F	0,000164	Y-пересечение	10,753	0,0001
				Rm-Rf	0,7773	8E-05
R-квадрат	0,7392	*F*	31,18204			

Таблица 25

Результаты моделирования зависимости доходности портфеля ценных бумаг от рынка в целом с учетом качества управления портфелем.

Регрессионная статистика		Дисперсионный анализ		Коэффициенты		P-Значение
Множественный R	0,95944	Значимость F	3,17E-06	Y-пересечение	4,05402	0,0262
				Rm-Rf	1,33451	1E-06
Нормированный R-квадрат	0,90462	F	57,9094	d*(Rm-Rf)	-1,194	0,0004

При использовании фиктивной переменной наблюдается более высокая точность описания регрессионной моделью эмпирических данных, чем без ее использования. Так скорректированный R^2 при значении 90,46% говорит о том, что регрессионная модель, отражающая качество управления портфелем ценных бумаг, описывает эмпирические данные более точно, чем это делает модель, в которой не придается значения качеству управления портфелем (R^2=73,92%). Обе модели оказались адекватными при уровне значимости $\alpha_{кр} = 3\%$, поскольку в первом случае 0,03>0,000164; во втором случае

0,03>3,17E-06. Все коэффициенты в обоих случаях при уровне значимости $\alpha_{кр} = 3\%$ оказались значимыми (отличными от нуля), так как 0,03>0,0001; 0,03>8E-05 и 0,03>0,026; 0,03>1E-06; 0,03>0,0004, что говорит о том, что качество управления портфелем влияет на конечную доходность портфеля ценных бумаг и деятельность управляющего может быть оценена положительно. Уравнение регрессии с использованием фиктивной переменной будет иметь вид: $Y = 4{,}054 + 1{,}3345 \times (R_m - R_f) - 1{,}194 \times (R_m - R_f) \times d$

7.2 Задания для самостоятельной работы по теме: «Фиктивные переменные»

1. Необходимо оценить влияние качественного признака d на результативную переменную У. Для этого сравним регрессионную модель, не учитывающую качественный признак с регрессионной моделью, учитывающей качественный признак.

Таблица 26

x	D (фикт)	y
1	0	6,191552
2	0	8,235765
3	0	11,48288
4	0	14,40931
5	0	16,97195
6	0	19,79557
7	0	22,6074
8	0	24,76633
9	0	26,98018
10	0	28,76968
11	0	30,52625
12	1	34,92994
13	1	36,18605
14	1	37,3445
15	1	38,5371
16	1	39,4185
17	1	40,52482
18	1	41,17279
19	1	41,96986
20	1	42,73582

2. Необходимо исследовать зависимость между результатами ЕГЭ и курсового экзамена по математике. Получены данные о числе решенных задач (задание – 18 задач) на ЕГЭ и на курсовом экзамене 16 студентами, а также распределение этих студентов по категории «пол»: y_i – число решенных задач на курсовом экзамене; x_i – число решенных задач на ЕГЭ:

Таблица 27

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x_i	18	15	9	18	12	12	18	6	18	12	16	12	18	6	17	15
y_i	16	15	3	18	6	6	16	4	18	6	16	6	17	2	15	9
пол	М.	М.	Ж.	М.	Ж.	Ж.	М.	Ж.	М.	Ж.	М.	Ж.	М.	Ж.	М.	Ж.

1. Составить модель множественной регрессии изменения зависимой переменной y_i от переменной x_i при учете пола студента с помощью фиктивной переменной, взяв женский пол в качестве эталонной категории.

2. Построить выборочное уравнение регрессии и с доверительной вероятностью 0,95 оценит значимость коэффициентов уравнения регрессии и самой регрессии.

3. Из найденного выборочного уравнения регрессии вывести отдельные уравнения для юношей и девушек и дать их графическую иллюстрацию.

4. С помощью критерия Чоу для уровня значимости $\alpha=0,05$ проверить, являются ли выборки для юношей и девушек однородными в регрессионном смысле и можно ли их объединить.

7.3 Вопросы для самопроверки

1. В чем преимущества фиктивных переменных?
2. Как фиктивные переменные включаются в модель регрессии?
3. В чем состоит правило применения фиктивных переменных?

8. МОДЕЛИРОВАНИЕ ОДНОМЕРНЫХ РЯДОВ

8.1 Основные понятия

Обычно эконометрические модели строятся на основе двух типов исходных данных:

- данные, характеризующие совокупность различных объектов в определенный момент (период) времени;
- данные, характеризующие один объект за ряд последовательных моментов (периодов) времени.

Модели, построенные по данным первого типа, называются **пространственными моделями**. Модели, построенные на основе второго типа данных, называются **моделями временных рядов.**

Временной ряд – совокупность значений какого-либо показателя за несколько последовательных моментов или периодов времени. Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на три группы:

- факторы, формирующие тенденцию ряда (например, инфляция влияет на увеличение размера средней заработной платы);
- факторы, формирующие циклические колебания ряда (например, уровень безработицы в курортных городах в зимний период выше по сравнению с летним);
- случайные факторы.

Очевидно, что реальные данные чаще всего содержат все три компоненты. Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется **аддитивной моделью** временного ряда. Если же временной ряд представлен как их произведение, то такая модель называется **мультипликативной.**

При наличии в временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционную зависимость между последовательными уровнями

временного ряда называют уровнями автокорреляцией уровней ряда. Количественно эту зависимость с помощью коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутого на несколько шагов во времени.

Пример 1. Пусть имеются условные данные о средних расходах на конечное потребление (y_t, денежных единиц) за 8 лет.

Таблица 28

t	y_t	y_{t-1}	$y_t - \bar{y}_1$	$y_{t-1} - \bar{y}_2$	$(y_t - \bar{y}_1)\cdot(y_{t-1} - \bar{y}_2)$	$(y_t - \bar{y}_1)^2$	$(y_{t-1} - \bar{y}_2)^2$
1	7	-	-	-	-	-	-
2	8	7	-3,39	-3	9,87	10,8241	9
3	8	8	-3,29	-2	6,58	10,8241	4
4	10	8	-1,29	-2	2,58	1,6641	4
5	11	10	-0,29	0	0,00	0,0841	0
6	12	11	0,71	1	0,71	0,5041	1
7	14	12	2,71	2	5,42	7,3441	4
8	16	14	4,71	4	18,84	22,1841	16
Σ	86	70	-0,03	0	44,0	53,4287	38

Решение.

По формулам

$$\bar{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1}; \quad \bar{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1}$$

вычисляем

$$\bar{y}_1 = \frac{8+8+10+11+12+14+16}{7} = \frac{79}{7} = 11{,}29,$$

$$\bar{y}_2 = \frac{7+8+8+10+11+12+14}{7} = \frac{70}{7} = 10.$$

Далее, заполняем таблицу и используя формулу для вычисления линейного коэффициента корреляции, получаем

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)\cdot(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2 \cdot (y_{t-1} - \bar{y}_2)^2}} = \frac{44}{\sqrt{53{,}4287 \cdot 38}} = 0{,}976.$$

Полученное значение свидетельствует об очень тесной зависимостью между расходами на конечное потребление текущего непосредственно

предшествующего годов и, следовательно, о наличии во временном ряде расходов на конечное потребление сильной линейной тенденции.

Нами был посчитан коэффициент автокорреляции для смещения на один год. Такой коэффициент называется коэффициентом первого порядка. При смещении на два года получим коэффициент второго порядка и так далее. Число периодов (в данном случае лет), по которым рассчитывается коэффициент автокорреляции, называется *лагом*.

Одним из наиболее распространенных способов моделирования тенденции временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени. Поскольку зависимость может принимать различные формы, то ее формализации можно использовать различные виды функций: линейную, гиперболическую, параболическую, степенную и т.п. Параметры каждой из перечисленных моделей могут быть найдены по МНК.

Пример 2. Известны статистические данные наблюдений за некоторое количество моментов или периодов времени.

Требуется:

1. Построить график динамики уровней ряда.
2. Рассчитать значения сезонных компонент методом скользящей средней.
3. Устранить сезонную компоненту из исходных уровней ряда. Построить уравнение, моделирующее динамику трендовой компоненты.
4. Найти прогноз фактора *y*.

Таблица 29

Номер квартала	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Объем продаж	3974	4134	4054	4200	3995	4162	4091	4224	4038	4191	4125	4259	4079	4240	4181	4320	4126	4285	4213	4359	4143	4311	4240	4390

Решение.

Построим график динамики уровней ряда.

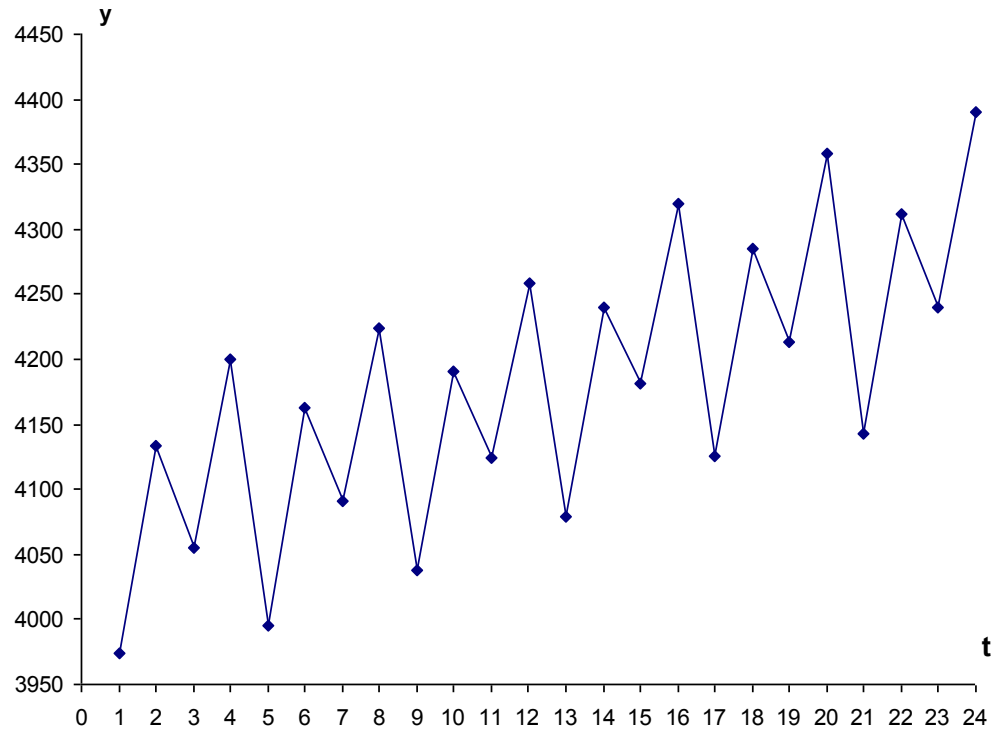


Рисунок 10 - График динамики

Изучая график, заметим наличие сезонных колебаний с периодом 4 квартала и возрастающую тенденцию в уровнях ряда.

Построим мультипликативную модель временного ряда $Y = T \cdot S \cdot E$**.** Рассчитаем компоненты.

1) Проведем выравнивание исходных уровней ряда, которое полностью совпадает выравниванием для построения мультипликативной модели (табл. 30).

Таблица 30

Расчет оценок сезонной компоненты в мультипликативной модели

Номер квартала, t	Объем продаж	Итого за четыре квартала	Скользящая средняя	Центрированная скользящая средняя	Оценка сезонной компоненты
1	3974				
2	4134	16362	4090,5		
3	4054	16383	4095,75	4093,125	0,990441289
4	4200	16411	4102,75	4099,25	1,024577667

5	3995	16448	4112	4107,375	0,972640677
6	4162	16472	4118	4115	1,011421628
7	4091	16515	4128,75	4123,375	0,992148422
8	4224	16544	4136	4132,375	1,02217248
9	4038	16578	4144,5	4140,25	0,975303424
10	4191	16613	4153,25	4148,875	1,010153355
11	4125	16654	4163,5	4158,375	0,991974028
12	4259	16703	4175,75	4169,625	1,021434781
13	4079	16759	4189,75	4182,75	0,975195744
14	4240	16820	4205	4197,375	1,010155156
15	4181	16867	4216,75	4210,875	0,992905275
16	4320	16912	4228	4222,375	1,023120874
17	4126	16944	4236	4232	0,974952741
18	4285	16983	4245,75	4240,875	1,010404692
19	4213	17000	4250	4247,875	0,991790013
20	4359	17026	4256,5	4253,25	1,02486334
21	4143	17053	4263,25	4259,875	0,972563749
22	4311	17084	4271	4267,125	1,010282099
23	4240				
24	4390				
Итого	100334	351131	87782,75	83602	19,99850143

2) Найдем оценки сезонной компоненты как частное от деления фактических уровней ряда на центрированные скользящие средние. Используем эти оценки для расчета значений сезонной компоненты S (табл. 31).

Таблица 31

Расчет сезонной компоненты

год	1 квартал	2 квартал	3 квартал	4 квартал		
1	0	0	0,990441289	1,024577667		
2	0,9726407	1,011421628	0,992148422	1,02217248		
3	0,9753034	1,010153355	0,991974028	1,021434781		
4	0,9751957	1,010155156	0,992905275	1,023120874		
5	0,9749527	1,010404692	0,991790013	1,02486334		
6	0,9725637	1,010282099	0	0		
итого	4,8706563	5,052416931	4,959259027	5,116169141		
Средняя оценка	0,8117761	0,842069488	0,826543171	0,852694857	Сумма	3,333083572
Скорректированная	0,9742043	1,010559106	0,991926129	1,023310503	Корректирующий коэффициент	1,200089921

Для данной модели имеем:

0,8117761+0,84206949+0,826543171+0,852694857=3,333083572

Определим корректирующий коэффициент:

$$k = \frac{4}{3{,}333083572} = 1{,}200089921$$

Рассчитаем скорректированные значения сезонной компоненты:

S_1=0,9742043

S_2=1,01055911

S_3=0,991926129

S_4=1,023310503

3) Разделим каждый уровень исходного ряда на соответствующие значения сезонной компоненты. Получим величины $T \cdot E = Y : S$.

4) Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $T \cdot E$ с помощью линейного тренда.

Получим линейный тренд: $T = 4061{,}189 + 9{,}534 \cdot t$.

Подставляя в это уравнение значения t =1, 2,…., 24 найдем уровни T для каждого момента времени.

5) Найдем уровни ряда по мультипликативной модели, умножив уровни T на значения сезонной компоненты для соответствующих кварталов, т.е. ряд $T \cdot S$.

6) Рассчитаем ошибку: $E = Y/(T \cdot S)$. Абсолютные ошибки определяются как $E' = Y - (T \cdot S)$. Доля объясненной дисперсии уровней ряда равна 99,53 %.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,993727
R-квадрат	0,987493
Нормированный R-квадрат	0,986925
Стандартная ошибка	7,761887
Наблюдения	24

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	1	104651,7	104651,7	1737,048	1,98E-22
Остаток	22	1325,431	60,24688		
Итого	23	105977,1			

	Коэффициен	*дартная о*	*татистик*	*P-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*рхние 95,0%*
Y-пересечение	4061,189	3,270475	1241,774	8,38E-55	4054,407	4067,972	4054,407	4067,972
t	9,53947	0,228886	41,6779	1,98E-22	9,06479	10,01415	9,06479	10,01415

Рисунок 11 - Протокол выполнения регрессионного анализа

Построим аддитивную модель временного ряда $Y = T + S + E$**.** Рассчитаем компоненты.

1) Проведем выравнивание исходных уровней ряда методом скользящей средней, найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и центрированными скользящими средними (табл. 32).

Таблица 32

Расчет оценок сезонной компоненты в аддитивной модели

Номер квартала, t	Объем продаж	Итого за четыре квартала	Скользящая средняя	Центрированная скользящая средняя	Оценка сезонной компоненты
1	3974				
2	4134	16362	4090,5		
3	4054	16383	4095,75	4093,125	-39,125
4	4200	16411	4102,75	4099,25	100,75
5	3995	16448	4112	4107,375	-112,375
6	4162	16472	4118	4115	47
7	4091	16515	4128,75	4123,375	-32,375
8	4224	16544	4136	4132,375	91,625
9	4038	16578	4144,5	4140,25	-102,25
10	4191	16613	4153,25	4148,875	42,125
11	4125	16654	4163,5	4158,375	-33,375
12	4259	16703	4175,75	4169,625	89,375
13	4079	16759	4189,75	4182,75	-103,75
14	4240	16820	4205	4197,375	42,625
15	4181	16867	4216,75	4210,875	-29,875
16	4320	16912	4228	4222,375	97,625
17	4126	16944	4236	4232	-106
18	4285	16983	4245,75	4240,875	44,125
19	4213	17000	4250	4247,875	-34,875
20	4359	17026	4256,5	4253,25	105,75
21	4143	17053	4263,25	4259,875	-116,875
22	4311	17084	4271	4267,125	43,875
23	4240				
24	4390				
Итого	100334	351131	87782,75	83602	-6

2) Используем эти оценки для расчета значений сезонной компоненты S (табл. 33).

Таблица 33

Расчет сезонной компоненты

год	1 квартал	2 квартал	3 квартал	4 квартал		
1	0	0	-39,125	100,75		
2	-112,375	47	-32,375	91,625		
3	-102,25	42,125	-33,375	89,375		
4	-103,75	42,625	-29,875	97,625		
5	-106	44,125	-34,875	105,75		
6	-116,875	43,875	0	0		
итого	-541,25	219,75	-169,625	485,125		
Средняя оценка	-90,208333	36,625	-28,27083333	80,85416667	Сумма	-1
Скорректированная	-89,958333	36,875	-28,02083333	81,10416667	Корректирующий коэффициент	-0,25

Для данной модели имеем:

-90,208333+36,625-28,27083333+80,85416667=-1

Определим корректирующий коэффициент:

$$k = \frac{-1}{4} = -0{,}25$$

Рассчитаем скорректированные значения сезонной компоненты:

S_1=-89,958333

S_2=36,875

S_3=-28,02083333

S_4=81,10416667

3) Элиминируем влияние сезонной компоненты, вычитая ее значение из каждого уровня исходного временного ряда. Получим величины $T + E = Y - S$.

4) Определим компоненту T данной модели. Для этого проведем аналитическое выравнивание ряда $T + E$ с помощью линейного тренда.

Получим линейный тренд: $T = 4058{,}397 + 9{,}775 \cdot t$.

Подставляя в это уравнение значения t =1, 2,…., 24 найдем уровни T для каждого момента времени.

5) Найдем значения уровней ряда, полученные по аддитивной модели. Для этого прибавим к уровням T значения сезонной компоненты для соответствующих кварталов, т.е. ряд $T+S$.

6) Рассчитаем ошибку: $E=Y-(T+S)$. Эта абсолютная ошибка составляет ≈0. Доля объясненной дисперсии уровней ряда равна 98,13 %.

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,976879							
R-квадрат	0,954293							
Нормированный R-квадрат	0,952216							
Стандартная ошибка	15,4667							
Наблюдения	24							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	1	109880,8	109880,8	459,3317	3,13E-16			
Остаток	22	5262,814	239,2188					
Итого	23	115143,6						
	Коэффициен	*дартная о*	*татистик*	*P-Значени*	*ижние 95%*	*ерхние 95%*	*ижние 95,0*	*рхние 95,0%*
Y-пересечение	4058,397	6,516901	622,7495	3,29E-48	4044,882	4071,912	4044,882	4071,912
t	9,774891	0,456088	21,43202	3,13E-16	8,829023	10,72076	8,829023	10,72076

Рисунок 12 - Протокол выполнения регрессионного анализа

ВЫВОД: Лучше всего описывает процесс мультипликативная модель, т.к. доля объясненной дисперсии больше, чем в аддитивной модели.

8.2 Задания для самостоятельной работы по теме: «Моделирование одномерных рядов»

1. Имеются поквартальные данные по розничному товарообороту (в % к предыдущему периоду). Постройте мультипликативную модель временного ряда. Рассчитайте прогноз розничного товарооборота на 1 квартал 2016 года.

Таблица 34

№ квартала	*t*	*yt*
1 кв. 2011 г.	1	113,1
2 кв. 2011 г.	2	95,9

3 кв. 2011 г.	3	98
4 кв. 2011 г.	4	101,8
1 кв. 2012 г.	5	107,8
2 кв. 2012 г.	6	96,3
3 кв. 2012 г.	7	95,7
4 кв. 2012 г.	8	99,8
1 кв. 2013 г.	9	104
2 кв. 2013 г.	10	95,8
3 кв. 2013 г.	11	95,5
4 кв. 2013 г.	12	99,3
1 кв. 2014 г.	13	104
2 кв. 2014 г.	14	96,2
3 кв. 2014 г.	15	95,1
4 кв. 2014 г.	16	97,5
1 кв. 2015 г.	17	101
2 кв. 2015 г.	18	93,5
3 кв. 2015 г.	19	92
4 кв. 2015 г.	20	94,6

2. Имеются поквартальные данные о разрешениях на строительство нового частного жилья, выданных в США в 1990 - 1994 гг в % к уровню 1987 года. Постройте аддитивную модель временного ряда. Рассчитайте прогноз на 1 квартал 1995 года.

Таблица 35

№ квартала	*t*	*yt*
1 кв.1990 г.	1	68,3
2 кв.1990 г.	2	61,9
3 кв.1990 г.	3	65,1
4 кв.1990 г.	4	74,1
1 кв. 1991 г.	5	67,5
2 кв. 1991 г.	6	66,8
3 кв. 1991 г.	7	65,5
4 кв. 1991 г.	8	73,6
1 кв. 1992 г.	9	73,7
2 кв. 1992 г.	10	61,0
3 кв. 1992 г.	11	71,4
4 кв. 1992 г.	12	82,3
1 кв. 1993 г.	13	79,0
2 кв. 1993 г.	14	72,7
3 кв. 1993 г.	15	73,9
4 кв. 1993 г.	16	85,3
1 кв. 1994 г.	17	83,4
2 кв. 1994 г.	18	76,2
3 кв. 1994 г.	19	81,6
4 кв. 1994 г.	20	89,3

8.3 Вопросы для самопроверки

1. Дайте понятие временного ряда. Перечислите его основные характеристики.

2. В чем экономическое содержание составляющих временного ряда?

3. Как выделяется случайная составляющая временного ряда?

4. Для чего строится математическая модель временного ряда?

5. Что такое автокорреляция уровней временного ряда и как ее можно оценить количественно?

6. Перечислите основные виды трендов.

7. Выпишите общий вид аддитивной и мультипликативной моделей временного ряда.

8. Перечислите этапы построения модели временного ряда.

9. С какими целями проводится выявление и устранение сезонного эффекта?

СПИСОК ЛИТЕРАТУРЫ

Основная литература:

1. Афанасьев, В.Н. Анализ временных рядов и прогнозирование [Электронный ресурс] / В.Н. Афанасьев, М.М. Юзбашев. — Электрон. текст. дан. — М.: Финансы и статистика, 2012. — 320 с. — Режим доступа: www.e.lanbook.com.

2.Тимофеев, В.С. Эконометрика /В.С. Тимофеев.- М.: Юрайт, 2014.- 328 с.

3.Плотников, А.Н. Элементарная теория анализа и статистическое моделирование временных рядов [Электронный ресурс] /А.Н. Плотников. — Электрон. текст. дан. — СПб.: Лань, 2016. — 220 с. — Режим доступа: www.e.lanbook.com.

4. Уткин, В.Б. Эконометрика [Электронный ресурс]: учебник /В.Б. Уткин. — Электрон. текст. дан. — М.: Дашков и К, 2013. — 562 с. — Режим доступа: www.e.lanbook.com.

5.Яковлев, В.П. Эконометрика [Электронный ресурс]: учебник /В.П. Яковлев. — Электрон. текст. дан. — М.: Дашков и К, 2016. — 384 с. — Режим доступа: www.e.lanbook.com.

Дополнительная литература:

1. Буховец, А.Г. Алгоритмы вычислительной статистики в системе R [Электронный ресурс] / А.Г. Буховец, П.В. Москалев. — Электрон. текст. дан. — СПб.: Лань, 2015. — 160 с. — Режим доступа: www.e.lanbook.com.

2.Воскобойников, Ю.Е. Основы вычислений и программирования в пакете MathCAD PRIME. [Электронный ресурс] / Ю.Е. Воскобойников, А.Ф. Задорожный. — Электрон. текст. дан. — СПб.: Лань, 2016. — 224 с. — Режим доступа: www.e.lanbook.com.

3.Иода, Е.В. Статистика: учеб. пособие / Е.В. Иода.— М.: Вузовский учебник: ИНФРА-М, 2016.— 303 с.

4.Шириков, В.Ф. Математическая статистика: учеб. пособие / В.Ф. Шириков, С.М. Зарбалиев.— М.: КолосС, 2009.— 480 с.: ил.

Савельева Екатерина Владимировна

Островская Ирина Эдуардовна

Моделирование и статистическая обработка результатов научных исследований: учебное пособие для обучающихся по направлениям подготовки 35.06.01 Сельское хозяйство; 35.06.02 Лесное хозяйство; 35.06.04 Технологии, средства механизации и энергетическое оборудование в сельском, лесном и рыбном хозяйстве; 36.06.01 Ветеринария и зоотехния; 38.06.01 Экономика ФГБОУ ВПО Приморская ГСХА

Подписано в печать _____Формат 60 х 84 1/16. Бумага писчая.

Печать офсетная. Уч. - изд.л. ___. Тираж ____экз. Заказ_____

ФГБОУ ВПО Приморская ГСХА

Адрес: 692510, г. Уссурийск, пр-т. Блюхера, 44

Участок оперативной полиграфии ФГБОУ ВПО Приморская ГСХА

692500, г. Уссурийск, ул. Раздольная, 8.