МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«ТЮМЕНСКИЙ ГОСУДАРСТВЕННЫЙ НЕФТЕГАЗОВЫЙ УНИВЕРСИТЕТ»

С. В. Овчинникова, Е. Н. Гура

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В УПРАВЛЕНИИ

Рекомендовано Советом Учебно-методического объединения по образованию в области менеджмента в качестве учебного пособия для студентов высшего профессионального образования, обучающихся по направлению 080200.68 «Менеджмент»

Тюмень
ТюмГНГУ
2012

УДК 519.2(075.8)
ББК 65в 6я73
О 355

Рецензенты:
доктор социологических наук, профессор А. Н. Силин
доктор экономических наук, профессор Т. Г. Линник
доктор технических наук, профессор В. М. Спасибов

Овчинникова, С. В.

О355 Статистический анализ данных в управлении : учебное пособие / С. В. Овчинникова, Е. Н. Гура. – Тюмень : ТюмГНГУ, 2012. – 126 с.

ISBN 978-5-9961-0629-5

В учебном пособии приводятся основные модели и методы анализа социально-экономических процессов и показатели по статистическим данным. Рассматриваются алгоритмы построения эконометрических моделей, методы оценки параметров модели и проверки их значимости. Приводятся примеры решения типовых задач и реализация их на компьютере с помощью Excel.

Пособие предназначено для студентов направления «Менеджмент» и рекомендуется при изучении дисциплин «Статистические методы обработки данных», «Эконометрика», «Математические методы и компьютерные технологии в науке и образовании», «Методы социально-экономического прогнозирования»

УДК 519.2(075.8)
ББК 65в6я73

ISBN 978-5-9961-0629-5

Содержание

Глава 1. Базовые понятия статистики

1.1. Выборочная и генеральная совокупности. Типы выборок

Установление статистических закономерностей, присущих массовым случайным явлениям, основано на изучении статистических данных – сведения о том, какие значения принял в результате наблюдений интересующий нас признак.

Различные значения признака называются *вариантами*.

В практике статистических наблюдений различают два вида: сплошное, когда изучаются все объекты совокупности, и выборочное, когда изучается часть объектов. Примером сплошного наблюдения является перепись населения, охватывающая все население страны. Выборочными наблюдениями являются, например, социологические исследования.

Генеральной совокупностью называют множество всех объектов некоторого наблюдения в совокупности с множеством всех значений этого наблюдения, соответствующих каждому объекту.

Выборкой объема n называют множество из n объектов, реально подвергшихся наблюдению, в совокупности с n значениями наблюдения для каждого объекта. Например, генеральная совокупность – группа студентов, присутствующих на паре; выборка объема n – n студентов, опрошенных преподавателем, вместе с оценками, полученными ими.

Основная задача статистики – получить обоснованные выводы о свойствах генеральной совокупности, анализируя, извлеченную из нее выборку.

Для того чтобы по данным выборки можно было достаточно уверенно судить об интересующем признаке генеральной совокупности, необходимо, чтобы выборка была *репрезентативной (представительной),* т.е. достаточно полно описываа изучаемые признаки генеральной совокупности. Условием обеспечения репрезентативности выборки является, согласно закону больших чисел, соблюдение случайности отбора, т.е. все объекты имеют одинаковую вероятность попасть в выборку.

Повторной называют выборку, при которой отобранный объект перед извлечением следующего возвращается в генеральную совокупность. *Бесповторной* называют выборку, при которой отобранный объект перед извлечением следующего не возвращается в генеральную совокупность. На практике чаще используется безповторная выборка.

1.2. Ряды распределения и их построение

Важнейшей частью статистического анализа является построение рядов распределения (структурной группировки) с целью выделения характерных свойств и закономерностей изучаемой совокупности. В зависимости от того, какой признак (количественный или качественный) взят за основу группировки данных, различают соответственно типы рядов распределения.

Если за основу группировки взят качественный признак, то такой ряд распределения называют *атрибутивным* (распределение по видам труда, по полу, по профессии, по религиозному признаку, национальной принадлежности и т.д.).

Статистическим распределением выборки (статистическим рядом) называется перечень вариант и соответствующим им частотами или частостями.

Если ряд распределения построен по количественному признаку в порядке возрастания, то такой ряд называют *вариационным*. Построить вариационный ряд – значит упорядочить количественное распределение единиц совокупности по значениям признака, а затем подсчитать числа единиц совокупности с этими значениями (построить групповую таблицу).

Выделяют две формы вариационного ряда: дискретный ряд и интервальный ряд.

Дискретный ряд – это такой вариационный ряд, в основу построения которого положены признаки с прерывным изменением (дискретные признаки). К последним можно отнести тарифный разряд, количество детей в семье, число работников на предприятии и т.д. Эти признаки могут принимать только конечное число определенных значений.

Дискретный вариационный ряд представляет таблицу, которая состоит из двух граф. В первой графе указывается конкретное значение признака, а во второй – число единиц совокупности с определенным значением признака. Частота (частота повторения) – число повторений отдельного варианта значений признака, обозначается n_i , а сумма частот, равна объему исследуемой совокупности n, т.е. $\sum_{i=1}^{k} n_i = n.$

Очень часто таблица дополняется графой, в которой подсчитываются частости w_i, выраженные в относительных числах (долях или процентах), причем $\sum_{i=1}^{k} w_i = 1$. Примером дискретного ряда является распределение 40 рабочих цеха по тарифному разряду (табл.1.1)

Таблица 1.1

Дискретный вариационный ряд

Тарифный разряд, x_i	1	2	3	4
Частота n_i	12	15	8	5
Частость $w_i=n_i/n$	0,3	0,375	0,2	0,125

Если признак имеет непрерывное изменение (размер дохода, стаж работы, стоимость основных фондов предприятия и т.д., которые в определенных границах могут принимать любые значения), то для этого признака нужно строить интервальный вариационный ряд.

Для построения интервального ряда шаг выборки определяют по формуле Стерджеса $h=(x_{max}-x_{min})/(1+3,322lgn)$, где n – объём выборки.

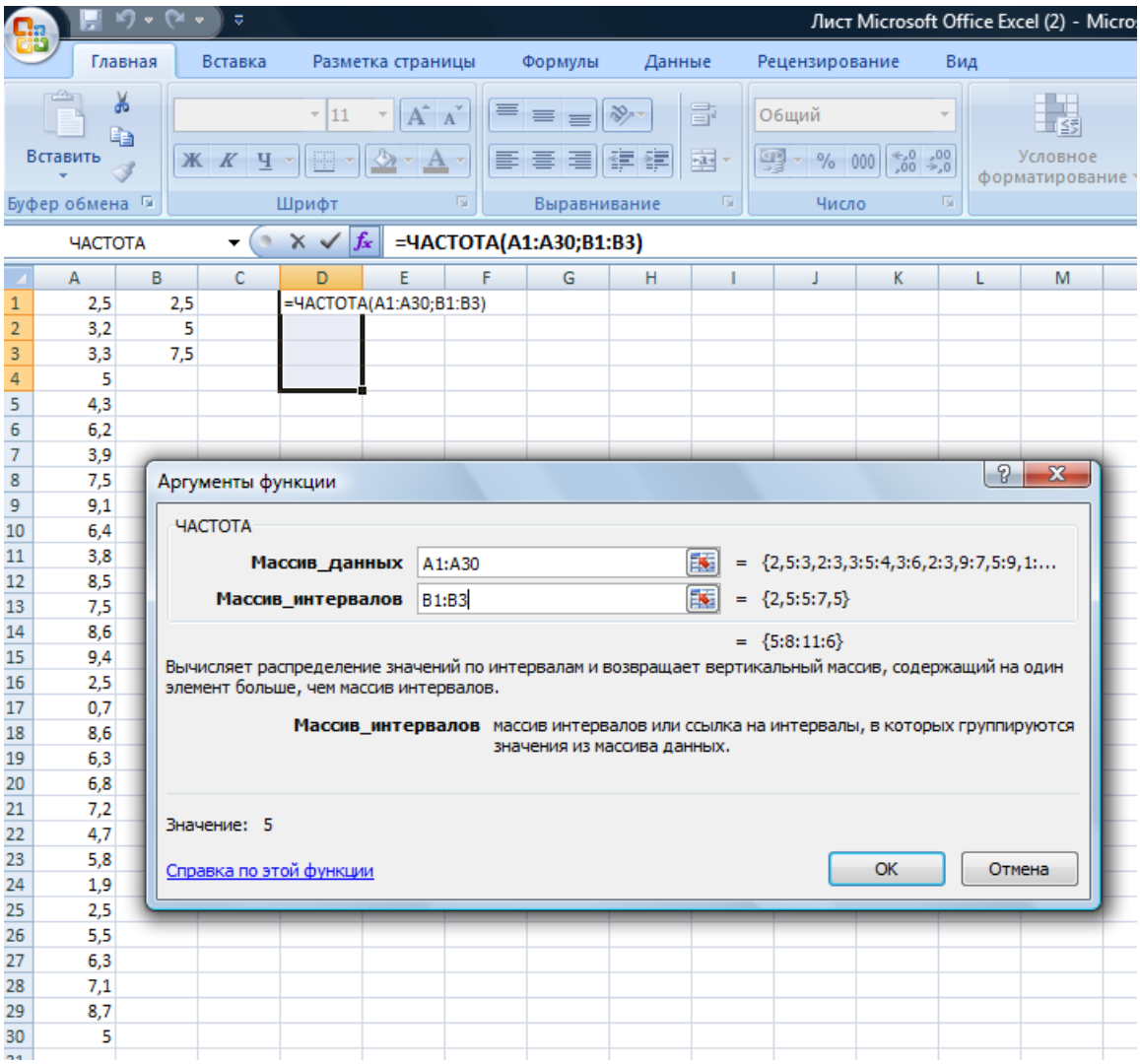
Интервал $[x_i;x_{i+1})$ наблюдённых значений	Частота n_i	Частость $w_i=n_i/n$
x_1 -x_2	n_1	w_1
x_2 –x_3	n_2	w_2
…	…	…
x_m -x_{m+1}	n_m	w_m

В **Excel** для построения интервальных временных рядов используется функция ЧАСТОТА.

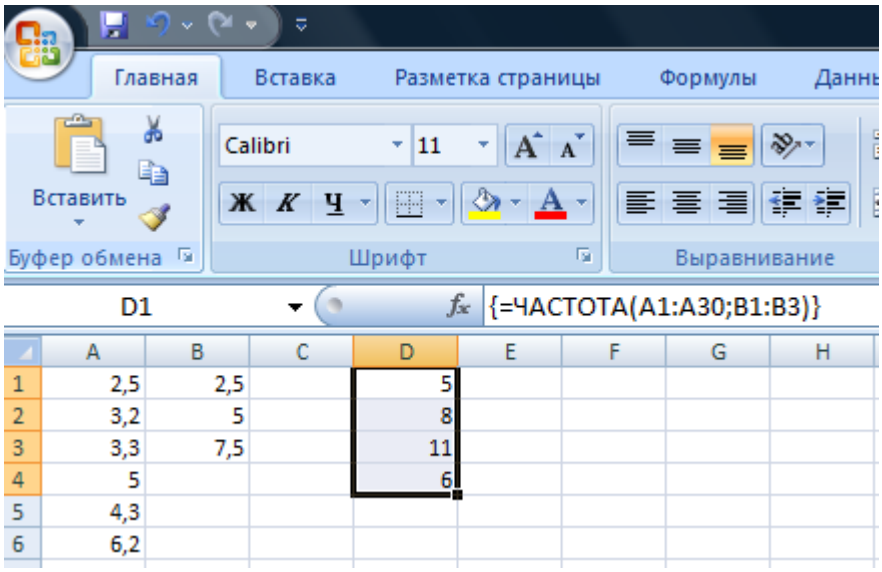
Алгоритм действий покажем на примере.

Дан массив данных: 2,5; 3,2; 3,3; 5; 4,3; 6,2; 3,9; 7,5; 9,1; 6,4; 3,8; 8,5 7,5 8,6; 9,4; 2,5; 0,7; 8,6; 6,3; 6,8; 7,2; 4,7; 5,8; 1,9; 2,5; 5,5; 6,3; 7,1; 8,7; 5,0

1. Задаем границы интервалов(массив карманов): 2,5;5; 7,5.
2. Формируем исходные данные в виде двух столбцов: массив данных и массив карманов.
3. Так как массив карманов содержит три значения, мышью выделяем четыре (3+1) смежные ячейки для вывода частот попадания значений из массива данных в заданные интервалы.
4. Вызываем **Мастер функции**.
5. В **Мастере функций** из категории *Статистические* выбираем функцию ЧАСТОТА.



6. Выделяем *Массив данных* и *Двоичный массив* и при помощи комбинации клавиш CTRL+SHIFT+ENTER выводим результат решения.



В результате получаем:

Интервал значений	0 – 2,5	2,5 – 5,0	5,0 – 7,5	7,5 – 10
Частота n_i	4	6	11	6

Замечание. Если диапазоны (карманы, в терминологии Excel) не указаны, процедура разбивки и определение границ диапазонов (карманов) производится автоматически на основании формулы Стерджесса. Такая предварительная разбивка может быть полезна перед заданием интервалов.

Накопленной частотой $n_i^{нак}$ называется число вариант выборки меньших данного числа x, а их отношения к объему выборки $w_i^{нак} = n_i^{нак}/n$ – *относительной накопленной частотой.* Относительные накопленные частоты – это статистические аналоги значений функции распределения $F(x_i)$ дискретной случайной величины.

Для наглядности строят различные графики статистического распределения.

Полигон частот (многоугольник распределения) – ломаная, соединяющая точки с координатами (x_i, n_i) или (x_i, w_i).

Гистограмма – это ступенчатая фигура, состоящая из прямоугольников. Их основаниями служат частичные интервалы, а высоты равны частотам (частостям).

Кумулята – это кривая накопленных частот (частостей). Для её построения находят $w^{нак}$.

1.3. Показатели центра распределения

Для определения структуры совокупности используют особые средние показатели, к которым относятся выборочное среднее, медиана и мода, или так называемые структурные средние.

Выборочное среднее $\bar{x}_в$ – это среднее арифметическое вариант выборки. Если объем выборки равен n, то $\bar{x}_в = \dfrac{\sum_{i=1}^{k} n_i x_i}{n} = \sum_{i=1}^{k} w_i x_i$, где k – число различных вариант в дискретном статистическом распределении; n_i – частота варианты x_i ($i = \overline{1,k}$). Если же выборка сгруппирована в интервальный статистический ряд, то в качестве вариант x_i берут середины соответствующих интервалов $x_i^* = (x_i + x_{i-1})/2$. Конечно, при такой замене возникает ошибка, так как варианты, попавшие в интервал,

не обязаны все совпадать с числом x_i^*. Но эта ошибка не может быть слишком большой, особенно при достаточно больших n. Действительно, в среднем половина вариант, попавших в этот интервал, будет меньше числа x_i^*, а половина – больше, поэтому ошибки будут иметь разные знаки, и, в сумме компенсируют друг друга.

Если выборочная средняя рассчитывается на основе использования всех вариантов значений признака, то медиана и мода характеризуют величину того варианта, который занимает определенное среднее положение в ранжированном вариационном ряду.

Модой Mo (для дискретного статистического ряда) называется варианта x_i с наибольшей частотой (относительной частотой). Если же выборка сгруппирована в интервальный статистический ряд, то сначала определяют модальный интервал, т.е. интервал с наибольшей частотой (относительной частотой). В качестве моды можно взять середину модального интервала или в пределах этого интервала найти то значение признака, которое может являться модой.

Мода имеет широкое распространение в маркетинговой деятельности при изучении покупательского спроса, особенно при определении пользующихся наибольшим спросом размеров одежды и обуви, при регулировании ценовой политики.

Медианой Me называют варианту, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно, т. е. $n = 2k+ 1$, то $Me = x_{k+1}$, при четном $n = 2k$ медиана $Me = (x_k + x_{k+1})/2$. (для дискретного статистического ряда).

Например, для ряда:

x_i	2	3	5	6	7
n_i	13	7	22	7	10

медиана равна 5;

для ряда:

x_i	2	3	5	6	7	9
n_i	13	7	22	7	10	11

медиана равна (5 + 6)/2 = 5,5.

В интервальном ряду распределения сначала указывается интервал, в котором находится медиана.

Медианным является первый интервал, в котором сумма накопленных частот превысит половину общего числа наблюдений.

Численное значение медианы обычно определяют по формуле

$$Me = x_{ме} + i\frac{\frac{n+1}{2} - S_{(-1)}}{n_{ме}}, \qquad (1.1)$$

где x_{Me} – нижняя граница медианного интервала; i – величина интервала; $S_{(-1)}$ – накопленная частота интервала, которая предшествует медианному; n – частота медианного интервала.

Моду и медиану можно определить на основе графического изображения ряда. Медиана определяется по кумуляте. Для ее определения высоту наибольшей ординаты делят пополам. Через полученную точку проводят прямую, параллельную оси абсцисс, до пересечения её с кумулятой. Абсцисса точки пересечения является медианной величиной.

Мода определяется по гистограмме распределения. Для этого правую вершину модального прямоугольника соединяют с правым верхним углом предыдущего прямоугольника, а левую вершину модального прямоугольника – с левым верхним углом последующего прямоугольника. Абсцисса точки пересечения этих прямых и будет модой распределения

1.4. Показатели вариации (колеблемости) признака

Для характеристики размера вариации признака используются абсолютные и относительные показатели. К абсолютным показателям вариации относятся: размах варьирования, среднее линейное отклонение, среднее квадратическое отклонение, дисперсия.

Размах варьирования R (размах вариации) – простейшая мера разброса значений данной выборки.

$R = x_{\max} - x_{\min}$, где $x_{\max}$ - максимальное, $x_{\min}$ - минимальное значение признака. Величина показателя зависит только от двух крайних вариант и не учитывает степень колеблемости основной массы членов ряда. Этой величиной пользуются при работе с маленькими выборками.

Среднее линейное отклонение d и среднее квадратическое отклонение σ показывают, на сколько в среднем отличаются индивидуальные значения признака от среднего его значения.

Среднее линейное отклонение определяется по формуле $d = \frac{\sum_{i=1}^{\kappa} |x_i - \bar{x}_в| \cdot n_i}{n}$, где k – число различных вариант в дискретном статистическом распределении; n_i – частота варианты x_i $(i = \overline{1,k})$, $\bar{x}_в$ - выборочная средняя.

Среднее квадратическое отклонение σ и выборочная дисперсия $D_в(\sigma^2)$ определяются так:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_в)^2}{n}}; \; D_в = \frac{\sum_{i=1}^{k} n_i (x_i - \bar{x}_в)^2}{n}. \qquad (1.2)$$

Формул для расчета дисперсии может быть преобразована:

$$D_в = = \frac{\sum_{i=1}^{k} n_i x_i^{\,2}}{n} - (\bar{x}_в)^2, \qquad (1.3)$$

где k – число различных вариант выборки в дискретном статистическом распределении; n_i – частота варианты x_i ($i = \overline{1,k}$). Если же выборка сгруппирована в интервальный статистический ряд, то в качестве вариант x_i берут середины соответствующих интервалов $x_i^* = (x_i + x_{i-1})/2$.

Для большинства унимодальных законов распределения и, следовательно, выборок из таких генеральных совокупностей выполняются:

- «правило двух сигм»: более 95% значений выборки лежат в интервале $(\bar{x}_B - 2\sigma_B, \bar{x}_B + 2\sigma_B)$;
- «правило трех сигм»: более 99% значений выборки лежат в интервале $(\bar{x}_B - 3\sigma_B, \bar{x}_B + 3\sigma_B)$.

Коэффициент вариации $V = \frac{\sigma_в}{\bar{x}_в} \cdot 100\%$ служит для сравнения стандартных отклонений нескольких выборок.

Если коэффициенты вариации оказались величинами одного порядка, то средние рассеяния данных относительного среднего в этих выборках можно считать примерно равными. Тот из рядов, у которого коэффициент вариации больше, имеет большее рассеяние по отношению к выборочной средней. Коэффициент вариации — безразмерная величина, поэтому он пригоден для сравнения рассеяний вариационных рядов, варианты которых имеют различную размерность. Также коэффициент вариации применяют для характеристики однородности совокупности. Совокупность считается однородной, если коэффициент вариации не превышает 33% (для распределений, близких к нормальному).

1.5. Показатели формы распределения

Для получения приблизительного представления о форме распределения строят графики распределения (полигон и гистограмму). В практике статистических исследований приходится встречаться с самыми различными распределениями. Однородные совокупности характеризуются, как правило, одновершинными распределениями. Многовершинность свидетельствует о неоднородности изучаемой совокупности. В этом случае необходима перегруппировка данных с целью выделения более однородных групп.

Выяснение общего характера распределения предполагает оценку степени его однородности, а также исчисления показателей асимметрии и эксцесса.

Для сравнительного анализа степени асимметрии нескольких распределений рассчитывается относительный показатель асимметрии:

$$As = \frac{\bar{x}_в - Mo}{\sigma}. \qquad (1.4)$$

Величина показателя асимметрии может быть положительной и отрицательной. Положительная величина указывает на наличие правосторонней асимметрии, отрицательная – левосторонней. Чем больше абсолютная величина коэффициента, тем больше степень скошенности. Принято считать, что если коэффициент асимметрии меньше 0,25, то асимметрия незначительна, если свыше 0,5, то асимметрия значительна.

Наиболее распространенным является показатель асимметрии исчисляемый по формуле:

$$As = \frac{\mu_3}{\sigma^3}, \qquad (1.5)$$

где μ_3 – центральный момент третьего порядка; $\mu_3 = \dfrac{\sum_{i=1}^{k}(x_i - \bar{x}_в)^3 n_i}{n}$

Этот показатель асимметрии не только определяет степень асимметрии, но и указывает на наличие или отсутствие асимметрии в распределении признака в генеральной совокупности. Оценка степени существенности этого показателя дается с помощью средней квадратической ошибки, рассчитываемой по формуле $\sigma_{As} = \sqrt{\dfrac{6(n-1)}{(n+1)(n+3)}}$, где n – число наблюдений.

Если $\dfrac{|As|}{\sigma_{As}} > 3$, асимметрия существенна и распределение признака в генеральной совокупности не является симметричным.

Если $\frac{|As|}{\sigma_{As}} < 3$, асимметрия несущественна, ее наличие объясняется влиянием случайных обстоятельств.

Для симметричных распределений рассчитывается показатель эксцесса (островершинности):

$$Ex = \frac{\mu_4}{\sigma^4} - 3, \qquad (1.6)$$

где μ_4 – центральный момент четвертого порядка; $\mu_4 = \frac{\sum_{i=1}^{k}(x_i - \bar{x}_в)^4 n_i}{n}$

Эксцесс может быть положительным и отрицательным. У высоковершинных распределений показатель эксцесса имеет положительный знак, а у низковершинных – отрицательный знак. Предельным значением отрицательного эксцесса является значение *Ex*=-2; величина положительного эксцесса является бесконечной величиной.

Средняя квадратическая ошибка эксцесса вычисляется по формуле $\sigma_{Ex} = \sqrt{\frac{24n(n-2)(n-3)}{(n-1)^2(n+3)(n+5)}}$, где n – число наблюдений.

1.6. Кривые распределения

Основной целью анализа вариационных рядов является выявление закономерности распределения, исключая при этом влияние случайных для данного распределения факторов. По мере увеличения количества наблюдений (в пределах тоже однородной совокупности) при одновременном уменьшении величины интервала закономерность, характерная для данного распределения, будет выступать более ясно, а представляющая полигон частот ломанная линия будет приближаться к некоторой плавной линии. Эту линию называют кривой распределения.

Иными словами, кривая распределения есть графическое изображение в виде непрерывной линии изменения частот в вариационном ряду, которое функционально связано с изменением вариант. Кривая распределения отражает закономерность изменения частот при отсутствии случайных факторов. Графическое изображение облегчает анализ рядов распределения.

Известно достаточно много форм кривых распределения, по которым может выравниваться вариационный ряд, но в практике статистических исследований наиболее часто используется нормальное распределение. Распределения, близкие к нормальному распределению,

были обнаружены при изучении самых различных явлений как в природе, так и развитии общества.

Нормальное распределение зависит от двух параметров: средней арифметической и среднего квадратического отклонения. Его кривая выражается уравнением

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

где *f(x)* – ордината кривой нормального распределения; σ – среднее квадратическое отклонение, *e*=2,71 и π=3,14 – математические постоянные; *x* – варианты вариационного ряда; *a* – их средняя величина.

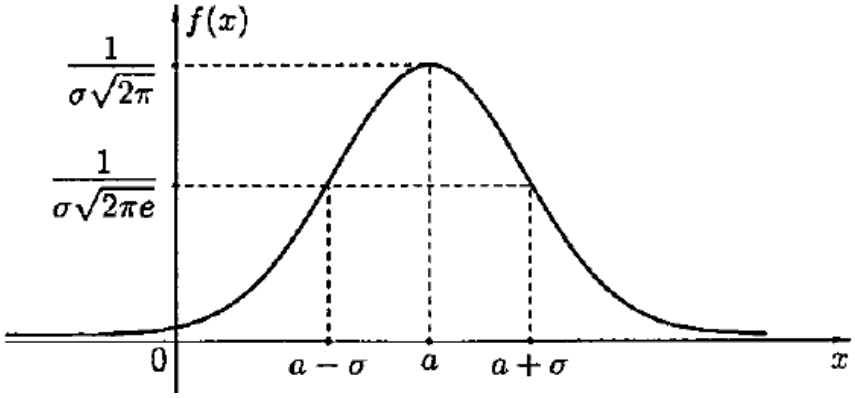


Рис. 1.1. Кривая плотности нормального распределения

В статистической практике большой интерес представляет решение вопроса о том, на сколько полученное в результате статистического наблюдения распределение признака в исследуемой совокупности, соответствует нормальному распределению.

1.7. Точечные и интервальные оценки параметров генеральной совокупности

Если закон распределения генеральной совокупности известен, а значения параметров, от которых этот закон зависит, неизвестны, возникает задача оценки значений этих параметров по имеющимся значениям $x_1, x_2,...,x_n$ извлеченной из генеральной совокупности выборки. Точечные оценки параметров – это оценки, полученные с помощью числовых значений подходящих статистик. При этом можно оценивать не только параметры, непосредственно входящие в формулу для закона распределения, но и числовые характеристики генеральной совокупности – математическое ожидание, дисперсию, асимметрию, эксцесс и т.д.

Для вычисления параметра θ исследовать все элементы генеральной совокупности не представляется возможным. Поэтому о параметре θ

можно пытаются судить по выборке $x_1, x_2,..., x_n$. Эти значения можно рассматривать как частные значения (реализации) *n* независимых случайных величин $X_1, X_2,..., X_n$, каждая из которых имеет тот же закон распределения, что и сама случайная величина *X*.

Оценкой θ_n^* параметра θ называют всякую функцию результатов наблюдений над случайной величиной X (иначе – статистику), с помощью которой судят о значении параметра θ: $\theta_n^* = \theta_n^*(x_1, x_2,..., x_n)$.

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т. е. приводить к грубым ошибкам. По этой причине при небольшом объеме выборки следует пользоваться интервальными оценками.

Пусть по данным выборки $x_1, x_2,..., x_n$ для оценки параметр известного распределения генеральной совокупности подобрана статистика $\theta_n^* = \theta_n^*(x_1, x_2,..., x_n)$. Заменяя неизвестное значение θ числом θ_n^*, мы совершаем ошибку. Тогда случайная величина $|\theta - \theta_n^*|$ - абсолютное значение ошибки. Если δ>0 и $|\theta - \theta_n^*|<\delta$, то чем меньше δ, тем оценка точнее. Таким образом, положительное число δ характеризует *точность оценки.* Однако статистические методы не позволяют категорично утверждать, что оценка θ_n^* удовлетворяет неравенству $|\theta - \theta_n^*|<\delta$; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется. Если известен закон распределения случайной величины $\theta_n^* = \theta_n^*(x_1, x_2,..., x_n)$, то эту вероятность можно найти $P(|\theta - \theta_n^*| < \delta) = \gamma$. Если для небольших δ вероятность γ достаточно велика, то число θ^* можно считать точной и надежной оценкой неизвестного параметра θ.

Надежностью (доверительной вероятностью) оценки θ *по* θ_n^* *называют вероятность γ, с которой осуществляется неравенство* $|\theta - \theta_n^*|<\delta$. Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Пусть $P(|\theta - \theta_n^*| < \delta) = \gamma.$

Доверительным называют интервал (θ_n^*-δ, θ_n^*+δ), который с заданной надежностью γ покрывает оцениваемый параметр.

1.8. Статистическая гипотеза.

Статистической гипотезой называют предположение либо о виде распределения генеральной совокупности (закон распределения неизвестен), либо о значениях неизвестных параметров известного закона распределения.

Статистические гипотезы делятся на параметрические (в них говорится о значениях параметров известного распределения) и непараметрические (в них высказывается предположение о виде закона распределения).

Наряду с проверяемой гипотезой (ее называют нулевой или основной H_0.) рассматривают и противоречащую ей гипотезу (ее называют конкурирующей или альтернативной H_1). Нулевая и альтернативная гипотеза взаимно исключают друг друга. Процедура проверки применяется к нулевой гипотезе H_0. Если в результате проверки нулевую гипотезу H_0 оказывается целесообразней отвергнуть, то принимается альтернативная гипотеза.

Простой называют гипотезу, содержащую только одно предположение. *Сложной* называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез.

Нулевая гипотеза – это простая гипотеза, в ней говорится о конкретных значениях параметров или о конкретном распределении. Альтернативная гипотеза сложная, в ней подразумевается бесконечно много возможностей.

Для проверки нулевой гипотезы используют специально подобранную случайную величину (выборочную статистику) K, точное или приближенное распределение которой известно в предположении справедливости нулевой гипотезы H_0. Эта случайная величина K называется статистическим критерием.

После выбора определенного критерия множество всех его возможных значений разбивают на два непересекающихся подмножества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается (оно называется *критической областью* или областью отвержения гипотезы), а другое — те значения, при которых она принимается (*область принятия гипотезы*).При справедливости нулевой гипотезы H_0 вероятность того, что случайная величина K примет значения из области принятия гипотезы, велика, а вероятность того, что случайная величина K примет значения из критической области, мала. Вероятность попадания в критическую область называется уровнем значимости и обозначается буквой α.

Тогда вероятность попадания в область принятия гипотезы H_0 равна $1-\alpha$.

По выборке, извлеченной из генеральной совокупности, вычисляют наблюдаемое значение критерия K – число $K_{\text{набл}}$. Если это число принадлежит критической области, то гипотеза H_0 отвергается, как противоречащая опытным данным. Справедливой в этом случае считается альтернативная гипотеза H_1. Если же число $K_{\text{набл}}$ принадлежит области принятия гипотезы H_0, то эта гипотеза считается согласующейся с опытными данными.

В зависимости от условия эксперимента различают разные виды критических областей:

- правостороннюю критическую область, определяемую неравенством $K_{\text{набл}}>k_{2\text{кр}}$ ($k_{2\text{кр}} > 0$). Интервал $(-\infty, k_{1\text{кр}})$ – область принятия нулевой гипотезы. Площадь под кривой на интервале $(k_{2\text{кр}}, \infty)$ равна α.

- левостороннюю критическую область, определяемую неравенством $K_{\text{набл}}<k_{1\text{кр}}$ ($k_{1\text{кр}}<0$). Интервал $(k_{2\text{кр}}, \infty)$ – область принятия нулевой гипотезы. Площадь под кривой на интервале $(-\infty, k_{1\text{кр}})$ равна α.

- двустороннюю критическую область, определяемую неравенствами $K_{\text{набл}} < k_{1\text{кр}}$, $K_{\text{набл}} > k_{2\text{кр}}$ ($k_{2\text{кр}}>k_{1\text{кр}}$). Интервал $(k_{1\text{кр}}, k_{2\text{кр}})$ – область принятия нулевой гипотезы. Площади под кривой на интервалах $(-\infty, k_{1\text{кр}})$ и $(k_{2\text{кр}}, \infty)$ равны $\alpha/2$ каждая.

Глава 2. Теоретические основы регрессионного анализа

2.1. Парная линейная регрессия

Термин «регрессия» (движение назад, возвращение в прежнее состояние) был введен Фрэнсисом Галтоном в конце XIX века при анализе зависимости между ростом родителей и ростом детей. Галтон заметил, что рост детей у очень высоких родителей в среднем меньше, чем средний рост родителей. У очень низких родителей, наоборот, средний рост детей выше. И в том и в другом случае средний рост детей стремится к среднему росту людей в данном регионе. Отсюда и выбор термина, отражающего такую зависимость.

В настоящее время под *регрессией* понимается функциональная зависимость между объясняющими переменными и условным математическим ожиданием (средним значением) зависимой переменной, которая строится с целью предсказания (прогнозирования) этого среднего значения при фиксированных значениях первых.

Так как реальные значения зависимой переменной не всегда совпадают с ее условными математическими ожиданиями и могут быть различными при одном и том же значении объясняющей переменной, фактическая зависимость должна быть дополнена некоторым слагаемым ε, которое является случайной величиной. Связи между зависимой и объясняющей(ими) переменными которые выражаются соотношениями

$$Y = M\left(Y/_{X}\right) + \varepsilon, \qquad (2.1)$$

$$Y = M\left(Y/_{x_1, x_2, \dots, x_m}\right) + \varepsilon, \qquad (2.2)$$

называют *регрессионными моделями (уравнениями)*.

Предположим, что истинная зависимость между x и y – линейная, т.е. существует некоторая прямая $Y = \beta_0 + \beta_1 x$, отражающая истинную зависимость. Задача регрессионного анализа состоит в получении оценок β_0, β_1 и положения прямой. Пусть имеется набор значений двух переменных $X_t, Y_t, t = 1,2, \dots, n$; можно отобразить пары (X_t, Y_t) точками на плоскости (рис.2.1)

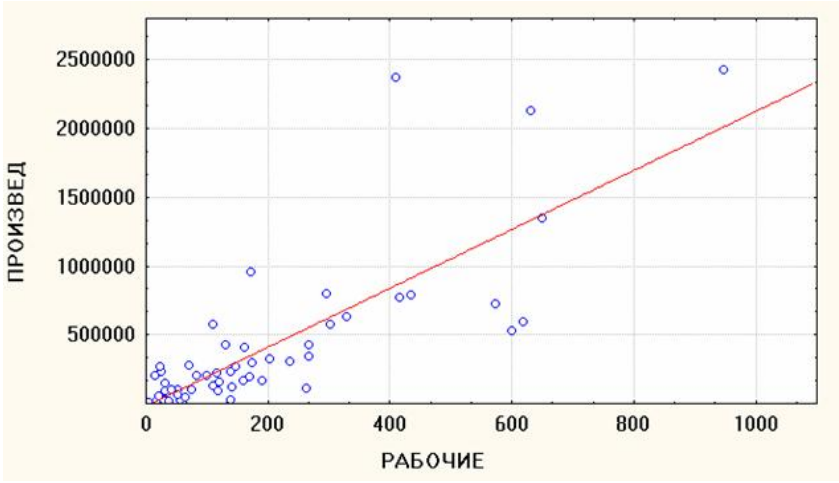


Рис. 2.1. Поле корреляции

Существование отклонений от прямой регрессии, т.е. случайных слагаемых ε, объясняется рядом причин. Среди таких причин выделим наиболее существенные.

1) Ошибки измерения. Например, при сборе данных об урожайности сельскохозяйственных культур, результаты работы в отчетах могут завышаться или занижаться в зависимости от экономической политики.

2) Невключение объясняющих переменных. Любая регрессионная модель является упрощением реальной ситуации. Возможно, что простая зависимость $Y = \beta_0 + \beta_1 x$ является очень большим упрощением. Наверняка существуют и другие влияющие на изменение Y, факторы, которые не включены в уравнение.

3) Неправильный выбор вида зависимости. Из-за слабой изученности исследуемого процесса может быть неверно подобрана функция, его моделирующая.

4) Ограниченность статистических данных. Часто строятся модели, выражаемые непрерывными функциями. Но для этого используется набор данных, имеющих дискретную структуру.

Решение задачи построения качественного уравнения регрессии, соответствующего эмпирическим данным и целям исследования, является достаточно сложным и многоступенчатым процессом. Его можно разбить на три этапа:

1) выбор формулы уравнения регрессии;
2) определение параметров выбранного уравнения;
3) анализ качества уравнения и проверка адекватности уравнения эмпирическим данным.

Выбор формы связи переменных называется спецификацией уравнения регрессии. Для парной регрессии выбор формулы обычно

осуществляется по графическому изображению эмпирических данных в виде точек в декартовой системе координат, которое называется полем корреляции.

В случае множественной регрессии определение подходящего вида зависимости является более сложной задачей.

Если функция линейна, то говорят о линейной регрессии. Линейная регрессия представляет собой линейную функцию между условным математическим ожиданием

$$y = M\left(Y/_X\right) = \beta_0 + \beta_1 x, \qquad (2.3)$$

где $M\left(Y/_X\right)$ - условное математическое ожидание зависимой переменной Y и одной объясняющей переменной X (x_i – значение независимой переменной в i-ом наблюдении, $i=1,2,\ldots n$), β_0 и β_1 - неизвестные параметры генеральной совокупности, которые подлежат оценке по результатам выборочных наблюдений. Так как каждое индивидуальное значение y_i отклоняется от соответствующего условного математического ожидания, в соотношении (2.3) вводится случайное слагаемое ε_i. В этом случае линейная модель регрессии имеет вид

$$y_i = M(Y/_X = x_i) + \varepsilon_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad (2.4)$$

где ε_i – взаимно независимые случайные величины с нулевым математическим ожиданием и дисперсией σ^2, т.е. $M(\varepsilon_i)=0$; $D(\varepsilon_i) = \sigma^2$ для всех $i = 1,2,3, \ldots, n$.

Соотношение (2.3) называют теоретической линейной регрессионной моделью; β_0, β_1 – теоретическими параметрами регрессии; ε_i – случайным отклонением. Для определения значений теоретических коэффициентов регрессии необходимо знать и использовать все значения переменных X и Y генеральной совокупности, что практически невозможно.

Задачи линейного регрессионного анализа состоят в том, чтобы по имеющимся статистическим данным (x_i, y_i), $i = 1,2, \ldots, n$, для переменных X и Y :

a) получить наилучшие оценки неизвестных параметров β_0 и β_1;

b) проверить статистические гипотезы о параметрах модели;

c) проверить достаточно ли хорошо модель согласуется со статистическими данными (адекватность модели данным наблюдений).

Следовательно, по выборке ограниченного объема строят эмпирическое уравнение регрессии

$$\hat{y}_i = a + bx_i, \qquad (2.5)$$

где $\hat{y}_i$ - оценка условного математического ожидания $M(Y/X = x_i)$;

$a\ и\ b$ – оценки неизвестных параметров $\beta_0\ и\ \beta_1$, называемые эмпирическими коэффициентами регрессии.

В конкретном случае

$$y_i = a + bx_i + e_i, \qquad (2.6)$$

где e_i - оценка теоретического случайного отклонения.

В силу несовпадения статистических данных генеральной совокупности и выборки оценки $a\ и\ b$ практически всегда отличаются от истинных значений $\beta_0\ и\ \beta_1$, что приводит к несовпадению эмпирической и теоретической линий регрессии. Различные выборки из одной и той же генеральной совокупности приводят к определению отличающихся друг от друга оценок. Задача состоит в том, чтобы по конкретной выборке (x_i, y_i), $i = 1,2, \dots, n$ найти оценки $a\ и\ b$ неизвестных параметров $\beta_0\ и\ \beta_1$ так, чтобы построенная линия регрессии являлась наилучшей среди всех других прямых. Таким образом, построенная прямая $\hat{y}_i = a + bx_i$ должна быть «ближайшей» к точкам наблюдений по их совокупности.

Самым распространенным является метод нахождения коэффициентов, при котором минимизируется сумма $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Он получил название метод наименьших квадратов (МНК). Этот метод является наиболее простым с вычислительной точки зрения. Оценки коэффициентов регрессии, найденные МНК при определенных предпосылках, обладают рядом оптимальных свойств.

2.2. Метод наименьших квадратов

Пусть по выборке (x_i, y_i), $i = 1,2, \dots, n$, требуется определить оценки $a\ и\ b$ эмпирического уравнения регрессии (2.3).

Метод наименьших квадратов позволяет получить такие оценки $a\ и\ b$ параметров $\beta_0\ и\ \beta_1$, при которых сумма квадратов отклонений ε_i - фактических значений признака y_i от расчетных (теоретических) $\hat{y} = a + bx + \varepsilon$ является минимальной:

$$Q(a,b) = \sum_i (y_i - \hat{y})^2 = \sum_i (y_i - a - bx_i)^2 = \sum_i \varepsilon_i^2 \to \min \to \qquad (2.7)$$

Функция Q дифференцируема по $a\ и\ b$, поэтому для отыскания минимума функции найдем частные производные и приравняем их к нулю:

$$\begin{cases} \frac{\partial Q}{\partial a} = -2\sum y_i + 2na + 2b\sum x_i = 0 \\ \frac{\partial Q}{\partial b} = -2\sum y_i x_i + 2a\sum x_i + 2b\sum x_i^2 = 0 \end{cases} \quad (2.8)$$

После преобразований получаем систему уравнений:

$$\begin{cases} na + b\sum x_i = \sum y_i \\ a\sum x_i + b\sum x_i^2 = \sum y_i x_i \end{cases} \quad (2.9)$$

Система (2.9) называют системой нормальных уравнений МНК.

Решая систему (2.9) относительно a и b получим:

$$b = \frac{\sum_{i=1}^n x_i y_i - \frac{1}{n}\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{\sum_{i=1}^n x_i^2 - \frac{1}{n}\left(\sum_{i=1}^n x_i\right)^2}; \quad a = \frac{1}{n}\sum_{i=1}^n y_i \cdot b\frac{1}{n}\sum_{i=1}^n x_i. \quad (2.10)$$

Относительно оценок (2.10) можно сделать следующие выводы:

1. Оценки МНК являются функциями от выборки, что позволяет их легко рассчитывать.
2. Оценки МНК являются точечными оценками теоретических коэффициентов регрессии.
3. Эмпирическая прямая регрессии обязательно проходит через точку $\bar{x}, \bar{y}$.
4. Эмпирическое уравнение регрессии построено таким образом, что сумма отклонений $\sum_{i=1}^n e_i^2 = 0$, а также среднее значение отклонений $\bar{e} = \left(\sum_{i=1}^n e_i / n\right) = 0$.
5. Отклонения e_i не коррелированны с наблюдаемыми значениями y_iпеременной Y.
6. Отклонения e_i не коррелированны с наблюдаемыми значениями x_i переменной X.

Для оценки коэффициентов a *и* b можно воспользоваться готовыми формулами, которые вытекают из системы (2.9):

$$a = \bar{y} - b\bar{x}, \quad b = \frac{cov(x,y)}{\sigma_x^2} = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}, \quad (2.11)$$

где $\bar{x} = \frac{\sum x_i}{n}$, $\bar{y} = \frac{\sum y_i}{n}$, $cov(x,y) = \overline{y \cdot x} - \bar{y} \cdot \bar{x}$,

$\overline{yx} = \frac{\sum x_i y_i}{n}$, $\sigma_x^2 = \overline{x^2} - \bar{x}^2$, $\overline{x^2} = \frac{\sum x_i^2}{n}$.

Коэффициент b при x называется выборочным коэффициентом регрессии, он показывает среднее изменение результата y при изменении фактора на единицу своего измерения. Коэффициент a указывает на значение результирующего признака при нулевом значении фактора. Это важный индикатор для выбора вида уравнения. Например, если в результате вычислений коэффициент a оказался отрицательным, а экономический смысл задачи диктует положительность или равенство нулю показателя a, значит, выбор вида уравнения был неудачным.

Рассмотрим следующий пример.

Пример 2.1. Для анализа зависимости объема потребления *Y*(у.е.) домохозяйств от располагаемого дохода *X* (у.е.) отобрана выборка объема 14, результаты которой приведены в таблице 2.1. Требуется:

1) определить вид зависимости;
2) по МНК оценить параметры уравнения регрессии *Y* на *X*;
3) оценить силу линейной зависимости между *X* и *Y*;
4) спрогнозировать потребление при *X*=155.

Таблица 2.1.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
x_i	106	107	108	109	110	112	114	116	120	126	130	136	142	145
y_i	103	102	104	107	109	110	112	114	118	124	127	133	134	141

Решение. Предположим, что зависимость между *X* и *Y* линейная: $\hat{y} = a + bx$.

Для наглядности вычислений по МНК построим таблицу 2.2.

Согласно МНК, имеем:

$$b = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{14206 - 117 \cdot 120{,}071}{14584{,}79 - 14417{,}15} = 0{,}94;$$

$$a = \bar{y} - b \cdot \bar{x} = 117 - 0{,}94 \cdot 120{,}071 = 4{,}087.$$

Таблица 2.2

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$	e_i	e_i^2
1	106	103	11236	10918	10609	103,77	-0,77	0,59
2	107	102	11449	10914	10404	104,71	-2,71	7,33
3	108	104	11664	11232	10816	105,65	-1,65	2,72
4	109	107	11881	11663	11449	106,59	0,41	0,17
5	110	109	12100	11990	11881	107,53	1,47	2,16
6	112	110	12544	12320	12100	109,41	0,59	0,35
7	114	112	12996	12768	12544	111,29	0,71	0,50
8	116	114	13456	13224	12996	113,17	0,83	0,69
9	120	118	14400	14160	13924	116,93	1,07	1,14
10	126	124	15876	15624	15376	122,58	1,42	2,03
11	130	127	16900	16510	16129	126,34	0,66	0,44
12	136	133	18496	18088	17689	131,98	1,02	1,04
13	142	134	20164	19028	17956	137,62	-3,62	13,11
14	145	141	21025	20445	19881	140,44	0,56	0,31
сумма	1681	1638	204187	198884	193754	1638,01	-0,01	32,59
среднее	120,071	117	14584,79	14206	13839,5	117		

Таким образом, уравнение парной линейной регрессии имеет вид:

$$\hat{y} = 4{,}087 + 0{,}94 \cdot \text{x}.$$

Рассчитаем по данному уравнению $\hat{y}$, а также $e_i = y_i - \hat{y}_i$. Результаты вычислений приведены в таблице 2.2.

Для анализа тесноты связи рассчитаем коэффициент корреляции:

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sqrt{\overline{x^2} - \bar{x}^2} \cdot \sqrt{\overline{y^2} - \bar{y}^2}} = \frac{14206 - 117 \cdot 120{,}07}{12{,}95 \cdot 12{,}27} = \frac{157{,}64}{158{,}87} = 0{,}992.$$

Значение коэффициента корреляции показывает сильную линейную зависимость между рассматриваемыми переменными *X* и *У*.

Спрогнозируем потребление при располагаемом доходе x=155, для этого подставим данное значение в уравнение регрессии:

$$\hat{y}(155) = 4{,}08747 + 0{,}940378 \cdot 155 = 149{,}85.$$

Эмпирические коэффициенты регрессии $a\ и\ b$ являются лишь оценками теоретических коэффициентов $\beta_0\ и\ \beta_1$, а само уравнение отражает лишь общую тенденцию в поведении рассматриваемых переменных. Индивидуальные значения переменных в силу различных причин могут отклоняться от модельных значений. Эти отклонения выражены через значения e_i, которые являются оценками отклонений ε_i для генеральной совокупности.

2.3. Матричная форма записи

Парное уравнение регрессии можно записать в матричной форме:

$$Y = BX + E, \qquad (2.12)$$

где Y – вектор-столбец размерности $(n \times 1)$ фактических значений результативного признака; B - вектор-столбец размерности (2×1) подлежащих оценке параметров модели, т.е. коэффициента регрессии «*b*» и свободного члена (параметра «*a*» в уравнении $\hat{y} = a + bx$); $X = (x_0, x_1)$ – матрица размерности $(n \times 2)$ значений факторов. При этом x_0=1 и связано с наличием в уравнении регрессии свободного члена, а x_1 – значения включенного в уравнение регрессии фактора; E – вектор-столбец случайной величины e_i размерности $(n \times 1)$.

Матрица исходных данных имеет вид:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix},\ \text{X}= \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}. \qquad (2.13)$$

Оценивая параметры уравнения регрессии, находим вектор B вектор случайной компоненты E, т.е.

$$\text{B}=\begin{pmatrix} a \\ b \end{pmatrix}, \quad E = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}. \qquad (2.14)$$

Оценка вектора B составит

$$B = (X^T X)^{-1} X^T Y. \qquad (2.15)$$

Пример 2.2. По группе предприятий, выпускающих один и тот же вид продукции, рассматривается функция издержек $y = a + bx + e$. Исходные данные приведены в таблице 2.3.

Таблица 2.3

Номер предприятия	Выпуск продукции, тыс. ед., x	Затраты на производство, млн. руб., y
1	2	40
2	3	35
3	5	67
4	6	70
5	8	100
6	10	125
7	12	140

Решение. Для определения вектора $b = \begin{pmatrix} a \\ b \end{pmatrix}$ найдем предварительно матрицу $X^T X$:

$$1)\ X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 6 & 8 & 10 & 12 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 1 & 3 \\ 1 & 5 \\ 1 & 8 \\ 1 & 10 \\ 1 & 12 \end{pmatrix} = \begin{pmatrix} 7 & 46 \\ 46 & 382 \end{pmatrix};$$

2) Найдем обратную матрицу $\left(X^T X\right)^{-1} = \begin{pmatrix} 0{,}685 & -0{,}082 \\ -0{,}082 & 0{,}013 \end{pmatrix}$;

3) Вектор $X^T Y$ имеет вид:

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 2 & 3 & 5 & 6 & 8 & 10 & 12 \end{pmatrix} \begin{pmatrix} 40 \\ 35 \\ 67 \\ 70 \\ 100 \\ 125 \\ 140 \end{pmatrix} = \begin{pmatrix} 577 \\ 4670 \end{pmatrix};$$

4) Вектор оценок параметров регрессии равен:

$$B = \begin{pmatrix} 0{,}685 & -0{,}082 \\ -0{,}082 & 0{,}013 \end{pmatrix} \begin{pmatrix} 577 \\ 4670 \end{pmatrix} = \begin{pmatrix} 10{,}025 \\ 11{,}018 \end{pmatrix}$$

Оценка уравнения регрессии имеет вид:

$$\hat{y} = 10{,}025 + 11{,}018 \cdot x.$$

2.4. Проверка качества уравнения регрессии. Предпосылки метода наименьших квадратов

Метод наименьших квадратов предполагает ряд ограничений на поведение случайного слагаемого ε - *условия Гаусса-Маркова*:

1. Математическое ожидание случайного отклонения ε_i: $M(\varepsilon_i)=0$, $i=1, 2,\ldots,n$.
2. Гомоскедастичность (постоянство дисперсии отклонений). Дисперсия случайных отклонений ε_i постоянна: $D(\varepsilon_i)=D(\varepsilon_j)=\sigma^2$ для любых наблюдений i и j.
3. Отсутствие автокорреляции.
4. Случайное отклонение должно быть независимо от объясняющих переменных.
5. Модель является линейной относительно параметров.
6. Отсутствие мультиколлинеарности.
7. Ошибки ε_i, $i=1, 2,\ldots,n$, имеют нормальное распределение ($\varepsilon_i \sim N(0,\sigma)$).

Наряду с выполнимостью указанных предпосылок при построении классических регрессионных моделей делаются еще некоторые предположения. Например:

- объясняющие переменные не являются СВ;
- число наблюдений существенно больше числа объясняющих переменных;
- отсутствуют ошибки спецификации.

2.5. Средняя ошибка аппроксимации

Фактические значения результативного признака отличаются от теоретических, рассчитанных по уравнению регрессии. Чем меньше эти отличия, тем лучше качество модели. Величина отклонений фактических и расчетных значений результативного признака $(y - \hat{y}_x)$ по каждому наблюдению представляет собой ошибку аппроксимации. Так как $(y - \hat{y}_x)$ может быть величиной как положительной, так и отрицательной, ошибки аппроксимации для каждого наблюдения принято определять в процентах по модулю. Отклонения $(y - \hat{y}_x)$ можно рассматривать как абсолютную

ошибку аппроксимации, а $\left|\frac{(y-\hat{y}_x)}{y}\right| \cdot 100$ – как относительную ошибку аппроксимации. Для того чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, находят среднюю ошибку аппроксимации как среднюю арифметическую простую.

$$\overline{A} = \frac{1}{n} \cdot \sum \left|\frac{(y-\hat{y}_x)}{y}\right| \cdot 100 \qquad (2.16)$$

Допустимый предел значений $\overline{A}$ - не более 8-10%.

2.6. Оценка статистической значимости коэффициентов регрессии и корреляции

Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитываются t – критерий Стьюдента и доверительные интервалы каждого из показателей. Выдвигается гипотеза H_0 о случайной природе показателей, т.е. о незначимом их отличие от нуля. Оценка значимости коэффициентов регрессии и корреляции с помощью t – критерий Стьюдента проводится путем сопоставления их значений с величиной случайной ошибки:

$$t_a = \frac{a}{m_a}; \quad t_b = \frac{b}{m_b}; \quad t_r = \frac{r}{m_r}. \qquad (2.17)$$

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

$$m_b = \sqrt{\frac{\sum(y-\hat{y}_x)^2}{\sum(x-\bar{x})^2(n-2)}} = \sqrt{\frac{S_{ост}^2}{\sum(x-\bar{x})^2}} = \frac{S_{ост}}{\sigma_x\sqrt{n}};$$

$$m_a = \sqrt{\frac{\sum(y-\hat{y}_x)^2}{(n-2)}\frac{\sum x^2}{n\sum(x-\bar{x})^2}} = \sqrt{S_{ост}^2\frac{\sum x^2}{n^2\sigma_x^2}} = S_{ост}\frac{\sqrt{\sum x^2}}{n\sigma_x}. \qquad (2.18)$$

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

Сравнивая фактическое и табличное значения t – статистики принимаем или отвергаем гипотезу H_0. Табличные значения критерия находятся из таблицы Стьюдента (приложение 2) при df=(n-2) степенях свободы и уровне значимости $\alpha = 0{,}05$.

Если $t_{табл} < t_{факт,}$ то H_0 отклоняется и признается статистическая значимость коэффициентов регрессии и показателя тесноты связи.

Для расчета доверительных интервалов определяют предельную ошибку Δ для каждого показателя:

$$\Delta_a = t_{табл} m_a, \ \Delta_b = t_{табл} m_b. \qquad (2.19)$$

Формулы для расчета доверительных интервалов имеют следующий вид:

$$\gamma_a = a \pm \Delta_a;\ \gamma_{a_{min}} = a - \Delta_a;\ \gamma_{a_{max}} = a + \Delta_a;$$
$$\gamma_b = b \pm \Delta_b;\ \gamma_{b_{min}} = b - \Delta_b;\ \gamma_{b_{max}} = b + \Delta_b. \qquad (2.20)$$

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым.

Прогнозное значение $\hat{y}_{прог}$ определяется путем постановки в уравнение регрессии $\hat{y} = a + bx$ соответствующего прогнозного значения $x_{прог}$. Вычисляется средняя ошибка прогноза

$$m_{\hat{y}_{прог}} = \sigma_{ост} \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_{прог} - \bar{x})^2}{\sum (x_{прог} - \bar{x})^2}}, \qquad (2.21)$$

где $\sigma_{ост} = \sqrt{\frac{\sum (y - \hat{y})^2}{n - m - 1}}$;

и строятся доверительные интервалы прогноза:

$$\gamma_{\hat{y}_{прог}} = \hat{y}_{прог} \pm \Delta_{\hat{y}_{прог}};$$
$$\gamma_{\hat{y}_{прог_{min}}} = \hat{y}_{прог} - \Delta_{\hat{y}_{прог}}; \qquad (2.22)$$
$$\gamma_{\hat{y}_{прог_{max}}} = \hat{y}_{прог} + \Delta_{\hat{y}_{прог}}.$$

Пример 2.3. По 12 предприятиям концерна изучается зависимость прибыли (тыс. руб.) y от выработки продукции на одного человека (единиц) x по следующим данным (табл.2.4).

Требуется:

1. Построить линейное уравнение парной регрессии $\hat{y} = f(x)$.

2. Рассчитать линейный коэффициент парной корреляции и среднюю ошибку аппроксимации.

3. Оценить статистическую значимость параметров регрессии и корреляции.

4. Дать точечный и интервальный прогноз прибыли с вероятность 0,95, принимая уровень выработки равным 92 единицам.

Таблица 2.4.

Номер предприятия	Выработка продукции на одного человека, x	Прибыль предприятия, тыс. руб., y
1	78	133
2	82	148
3	87	134
4	79	154
5	89	162

Продолжение табл. 2.4

6	106	195
7	67	139
8	88	158
9	73	152
10	87	162
11	76	159
12	115	173

Решение

1. Для расчета параметров уравнения регрессии строим таблицу 2.5.

$$b = \frac{\overline{y \cdot x} - \bar{y} \cdot \bar{x}}{\frac{\sum x^2}{n} - (\bar{x})^2} = \frac{13484 - 85{,}6 \cdot 155{,}8}{\frac{7492{,}3}{12} - 85{,}6^2} = \frac{151{,}8}{164{,}94} = 0{,}92$$

$$a = \bar{y} - b \cdot \bar{x} = 155{,}8 - 0{,}92 \cdot 85{,}6 = 77$$

Получаем уравнение парной регрессии $y = 77 + 0{,}92 \cdot x$.

С увеличением выработки на 1 единицу прибыль возрастает в среднем на 0,92 тыс. руб.

Таблица 2.5.

№	x	y	xy	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	A_i	$(y - \hat{y})^2$
1	78	133	10374	6084	17689	149	-16	12,0	256
2	82	148	12136	6724	21904	152	-4	2,7	16
3	87	134	11658	7569	17956	157	-23	17,2	529
4	79	154	12166	6241	23716	150	4	2,6	16
5	89	162	14418	7921	26244	159	3	1,9	9
6	106	195	20670	11236	38025	174	21	10,8	441
7	67	139	9313	4489	19321	139	0	0,0	0
8	88	158	13904	7744	24964	158	0	0,0	0
9	73	152	11096	5329	23104	144	8	5,3	64
10	87	162	14094	7569	26244	157	5	3,1	25
11	76	159	12084	5776	25281	147	$\frac{1}{2}$	7,5	144
12	115	173	19895	13225	29929	183	-10	5,8	100
Итого	1027	1869	161808	89907	294377	1869	0	68,8	1600
Среднее значение	85,58	155,75	13484,0	7492,3	24531,4	x	x	5,7	
σ	12,95	16,53	x	x	x	x	x	x	
σ^2	167,7	273,4	x	x	x	x	x	x	

2. Тесноту линейной связи измеряет коэффициент корреляции:

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = 0{,}92 \cdot \frac{12{,}95}{16{,}53} = 0{,}721,$$

который можно также рассчитать по формуле

$$r_{xy} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\sigma_x \sigma_y} = \frac{13484 - 85{,}58 \cdot 155{,}75}{12{,}953 \cdot 16{,}533} = 0{,}722 .$$

Величина коэффициента корреляции означает достаточно тесную связь рассматриваемых признаков. Коэффициент детерминации $r_{xy}^2 = 0{,}52$ показывает, что 52% вариации прибыли связано с вариацией выработки продукции на одного работника.

Качество модели оценивается как хорошее, так как $\overline{A}$ не превышает 8-10%.

3. Оценку статистической значимости параметров регрессии проведем с помощью *t*-статистики Стьюдента и вычислим доверительные интервалы для каждого из показателей.

Выдвигаем гипотезу H_0 о статистически незначимых отличиях от нуля значении показателей: $a = b = r_{xy} = 0$.

$t_{таб} = 2{,}23$ для числа степеней свободы $df = n - 2 = 12 - 2 = 10$ и $\alpha = 0{,}05$.

Определим случайные ошибки параметров m_a, m_b и коэффициента корреляции $m_{r_{xy}}$ по формулам (2.18):

$$S = \sqrt{\frac{1600}{12-2}} = 12{,}6; \quad m_a = 12{,}6\frac{\sqrt{89907}}{12 \cdot 12{,}973} = 24{,}3; \quad m_b = S\sqrt{\frac{1}{\sum (x - \bar{x})^2}} = \frac{S}{\sigma_x \sqrt{n}};$$

$$m_b = \frac{12{,}6}{12{,}973 \cdot \sqrt{12}} = 0{,}281; \quad m_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}} = \sqrt{\frac{1 - 0{,}52}{12-2}} = 0{,}219.$$

Далее вычисляем значения *t*-критерия Стьюдента по формулам (2.17):

$$t_a = \frac{77}{24{,}3}; t_b = \frac{0{,}92}{0{,}281}; t_r = \frac{0{,}721}{0{,}219} = 3{,}3 .$$

Фактические значения *t*-статистики превосходят табличное значение на 5%-м уровне значимости при числе степеней свободы *n-2*=10: $t_a = 2{,}228$. Поэтому гипотеза H_0 отклоняется, т.е. a, b и r_{xy} отличаются от нуля не случайно и их значения статистически значимы.

Рассчитаем доверительный интервал для a и b, для чего определим предельную ошибку для каждого параметра:

$$\Delta_a = 2{,}23 \cdot 24{,}3 = 54; \; \Delta_b = 2{,}23 \cdot 0{,}281 = 0{,}62 .$$

Доверительные интервалы:

$\gamma_a = a \pm \Delta_a = 77 \pm 5; \gamma_{a\min} = 77 - 54 = 23; \gamma_{a\max} = 77 + 54 = 131;$

$\gamma_b = b \pm \Delta_b = 0{,}92 \pm 0{,}62; \gamma_{b\min} = 0{,}92 - {,}062 = 0{,}3; \gamma_{b\max} = 0{,}92 + 0{,}62 = 1{,}54.$

Анализ верхней и нижней границ доверительных интервалов приводит к выводу о том, что с вероятностью $p = 1 - \alpha = 0{,}95$ параметры a и

b, находясь в указанных границах, не принимают нулевых значений, т.е. не являются статистически незначимыми и существенно отличны от нуля.

4. Полученные оценки уравнения регрессии позволяют использовать его для прогноза. Если примем прогнозное значение выработки x=92, то точечный прогноз прибыли составит: $y_p = 77 + 0{,}92 \cdot 92 = 161{,}6$ тыс. руб. Чтобы получить интервальный прогноз, найдем стандартную ошибку предсказываемого значения прибыли m_{y_p} (2.21):

$$m_{y_p} = 12{,}6\sqrt{1 + \frac{1}{12} + \frac{(92 - 85{,}58)^2}{12 \cdot 12{,}973^2}} = 13{,}2 \; тыс.руб.$$

Предельная ошибка прогнозируемой прибыли составит: $y_p = 161{,}6 \pm 29{,}4$, т.е. при выработки, равной 92 ед., получим значение прибыли не меньше, чем $y_{p\min} = 161{,}6 - 29{,}4 = 132{,}2 \; тыс.руб.$, и не более чем $y_{p\max} = 161{,}6 + 29{,}4 = 191{,}0 \; тыс.руб.$

2.7. Множественная регрессия

На любой экономический показатель чаще всего оказывают влияние не один, а несколько факторов. Например, при построении модели потребления того или иного товара от дохода населения предполагается, что в каждой группе дохода одинаково влияние на потребление таких факторов, как цена товара, размер семьи и ее состав. Исследователь не может быть уверен в справедливости данного предположения. Чтобы иметь правильное представление о влияние дохода на потребление, необходимо изучить их корреляцию при неизменном уровне других факторов. Решение такой задачи предполагает отбор единиц совокупности с одинаковыми значениями всех других факторов, кроме дохода. Следует попытаться выявить влияние других факторов, введя их в модель, т.е. построить уравнение множественной регрессии.

Множественная регрессия – это уравнение связи с несколькими независимыми переменными:

$$y = f(x_1, x_2, \cdots, x_p),$$

где y – зависимая переменная (результативный признак);
x_1, x_2, …x_p – независимые переменные (факторы).

Построение модели множественной регрессии включает этапы:

1) выбор формы связи (уравнение регрессии);
2) отбор факторных признаков;
3) обеспечение достаточного объема совокупности для получения несмещенных оценок.

Для построения уравнения множественной регрессии чаще используются следующие функции:

- линейная – $y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_p \cdot x_p + \varepsilon;$
- степенная – $y = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot \cdots \cdot x_p^{b_p} \cdot \varepsilon;$
- экспонента – $y = e^{a+b_1 \cdot x + b_2 \cdot x_2 + \cdots + b_{p_1} \cdot x_p + \varepsilon};$
- гипербола – $y = \dfrac{1}{a + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_p \cdot x_p + \varepsilon}.$

Нелинейные формы зависимости приводятся к линейным путем линеаризации.

Ввиду четкой интерпретации параметров наиболее широко используется линейная и степенная функция.

Уравнение линейной множественной регрессии имеет вид:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_p \cdot x_p + \varepsilon\ , \qquad (2.23)$$

где ε – случайное отклонение.

Параметры $b_1, b_2, \ldots, b_p$ при x называют *коэффициентами «чистой» регрессии*. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизменном значение других факторов, закрепленных на среднем уровне.

Пример 2.4. Пусть зависимость расходов на продукты питания по совокупности семей характеризуется следующим уравнением:

$$\hat{y}_x = 0{,}6 + 0{,}47 \cdot x_1 + 0{,}82 \cdot x_2,$$

где y – расходы семьи за месяц на продукты питания, тыс. руб.; x_1- месячный доход на одного члена семьи, тыс. руб.; x_2- размер семьи, человек.

Анализ данного уравнения – с ростом дохода на одного члена семьи на 1000 рублей расходы на питание возрастут в среднем на 470 рублей при том же среднем размере семьи. Увеличение размера семьи при тех же ее доходах предполагает дополнительный рост расходов на питание на 820 рублей. Параметр a не имеет экономической интерпретации.

В степенной функции $y = a \cdot x_1^{b_1} \cdot x_2^{b_2} \cdot \cdots \cdot x_p^{b_p} \cdot \varepsilon$ коэффициенты b_j являются коэффициентами эластичности.

Пример 2.5. Предположим, что зависимость урожайности озимой пшеницы (y) от количества внесенных азотных (x_1) и фосфорных (x_2) удобрений на сельскохозяйственном предприятие задается следующим уравнением Кобба-Дугласса:

$$y = 27{,}4946 \cdot x_1^{-0{,}015221} \cdot x_2^{0{,}350537}.$$

Коэффициенты регрессии показывают, что внесение азотных удобрений под пшеницу на сельскохозяйственном предприятии

неэффективно, а увеличение дозы фосфорных удобрений на 1% обусловливает рост урожайности в среднем примерно на 0,35%.

2.8. Оценка параметров уравнения множественной регрессии

Для оценки параметров уравнения множественной регрессии применяют метод наименьших квадратов (МНК).

Для линейного уравнения (2.23) строится система нормальных уравнений, решение которой позволяет получить оценки параметров регрессии:

$$\begin{cases} \sum y = \mathrm{na} + \mathrm{b}_1 \sum x_1 + b_2 \sum x_2 + \cdots + b_p \sum x_p, \\ \sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1 x_2 + \cdots + b_p \sum x_p x_1, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \sum yx_p = a \sum x_p + b_1 \sum x_1 x_p + b_2 \sum x_2 x_p + \cdots + b_p \sum x_p^2. \end{cases} \qquad (2.24)$$

Ее решение может быть осуществлено методом определителей.

Возможен и другой подход к определению параметров множественной регрессии, когда на основе матрицы парных коэффициентов корреляции строится уравнение регрессии в стандартизованном масштабе:

$$t_y = \beta_1 t_{x1} + \beta_2 t_{x2} + \cdots + \beta_p t_{xp}, \qquad (2.25)$$

где $t_y = \dfrac{y - \bar{y}}{\sigma_y}$, $t_{x_i} = \dfrac{x_i - \bar{x}_i}{\sigma_{x_i}}$ – стандартизованные переменные для которых среднее значение равно нулю, т.е. $\bar{t}_y = \bar{t}_{xi} = 0$, а среднее квадратическое отклонение равно единице, т.е. $\sigma_{t_y} = \sigma_{t_x} = 1$, β_i – стандартизованные коэффициенты регрессии.

Стандартизованные коэффициенты регрессии определяются из следующей системы:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + \ldots + \beta_p r_{x_px_1}, \\ r_{yx_2} = \beta_1 r_{x_2x_1} + \beta_2 + \beta_3 r_{x_3x_2} + \ldots + \beta_p r_{x_px_2}, \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ r_{yx_p} = \beta_1 r_{x_px_1} + \beta_2 r_{x_px_2} + \beta_3 r_{x_px_3} + \ldots + \beta_p. \end{cases} \qquad (2.26)$$

Связь коэффициентов множественной регрессии b_i со стандартизованными коэффициентами β_i описывается соотношением:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{x_i}}. \qquad (2.27)$$

Параметр *a* определяется как

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_p\bar{x}_p. \qquad (2.28)$$

Стандартизованные коэффициенты регрессии показывают, на сколько сигм изменится в среднем результат, если соответствующий фактор x_i изменится на одну сигму при неизменном среднем уровне других факторов. Коэффициенты регрессии β_i сравнимы между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат.

Пример 2.6. По 25 территориям страны изучается влияние климатических условий на урожайность зерновых y (ц/га). Для этого были отобраны две объясняющие переменные:

x_1 – количество осадков в период вегетации (мм);

x_2 – средняя температура воздуха.

Матрица парных коэффициентов имеет следующий вид:

	y	x_1	x_2
y	1		
x_1	0,6	1	
x_2	−0,5	−0,9	1

Решение.

Линейное уравнение множественной регрессии y от x_1 и x_2 имеет вид:

$$y = a + b_1 \cdot x_1 + b_2 \cdot x_2.$$

Для расчета его параметров, применим метод стандартизации переменных, и построим искомое уравнение в стандартизованном масштабе: $t_y = \beta_1 \cdot t_{x1} + \beta_2 \cdot t_{x2}.$

По условию $r_{yx1} = 0,6, \quad r_{yx2} = -0,5, \quad r_{x1x2} = -0,9.$

Расчет β - коэффициентов выполним по формулам

$$\beta_1 = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{1 - r_{x1x2}^2} = \frac{0,6 - (-0,5)\cdot(-0,9)}{1 - (-0,9)^2} = \frac{0,15}{0,19} = 0,7895;$$

$$\beta_2 = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{1 - r_{x1x2}^2} = \frac{-0,5 - 0,6\cdot(-0,9)}{1 - (-0,9)^2} = \frac{0,04}{0,19} = 0,2105.$$

Получим уравнение $t_y = 0,7895 t_{x1} + 0,2105 t_{x2}.$

Пример.2.7. Изучается зависимость по 25 предприятиям концерна потребления материалов y (т) от энерговооруженности труда x_1 (кВт.ч на одного рабочего) и объема произведенной продукции x_2 (тыс. ед.). Данные приведены в таблице 2.6.

Таблица 2.6.

Признак	Среднее значение	Среднее квадратическое отклонение	Парный коэффициент корреляции
y	12,0	2,0	$r_{yx1} = 0{,}52$
x_1	4,3	0,5	$r_{yx2} = 0{,}84$
x_2	10,0	1,8	$r_{x1x2} = 0{,}43$

Требуется:

1. Построить уравнение множественной регрессии и пояснить экономический смысл его параметров.
2. Определить частные коэффициенты эластичности.
3. Определить частные и множественный коэффициенты корреляции.
4. Оцените уравнение регрессии с помощью F-критерия Фишера.

Решение.

1. Линейное уравнение множественной регрессии y от x_1 и x_2 имеет вид: $y = a + b_1 \cdot x_1 + b_2 \cdot x_2$.

Для расчета его параметров, применим метод стандартизации переменных, и построим искомое уравнение в стандартизованном масштабе: $t_y = \beta_1 \cdot t_{x1} + \beta_2 \cdot t_{x2}$.

Расчет β - коэффициентов выполним по формулам:

$$\beta_1 = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{1 - r_{x1x2}^2} = \frac{0{,}52 - 0{,}84 \cdot 0{,}43}{1 - 0{,}43^2} = \frac{0{,}52 - 0{,}3612}{1 - 0{,}1849} = \frac{0{,}1588}{0{,}8151} = 0{,}1948;$$

$$\beta_2 = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{1 - r_{x1x2}^2} = \frac{0{,}84 - 0{,}52 \cdot 0{,}43}{1 - 0{,}43^2} = \frac{0{,}84 - 0{,}2236}{1 - 0{,}1849} = \frac{0{,}6164}{0{,}8151} = 0{,}7562.$$

Получим уравнение

$$t_y = 0{,}1948 t_{x1} + 0{,}7562 t_{x2}.$$

Для построения уравнения в естественной форме рассчитаем b_1 и b_2, используя формулы для перехода от β_i к b_i:

$$\beta_i = b_i \frac{\sigma_{xi}}{\sigma_y}; \qquad b_i = \beta_i \frac{\sigma_y}{\sigma_{xi}};$$

$$b_1 = 0{,}1948 \frac{2{,}0}{0.5} = 0{,}7792; \qquad b_2 = 0{,}7562 \frac{2{,}0}{1{,}8} = 0{,}841.$$

Значение a определим из соотношения

$$a = \bar{y} - b_1 \cdot \bar{x}_1 - b_2 \cdot \bar{x}_2 = 12 - 0{,}7792 \cdot 4{,}3 - 0{,}841 \cdot 10 = 0{,}23944,$$

$$y_{x_1 x_2} = 0{,}23944 + 0{,}7792 x_1 + 0{,}841 x_2.$$

2. Для характеристики относительной силы влияния x_1 и x_2 на y рассчитаем средние коэффициенты эластичности:

$$\overline{Э}_{yx_j} = b_j \frac{\bar{x}_j}{\bar{y}}; \overline{Э}_{yx1} = 0{,}7792 \cdot \frac{4{,}3}{12} = 0{,}2792\%, \qquad \overline{Э}_{yx2} = 0{,}841 \cdot \frac{10}{12} = 0{,}70083\%.$$

С увеличением энерговооруженности труда на 1% от его среднего уровня потребление материалов у возрастает на 0,2792% от своего среднего уровня; при повышении объема произведенной продукции на 1% потребление материалов возрастает на 0,70083% от своего среднего уровня. Сила влияния среднего объема произведенной продукции x_1 на средний уровень потребления материалов у оказалась большей, чем сила влияния средней энерговооруженности труда.

3. Линейные коэффициенты частной корреляции рассчитаем по рекуррентной формуле:

$$r_{yx1 \cdot x2} = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{\sqrt{\left(1 - r_{yx2}^2\right)\ \left(1 - r_{x1x2}^2\right)}} = \frac{0{,}52 - 0{,}84 \cdot 0{,}43}{\sqrt{\left(1 - 0{,}84^2\right)\ \left(1 - 0{,}43^2\right)}} = \frac{0{,}1588}{\sqrt{0{,}2399}} = 0{,}3242;$$

$$r_{yx2 \cdot x1} = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x1}}{\sqrt{\left(1 - r_{yx1}^2\right)\ \left(1 - r_{x1x1}^2\right)}} = \frac{0{,}84 - 0{,}52 \cdot 0{,}43}{\sqrt{\left(1 - 0{,}52^2\right)\ \left(1 - 0{,}43^2\right)}} =$$

$$= \frac{0{,}6164}{\sqrt{(0{,}7296)\ \ (0{,}8151)}} = 0{,}7993;$$

$$r_{x1x2} = \frac{r_{x1x1} - r_{yx1} \cdot r_{yx2}}{\sqrt{\left(1 - r_{yx1}^2\right)\ \left(1 - r_{yx2}^2\right)}} = \frac{0{,}43 - 0{,}52 \cdot 0{,}84}{\sqrt{\left(1 - 0{,}52^2\right)\ \left(1 - 0{,}84^2\right)}} = \frac{0{,}0068}{\sqrt{0{,}7296 \cdot 0{,}2944}} =$$

$$= \frac{0{,}0068}{0{,}4635} = 0{,}0147.$$

Рассчитаем линейный коэффициент множественной корреляции по формуле:

$$R_{yx1x2} = \sqrt{r_{yx1} \cdot \beta_1 + r_{yx2} \cdot \beta_2} = \sqrt{0{,}52 \cdot 0{,}1948 + 0{,}84 \cdot 0{,}7562} = \sqrt{0{,}101296 + 0{,}635208} =$$

$$= 0{,}8582$$

Зависимость y от x_1 и x_2 характеризуется как тесная, в которой 74% вариации ($R^2 = 0{,}74$) среднего потребления материалов определяются вариацией учтенных в модели факторов. Прочие факторы, не включенные в модель, составляют соответственно 36% от общей вариации у.

4. Общий F-критерий проверяет гипотезу H_o о статистической значимости уравнения регрессии и показателя тесноты связи (R^2=0):

$$F_{факт} = \frac{R_{yx1x2}^2}{1 - R_{yx1x2}^2} : \frac{m}{n - m - 1} = \frac{R_{yx1x2}^2}{1 - R_{yx1x2}^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}7365}{0{,}2635} \cdot \frac{25 - 2 - 1}{2} = 30{,}75;$$

$F_{табл.} = 3{,}44; \quad \alpha = 0{,}5;$

R^2 - коэффициент множественной детерминации; m – число параметров при переменных x; n – число наблюдений.

Сравнивая $F_{\text{табл.}}$ и $F_{\text{факт.}}$ приходим к выводу о необходимости отклонить гипотезу H_0, так как $F_{\text{табл.}}=3,44<F_{\text{факт}}=30,75$. С вероятностью 1-α=1-0,05=0,95 делаем заключение о статистической значимости уравнения в целом и показателя тесноты связи R_{yx1x1}.

Частные *F*-критерии – F_{x1} и F_{x2} оценивают статистическую значимость присутствия факторов x_1 и x_2 в уравнении множественной регрессии. Критерий F_{x1} оценивает целесообразность включения в уравнение фактора x_1 после того, как в него был включен фактор x_2. Соответственно F_{x2} указывает на целесообразность включения в модель фактора x_2. после фактора x_1.

$$F_{x1факт} = \frac{R^2_{yx1x2} - r^2_{yx2}}{1 - R^2_{yx1x2}} \cdot \frac{n-m-1}{1} = \frac{0,8582^2 - 0,84^2}{1 - 0,8582^2} \cdot \frac{25-2-1}{1} = 2,58;$$

$$F_{табл} = 4,3; \qquad \alpha = 0,05.$$

Сравнивая $F_{\text{табл.}}$ и $F_{\text{факт}}$ приходим к выводу о нецелесообразности включения в модель фактора x_1 после фактора x_2, так как $F_{x1факт} = 2,58 < F_{табл} = 4,3$.

Целесообразность включения в модель фактора x_2 после фактора x_1 проверяет F_{x2}:

$$F_{x2факт} = \frac{R^2_{yx1x2} - r^2_{yx1}}{1 - R^2_{yx1x2}} \cdot \frac{n-m-1}{1} = \frac{0,8582^2 - 0,52^2}{1 - 0,8582^2} \cdot \frac{25-2-1}{1} = \frac{0,4661}{0,2635} \cdot 22 = 38,9.$$

Сравнивая $F_{\text{табл.}}$ и $F_{\text{факт}}$ приходим к выводу о целесообразности включения в модель фактора x_2 после фактора x_1, так как $F_{x2факт} = 38,9 > F_{табл} = 4,3$.

2.9. Нелинейная регрессия

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью нелинейных функций. Нелинейными моделями, как правило, оказываются производственные функции (зависимости между объемом производственной продукции и основными факторами производства – трудом, капиталом и т.п.), функции спроса (зависимость между спросом на товары или услуги и их ценами или доходом) и другие. Отдельные производственные функции получили известность благодаря разработке и применения их в каких-то специальных целях (для планирования, прогнозирования). Но более удачные из них использовались и в других целях. Изначальное же название таких производственных функций (по направлению использования, по автору) сохранилось. Таким образом, сформировалась целая группа производственных функций, называемых специальными.

Ниже приведены основные специальные функции, которые находят применение в анализе, прогнозировании, планировании и в экономических исследованиях.

Широкое распространение получила, например, производственная функция Кобба – Дугласа:

$$Y = a_0 K^{a_1} L^{1-a_1} . \quad (2.29)$$

В настоящее время известно несколько ее модификаций. В частности, применяют кинематическую производственную функцию:

$$Y = a_0 K^{a_1} L^{a_2} e^{a_3 t}, \quad (2.30)$$

где Y – валовой продукт, K – производственные фонды, L – трудовые ресурсы, e – основание натуральных логарифмов, t - временной фактор.

Большой известностью характеризуется функция Гомперца. Используется несколько ее модификаций. Наиболее простой из них является следующая:

$$y = a_1^{a_2 t} \quad (2.31)$$

Иногда зависимость (2.31) представляют в несколько другом виде:

$$lgy = a_2 t \cdot lga_1. \quad (2.32)$$

В научных исследованиях широкое применение имеет логистическая производственная функция (S-образная кривая, или кривая П. Рида). Она имеет следующий вид:

$$y = \frac{p}{1+a_1 e^{-a_2 t}}. \quad (2.33)$$

Используется также квадратичная логистическая функция

$$y = \frac{p}{(1+a_1 e^{-a_2 t})^2}. \quad (2.34)$$

Особого внимания заслуживают производственные функции, предложенные шведским экономистом Торнквистом. Первая функция моделирует зависимость потребления продуктов питания (y) от суммы получаемого дохода (x):

$$y = \frac{a_1 x}{a_2+x}. \quad (2.35)$$

Но в прогнозировании и планировании применяется следующая модификация этой зависимости:

$$y = \frac{pt}{a+t} \quad (2.36)$$

Вторая функция Торнквиста предназначена для изучения влияния суммы дохода на приобретение предметов относительной необходимости (одежды, жилья, мебели и т.д.). Она имеет вид:

$$y = \frac{(a_1(a_2+x))}{a_3+x} \quad (2.37)$$

В прогнозировании используется иной вариант второй функции Торнквиста:

$$y = \frac{p(a_1+t)}{a_2+t}. \quad (2.38)$$

Зависимость приобретения предметов роскоши населением от суммы получаемого им дохода выражается третьей функцией Торнквиста:

$$y = \frac{a_1 x(a_2+x)}{a_3+x}. \quad (2.39)$$

Эта зависимость в экономических исследованиях находит применение в виде функции насыщения:

$$y = \frac{pt(a_1+t)}{a_2+t}. \quad (2.40)$$

Во всех приведенных выше производственных функциях параметр р предварительно задается как предел исследуемой зависимости или рассчитывается на основе имеющейся информации в процессе построения конкретной производственной модели.

В демографических и некоторых других прогнозах приемлемой является следующая комбинированная функция:

$$y = a_o x^{a_1+a_2 t}. \quad (2.41)$$

Кривыми насыщения называются показательная, логарифмическая и экспоненциальная функции, т. к. будущий прирост результативной переменной зависит от уже достигнутого уровня функции.

Кривые насыщения применяются для характеристики явлений и процессов, величина роста которых является ограниченной величиной (например, в демографии).

Определение. S-образными кривыми называются кривая Гомперца и кривая Перла-Рида. Данные кривые представляют собой кривые насыщения с точкой перегиба.

S-образные кривые применяются для характеристики явлений, включающий в себя два последовательных процесса – ускорения и замедления достигнутого уровня развития. Подобные явления характерны для демографии, страхования и других областей.

Модели регрессии, нелинейные по оцениваемым коэффициентам, делятся на два класса:

1) модели регрессии, которые можно с помощью преобразований привести к линейному виду;

2) модели регрессии, которые невозможно привести к линейному виду.

Рассмотрим первый класс моделей регрессии.

Показательная функция вида

$$y_i = \beta_0 \beta_1^{x_i} \varepsilon_i \quad (2.42)$$

является нелинейной по коэффициенту β_l и относится к классу моделей регрессии, которые можно с помощью преобразований привести к линейному виду. Данная модель характеризуется тем, что случайная ошибка ε_i мультипликативно связана с факторной переменной x_i.

Данную модель можно привести к линейному виду с помощью логарифмирования:

$$\log y_i = \log \beta_0 + x_i \cdot \log \beta_1 + \log \varepsilon_i. \quad (2.43)$$

Для более наглядного представления данной модели регрессии воспользуемся методом замен:

$$\log y_i = Y_i, \ \log \beta_0 = A, \log \beta_1 = B, \ \log \varepsilon_i = E.$$

В результате произведённых замен получим окончательный вид показательной функции, приведённой к линейной форме:

$$Y_i = A + Bx_i. \quad (2.44)$$

Таким образом, можно сделать вывод, что рассмотренная показательная функция является внутренне линейной, поэтому оценки неизвестных параметров её линеаризованной формы можно рассчитать с помощью классического метода наименьших квадратов.

Другим примером моделей регрессии первого класса является степенная функция вида:

$$y_i = \beta_0 x_i^{\beta_1} \varepsilon_i. \quad (2.45)$$

Данная модель характеризуется тем, что случайная ошибка βi мультипликативно связана с факторной переменной xi.

Данную модель можно привести к линейному виду с помощью логарифмирования

$$\ln y_i = \ln \beta_0 + \beta_1 \ln x_i + \ln \varepsilon_i. \quad (2.46)$$

Для более наглядного представления данной модели регрессии воспользуемся методом замен:

$$\ln y_i = Y_i, \ \ln \beta_0 = A, \ \ln x_i = X_i, \ \ln \varepsilon_i = E.$$

В результате произведённых замен получим окончательный вид показательной функции, приведённой к линейной форме:

$$Y_i = A + \beta_1 X_i + E. \quad (2.47)$$

Таким образом, можно сделать вывод, что рассмотренная степенная функция является внутренне линейной, поэтому оценки неизвестных параметров её линеаризованной формы можно рассчитать с помощью классического метода наименьших квадратов.

Рассмотрим второй класс моделей регрессии, нелинейных по оцениваемым коэффициентам.

Показательная функция вида $y_i = \beta_0 \cdot \beta_1^{x_i} + \varepsilon_i$ относится к классу моделей регрессии, которые невозможно привести к линейной форме

путём логарифмирования. Данная модель характеризуется тем, что случайная ошибка βi аддитивно связана с факторной переменной xi.

Степенная функция вида $y_i = \beta_0 x_i^{\beta_1} \varepsilon_i$ относится к классу моделей регрессии, которые невозможно привести к линейной форме путём логарифмирования. Данная модель характеризуется тем, что случайная ошибка ε_i аддитивно связана с факторной переменной x_i.

Таким образом, для оценки неизвестных параметров моделей регрессии, которые нельзя привести к линейному виду, нельзя применять классический метод наименьших квадратов. В этом случае используются итеративные процедуры оценивания (квази-ньютоновский метод, симплекс-метод, метод *Хука-Дживса*, метод *Розенброка* и др.).

Индексом корреляции для нелинейных форм связи называется коэффициент корреляции, который вычисляется для оценки качества построенной нелинейной модели регрессии.

Индекс корреляции для нелинейных форм вычисляется с помощью теоремы о разложении дисперсий по формуле:

$$R = \sqrt{1 - \frac{\sigma_{ост}^2}{\sigma_y^2}}, \qquad (2.48)$$

где σ_y^2– это общая дисперсия зависимой переменной;

$\sigma_{ост}^2$ −остаточная дисперсия, определяемая из уравнения регрессии.

Индекс корреляции можно выразить как

$$R = \sqrt{1 - \frac{\sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}}. \qquad (2.49)$$

Величина данного показателя находится в границах: $0 \leq R \leq 1$.

С помощью корреляции нельзя охарактеризовать направление связи между результативной и факторными переменными. Чем ближе значение индекса корреляции для нелинейных форм связи к единице, тем сильнее взаимосвязь между результативной и независимыми переменными, и наоборот, чем ближе значение индекса корреляции для нелинейных форм связи к нулю, тем слабее взаимосвязь между результативной и независимыми переменными.

Индексом детерминации называется квадрат индекса корреляции для нелинейных форм связи.

Индекс детерминации характеризует, на сколько процентов построенная модель регрессии объясняет вариацию значений результативной переменной относительно своего среднего уровня, т. е. показывает долю общей дисперсии результативной переменной, объяснённой вариацией факторных переменных, включённых в модель регрессии.

Коэффициент множественной детерминации также называется количественной характеристикой объяснённой построенной моделью регрессии дисперсии результативной переменной. Чем больше значение коэффициента множественной детерминации, тем лучше построенная модель регрессии характеризует взаимосвязь между переменными.

На нелинейные модели регрессии, которые являются внутренне линейными, т. е. сводимыми к линейному виду, распространяются все методы проверки гипотез, используемые для классических линейных моделей регрессии.

Таким образом, если внутренне линейную модель регрессии можно свести к линейной модели парной регрессии, то на эту модель будут распространяться все методы проверки гипотез, используемые для парной линейной зависимости.

2.10. Гетероскедастичность

Для применения МНК требуется, чтобы дисперсия остатков была *гомоскедастичной.* Это значит, что для каждого значения фактора x_j остатки ε_i имеют одинаковую дисперсию. Если это условие не выполняется, то имеет место *гетероскедастичность*. При нарушении гомоскедастичности мы имеем неравенства $\sigma_{\varepsilon i}^2 \neq \sigma_{\varepsilon j}^2 \neq \sigma^2, \quad j \neq i.$

При малом объеме выборки для оценки гетероскедастичности используют метод Гольдфельда-Квандта. Тест Гольдфельда-Квандта состоит в следующем:

1. упорядочение n наблюдений по мере возрастания переменной x;
2. исключение из рассмотрения C центральных наблюдений; при этом $(n\text{-}C):2>p$, где p-число оцениваемых параметров.
3. разделение совокупности из $(n\text{-}C)$ наблюдений на две группы (соответственно с малыми и большими значениями фактора x) и определение по каждой из групп уравнений регрессии;
4. определение остаточной суммы квадратов для первой (S_1) и второй (S_2) групп и нахождение их отношения: $R=S_2:S_1$.

При выполнении нулевой гипотезы о гомоскедастичноти отношение R будет удовлетворять F-критерию с $(n\text{-}C\text{-}2p):2$ степенями свободы для каждой остаточной суммы квадратов.

Если $R = \frac{S_2}{S_1} \succ F_{кр} = F_{\alpha, v1, v2}$, то гипотеза об отсутствии гетероскедастичности отклоняется (α – выбранный уровень значимости, $v_1=v_2=(n\text{-}C\text{-}2p):2$ степенями свободы).

Для применения теста Гольдфельда-Квандта, необходимо определить число исключаемых центральных наблюдений *C*. Из экспериментальных расчетов, рекомендовано при *n*=30 принимать *C*=8, при *n*=60 - соответственно 16.

Пример 2.8. Поступление доходов в консолидированный бюджет Санкт-Петербурга (*y* - млрд руб.) в зависимости от численности работающих на крупных и средних предприятиях (*x* – тыс.чел.) и экономики районов за 20**г. (табл. 2.7.)

Таблица 2.7.

№ п/п	Район города	x_i	y_i	$\hat{y}_i$	ε_i
1	Павловский	3	4,4	-1,0	5,4
2	Кронштадт	6	8,1	2,5	5,6
3	Ломоносовский	8	12,9	4,9	8,0
4	Курортный	18	20,8	16,6	4,2
5	Петродворец	20	15,5	19,0	-3,5
6	Пушкинский	23	28,8	22,5	6,3
7	Красноссельский	39	37,5	41,4	-3,9
8	Приморский	49	48,7	53,2	-4,5
9	Колпинский	60	68,6	66,1	2,5
10	Фрунзенский	74	104,6	82,6	22,0
11	Крсногвардейский	79	90,5	88,5	2,0
12	Василеостровский	95	88,3	107,4	-19,1
13	Невский	106	132,4	120,4	12,0
14	Петроградский	112	122,0	127,4	-5,4
15	Калининский	115	99,1	131,0	-31,9
16	Выборгский	125	114,2	142,7	-28,5
17	Кировский	132	150,6	151,0	-0,4
18	Московский	149	156,1	171,0	-14,9
19	Адмилартейский	157	209,5	180,5	29,0
20	Центральный	282	342,9	327,8	15,1
	Итого:	1652	1855,5	1855,5	0

В соответствии с уравнением

$$\hat{y}_x = -4{,}565 + 1{,}178 \cdot x; r = 0{,}9828, F = 510{,}7$$

Найдены теоретические значения $\hat{y}_x$ и отклонения от них фактических значений *y*, т.е. ε_i. Итак, остаточные величины ε_i обнаруживают тенденцию к росту по мере увеличения *x* и *y* (рис.2.2.)

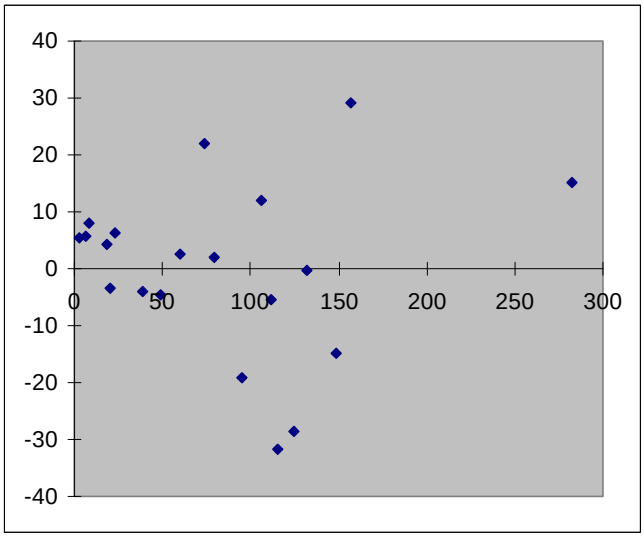


Рис.2.2. График остатков

Этот вывод подтверждается и по критерию Гольдфельда – Квандта. Для его применения сначала необходимо определить число исключаемых центральных наблюдений *C*. Из экспериментальных расчетов, проведенных авторами метода для случая одного фактора, рекомендовано при *n*=30 принимать *C*=8, а при *n*=60- соответственно *C*=16. В рассматриваемом примере при *n*=20 было обратно *C*=4. Тогда в каждой группе будет по 8 наблюдений [(20-4):2]. Результаты расчетов представлены в таблице 2.8.

Проверка линейной регрессии на гетероскедастичность.

Величина *R*=38,6 (2638,4:68,34=38,6), что превышает табличное значение *F*-критерия=4,28 при 5 %-ном и 8,47 при 1 %-ном уровне значимости для числа степеней свободы 6 для каждой остаточной суммы квадратов [(20-4-2·2):2], подтверждая тем самым наличие гетеросскедостичности.

Критерий Гольдфельда – Квандта используется и при проверке остатков множественной регрессии на гетерскедостичность

Табл.2.8.

	x	y	$\hat{y}_i$	ε_i	ε^2
Первая группа с первыми 8 районами: $\hat{y}_x = -2{,}978 + 0{,}921 \cdot x;$ $r = 0{,}979,$ $F = 136{,}4$	3	4,4	5,7	-1,3	1,69
	6	8,1	8,5	-0,4	0,16
	8	12,9	10,3	2,6	6,76
	18	20,8	19,6	1,2	1,44
	20	15,5	21,4	-5,9	34,81
	23	28,8	24,2	4,6	21,16
	39	37,5	38,9	-1,4	1,96
	49	48,7	48,1	0,6	0,36
Сумма					68,34

Вторая группа с последними 8 районами: $\hat{y}_x = 31{,}142 + 1{,}338 \cdot x;$ $r = 0{,}969,$ $F = 93{,}4$	106	132,4	110,7	21,7	470,89
	112	122,0	118,7	3,3	10,89
	115	99,1	122,7	-23,6	556,96
	125	114,2	136,1	-21,9	479,61
	132	150,6	145,4	5,2	27,04
	149	156,1	168,2	-12,1	146,41
	157	209,5	178,9	30,6	936,36
	282	342,9	346,1	-3,2	10,224
Сумма					2638,40

2.11. Мультиколлинеарность

При построении моделей множественной линейной регрессии по МНК возникает проблема мультиколлинеарности – линейной взаимосвязи двух или нескольких объясняющих переменных.

Считается, что две переменных явно коллинеарны, т.е. находятся между собой в линейной зависимости, если $r_{xixj} \geq 0{,}7$. Если факторы явно коллинеарны, то они дублируют друг друга и один из них следует исключить из регрессии. Для этого проводится анализ матрицы парных коэффициентов корреляции, что позволяет произвести отбор факторов, включаемых в модель множественной зависимости. Матрица имеет следующий вид:

Признак	y_o	X_1	X_2	…	X_k
y_o	1	r_{01}	r_{02}	…	r_{ok}
X_1	r_{01}	1	r_{21}	…	r_{k1}
X_2	r_{02}	r_{12}	1	…	r_{k2}
…	…	…	…	1…	…
X_k	r_{ok}	r_{1k}	r_{2k}		1

Анализ первой строки матрицы позволяет выявить факторы, у которых степень тесноты связи с результативным показателем значительна, а поэтому они могут быть включены в модель. В качестве критерия мультиколлинеарности может быть принято соблюдение следующих неравенств:

$$r_{xiy} > r_{xjxk}\ ; \quad r_{xky} > r_{xjxk}\ . \qquad (2.50)$$

Если приведенные неравенства (или хотя бы одно из них) не выполняются, то исключается тот фактор x_j или x_k, связь которого с

результативным признаком y будет менее тесной. По величине парных коэффициентов обнаруживается лишь явная коллинеарность факторов.

Для оценки мультиколлинеарности факторов можно использовать определитель парных коэффициентов корреляции между факторами.

Так, для включающего три объясняющих переменных уравнения $y = a + b_1 x_1 + b_2 x_2 + b_3 x_3 + \varepsilon$ матрица коэффициентов корреляции между факторами имеет определитель $Det|R| = \begin{vmatrix} r_{x1x1} & r_{x2x1} & r_{x3x1} \\ r_{x1x2} & r_{x2x2} & r_{x3x2} \\ r_{x1x3} & r_{x2x3} & r_{x3x3} \end{vmatrix}$.

Чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И наоборот, чем ближе к 1 определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

Проверка мультиколлинеарности факторов можно провести методом испытания гипотезы о независимости переменных $H_o : Det|R| = 1$. Доказано, что величина $\left[n - 1 - \frac{1}{6}(2m+5)\lg DetR \right]$ имеет приближенное распределение χ^2 с $\left(\frac{1}{2} \cdot n \cdot (n-1) \right)$ степенями свободы. Если фактическое значение χ^2 превосходит табличное (критическое) $\chi^2_{\text{факт.}} > \chi^2_{\text{табл.(df,α)}}$, то гипотеза H_0 отклоняется. Мультиколлинеарность считается доказанной.

Пример 2.9. Для линейного трехфакторного уравнения регрессии $y=a_0+a_1x_1+a_2x_2+a_3x_3+\varepsilon$ имеются следующие данные (табл. 2.9):

Таблица 2.9.

y	x_1	x_2	x_3
40	10,3	20,8	4,1
80	14,6	28	20,3
55	11,4	23	9,8
58	17,1	30,5	8,1
70	10,6	21,7	17,7

Требуется:

1. Определить корреляционную матрицу ***R*** и содержащийся в этих данных размер коллинеарности как ***det(R)***.
2. Рассчитать размер коллинеарности, в случае если из уравнения выводится переменная x_2.

Решение

1. Матрицу коэффициентов корреляции переменных можно рассчитать, используя инструмент анализа данных **Корреляция**. Для этого:

1) в главном меню последовательно выберите пункты **Сервис / Анализ данных / Корреляция**;

2) заполните диалоговое окно ввода данных и параметров вывода;

3) результаты вычислений – матрица коэффициентов парной корреляции (R_1)

4)

	*x*1	*x*2	*x*3
*x*1	1		
*x*2	0,99428	1	
*x*3	0,1058	0,193699	1

Определитель этой матрицы $\det(R_1)$=0,0034465. Поскольку $\det(R_1)$ близок к нулю, можно сделать вывод о наличии высокой коллинеарности.

Если вывести из уравнения переменную x_2, то получим следующую корреляционную матрицу (R_2):

	*x*1	*x*3
*x*1	1	0,1058
*x*3	0,1058	1

$\det(R_2)$=0,98806.

Коллинеарность значительно уменьшилась.

Методы устранения коллинеарности:

1. Исключение переменной (ых) из модели.
2. Получение дополнительных данных или новой выборки.
3. Изменение спецификации модели.
4. Использование предварительной информации о некоторых параметрах.
5. Преобразование переменных.

Пример 2.10. По 20 предприятиям региона изучается зависимость выработки продукции на одного работника *y* (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%). Данные приведены в таблице 2.10.

Таблица 2.10

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7,1	3,8	10	11	9,0	6,0	21,0
2	7,3	3,9	13	12	10,0	6,2	22,0
3	7,2	3,8	16	13	9,3	6,9	22,1
4	7,6	4,0	17	14	11,0	7,5	25,0
5	7,5	3,9	19	15	12,0	8,0	28,2
6	8,0	4,5	18	16	12,3	8,3	29,0
7	8,1	5,6	20	17	12,6	8,5	30,5
8	8,3	4,8	21	18	12,8	8,6	31,0
9	8,6	5,9	22	19	14,1	9,5	33,0
10	10,0	6,1	21	20	14,5	9,0	36,0

Требуется:

1. Оценить показатели вариации каждого признака и сделать вывод о возможности применения МНК для их изучения.
2. Проанализировать линейные коэффициенты парной корреляции.
3. Написать уравнение множественной регрессии, оценить значимость его параметров, пояснить экономический смысл.
4. С помощью F- критерия Фишера оценить статистическую надежность уравнения регрессии и R^2_{yx1x2}. Сравнить значимость скорректированного и нескорректированного линейных коэффициентов множественной детерминации.

Решение

1. Сводную таблицу основных статистических характеристик для одного или нескольких массивов данных можно получить с помощью инструмента анализа данных **Описательная статистика**. Для этого выполните следующие шаги:

1) введите исходные данные или откройте существующий файл, содержащий анализируемые данные;

2) в главном меню выберите последовательно пункты **Сервис / Анализ данных / Описательная статистика**, после чего щелкните по кнопке **ОК**;

3) заполните диалоговое окно ввода данных и параметров вывода.

Входной интервал – диапазон, содержащий анализируемые данные, это может быть одна или несколько строк (столбцов);

Группирование – по столбцам или по строкам – необходимо указать дополнительно;

Метки – флажок, который указывает, содержит ли первая строка названия строк или столбцов;

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона;

Получим следующие результаты:

y		*x*1		*x*2	
Среднее	9,865	Среднее	6,24	Среднее	22,7
Стандартная ошибка	0,541	Стандартная ошибка	0,43	Стандартная ошибка	1,52
Медиана	9,15	Медиана	6,05	Медиана	21,5
Мода	10	Мода	3,8	Мода	21
Стандартное отклонение	2,418	Стандартное отклонение	1,94	Стандартное отклонение	6,78
Дисперсия выборки	5,847	Дисперсия выборки	3,78	Дисперсия выборки	46
Эксцесс	-0,98	Эксцесс	-1,38	Эксцесс	-0,4
Асимметрич-ность	0,606	Асимметрич-ность	0,18	Асимметрич-ность	0,22
Интервал	7,4	Интервал	5,7	Интервал	26
Минимум	7,1	Минимум	3,8	Минимум	10
Максимум	14,5	Максимум	9,5	Максимум	36
Сумма	197,3	Сумма	125	Сумма	455

Сравнивая значения средних квадратических отклонений и средних величин и определяя коэффициенты вариации:

$$v_y = \frac{\sigma_y}{\bar{y}} \cdot 100\% = \frac{2{,}418}{9{,}865} \cdot 100\% = 24{,}5\%;$$

$$v_{x1} = \frac{\sigma_{x1}}{\bar{x}_1} \cdot 100\% = \frac{1{,}94}{6{,}24} \cdot 100\% = 31{,}1\%;$$

$$v_{x2} = \frac{\sigma_{x2}}{\bar{x}_2} \cdot 100\% = \frac{6{,}78}{22{,}7} \cdot 100\% = 29{,}9\%,$$

приходим к выводу о повышенном варьирования признаков, хотя и в допустимых пределах, не превышающих 35%. Совокупность предприятий однородна, и для ее изучения могут использоваться метод наименьших квадратов и вероятностные методы оценки статистических гипотез.

2. Значения линейных коэффициентов парной корреляции определяют тесноту попарно связанных переменных, использованных в данном уравнении множественной регрессии. Линейные коэффициенты частной корреляции оценивают тесноту связи значений двух переменных, исключая влияние всех других переменных, представленных в уравнении множественной регрессии. Используя, инструмент анализа данных **Корреляция,** получим следующую матрицу коэффициентов парной корреляции:

	y	*x*1	*x*2
y	1		
*x*1	0,969643	1	
*x*2	0,958858	0,947323	1

Значения коэффициентов парной корреляции указывают на тесную связь выработки y как с коэффициентом обновления основных фондов x_1, так и с долей рабочих высокой квалификации x_2 ($r_{yx1} = 0{,}969643$ *и* $r_{yx2} = 0{,}95885$).

Межфакторная связь тесная $r_{x1x2} = 0{,}947323 > 0{,}7$. Для улучшения модели можно исключить фактор x_2 как малоинформативный, недостаточно статистически надежный.

3. Вычислим параметры линейного уравнения множественной регрессии. Для этого используем инструмент анализа данных **Регрессия**. Заполните диалоговое окно ввода данных и параметров вывода:

Входной интервал Y – диапазон, содержащие данные результативного признака;

Входной интервал X – диапазон, содержащий данные факторов независимого признака (следует указывать все столбцы, содержащие значения факторных признаков);

Метки – флажок, который указывает, содержит ли первая строка названия строк или столбцов;

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона.

Результаты анализа следующие:

ВЫВОД ИТОГОВ

Регрессионная статистика

Множественный R	0,97777
R-квадрат	0,956035
Нормированный R-квадрат	0,950862
Стандартная ошибка	0,535993
Наблюдения	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*
Регрессия	2	106,2016	53,1008	184,8342
Остаток	17	4,883909	0,287289	
Итого	19	111,0855		

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*
Y-пересечение	2,044462	0,429579	4,759218	0,000182
*x*1	0,742896	0,197406	3,763292	0,00155
*x*2	0,140056	0,056615	2,473805	0,02421

По результатам вычислений составим уравнение множественной регрессии вида $\hat{y} = 2{,}044462 + 0{,}742896 \cdot x_1 + 0{,}140056 \cdot x_2$.

Значения случайных ошибок параметров b_0, b_1 *и* b_2 с учетом округления: $m_{b_0} = 0{,}4296$; $m_{b_1} = 0{,}1974$; $m_{b_2} = 0{,}0566$.

Они показывают, какое значение данной характеристики сформировалось под влиянием случайных факторов. Эти значения используются для расчета t-критерия Стьюдента: $t_{b_0} = 4{,}76$; $t_{b_1} = 3{,}76$; $t_{b_2} = 2{,}47$.

Если значения t-критерия больше 2-3, можно сделать вывод о существенности данного параметра, который формируется под воздействием неслучайных причин. Здесь статистически значимы все параметры.

На это же указывает показатель вероятности случайных значений параметров регрессии: если α меньше принятого нами уровня (обычно 0,1; 0,05или 0,01; это соответствует 10%; 5% или 1% вероятности), делают вывод о неслучайной природе данного значения параметра, т.е. он статистически значим и надежен. В данном примере:

$$\alpha_{b_0} = 0\% < 5\%; \alpha_{x1} = 0{,}1\% < 5\%; \alpha_{x2} = 2{,}4\% < 5\%.$$

4. Оценку надежности уравнения регрессии в целом и показателя тесноты связи R_{yx1x2} дает *F*-критерий Фишера:

$$F_{факт} = \frac{\sum(\hat{y}_{x1x2} - \bar{y})^2}{m} : \frac{\sum(y - \hat{y}_{x1x2})^2}{n - m - 1}.$$

По данным таблиц дисперсионного анализа $F_{факт}$=184,83>$F_{табл}$=3,59, т.е. подтверждается статистическая значимость всего уравнения и показателя тесноты связи R^2_{yx1x2}.

Значения скорректированного и нескорректированного линейных коэффициентов множественной детерминации приведены в рамках регрессионной статистике.

Нескорректированный коэффициент множественной детерминации R^2_{yx1x2}=0,956 оценивает долю вариации результата за счет представленных в уравнении факторов и общей вариации результата. Эта доля составляет 95,6%, и указывает на весьма тесную связь факторов с результатом.

Скорректированный коэффициент множественной детерминации $\hat{R}^2_{yx1x2} = 0{,}95$ определяет тесноту связи с учетом степеней свободы общей и остаточной дисперсий. Он дает оценку тесноты связи, которая не зависит от числа факторов в модели. Поэтому его можно сравнивать по разным моделям с разным числом факторов.

Оба коэффициента указывают на весьма высокую (более 90%) детерминированность результата *y* в модели факторами x_1 и x_2.

2.12. Фиктивные переменные в регрессионных моделях

Уравнения множественной регрессии могут включать в качестве независимых переменных качественные признаки (например, профессия, пол, образование, отдельные регионы и т.д.). Чтобы ввести такие переменные в модель, их необходимо упорядочить, и присвоить им цифровые метки. Такие переменные в эконометрике называют фиктивными или структурные переменные.

Например, включать в модель фактор D в виде фиктивной переменной можно в следующем виде:

$$D = \begin{cases} 0, & \textit{фактор не действует}, \\ 1, & \textit{фактор действует}. \end{cases} \qquad (2.51)$$

Коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение зависимой переменной при переходе от одной категории к другой при неизменных значениях остальных параметров.

2.13. Типологическая регрессия

Типологические группировки, являясь одним из важнейших и самостоятельных приемов изучения социально-экономических явлений и процессов, составляют неотъемлемую часть большинства эконометрических исследований в различных отраслях знания.

Регрессионные модели корректно строить только по однородным группам данных, так как для объектов из неоднородных групп могут существовать различные закономерности изменения зависимой переменной y при изменении регрессора.

Пусть каждое i- е наблюдение (фирма) характеризуется набором k-признаков: прибылью, объемом выпущенной продукции, численностью работников, себестоимостью продукции и т.д. Абсолютная величина перечисленных показателей часто связана не с эффективностью работы фирмы, а с ее масштабностью, так как крупное предприятие имеет больше ресурсов. Прежде чем переходить к построению регрессионных моделей, необходимо выделить из n рассматриваемых фирм однородные по комплексу признаки (кластеры), используя типологические группировки (например, малые, средние, крупные предприятия и предприятия гиганты).

Регрессия типологическая – вид статистической модели, при построении которой используется сочетание методов классификации многомерных наблюдений и множественной регрессии. Последовательное

применение этих методов обеспечивает выделение однородных классов объектов и построение в каждом из выделенных кластеров регрессионных зависимостей.

Классификация объектов может проводиться различными способами. Один из простейших методов – группировка по одному или нескольким признакам.

Рассмотрим следующую задачу. Пусть исследуется совокупность n объектов, каждый из которых характеризуется по k замеренным на нем признакам X. Требуется разбить эту совокупность на однородные в некотором смысле группы. При этом почти отсутствует априорная информация о характере распределения измерений X внутри классов. Полученные в результате разбиения группы называют кластерами (таксонами, образами). Методы их нахождения – кластер-анализом (численной таксономией, или распознаванием образов с самообучением). В пространстве переменных кластеры представляют собой скопления точек различной формы. Решение задачи классификации заключается в определении естественного расслоения исходных наблюдений на четко выраженные кластеры, лежащие друг от друга на некотором расстоянии. Обычной формой представления исходных данных в задачах кластерного анализа служит прямоугольная таблица, каждая строка которой представляет результат измерений k рассматриваемых признаков на одном из обследованных объектов:

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1k} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ik} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nk} \end{pmatrix}.$$

Алгоритмы кластерного анализа исходят из матрицы расстояний (или близостей), поэтому первым этапом решения задачи поиска кластеров является выбор способа вычисления расстояний, или близости, между объектами или признаками.

Виды расстояний между объектами

На практике используются следующие виды расстояний между объектами:

Расстояние Махаланобиса

Если компоненты $\left(x^{(1)}, x^{(2)}, \dots, x^{(p)}\right)$ вектора наблюдений X зависимы и их значимость в решении вопроса о классификации объекта различна, то пользуются расстоянием типа $d_0\left(X_i, X_j\right) = \sqrt{\left(X_i - X_j\right)^T \Delta^T \Sigma^{-1} \left(X_i - X_j\right)}$, где Σ - ковариационная матрица генеральной совокупности, из которой

извлекаются наблюдения X_i; Δ – некоторая симметричная неотрицательно определенная матрица весовых коэффициентов, которая чаще всего выбирается диагональной. Другие расстояния являются частными случаями расстояния Махаланобиса.

Обычное евклидово расстояние $d_E = (X_i, X_j) = \sqrt{\sum_{l=1}^{p} \left(x_i^{(l)} - x_j^{(l)}\right)^2}$. Оно используется, если наблюдения извлекаются из генеральной совокупности, которая описывается многомерным нормальным законом распределения, причем X должны быть взаимно независимыми и иметь одинаковую дисперсию. Компоненты этих наблюдений $\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}\right)$ должны быть однородными, т.е. одинаково важными для классификации.

Взвешенное евклидовое расстояние $d_{BE}(X_i, X_j) = \sqrt{\sum_{l=1}^{p} w_l \left(x_i^{(l)} - x_j^{(l)}\right)^2}$. Компоненты $\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}\right)$ неоднородны и степень их важности задают веса w_l, причем $0 < w_l < 1, \quad l = 1, \ldots, p.$

Определение весов требует дополнительного исследования (экспертные опросы, специальные модели и др.), определение их по исходным данным не дает желаемого эффекта.

Расстояние Минковского $d_M(X_i, X_j) = \left(\sum_{l=1}^{p} \left| x_i^{(l)} - x_j^{(l)} \right|^q \right)^{1/q}$

Частные его случаи:

При q=1 Хеммингово расстояние (манхеттенское или расстояние городских кварталов city-block) $d_H = (X_i, X_j) = \sum_{l=1}^{p} \left| x_i^{(l)} - x_j^{(l)} \right|$, которое используется в случае объектов, характеризуемых дихотомическими признаками.

При q=2 получаем евклидово расстояние.

Решение задач классификации многомерных данных предусматривает в качестве предварительного этапа исследования реализацию методов, позволяющих выбрать из компонент $\left(x^{(1)}, x^{(2)}, \ldots, x^{(p)}\right)$ наблюдаемых векторов X сравнительно небольшое число наиболее информативных, т.е. уменьшить размерность наблюдаемого пространства.

В ряде процедур классификации (кластер-процедур) используют понятия расстояния между группами объектов и меры близости двух групп объектов.

Пусть S_i – i-я группа (класс, кластер),состоящая из n_i объектов;

$\overline{X}_i$ - вектор средних арифметических значений для S_i группы, т.е. «центр тяжести» i-й группы;

$d_{\min}(S_l, S_m)$ - расстояние между группами S_l и S_m.

Наиболее употребительными расстояниями между классами объектов являются:

- расстояние, измеряемое по принципу «ближайшего соседа»:

$$d_{\min}(S_l, S_m) = \min_{X_i \in S_l, X_j \in S_m} d(X_i, X_j);$$

- расстояние, измеряемое по принципу «дальнего соседа»:

$$d_{\max}(S_l, S_m) = \max_{X_i \in S_l, X_j \in S_m} d(X_i, X_j);$$

- расстояние, измеряемое по «центрам тяжести» групп:

$$d_{ЦТ}(S_l, S_m) = d(\overline{X}_l, \overline{X}_m);$$

- расстояние, измеряемое по принципу «средней связи», определяется как среднее арифметическое всех попарных расстояний между представителями рассматриваемых групп:

$$d_{cp}(S_l, S_m) = \frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d(X_i, X_j).$$

Академиком А.Н. Колмогоровым было предложено «обобщенное расстояние» между классами, которое включает в себя в качестве частных случаев все рассмотренные виды расстояний. Расстояние между классами s_l и $s_{(m,q)}$, являющееся объединением двух классов s_m и s_q, можно определить по формуле

$$d_{l,(m,q)} = d(S_l, S_{(m,q)}) = \alpha d_{lm} + \beta d_{lq} + \gamma d_{mq} + \delta(d_{lm} - d_{lq}), \qquad (2.52)$$

где $d_{lm} = d(S_l, S_m)$; $d_{lq} = d(S_l, S_q)$ и $d_{mq} = d(S_m, S_q)$ – расстояние между классами s_l, s_m и s_q; α, β, δ и γ – числовые коэффициенты, значения которых определяют специфику процедуры, ее алгоритм.

Например, при $\alpha = \beta = -\delta = \frac{1}{2}$ и $\gamma = 0$ приходим к расстоянию, построенному по принципу «ближайшего соседа». При $\alpha = \beta = \delta = \frac{1}{2}$ и $\gamma = 0$ - расстояние между классами определяется по принципу «дальнего соседа», т.е. как расстояние между двумя самыми дальними элементами этих классов. И наконец, при $\alpha = \frac{n_m}{n_m + n_q}$, $\beta = \frac{n_q}{n_m + n_q}$, $\gamma = \delta = 0$ соотношение (1) приводит к расстоянию d_{cp} между классами, вычисленному как среднее из расстояний между всеми парами элементов, один из которых берется из одного класса, а другой из другого.

Пример 2.11. По данным представленным в табл. , провести классификацию 5 фирм, каждая из которых характеризуется тремя переменными:

x_1 - среднегодовая величина оборотных средств, млн руб.,

x_2 – материальные затраты на 1 руб. произведенной продукции, коп., x_3 – объем произведенной продукции, млн руб.

Таблица 2.11.

№ п.п.	x_1	x_2	x_3
1	130	90	165
2	80	75	90
3	140	80	100
4	70	76	80
5	60	66	110
среднее	96	77,4	109
среднеквадратическое отклонение («исправленное»)	36,5	8,7	33,2

Решение.

Нормируем исходные данные. Для этого из каждого значения вычитаем среднее и делим на среднее квадратическое отклонение. Получим нормированные переменные:

№ п.п.	z_1	z_2	z_3
1	0,93	1,45	1,69
2	-0,44	-0,28	-0,57
3	1,21	0,30	-0,27
4	-0,71	-0,16	-0,87
5	-0,99	-1,31	0,03

Классификацию проведем при помощи иерархического агломеративного метода. Согласно обычной евклидовой метрике, расстояние между первым и вторым объектами равно:

$$d_{1,2} = \sqrt{\sum_{j=1}^{k}\left(x_{1j} - x_{2j}\right)^2} = \sqrt{(0{,}93-(-0{,}44))^2 + (1{,}45-(-0{,}28))^2 + (1{,}69-(-0{,}57))^2} = 3{,}15,$$

Матрица расстояний имеет вид:

$$D_1 = \begin{array}{c} \\ S_1 \\ S_2 \\ S_3 \\ S_4 \\ S_5 \end{array} \begin{array}{c} \begin{array}{ccccc} S_1 & S_2 & S_3 & S_4 & S_5 \end{array} \\ \begin{pmatrix} 0 & 3{,}15 & 2{,}29 & 3{,}44 & 3{,}75 \\ 3{,}15 & 0 & 1{,}77 & 0{,}42 & 1{,}32 \\ 2{,}29 & 1{,}77 & 0 & 2{,}06 & 2{,}74 \\ 3{,}44 & 0{,}42 & 2{,}06 & 0 & 1{,}49 \\ 3{,}75 & 1{,}32 & 2{,}74 & 1{,}49 & 0 \end{pmatrix} \end{array}$$

Из матрицы расстояний следует, что объекты 2 и 4 наиболее близки $d_{2,4}=0{,}42$ и поэтому объединим их в один кластер. После объединения объектов имеем четыре кластера: $S_1, S_{2,4}, S_3, S_5$.

Воспользуемся принципом «дальнего соседа» для вычисления расстояния между объектами. Расстояние между кластером S_2 и кластером S_4 равно: $d_{S_2,S_4}=\max\{d_{1,2},d_{1,4}\}=\max\{3{,}15;3{,}44\}=3{,}44$ и т.д.

Получим матрицу расстояний $D_2=\begin{pmatrix} 0 & 2{,}29 & 3{,}44 & 3{,}75 \\ & 0 & 2{,}06 & 2{,}74 \\ & & 0 & 1{,}49 \\ & & & 0 \end{pmatrix}$

$S_1 \quad S_3 \quad S_{2,4} \quad S_5$

В матрице D_2 опять находим самые близкие кластеры. Так как $d_{45}=1{,}49$, то объединяем кластеры S_4 и S_5, получим новый кластер S_4, содержащий объекты 2,4 и 5. Пересчитываем расстояния

$$d_{S_{2,4},S_5}=\max\{d_{2,4},d_5\}=\max\{3{,}44;3{,}75\}=3{,}75,$$

$d_{S_3,S_{2,4}}=\max\{d_3,d_{2,4}\}=\max\{2{,}06;2{,}74\}=2{,}74,$ получаем матрицу

$$D_3=\begin{pmatrix} 0 & 2{,}29 & 3{,}75 \\ 2{,}29 & 0 & 2{,}74 \\ 3{,}75 & 2{,}74 & 0 \end{pmatrix}.$$

Объединяем кластеры S_1 и S_3, получим новый кластер S_1, содержащий объекты 1 и 3. Имеем два кластера $S_1\{1.3\}$ и $S_4\{2,4,5\}$:

$d_{S_1S_4}=\max\{d_{1,3},d_{2,4,5}\}=\max\{3{,}75;2{,}74\}=3{,}75$ и матрица D_4 имеет следующий вид: $D_4=\begin{pmatrix} 0 & 3{,}75 \\ 3{,}75 & 0 \end{pmatrix}$.

На последнем шаге объединяются кластеры S_1 и S_4. Результаты классификации представлены графически на рис.7 в виде следующей дендрограммы:

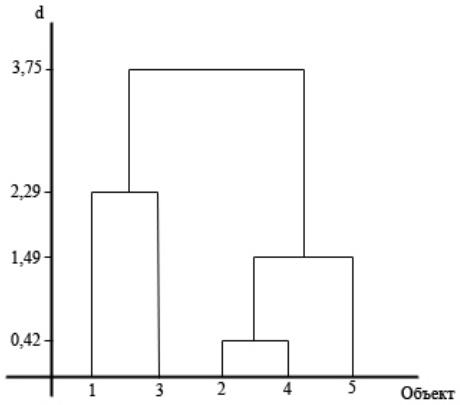


Рис. 2.3 . Дендрограмма кластеризации пяти объектов

Глава 3. Ряды динамики

3.1. Основные элементы временного ряда

Социально-экономические явления общественной жизни находятся в непрерывном развитии. Их изменение во времени статистический анализ изучает при помощи построения и анализа рядов динамики.

Ряд динамики – это числовые значения статистического показателя во временной последовательности. Он состоит из двух граф: в первой указываются периоды, во второй – показатели, характеризующие изучаемый объект за эти периоды.

Показатели второй графы носят название уровней ряда: первый показатель называется начальным уровнем, последний – конечным. Уровни могут быть выражены абсолютными, средними или относительными величинами.

Всякий ряд динамики может быть представлен в следующем виде:

$$y_t = f(t) + s(t) + \varepsilon(t), \qquad (3.1)$$

где $f(t)$ – тренд (основная тенденция развития динамического ряда); $s(t)$ – циклические (периодические) колебания, в том числе и сезонные; $\varepsilon(t)$ – случайные колебания.

Изучение ряда динамики включает два основных этапа:

- ряд динамики проверяется на наличие тренда;
- производится выравнивание временного ряда и непосредственное выделение тренда с экстраполяцией полученных результатов.

В рядах динамики наблюдаются тенденции трех видов:

- тенденция среднего уровня, которую можно представить графиком временного ряда. Аналитически она выражается в виде функции $f(t)$, вокруг которой варьируются фактические значения изучаемого явления;
- тенденция дисперсии – это изменения отклонений эмпирических значений временного ряда от значений, вычисленных по уравнению тренда;
- тенденция автокорреляции – это тенденция изменения связи между отдельными уровнями временного ряда

В статистическом анализе разработан ряд метод выявления перечисленных видов тенденции. На практике широкое распространение получили методы Фостера и Стюарта и сравнения средних уровней ряда динамики.

Метод средних

Ряд динамики разбивается на две равные или почти равные части, каждая из которых рассматривается как некоторая самостоятельная выборочная совокупность, имеющая нормальное распределение. Если временной ряд имеет тенденцию, то средние вычисленные для каждой совокупности, должны существенно (значимо) различаться между собой. Если же расхождение будет незначимым, несущественным (случайным), то временной ряд не имеет тенденции. Таким образом, проверка наличия тренда в исследуемом ряду сводится к проверке гипотезы о равенстве средних двух нормально распределенных совокупностей.

Процедура проверки гипотезы о постоянстве средних значений по двум выборкам ряда определяется предположением относительно дисперсии распределения.

Проверка гипотезы о равенстве дисперсий, реализуется с помощью F-критерия.

H_0: $\sigma_1=\sigma_2$; H_1: $\sigma_1\neq\sigma_2$;

$$F_{расч}=\frac{S_2^2}{S_1^2}, \qquad (3.2)$$

где $S_1^2=\frac{\sum_{i=1}^{n_1}(y_i-\bar{y}_1)^2}{n_1-1}$, $S_2^2=\frac{\sum_{i=1}^{n_{12}}(y_i-\bar{y}_2)^2}{n_2-1}$, $\bar{y}_1,\bar{y}_2$ - средние для первой и второй половины временного ряда, n_1,n_2 - число наблюдений в этих частях ряда.

$F_{крит}=F(\alpha,\ n_1-1,\ n_2-1)$, где $F_{крит}$ табличные значения критерия Фишера-Снедекора.

Если $F_{расч}<F_{крит}$, то нулевая гипотеза о равенстве дисперсий не отвергается, дисперсии различаются незначительно, расхождения между ними носят случайный характер. Если же $F_{расч}\geq F_{крит}$, то гипотеза о равенстве дисперсий отклоняется и проверка гипотез о равенстве средних не может быть применена.

Проверка основной гипотезы о равенстве средних уровней двух нормально распределенных совокупностей n_1 и n_2 осуществляется на основе t-критерия Стьюдента:

H_0: $\bar{y}_1=\bar{y}_2$; H_1: $\bar{y}_1\neq\bar{y}_2$;

$$t_{расч}=\frac{\bar{y}_1-\bar{y}_2}{\sqrt{(n_1-1)\sigma_1^2+(n_2-1)\sigma_2^2}}\sqrt{\frac{n_1n_2(n_1+n_2-2)}{n_1+n_2}}; \qquad (3.3)$$

$$t_{крит}=t(\alpha,\ n_1+n_2-2). \qquad (3.4)$$

Если $|t_{расч}|<t_{крит}$, то нет основания отвергать нулевую гипотезу, т.е. тенденция отсутствует.

Замечание. Данный метод применим в случае рядов с монотонной тенденцией. Если же ряд динамики меняет общее направление развитии, то точка тенденции оказывается близкой к середине ряда поэтому средние двух отрезков будут близки, а проверка может не показать тенденции.

Метод Фостера-Стюарта

Данный метод разработан Ф.Фостером и А. Стюартом. Они предложили по данным исследуемого ряда определить величины u_t и v_t путем последовательного сравнения уровней ряда.

Если какой-либо уровень ряда превышает по своей величине каждый из предыдущих уровней, то u_t=1, в остальных случаях 0, т.е.

$$u_t = \begin{cases} 1, \textit{если } y_t > y_t, y_{t-1}, y_{t-3}, \ldots, y_1 \\ 0, \textit{в остальных случаях} \end{cases}; \qquad (3.5)$$

и наоборот, если уровень ряда меньше всех предыдущих, то v_t=1, в остальных случаях 0, т.е.

$$v_t = \begin{cases} 1, \textit{если } y_t < y_t, y_{t-1}, y_{t-3}, \ldots, y_1 \\ 0, \textit{в остальных случаях} \end{cases}. \qquad (3.6)$$

Затем находим еще две величины s и d следующим образом:

$$\begin{aligned} s &= \sum s_t, \textit{ где } s_t = u_t + v_t, \\ d &= \sum d_t, \textit{ где } d_t = u_t - v_t \end{aligned}. \qquad (3.7)$$

Суммирование производится по всем членам ряда.

Величины s и d асимптотически нормальны и имеют независимые распределения. Они существенно зависят от порядка расположения уровней во времени.

С помощью s можно проверить существует ли тенденция в дисперсиях, а d позволяет обнаружить тенденцию в средней.

Проверяют две гипотезы существенно ли отличается d от 0 и s от μ, где μ – математическое ожидание s:

$$T_s = \frac{s-\mu}{\sigma_1} \text{ и } T_d = \frac{d-0}{\sigma_2}, \qquad (3.8)$$

где σ_1 – средняя квадратическая ошибка s; σ_2 – средняя квадратическая ошибка d; значения σ_1, σ_2, μ – табуированы для различных n.

$T_d > t_{крит}(\alpha, n-1)$, то гипотеза об отсутствии тенденции в среднем отклоняется; в противном случае нет основания отвергать гипотезу. Аналогично $T_s > t_{крит}(\alpha, n-1)$, то тенденция есть и описывается некоторым трендом.

Пример 3.1. Определим наличие основной тенденции по данным табл.3.1

Таблица 3.1

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y_t	14,1	9,3	19,4	19,7	5,4	24,2	13,8	24,5	14,7	16,6	5,6	16,2	25,3	11,9	18,5

Решение.

Проверим наличие тенденции с помощью метода серий. Делим ряд на две части: n_1=7, n_2=8. По каждой вычисляем средние и дисперсии:

$$\bar{y}_1 = 15{,}13; \quad \bar{y}_2 = 16{,}66; \quad S_1^{\,2} = 36{,}08; \quad S_2^{\,2} = 36{,}15.$$

Проверяем гипотезу о равенстве дисперсий при уровне значимости α=0,05.

$$H_0: \sigma_1=\sigma_2;\ H_1: \sigma_1\neq\sigma_2;$$

$$F_{расч} = \frac{S_2^2}{S_1^2} = \frac{36{,}15}{36{,}08} \approx 1{,}002;$$

$$F_{крит}=F(0{,}05,\ 7;\ 6)=4{,}21.$$

Так как $F_{расч}<F_{крит}$, то нет основания отвергать нулевую гипотезу. По данным наблюдения дисперсии генеральных совокупностей равны $\sigma_1^{\,2}=\sigma_2^{\,2}$, исправленные выборочные дисперсии $S_1^{\,2}, S_2^{\,2}$ различаются незначимо. Тогда проверяем основную гипотезу:

$$t_{расч} = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{(n_1-1)\sigma_1^2 + (n_2-1)\sigma_2^2}} \sqrt{\frac{n_1 n_2 (n_1+n_2-2)}{n_1+n_2}} \approx -0{,}49;$$

$$t_{крит} = t(0{,}05; 13) = 2{,}16$$

Так как $|t_{расч}| < t_{крит}$, то нет оснований отвергать нулевую гипотезу о равенстве средних, расхождения между вычисленными средними незначимо. Отсюда вывод, что тренд в данной выборке отсутствует.

1. Проверим наличие тренда в данном ряду по методу Фостера-Стюарта. Строим дополнительную таблицу:

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
y_t	14,1	9,3	19,4	19,7	5,4	24,2	13,8	24,5	14,7	16,6	5,6	16,2	25,3	11,9	18,5
u_t	0	0	1	1	0	1	0	1	0	0	0	0	1	0	0
v_t	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0
s_t	0	1	1	1	1	1	0	1	0	0	0	0	1	0	0
d_t	0	-1	1	1	-1	1	0	1	0	0	0	0	1	0	0

Из таблицы находим: s=7; d=3; при n=15 имеем μ=4,636; σ_1=1,521; σ_2=2,153 (приложение 8). Тогда

$$T_s = \frac{3-0}{2{,}153} \approx 1{,}39; \quad T_d = \frac{7-4{,}636}{1{,}521} \approx 1{,}55; \quad t_{крит}(0{,}01, 14) = 2{,}62.$$

$T_d < t_{крит}$, т.е. нет оснований отвергать гипотезу об отсутствии тенденции в среднем. Метод Фостера-Стюарта еще раз подтвердил, что тренд в ряду динамики отсутствует.

$T_s < t_{крит}$, то гипотеза об отсутствии тенденции в дисперсиях не отвергается, тенденция не существует.

3.2. Сглаживание временных рядов

Сглаживание временного ряда, т.е. замена фактических уровней расчетными значениями, имеющими меньшую колеблемость, чем исходные данные является простым методом выявления тенденции развития. Соответствующее преобразование называется фильтрованием.

Наиболее часто на практике используются линейные фильтры. Общая формула линейного фильтра такова:

$$\widetilde{y}_t = \sum_{r=-1}^{k} a_r y_{t+r}, \qquad (3.7)$$

где $\widetilde{y}_t$ - сглаженное (отфильтрованное) значение временного ряда в момент времени t; a_r – вес присваиваемый значению исходного ряда, находящемуся на расстоянии r от рассматриваемого момента t.

Сглаживание временных рядов проводится по следующим причинам:

- В ряде случаев при графическом изображении временного ряда тренд прослеживается недостаточно отчетливо. Поэтому ряд сглаживают, на график наносят сглаженные значения и, как правило, тенденция проявляется более четко.
- Некоторые методы анализа и прогнозирования требуют в качестве предварительного условия сглаживание временного ряда.
- Сглаживание временных рядов используется при устранении аномальных наблюдений.
- Методы сглаживания в настоящее время применяются для непосредственного прогнозирования экономических показателей.

Существующие методы сглаживания делят на две группы:

1. *Методы первого типа (аналитические)*. Сглаживание с использованием кривой, проведенной относительно фактических значений ряда так, чтобы эта кривая отображала тенденцию, присущую ряду и одновременно освобождала его от мелких незначительных колебаний. Такие кривые называют еще *кривыми роста,* и они используются главным образом для прогнозирования экономических показателей.

2. *Методы механического сглаживания*. При использовании этих методов производится сглаживание каждого отдельного уровня ряда с использованием фактических значений соседних с ним уровней. Для сглаживания временных рядов часто используются методы простой и взвешенной скользящей средней, экспоненциального сглаживания.

Метод простой скользящей средней

Если $\sum a_{r=1}$ и a_r=const, то фильтр (3.7) означает вычисление средней арифметической, которую называют скользящей средней.

Цель сглаживания временного ряда заключается в получении ряда с меньшим разбросом уровней, что в ряде случаев позволяет на основе визуального анализа сделать вывод о наличии тенденции и ее характерных особенностях.

Согласно этому методу определяется количество наблюдений, входящих в интервал сглаживания. При этом используют правило:

если необходимо сгладить мелкие, беспорядочные колебания, то интервал сглаживания берут по возможности большим и, наоборот, интервал сглаживания уменьшают, когда нужно сохранить более мелкие волны и освободиться от периодически повторяющихся колебаний, возникающих, например, из-за автокорреляций уровней.

Для удобства сопоставления сглаженного и исходного рядов ширину интервала сглаживания чаще выбирают нечетным числом $m=2k+1$. Тогда $a_r = \frac{1}{m}$ и из (3.7) получаем:

$$\tilde{y}_t = \frac{1}{2k+1} \sum_{r=-k}^{k} y_{t+r} \,. \qquad (3.8)$$

Интервал сглаживания сдвигается на один член вправо и по формуле (3.8) находится сглаженное значение для t+1 наблюдения. Затем снова производят сдвиг и т.д. Процедура продолжается до тех пор, пока в интервал сглаживания не войдет последнее наблюдение временного ряда.

Недостатком метода является невключение в процедуру сглаживания первых и последних k наблюдений временного ряда.

Метод простой скользящей средней возможно использовать, если графическое изображение ряда напоминает прямую линию. В этом случае не искажается динамика развития исследуемого процесса. Однако когда тренд выравниваемого ряда имеет изгибы и к тому же желательно сохранить мелкие волны, использовать для сглаживания ряда метод простой скользящей средней нецелесообразно, поскольку при этом:

- выравниваются и выпуклые, и вогнутые линии;
- происходит сдвиг волны вдоль ряда;

- изменяется знак волны, т.е. на кривой, соединяющей сглаженные точки, вместо выпуклого участка образуется вогнутый и наоборот. Последнее имеет место в случаях, когда интервал сглаживания в полтора раза превышает длину волны.

Таким образом, если развитие процесса носит нелинейный характер, то применение метода простой скользящей средней может привести к значительным искажениям исследуемого процесса. В таких случаях более надежным является использование других методов сглаживания, например, метода взвешенной скользящей средней.

Метод взвешенной скользящей средней

Суть методов взвешенных скользящих средних заключается в том, что значениям исходного ряда приписывается вес a_r, зависящий от расстояния до середины интервала сглаживания, т.е. от $|a_r|$. Для определения весов прибегают к различным подходам.

Рассмотрим первый подход. Пусть весами являются члены разложения бинома $(0{,}5 + 0{,}5)^{2k}$, m=2k+1. Тогда $a_r = C_{2k}^{k-r}(0{,}5)^{2k}$, $a_{-r} = C_{2k}^{k+r}(0{,}5)^{2k}$

Получаем

при m=3 (k=1) $a_{-1} = \frac{1}{4},\ a_0 = \frac{1}{2},\ a_1 = \frac{1}{4}$;

при m =5 (k=2) $a_{-2} = \frac{1}{16},\ a_{-1} = \frac{1}{4},\ a_0 = \frac{3}{8},\ a_1 = \frac{1}{4},\ a_2 = \frac{1}{16}$;

при m=7 (k=3) $a_{-3} = \frac{1}{64},\ a_{-2} = \frac{3}{32},\ a_{-1} = \frac{15}{64}, a_0 = \frac{5}{16}, a_1 = \frac{15}{64};\ a_2 = \frac{3}{32},\ a_3 = \frac{1}{64}$;

Второй подход заключается в подборе полинома регрессии к данным, содержащимся в интервале сглаживания. Если сглаживание производится с помощью полинома (многочлена) второго или третьего порядка, то веса берутся следующие:

при m=5 – веса $\frac{-3}{35}; \frac{12}{35}; \frac{17}{35}; \frac{12}{35}; \frac{-3}{35};$

для m=7 – веса $\frac{-2}{21}; \frac{3}{21}; \frac{6}{21}; \frac{7}{21}; \frac{6}{21}; \frac{3}{21}; \frac{-2}{21}.$

Особенности весов:

- симметричны относительно центрального члена;
- сумма весов с учетом общего множителя равна 1.

Недостаток метода: первые и последние p наблюдений ряда остаются не сглаженными.

Метод экспоненциального сглаживания

Рассмотренные методы простой и взвешенной скользящей средней не дают возможности сгладить первые и последние p наблюдений

временного ряда. Отсутствие сглаженных первых наблюдений не так важно по сравнению с последними наблюдениями, особенно если целью исследования является прогнозирование развития процесса. Есть методы, позволяющие получить сглаженные значения последних уровней так же, как и всех остальных. К их числу относится метод экспоненциального сглаживания.

Особенность этого метода заключена в том, что в процедуре выравнивания каждого наблюдения используются только значения предыдущих уровней, взятых с определенным весом. Вес каждого наблюдения уменьшается по мере его удаления от момента, для которого определяется сглаживаемое значение. Сглаженное значение наблюдения ряда S_t на момент времени t определяется по формуле:

$$S_t = \alpha x_t + (1-\alpha) S_{t-1}, \qquad (3.9)$$

где S_t – значение экспоненциальной средней в момент t;

S_{t-1} – значение экспоненциальной средней в момент t-1;

α - параметр сглаживания, т.н. сглаживающий фильтр, $0 < \alpha < 1$.

Вариации α имеют серьезное влияние на характеристики самого сглаживания, и выбор оптимального значения зависит сразу от нескольких из них, причем противоречащих друг другу.

Если записать значение сглаженного ряда S_t и последовательно раскрывать значения S_{t-1}, S_{t-2}, …, через предыдущие уровни ряда и так до $y_0 = S_0$, используя рекуррентное соотношение (3.9), то в итоге легко получаем следующее представление исходного соотношения:

$$S_t = \alpha y_t + (1-\alpha) S_{t-1} = \alpha y_t + (1-\alpha)\left[\alpha y_{t-1} + (1-\alpha) S_{t-2}\right] = \alpha y_t + a(1-\alpha) y_{t-1} + \alpha(1-\alpha)^2 y_{t-2} +$$
$$\ldots + \alpha(1-\alpha)^k y_{t-k} + \ldots + (1-\alpha)^t y_0,$$

где y_0 - является начальным уровнем временного ряда.

Относительный вес каждого предшествующего уровня снижается по экспоненте по мере его удаления от момента, для которого вычисляется сглаженное значение (отсюда произошло название этого метода сглаживания).

В качестве нулевого уровня может быть использована средняя арифметическая нескольких начальных значений исходного ряда.

Выбор величины постоянной сглаживания требует особого внимания. Рассмотрим критические значения α, чтобы пронаблюдать, что будет происходить с процессом в этих крайних точках. Если взять $\alpha = 0$, то получим $S_t = S_0$, то есть адаптация модели отсутствует. Если принять $\alpha = 1$, то получим $S_t = y_t$, то есть модель, в которой сглаженное значение равно фактическому уровню временного ряда. От численного значения параметра α зависит, насколько быстро будет уменьшаться вес

предшествующих наблюдений и в соответствии с этим степень их влияния на сглаживаемый уровень. На практике подбор допустимого значения параметра сглаживания рекомендуется производить эмпирическим путем, то есть, итеративно перебирая его возможные значения и выбирая оптимальный уровень коэффициента по критерию минимизации дисперсии остатков на тестовом наборе данных. Следует отметить, что в случае, когда параметр принимает значения близкие к 1, следует подвергнуть сомнению законность выбора данной модели. Так как это может свидетельствовать о наличии в ряду ярко выраженных тенденций или сезонных колебаний. Для таких рядов следует использовать другие модели, более эффективные.

При практическом использовании метода экспоненциального сглаживания возникают следующие затруднения: выбор сглаживающего параметра α и определение начального условия y_0. Чем больше значение параметра α, тем меньше сказывается влияние предшествующих уровней и соответственно меньшим оказывается сглаживающее воздействие экспоненциальной средней.

Задачу выбора параметра y_0, определяющего начальные условия, предлагается решать следующим образом: если есть данные о развитии процесса в прошлом, то их среднее значение можно принять в качестве y_0, если таких сведений нет, то в качестве y_0 используют исходное (первое) значение наблюдения временного ряда y_1.

3.3 Показатели ряда динамики

Абсолютные показатели динамики

Анализ скорости и интенсивности развития явления во времени осуществляется с помощью статистических показателей, которые получаются в результате сравнения уровней между собой. К таким показателям относятся:

- абсолютный прирост;
- темп роста и прироста;
- абсолютное значение одного процента прироста.

При этом принято сравниваемый уровень называть отчётным, а уровень, с которым производится сравнение, - базисным.

Приняты следующие условные обозначения:

y_i – уровень любого периода (кроме первого), называемый уровнем текущего периода;

y_{i-1} – уровень периода, предшествующего текущему;

y_k – уровень, принятый за постоянную базу сравнения (часто начальный уровень).

Методы расчета показателей представлены в таблице 3.1.

Таблица 3.1. Показатели динамики

Наименование показателя	Метод расчета	
	С переменной базой (цепные)	С постоянной базой (базисные)
1. Абсолютный прирост (Δ)	$\Delta = y_i - y_{i-1}$	$\Delta' = y_i - y_k$
2. Коэффициент роста ($К_р$)	$K_p = \frac{y_i}{y_{i-1}}$	$K'_p = \frac{y_i}{y_k}$
3.Темп роста ($Т_р$)	$T_p = K_p \cdot 100$	$T'_p = K'_p \cdot 100$
4. Темп прироста ($Т_п$)	$T_n = (K_p - 1) \cdot 100$ $T_n = T_p - 100$ $T_n = \frac{\Delta}{y_{i-1}} \cdot 100$	$T'_n = (K'_p - 1) \cdot 100$ $T'_n = T'_p - 100$ $T'_n = \frac{\Delta'}{y_k} \cdot 100$
5. Абсолютное значение 1% прироста (А)	$A = \frac{\Delta}{T_n};\ A = \frac{y_{i-1}}{100}$	$A' = \frac{\Delta'}{T'_n};\ A' = \frac{y_k}{100}$

Абсолютный прирост характеризует размер увеличения (или уменьшения) уровня ряда за определённый промежуток времени.

Если k=1, то уровень y_{i-1} является предыдущим для данного ряда, а абсолютные приросты изменения уровня будут цепными. Если же k постоянно для данного ряда, то абсолютные приросты будут базисными.

Коэффициент роста или *темп роста* – это показатель интенсивности изменения уровня ряда в зависимости от того, выражается ли он в виде коэффициента или в процентах. Коэффициент роста показывает, во сколько раз данный уровень ряда больше базисного уровня (если этот больше единицы) или какую часть базисного уровня составляет уровень текущего периода за некоторый промежуток времени (если он меньше единицы). В качестве базисного уровня в зависимости от цели исследования может приниматься какой – то постоянный для всех уровень (часто начальный уровень ряда) либо для каждого последующего предшествующий ему.

Темп прироста – это показатель, характеризующий относительную скорость изменения уровня ряда в единицу времени. Темп прироста показывает, на какую долю (или процент) уровень данного периода или момента времени больше (или меньше) базисного уровня.

Если темп роста всегда положительное число, то темп прироста может быть положительным, отрицательным и равным нулю.

Абсолютное значение одного процента прироста служит косвенной мерой базисного уровня и вместе с темпом прироста позволяет рассчитывать абсолютный прирост уровня за рассматриваемый период.

Для характеристики динамики явлений в ряде случаев используют: абсолютное ускорение, относительное ускорение.

Абсолютным ускорением называют разность между последующим и предыдущим абсолютными приростами.

$$\Delta'' = \Delta y_i - \Delta y_{i-1}$$

Ускорение показывает, насколько данная скорость больше (меньше) предыдущей. Абсолютное ускорение есть скорость изменения скорости. Оно может быть положительным и отрицательным числом.

Относительным ускорением называют отношение абсолютного ускорения к абсолютному приросту, принятому за базу сравнения ($\frac{\Delta''}{\Delta y_i} \cdot 100$), т.е. относительное ускорение есть темп прироста абсолютного прироста. Оно вычисляется в том случае, если абсолютный прирост, принятый за базу сравнения, число положительное.

Например, для ряда 40, 45, 48, 52, 57

абсолютные приросты составят 5,3,4,5;

абсолютные ускорения: -2, 1, 1

относительные ускорения: $\frac{-2}{5} \cdot 100 = -40\%;\ \ \frac{1}{3} \cdot 100 = 33{,}3\%;\ \ \frac{1}{4} \cdot 100 = 25\%.$

Средние показатели динамики

Для характеристики интенсивности развития за длительный период рассчитываются средние показатели динамики. Приняты следующие условные обозначения:

$y_1, y_2, \ldots, y_n$ – все уровни последовательных периодов (дат);

n – число уровней ряда;

t – продолжительность периода, в течение которого уровень не изменялся.

Для характеристики интенсивности развития за длительный период рассчитываются средние показатели динамики.

Средний уровень ряда динамики ($\bar{y}$) рассчитывается по средней хронологической. *Средней хронологической* называется средняя, исчисленная из значений, изменяющихся во времени. В хронологической средней отражается совокупность тех условий, в которых развивалось изучаемое явление в данном промежутке времени.

Методы расчёта среднего уровня интервального и моментального рядов динамики различны. Для интервальных рядов с равноотстоящими уровнями средний уровень находится по формуле средней

арифметической простой, а для неравноотоящих уровней – по средней арифметической взвешенной:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}; \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i \cdot t_i}{\sum t_i},$$

где y_i – уровень ряда динамики; n – число уровней; t_i –длительность интервала времени между уровнями.

Для моментного ряда с равными интервалами средгий уровень ряда вычисляется по следующей формуле:

$$\bar{y} = \frac{\frac{1}{2} \cdot y_1 + y_2 + \cdots + y_{n-1} + \frac{1}{2} \cdot y_n}{n-1}$$

Обобщающим показателем скорости изменения во времени является *средний абсолютный прирост* ($\bar{\Delta}$).

Этот показатель дает возможность установить, насколько в среднем за единицу времени должен увеличиться уровень ряда (в абсолютном выражении), чтобы, отправляясь от начального уровня за данное число периодов (например, лет), достигнуть конечного уровня. Для его определения используют формулу средней арифметической простой:

$$\bar{\Delta} y = \frac{\sum_{i=1}^{n-1} \Delta_{i/i-1}}{n-1} \text{ или } \bar{\Delta} y = \frac{y_n - y_1}{n-1}.$$

Возможен и другой способ расчёта среднего абсолютного прироста исходя из кумулятивных данных:

$$\bar{\Delta} y = \frac{2\left(\sum_{i=1}^{n} y_i - n y_1\right)}{n(n+1)}.$$

Обе формулы применяются в зависимости от цели исследования.

Средний коэффициент роста $(\bar{K}_p)$ вычисляется по формулам:

$$\bar{K}_p = \sqrt[n-1]{K_{p1} \cdot K_{p2} \cdot \cdots \cdot K_{n-1}}, \text{ или } \bar{K}_p = \sqrt{\frac{y_n}{y_1}}.$$

Сводной обобщающей характеристикой интенсивности изменения уровней ряда динамики служит *средний темп роста*, показывающий, во сколько раз в среднем за единицу времени изменился уровень динамического ряда.

Необходимость исчисления среднего темпа роста возникает вследствие того, что темпы роста из года в год колеблются. Кроме того, средний темп роста часто следует определить в тех случаях, когда имеются данные об уровне в начале какого-либо периода и в конце его, а промежуточные данные отсутствуют. Обычно средний темп роста

вычисляется по формуле средней геометрической из цепных коэффициентов роста:

$$\overline{T}_{p_y} = \sqrt[m]{K_{2/1} \cdot K_{3/2} \cdot \ldots \cdot K_{n/n-1}} = \sqrt[m]{ПК_{pi/i-1}}.$$

Средний темп роста может быть также выражен формулой:

$$\overline{T}_{p_y} = \sqrt[n-1]{\frac{y_n}{y_1}}.$$

При расчёте средних темпов роста по периодам различной продолжительности (разноотстоящие ряды динамики) пользуются средними геометрическими взвешенными по продолжительности периодов. Формула средней геометрической взвешенной будет иметь вид:

$$\overline{T}_{p_y} = \sqrt[\sum t]{\left(K_{2/1}\right)^{t_1} \cdot \left(K_{3/2}\right)^{t_2} \cdot \ldots \left(K_{n/n-1}\right)^{t_n}},$$

где t – интервал времени, в течении которого сохраняется данный темп роста; $\sum t$ - сумма отрезков времени периода.

Средний темп прироста не может быть определен непосредственно на основании последовательных темпов прироста или показателей среднего абсолютного прироста. Для его вычисления необходимо вначале найти средний темп роста, а затем уменьшить его на единицу, или на 100%:

$$\overline{T}_{np_y} = \overline{T}_p - 100.$$

Средняя величина абсолютного значения 1% прироста $\left(\overline{A}\right)$ вычисляется по формуле:

$$\overline{A} = \frac{\overline{\Delta}}{\overline{T}_n}.$$

3.4. Метод аналитического выравнивания

Аналитическим выравниванием временного ряда называют нахождение аналитической функции $\hat{y}=f(t)$, характеризующей основную тенденцию изменения уровней ряда с течением времени.

При аналитическим выравнивании исходят из предположения, что аддитивная модель временного ряда может быть представлена как сумма двух компонент:

$$y(t)=f(t)+\varepsilon_t,$$

где ε_t – случайная компонента с нулевой средней и постоянной дисперсией выражает ошибку модели из-за действия случайных факторов.

Чаще всего в качестве кривой роста применяются следующие функции:

- линейная $y_t=a_0+a_1t$;
- парабола второго и более высоких порядков k

$$y_t=a_0+a_1t+a_2t^2+...+a_kt^k;$$

- гиперболическая $y_t = a_0 + \frac{a_1}{t}$;
- экспонента $y_t = e^{a_0+a_1t}$;
- потенциальная $y_t = a_0 \cdot a_1^{\,t}$;
- степенная $y_t = a_0 t^{a_1}$;
- логистическая кривая $y_t = \frac{K}{1+a_0e^{-a_1t}}$;
- кривая Гомперца $y_t = K \cdot a_0^{\,a_1^{\,t}}$.

Построение таких функций ничем не отличается от построения уравнений парной регрессии (линейной или нелинейной) с учетом того, что в качестве зависимой переменной используются фактические уровни временного ряда y_t, а в качестве независимой переменной моменты времени t=1,2, ..., n.

Для построения кривой роста необходимо выбрать вид аналитической зависимости и затем оценить значения ее параметров.

Для определения вида аналитической зависимости применяются такие методы, как

– построение и визуальный анализ графика зависимости уровней ряда от времени. При этом целесообразнее использовать графическое изображение сглаженных уровней, в которых случайные колебания погашены.

– расчет и анализ вида уравнения, основанный на рассчитанных показателях динамики:

- если относительно стабильны абсолютные приросты, сглаживание может быть выполнено по прямой;
- если абсолютные приросты равномерно увеличиваются (вторые разности уровней приблизительно равны), сглаживание может быть выполнено по параболе второго порядка;
- при ускоренно возрастающих (убывающих)абсолютных приростах – параболу третьего порядка;
- при относительно стабильных темпах роста – показательную функцию.

– анализ автокорреляционной функции исходного и преобразованного временного ряда;

– метод перебора, при котором строятся кривые роста различного вида с последующим выбором наилучшей на основании значения скорректированного коэффициента детерминации R^2.

Следует отметить, что свойства изучаемого явления должны соответствовать свойствам функций, используемых для построения моделей.Надо иметь ввиду, что отдельные кривые выражают следуюший тип динамики.

Монотонное возрастание и убывание процесса характеризуют функции: 1) линейная; 2) параболическая; 3) степенная; 4) простая экспоненциальная кривая; 5) гиперболическая (главным образом убывающих процессов); 6) комбинация их видов.

Для моделирования динамических рядов, в которых появляется быстрое развитие в начале и затухание к концу ряда, т.е. которые характеризуются стремлением к некоторой предельной величине, насыщению, применяются логистические кривые.

Тип процессов, характеризующихся наличием экстремальных значений, описывается кривой Гомперца.

Однако процедура построения модели и разработки прогноза с использованием аналитического выравнивания тренда состоит не только из предварительного выбора одной или нескольких кривых, которые наилучшим образом соответствуют характеру изменения ряда динамики, но и оценки параметров выбранных кривых, проверки адекватности выбранных кривых рассматриваемому явлению; окончательного выбора кривой роста; расчета точечного и интервального прогнозов.

3.5. Критерии адекватности моделей временных рядов

Проверка адекватности модели реальному явлению является важным этапом в статистическом анализе, так как только при правильном выборе модели возможна процедура прогнозирования. Для ее осуществления исследуют ряд остатков $\varepsilon_t = y_t - \hat{y}_t$, т.е. отклонений расчетных значений от фактических. Если модель выбрана правильно, то для остатков характерны:

1. равенство нулю математического ожидания;
2. случайный характер отклонений от математического ожидания;
3. отсутствие автокорреляции и неизменность дисперсии остатков во времени;
4. нормальный закон распределения.

1. Проверка равенства математического ожидания уровней ряда остатков нулю осуществляется в ходе проверки соответствующей нулевой гипотезы H_0:|ε|=0. С этой целью строится *t*-статистика

$$t_{набл} = \frac{|\bar{\varepsilon}|}{S_\varepsilon}\sqrt{n},$$

где $\bar{\varepsilon}$ – среднее арифметическое значение уровней ряда остатков ε_t;

$S_\varepsilon = \sqrt{\frac{\sum_{t=1}^{n}(\varepsilon_t - \bar{\varepsilon})^2}{n-1}}$ – среднеквадратическое отклонение для этой последовательности.

На уровне значимости α гипотеза отклоняется, если $t_{набл} > t_{\alpha,\nu}$, где $t_{\alpha,\nu}$ – критерий распределения Стьюдента с доверительной вероятностью (1 - α) и $\nu = n$-1 степенями свободы.

2. Для проверки условия случайности возникновении отдельных отклонений от тренда часто используется критерий, основанный на поворотных точках. Значение случайной переменной считается поворотной точкой, если оно одновременно больше соседних с ним элементов или, наоборот, меньше значений предыдущего и последующего за ним члена. Если остатки случайны, то поворотная точка приходится примерно на каждые 1,5 наблюдения. Если их больше, то возмущения быстро колеблются, и это не может быть объяснено только случайностью. Если же их меньше, то последовательные значения случайного компонента положительно коррелированны.

Существует определенная зависимость между средней арифметической $\bar{p}$, дисперсией $\sigma_{\bar{p}}^2$; количества поворотных точек *p* и числом членов исходного ряда наблюдений *n*. В случайной выборке средняя арифметическая (математическое ожидание) числа поворотных точек равна $\bar{p} = \frac{2}{3}(n-2)$, а их дисперсия вычисляется по формуле $\sigma_{\bar{p}}^2 = \frac{16n-29}{90}$.

Учитывая эти соотношения, критерий случайности отклонений от тренда при уровне вероятности 0,95 можно представить, как

$$p > \left[\frac{2}{3}(n-2) - 1{,}96\sqrt{\frac{16n-29}{90}}\right],$$

где *p* — фактическое количество поворотных точек в случайном ряду;

1,96 — квантиль нормального распределения для 5%-го уровня значимости; квадратные скобки означают, что от результата вычисления следует взять целую часть (не путать с процедурой округления!).

Если неравенство не соблюдается, то ряд остатков нельзя считать случайным (т.е. он содержит регулярную компоненту), и стало быть, модель не является адекватной.

Кроме критерия поворотных точек можно использовать метод серий, основанный на медиане выборки. Суть его в следующем. Расположим отклонения от тренда в порядке возрастания вариационный ряд ε_1; ε_2; ε_3;…;ε_n; где ε_1 – наименьшее отклонение. В данном вариационном ряду находим медиану $\varepsilon_{ме}$, т.е. берем среднее (по расположению) значение вариационного ряда: $\varepsilon_{ме} = \varepsilon_{\frac{m+1}{2}}$, если n – нечетное; $\varepsilon_{ме} = \frac{1}{2}\left(\varepsilon_{\frac{n}{2}} + \varepsilon_{\frac{n}{2}+1}\right)$, если n – четно. Затем возвращаемся к исходному ряду динамики отклонений от тренда и будем для вместо каждого ε_i ставит плюс, если $\varepsilon_i > \varepsilon_{ме}$, и минус, если $\varepsilon_i < \varepsilon_{ме}$ (отклонения от тренда, равные $\varepsilon_{ме}$, в полученной таким образом последовательности плюсов и минусов опускаются). Последовательность пюсов и минусов характеризуется общим числом серий V_n и продолжительностью самой длинной серии K_n. Под “серией” понимается последовательность подряд идущих плюсов или минусов. Иногда серия может состоять только из одного плюса или минуса, и тогда протяженность равна единице. Если отклонения от тренда стохастически независимы, то чередование плюсов и минусов в последовательности должно быть более или менее “случайным”, т.е. такая последовательность не должна содержать слишком длинных серий подряд идущих плюсов и подряд идущих минусов, а общее число серий не дожно быть слишком малым.

Отконения от тренда будут случайными, если выполнены следующие неравенства при 5%-ном уровне значимости:

$$K_{\max(n)} < [3{,}3(\lg n + 1)];$$
$$V_{(n)} > \left[\frac{1}{2}\left(n + 1 - 1{,}96\sqrt{n-1}\right)\right]. \qquad (3.10)$$

3. Наличие (отсутствие) автокорреляции в отклонениях от модели роста проще всего проверить с помощью критерия Дарбина—Уотсона. С этой целью строится статистика Дарбина— Уотсона (d-статистика), в основе которой лежит расчетная формула

$$d = \frac{\sum_{t=1}^{n} (\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}. \qquad (3.11)$$

Теоретическое основание применения этого критерия обусловлено тем, что в динамических рядах, как сами наблюдения, так и отклонения от них распределяются в хронологическом порядке.

При отсутствии автокорреляции значение d примерно равно 2, а при полной автокорреляции - 0 или 4. Следовательно, оценки, получаемые по критерию, являются не точечными, а интервальными. Верхние (d_2) и нижние (d_1) критические значения, позволяющие принять или отвергнуть гипотезу об отсутствии автокорреляции, зависят от количества уровней динамического ряда и числа независимых переменных модели. Значения этих границ даны в специальных таблицах. При сравнении расчетного значения d-статистики с табличным могут возникнуть такие ситуации: $d > d_2$ — ряд остатков не коррелирован; $d < d_1$ — остатки содержат автокорреляцию; $d_1 < d < d_2$ — область неопределенности, когда нет оснований ни принять, ни отвергнуть гипотезу о существовании автокорреляции. Если d превышает 2, то это свидетельствует о наличии отрицательной корреляции. Перед входом в таблицу такие значения следует преобразовать по формуле $d' = 4 - d$.

Установив наличие автокорреляции остатков, надо улучшать модель. Если же ситуация оказалась неопределенной, применяют другие критерии. В частности, можно воспользоваться первым коэффициентом автокорреляции:

$$r_1 = \frac{\sum_{t=2}^{n} \varepsilon_t \varepsilon_{t-1}}{\sum_{t=1}^{n} \varepsilon_t^2} \qquad (3.12)$$

Для суждения о наличии или отсутствии автокорреляции в исследуемом ряду фактическое значение коэффициента автокорреляции сопоставляется с табличным (критическим) для 5%-го или 1%-го уровня значимости. Если фактическое значение коэффициента автокорреляции меньше табличного, то гипотеза об отсутствии автокорреляции в ряду может быть принята. Когда же фактическое значение больше табличного, делают вывод о наличии автокорреляции в ряду динамики.

4. Для определения того, что отклонения от тренда подчиняется закону нормального распределения можно рассчитать показатели асимметрии, эксцесса, а также их квадратические ошибки.

При нормальном распределении показатели асимметрии и эксцесса равны нулю, но поскольку мы используем предположение, что исследуемый ряд динамики является выборкой из более длинного ряда динамки, то в этом случае показатели асимметрии и эксцесса характеризуют выборочную совокупность, являются выборочными оценками. Поэтому уровни ряда являются нормально распределенными, если выполняются следующие условия:

$$|As| < 1{,}5\sigma_{As}; \quad \left|Ex + \frac{6}{n+1}\right| < 1{,}5\sigma_{Ex}, \tag{3.13}$$

где среднеквадратические ошибки коэффициентов асимметрии и эксцесса определяются по формулам:

$$\sigma_{As} = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}; \quad \sigma_{Ex} = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}. \tag{3.14}$$

Если выполняется хотя бы одно из неравенств:

$$|As| \geq 2\sigma_{As}; \quad \left|Ex + \frac{6}{n+1}\right| \geq 2\sigma_{Ex}, \tag{3.15}$$

то данные не являются даже приблизительно нормальными и их применение в дальнейшем анализе не рекомендуется.

Если все четыре пункта проверки 1-4 дают положительный результат, делается вывод о том, что выбранная модель является адекватной реальному ряду динамики. Только в этом случае ее можно использовать для построения прогнозных оценок. В противном случае модель надо улучшать.

3.6. Оценка точности модели

Оценка точности модели имеет смысл только для адекватных моделей. В случае временных рядов точность модели определяется как разность между фактическим и расчетным значениями. В качестве статистических показателей точности чаще всего применяют стандартную ошибку прогнозируемого показателя или *среднеквадратическое отклонение от линии тренда*:

$S_\varepsilon = \sqrt{\frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n-m}} = \sqrt{\frac{\sum_{t=1}^{n}\varepsilon_t^2}{n-m}}$, где m — число параметров модели,

и *среднюю относительную ошибку аппроксимации*:

$$E_{отн} = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{\varepsilon_t}{y_t}\right| \cdot 100\%. \tag{3.16}$$

Если ошибка не превосходит 15%, точность модели считается приемлемой. В общем случае допустимый уровень точности, а значит и надежности устанавливает пользователь модели, который в результате содержательного анализа проблемы выясняет, насколько она чувствительна к точности решения и насколько велики потери из-за неточного решения.

3.7 Построение точечного и интервального прогнозов

Если в ходе проверки разрабатываемая модель признана достаточно надежной, то процесс экстраполяции заключаются в подстановке соответствующей величины периода упреждения в формулу, описывающую тренд. Экстраполяция дает точечную прогностическую оценку. Поэтому одна из основных задач, возникающий при экстраполяции тренда, заключается в определении доверительных интервалов прогноза. В общем виде доверительный интервал для тренда определяется как

$$\hat{y}_t \pm t_\alpha S_{\hat{y}}, \qquad (3.17)$$

где $S_{\hat{y}}$ - средняя квадратическая ошибка тренда; $\hat{y}_t$ - расчетное значение y_t; t_α – значение t-статистики Стьюдента.

Если $t=i+L$, то уравнение определит значение доверительного интервала для тренда, определенного на L единиц времени.

Доверительный интервал для прогноза также должен учитывать не только неопределенность, связанную с положением тренда, но возможность отклонения от этого тренда. Обозначим соответствующую среднюю квадратическую ошибку как S_p, тогда доверительный интервал прогноза составит

$$\hat{y}_{i+L} \pm t_\alpha S_p. \qquad (3.18)$$

Если тренд характеризуется прямой, то величина S_p определяется следующим образом:

$$S_p = S_{\hat{y}} \sqrt{\frac{n+1}{n} + \frac{(t_L - \bar{t})^2}{\sum_{t=1}^{n} (t_i - \bar{t})^2}}, \qquad (3.19)$$

где $S_{\hat{y}}$ - среднее квадратическое отклонение фактических наблюдений от расчетных значение y; n – число наблюдений (длина ряда динамики); t_L – время на которое делается экстраполяция; $\bar{t}$ - значение порядкового номера уровня, стоящего в середине ряда.

Рассмотрим процедуру построение модели по следующим данным (табл.3.2)

Таблица 3.2

Среднегодовая стоимость промышленных фондов предприятия за 1995-2010 г.г.

Годы	1995	1996	1997	1998	1999	2000	2001	2002
Усл. ед.	153	160	170	179	187	187	202	219
Годы	2003	2004	2005	2006	2007	2008	2009	2010
Усл. ед.	229	248	260	277	291	312	334	352

Визуальный анализ графика зависимости уровней ряда от времени позволил отобрать три функции:

$$\hat{y}_t = a_0 + a_1 t;$$

$$\hat{y}_t = a_0 + a_1 t + a_2 t^2;$$

$$\hat{y}_t = a_0 \cdot a_1^t.$$

Параметры функции и коэффициенты аппроксимации таковы:

$$\hat{y}_t = 123 + 13{,}18t \quad (R^2 = 0{,}9707);$$

$$\hat{y}_t = 150{,}48 + 4{,}02t + 0{,}54t^2 \quad (R^2 = 0{,}998);$$

$$\hat{y}_t = 140{,}85 \cdot 1{,}06^t \quad (R^2 = 0{,}9937).$$

Сравнивая коэффициенты аппроксимации был сделан вывод, что в наибольшей степени к фактическим данным приближается тренд рассчитанный по параболе второго порядка. Но для того, чтобы использовать в прогнозе параболу второго порядка, необходимо проверить правильность её выбора.

1. Проверяем равенств нулю математического ожидания уровней ряда остатков:

$$t_{набл} = \frac{|-0{,}00145|}{3{,}69718}\sqrt{16} = 0{,}00067,$$

$$t_{крит}(0{,}05;15) = 1{,}75.$$

На уровне значимости α гипотеза о равенстве нулю математического ожидания принимается.

2. С помощью критерия серий проверяем случайность отклонений от тренда ряда динамики среднегодовой стоимости промышленных фондов предприятия. Получим

$$\varepsilon_{ме} = -0{,}093;\ K_{max(n)} = 3;\ V_{(n)} = 9;$$

$$3 < [3{,}3(\lg 16 + 1)] = 7.3;$$

$$9 > \left[\frac{1}{2}\left(n + 1 - 1{,}96\sqrt{n-1}\right)\right] = 4{,}7.$$

Следовательно, ряд динамики отклонений от тренда состоит из случайных независимых величин.

3. Наличие (отсутствие) автокорреляции в отклонениях от модели роста проверяем с помощью критерия Дарбина—Уотсона. *d*-статистика равна

$$d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n}\varepsilon_t^2} = 1{,}92 .$$

Из таблицы распределения критерия Дарбина – Уотсона (Приложение VI) находим d_2=1,54 и d_1=0,95 Так как $d > d_2$, то ряд остатков не коррелирован и гипотеза об отсутствии автокорреляции принимается, автокорреляции в ряду остатков нет.

4. Для определения того, что отклонения от тренда подчиняются нормального распределения, рассчитываем показатели асимметрии и эксцесса и их средние квадратические ошибки по формулам 3.13. В данном случае они равны: As=-0,76; Ex=0,81; σ_{As}=0,51; σ_{Ex}=0,78. Для проверки нормальности получим следующие неравенства 3.14:

$$|-0{,}76| < 0{,}765; \quad |1{,}166| < 1{,}17.$$

Требования неравенств выполнены.

Следовательно, при аппроксимации ряда динамики показателя среднегодовой стоимости промышленных фондов предприятия предпочтения следует отдать параболе второго порядка $\hat{y}_t = 150{,}48 + 4{,}02t + 0{,}54t^2$, так как коэффициент аппроксимации этой кривой ближе к 1, и ряд остатков, образованный после исключения тренда, вычисленного по параболе отвечает всем гипотезам.

3.8 Моделирование сезонных колебаний

Аддитивная модель имеет следующий вид: *Y=T+S+E*, *мультипликативная модель* представляет собой произведение перечисленных компонент и имеет вид: *Y=T*S*E*, где *T* - трендовая компонента; *S* – циклическая компонента; *E* – случайная компонента.

При наличии тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих.

Автокорреляция временного ряда – корреляционная зависимость между последовательными уровнями временного ряда. Количественно она измеряется с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на несколько шагов во времени. Формула для расчета коэффициента корреляции имеет вид:

$$r_{xy} = \frac{\Sigma(x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\Sigma(x_j - \bar{x})^2 \cdot \Sigma(y_j - \bar{y})^2}} \qquad (3.20)$$

В качестве переменной x рассмотрим ряд $y_2, y_3, \ldots, y_n$; в качестве переменной y ряд $y_1, y_2, \ldots, y_{n-1}$. Тогда формула (3.20) примет вид:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)\cdot(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2 \cdot \sum_{t=2}^{n}(y_{t-1} - \bar{y}_2)^2}}, \qquad (3.21)$$

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

Величину (3.21) называют коэффициентом автокорреляции уровней ряда первого порядка, так как он измеряет зависимость между соседними уровнями ряда *t* и *t-1*, т.е. при лаге 1.

Коэффициент автокорреляции второго порядка характеризует тесноту связи между уровнями ряда y_t и y_{t-2} и определяется по формуле:

$$r_2 = \frac{\sum_{t=3}^{n}(y_t - \bar{y}_3)\cdot(y_{t-2} - \bar{y}_4)}{\sqrt{\sum_{t=3}^{n}(y_t - \bar{y}_3)^2 \cdot \sum_{t=3}^{n}(y_{t-2} - \bar{y}_4)^2}} \qquad (3.22)$$

где $\bar{y}_3 = \frac{\sum_{t=3}^{n} y_t}{n-2}$; $\bar{y}_4 = \frac{\sum_{t32}^{n} y_{t-2}}{n-2}$.

Лагом называется число периодов, по которым рассчитывается коэффициент автокорреляции.

Автокорреляционной функцией временного ряда называется последовательность коэффициентов автокорреляции уровней временного ряда.

Коррелограмма – график зависимости значений автокорреляционной функции от величины лага.

Анализ автокорреляционной функции и коррелограммы:

• если r_1 наиболее высокий, то исследуемый ряд содержит только тенденцию;

• если r_t наиболее высокий, то ряд содержит циклические колебания с периодичностью в t моментов времени;

• если ни один из коэффициентов автокорреляции не является значимым, то либо ряд не содержит тенденции и циклических колебаний, либо ряд содержит сильную нелинейную тенденцию.

Алгоритм построения аддитивной и мультипликативной моделей

Построение аддитивной и мультипликативной моделей сводится к расчету значений *T*,*S* и *E* для каждого уровня ряда. Процесс построения модели включает в себя следующие шаги.

Шаг 1. Выравнивание исходных уровней ряда методом скользящей средней.

Шаг 2. Оценка сезонной компоненты S.

Шаг 3. Устранение сезонной компоненты из исходных уровней ряда и получение выровненных данных (T+S) в аддитивной или (T·S) в мультипликативной модели.

Шаг 4. Аналитическое выравнивание уровней (T+S) или (T·S) и расчет значений T с использованием полученного уравнения тренда.

Шаг 5. Расчет полученных по модели значений (T+S) или (T·S).

Шаг 6. Расчет абсолютных и относительных ошибок.

Рассмотрим методику построения аддитивной модели на примере.

Пример 3.2. Имеются условные данные об объемах потребления электроэнергии жителями региона за 16 кварталов (табл. 3.3)

Таблица 3.3

Потребление электроэнергии жителями региона, млн кВт.ч

t	y_t	y_{t-1}	y_{t-2}	y_{t-3}	y_{t-4}
1	6,0	-	-	-	-
2	4,4	6,0	-	-	-
3	5,0	4,4	6,0	-	-
4	9,0	5,0	4,4	6,0	-
5	7,2	9,0	5,0	4,4	6,0
6	4,8	7,2	9,0	5,0	4,4
7	6,0	4,8	7,2	9,0	5,0
8	10,0	6,0	4,8	7,2	9,0
9	8,0	10,0	6,0	4,8	7,2
10	5,6	8,0	10,0	6,0	4,8
11	6,4	5,6	8,0	10,0	6,0
12	11,0	6,4	5,6	8,0	10,0
13	9,0	11,0	6,4	5,6	8,0
14	6,6	9,0	11,0	6,4	5,6
15	7,0	6,6	9,0	11,0	6,4
16	10,8	7,0	6,6	9,0	11,0

Нанесем значения y_t на график (рис.3.1)

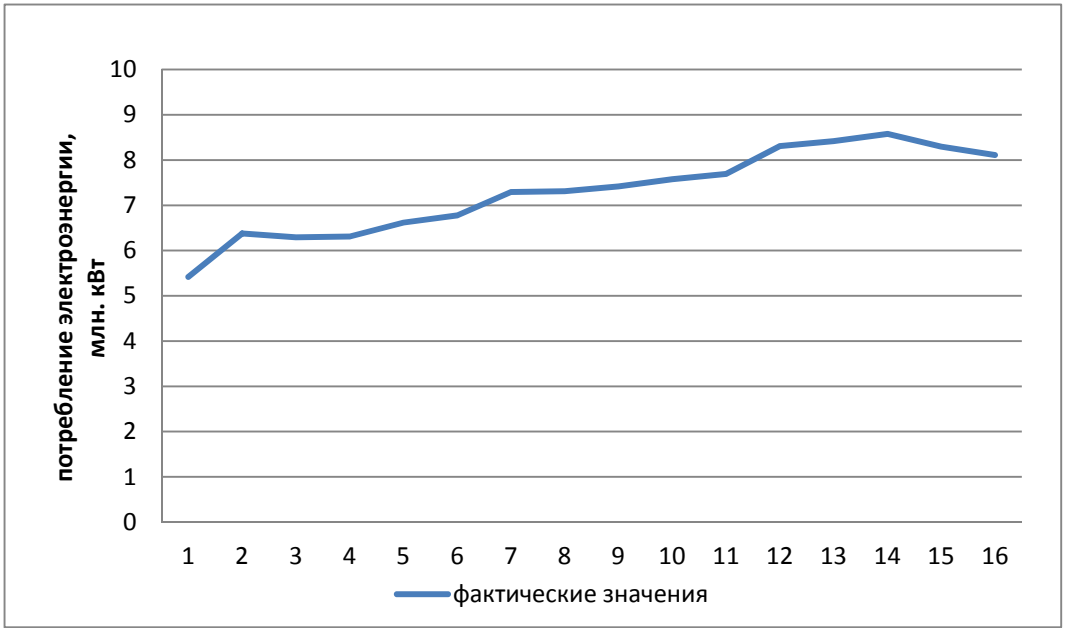


Рис.3.1. Потребление электроэнергии жителями региона,

Определим коэффициент корреляции между рядами y_t и y_{t-1}. Составим расчетную таблицу 3.4.

Таблица 3.4

Расчет коэффициентов автокорреляции I порядка для временного ряда потребления электроэнергии

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	6						
2	4,4	6,0	-2,987	-1,067	3,186	8,920	1,138
3	5	4,4	-2,387	-2,667	6,364	5,696	7,111
4	9	5,0	1,613	-2,067	-3,334	2,603	4,271
5	7,2	9,0	-0,187	1,933	-0,361	0,035	3,738
6	4,8	7,2	-2,587	0,133	-0,345	6,691	0,018
7	6	4,8	-1,387	-2,267	3,143	1,923	5,138
8	10	6,0	2,613	-1,067	-2,788	6,830	1,138
9	8	10,0	0,613	2,933	1,799	0,376	8,604
10	5,6	8,0	-1,787	0,933	-1,668	3,192	0,871
11	6,4	5,6	-0,987	-1,467	1,447	0,974	2,151
12	11	6,4	3,613	-0,667	-2,409	13,056	0,444
13	9	11,0	1,613	3,933	6,346	2,603	15,471
14	6,6	9,0	-0,787	1,933	-1,521	0,619	3,738
15	7	6,6	-0,387	-0,467	0,180	0,150	0,218
16	10,8	7,0	3,413	-0,067	-0,228	11,651	0,004

$$r_1 = \frac{\sum_{t=2}^{n}(y_t-\bar{y}_1)\cdot(y_{t-1}-\bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t-\bar{y}_1)^2\cdot\sum_{t=2}^{n}(y_{t-1}-\bar{y}_2)^2}} = \frac{9{,}813333}{\sqrt{65{,}31733\cdot 54{,}05333}} = 0{,}165;$$

где $\bar{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1} = \frac{110,8}{16-1}$ =7,386667;

$\bar{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1} = \frac{106}{16-1}$ =7,066667.

Аналогично рассчитываем остальные коэффициенты автокорреляции, результаты отражены в таблице 3.5.

Таблица 3.5

Лаг	Коэффициент автокорреляции уровней
1	0,165
2	0,567
3	0,114
4	0,983
5	0,119
6	0,722
7	0,003
8	0,974

Анализ значений коэффициентов автокорреляции позволяет сделать вывод о наличие сезонных колебаний периодичностью в четыре квартала. Данный вывод подтверждается и графическим анализом структуры ряда (см. рис. 1). Объемы потребления электроэнергии в осенне-зимний период времени (I и IV кварталы) выше, чем весной и летом (I и III кварталы). Рассчитаем компоненты аддитивной модели.

Шаг 1. Проведем выравнивание исходных уровней ряда методом скользящей средней. Для этого:

- просуммируем уровни ряда последовательно за каждые четыре квартала со сдвигом на один момент времени и определим условные годовые объемы потребления электроэнергии (гр. 3 таб.3.6);

Таблица 3.6

Расчет оценок сезонной компоненты в аддитивной модели

Номер квартала t	Потребление электроэнергии, y_t	Итого за четыре квартала	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной компоненты
1	2	3	4	5	6
1	6,0	-	-	-	-
2	4,4	24,4	6,10	-	-
3	5,0	25,6	6,40	6,250	-1,250
4	9,0	26,0	6,50	6,450	2,550
5	7,2	27,0	6,75	6,625	0,575
6	4,8	28,0	7,00	6,875	-2,075
7	6,0	28,8	7,20	7,100	-1,100
8	10,0	29,6	7,40	7,300	2,700
9	8,0	30,0	7,50	7,450	0,550
10	5,6	31,0	7,75	7,625	-2,025
11	6,4	32,0	8,00	7,875	-1,475

12	11,0	33,0	8,25	8,125	2,875
13	9,0	33,6	8,40	8,325	0,675
14	6,6	33,4	8,35	8,375	-1,775
15	7,0	-	-	-	-
16	10,8			-	-

- разделим полученные суммы на 4, найдем скользящие средние (гр.4 таб.3.6);
- найдем средние значения из двух последовательных скользящих средних – центрированные скользящие средние (гр. 5 таб.3.6).

Шаг 2. Найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и центрированными скользящими средними (гр. 6 табл. 3.6). Эти оценки, исрользуем для расчета значений сезонной компоненты S (табл. 3.7). Для этого находим средние за каждый квартал (по всем годам) оценки сезонной компоненты S_i. В моделях с сезонной компонентой предполагается, что сезонные воздействия за период взаимопогашаются. Т.е. сумма значений сезонной компоненты по всем кварталам должна быть равна нулю. Для данной модели имеем:

0,6-1,958-1,275+2,708=0,075.

Таблица 3.7

Расчет значений сезонной компоненты в аддитивной модели

Показатель	Год	Номер квартала, i			
		I	II	III	IV
	1	-	-	-1,250	2,550
	2	0,575	-2,075	-1,100	2,700
	3	0,550	-2,025	-1,475	2,875
	4	0,675	-1,775	-	-
Итого за *i*-тый квартал (за все годы)		1,800	-5,875	-3,825	8,125
Средняя оценка сезонной компоненты для за i-того квартала, $\bar{S}_i$		0,600	-1,958	-1,275	2,708
Скорректированная сезонная компонента S_i		0,581	-1,977	-1,294	2,690

Определим корректирующий коэффициент:

$$k = \frac{0,075}{4} = 0,01875.$$

Скорректированные значения сезонной компоненты находятся как разность между ее средней оценкой и корректирующим коэффициентом *k:*

$$S_i = \bar{S}_i - k, \text{ где } i=1,..,4. \quad (3.23)$$

Проверим условие равенства нулю суммы значений сезонной компоненты:

0,581-1,977-1,294+2,690=0.

Получены следующие значения сезонной компоненты:

I квартал: S_1=0,581; II квартал: $S_2 = -1{,}979$;

III квартал: $S_3 = -1{,}294$; IV квартал: $S_4 = 2{,}690$.

Шаг 3. Исключим влияние сезонной компоненты, вычитая ее значения из каждого уровня временного ряда. Получим: *T+E=Y-S* (гр.4 табл. 3.8).

Шаг 4. Для определения компоненты *T* проведем аналитическое выравнивание ряда (T+E) с помощью линейного тренда.

В ППП MS Excel линия тренда может быть добавлена в диаграмму с областями гистограммы или в график. Для этого:

1) выделите область построения диаграммы; в меню вставка выберите **Диаграмма/Добавить линию тренда**;

2) в появившемся диалоговом окне (рис. 3.2) выберите вид линии тренда и задайте соответствующие параметры. Для полиномиального тренда необходимо задать степень аппроксимирующего полинома, для скользящего среднего – количество точек усреднения.

В качестве дополнительной информации на диаграмме можно отобразить уравнение регрессии и значение среднеквадратического отклонения, установив соответствующие флажки на закладке Параметры (рис.3.3). Щелкните по кнопке **Ок**.

Таблица 3.8

Расчет выравненных значений T и ошибок E в аддитивной модели

t	y_t	S_i	$T+E=y_t - S_i$	T	$T+S$	$E=y_t-(T+S)$	E^2
1	6,0	0,581	5,419	5,902	6,483	-0,483	0,233
2	4,4	-1,977	6,377	6,088	4,111	0,289	0,083
3	5,0	-1,294	6,294	6,275	4,981	0,019	0,001
4	9,0	2,690	6,310	6,461	9,151	-0,151	0,023
5	7,2	0,581	6,619	6,648	7,229	-0,029	0,001
6	4,8	-1,977	6,777	6,834	4,857	-0,057	0,003
7	6,0	-1,294	7,294	7,020	5,726	0,274	0,075
8	10,0	2,690	7,310	7,207	9,897	0,103	0,011
9	8,0	0,581	7,419	7,393	7,974	0,026	0,001
10	5,6	-1,977	7,577	7,580	5,603	-0,003	0,000
11	6,4	-1,294	7,694	7,766	6,472	-0,072	0,005
12	11,0	2,690	8,310	7,952	10,642	0,358	0,128
13	9,0	0,581	8,419	8,139	8,720	0,280	0,078
14	6,6	-1,977	8,577	8,325	6,348	0,252	0,063
15	7,0	-1,294	8,294	8,512	7,218	-0,218	0,047
16	10,8	2,690	8,110	8,698	11,388	-0,588	0,346
							$\sum E^2 = 1{,}1$

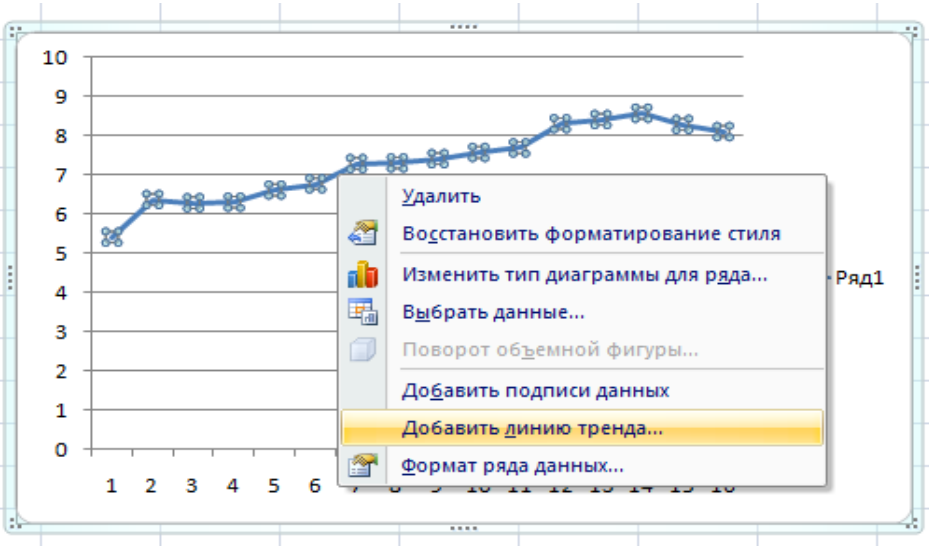


Рис.3.2. Контекстное меню Мастера диаграмм

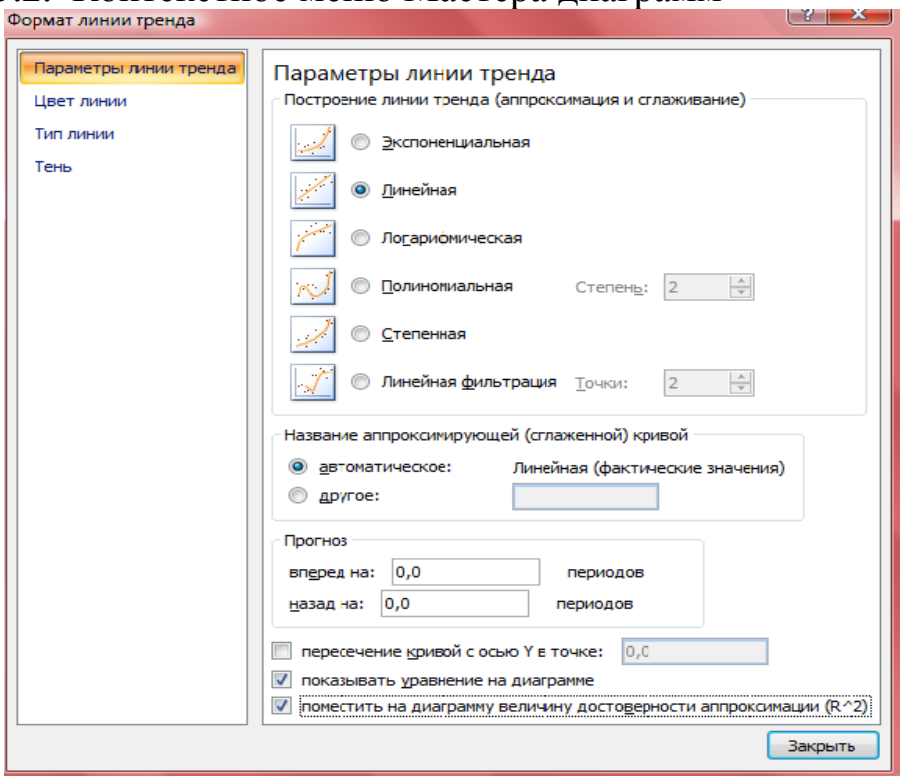


Рис.3.3. Диалоговое окно типов и параметров линий тренда

На рис.3.4 представлен тренд, описывающий данные из гр.4., табл.3.8.

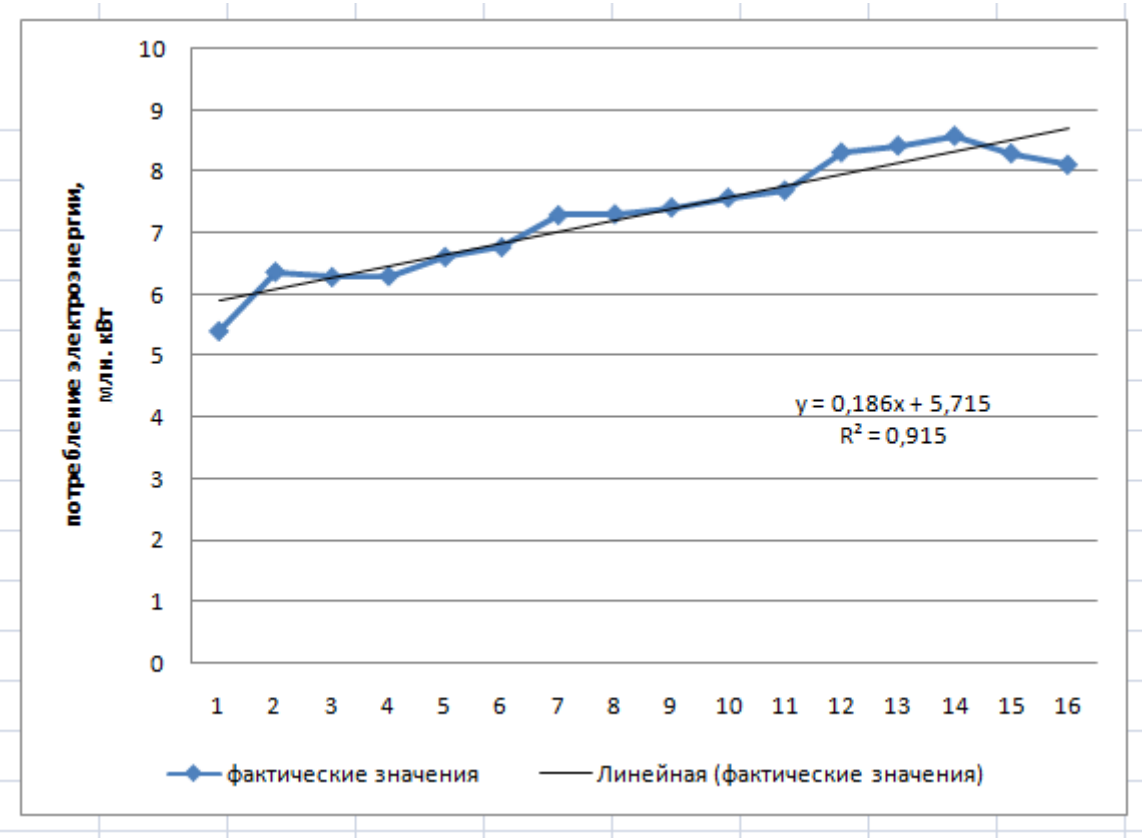


Рис.3.4. Линейный тренд

Подставим в это уравнение значения t=1,…,16, найдем уровни ряда T для каждого момента времени (гр. 5 табл.3.8).

Шаг 5. Рассчитаем значения уровней ряда, полученные по аддитивной модели. Для этого прибавим к уровням Т значения сезонной компоненты для соответствующих кварталов. Графически изображения (Т+S) представлены на рис. 3.5.

Шаг 6. Рассчитаем абсолютную ошибку $E = Y - (T + S)$.

Численные значения приведены в гр. 7 табл.3.8.

Оценим качество построенной модели. Для этого рассчитаем общую сумму квадратов отклонений уровней ряда от его среднего уровня (табл.3.8).

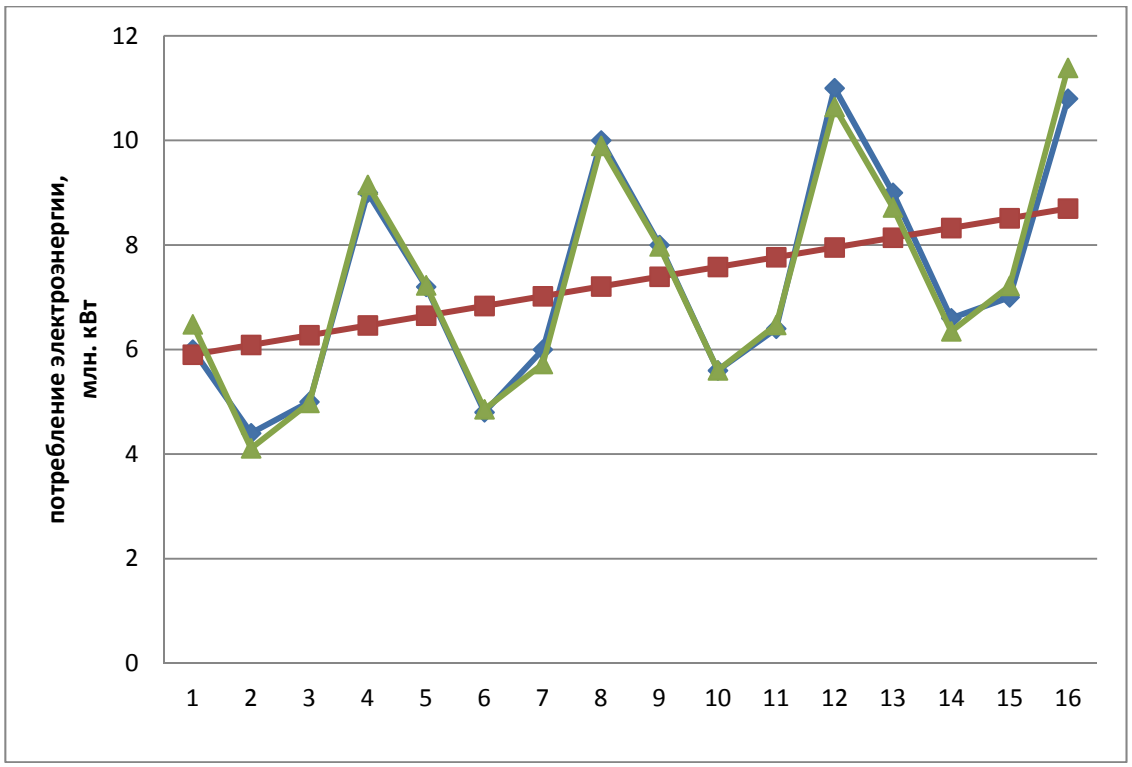


Рис. 3.5. Потребление электроэнергии жителями района (фактические, выравненные и полученные по аддитивной модели значения уровней ряда)

Таблица 3.9

Расчет качества аддитивной модели

t	y_t	$y_t - \bar{y}$	$(y_t - \bar{y})^2$
1	6,0	-1,3	1,69
2	4,4	-2,9	8,41
3	5,0	-2,3	5,29
4	9,0	1,7	2,89
5	7,2	-0,1	0,01
6	4,8	-2,5	6,25
7	6,0	-1,3	1,69
8	10,0	2,7	7,29
9	8,0	0,7	0,49
10	5,6	-1,7	2,89
11	6,4	-0,9	0,81
12	11,0	3,7	13,69
13	9,0	1,7	2,89
14	6,6	-0,7	0,49
15	7,0	-0,3	0,09
16	10,8	3,5	12,25
Сумма	116,8		67,12
Среднее значение	7,3		

Качество модели: $R^2 = \left(1 - \frac{E^2}{\sum(y_t - \bar{y})^2}\right) \cdot 100\% = \left(1 - \frac{1,1}{67,12}\right) \cdot 100\% = 98,36\%$

Следовательно, аддитивная модель объясняет 98,36% общей вариации уровней временного ряда потребления электроэнергии за последние 16 кварталов.

Прогнозирование значений уровня временного ряда для аддитивной модели.

Пронозные значения F_t уровня временного ряда в аддитивной модели это сумма трендовой и сезонной компонент.

Рассчитаем прогнозные значения потребления электроэнергии жителями региона в течение первого полугодия ближайшего следующего года.

Объем электроэнергии, потребленной в течение первого полугодия, рассчитаем как сумму объемов потребления электроэнергии в I и во II кварталах, соответственно F_{17} и F_{18}. Для определения трендовой компоненты воспользуемся уравнением тренда

$$T = 5,715 + 0,186 \cdot t.$$

Получим:

$$F_{17} = 5,715 + 0186 \cdot 17 = 8,877;$$
$$F_{18} = 5,715 + 0,186 \cdot 18 = 9,063.$$

Значения сезонной компоненты равны: S_1=0,581 (I квартал); $S_2 = -1,979$ (II квартал).

Таким образом,

$$F_{17} = T_{17} + S_1 = 8,877 + 0,581 = 9,458;$$
$$F_{18} = T_{18} + S_2 = 9,063 - 1,977 = 7,086.$$

Прогноз объема потребления электроэнергии на первое полугодие следующего (пятого) года составит:

(9,458+7,086)=16,544 млн кВт·ч.

Методику построения мультипликативной модели рассмотрим на следующем примере.

Пример 3.3. Имеются поквартальные данные о прибыли компании за последние четыре года табл. 3.10.

График данного временного ряда (рис.3.6) свидетельствует о наличии сезонных колебаний (период равен 4) и общей убывающей тенденции уровней ряда. Прибыль компании в весенне-летний период выше, чем осенне-зимний период. Поскольку амплитуда сезонных колебаний уменьшается, можно предположить существование мультипликативной модели. Определим ее компоненту.

Таблица 3.10

Расчет оценок сезонной компоненты в мультипликативной модели

Номер квартала t	Прибыль компании y_t	Итого за четыре квартала	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной компоненты
1	75	-	-	-	-
2	110	337	84,25	-	-
3	90	332	83	83,625	1,076233
4	62	314	78,5	80,75	0,767802
5	70	304	76	77,25	0,906149
6	92	298	74,5	75,25	1,222591
7	80	290	72,5	73,5	1,088435
8	56	278	69,5	71	0,788732
9	62	264	66	67,75	0,915129
10	80	256	64	65	1,230769
11	66	244	61	62,5	1,056
12	48	224	56	58,5	0,820513
13	50	206	51,5	53,75	0,930233
14	60	190	47,5	49,5	1,212121
15	48	-	-	-	-
16	32	-	-	-	-

Рис.3.6. Прибыль компании

Шаг 1. Проведем выравнивание исходных уровней ряда методом скользящей средней. Методика, применяемая на этом шаге, совпадает с методикой аддитивной модели.

Шаг 2. Оценки сезонной компоненты находятся как частное от деления фактических уровней ряда на центрированные скользящие средние. Взаимопогашаемость сезонных воздействий в мультипликативной модели выражается в том, что сумма значений

сезонной компоненты по всем кварталам должна быть равна числу периодов в цикле т.е. четырем (табл.3.10).

Для данной модели имеем:

0,91717+1,221827+1,073556+0,792349=4,004903

Определим корректирующий коэффициент:

$k = \frac{4}{4{,}004903}$=0,998776

Таблица 3.11

Расчет значений сезонной компоненты в мультипликативной модели

Показатель	Год	Номер квартала, i			
		I	II	III	IV
	1	-	-	1,076233	0,767802
	2	0,906149	1,222591	1,088435	0,788732
	3	0,915129	1,230769	1,056000	0,820513
	4	0,930233	1,212121	-	-
Итого за i-тый квартал (за все годы)		2,751511	3,665482	3,220669	2,377047
Средняя оценка сезонной компоненты для за i-того квартала, $\bar{S}_i$		0,91717	1,221827	1,073556	0,792349
Скорректированная сезонная компонента S_i		0,916047	1,220332	1,072242	0,791379

Скорректированные значения сезонной компоненты находят умножив ее средние значения на корректирующий коэффициент:

$$S_i = \bar{S}_i \cdot k, \text{ где i=1:4.} \qquad (3.29)$$

Проверим условие равенства четырем суммы значений сезонной компоненты:

0,916047+1,220332+1,072242+0,791379=4.

Получены следующие значения сезонной компоненты:

I квартал: S_1= 0,916047; II квартал: $S_2 = 1{,}220332$;

III квартал: $S_3 = 1{,}072242$; IV квартал: $S_4 = 0{,}791379$.

Шаг 3. Исключим влияние сезонной компоненты из каждого уровня временного ряда. Для этого разделим каждый уровень временного ряда на соответствующие значения сезонной компоненты. Получим: T·E=Y/S (гр.4 табл. 3.12), которые содержат только тенденцию и случайную компоненту.

Шаг 4. Для определения компоненты Т проведем аналитическое выравнивание ряда (Т·Е) с помощью линейного тренда. Уравнение тренда имеет вид:

$$T = 94{,}65 - 3{,}19 \cdot t.$$

Подставим в это уравнение значения t=1,…,16, найдем уровни ряда Т для каждого момента времени (гр. 5 табл. 3.12).

Шаг 5. Рассчитаем значения уровней ряда, полученные по мультипликативной модели. Для этого умножим уровни Т на значения сезонной компоненты для соответствующих кварталов. Графически изображения (Т·S) представлены на рис. 3.12.

Таблица 3.12

Расчет выравненных значений Т и ошибок Е в мультипликативной модели

t	y_t	S_i	Т·Е=y_t / S_i	Т	Т·S	Е=$y_t - (T \cdot S)$	E^2
1	2	3	4	5	6	7	8
1	75	0,916047	81,87349	91,46324	83,78466	-8,78466	77,17027
2	110	1, 220332	90,13944	88,27647	107,7266	2,273438	5,168521
3	90	1,072242	83,93628	85,08971	91,23675	-1,23675	1,529561
4	62	0,791379	78,34425	81,90294	64,81627	-2,81627	7,93139
5	70	0,916047	76,41526	78,71618	72,10775	-2,10775	4,442612
6	92	1,220332	75,38935	75,52941	92,17092	-0,17092	0,029215
7	80	1,072242	74,61002	72,34265	77,56882	2,431177	5,910623
8	56	0,791379	70,76255	69,15588	54,72852	1,271483	1,616669
9	62	0,916047	67,68209	65,96912	60,43084	1,56916	2,462263
10	80	1,220332	65,55595	62,78235	76,61529	3,384714	11,45629
11	66	1,072242	61,55327	59,59559	63,90089	2,099109	4,406258
12	48	0,791379	60,65361	56,40882	44,64076	3,359238	11,28448
13	50	0,916047	54,58233	53,22206	48,75393	1,24607	1,552691
14	60	1,220332	49,16697	50,03529	61,05965	-1,05965	1,122854
15	48	1,072242	44,76601	46,84853	50,23296	-2,23296	4,986109
16	32	0,791379	40,43574	43,66176	34,55301	-2,55301	6,51784

Рис. 3.7. Прибыль компании (фактические и выравненные по мультипликативной модели значения уровней ряда)

Шаг 6. Рассчитаем абсолютную ошибку $E = Y - (T \cdot S)$.

Численные значения приведены в гр. 7 табл.3.12.

Оценим качество построенной модели. Для этого рассчитаем общую сумму квадратов отклонений уровней ряда от его среднего уровня.

Таблица 3.13

Расчет качества мультипликативной модели

t	y_t	$y_t - \bar{y}$	$(y_t - \bar{y})^2$
1	75	7,4375	55,31641
2	110	42,4375	1800,941
3	90	22,4375	503,4414
4	62	-5,5625	30,94141
5	70	2,4375	5,941406
6	92	24,4375	597,1914
7	80	12,4375	154,6914
8	56	-11,5625	133,6914
9	62	-5,5625	30,94141
10	80	12,4375	154,6914
11	66	-1,5625	2,441406
12	48	-19,5625	382,6914
13	50	-17,5625	308,4414
14	60	-7,5625	57,19141
15	48	-19,5625	382,6914
16	32	-35,5625	1264,691
Сумма	1081		5865,938
Среднее значение	67,56		

Сумма квадратов остатков равна 147,5876.

Качество модели: $R^2 = \left(1 - \frac{E^2}{\sum(y_t - \bar{y})^2}\right) \cdot 100\% = \left(1 - \frac{147,5876}{5865,938}\right) \cdot 100\% = 97,48\%$

Следовательно, мультипликативная модель объясняет 97,48% общей вариации уровней временного ряда прибыли компании за последние 16 кварталов.

Прогнозирование значений уровня временного ряда для мультипликативной модели

Пронозные значения F_t уровня временного ряда в мультипликативной модели это произведение трендовой и сезонной компонент.

Рассчитаем прогнозные значения ожидаемой прибыли компании в течение первого полугодия ближайшего следующего года.

Прибыль компании в течение первого полугодия равна $F_{17} + F_{18}$. Для определения трендовой компоненты воспользуемся уравнением тренда

$$T = 94,65 - 3,19 \cdot t.$$

Получим:

$$F_{17} = 94,65 - 3,19 \cdot 17 = 40,42;$$
$$F_{18} = 94,65 - 3,19 \cdot 18 = 37,23.$$

Значения сезонной компоненты равны: S_1= 0,916047 (I квартал); $S_2 = 1,220332$ (II квартал)

Таким образом,

$$F_{17} = T_{17} * S_1 = 40,420 * 0,916 = 37,02;$$
$$F_{18} = T_{18} * S_2 = 37,23 * 1,22 = 45,42.$$

Прогноз ожидаемой прибыли на первое полугодие следующего (пятого) года составит:

(37,02+45,42)=82,44тыс. долл. США.

Глава 4. Адаптивные методы прогнозирования

При прогнозировании сложных систем наиболее важным является последний период ее функционирования, а не тенденции, сложившиеся в среднем на всем периоде предыстории. В таких случаях значимость свойства динамичности развития экономических систем должна преобладать над значимостью свойства инерционности. Поэтому более эффективными оказываются методы и модели, в которых значимость уровней временного ряда убывает по мере их удаления от прогнозируемого периода. Для повышения качества прогнозов на основе таких методов и моделей в их алгоритмах, как правило, предусмотрены процедура постоянного сопоставления прогнозных оценок, рассчитанных по модели с фактическими данными, и корректировка параметров модели с учетом полученных расхождений. Практически все существующие методы прогнозирования в разной степени тем или иным способом реализуют такие сопоставления, т.е. приспосабливают модель к новой информации, характеризующей фактическое развитие процесса. Такое приспособление называют адаптацией.

В наибольшей степени в процедурах экономического прогнозирования процесс приспособления реализуются в специальных методах, называются адаптивными. К ним относятся, прежде всего, следующие методы: экспоненциального сглаживания, гармонических весов, Хольта-Уинтерса.

Адаптивные методы позволяют строить саморегулирующиеся модели, которые, учитывая результат прогноза, сделанного на предыдущем шаге, и различную ценность членов временного ряда, способны оперативно реагировать на изменяющиеся условия.

4.1. Экспоненциальное сглаживание Брауна

Простое экспоненциальное сглаживание

Ранее было рассмотрено экспоненциальное сглаживание, которое осуществляется по формуле:

$$S_t = \alpha x_t + (1-\alpha) S_{t-1} \qquad (4.1)$$

где S_t – значение экспоненциальной средней в момент t; S_{t-1} – значение экспоненциальной средней в момент t-1; α- параметр сглаживания. Используя его мы можем получить модель дя прогнозирования.

Пусть дан динамический ряд, который генерируется процессом, не содержащим тренда: $y_t=a_t+\varepsilon_t$, где a_t – изменяющийся во времени средний уровень ряда; ε_t – случайные неавтокоррелированные величины.

Прогнозная модель имеет вид

$$\hat{y}_t(\tau) = \hat{a}_t, \qquad (4.2)$$

где $\hat{y}_t(\tau)$ - прогноз, сделанный в момент t на τ шагов вперед; $\hat{a}_t$ - оценка a_t.

Экспоненциальная средняя $S_t = \hat{a}_t$ служит средством оценки единственного параметра модели.

Интервальный прогноз, проведенный методом простого экспоненциального сглаживания можно оценить следующим образом:

$$\hat{y}_n(1) \pm t_q S \sqrt{\frac{2}{2-\alpha}}.$$

Приведение модели Брауна к виду (4.1) позволяет определить процедуру многократного экспоненциального сглаживания. Процедура многократного экспоненциального сглаживания фактически является применением простого экспоненциального сглаживания к результатам сглаживания порядка *p-1*. Ее можно записать так:

$$S_t^{[p]} = \alpha S_t^{p-1} + (1-\alpha) S_{t-1}^{[p]},$$

где $S_t^{[0]} = y_t$, *p = 1, 2, ..., n* – порядок сглаживания, $S_0, S_0^{[2]}, ..., S_0^{[n]}$ - начальные значения экспоненциальных средних соответствующего порядка.

Линейное экспоненциальное сглаживание

Пусть модель сглаживающего прогноза на основе модели Брауна имеет вид:

$$\hat{y}_t(\tau) = \hat{a}_t + \hat{b}_t \tau + \varepsilon_t \qquad (4.3),$$

а начальные условия для сглаживающего полинома определены как:

$$S_0 = a_0 - \frac{1-\alpha}{\alpha} b_0 \quad S_0^{[2]} = a_0 - \frac{2(1-\alpha)}{\alpha} b_0. \qquad (4.4)$$

Для того чтобы выразить коэффициенты a_0 и b_0 необходимо воспользоваться коэффициентами уравнения тренда $y_t = a_0 + b_0 t$, полученными методом наименьших квадратов.

Тогда экспоненциальные средние моделей первого и второго порядков могут быть оценены как:

$$S_t = \alpha y_t + (1-\alpha) S_{t-1}, \; S_t^{[2]} = \alpha S_t + (1-\alpha) S_{t-1}^{[2]}. \qquad (4.5)$$

Оценки параметров коэффициентов модели (3) составят:

$$\hat{a}_t = 2S_t - S_t^{[2]}, \; \hat{b}_t = \frac{\alpha}{1-\alpha}(S_t - S_t^{[2]}). \qquad (4.6)$$

Окончательно точечный прогноз по модели экспоненциального среднего первого порядка на момент времени t:

$$\hat{y}_t(\tau) = \hat{a}_t + \tau\hat{b}_t = (2 + \frac{\alpha}{1-\alpha}\tau)S_t - (1 + \frac{\alpha}{1-\alpha}\tau)S_t^{[2]}. \qquad (4.7)$$

Модельная дисперсия находится по формуле:

$$s_{\hat{y}} = s_y\sqrt{\frac{\alpha}{(2-\alpha)}\left[1 - 4(1-\alpha) + 5(1-\alpha)^2 + 2\alpha(4-3\alpha)t + 2\alpha^2t^2\right]}, \qquad (4.8)$$

где s_y - среднеквадратическая ошибка отклонения от линейного тренда, которую определяем из формулы: $s_y = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n-2}}$.

Квадратичное экспоненциальное сглаживание

Пусть модель сглаживания прогноза по модели Брауна имеет вид:

$$\hat{y}_t = a_t + b_t\tau + \frac{1}{2}c_t\tau^2 + \varepsilon_t, \qquad (4.9)$$

а начальные условия для сглаживающего полинома заданы следующим образом:

$$\begin{aligned} S_0 &= a_0 - \frac{1-\alpha}{\alpha}b_0 + \frac{(1-\alpha)(2-\alpha)}{2\alpha^2}c_0, \\ S_0^{[2]} &= a_0 - \frac{2(1-\alpha)}{\alpha}b_0 + \frac{(1-\alpha)(3-2\alpha)}{\alpha^2}c_0, \\ S_0^{[3]} &= a_0 - \frac{3(1-\alpha)}{\alpha}b_0 + \frac{3(1-\alpha)(4-3\alpha)}{2\alpha^2}c_0. \end{aligned} \qquad (4.10)$$

Тогда экспоненциальные средние первого, второго и третьего порядков могут быть подсчитаны по следующим формулам:

$$S_t = \alpha y_t + (1-\alpha)S_{t-1},\ S_t^{[2]} = \alpha S_t + (1-\alpha)S_{t-1}^{[2]},\ S_t^{[3]} = \alpha S_t^{[2]} + (1-\alpha)S_{t-1}^{[3]},$$

а оценки коэффициентов модели могут быть оценены из следующих соотношений:

$$\begin{aligned} \hat{a}_t &= 3S_t - 3S_t^{[2]} + S_t^{[3]}, \\ \hat{b}_t &= \frac{\alpha}{2(1-\alpha)^2}\left[(6-5\alpha)S_t - 2(5-4\alpha)S_t^{[2]} + (4-3\alpha)S_t^{[3]}\right], \\ \hat{c}_t &= \frac{\alpha}{(1-\alpha)^2}\left[S_t - 2S_t^{[2]} + S_t^{[3]}\right]. \end{aligned} \qquad (4.11)$$

Окончательно точечный прогноз по модели экспоненциального среднего второго порядка на момент времени t:

$$\begin{aligned} \hat{y}_t(\tau) = \hat{a}_t\tau + \tau\hat{b}_t + \frac{1}{2}\tau^2\hat{c}_t = \left[6(1-\alpha)^2 + (6-5\alpha)\alpha\tau + \alpha^2\tau^2\right]\frac{S_t}{2(1-\alpha)^2} - \\ -\left[6(1-\alpha)^2 + 2(5-4\alpha)\alpha\tau + 2\alpha^2\tau^2\right]\frac{S_t^{[2]}}{2(1-\alpha)^2} + \left[2(1-\alpha)^2 + (4-3\alpha)\alpha\tau + \alpha^2\tau^2\right] \end{aligned}. \qquad (4.12)$$

Ошибка модели прогноза находится по формуле:

$$s_{\hat{y}} = s_y \sqrt{2\alpha + 3\alpha^2 + 3\alpha^3 t}, \qquad (4.13)$$

где s_y- среднеквадратическая ошибка отклонения от квадратичного тренда, которую определяем по формуле: $s_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-3}}$, где n – количество членов в исследуемом ряду.

Продемонстрируем применение метода экспоненциального сглаживания ряда y_t, структура которого описывается моделью (4.3) рис.4.1.

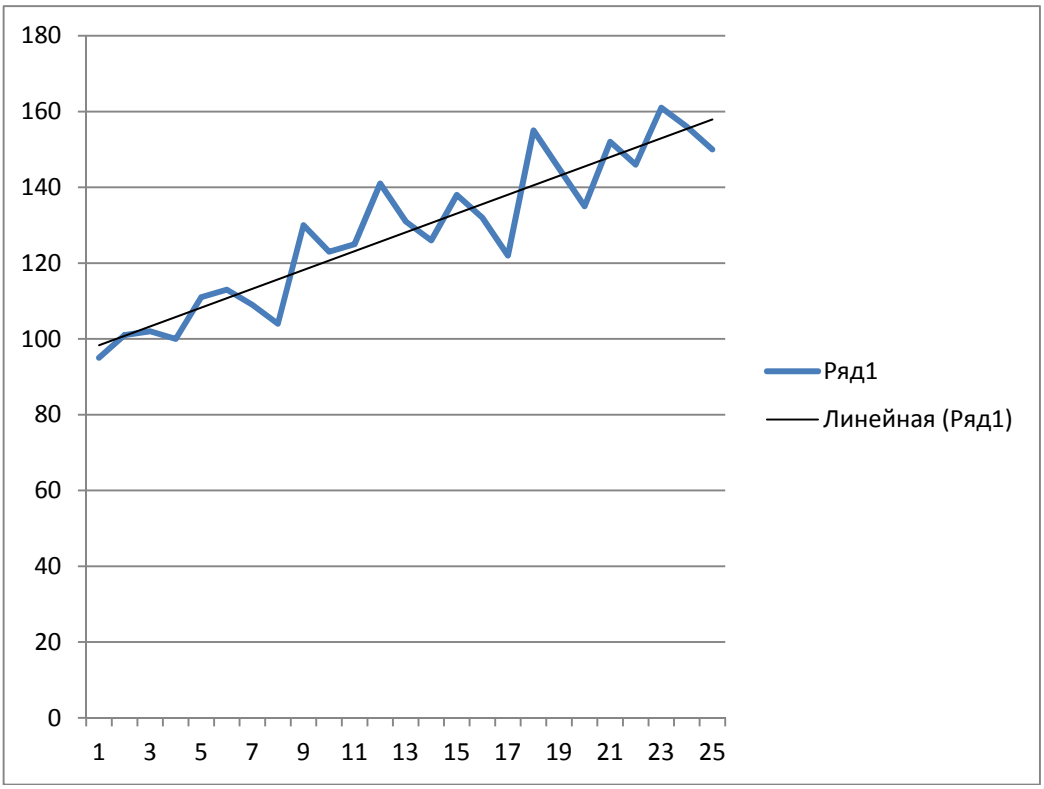


Рис. 4.1. Аппроксимация эмпирического временного ряда линейной функцией

Анализ графика свидетельствует о наличии тренда, который в соответствии с критерием минимума дисперсии будет лучшим образом отражаться линейной моделью. В табл. 1 приведены результаты такого моделирования. Начальные оценки a_0=90,24 и b_0=3,48 получены МНК по первым 12 уровням ряда. Тот же результат можно получить используя мастер-диаграмму, ППС Excel. Используя *Поиск решения* ППС Excel определили значение параметра сглаживания, α=0,1273, при котором $s_y = 9{,}1$.

Таблица 4.1

Расчет параметров линейной модели

t	y_t	a_t	b_t	$S_t^{(1)}$	$S_t^{(2)}$	$\hat{y}_t$	ε_t
0	-	90,24	3,48	66,43	42,61	-	-
1	95	94,02	3,50	70,07	46,11	93,72	1,28
2	101	98,35	3,55	74,01	49,66	97,52	3,48
3	102	101,93	3,55	77,57	53,22	101,90	0,10
4	100	104,17	3,47	80,43	56,68	105,48	-5,48
5	111	108,44	3,52	84,32	60,20	107,64	3,36
6	113	112,21	3,54	87,97	63,74	111,96	1,04
7	109	114,14	3,43	90,65	67,17	115,75	-6,75
8	104	114,33	3,21	92,35	70,37	117,56	-13,56
9	130	120,51	3,41	97,15	73,78	117,54	12,46
10	123	123,70	3,39	100,44	77,18	123,92	-0,92
11	125	126,59	3,36	103,57	80,54	127,09	-2,09
12	141	132,59	3,54	108,33	84,08	129,96	11,04
13	131	134,91	3,46	111,22	87,54	136,13	-5,13
14	126	135,41	3,26	113,10	90,79	138,36	-12,36
15	138	138,51	3,25	116,27	94,04	138,67	-0,67
16	132	139,43	3,09	118,28	97,12	141,76	-9,76
17	122	137,62	2,75	118,75	99,88	142,52	-20,52
18	155	143,86	2,99	123,37	102,87	140,38	14,62
19	145	146,41	2,96	126,12	105,83	146,86	-1,86
20	135	145,95	2,73	127,25	108,56	149,37	-14,37
21	152	149,47	2,78	130,41	111,34	148,67	3,33
22	146	150,76	2,68	132,39	114,02	152,25	-6,25
23	161	155,24	2,80	136,03	116,83	153,44	7,56
24	156	157,56	2,77	138,58	119,60	158,05	-2,05
25	150	157,87	2,60	140,03	122,20	160,33	-10,33

В результате имеем прогнозную модель $\hat{y}_{25}(\tau) = 157{,}87 + 2{,}60\tau$.

При статистической оценке адекватности модели необходимо проверить, что для ряда остатков ε_t характерны: равенство нулю математического ожидания, случайность отклонений от математического ожидания, независимость уровней ряда остатков и нормальный закон распределения.

4.2. Модель Брауна и Хольта

С развитием экспоненциального сглаживания стали появляться новые модели, основанные на тех же принципах адаптации, что и модели экспоненциального сглаживания. За исходную гипотезу построения моделей Брауна и Хольта берется представление о том, что имеется не только медленно развивающийся местный уровень, но также и тенденцию с медленно развивающимся наклоном.

Для этой ситуации Ч. Хольтом была предложена модель, в которой прогноз осуществляется путем экстраполяции тенденции линейным трендом на τ тактов вперед:

$$\hat{y}_t(\tau) = \hat{a}_t + \tau\hat{b}_t . \qquad (4.14)$$

Простейшим способом обновления коэффициентов $\hat{a}_t$, $\hat{b}_t$ в каждый момент времени является простое экспоненциальное сглаживание. Поэтому параметры линейной модели Брауна корректируются следующим образом:

$$\begin{aligned} \hat{a}_t &= \hat{a}_{t-1} + \hat{b}_{t-1}\cdot 1 + \left(1-\beta^2\right)\cdot\varepsilon_t ; \\ \hat{b}_t &= \hat{b}_{t-1} + \left(1-\beta\right)^2\varepsilon_t , \end{aligned} \qquad (4.15)$$

где $\varepsilon_t = y - \hat{y}_t$ – ошибка прогнозирования.

Более общей формой является линейная модель Хольта, в которой оценка коэффициентов осуществляется на основе двух параметров сглаживания. Адаптация данных параметров линейного тренда проводится по следующим формулам:

$$\begin{aligned} \hat{a}_t &= \alpha_1 y(t) + (1-\alpha_1)\left(\hat{a}_{t-1} + \hat{b}_{t-1}\right) \\ \hat{b}_t &= \alpha_2\left(\hat{a}_t - \hat{a}_{t-1}\right) + (1-\alpha_2)\hat{b}_{t-1} . \end{aligned} \qquad (4.16)$$

Начальные уровни процедуры сглаживания также рекомендуется подбирать эмпирическим путем.

4.3. Модель Уинтерса

Для прогнозирования сезонных процессов широко используется модель Уинтерса . Мультипликативная модель Уинтерса с линейным ростом опирается на экспоненциальную схему и для шага прогнозирования, равного единице имеет вид:

$$\hat{y}_t(1) = (\hat{a}_t + \hat{b}_t) \cdot F_{t-L+1};$$
$$\hat{a}_t = \alpha_1 \frac{y_t}{F_{t-L}} + (1-\alpha_1)(\hat{a}_{t-1} + \hat{b}_{t-1});$$
$$\hat{b}_t = \alpha_3 (\hat{a}_t - \hat{a}_{t-1}) + (1-\alpha_3)\hat{b}_{t-1}; \qquad (4.17)$$
$$F_t = \alpha_2 \frac{y_t}{\hat{a}_t} + (1-\alpha_2) F_{t-L}.$$

Здесь L – период сезонности (так, для квартальных данных L=4, для месячных L=12); F_{t-L+1} – значение коэффициента сезонности, которое приписывают моменту (t+1), а вычисляют сезон назад, т.е. в момент ($t+1-L$); $\hat{a}_t$, $\hat{b}_t$ – параметры, имеющие тот же смысл, что и в моделях Брауна и Хольта; четвертое равенство задает правило обновления коэффициентов сезонности на последующий сезон.

Следует обратить внимание, что при расчете $\hat{y}_t(1) = (\hat{a}_t + \hat{b}_t) \cdot F_{t-L+1}$, например в момент времени t=1 и L=4, мы должны иметь $F_{0-4+1} = F_{-3}$. Очевидно, что F_{-3} следует понимать как коэффициент сезонности, относящийся к первому кварталу года, предыдущего к первому году ряда наблюдений, так сказать к предыстории анализируемого процесса. Аналогичный смысл имеют исходные коэффициенты сезонности F_{-2}, F_{-1} и F_0.

Следует отметить, что при относительном постоянстве амплитуды сезонной волны целесообразно использовать аддитивную модель Тейла-Вейджа, которая будет рассмотрена далее.

В табл. 4.2 приведены результаты построения трехпараметрической мультипликативной модели для 20 точек эмпирического ряда при τ=1.

Таблица 4.2

Построение модели Уинтерса

t	y_t	a_t	b_t	F_t	$\hat{y}_t$	ε_t
-3	-	-	-	0,5838		
-2	-	-	-	1,0775		
-1	-	-	-	1,8265		
0	-	200,65	5,69	0,5113		
1	124	206,95	5,75	0,5961	120,46	-3,54
2	236,5	213,38	5,82	1,1022	229,18	-7,32
3	409	219,67	5,87	1,8548	400,36	-8,64
4	115	225,47	5,86	0,5103	115,32	0,32
5	129,8	229,99	5,73	0,5708	137,80	8,00
6	244,1	234,31	5,58	1,0540	259,60	15,50

7	426,7	238,92	5,49	1,8000	444,72	18,02
8	125	244,46	5,49	0,5112	124,73	-0,27
9	133,2	248,26	5,32	0,5434	142,85	9,65
10	251,3	252,04	5,17	1,0084	267,63	16,33
11	498,6	259,17	5,36	1,8988	463,35	-35,25
12	148,2	267,07	5,62	0,5461	135,20	-13,00
13	144,5	271,96	5,55	0,5336	148,48	3,98
14	273,5	276,83	5,48	0,9916	280,38	6,88
15	535,2	282,30	5,48	1,8959	535,23	0,03
16	152,4	286,97	5,40	0,5340	156,82	4,42
17	157,5	292,61	5,42	0,5372	156,22	-1,28
18	286,8	297,10	5,33	0,9703	295,97	9,17
19	588,2	303,22	5,41	1,9305	573,37	-14,83
20	162,9	308,25	5,37	0,5294	164,87	1,97

Начальные оценки a_0=200,65 и b_0=5,69, взяты из достаточно грубой модели $y_t^{л}$= 200,65+5,69t, построенной по первым 8 точкам. Оценка коэффициентов “предыстории”, необходимых для вычисления коэффициентов сезонности первого года “истории”, осуществляется делением первых восьми фактических уровней временного ряда на их расчетные значения, вычисленные по линейной модели с последующим усреднением по одноименным кварталам:

$$F_{-3}=\frac{1}{2}\left(\frac{y_1}{y_1^{л}}+\frac{y_5}{y_5^{л}}\right);\quad F_{-2}=\frac{1}{2}\left(\frac{y_2}{y_2^{л}}+\frac{y_6}{y_6^{л}}\right);$$

$$F_{-1}=\frac{1}{2}\left(\frac{y_3}{y_3^{л}}+\frac{y_7}{y_7^{л}}\right);\quad F_{0}=\frac{1}{2}\left(\frac{y_4}{y_4^{л}}+\frac{y_8}{y_8^{л}}\right).$$

Корректировка параметров модели проводилась при α_1=0,1, α_2=0,77, α_3=0,1.

На последнем шаге имеем прогнозную модель

$$\hat{y}_t(\tau)=(308{,}25+5{,}37\cdot\tau)\cdot F_{20-4+\tau},$$

где F принимает значения 0,5372; 0,9703; 1,9394; 0,5294 соответственно при τ=1, 2, 3, 4.

Динамический фильтр

Практика использования мультипликативных моделей показывает, что в случае, когда колебания исследуемого процесса велики, не всегда получаются хорошие результаты. Причина этого кроется в том, что текущая оценка $\hat{a}_t$ формируется на основе двух взвешенных в

соответствии со значениями параметра сглаживания α_1 компонентов – очищенного от сезонных колебаний фактического уровня в данный момент и его значения в предыдущий период времени. Учитывая, что истинное значение уровня всегда определяется с некоторой ошибкой, первую компоненту можно представить в виде: $\frac{y_t}{F_{t-L}} = \frac{\hat{y}_t + \varepsilon_t}{F_{t-L}} = \frac{\hat{y}_t}{F_{t-L}} + \frac{\varepsilon_t}{F_{t-L}}$.

Очевидно, что при малых значениях сезонных коэффициентов влияние случайного колебания уровня процесса может быть велико, причем оно становится тем сильнее, чем ближе значение этих коэффициентов к нулю. Модель, полученная в этот момент, будет заведомо неадекватной процессу и потому не может быть использована для прогнозирования. Для устранения мультипликативного эффекта случайной ошибки используют параметр сглаживания уровня процесса α_{1t}, сформированного на основе динамического фильтра (фильтра Левандовского): $\alpha_{1t} = \alpha_1 \alpha_t(\tau_t)$, где $\alpha_t(\tau_t)$ – значение дополнительного параметра сглаживания, учитывающего структуру сезонных коэффициентов в зависимости от аргумента $\tau_t = \frac{F_t}{F_{\max}}$.

Дополнительный параметр сглаживания принимает значение, равное единице, если существенных колебаний в структуре сезонных коэффициентов нет. В этом случае первоначальное значение параметра сглаживания α_{1t} сохраняется и используется для корректировки параметров модели.

Когда сезонные коэффициенты F_t отличаются от наибольшего из них $F_{\max}$, например, более чем в 4 раза, т.е. $\tau_t = \frac{F_t}{F_{\max}} < 0{,}25$, величина $\alpha_t(\tau_t)$ принимает значения, меньшие единицы. В простейшем случае это может быть обеспечено использованием прямой, проходящей через две точки с координатами (0;0) и (0,25; 1):

$$\alpha_t(\tau_t) = \begin{cases} 1, \text{ где } \tau_t \geq 0{,}25 \\ 4\tau, \text{ где } 0 \leq \tau_t \leq 0{,}25. \end{cases} \qquad (4.18)$$

Таким образом, параметр $\alpha_t(\tau_t)$ будет равен единице в случае обработки обычных временных рядов и стремиться, например линейно, к нулю при прогнозировании рядов с сильной сезонностью.

Использование динамического фильтра в мультипликативной модели обычно приводит к уменьшению средней относительной ошибки, что свидетельствует об эффективности данного подхода и целесообразности

его использования в исследованиях процессов с относительно сильной сезонностью.

4.4. Метод гармонических весов

В методе гармонических весов, который разработан Хельдвигом, идея дисконтирования данных реализована иначе, чес в рассмотренных методах. Параметр прироста линейной модели, используемой для прогнозирования, находится путем взвешивания на основе гармонических весов приростов сглаженного ряда. Сглаживание осуществляется линейной функцией для каждого из (N-k+1) перекрывающихся сегментов одинаковой длины k.

Метод гармонических весов базируется на следующих предпосылках:

1) период времени, за который изучается экономический процесс, должен быть достаточно длительным, чтобы можно было определить его закономерность;

2) исходный ряд динамики не должен иметь скачкообразных изменений;

3) прогнозируемое социально-экономическое явление должно обладать инерционностью, т.е. для наступления большого изменения в характеристиках процесса необходимо, чтобы прошло значительное время;

4) отклонения от скользящего тренда носят случайный характер;

5) автокорреляционная функция, рассчитанная на основе последовательных разностей, должна уменьшаться с ростом t, т.е. влияние более поздней информации должно сильнее отражаться на прогнозируемой величине, чем на ранней информации. Для получения точного прогноза методом гармонических весов необходимо выполнение всех вышеуказанных предпосылок для исходного ряда динамики.

Для осуществления прогноза данным методом исходный ряд динамики разбивается на фазы k. Число фаз должно быть меньше числа членов ряда n, т.е. $k<n$. Обычно фаза равна 3-5 уровням. Для каждой фазы рассчитывается линейный тренд, т.е.

$$y_i(t) = a_i + b_i t \ \ (i=1,2,\ldots,n-k+1). \qquad (4.20)$$

Коэффициенты всех моделей скользящего тренда определяются с помощью метода наименьших квадратов или по формулам:

$$b_i = \frac{\sum y_i t - \frac{1}{k}\sum y_i \sum t}{\sum t^2 - \frac{1}{k}\left(\sum t\right)^2}; \qquad (4.21)$$

$$a_i = \frac{1}{k}\sum y_i - \frac{1}{k} b_i \sum t.$$

Здесь суммирование по t производится от i до (i+k-1). После получения всех (n-k+1) оценок определяются сглаженные значения ряда:

$$\hat{y}_t = \frac{1}{t}\sum_{i=1}^{t} y_i(t), \quad t < k;$$

$$\hat{y}_t = \frac{1}{k}\sum_{i=t-k+1}^{t} y_i(t), \quad k \le t \le n-k+1; \qquad (4.22)$$

$$\hat{y}_t = \frac{1}{n-k+1}\sum_{i=t-k+1}^{n-k+1} y_i(t), \quad t > n-k+1.$$

После этого необходимо проверить гипотезу о том, что отклонения от тренда представляют собой стационарный процесс. С этой целью рассчитывается автокорреляционная функция. Если значения автокорреляционной функции уменьшаются от периода к периоду, то пятая предпосылка данного метода выполняется.

Предположим, что приросты сглаженного ряда являются случайной величиной, для которой оценкой математического ожидания является средний уровень $\bar{p}$ с дисперсией $S_{\bar{p}}^2$. В этом случае средний прирост используется для получения прогноза на 1, 2 и более шагов вперед, прибавляя его к уровню процесса, в качестве которого можно принять последний уровень сглаженного ряда. Учитывая информационную неравноценность данных, Хельвиг предложил использовать гармоническую среднюю вида $\bar{p} = \sum_{t=1}^{n-1} C_{t+1}^n p_{t+1}; \quad S_{\bar{p}}^2 = \sum_{t=1}^{n-1} C_{t+1}^n (p_{t+1} - \bar{p})^2,$ где $p_{t+1} = \hat{y}_{t+1} - \hat{y}_t$ – приросты сглаженного ряда; $C_{t+1}^n = \frac{1}{n-1}\sum_{i=1}^{t}\frac{1}{n-i}$ – гармонические коэффициенты, удовлетворяющие следующим условиям:

$$C_{t+1}^n > 0, \qquad \sum_{t=1}^{n-1} C_{t"+1}^n = 1. \qquad (4.23)$$

Данное выражение позволяет более поздней информации придавать большие веса, так как приросты обратно пропорциональны времени, которое отделяет раннюю информацию от поздней для момента t=n.

Прогноз на τ шагов вперед получается по формуле:

$$\hat{y}_t = a + b\cdot(t-n) \quad (t = n+1,\ldots,n+\tau), \text{ где } a = \hat{x}_n; b = \bar{p} \quad (4.24)$$

Доверительные интервалы рассчитываются по формуле:

$$\hat{y}_t \pm S_{\bar{p}} d\,, \qquad (4.25)$$

где d – целое положительное число, задаваемое в интервале от 2 до 4.

Получим прогнозные оценки на основе этого метода для следующего показателя, используя первые 25 уровней временного ряда.

Автоковариационная функция ряда убывает, что свидетельствует о большой зависимости последних уровней, т.е. о целесообразности дисконтирования данных. Выберем длину сегмента в пять уровней, т.е. *к*=5. Значит, надо построить *25-5+1=21* уравнение.

В табл.4.3 приведены результаты оценки параметров этих уравнений (a_i, b_i =*1,...21*), сглаженные уровни, их приросты и веса.

На примере одного, пусть шестого (*i*=6) уравнения $y_6(t) = a_6 + b_6 t$ (5<t<9) покажем технологию расчета:

$$b_6 = \frac{\sum_{t=6}^{10} y_t t - \frac{1}{k}\sum_{t=6}^{10} t \sum_{t=6}^{10} y_t}{\sum_{t=6}^{10} t^2 - \frac{1}{k}\left(\sum_{t=6}^{10} t\right)^2} = \frac{4673 - \frac{1}{5}\cdot 40 \cdot 579}{330 - \frac{1}{5}\cdot 1600} = 4.1;$$

$$a_6 = \frac{1}{k}\sum_{t=6}^{10} y_t - \frac{1}{k} b_6 \sum_{t=6}^{10} t = \frac{1}{5}\cdot 579 - \frac{1}{5}\cdot 4{,}1 \cdot 40 = 83{,}0.$$

Таблица 4.3

Прогнозирование ряда на основе метода гармонических весов (*k*=5)

t	y_t	a_t	b_t	y_t	P_{t+1}	C_{t+1}	P_{t+1} C_{t+1}	ε_t
1	95	92,5	3,1	95,6	-	-	-	-
2	101	92,2	3,3	98,8	3,2	0,002	0,006	-0,6
3	102	93,5	2,7	101,8	3,1	0,004	0,012	2,2
4	100	103,8	0,6	105,2	3,4	0,005	0,017	0,2
5	111	93,1	2,9	107,6	2,4	0,007	0,017	-5,2
6	113	83	4,1	109,4	1,8	0,010	0,018	3,4
7	109	72,3	5,1	110,7	1,3	0,012	0,016	3,6
8	104	55,6	6,9	112,9	2,2	0,014	0,031	-1,7
9	130	108	2	120,2	7,2	0,016	0,115	-8,9
10	123	114,8	1,2	125,3	5,1	0,019	0,097	9,8
11	125	117,9	1,1	126,6	4,2	0,022	0,092	-2,3
12	141	149	-1,1	133,3	3,7	0,025	0,093	-1,6
13	131	147	-1,2	132,7	-0,6	0,028	-0,168	7,7
14	126	67,4	4,2	131,1	-1,6	0,032	-0,051	-1,7
15	138	75,5	3,7	131,6	0,5	0,035	0,018	-5,1
16	132	85,6	2,9	132,3	0,6	0,039	0,023	6,4
17	122	65,8	4	134,7	2,4	0,044	0,106	-0,3
18	155	168,6	-1,1	141,9	7,2	0,049	0,353	-12,7
19	145	57,5	4,3	143	1,1	0,055	0,061	13,1
20	135	37,8	5,1	143,9	0,8	0,062	0,050	2
21	152	139,2	0,6	148	4,1	0,071	0,291	-8,9
22	146	-	-	149,7	1,8	0,081	0,146	4
23	161	-	-	154,8	5,1	0,095	0,485	-3,7
24	156	-	-	156,9	2,1	0,116	0,244	6,2
25	150	-	-	154,2	-2,7	0,157	-0,424	-0,9
					58,4	1	1,648	

Для получения сглаженного, например, шестого (t=6) воспользуемся формулой (4.25), из которой имеем:

$$\hat{y}_6 = \frac{1}{5}\sum_{i=6-5+1}^{6} y_i(t) = \frac{1}{5}\left(y_2(6) + y_3(6) + y_4(6) + y_5(6) + y_6(6)\right),$$

где

$$y_2(6) = 92{,}2 + 3{,}3 \cdot 6 = 112{,}0; \quad y_3(6) = 93{,}5 + 2{,}7 \cdot 6 = 104{,}3;$$
$$y_4(6) = 103{,}8 + 0{,}6 \cdot 6 = 105{,}6; \quad y_5(6) = 93{,}1 + 2{,}9 \cdot 6 = 110{,}5;$$
$$y_6(6) = 83{,}0 + 4{,}1 \cdot 6 = 107{,}6.$$

Таким образом, $\hat{y}_6 = \frac{1}{5} \cdot 545{,}4 = 109{,}4$.

Аналогично рассчитываются и другие сглаженные уровни ряда за исключением крайних первых и последних четырех уровней, при вычислении которых используется не 5, а меньшее количество значений.

Определив взвешенные приросты сглаженных уровней, получим модель

$$\hat{y}_t = 154{,}2 + 1{,}648 \cdot (t - 25) \quad (t = 26, 27, \ldots).$$

Глава 5. Дискриминантный анализ

5.1. Постановка задачи дискриминантного анализа

Дискриминантный анализ является разделом многомерного статистического анализа, который включает в себя методы классификации многомерных наблюдений по принципу максимального сходства при наличии обучающих признаков.

В дискриминантном анализе формулируется правило, по которому объекты подмножества подлежащего классификации относятся к одному из уже существующих (обучающих) подмножеств (классов) на основе сравнения величины дискриминантной функции классифицируемого объекта, рассчитанной по дискриминантным переменным, с некоторой константой дискриминации.

Классификация – разделение рассматриваемой совокупности объектов или явлений на однородные в определенном смысле группы, либо отнесение каждого из заданного множества объектов к одному из заранее известных классов.

Класс – генеральная совокупность, описываемая одномодальной функцией плотности $f(x)$.

Дискриминантная функция (решающее правило, процедура классификации) – статистика, служащая для построения правила классификации объектов по группам.

Дискриминантная функция может быть как линейной, так и нелинейной. Выбор ее вида зависит от геометрического расположения разделяемых классов в пространстве дискриминантных переменных.

Предположим, что существуют две или более совокупности (группы) и что мы располагаем множеством выборочных наблюдений над ними. Основная задача дискриминантного анализа состоит в построении с помощью этих выборочных наблюдений правила, позволяющего отнести новое наблюдение к одной из совокупностей.

Пусть имеется множество M единиц N объектов наблюдения, каждая i-я единица которого описывается совокупностью p значений дискриминантных переменных (признаков) x_{ij}, ($i = 1, 2, ..., N$; $j = 1, 2, ..., p$). Причем все множество M объектов включает q обучающих подмножеств ($q \geq 2$) M_k размером n_k каждое и подмножество M_0 объектов подлежащих дискриминации (под дискриминацией понимается различие). Здесь k — номер подмножества (класса), $k = 1, 2, ..., q$.

Требуется установить правило (линейную или нелинейную дискриминантную функцию $f(X)$) распределения m-объектов подмножества M_0 по подмножествам M_k.

Наиболее часто используется линейная форма дискриминантной функции, которая представляется в виде скалярного произведения векторов $A = (a_1, a_2, ..., a_p)$ дискриминантных множителей и вектора $X_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,p})$ дискриминантных переменных:

$$F_i = A \times X_i \qquad (5.1)$$

или

$$F_i = a_1 x_{i,1} + a_2 x_{i,2} + \ldots a_p x_{i,p}, \qquad (5.2)$$

где X_i — транспонированный вектор дискриминантных переменных x_{ij} — значений j-х признаков у i-го объекта наблюдения.

5.2. Алгоритм выполнения дискриминантного анализа

Алгоритм выполнения дискриминантного анализа рассмотрен применительно к линейной дискриминантной функции вида (5.1). Его основные этапы.

1. Исходные данные представляются либо в табличной форме в виде q подмножеств (обучающих выборок) M_k и подмножества M_0 объектов подлежащих дискриминации, либо сразу в виде матриц $X^{(1)}$, $X^{(2)}$, ..., $X^{(q)}$, размером $(n_k \times p)$.

Таблица 5.1.

Номер подмножества M_k ($k = 1, 2, ..., q$)	Номер объекта, i ($i = 1, 2, ..., n_k$)	Свойства (показатель), j ($j = 1, 2, ..., p$)			
		x_1	x_2	…	x_0
Подмножество M_1 ($k = 1$)	1	$x_{1,1}^{(1)}$	$x_{1,2}^{(1)}$	…	$x_{1,p}^{(1)}$
	2	$x_{2,1}^{(1)}$	$x_{2,2}^{(1)}$	…	$x_{2,p}^{(1)}$
	…	…	…	…	…
	n_1	$x_{n_1,1}^{(1)}$	$x_{n_1,2}^{(1)}$	…	$x_{n_1,p}^{(1)}$
Подмножество M_2 ($k = 2$)	1	$x_{1,1}^{(2)}$	$x_{1,2}^{(2)}$	…	$x_{1,p}^{(2)}$
	2	$x_{2,1}^{(2)}$	$x_{2,2}^{(2)}$	…	$x_{2,p}^{(2)}$
	…	…	…	…	…
	n_2	$x_{n_2,1}^{(2)}$	$x_{n_2,2}^{(2)}$	…	$x_{n_2,p}^{(2)}$
…	…	…	…	…	…
Подмножество M_q	1	$x_{1,1}^{(q)}$	$x_{1,2}^{(q)}$	…	$x_{1,p}^{(q)}$

$(k = q)$	2	$x_{2,1}^{(q)}$	$x_{2,2}^{(q)}$	…	$x_{2,p}^{(q)}$
	…	…	…	…	…
	n_q	$x_{n_q,1}^{(q)}$	$x_{n_q,2}^{(q)}$	…	$x_{n_q,p}^{(q)}$
Подмножество M_0, подлежащее дискриминации	1	$x_{1,1}^{(0)}$	$x_{1,2}^{(0)}$	…	$x_{1,p}^{(0)}$
	2	$x_{2,1}^{(0)}$	$x_{2,2}^{(0)}$	…	$x_{2,p}^{(0)}$
	…	…	…	…	…
	m	$x_{m,1}^{(0)}$	$x_{m,2}^{(0)}$	…	$x_{m,p}^{(0)}$

$$X^{(1)} = \begin{pmatrix} x_{1,1}^{(1)} & x_{1,2}^{(1)} & \dots & x_{1,p}^{(1)} \\ x_{2,1}^{(1)} & x_{2,2}^{(1)} & \dots & x_{2,p}^{(1)} \\ \dots & \dots & \dots & \dots \\ x_{n1,1}^{(1)} & x_{n1,2}^{(1)} & \dots & x_{n1,p}^{(1)} \end{pmatrix}; \quad X^{(2)} = \begin{pmatrix} x_{1,1}^{(2)} & x_{1,2}^{(2)} & \dots & x_{1,p}^{(2)} \\ x_{2,1}^{(2)} & x_{2,2}^{(2)} & \dots & x_{2,p}^{(2)} \\ \dots & \dots & \dots & \dots \\ x_{n2,1}^{(2)} & x_{n2,2}^{(2)} & \dots & x_{n2,p} \end{pmatrix} ;\dots;$$

$$X^{(q)} = \begin{pmatrix} x_{1,1}^{(q)} & x_{1,2}^{(q)} & \dots & x_{1,p}^{(q)} \\ x_{2,1}^{(q)} & x_{2,2}^{(q)} & \dots & x_{2,p}^{(q)} \\ \dots & \dots & \dots & \dots \\ x_{nq,1}^{(q)} & x_{nq,2}^{(q)} & \dots & x_{nq,p}^{(q)} \end{pmatrix}; \quad X^{(0)} = \begin{pmatrix} x_{1,1}^{(0)} & x_{1,2}^{(0)} & \dots & x_{1,p}^{(0)} \\ x_{2,1}^{(0)} & x_{2,2}^{(0)} & \dots & x_{2,p}^{(0)} \\ \dots & \dots & \dots & \dots \\ x_{m,1}^{(0)} & x_{m,2}^{(0)} & \dots & x_{m,p}^{(0)} \end{pmatrix}$$

где $X^{(k)}$ - матрицы с обучающими признаками (k = 1, 2, ..., q), $X^{(0)}$ - матрица новых m-объектов, подлежащих дискриминации (размером $m \times p$), p — количество свойств, которыми характеризуется каждый i-й объект.

Здесь должно выполняться условие: общее количество объектов N множества M должно быть равно сумме количества объектов m (в подмножестве M_0), подлежащих дискриминации, и общего количества объектов $\sum_{k=1}^{q} n_k$ в обучающих подмножествах: $N = m + \sum_{k=1}^{q} n_k$, где q - количество обучающих подмножеств ($q \geq 2$). В реальной практике наиболее часто реализуется случай q=2, поэтому и алгоритм дискриминантного анализа приведен для данного варианта.

2. Определяются $\overline{X}_j^{(k)}$ элементы векторов $\overline{X}^{(\kappa)}$ средних значений по каждому j-му признаку для i объектов внутри k-го подмножества (k = 1, 2):

$$\overline{X}_j^{(k)} = \frac{\sum_{i=1}^{n_1} x_{ij}^{(k)}}{n_k}, \quad j = 1,2,..., p. \qquad (5.3)$$

Результаты расчета представляются в виде векторов столбцов $\overline{X}^{(\kappa)}$:

$$\overline{X}^{(k)} = \begin{pmatrix} \overline{X}_1^{(k)} \\ \overline{X}_2^{(k)} \\ \dots \\ \overline{X}_p^{(k)} \end{pmatrix}$$

3. Для каждого обучающего подмножества рассчитываются ковариационные матрицы $S^{(k)}$ (размером $p \times p$):

$$S^{(k)} = \left(\frac{1}{n_k} \sum_{k=1}^{n_1} (X_{ik}^{(k)} - \overline{X}_i^{(k)}) \right) \left(X_{jk}^{(k)} - \overline{X}_j^{(k)} \right)_{p \times p}. \qquad (5.4)$$

4. Рассчитывается объединенная ковариационная матрица $\hat{S}$ по формуле:

$$\hat{S} = \frac{1}{n_1 + n_2 - 2} \left(n_1 \times S^{(1)} + n_2 \times S^{(2)} \right). \qquad (5.5)$$

5. Рассчитывается матрица $\hat{S}^{-1}$ обратная к объединенной ковариационной матрице $\hat{S}$:

$$\hat{S}^{-1} = \frac{1}{\left| \hat{S} \right|} \times \overline{\hat{S}}, \qquad (5.6)$$

где $\left| \hat{S} \right|$ — определитель матрицы $\hat{S}$, (причем $\left| \hat{S} \right| \neq 0$), $\overline{\hat{S}}$ — присоединенная матрица, элементы которой являются алгебраическими дополнениями элементов матрицы $\hat{S}$.

6. Рассчитывается вектор-столбец $A = \begin{pmatrix} A_1 \\ A_2 \\ \dots \\ A_p \end{pmatrix}$ дискриминантных множителей с учетом всех элементов обучающих подмножеств по формуле: $A = \hat{S}^{-1} \left(\overline{X}^{(1)} - \overline{X}^{(2)} \right)$.

Данная расчетная формула получена с помощью метода наименьших квадратов из условия обеспечения наибольшего различия между дискриминантными функциями. Наилучшее разделение двух обучающих подмножеств обеспечивается сочетанием минимальной внутригрупповой вариации и максимальной межгрупповой вариации.

7. По каждому *i*-му объекту (*i* = 1, 2, ..., *N*) множества *M* определяется соответствующее значение дискриминантной функции:

$$F_i^{(k)} = A_1 x_{i,1}^{(k)} + A_2 x_{i,2}^{(k)} + \dots + A_p x_{i,p}^{(k)}. \qquad (5.7)$$

8. По совокупности найденных значений $F^{(k)}$ рассчитываются средние значения для каждого подмножества M_k:

$$\overline{F}^{(k)} = \frac{\sum_{i=1}^{n_1} F_i^{(k)}}{n_k}, \quad k = 1,2. \qquad (5.8)$$

9. Определяется общее среднее (константа дискриминации) для дискриминантных функций

$$\overline{F} = \frac{\sum_{k=1}^{q} \overline{F}^{(k)}}{q}. \qquad (5.9)$$

10. Выполняется распределение (дискриминация) объектов подмножества M_0 подлежащих дискриминации по обучающим выборкам M_1 и M_2. С этой целью рассчитанные и п. 7 по каждому i-му объекту значения дискриминантных функций

$$F_i^{(0)} = A_1 X_{i,1}^{(0)} + A_2 X_{i,2}^{(0)} + ... + A_p X_{i,p}^{(0)}, \quad i = 1,2,...,m$$

сравниваются с величиной $\overline{F}$ общего среднего. На основе сравнения данный объект относят к одному из обучающих подмножеств.

Если $\overline{F}^{(1)} > \overline{F}^{(2)}$, то i-й объект подмножества M_0 относят к подмножеству M_1, при $F_i^{(0)} - \overline{F} > 0$ и к подмножеству M_2 при $F_i^{(0)} - \overline{F} < 0$. Если же $\overline{F}^{(1)} < \overline{F}^{(2)}$, то заданный объект относят к подмножеству M_1, при $F_i^{(0)} - \overline{F} < 0$ и к подмножеству M_2 в противном случае.

11. Далее делается оценка качества распределения новых объектов, для чего оценивается вклад переменных в дискриминантную функцию.

Влияние признаков на значение дискриминантной функции и результаты классификации может оцениваться по дискриминантным множителям (коэффициентам дискриминации), по дискриминантным нагрузкам признаков или по дискриминантной матрице.

Дискриминантные множители зависят от масштабов единиц измерения признаков, поэтому они не всегда удобны для оценки.

Дискриминантные нагрузки более надежны в оценке признаков, они вычисляются как парные линейные коэффициенты корреляции между рассчитанными уровнями дискриминантной функции F и признаками, взятыми для ее построения.

Дискриминантная матрица характеризует меру соответствия результатов классификации фактическому распределению объектов по подмножествам и используется для оценки качества анализа. В этом случае дискриминантная функция F формируется по данным объектов (с измеренными p признаками) обучающих подмножеств, а затем проверяется качество этой функции путем сопоставления фактической

классовой принадлежности объектов с той, что получена в результате формальной дискриминации.

Пример5.1. Применения дискриминантного анализа при наличии двух обучающих выборок (q=2).

Имеются данные по двум группам промышленных предприятий отрасли: X_1 - среднегодовая стоимость основных производственных фондов, млн. д.ед.; X_2 — среднесписочная численность персонала, тыс. чел.; X_3 — балансовая прибыль млн. д.ед.

Исходные данные представляются в табличной форме

Таблица 5.2

Номер группы M_k (k =1, 2)	Номер предприятия, i (i = 1, 2, ..., n_k)	Свойства (показатель), j (j = 1, 2, ..., p)		
		X_1	X_2	X_3
Группа 1, M_1 (k = 1)	1	224,228	17,115	22,981
	2	151,827	14,904	21,481
	3	147,313	13,627	28,669
	4	152,253	10,545	10,199
Группа 2, M_2 (k = 2)	1	46,757	4,428	11,124
	2	29,033	5,51	6,091
	3	52,134	4,214	11,842
	4	37,05	5,527	11,873
	5	63,979	4,211	12,860
Группа предприятий M_0, подлежащих дискриминации	1	55,451	9,592	12,840
	2	78,575	11,727	15,535
	3	98,353	17,572	20,458

Необходимо провести классификацию (дискриминацию) трех новых предприятий, образующих группу M_0 с известными значениями исходных переменных.

***Решение*:**

1. Значения исходных переменных для обучающих подмножеств M_1 и M_2 (групп предприятий) записываются в виде матриц $X^{(1)}$ и $X^{(2)}$:

$$X^{(1)} = \begin{pmatrix} 224,228 & 17,115 & 28,981 \\ 151,827 & 14,904 & 21,481 \\ 147,313 & 13,627 & 28,669 \\ 152,253 & 10,545 & 10,199 \end{pmatrix};$$

$$X^{(2)} = \begin{pmatrix} 46{,}757 & 4{,}428 & 11{,}124 \\ 29{,}033 & 5{,}510 & 6{,}091 \\ 52{,}134 & 4{,}214 & 11{,}842 \\ 37{,}050 & 5{,}527 & 11{,}873 \\ 63{,}979 & 4{,}211 & 12{,}86 \end{pmatrix}.$$

и для подмножества M_0 группы предприятий, подлежащих классификации в виде матрицы $X^{(0)}$:

$$X^{0} = \begin{pmatrix} 55{,}451 & 9{,}592 & 12{,}840 \\ 78{,}575 & 11{,}727 & 15{,}535 \\ 98{,}353 & 17{,}572 & 20{,}458 \end{pmatrix}.$$

Общее количество предприятий, составляющих множество M, будет равно $N = 3+4+5 = 12$ ед.

2. Определяются элементы векторов $\overline{X}_j^{(k)}$ средних значений по j признакам для i-х объектов по каждой k-й выборке (k = 1, 2), которые представляются в виде двух векторов $\overline{X}^{(k)}$ (по количеству обучающих выборок):

$$\overline{X}^{(1)} = \begin{pmatrix} 168{,}9053 \\ 14{,}04775 \\ 20{,}8325 \end{pmatrix}; \quad \overline{X}^{(2)} = \begin{pmatrix} 45{,}7906 \\ 4{,}778 \\ 10{,}758 \end{pmatrix}.$$

3. Для каждого обучающего подмножества M_1 и M_2 рассчитываются ковариационные матрицы S_k (размером $p \times p$):

$$S^{(1)} = \begin{pmatrix} 1023{,}949 & 55{,}61977 & 29{,}91243 \\ 55{,}61977 & 5{,}646869 & 10{,}27364 \\ 28{,}91243 & 10{,}27364 & 44{,}87966 \end{pmatrix};$$

$$S^{(2)} = \begin{pmatrix} 145{,}8412 & -6{,}6048 & 22{,}78478 \\ -6{,}6084 & 0{,}371782 & -0{,}90248 \\ 22{,}78478 & -0{,}90248 & 5{,}750306 \end{pmatrix}$$

4. Рассчитывается объединенная ковариационная матрица:

$$\widehat{S} = \frac{1}{4+5-2}\left[4 \times \begin{pmatrix} 1023{,}949 & 55{,}61977 & 29{,}91243 \\ 55{,}61977 & 5{,}646869 & 10{,}27364 \\ 28{,}91243 & 10{,}27364 & 44{,}87966 \end{pmatrix} + \right.$$

$$\left. + 5 \times \begin{pmatrix} 145{,}8412 & -6{,}6084 & 22{,}78478 \\ -6{,}6084 & 0{,}371782 & -0{,}90248 \\ 22{,}78478 & -0{,}90248 & 5{,}750306 \end{pmatrix} \right] = \begin{pmatrix} 689{,}286 & 27{,}06244 & 32{,}79623 \\ 27{,}06244 & 3{,}492341 & 5{,}22602 \\ 32{,}79623 & 5{,}22602 & 29{,}75288 \end{pmatrix}$$

5. Рассчитывается матрица $\widehat{S}^{-1}$ обратная к объединенной ковариационной матрице:

$$\widehat{S}^{-1} = \begin{pmatrix} 0{,}002097 & -0{,}01735 & 0{,}000736 \\ -0{,}01735 & 0{,}532023 & -0{,}07432 \\ 0{,}000736 & -0{,}07232 & 0{,}045853 \end{pmatrix}$$

6. Рассчитываются дискриминантные множители (коэффициенты дискриминантной функции) по всем элементам обучающих подмножеств:

$$A = \begin{pmatrix} 0{,}104743 \\ 2{,}046703 \\ -0{,}13635 \end{pmatrix}$$

7. Для каждого *i*-го объекта *k*-го подмножества *M* определяется значение дискриминантной функции:

$F_1^{(1)}$=0,104743×224,228+2,046703×17,115+(-0,13635)×22,981=55,38211;
$F_2^{(1)}$=0,104743×151,827+2,046703×14,904+(-0,13635)×21,481=43,47791;
$F_3^{(1)}$=0,104743×147,313+2,046703×13,627+(-0,13635)×28,669=39,41138;
$F_4^{(2)}$=0,104743×152,253+2,046703×10,545+(-0,13635)×10,199=36,13924;
$F_1^{(2)}$=0,104743×46,757+2,046703×4,428+(-0,13635)×11,124=12,44351;
……..
$F_5^{(2)}$=0,104743×63,979+2,046703×4,211+(-0,13635)×12,860=13,56655.

8. По совокупности найденных значений $F^{(k)}$ рассчитываются средние значения $\overline{F}^{(k)}$ для каждого подмножества M_k:

$$\overline{F}^{(1)} = 43{,}60266;$$
$$\overline{F}^{(2)} = 13{,}10853.$$

9. Определяется общее среднее (константа дискриминации) для дискриминантных функций:

$$\overline{F} = \frac{43{,}60266 + 13{,}10853}{2} = 28{,}3556$$

10. Выполняется распределение объектов подмножества M_0 по обучающим подмножествам M_1 и M_2, для чего по каждому объекту (i = 1, 2, 3) рассчитываются дискриминантные функции:

$F_1^{(0)}$=0,104743×55,451+2,046703×9,592+(-0,13635)×12,840=23,68661
$F_2^{(0)}$=0,104743×78,575+2,046703×11,727+(-0,13635)×15,535=30,11366
$F_3^{(0)}$=0,104743×98,353+2,046703×17,572+(-0,13635)×20,458=23,68661

Затем рассчитанные значения дискриминантных функций $F^{(0)}$ сравниваются с общей средней F=28,3556.

Поскольку $\overline{F}^{(1)} > \overline{F}^{(2)}$, то *i*-й объект подмножества M_0 относят к подмножеству M_1 при $F_i^{(0)} - \overline{F} > 0$ и к подмножеству M_2 при $F_i^{(0)} - \overline{F} < 0$. С

учетом этого в данном примере предприятия 2 и 3 подмножества M_0 относятся к M_1, а предприятие 1 относится к M_2.

Если бы выполнялось условие $\overline{F}^{(1)} < \overline{F}^{(2)}$, то объекты M_0 относились к подмножеству M_1, при $F_i^{(0)} - \overline{F} \prec 0$ и к подмножеству M_2 в противном случае.

11. Оценку качества распределения новых объектов выполним путем сравнения с константой дискриминации F значений дискриминантных функций $F_i^{(k)}$=обучающих подмножеств M_1 и M_2. Поскольку для всех найденных значений выполняются неравенства $F_i^{(1)} > \overline{F}$, и $\overline{F}_i^{(2)} < \overline{F}$, то можно предположить о правильном распределении объектов и уже существующих двух классах и верно выполненной классификации объектов подмножества M_0.

Приложение I

Значения функции Лапласа $\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-t^2/2}\, dt$

Целые и десятые доли x	Сотые доли x									
A	0	1	2	3	4	5	6	7	8	9
0,0	0,00000	0,00400	0,00800	0,01195	0,01595	0,01995	0,02390	0,02790	0,03190	0,03585
0,1	0,03985	0,04380	0,04775	0,05170	0,05565	0,05960	0,06355	0,06750	0,07140	0,07535
0,2	0,07925	0,08315	0,08705	0,09095	0,09485	0,09870	0,10255	0,10640	0,11025	0,11410
0,3	0,11790	0,12170	0,12550	0,12930	0,13305	0,13685	0,14060	0,14430	0,14800	0,15175
0,4	0,15540	0,15910	0,16275	0,16640	0,17005	0,17365	0,17725	0,18080	0,18440	0,18795
0,5	0,19145	0,19495	0,19845	0,20195	0,20540	0,20885	0,21225	0,21565	0,21905	0,22240
0,6	0,22575	0,22905	0,23235	0,23565	0,23890	0,24215	0,24535	0,24855	0,25175	0,25490
0,7	0,25805	0,26115	0,26425	0,26730	0,27035	0,27335	0,27635	0,27935	0,28230	0,28525
0,8	0,28765	0,29105	0,29390	0,29675	0,29955	0,30235	0,30510	0,30785	0,31055	0,31325
0,9	0,31595	0,31860	0,32120	0,32380	0,32640	0,32895	0,33145	0,33395	0,33645	0,33890
1,0	0,34135	0,34375	0,34615	0,34850	0,35085	0,35165	0,35545	0,35770	0,35995	0,36215
1,1	0,36435	0,36650	0,36865	0,37075	0,37285	0,37495	0,37700	0,37900	0,38100	0,38300
1,2	0,38495	0,38685	0,38875	0,39065	0,39250	0,39435	0,39615	0,39795	0,39920	0,40145
1,3	0,40270	0,40480	0,40660	0,40825	0,40990	0,41150	0,41310	0,41465	0,41620	0,41775
1,4	0,41925	0,42075	0,42220	0,42365	0,32505	0,42645	0,42785	0,42920	0,43055	0,43440
1,5	0,43320	0,43450	0,43575	0,43700	0,43820	0,43945	0,44060	0,44180	0,44295	0,44410
1,6	0,44520	0,44630	0,44740	0,44845	0,44950	0,45055	0,45155	0,45255	0,45350	0,45450
1,7	0,45545	0,45635	0,45730	0,45820	0,45905	0,45995	0,46080	0,46165	0,46245	0,46325
1,8	0,46405	0,46485	0,46560	0,46635	0,46710	0,46785	0,46855	0,46925	0,46960	0,47060
1,9	0,47130	0,47195	0,47255	0,47320	0,47380	0,47440	0,47500	0,47560	0,47615	0,47665
2,0	0,47725	0,47775	0,47825	0,47880	0,47930	0,47980	0,48030	0,48080	0,48125	0,48170
2,1	0,47715	0,48255	0,48300	0,48340	0,47880	0,48420	0,48460	0,48500	0,48535	0,48575
2,2	0,48610	0,48645	0,48680	0,48715	0,48745	0,48780	0,48810	0,48790	0,48870	0,48900
2,3	0,48930	0,48955	0,48985	0,49010	0,49035	0,49060	0,49085	0,49110	0,49135	0,49160
2,4	0,49180	0,49205	0,49225	0,49245	0,49265	0,49285	0,49305	0,49325	0,49345	0,49360
2,5	0,49380	0,49395	0,49415	0,49430	0,49445	0,49460	0,49475	0,49490	0,49505	0,49520
2,6	0,49535	0,49550	0,49560	0,49575	0,49585	0,49600	0,49610	0,49620	0,49630	0,49640
2,7	0,49655	0,49665	0,49675	0,49685	0,49695	0,49700	0,49710	0,49720	0,49730	0,49735
2,8	0,49745	0,49755	0,49760	0,49765	0,49775	0,49780	0,49790	0,49795	0,49800	0,46805
2,9	0,49815	0,49770	0,49825	0,49830	0,49835	0,49840	0,49845	0,49850	0,49855	0,49860
3,0	0,49865	0,49870	0,49875	0,49880	0,49880	0,49885	0,49890	0,49895	0,49895	0,49900
3,1	0,49905	0,49905	0,49910	0,49915	0,49915	0,49920	0,49920	0,49925	0,49925	0,49930
3,2	0,49925	0,49935	0,49935	0,49940	0,49940	0,49945	0,49945	0,49945	0,49950	0,49950
3,3	0,49950	0,49955	0,49955	0,49955	0,49960	0,49960	0,49960	0,49960	0,49965	0,49965
3,4	0,49965	0,49970	0,49970	0,49970	0,49970	0,49970	0,49975	0,49975	0,49975	0,49975
3,5	0,49975	0,49975	0,49975	0,49980	0,49980	0,49980	0,49980	0,49980	0,49985	0,49985
3,6	0,49985	0,49985	0,49985	0,49985	0,49985	0,49985	0,49985	0,49990	0,49990	0,49990
3,7	0,49990	0,49990	0,49990	0,49990	0,49990	0,49990	0,49990	0,49990	0,49990	0,49990
3,8	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995
3,9	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995
4,0	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995	0,49995

Приложение II

Значения $t_{\alpha,k}$ – критерия Стьюдента

Число степеней свободы k	Уровень значимости α													
	0,9	0,8	0,7	0,6	0,5	0,4	0,3	0,2	0,1	0,05	0,02	0,01	0,002	0,001
1	0,16	0,32	0,51	0,73	1,00	1,38	1,96	3,08	6,31	12,71	31,82	63,7	318,3	637,0
2	0,14	0,29	0,44	0,62	0,82	1,06	1,34	1,89	2,92	4,30	6,96	9,92	22,33	31,6
3	0,14	0,28	0,42	0,58	0,76	0,98	1,25	1,64	2,35	3,18	4,54	5,84	10,22	12,9
4	0,13	0,27	0,41	0,57	0,74	0,94	1,19	1,53	2,13	2,78	3,75	4,60	7,17	8,61
5	0,13	0,27	0,41	0,56	0,73	0,92	1,16	1,48	2,01	2,57	3,36	4,03	5,89	6,86
6	0,13	0,26	0,40	0,55	0,72	0,91	1,13	1,44	1,94	2,45	3,14	3,71	5,21	5,96
7	0,13	0,26	0,40	0,55	0,71	0,90	1,12	1,41	1,89	2,36	3,00	3,50	4,79	5,40
8	0,13	0,26	0,40	0,55	0,70	0,89	1,11	1,40	1,86	2,31	2,90	3,36	4,50	5,04
9	0,13	0,26	0,40	0,54	0,70	0,88	1,10	1,38	1,83	2,26	2,82	3,25	4,30	4,78
10	0,13	0,26	0,40	0,54	0,70	0,88	1,09	1,37	1,81	2,23	2,76	3,17	4,14	4,59
11	0,13	0,26	0,40	0,54	0,70	0,88	1,09	1,36	1,80	2,20	2,72	3,11	4,03	4,44
12	0,13	0,26	0,39	0,54	0,69	0,87	1,08	1,36	1,78	2,18	2,68	3,05	3,93	4,32
13	0,13	0,26	0,39	0,54	0,69	0,87	1,08	1,35	1,77	2,16	2,65	3,01	3,85	4,22
14	0,13	0,26	0,39	0,54	0,69	0,87	1,08	1,34	1,76	2,14	2,62	2,98	3,79	4,14
15	0,13	0,26	0,39	0,54	0,69	0,87	1,07	1,34	1,75	2,13	2,60	2,95	3,73	4,07
16	0,13	0,26	0,39	0,53	0,69	0,86	1,07	1,34	1,75	2,12	2,58	2,92	3,69	4,01
17	0,13	0,26	0,39	0,53	0,69	0,86	1,07	1,33	1,74	2,11	2,57	2,90	3,65	3,96
18	0,13	0,26	0,39	0,53	0,69	0,86	1,07	1,33	1,73	2,10	2,55	2,88	3,61	3,92
19	0,13	0,26	0,39	0,53	0,69	0,86	1,07	1,33	1,73	2,09	2,54	2,86	3,58	3,88
20	0,13	0,26	0,39	0,53	0,69	0,86	1,06	1,32	1,72	2,09	2,53	2,85	3,55	3,85
21	0,13	0,26	0,39	0,53	0,69	0,86	1,06	1,32	1,72	2,08	2,52	2,83	3,53	3,82
22	0,13	0,26	0,39	0,53	0,69	0,86	1,06	1,32	1,72	2,07	2,51	2,82	3,51	3,79
23	0,13	0,26	0,39	0,53	0,68	0,86	1,06	1,32	1,71	2,07	2,50	2,81	3,49	3,77
24	0,13	0,26	0,39	0,53	0,68	0,86	1,06	1,32	1,71	2,06	2,49	2,80	3,47	3,74
25	0,13	0,26	0,39	0,53	0,68	0,86	1,06	1,32	1,71	2,06	2,48	2,79	3,45	3,72

Окончание приложения II

26	0,13	0,26	0,39	0,53	0,68	0,86	1,06	1,31	1,71	2,06	2,48	2,78	3,44	3,71
27	0,13	0,26	0,39	0,53	0,68	0,85	1,06	1,31	1,70	2,05	2,47	2,77	3,42	3,69
28	0,13	0,26	0,39	0,53	0,68	0,85	1,06	1,31	1,70	2,05	2,47	2,76	3,40	3,66
29	0,13	0,26	0,39	0,53	0,68	0,85	1,05	1,31	1,70	2,04	2,46	2,76	3,40	3,66
30	0,13	0,26	0,39	0,53	0,68	0,85	1,05	1,31	1,70	2,04	2,46	2,75	3,39	3,65
40	0,13	0,25	0,39	0,53	0,68	0,85	1,05	1,30	1,68	2,02	2,42	2,70	3,31	3,55
60	0,13	0,25	0,39	0,53	0,68	0,85	05	1,30	1,67	2,00	2,39	2,66	3,23	3,46
120	0,13	0,25	0,39	0,53	0,68	0,84	1,04	1,29	1,66	1,98	2,36	2,62	3,17	3,37
∞	0,13	0,25	0,38	0,52	0,67	0,84	1,04	1,28	1,64	1,96	2,33	2,58	3,09	3,29

Приложение III

Значения средней μ и стандартных ошибок σ_1, σ_2 для n от 10 до 50

n	μ	σ_1	σ_2
10	3,858	1,288	1,964
15	4,636	1,521	2,153
20	5,195	1,677	2,279
25	5,632	1,791	2,373
30	5,99	1,882	2,447
35	6,294	1,956	2,509
40	6,557	2,019	2,561
45	6,79	2,072	2,606
50	6,998	2,121	2,645

Приложение IV

Значения $\chi^2_{\alpha;k}$ критерия Пирсона

Число степеней свободы k	Вероятность α												
	0,99	0,98	0,95	0,90	0,80	0,70	0,50	0,30	0,20	0,10	0,05	0,02	0,01
1	0,00	0,00	0,00	0,02	0,06	0,15	0,45	1,07	1,64	2,71	3,84	5,41	6,64
2	0,02	0,04	0,10	0,21	0,45	0,71	1,39	2,41	3,22	4,60	5,99	7,82	9,21
3	0,11	0,18	0,35	0,58	1,00	1,42	2,37	3,66	4,64	6,25	7,82	9,84	11,3
4	0,30	0,43	0,71	1,06	1,65	2,20	3,36	4,88	5,99	7,78	9,49	11,7	13,3
5	0,55	0,75	1,14	1,61	2,34	3,00	4,35	6,06	7,29	9,24	11,1	13,4	15,1
6	0,87	1,13	1,63	2,20	3,07	3,83	5,35	7,23	8,56	10,6	12,6	15,0	16,8
7	1,24	1,56	2,17	2,83	3,82	4,67	6,35	8,38	9,80	12,0	14,1	16,6	18,5
8	1,65	2,03	2,73	3,49	4,59	5,53	7,34	9,52	11,0	13,4	15,5	18,2	20,1
9	2,09	2,53	3,32	4,17	5,38	6,39	8,34	10,7	12,2	14,7	16,9	19,7	21,7
10	2,56	3,06	3,94	4,86	6,18	7,27	9,34	11,8	13,4	16,0	18,3	21,2	23,2
11	3,05	3,61	4,58	5,58	6,99	8,15	10,3	12,9	14,6	17,3	19,7	22,6	24,7
12	3,57	4,18	5,23	6,30	7,81	9,03	11,3	14,0	15,8	18,5	21,0	24,1	26,2
13	4,11	4,76	5,89	7,04	8,63	9,93	12,3	15,1	17,0	19,8	22,4	25,5	27,7
14	4,66	5,37	6,57	7,79	9,47	10,8	13,3	16,2	18,1	21,1	23,7	26,9	29,1
15	5,23	5,98	7,26	8,55	10,3	11,7	14,3	17,3	19,3	22,3	25,0	28,3	30,6
16	5,81	6,61	7,96	9,31	11,1	12,6	15,3	18,4	20,5	23,5	26,3	29,6	32,0
17	6,41	7,26	8,67	10,1	12,0	13,5	16,3	19,5	21,6	24,8	27,6	31,0	33,4
18	7,02	7,91	9,39	10,9	12,9	14,4	17,3	20,6	22,8	26,0	28,9	32,3	34,8
19	7,63	8,57	10,1	11,6	13,7	15,3	18,3	21,7	23,9	27,2	30,1	33,7	36,2
20	8,26	9,24	10,8	12,4	14,6	16,3	19,3	22,8	25,0	28,4	31,4	35,0	37,6
21	8,90	9,92	11,6	13,2	15,4	17,2	20,3	23,9	26,2	29,6	32,7	36,3	38,9
22	9,54	10,6	12,3	14,0	16,3	18,1	21,3	24,9	27,3	30,8	33,9	37,7	40,3
23	10,2	11,3	13,1	14,8	17,2	19,0	22,3	26,0	28,4	32,0	35,2	39,0	41,6
24	10,9	12,0	13,8	15,7	18,1	19,9	23,3	27,1	29,6	33,2	36,4	40,3	43,0
25	11,5	12,7	14,6	16,5	18,9	20,9	24,3	28,2	30,7	34,4	37,7	41,7	44,3

Окончание приложения IV

26	12,2	13,4	15,4	17,3	19,8	21,8	25,3	29,2	31,8	35,6	38,9	42,9	45,6
27	12,9	14,1	16,1	18,1	20,7	22,7	26,3	30,3	32,9	36,7	40,1	44,1	47,0
28	13,6	14,8	16,9	18,9	21,6	23,6	27,3	31,4	34,0	37,9	41,3	45,4	48,3
29	14,3	15,6	17,7	19,8	22,5	24,6	28,3	32,5	35,1	39,1	42,6	46,7	49,6
30	14,9	16,3	18,5	20,6	23,4	25,5	29,3	33,5	36,2	40,3	43,8	48,0	50,9

Приложение V

Значения $F_{\alpha,k1:k2}$ критерия Фишера–Снедекора

k_2 \ k_1	α=0,05																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161	200	216	225	230	234	237	239	240	242	244	246	248	249	250	251	252	253	254
2	18,5	19,0	19,2	19,2	19,3	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5	19,5
3	10,1	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	659	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,40
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	,98	1,94	,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	4,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85	1,80	1,75	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84	1,79	1,73	1,67
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,12	2,04	1,96	1,91	1,87	1,82	1,77	1,71	1,65
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81	1,75	1,70	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62

40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02	1,96	1,91	1,83	1,75	0,66	1,61	1,55	1,50	1,43	1,35	1,25
∞	3,84	3,00	3,60	2,37	2,21	2,10	2,01	1,94	1,83	1,83	1,75	1,67	1,57	1,52	1,46	1,39	1,32	1,22	1,00
α=0,01																			
1	4052	4999,5	5403	5625	5764	5859	5928	5982	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,00
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09	2,96	2,81	2,66	2,58	2,50	2,42	2,33	2,23	2,13
27	7,68	5,49	4,60	4,11	3,78	3,56	3,39	3,26	3,15	3,06	2,93	2,78	2,63	2,55	2,47	2,38	2,29	2,20	2,10
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03	2,90	2,75	2,60	2,52	2,44	2,35	2,26	2,17	2,06
29	7,60	5,42	4,54	4,04	3,73	3,50	3,33	3,20	3,09	3,00	2,87	2,73	2,57	2,49	2,41	2,33	2,23	2,14	2,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,80

60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
∞	6,63	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,18	2,04	1,88	1,79	1,70	1,59	1,47	1,32	1,00

k_1 – число степеней свободы для большой дисперсии, k_1 – для меньшей дисперсии.

Приложение VI

Значения статистик Дарбина-Уотсона при 5%-ном уровне значимости
(K - число независимых переменных уравнения регрессии)

n	K=1		K=2		K=3		K=4		K=5	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
6	0,61	1,40	-	-	-	-	-	-	-	-
7	0,7	1,36	0,47	1,9	-	-	-	-	-	-
8	0,76	1,33	0,56	1,78	0,37	2,29	-	-	-	-
9	0,82	1,32	0.63	1,7	0,46	2,13	-	-	-	-
10	0,88	1,32	0,7	1,64	0,53	2,02	-	-	-	-
11	0,93	1,32	0,66	1,6	0,6	1,93	-	-	-	-
12	0,97	1,33	0,81	1,58	0,66	1,86	-	-	-	-
13	1,01	1,34	0,86	1,56	0,72	1,82	-	-	-	-
14	1,05	1,35	0,91	1,55	0,77	1,78	-	-	-	-
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,47
100	1,65	1,68	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Литература

1. Айвазян С. А., Мхитарян В. С. Прикладная статистика и основы эконометрики. Юнити, 1998, - 1022 стр.
2. Айвазян С.А, Енюков И.О., Мешалкин Л.Д. Прикладная статистика и первичная обработка данных: Справ. изд. – М.: Финансы и статистика, 1983. – 471с.
3. Бородич С.А. Эконометрика: Учеб. пособие/ С.А. Бородич. – 2-е изд., испр. – Мн.: Новое знание, 2004. – 416 с. – (Экономическое образование).
4. Добров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. – М.: Финансы и статистика, 1998. – 352с.
5. Гришин А.Ф. Статистические методы в экономике/ А.Ф. Гришин, С.Ф. Коротов-Дарти, В.Н. Ягунов. – Ростов н/Д: «Феникс», 2005. – 344с. – (Высшее образование)
6. Козлов А.Ю., Мхитарян В.С., Шишов В.Ф. Статистические функции MS Excel в экономико-статистических расчетах: Учеб. пособие для вузов/ Под ред. проф. В.С. Мхитаряна. – М.: ЮНИТИ-ДАНА, 2003. – 231.
7. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебник для вузов. – М.: ЮНИТИ-ДАНА, 2001. – 543с.
8. Методы прогнозирования развития социально-экономических систем: Учебное пособие / О.М. Писарева. — М.: Высшая школа, 2007. - 591 с.
9. Практикум по эконометрике: Учеб. пособие / И.И. Елисеева, С.В. Курышева, Н.М. Гордеенко и др.; Под ред. И.И. Елисеевой. - М.: Финансы и статистика, 2005. – 192с.
10. Рябушкин Т.В., Симчера В.М., Машихин Е.А. Статистические методы и анализ социально-экономических процессов. – М.: Наука, 1990. – 293с.
11. Статистическое моделирование и прогнозирование: Учеб. пособие / Г.М. Гамбаров, Н.М. Журавель, Ю.Г. Королев и др.; Под ред.
А.Г. Гранберга. – М.: Финансы и статистика, 1990. – 383с.
12. Финансовая математика: математическое моделирование финансовых операций: Учеб.пособие/ Под ред. В.А. Половникова, А.И. Пилипенко. – М.: Вузовский учебник, 2004. – 360с.
13. Эконометрика: Учебник / Под ред. И. И. Елисеевой. - М.: Финансы и статистика, 2004. - 344 с

Учебное издание

Овчинникова Светлана Валерьевна
Гура Елена Николаевна

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ В УПРАВЛЕНИИ

Редактор А. В. Давтян

Дизайн обложки А. В. Клеменко

Подписано в печать 28.12.2012. Формат 60x90 1/16. Печ. л. 7,8.
Тираж 500 экз. Заказ № 2614.

Библиотечно-издательский комплекс
федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Тюменский государственный нефтегазовый университет».
625000, Тюмень, ул. Володарского, 38.

Типография библиотечно-издательского комплекса.
625039, Тюмень, ул. Киевская, 52.