Белгородский государственный университет

Е.Г.Жиляков Ю.М.Перлов
Е.П.Ревтова

ОСНОВЫ ЭКОНОМЕТРИЧЕСКОГО АНАЛИЗА ДАННЫХ

Учебное пособие

БЕЛГРОД 2004

ЧАСТЬ I. ОСНОВЫ СТАТИСТИЧЕСКОГО АНАЛИЗА

1. ВВЕДЕНИЕ В ПРЕДМЕТ. ЧТО ТАКОЕ СТАТИСТИЧЕСКИЙ АНАЛИЗ

Методы обработки экспериментальных данных либо других данных, связанных с регистрацией значений некоторой наблюдаемой величины принято называть *статистическими*, если при этом вычисляются некоторые усреднённые характеристики. В этом смысле термин «статистический» означает -*усреднённый.*

Например: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} Xi$; $X_+ = \max\{x_1, \ldots, x_n\}$;

$X_- = \min\{x_1, \ldots, x_n\}$;

$X_M = \text{med}\{x_1, \ldots, x_n\}$.

$X_{(1)} < X_{(2)} < \ldots < X_{(n)}$

Когда наблюдение представляет собой последовательность зависящую от времени, они называются *временным рядом* – этому понятию соответствует понятие *случайного процесса;* если нет оснований считать наблюдение зависимым от времени, то статистические данные соответствуют понятию *случайной величины*.
Экономическая информация – представляет собой сведения, характеризующие экономические процессы, в частности отражающая эффективность деятельности отдельных предприятий и других субъектов экономики.
Информация об успешности деятельности какого-либо субъекта экономики содержится в *статистических данных*, характеризующих их финансовое состояние и факторы производства, поэтому необходимо использовать статистические методы обработки этих данных, чтобы получить надежные выводы для принятия управленческих решений.
Статистический анализ связан с реализацией по средствам обработки как можно большего количества данных, так как надежность выводов повышается при увеличении количества обрабатываемых данных. Ограничениями на количество данных служат опасения, что условия меняются и данные будут характеризовать разное состояние исследуемого объекта.
В любом случае реализация методов статистического анализа в современных условиях предполагает использование компьютерных технологий. В настоящее время разработано много средств программной поддержки методов статистических исследований.

2. ОСНОВНЫЕ ЭТАПЫ И МОДЕЛИ СТАТИСТИЧЕСКОГО АНАЛИЗА

2.1. Общие понятия

В статистическом анализе можно выделить 4 основных этапа:

1. Формулирование цели исследований;
2. Наблюдение за исследуемым объектом и сбор данных;
3. Анализ статистических данных;

4. Предсказание.

2.1.1. Формулирование целей

Целью статистических исследований – является установления закономерности поведения исследуемых объектов или процессов, чтобы имелась возможность предсказать их поведение в каких-либо условиях.
Исследовать поведение объектов при всех возможных ситуациях не возможно, так как их слишком много; и только установленные закономерности позволяют с некоторой степенью доверия предсказать поведение в ещё не встречавшихся ситуациях. Очевидно, что предсказание абсолютно точно осуществить нельзя, поэтому следует всегда стремиться осуществить такое описание закономерностей, которое дает наименьшую погрешность предсказания.

2.1.2. Наблюдение и сбор данных

Наблюдения связаны с регистрацией, с помощью некоторых приборов, событий, процессов, явлений и тому подобное, которые подвергаются изучению. Обычно приборы регистрируют в количественном виде значение отдельных параметров, характеризующих объект исследования. Статистический анализ в подавляющем большинстве случаев оперирует с количественными данными. В ряде случаев, делаются попытки изучения поведения объектов с помощью качественных данных, например:

- интенсивность цвета;
- тепловые излучения (тепло, горячо и т.д.);

Сбор данных предлагает создание некоторых хранилищ, в которых данные организованны в определённые структуры. При этом структура баз данных и способы работы с ними во многом определяются целями статистического анализа и степенью априорных знаний о свойствах исследуемых объектов.

2.1.3. Анализ статистических данных

На этом этапе происходит описание закономерности поведения изучаемых объектов. Это осуществляется с помощью моделей различной природы.
Модель – это описание поведения объектов, отражающее основные их свойства с некоторой точки зрения.
С использованием модели тесно связано понятие *адекватность* – соответствие чему-либо. Принято различать 2 рода адекватности:

- сильная адекватность означает – установление законов, которым подчиняется поведение объекта. Примером может служить физические законы (Ньютона и др.);
- слабая адекватность предлагает, что используемая модель позволяет решить прикладную задачу (предсказания) с достаточной для исследователя точностью, при этом такая модель не обязана соответствовать реальным закономерностям поведения объекта.

В статистическом анализе в большинстве случаев используются модели слабой адекватности, особенно это касается исследования экономических процессов.

Принято различать классы моделей:

- информационные;
- физические (вещественные);
- вещественно-математические;
- логико-математические;

- имитационные

Информационные модели – описание свойств объектов средствами обычного разговорного языка. Любое понятие или определение – это вербальные модели.
Вещественная модель- уменьшенные копии реальных объектов. Таким образом можно моделировать только реально существующие объекты.
Вещественно – математические модели – физические объекты иной природы, чем исследуемые, но их поведение описывается одинаковыми математическими зависимостями. Часто возможность использования этой модели возникает при реализации решений дифференциальных уравнений.
Логико-математические модели – математические зависимости, с помощью которых используются свойства объектов чисто математическими средствами, или с помощью вычислительных экспериментов на компьютере. Иногда используются компьютерные модели, называющиеся *алгоритмическими* или *имитационными.*
Имитационные модели используются с целью воспроизведения поведения наблюдаемой характеристики объекта в каких-либо условиях. Как правило, при этом используется моделирование не детерминированных событий, появления или отсутствие которых изменяет течение процесса. Одним из названий класса имитационных моделей являются *метод Монте-Карло*, который отражает некий удачный случай.
Принято различать модели:

- детерминированные;
- вероятностные.

Использование *детерминированной модели* основывается на уверенности, что модель по значениям одних переменных позволяет точно вычислить значение других, при этом предлагается полный контроль за условиями внешней среды.
Детерминированные модели – есть предельный случай вероятностных моделей.
Вероятностные модели описывают поведение объектов как возможность реализации каких-либо событий с указанием вероятности таких исходов.
Построение модели – означает выбор вида модели и её уточнение с помощью полученных статистических данных. *Уточнение* – это вычисление значений некоторых параметров модели.

2.1.4. Предсказание

Предсказание связано с использованием построенных моделей для получения прогноза развития исследуемых явлений и поведения объектов.
При осуществлении предсказания указывается возможное будущее значение предсказуемой величины и интервал, в который случайная прогнозируемая величина попадает с заданной вероятностью. Эта вероятность и соответствующий ей интервал называется *доверительными интервалом и вероятностью.*

2.2. Основные понятия и модели статистического анализа

Основными понятиями статистического анализа являются:

1. Гипотеза;
2. Решающая функция и решающее правило;
3. Генеральная совокупность и её свойства;
4. Выборка из генеральной совокупности;
5. Оценка характеристик генеральной совокупности;

6. Доверительный интервал;
7. Доверительная вероятность;
8. Тренд;
9. Статистическая взаимосвязь.

Гипотеза. При проведении любых исследований, прежде всего, делаются некоторые предположения о свойствах изучаемых объектов. Такие предположения в статистическом анализе принято называть *гипотезами*. По своей природе гипотезы имеют априорный смысл, то есть предположения делаются до начала наблюдений, либо на основе их некоторого предварительного количества. Имеется в виду, что эксперименты, подтверждающие или опровергающие эту гипотезу, будут проведены позже. Результат проверки гипотезы является *апостериорным*, то есть послеопытным.
Решающая функция и решающее правило. Для проверки справедливости или несправедливости выдвинутой гипотезы разрабатываются специальные *решающие правила*, которые задают алгоритм проведения наблюдений и обработки их результатов. При обработке результатов вычисляются некоторые функции от наблюдений, которые часто называются статистическими. Так если $X_1, X_2, \ldots, X_n$ – некоторые экспериментальные значения, имеющие количественную природу, то функция от наблюдений $g(\vec{x}) = g(x_1 \ldots x_n)$ - *статистика.* В тех случаях, когда статистика предназначена для проверки справедливости выдвинутой гипотезы, она называется *решающей функцией*. При этом она полностью определяется решающим правилом и стремлением получить наиболее достоверный ответ на вопрос о справедливости гипотезы. Как правило наибольшую определенность имеет результат проверки справедливости гипотезы, когда ответом является вывод о противоречивости гипотезы наблюдаемым данным. Типичная формулировка такого исхода имеет вид: результаты наблюдений противоречат первоначальной гипотезе с уравнением значимости:

$$P_\alpha = 1 - \alpha; \quad 0 \le \alpha < 1 \qquad \textbf{(1)}$$

Чем ближе P_α к 1, тем более определённым (достоверным) является сформулированный вывод. Вероятность P_α - принято называть *доверительной вероятностью.*
Генеральная совокупность и её свойства. Понятие *генеральной совокупности* используется для обозначения исследуемого объекта в целом во всей совокупности его свойств. Обычно исследуемые объекты достаточно сложны и могут быть в бесконечно большом числе состояний. Таким образом, *генеральная совокупность* – есть нечто такое, что во всём многообразии его проявлений наблюдать невозможно. При этом почти всегда необходимо исследовать именно свойства генеральной совокупности. Эти свойства формулируются (описываются) на *абстрактном уровне*, например, генерируемое объектом наблюдение является значением случайной величины с некоторым законом распределения вероятности. Легко понять, что такое описание является гипотезой, так как имеет априорную природу.
Выборка из генеральной совокупности. В результате наблюдений за исследуемым объектом можно получить только ограниченное число значений регистрируемой величины, выбор которой обусловлен решающим правилом. Этот ограниченный набор значений называется *выборкой из генеральной совокупности*. Для обозначения генеральной совокупности обычно используются греческие буквы: $\mathcal{E}, X_1, \ldots, X_n$, где $\mathcal{E}$ - это генеральная совокупность, а $X_1, \ldots, X_n$ – некоторые идентификаторы, которым в результате изменений присваиваются конкретные числовые значения. Только выборочные значения из генеральной совокупности X_i, $i=1..n$, могут быть использованы для проверки справедливости гипотез о некоторых характеристиках, описывающих свойства генеральной совокупности. Свойства любых объектов возможно описать только с

помощью ограниченного количества понятий и формул, которые выражают некоторые совокупные свойства. Этот ограниченный набор понятий и формул называют *характеристиками генеральной совокупности.* Очевидно, что это совпадает с понятием модели.

Оценка характеристик генеральной совокупности. Используя выборку можно с некоторой точностью определить эти характеристики генеральной совокупности. В этом случае говорят об оценивании характеристик, а полученные в результате значения называют *оценками.* Значения оценки никогда не совпадает с точными значениями гипотетических характеристик.

Доверительный интервал. Так как по определению свойства генеральной совокупности являются неизвестными, в том смысле, что заранее выборочное значение точно предсказать нельзя, то для описания свойств генеральной совокупности естественно использовать вероятностные понятия, в частности саму генеральную совокупность в некоторых случаях естественно отождествлять со случайной величиной, а выборку – с конкретными её значениями, которые она принимает в результате проведения опытов с номерами *1, 2, ..., n*. Оценки характеристик генеральной совокупности, получаемые в результате вычислений, имеют вид статистик, то есть некоторой функции от выборочных значений:

$$g(\vec{X}) = g(x_1, \ldots, x_n) \qquad \textbf{(2)}$$

Так как до опыта выборочные значения неизвестны и являются случайными, то любая статистика как функция случайной величины, тоже будет случайной величиной. Её вероятностные свойства (функция плотности вероятности) полностью определяются вероятностными свойствами генеральной совокупности и видом формулы (функции от наблюдений). В некоторых случаях имеется возможность указать на числовой оси интервал, (возможно не сплошной) такой, что выполняется требование:

$$P\{\xi \in [a, b]\} = P_\alpha, \qquad \textbf{(3)}$$

где $P_\alpha = 1 - \alpha$;

P – вероятность события, заключающаяся в том, что случайная величина принимает значения из заданного интервала [a, b];

$P\alpha$ - конкретное число, удовлетворяющее условию :

$$0 \le P\alpha \le 1, \qquad \textbf{(4)}$$

а интервал [a, b] называется доверительным интервалом $D\alpha = [a, b]$.

Эти определения чаще всего используются для описания точности вычисления оценок с помощью некоторых статистик;

$$P\{ g(\vec{X}) \in [a; b]\} = P\alpha = 1 - \alpha. \qquad \textbf{(5)}$$

Допустим, что в результате вычислений получено некоторое число $g(\vec{X})$, тогда имеет смысл говорить о том, что «истинное» значение оцениваемой характеристики лежит в пределах этого интервала с вероятностью $P_{\alpha 1}$. Чтобы погрешности оценивания были наименьшими, необходимо использовать статистики с наименее возможным диапазоном изменения значений.

Таким образом, главной проблемой теории статистического анализа является получение формул для статистик с минимальным разбросом значений при оценивании и при проверке гипотез. Следует иметь в виду, что доверительные интервалы и доверительные вероятности определяются на основе выдвигаемых гипотез, тогда, если $g(\vec{X})$- решающая функция, выбранная на основе решающего правила, то попадание её вычисленных значений вне доверительного интервала можно признать за факт несправедливости выдвинутой гипотезы.

3. ВЕРОЯТНОСТНЫЕ МОДЕЛИ ОДНОМЕРНОЙ ГЕНЕРАЛЬНОЙ СОВОКУПНОСТИ

3.1. Адекватность вероятностных представлений

До проведения исследований основные свойства генеральной совокупности являются неизвестными, а сами исследования предназначены для установления этих свойств. Поэтому *адекватными* являются *вероятностные представления*, основу которых составляет понятие случайных событий, величин, вероятностей и их распределений.

3.2. Случайные события, вероятности, частоты

Случайное событие – такое событие, которое в результате некоторого эксперимента (наблюдения) может появиться или нет.
Если событие появляется всегда, то оно называется *достоверным*; если не появляется никогда – *невозможным*.
Следует иметь в виду, что событие появляется или нет при выполнении того, или иного комплекса условий.
- Если выполнение некоторого комплекса условий приводит к однозначному результату (появление или не появление некоторого события), то такое событие называется *детерминированным* (вполне определённым);
- если при неизменных условиях, которые не полностью определяют поведение некоторого события, регистрируются факты его появления в цепи экспериментов

1, 2,3, ...,n, то можно получить некоторое количество $\nu_A(n)$ этих фактов. Рассмотрим отношение:

$$\tilde{P}_A(n) = \nu_A(n)/n \qquad \textbf{(6)}$$

Очевидно, что выполняется неравенство:

$$0 \le \tilde{P}_A(n) \le 1 \qquad \textbf{(7)}$$

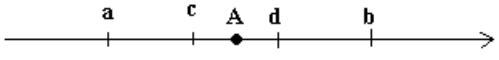
Если при выполнении данного комплекса условий и $n\to\infty$ $\tilde{P}_A(n) = \nu_A(n)/n$ стремится к некоторому пределу P_A, то этот предел называется *вероятностью*, $\tilde{P}_A(n)$ *–частота.*
Очевидно, что её всегда можно подсчитать для любой серии экспериментов. Вероятность - является неизвестной и часто даже неизвестно, существует ли этот предел. Поэтому существование вероятности некоторого события в виде определённого числа удовлетворяющего неравенству:

$$P_A = P, \quad 0 \le P \le 1 \qquad \textbf{(8)}$$

является одной из гипотез, используемых при статистическом анализе.

3.3. Случайные величины и связанные с ними случайные события

Случайная величина – переменная, принимающая значение из некоторого диапазона числовой оси [a, b].
Имея в виду <u>главную цель</u>: предсказание поведения случайной величины, наиболее часто рассматривают событие попадания её значений в некоторый интервал числовой оси.



Тогда событие A означает, что $\varepsilon \in [c, d]$, где ε - случайная величина. С этими событиями связана вероятность

$$P_A = P\ \{\varepsilon \in [c, d]\}. \quad (9)$$

Основной проблемой является поиск такого интервала минимальной длины, когда выполняется условие:

$$|d - c| \to \min \quad (10)$$

$$P_A = P\ \{\varepsilon \in [c, d]\} = P \quad (11)$$

Часто используется альтернативная формулировка проблемы: необходимо указать границы интервала фиксируемой ширины ∇:

$$d - c = \Delta, \quad (12)$$

такого, что имеет место

$$P\ \{\varepsilon \in [c, d]\} \to \max. \quad (13)$$

При предсказании значений случайной величины ошибки предсказания характеризуются двумя параметрами:
-Шириной задаваемого интервала возможных значений, которые характеризуют абсолютную погрешность предсказания;
-вероятностью попадания значения случайной величины в этот интервал, которая характеризует надёжность прогноза.
Понятие случайной величины формулируется в теории вероятностей и является вероятностной моделью генеральной совокупности.

3.4. Функция распределения вероятности и функция плотности вероятности – как количественные модели генеральной совокупности

Вероятности событий заключающихся в попадании значений случайной величины на некоторые интервалы числовой оси, чаще всего вычисляются с помощью *функции плотности вероятности (ФПВ)* или *функции распределения вероятности (ФРВ)*.
ФПВ – является количественной моделью, предназначенной для описания поведения генеральной совокупности.
ФРВ – называется числовая функция вида

$$F_\xi(x) = P\{\ \xi\ < x\}, \quad (14)$$

где ξ - случайная величина,
x – некоторая точка числовой оси,
$P\{\}$ – вероятность события.
Из определения (14) и свойств вероятности событий следуют свойства ***ФРВ***:

$$0 \le F_\xi(x) \le 1 \quad (15)$$

$$F_\xi(x_1) = F(x_2) \quad \text{при } x_2 > x_1 \quad (16)$$

- не убывающая,

$$F_\xi(-\infty) = 0 \quad (17)$$

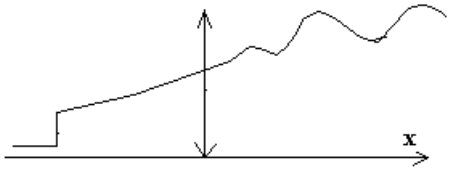
– вероятность невозможного события,

$$F_\xi(\infty) = 1 \quad (18)$$

– вероятность достоверного события.
Чаще в вероятностных расчетах и в теоретических исследованиях используется понятие ***ФПВ***:

$$W_\xi(x) = \frac{dF_\xi(x)}{dx} \quad (19)$$

Предполагается, что ***ФПВ*** – дифференцируема хотя бы в смысле δ - функции, необходимость использования которой возникает когда ***ФРВ*** имеет разрывы



$$W_\xi(x) = \frac{dF_\xi(x)}{dx} = \lim \frac{F_\xi(x+\Delta) - F_\xi(x)}{\Delta} \qquad \textbf{(20)}$$

Из (20) и (15 – 18) следуют свойства ***ФПВ***:

$$W_\xi(x) \geq 0 \qquad \textbf{(21)}$$

- производная неубывающей функции - неотрицательна

$$\int_{-\infty}^{\infty} W_\xi(x)dx = 1 \qquad \textbf{(22)}$$

– это соотношение легко вывести из представления:

$$F_\xi(x) = \int_{-\infty}^{x} W_\xi(t)dt\,, \qquad \textbf{(23)}$$

с учетом свойства (18).

Непосредственно из определений (14) и (23) имеем:

$$P\{\xi \in [c, d]\} = \mathbf{F}_\xi(d) - F_\xi(c) \qquad \textbf{(24)}$$

$$P\{\xi \in [c, d]\} = \int_{c}^{d} W_\xi(t)dt \qquad \textbf{(25)}$$

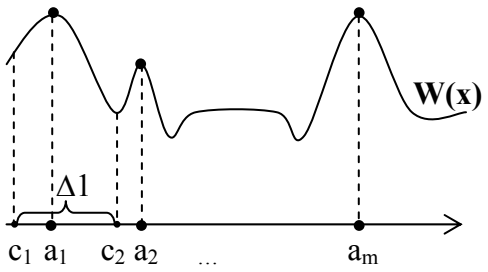
- последнее соотношение чаще всего используется для решения задач вида (10), (11) и (12), (13).

3.5. Числовые характеристики генеральных совокупностей

Характеристики положения случайной величины

В поведении случайной величины часто оказывается целесообразным указать некоторые точки числовой оси, которые имеют значимость в аспекте характерности этого поведения, так что, зная эти точки, можно до определённой степени составить представление о свойствах случайной величины. Эти характеристики положения так или иначе описываются с помощью ФПВ. Основными из них являются следующие:

Мода – точка на числовой оси, в которой функция плотности вероятности принимает локальное или глобальное экстремальное значение типа “максимум”.



Распределение вероятности с одной модой называется *унимодальным*, в противном случае – *полимодальным.*
Положение моды (максимум) определяет участки повышенной вероятности попадания в них значений случайной величины так как имеет место приближенное равенство:

$$P\ (x \in \Delta_1) = \int_{c_1}^{c_2} W(t)\ dt \approx W(a_1) \cdot \Delta_1,\ a_2 \in [c_1, c_2] \qquad \textbf{(26)}$$

Медиана – точка, которая делит область определения ФПВ следующим образом:

$$\int_{-\infty}^{M} W_\xi(x)dx = \int_{M}^{\infty} W_\xi(x)dx = F_\xi(M) = 0.5, \qquad \textbf{(27)}$$

то есть определяет два интервала (слева и справа), вероятности попадания в которые значения случайной величины одинаковы и равны 0,5.
Определение некоторых характеристик положения связано с понятиями моментов случайной величины. *Моментом к*-го *порядка случайной величины* относительно точки «***c***» числовой оси называется:

$$E\ [(\xi - c)^k] = \int_{-\infty}^{\infty} (x-c)^k \cdot W_\xi(x)dx\,. \qquad \textbf{(28)}$$

Абсолютные моменты:

$$E\ [|\xi - c|^k] = \int_{M}^{\infty} |x-c|^k * W_\xi(x)dx \qquad \textbf{(29)}$$

Соотношения (28) и (29) являются частными случаями понятия математического ожидания некоторой функции случайной величины

$$E\ [\varphi(\xi)] = \int_{-\infty}^{\infty} \varphi(x) W_\xi(x)dx \qquad \textbf{(30)}$$

В статистическом анализе часто приходится рассматривать различные функциональные преобразования, поэтому соотношение вида (30) используется достаточно часто. Если в соотношении (28) с = 0, то определяемые им моменты называются *начальными.* Часто используется понятие 1–го начального момента, когда к =1, с =0 и имеет место следующее соотношение:

$$m_\xi = E\ [\xi] = \int_{-\infty}^{\infty} W_\xi(x)dx \qquad \textbf{(31)}$$

Параметр m_ξ - называется математическим ожиданием случайной величины.
Если в соотношении (28) «***c***» заменить на m_ξ, то есть положить:

$$E[(\xi - m_\xi)^k] = \int_{-\infty}^{\infty} (x-m_\xi)^k W_\xi(x)dx\,, \qquad \textbf{(32)}$$

то это соотношение определяет *центральные моменты.*
Наименование «момент» ведёт своё происхождение, по аналогии с механикой. В частности соотношение (31) определяет центр тяжести фигуры, ограниченной числовой осью и ФПВ, поэтому называется *центральным моментом*.

Важнейшее значение в вероятностном моделировании имеет понятие *второго момента*, когда в соотношении (28) к = 2

$$E = [(\xi - c)^2] = \int_{-\infty}^{\infty} (x-c)^2 W_\xi(x)dx \qquad \textbf{(33)}$$

Этот момент определяет степень рассеивания значения случайной величины относительно точки «с», так как $(x - c)^2$ – квадрат расстояния от x до с. Таким образом, соотношение (33) определяет собой средний квадрат расстояния значений случайной величины от избранной точки «с».

3.6. Экстремальное свойство математического ожидания

С точки зрения наиболее надежного предсказания будущего значения случайной величины очень важно указать такую точку «с», относительно которой квадрат расстояния вида (33) будет минимальным, то есть в смысле этой меры значение случайной величины имеет наименьшее отклонение от этой точки. Найдём такую точку. Для этого левую часть соотношения (33) преобразуем к виду:

$$E[(\xi - c)^2] = E[(c- m_\xi+m_\xi-c)^2] = E[(\xi -m_\xi)^2 + 2(m_\xi-c)E(\xi-m_\xi) + (m_\xi-c)^2] =$$

$$= E[(\xi-m_\xi)^2] + (m_\xi-c)^2 \qquad \textbf{(34)}$$

Здесь от «с» зависит последнее слагаемое, которое положительно и равно 0, когда выполняется условие:

$$c = m_\xi. \qquad \textbf{(35)}$$

Очевидно, что только в этом случае правая часть выражения (34) достигает минимума, то есть имеет место равенство:

$$E[(\xi-m_\xi)^2] = \min E[(\xi - c)^2] \qquad \textbf{(36)}$$

То есть относительно математического ожидания будет минимальным среднеквадратический разброс значений случайной величины.

3.7. Дисперсия, среднеквадратическое отклонение и другие моменты, как меры разброса значений случайных величин

Левая часть (36) получила название *дисперсии.* Для неё часто используется специальное обозначение:

$$\sigma_\xi^2 = E[(\xi -m_\xi)^2] = \int_{\infty}^{\infty} (x- m_\xi)^2 W_\xi(x)dx \qquad \textbf{(37)}$$

В качестве меры разброса используется понятие *среднеквадратического отклонения (СКО)*

$$\sigma = \sqrt{E[(\xi - m_\xi)^2]} \qquad \textbf{(38)}$$

Размеры интервалов, определяющих разброс значений случайной величины часто задаётся в единицах σ.

Полученное выше экстремное свойство математического ожидания является одним из самых главных причин использования его в качестве прогноза будущих значений случайной величины, то есть $\hat{\xi} = m_\xi$. Кроме такого прогноза часто задаётся интервал:

$$\xi : \hat{\xi}, \xi \in [m_\xi - a\sigma_\xi, m_\xi + b\sigma_\xi] \quad \textbf{(39)}$$

при этом a и b выбираются, таким образом, вероятность выполнения условия (39) была задана какому-либо числу: P{ (39) } = Pq

$$Pq = P\ \{\xi \in [m_\xi - b\sigma_\xi, m_\xi + a\sigma_\xi]\}$$

“a” и “b” в общем случае определяются неоднозначно. Для доверительной вероятности:

$$Pq = \int_{m_\xi - a\sigma_\xi}^{m_\xi + b\sigma_\xi} W\xi(x)dx \quad \textbf{(40)}$$

Так что, если границы интервала заданы, то можно её вычислить. Но функция плотности вероятности не всегда известна. В этих условиях для оценки доверительных вероятностей можно использовать неравенство Чебышева:

$$P\{|\xi - m_\xi| \le k\sigma_\xi\} \ge 1 - \frac{1}{k^2} \quad \textbf{(41)}$$

То есть:

$$Pq \ge 1 - \frac{1}{k^2} \quad \textbf{(42)}$$

Для интервала:

$$[m_\xi - k\sigma_\xi,\ m_\xi + \sigma_\xi] \quad \textbf{(43)}$$

Если известно, что ФПВ обладает симметрией относительно математического ожидания и является одновершинной, то справедливо *неравенство Гаусса*:

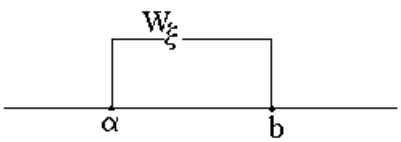
$$P\{|\xi - m_\xi| \ge k\ \sigma_\xi\} \le \frac{4}{9k^2} \quad \textbf{(44)}$$

Неравенство Чебышева и Гаусса подчеркивают значение математического ожидания и среднеквадратического отклонения для решения задач статистического анализа, так как с их помощью можно получить ответ на основной вопрос: «как будет себя вести генеральная совокупность?»

3.8. Основные виды функции плотности вероятности

1) *Равномерное распределение*:

$$W_\xi(x) = \begin{cases} \frac{1}{b-a}; a \le x \le b \\ 0, (x \le a) \cup (x \supset b) \end{cases}$$



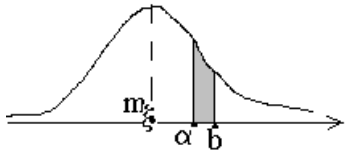
$$m_\xi = \frac{b+a}{2} \quad \textbf{(45)}$$

$$\sigma^2_\xi = \frac{(b+a)^2}{12} \quad \textbf{(46)}$$

Модель равномерного распределения выражает наибольшую неопределённость относительно протекающих процессов, так как имеется в виду равновероятность попадания значений случайной величины в интервалы одинаковой длины, если они находятся в области допустимых значений [a, b].

2) *Гауссово распределение*, используется в статистических исследованиях гораздо чаще:

$$W_\xi(x) = \frac{1}{\sqrt{2\pi}\ \sigma_\xi} \exp\left[-\frac{(\varepsilon - m_\xi)^2}{2\sigma_\xi^2}\right] \quad \textbf{(47)}$$



$$P\{\ \xi \in [a;\ b]\} = \int_a^b W_\xi(x)dx\ = F\left[\frac{b - m_\xi}{\sigma_\xi}\right] - F\left[\frac{a - m_\xi}{\sigma_\xi}\right] \quad \textbf{(48)}$$

$$F\ (c) = \frac{1}{\sqrt{2\pi}}\ \int_{-\infty}^{c} \ell^{\frac{-t^2}{2}} dt \quad \textbf{(49)}$$

Определённый отношением (49) интеграл называется *интегралом вероятности*. Его значения табулированы. Гауссово распределение является наиболее распространенной моделью. Во многом это является следствием центральных теорем

$$\xi = \sum_{i=1}^{n} \xi\ i \quad \textbf{(50)}$$

3) *Экспоненциальное распределение.*

$$W_\xi\ (x) = \begin{cases} \lambda e^{-\lambda x}, x \geq 0 \\ 0, x \subset 0 \end{cases} \quad \textbf{(51)}$$

Это распределение является хорошей моделью распределения длин интервалов между сбоями в работе некоторых приборов, между запросами на выполнение некоторых операций и т. д.

Основной моделью поведения генеральной совокупности в статистическом анализе является Гауссово распределение. Именно в предположении гауссовости удается построить вычислительные формулы, позволяющие осуществить обработку статистических данных с целью ответа на основные вопросы, и прежде всего каково будет поведение исследуемого объекта в будущем.

С помощью введенных выше понятий и моделей можно уточнить содержание этапов статистического анализа.

Для описания закономерностей поведения исследуемых объектов целесообразно использовать вероятностные модели, в том числе такие понятия как характеристики положений и моменты.

Сбор эмпирических данных должен быть спланирован таким образом, чтобы обеспечивалась возможность отобрать наиболее адекватную модель (описание) их

генерации. В частности существенное значение имеет достоверность выводов статистического анализа, которая определяется объемом данных.
Отбор наиболее адекватных моделей осуществляется на этапе анализа данных. Конкретные алгоритмы анализа во многом определяются видом гипотетической модели. Поэтому, прежде всего, необходимо убедиться в том, что имеющиеся данные не противоречат самым простым исходным предположениям. Этой цели служит так называемый разведочный анализ данных. Такое наименование подчеркивает также возможность получить некоторый представления о свойствах данных, с тем, чтобы более определенно сформулировать гипотезы о конкретных моделях их порождения.

Выше были перечислены некоторые из основных гипотез статистического анализа:

- генеральная совокупность является случайной величиной, свойства которой с течением времени не меняется (*стационарность*);
- случайная величина подчиняется некоторому закону распределения вероятности, например Гауссовому;
- генеральная совокупность генерирует однородный набор данных.

Задача статистического анализа: уточнение описания свойств генеральной совокупности, в частности определение вида ФПВ и значений параметров, которым она определяется.
Например: для равномерного распределения этими параметрами являются границы [a;b], для гауссового распределения - математическое ожидание и дисперсия.
Основные процедуры оценивания (то есть определения характеристик по экспериментальным данным) разработаны, и, следовательно, являются адекватными только при выполнени перечисленных выше условий, прежде всего стационарности и однородности.
Поэтому прежде чем использовать процедуры оценивания следует убедиться в справедливости этих гипотез, то есть убедиться в том, что они не противоречат имеющимся данным. Для этих целей служит *разведочный анализ данных*

4. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

Инструменты разведочного анализа данных:

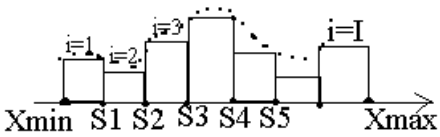
1. **Гистограмма и полигоны частот.**
2. **Графический анализ.**
3. **Отбраковка аномальных измерений.**
4. **Фильтрация с целью выделения трендов, которые не являются постоянными величинами.**

Построение диаграмм способствует визуальной оценки вида вероятностных распределений.
Гистограмма – набор столбцов, высота которых равна отношению:

$$Pi \approx \upsilon_i/n, \qquad \textbf{(52)}$$

$$x_{max} = x_{max}\{x_k\}$$

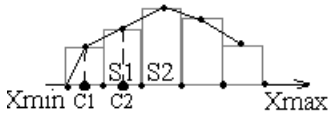
где υ_i - количество попаданий значений выборки из n – элементов в i – ый интервал.



$$X_{min} = \min\{x_k\};\ k = 1,2,\ldots,\ n$$

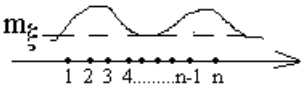
По виду гистограммы можно сделать предварительное заключение о типе закона распределения вероятностей.

Полигон частот – ломаная линия, соединяющая середины столбцов гистограмм:



$$Pi = \int_{-\infty}^{\infty} W\xi(x)dx \approx W_{\xi}\ (c_i) * Di$$

Но этот вывод обоснован в том случае, когда свойства генеральной совокупности с течением временем не меняется, и когда выборка однородна:

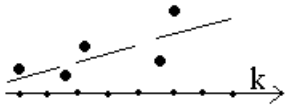


Простейшие признаки нестационарности заключаются в том, что изменяется некоторый параметр, относительно которого значения случайной величины имеют разброс. Этот параметр называется *трендом*. В стационарном случае должна иметь место некоторая постоянная величина (математическое ожидание), относительно которой случайная величина имеет разброс. Мерой разброса является среднеквадратическое отклонение, которое должно быть неизменно, то есть должно иметь место.

$$\sigma_{\xi} = \text{const}; \qquad \textbf{(53)}$$

$$m_{\xi} = \text{const} \qquad \textbf{(54)}$$

Графический анализ заключается в нанесении на график значений случайной величины, зарегистрированных при проведении экспериментов. При этом может оказаться, что некоторые значения сильно отличаются от остальных, а это может свидетельствовать об аномальности такого наблюдения.



С другой стороны график может выражать некоторую тенденцию к увеличению или к уменьшению (систематическому) значений наблюдений с увеличением их номера. В обоих случаях можно высказать гипотезу о неоднородности результатов наблюдения и следует применять специальные приёмы для устранения этой неоднородности. В частности, наличие тенденции к увеличению или к уменьшению принято характеризовать как *тренд значений* и это относится к разделу «Временные ряды».

Отбраковка аномальных измерений может быть осуществлена на базе получения оценок числовых характеристик случайной величины.

5. ОЦЕНИВАНИЕ ХАРАКТЕРИСТИК СЛУЧАЙНОЙ ВЕЛИЧИНЫ ПРИ ОДНОРОДНЫХ ДАННЫХ

Основные гипотезы:

- поведение случайной величины с течением времени не изменяется;
- случайная величина (генеральная совокупность) обладает некоторой функцией плотности вероятности;
- существуют моменты: математическое ожидание, дисперсия и, возможно, высшие моменты.

Задача: оценить эти характеристики по набору наблюдаемых значений, которые называются *выборкой*. При этом важное значение имеет ответ на вопрос: «с какой погрешностью оцениваются характеристики и каким образом можно эти погрешности уменьшить?»
Для определения вида функции плотности вероятности используется *метод подгонки к гистограмме* некоторой гипотетической функции из наиболее часто используемых классов функции плотности вероятности (гауссово, равномерное, экспоненциальное и т. д.). Процедура подгонки будет рассмотрена в разделе «Критерии согласия», так как эта задача относится к процедурам проверки гипотез.
Предполагается, что для оценивания характеристик используется выборка конечного объема $X_1 \ldots X_n$, где X_i- выборочное значение случайной величины при *i*- том наблюдении.

5.1. Задача оценивания характеристик

Однородность выборки в данном случае означает выполнение следующих условий:

$$1) \quad E[\xi_1] = E[\xi_2] = \ldots = E[\xi_n] = m_\xi \qquad \textbf{(55)}$$

$$2) \; E[(\xi_i - m_\xi)^2] = \sigma_\xi^2 = \text{const}, \; i = 1, 2, \ldots, n \qquad \textbf{(56)}$$

Здесь ξ_i- будущее значение случайной величины в *i*- тый момент наблюдения.
Любое оценивание предполагает осуществление некоторых вычислений с использованием выборочных значений, то есть, если оценивается некоторая характеристика «а», то предполагается получение её оценки как некоторой функции от наблюдений.

$$a \Rightarrow \hat{a}_m = q(x_1, \ldots, x_n) \qquad \textbf{(57)}$$

Эти функции называются *статистиками*. Класс статистик не может быть произвольным, так как при оценивании желательно достичь наилучшего соответствия с оцениваемой величиной. Понятие *соответствие* не выражается одним каким-либо термином и по существу означает (имеет смысл) качество оценки. В процессе становления статистического анализа, как способа исследования явлений и процессов на основе генерируемых ими данных, выработались определённые требования, которые характеризуют качество получения оценок. Эти требования относятся не только к оцениванию математического ожидания и дисперсии. По существу, они формулируются в виде условий, которым должны удовлетворять *оценивающие функции*. В некоторых случаях на основе этих условий можно построить оценивающие функции в явном виде.

5.2. Общие требования к оценивающим функциям

Иногда в литературе понятие «оценки» и «оценивающей функции» отождествляются и поэтому говорят о *требованиях к оценкам*. Вместе с тем, следует иметь ввиду, что оценивающие функции получаются (выбираются) априорно, то есть, ещё до проведения опыта, а оценка – некоторое число, которое получается с использованием оценивающей функции от выборочных значений.

Обычно, оценивающим функциям предъявляются следующие требования:
1) *Требование несмещённости*. Так как оценка вычисляется как функция случайной величины, то она до опыта должна рассматриваться как случайная величина с некоторой функцией плотности вероятности. Обычно, оценивающие функции обладают математическим ожиданием и дисперсией. Требование означает выполнение условия:

$$E[g(x_1, \ldots, x_n)] = E[\hat{a}_n] = a, \quad \textbf{(58)}$$

то есть математическое ожидание оценки должно быть равно оцениваемой величине. Это требование в частности отражает тот факт, что математическое ожидание обладает экстремальными свойствами (35), и следовательно разброс значений оценки (погрешность) в среднеквадратическом смысле будет наименьшим.

2) *Требование состоятельности* – означает, что при

$$\mathbf{n \to \infty \Rightarrow \ \hat{a}_n \to a,} \quad (59)$$

то есть оценка сходится к оцениваемой величине при неограниченном увеличении длины выборки, часто это понимается в смысле сходимости по вероятности:

$$\lim_{n\to\infty} P\left\{|\hat{a}_n - a| \subset \varepsilon\right\} = \mathbf{1}, \ \varepsilon > 0 \quad (60)$$

Это соотношение читается следующим образом: каково бы ни было $\varepsilon > 0$, при неограниченном увеличении числа наблюдений вероятность того, что разность $|\hat{a}_n - a|$ станет $< \varepsilon$ стремиться к 1. Более сильный вид сходимости – это сходимость среднеквадратическая, которая имеет смысл для несмещённых оценок, для которых можно определить дисперсию оценки:

$$\sigma^2 a_n = E[(\hat{a}_n - a)^2]; \ a = E[\hat{a}_n] \quad \textbf{(61)}$$

$$\lim_{n\to\infty} \sigma^2 a_n = 0 \quad \textbf{(62)}$$

То есть, если дисперсия оценки стремится к нулю, то это называется *среднеквадратической сходимостью.* Из неравенства Чебышева следует, что из сходимости в среднеквадратическом смысле следует сходимость и по вероятности.
3) *Требование эффективности.* Качество оценивания можно охарактеризовать в среднеквадратическим отклонением, если оценки несмещённые. Если имеется несколько оценочных функций для одной и той же величины «а» и при этом выполняется требование несмещённости, то можно вычислить и соответствующие дисперсии.
Оценка называется *эффективной*, если оценочная функция обладает наименьшей дисперсией. Таким образом, среди всех оценочных функций следует выбирать такую, которая удовлетворяет требованию:

$$E[(g_i(x_1, x_2, \ldots, x_n) - a)^2] = \min \quad \textbf{(63)}$$

На практике не всегда удается выполнить это требование. Но требование несмещённости и требование самостоятельности часто можно проверить непосредственно, особенно если предположить, что генеральная совокупность является гауссовой.

5.3. Основные методы построения оценочных функций

1. *Метод аналогий* – предполагает использование некоторых смысловых особенностей. Например:

- основные моменты – это результаты усреднений знаний случайной величины с весом в виде функции плотности вероятности, например

$$E[\xi] = \int_{-\infty}^{\infty} x W\xi(x) dx .$$

Аналогией может служить:

$$\bar{x} = 1/n \sum_{i=1}^{n} x_i \rightarrow g(x_1, x_2, \ldots, x_n) = 1/n \sum_{i=1}^{n} x_i \quad . \qquad \textbf{(64)}$$

- В свою очередь дисперсия = среднее значение квадрата отклонения:

$$\sigma^2{}_n = 1/n \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \qquad \textbf{(65)}$$

Аналогия заключается в выполнении процедуры усреднения, причем усреднение с функцией плотности вероятности заменяется на усреднение по количеству наблюдений. При этом сохраняется вид функционального преобразования данных (для дисперсии это квадрат разности).
Свойства оценки в виде среднего (64)
Для оценочной функции вида (64) нетрудно получить:

а) Смещение: $E[\bar{x}] = E[1/n \sum_{i=1}^{n} x_i] = 1/n \sum_{i=1}^{n} E[x_1] = 1/n \sum_{i=1}^{n} m_\xi = m_\xi$,

то есть

$$E[\bar{x}] = m_\xi \qquad \textbf{(66)}$$

Таким образом, оценка вида (64) является несмещённой.
б) Дисперсия оценки:

$$\sigma_{\bar{x}} = E[(\bar{x} - E[\bar{x}])^2] = |\text{из}(66)| = E[(\bar{x} - m_\xi)^2] = E[(1/n \sum_{i=1}^{n} (x_i - m_\xi)^2)] = 1/n^2 \sum_{i,k=1}^{n} R_{ki} \qquad \textbf{(67)}$$

R_{ik} – корреляционный момент

$$R_{ik} = E[(x_k - m_\xi)(x_i - m_\xi)]. \qquad \textbf{(68)}$$

Если ввести ковариационный момент

$$\rho_{i\,k} = R_{ik} / \sigma_\xi^2, \qquad \textbf{(69)}$$

который обладает свойством

$$-1 \le \rho_{i\,k} \le 1, \qquad \textbf{(70)}$$

то (67) дает:

$$\sigma^2{}_{\bar{x}} = \sigma^2{}_\xi \cdot \sum_{i,k=1}^{n} \rho_{i,k} \,/ n^2. \qquad \textbf{(71)}$$

Величина называется *некоррелированной*, когда выполняется условие:

$$\rho_{i\,k} = \begin{cases} 1, i = k \\ 0, i \neq k \end{cases} \qquad \textbf{(72)}$$

Для некоррелированной величины из соотношения (71) получаем:

$$\sigma^2{}_{\bar{x}} = \sigma^2{}_\xi / n \qquad \textbf{(73)}$$

Очеквидно, что при $n \rightarrow \infty$ имеет место $\sigma^2{}_{\bar{x}} \rightarrow 0$, то оценка (64) – состоятельна.
В любом случае соотношение (71) даёт:

$$\sigma_{\bar{x}} = \sigma_{\xi} / \mathrm{n} \cdot \sqrt{\sum_{i,k=1}^{n} \rho_{|i-k|}} , \qquad (75)$$

так как

$$\rho_{ik} = \rho_{|i-k|}. \qquad (76)$$

Соотношение (76) – является следствием того, что значения x_k, x_i - получены из одной и той же генеральной совокупности. Так как в обычных условиях предел

$$\lim_{|i-k|\to\infty} \rho_{|i-k|} = 0, \qquad (77)$$

то оценка математического ожидания вида (64) является состоятельной не только для некоррелированных значений, то есть когда (72) – не обязательно выполняется.

5.4. Оценивание дисперсии

Естественной аналогией понятия дисперсии случайной величины является:

$$\tilde{\sigma}_{\xi}^{\,2} = 1/\mathrm{N} \sum_{i=1}^{n} (x_i - \tilde{m}_n)^2 , \qquad (78)$$

где $\tilde{m}_n$ - оценка математического ожидания;

N – целое число.

В качестве оценки математического ожидания можно использовать любую функцию, любое число, которое удовлетворяет определённым требованиям. Естественно использовать в качестве оценки математического ожидания функцию (64), то есть:

$$\tilde{m} = \bar{x} \qquad (79)$$

Определим N из условия достижения несмещённости, то есть потребуем выполнения:

$$\mathrm{E}[\tilde{\sigma}_{\xi}^{\,2}] = \mathrm{E}[1/\mathrm{N} \sum_{i=1}^{n} (x_i - \bar{x})^2] = \tilde{\sigma}_{\xi}^{\,2} . \qquad (80)$$

Покажем, что для выполнения условия (80) при некоррелированных значениях, должно иметь место N = n-1; то есть ***несмещённая оценка дисперсии*** следовательно должна иметь вид:

$$\tilde{\sigma}_{\xi}^{\,2} = 1/(\mathrm{n}-1) \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad (81)$$

Действительно, соотношение (80) с учетом определения (64) и (72) дает

$$E\left[\sigma_{\xi}^2\right] = \frac{n-1}{N} \sigma_{\xi}^2 .$$

Из условия $\frac{n-1}{N} = 1$ и следует справедливость утверждения, что: дисперсия оценки дисперсии равна

$$\sigma^2_{\tilde{\sigma}_{\xi}^2} = \mathrm{E}[(\tilde{\sigma}_{\xi}^{\,2} - \sigma^2)]^2 = \frac{2\sigma_{\xi}^4}{n} \qquad (82)$$

– для некоррелированных гауссовых случайных величин.

Погрешности оценивания практически всех характеристик случайных величин имеют ***асимптотический закон изменения*** вида:

$$\sigma_n \sim 1/\sqrt{n}, \quad n>>1 \qquad \textbf{(84)}$$

если под *мерой погрешности* понимать среднеквадратическое отклонение оценки.
Этот закон демонстрируют соотношения (73) и (82). Таким образом, для уменьшения погрешности в 10 раз необходимо увеличить количество обрабатываемых значений в 100 раз.
Этот аспект статистического анализа является его главным недостатком.
Использование нескольких значений случайной величины для оценивания её характеристик требует для полного исследования свойств, получаемых оценок, рассмотрение так называемых ***многомерных функций плотности вероятности:*** $W_{n\xi}(x_1, x_2,\ldots, x_n)$.
Кроме того, это понятие используется при построении вычислительных алгоритмов, для оценивания параметров генеральной совокупности на основе так называемого ***метода максимального правдоподобия***.
При проведении статистического анализа, в основном, используется одна из двух гипотез:
- предположение о независимости значений генеральной совокупности в отдельных экспериментах, при этом многомерная функция плотности вероятности выражается в виде произведения:

$$W_{n\xi}(x_1, x_2,\ldots, x_n) = \prod_{i=1}^{n} W_{\xi}(x_i) \qquad \textbf{(85)}$$

- значение генеральной совокупности считается зависимым, но функция плотности вероятности является гауссовой:

$$W_{n\xi}(\vec{x}) = \frac{1}{2\pi^{n/2} \cdot \sqrt[n]{\det R} \cdot \exp\left[-\frac{1}{2}(\vec{x}-\vec{m})\right] \cdot R^{-1}(\vec{x}-\vec{m})}, \qquad \textbf{(86)}$$

где $\vec{x} = (x_1, x_2,\ldots, x_n)$,
$\vec{m} = (m_1, m_2,\ldots, m_n)$,
$m_i = E[\xi_i]$,
$R = \{R_{ik}\}$, i, k = 1, …, n.
Замечание: ещё один способ описать многомерную функцию плотности вероятности заключается в использовании условной функции плотности вероятности:

$$W_{n\xi}(\vec{x}) = W_{\xi}(x_1)\, W(x_2 \mid x_1) \ldots W(x_n \mid x_1, x_2,\ldots, x_{n-1}), \qquad \textbf{(87)}$$

и в предположение выполнимости марковости свойства:

$$W(x_к \mid x_{k-1} \ldots x_1) = W(x_k \mid x_{k-1}), \qquad \textbf{(88)}$$

такие процессы называются ***процессы без последствия***, так как условие (87) говорит о том, что будущее определяется только настоящим.

5.5. Метод максимального правдоподобия

Этот метод используется широко как в теоретических, так и в прикладных исследованиях. Он разработан Р.Фишером.
Метод предназначен для оценивания численных значений некоторых параметров, от которых зависит функция плотности вероятности.
Следует иметь в виду, что не для всех видов функций плотности вероятности метод максимального правдоподобия применим. Одно из главных условий заключается в том, что производные функции плотности вероятности по каждому из оцениваемых параметров должны существовать во всём диапазоне изменений значений случайной величины. Например, в равномерной функции плотности вероятности производные не существуют на концах интервалов, поэтому этот метод здесь не применяется.

<u>Основная идея Фишера заключается в следующем</u>. Пусть известна n – мерная функция плотности вероятности, которая зависит от m – параметров, в том смысле, что задана формула, описывающая эту зависимость, в которую входят символы, означающие эти параметры (речь идет о многомерной функции плотности вероятности)

$$W_n (x_1, x_2, \ldots, x_n / a_1, a_2, \ldots, a_m) \quad \textbf{(89)}$$

При использовании этой формулы (функции) для описания поведения случайной величины x_i считаются неизвестными величинами, а вектор параметров $\vec{a} = (a_1, \ldots, a_m)$ – задается, например, с целью вычисления функции плотности вероятности в некоторой точке:

$$\vec{x} = (x_1, \ldots, x_n).$$

В задачах статистического оценивания параметры являются неизвестными и подлежат оценке по зарегистрированным наблюдениям $x_1, \ldots, x_n$. То есть в данном случае $a_1, \ldots, a_m$ – неизвестны, а $x_1, \ldots, x_n$ – известны.

Для оценки параметров Фишер ввел в рассмотрение ***функцию правдоподобия***

$$L (a_1, \ldots, a_m / x_1, \ldots, x_n) = W_n (x_1, x_2, \ldots, x_n / a_1, a_2, \ldots, a_m) \quad \textbf{(90)}$$

В качестве наиболее подходящей оценки Фишер предложил использовать такой вектор: $\hat{\vec{a}} = (\hat{a}_1, \ldots, \hat{a}_m)$, который при заданном векторе наблюдений удовлетворяет условию:

$$L (\vec{a}_n / \vec{x}) = \max L(\vec{a} / \vec{x}) \quad \textbf{(91)}$$

при любом $\vec{a}$ из некоторого множества.

Таким образом, в качестве наиболее подходящей оценки вектора неизвестных параметров предполагается использовать такой, который при заданном векторе значений выборки даёт максимум *функции правдоподобия* на множестве допустимых векторов значений параметров $\vec{a}$.

В целях удобства построения вычислительных формул чаще рассматривается ***<u>логарифмическая функция правдоподобия</u>***:

$$l(\vec{a} / \vec{x}) = \ln L(\vec{a} / \vec{x}) \quad \textbf{(92)}$$

Так как ln – является монотонной функцией аргумента, то ln(z) = max - в точке - z = max, то есть значение векторов оценок, дающих максимум функции правдоподобия и логарифмической функции правдоподобия, совпадают:

$$l(\vec{a} / \vec{x}) = \max l(\vec{a} / \vec{x}) \quad \textbf{(93)}$$

Удобство перехода к ln главным образом связано с понятием независимости, когда функция плотности вероятности представима в виде произведения:

$$W_n (x_1, x_2, \ldots, x_n / a_1, a_2, \ldots, a_m) = \prod_{i=1}^{n} W_\xi(x_i/a_1, a_2, \ldots, a_m) \quad \textbf{(94)}$$

Тогда имеет место представление:

$$\boldsymbol{l (\vec{a} / \vec{x}) = \ln L (\vec{a} / \vec{x}) = \sum_{i=1}^{n} \ln W_\xi(x_i / \vec{a}).} \quad (95)$$

Пример гауссовой генеральной совокупности с независимыми значениями. Искомые параметры

$$\mathbf{a_1 = m_\xi; \quad a_2 = \sigma^2_\xi;}$$

$$\boldsymbol{l(m_\xi, \sigma^2_\xi / x_n) = } -\frac{n}{2}\ln(2n) - \frac{n}{2}\ln(\sigma^2_\xi) - \frac{1}{2\sigma^2_\xi}\sum_{i=1}^{n}(x_i - m_\xi)^2 .$$

Дифференцирование по m_ξ дает:

$$\frac{\partial \ell}{\partial m_\xi} = \frac{1}{\sigma_\xi^2} \sum_{i=1}^{n} (x_i - m_\xi), \tag{98}$$

$$\frac{d\ell}{d\sigma_\xi^2} = -\frac{n}{2\sigma_\xi^2}\left[1 - \frac{1}{n\sigma_\xi^2}\sum_{i=1}^{n}(x_i - m_\xi)^2\right]$$

Приравнивая обе частные производные к нулю для поиска экстремума, получаем так называемые уравнения правдоподобия для искомых оценок

$$\sum_{i=1}^{n}(x_i - \hat{m}_\xi) = \mathbf{0},$$

$$\frac{n}{\sigma_\xi^2}\sum_{i=1}^{n}(x_i - \tilde{m}_\xi)^2 = \mathbf{0}$$

Решением этой системы служат следующие представления:

$$\hat{m}_\xi = \mathbf{1/n} \sum_{i=1}^{n} x_i \tag{99}$$

$$\hat{\sigma}_\xi^2 = \mathbf{1/n} \sum_{i=1}^{n}(x_i - \tfrac{1}{n}\sum_{i=1}^{n} x_k)^2 \tag{100}$$

Таким образом, оценка максимального правдоподобия математического ожидания совпадает с обычным среднем, и поэтому будет иметь такие же свойства.
Оценка максимального правдоподобия для дисперсии будет смещённой, так как согласно (100) имеет место:

$$\mathbf{E}\left[\hat{\sigma}_\xi\right] = \frac{n-1}{n}\boldsymbol{\sigma}^2_\xi. \tag{101}$$

Как и в случае использования аналогий в этом случае оценки максимального правдоподобия будут состоятельны, т.к. их дисперсии подчиняются закону (84). Оценка математического ожидания несмещённой и эффективной, а оценка дисперсии имеет смещение, которое стремится к нулю при увеличении объема выборки. В самом деле, с учетом (100) нетрудно получить:

$$\boldsymbol{\sigma}^2_\xi - \mathbf{E}\left[\hat{\sigma}_\xi\right] = \boldsymbol{\sigma}^2_\xi - \frac{n-1}{n}\boldsymbol{\sigma}^2_\xi = \frac{1}{n}\boldsymbol{\sigma}^2_\xi \to \mathbf{0} \quad \textbf{при} \quad \mathbf{n >> 1} \tag{102}$$

Такие оценки принято называть асимптотически несмещенными.

6. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Это название имеет исторические корни и поэтому сохранилось в наше время как отдельный отдел статистического анализа, хотя при любом исследовании – обязательно выдвигаются гипотезы, затем проверяются на непротиворечивость наблюдаемым данным. К области проверки гипотез принято относить более определённые их разновидности. Прежде всего, сюда относится гипотеза о равенств (одинаковости) некоторых параметров в различные моменты времени, либо об одинаковости условий в которых протекают наблюдаемые процессы.

Таким образом, обычно предполагается известным некоторое состояние с которым сравниваются все остальные.

6.1. Общая схема проверки гипотез

1) Выдвигается гипотеза о некотором свойстве генеральной совокупности, эта гипотеза называется **основной** или **нулевой**, обозначается $\mathbf{H_0}$. Формулируется также **альтернативные гипотезы**, которые в какой-либо мере являются отрицанием основной гипотезы. Если альтернативная гипотеза единственная, то она обозначается $\mathbf{H_1}$.
2) Выбирается **критерий**, который позволяет сравнить выдвинутые гипотезы на предмет их противоречивости выборочным данным. Речь идёт о сравнении степени их противоречивости, то есть, например, в качестве вывода может быть заключение о том, что гипотеза H_0 больше согласуется с выборочными данными (менее противоречива), чем гипотеза H_1. В качестве критерия обычно выступают некоторые статистики, то есть функции от выборочных значений: $Z = \mathbf{g}(\vec{x}), \vec{x} = (x_1, \ldots, x_n)'$

Требования к статистикам:

а) должна быть известна функция плотности вероятности такой статистики при нулевой гипотезе, то есть: $Wg\ (Z/H_0)$;
б) статистика должна быть несмещённой, состоятельной и эффективной.
3) Определяется область (интервал) $D = [a_\alpha, b_\alpha]$, такой, что при выполнении основной гипотезы вероятность попадания статистики в заданный интервал не меньше заданной величины, т.е.:

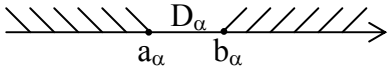
$$\mathbf{P\{(g(\vec{x}) \in D_\alpha)/H_0\} = 1 - \alpha}, \qquad (103)$$

где $0 \le \alpha \le 1$; $0{,}001 \le \alpha \le 0{,}1$

$$\int_{a_\alpha}^{b_\alpha} \omega_g(z/H_0)dz \ge 1-\alpha$$

$$\mathbf{P\{(g(\vec{x}) \notin D_\alpha)/H_0\} \le \alpha} \qquad (104)$$

Область $D\alpha$ **-** называется ***областью принятия основной гипотезы.*** Вся остальная часть числовой оси называется ***критической областью***, при попадании в которую значение критерия g отвергается H_0, и следовательно принимается альтернативная гипотеза.



критическая область

1) Осуществляется регистрация выборочных значений и вычисляется значение критерия $(g(\vec{x}))$.
2) Принимаются решения:

- если $a_\alpha \le g(\vec{x}) \le b_\alpha$, то H_0 –не противоречит данным;
- если $g \notin [a_\alpha, b_\alpha]$, то H_0 противоречит данным и поэтому отвергается на уровне значимости α.

Примечание: в проверке гипотез используется следующие термины:

α - уровень значимости;

a_α, b_α - *квантили* – границы интервала принятия решений.
Дополнительные требования к решающим функциям:
а) наибольшая мощность критерия **1 - β**, где β - вероятность ошибочного принятия H_0, когда на самом деле справедливо H_1.

6.2. Возможные ошибки при проверке гипотез

Ошибки I рода.
Принятие гипотезы H_0, когда на самом деле справедлива H_1. вероятность этой ошибки равна β.
Ошибки II рода.
Принятие H_1, когда имеет место H_0. вероятность ошибки этого рода равна α.
Невозможно сделать α - слишком маленьким, так как при этом будет расширяться область принятия гипотезы H_0, следовательно увеличится β, то есть вероятность отвергнуть гипотезу H_1, когда H_0 – неверна.
Всегда уменьшая вероятность ошибки одного вида, придем к увеличению вероятности ошибки другого вида. Поэтому обычно поступают следующим образом: фиксируется α, и вычисляется мощность критерия $1 - \beta$, то есть:

$$1 - \beta_\alpha = P\{g \notin D_\alpha / H_1\} \qquad \textbf{(105)}$$

Если величина $1 - \beta_\alpha$ - устраивает, то ограничиваемся значениями α, иначе изменяем α так, чтобы получить приемлемый уровень $1 - \beta_\alpha$.
Рассмотрим две наиболее часто встречающихся в статистических исследованиях процедуры проверки гипотез о значениях дисперсий и математических ожиданий.

6.3. Сравнение дисперсий в двух выборках (критерий Фишера)

Формулировка задачи: регистрируется n – значений генеральной совокупности (x_1, …, x_n) и m – значений возможно другой генеральной совокупности (y_1, …, y_m).
Предполагается, что исходная генеральная совокупность имеет гауссово распределение и отсчеты независимы.
Задача: установить одинаковы или нет дисперсии генеральных совокупностей выборками из которых являются векторы

$$\vec{x} = (x_1,\ldots,x_n)' \text{ и } \vec{y} = (y_1,\ldots,y_m)'.$$

Исходная гипотеза H_0: дисперсии одинаковы, т.е.

$$\sigma^2_1 = \sigma^2_2,$$

Альтернатива: (106)

$$\mathbf{H_1: \sigma^2_1 \neq \sigma^2_2.}$$

Решающая статистика представляет собой отношение:

$$\mathbf{g(x)} = \frac{S^2_1}{S^2_2}; \qquad (107)$$

$$\mathbf{S^2_1 = \max\{\tilde{\sigma}^2_1 : \tilde{\sigma}^2_2\}; \; S^2_2 = \min\{\tilde{\sigma}^2_1 : \tilde{\sigma}^2_1\};} \qquad (108)$$

$$\tilde{\sigma}^2_1 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}); \quad \tilde{\sigma}^2_2 = \frac{1}{m-1}\sum_{i=1}^{n}(y_i - \bar{y});$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \;, \; \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

Критерий вида (107) обычно обозначается

$$F = \frac{S^2_1}{S^2_2}; \qquad F \geq \mathbf{1.}$$

Квантили распределения этого критерия:

h_α: $P\{g \le h_\alpha\} = 1-\alpha$, являются функциями длин выборок *m* и *n* ; так что говорят о числе степеней свободы N_1, N_2:

N_1- количество отсчетов в выборке с максимальной оценкой дисперсии (S_1^2);

N_2 –длина выборки с меньшей оценкой дисперсии (S_2^2)

Значение $h_\alpha(N_1, N_2)$ – затабулированы.

Тогда, если $F > h_\alpha$, то принимается решение в пользу H_1 (т.е. гипотеза о равенстве дисперсий отвергается). Если $F \le h_\alpha$, то не отвергается.

6.4. Сравнение математических ожиданий в двух выборках

Постановка задачи. Регистрируются выборки:

$$\mathbf{X:\ x_1, \ldots, x_n;}$$

$$\mathbf{Y:\ y_1, \ldots, y_m,}$$

причем предполагается, что генеральные совокупности являются гауссовыми с независимыми значениями.

Задача заключается в проверке гипотезы о равенстве математических ожиданий генеральных совокупностей на основе этих выборок. Иначе говоря, искомая (нулевая) гипотеза заключается в предположении равенства

$$\mathbf{H_0:\ m_{\xi 1} = m_{\xi 2} = m_\xi;}$$

$$\mathbf{m_{\xi 1} = E[\xi_1],\ m_{\xi 2} = E[\xi_2].}$$

А гипотеза

$$\mathbf{H_1:\ m_{\xi 1} \neq m_{\xi 2}.} \qquad (109)$$

Прежде чем переходить к проверке справедливости этой гипотезы, необходимо убедиться в справедливости гипотезы о равенстве дисперсии. Если эта гипотеза не отвергается, то возможна дальнейшая проверка выполнения гипотезы о равенстве математических ожиданий.

Для проверки исходного предположения о равенстве математического ожидания используется статистика вида:

$$\mathbf{t_\nu} = \frac{\left|\bar{x} - \bar{y}\right|}{\sqrt{(m-1)\tilde{\sigma}^2_{\xi 2} + (n-1)\tilde{\sigma}^2_{\xi 1}}} \cdot \sqrt{\frac{(m+n-2)}{m+n} \cdot m \cdot n} \qquad (110)$$

Эта статистика называется ***статистикой Стьюдента*** с $\nu = m + n - 2$ - степенями свободы. Для неё составлены таблицы квантилей и соответствующих им уровней значимости:

$$\mathbf{P\{t_\nu > h_\alpha\} = \alpha;}$$

$$\mathbf{\alpha \text{ - уровень значимости,}}$$

$$\mathbf{h_\alpha\text{-квантиль (порог)}}$$

Таким образом, если вычисленная согласно соотношению (110) статистика превышает значение квантиля, соответствующего выбранному уровню значимости, т.е. $t_\nu \ge h_\alpha$, то делается вывод о том, что исходная гипотеза о равенстве математических ожиданий двух генеральных совокупностей отвергается на уровне значимости - α. Например, исходная гипотеза отвергается на $\alpha = 0{,}05$ и тому подобное. Обычно α выбирается из диапазона $0{,}001 \le \alpha \le 0{,}1$.

Предупреждение: таблицы квантилей рассчитаны в предположении, что исследуемые гауссовы генеральные совокупности имеют одинаковые дисперсии:

$$\sigma^2_{\xi 1} = \sigma^2_{\xi 2},$$

то есть только при этих условиях обоснованно может быть использовано предположение о том, что статистика (110) имеет распределение Стьюдента с ν -степенями свободы.

6.5. Проверка гипотезы о виде законов распределения вероятности генеральной совокупности. Критерии согласия

Важнейшей характеристикой случайной величины является *функция плотности вероятности* или *функция распределения вероятности*, поэтому в приложениях часто возникает необходимость определения закона распределения вероятности в виде некоторой формулы для $W_\xi(x)$.
Нулевая гипотеза заключается в том, что делается предположение о конкретном типе функции плотности вероятности.
Вид закона распределения вероятности можно до определённой степени установить используя гистограмму. В силу закона Бернулли это отношение по вероятности сходится к истинному значению:

$$\lim_{n\to\infty} P\{\nu_1 / n - P_i < \varepsilon\} = 1,$$

поэтому проверка нулевой гипотезы о виде закона распределения может быть осуществлена на сопоставлении имперических и теоретических вероятностей, описываемых формулой:

$$P_i = P\{\xi \in D_i\} = \int_{a_i}^{b_i} W_\xi(x)dx \qquad \textbf{(111)}$$

$$D_i = [a_i; b_i]$$

В качестве меря расхождения между P_i и $\widetilde{P}_i$ используется статистика:

$$\chi^2_{L-1} = \sum_{i=1}^{L} P_i \left(\frac{P_i - \nu_i / n}{P_i} \right)^2 \qquad \textbf{(112)}$$

где L – количество интервалов в гистограмме.
Эту статистику называют ***критерием согласия Пирсона***, то есть критерием того, что исходное предположение о виде закона распределения согласуется с экспериментальными данными. Пирсон показал, что если генеральная совокупность является гауссовой, то при достаточно большом количестве испытаний (n>>1) χ^2_{L-1} – имеет распределение типа χ^2 с L-1 степенью свободы. Для распределения χ^2 также составлены таблица квантилей, которые соответствуют выбранным уровням значимости.
Схема проверки:

1. Выдвигается гипотеза и воде закона распределения.
2. Проводится n – измерений значения случайной величины.
3. Строится гистограмма.
4. Оцениваются необходимые параметры закона распределения.
5. Вычисляются значения P_i, согласно формуле (111), в которой подставляются значения границ интервалов гистограммы и оцененных параметров.
6. Проводится вычисление по формуле (112).

7. Выбирается уровень значимости α и определяется по таблицам h_α.

8. Вычисленные значения сравниваются с h_α и выносится решение о том, что значение выборки противоречит исходной гипотезе, если выполняется неравенство:

$\chi^2_{L-1} > h_\alpha$,
иначе говорят, что исходная гипотеза о виде закона распределения противоречит выборочным данным на уровне значимости α.

7. ОЦЕНКА СТАТИСТИЧЕСКОЙ ВЗАИМОСВЯЗИ ГЕНЕРАЛЬНЫХ СОВОКУПНОСТЕЙ

Достаточно часто, особенно в экономике, приходится рассматривать несколько генеральных совокупностей, которые могут оказывать взаимное влияние. Возникает проблема описания этого влияния с помощью математических моделей. В противном случае речь идет о проверке гипотезы об отсутствии взаимосвязи между двумя генеральными совокупностями на основе статистических данных. Если нулевая гипотеза отвергается, то принимается гипотеза о наличии влияния одной генеральной совокупности на другую (о их взаимосвязи).
Итак, гипотеза имеет вид: $\{$ H_0 : ξ_2 не зависит от ξ_1 и/или наоборот ξ_1 не зависит от ξ_2, но

$$\xi_1: x_1, \ldots, x_n$$

$$\xi_2: y_1, \ldots, y_n.$$

Для проверки ее справедливости предполагается использовать выборочные значения. Так как по предположению генеральные совокупности – являются случайными величинами, то можно говорить о вероятностных моделях, описывающих влияние генеральной совокупности. То есть следует **взаимозависимость** понимать в вероятностном смысле.
Определение: события A и B называются **независимыми**, если вероятность появления одного из них не зависит от того, появится или нет другое событие.

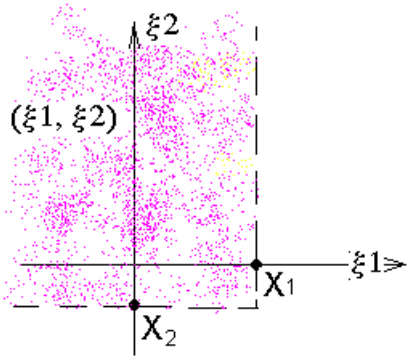
$$P(A/B) = P(A)$$

$$P(B/A) = P(B)$$

$P(A \cap B) = P(A) \cdot P(B/A) = P(B) \cdot P(A/B) \Rightarrow P(A)\,P(B) \rightarrow$ независимые
Моделирование взаимосвязей случайной величины требует рассмотрения многомерных распределений вероятностей и введение соответствующих понятий, в том числе: *ковариационных и корреляционных моментов, и условных математических ожиданий.*

7.1. Двумерные функции плотности распределения и их модели

По аналогии с одномерным случаем можно ввести функцию плотности вероятности совокупности двух случайных величин (генеральных совокупностей) (ξ_1, ξ_2):



$$\mathbf{F_2\,(x_1,\,x_2) = P\{\xi_1 < X_1 \cap \xi_2 < X_2\}}$$

Для независимых случайных величин:

$$\mathbf{P\{\xi_1 < X_1 \cap \xi_2 < X_2\} = P\{\xi_1 < X_1\}P\{\xi_2 < X_2\}} \qquad (113)$$

$$\mathbf{F_2(x_1,\,x_2) = F_{\xi 1}(x_1)F_{\xi 2}(x_2)} \qquad (114)$$

Двумерной функцией плотности вероятности – называется

$$W_2(x_1,x_2) = \frac{\partial^2 F_2(x_1,x_2)}{\partial x_1 \partial x_2}. \qquad \mathbf{(115)}$$

Для независимых случайных величин

$$W_2\,(x_1,\,x_2) = W_{\xi 1}(x_1)\,W_{\xi 2}(x_2). \qquad \mathbf{(116)}$$

Кроме понятия независимости, которое сводится к произведению функции плотности вероятности, имеется несколько (очень мало) конкретных моделей двумерной функции плотности вероятности, которые позволяют описать в том числе зависимость между случайными величинами. Основная модель – гауссова функция плотности вероятности:

$$W_2\,(x_1,\,x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{-\frac{1}{2(1-\rho_{12}^2)}\left[\frac{(x_1-m_1)^2}{\sigma_1^2} - 2\rho_{12}\frac{(x_1-m_1)(x_2-m_2)}{\sigma_1\sigma_2} + \frac{(x_2-m_2)^2}{\sigma_2^2}\right]\right\} \qquad \mathbf{(117)}$$

где $m_i = E[\xi_i]$, $i = 1, 2$;

$\sigma^2 = E\,[\,(\xi_i - m_i)^2\,]$.

$$\rho_{12} = \frac{E[(\xi_1 - m_1)(\xi_1 - m_2)]}{\sigma_1\sigma_2} = \frac{R_{12}}{\sigma_1\sigma_2} \qquad \mathbf{(118)}$$

В случае $\rho_{12} = 0 \Rightarrow W_2\,(x_1,\,x_2) = W_{\xi 1}(x_1)\,W_{\xi 2}(x_2)$; $\qquad W_{\xi 1}(x_1) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left[-\frac{(x_1-m_1)^2}{2\sigma_1^2}\right]$.

Таким образом, в данной модели признаком зависимости или независимости является *невыполнение или выполнение равенства:*

$$\rho_{12} = 0 \qquad \mathbf{(119)}$$

Параметр ρ_{12}- принято называть ***коэффициентом корреляции***, либо ***ковариацией.***
Всегда выполняется неравенство:

$$-1 \le \rho_{12} \le 1 \qquad \mathbf{(120)}$$

Можно сказать, что здесь для гауссовых случайных величин равенства ± 1 выполняются тогда, когда случайные величины связаны ***линейной зависимостью***, т.е. имеют место равенства:

$$\xi_2 = a + в\xi_1 \qquad \text{или} \qquad \xi_2 = c + d\xi_2 \qquad \mathbf{(121)}$$

Таким образом, наблюдается строгая функциональная зависимость.
Замечание: понятие коэффициента корреляции относится не только к гауссовым случайным величинам. *Корреляция* (от лат.) – correlatio – отношение, взаимосвязь.

То, что $|\rho_{12}| \le 1$, позволяет использовать его как характеристику тесноты связи.

Как для вероятности событий, так и для функций плотности вероятности можно ввести понятие ***условной функции плотности вероятности***:

$$W(x_1,x_2) = W_{\xi 1}(x_1)\, W_{\xi 2}(x_2 / \xi_1 = x_1) = W_{\xi 2}(x_2) = W_{\xi 1}(x_1/\xi_1 = x_2) \qquad \textbf{(122)}$$

$$P(A \cap B) = P(A) * P(B/A) = P(B) * P(A/B)$$

Исследование взаимосвязи случайной величины с прикладной точки зрения целесообразно потому, что может позволить предсказывать значение одних случайных величин по значению других. Особенно это полезно тогда, когда некоторые величины могут быть ненаблюдаемыми.

Возникает вопрос: на какой основе предсказания будут в каком-либо смысле наилучшими? В силу различных причин, в качестве меры ошибок прогноза используются среднеквадратические отклонения.

Известно, что математическое ожидание случайной величины обладает свойством: $E[(\xi - m_\xi)^2] \le E[(\xi - c)^2]$; где $c \neq m_\xi$.

7.2. Условное математическое ожидание

Можно ввести понятие условных математических ожиданий следующим образом

$$m_{\xi 1/\xi 2} = \int_{-\infty}^{\infty} x_1 W_{\xi_1}(x_1 / \xi_2 = x_2)dx_1 ,$$

$$m_{\xi 2/\xi 1} = \int_{-\infty}^{\infty} x_2 W_{\xi_2}(x_2 / \xi_1 = x_1)dx_2 . \qquad \textbf{(123)}$$

$$m_\xi = \int_{-\infty}^{\infty} x W_\xi(x)dx$$

Тогда можно показать, что условная дисперсия:

$$E[(\xi_1 - m_{1/2})^2 |\xi_2] \le E[(\xi_1 - c)^2 |\xi_2], \qquad \textbf{(124)}$$

где $m_{1/2} \neq c$.

Условное математическое ожидание обладает оптимальными свойствами, которые заключаются в том, что относительно него будет минимальным среднеквадратический разброс. В случае независимости, когда условная функция плотности вероятности совпадает с безусловной, имеют место равенства:

$$\sigma^2_{1|2} = \sigma^2_{\xi 1} ; \quad m_{1|2} = m_{\xi 1}$$

$$\sigma^2_{2|1} = \sigma^2_{\xi 2} ; \quad m_{2|1} = m_{\xi 2} \qquad \textbf{(125)}$$

В любом другом случае среднеквадратическое отклонение от условного математического ожидания меньше среднеквадратического отклонения от безусловного математического ожидания.

Если имеет место функциональная зависимость $\xi_1 = f(\xi_2)$, то $\sigma^2_{1|2} = 0$.

Условное математическое ожидание в гауссовом случае (117) из (122):

$$W_{\xi 1}(x_1 | \xi_2 = x_2) = \frac{W_2(x_1,x_2)}{W_{\xi 2}(x_2)}.$$

$$W_{\xi 2}(x_2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(x_2 - m_2)^2}{2\sigma^2_2}],$$

тогда получим:

$$W_{\xi 1}(x_1|\ \xi_2 = x_2) = \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{1-\rho^2_{12}}} \exp\{ \frac{[(x_1-m_1) - \rho_{12}\frac{\sigma_1}{\sigma_2}(x_2-m_2)]^2}{2\sigma^2_2(1-\rho^2_{12})} \} \qquad \textbf{(126)}$$

$$m_{1|2} = E[\xi_1|\ \xi_2 = x_2] = m_1 + \rho_{1\,2}\ \frac{\sigma_1}{\sigma_2}(x_2 - m_2) \qquad \textbf{(127)}$$

$$m_{2|1} = E[\xi_2|\ \xi_1 = x_1] = m_2 + \rho_{1\,2}\ \frac{\sigma_2}{\sigma_1}(x_1 - m_1) \qquad \textbf{(128)}$$

$$\sigma^2_{1|2} = \sigma^2_1(1 - \rho^2_{12}) \qquad \textbf{(129)}$$

$$\sigma^2_{2|1} = \sigma^2_2(1 - \rho^2_{12}) \qquad \textbf{(130)}$$

Условные математические ожидания часто называют ***регрессией***.
Соотношения (127):(130) – показывают, что коэффициент корреляции является *важнейшей характеристикой взаимосвязи* двух гауссовых случайных величин.
Соотношения (127):(128) определяют линейные функциональные зависимости условных математических ожиданий от безусловных и значений второй случайной величины, которые предполагаются известными.

8. КОРРЕЛЯЦИОННЫЕ ЗАВИСИМОСТИ (УРАВНЕНИЯ СВЯЗИ)

8.1. Графический анализ взаимосвязи

Корреляционные зависимости определяются ***коэффициентом корреляции (ρ)***, если $\rho_{1\,2} = 0$, то говорят об отсутствии связи, $|\rho_{1\,2}| = 1$ – соответствует максимально тесной корреляционной связи.
Предполагается, что в результате наблюдений получено n – пар значений двух генеральных совокупностей в одни и те же моменты времени.

$$\xi_1: x_1, x_2, \ldots, x_n$$

$$\xi_2: y_1, y_2, \ldots, y_n$$

$$(x_i, y_i),\ I = 1, 2, \ldots, n$$

Очевидно, что точки с координатами (x_i, y_i) можно нанести на рисунок, характерные примеры которых приведены ниже.

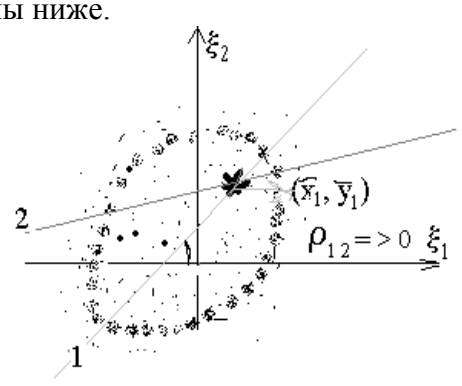


Рис.1

Рис.1 по форме напоминает эллипс, вытянутый вдоль некоторой прямой , угол которой с осью «x» является острым, это является признаком того, что коэффициент корреляции больше нуля.

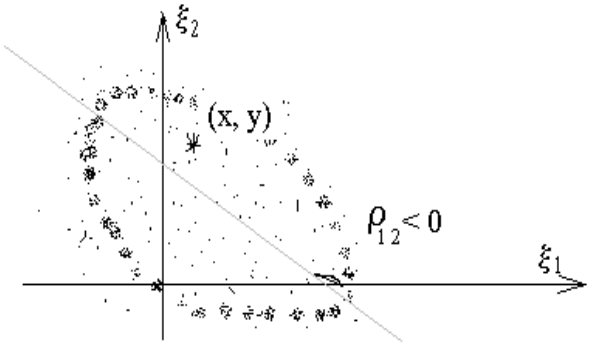


Рис.2

Рис.2 по форме напоминает эллипс, вытянутый вдоль кривой, которая с осью «x» составляет тупой угол, то есть коэффициент корреляции меньше нуля.

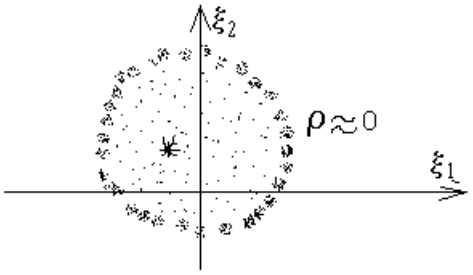


Рис.3

Рис.3 по форме напоминает ≈ окружность, коэффициент корреляции ≈ 0.

Совокупность точек на плоскости называется ***корреляционным полем***, использование которого для предварительного установления наличия корреляционной связи относится к разведочному анализу данных.

8.2. Модели корреляционных зависимостей

Корреляционные соотношения имеют в своей основе представление:

$$\xi_2 = E[\xi_2 \mid \xi_1] + \varepsilon_2 = f_2(\xi_1) + \varepsilon_2 \qquad \textbf{(131)}$$

$$\xi_1 = E[\xi_1 \mid \xi_2] + \varepsilon_1 = f_1(\xi_2) + \varepsilon_1 \qquad \textbf{(132)}$$

Суть в том, что желательно иметь необходимую информацию, с помощью которой можно было бы вычислить значения генеральной совокупности. Эта функция ($\hat{\xi}_2 = \mathbf{f_1(\xi_1)}$) должна обладать оптимальными свойствами в том смысле, что минимизирует погрешность предсказания. Если в качестве меры погрешности используется среднеквадратическое отклонение, то в смысле минимума этого критерия наилучшим является *условное математическое ожидание*.

На рис.1 проведена линия, которая является графическим представлением ***линейной модели условного математического ожидания*** (корреляционная зависимость) линия 1:

$$\hat{f}_2(\xi_1) = a + b\xi_1 \qquad \textbf{(133)}$$

линия 2 – линейная аппроксимация линии 1:

$$\hat{f}_1(\xi_2) = c + d\xi_2 \quad \textbf{(134)}$$

Эти линии пересекаются в точке с координатами ($\bar{x}$, $\bar{y}$)

Выигрыш в точности предсказания значений одной случайной величины по значениям другой на основе линейных моделей (133) и (134) можно оценить с помощью отношения:

$$\eta_1 = \frac{\tilde{\sigma}_1}{\tilde{\sigma}_{1/2}} \quad (135)$$

$$\eta_2 = \frac{\tilde{\sigma}_2}{\tilde{\sigma}_{2/1}} \quad (136)$$

$$\tilde{\sigma}^2_{1/2} = \frac{1}{n-1} \sum_{i=1}^{n} [x_i - \bar{x} - c - d(y_1 - \bar{y})]^2 \quad (137)$$

$$\tilde{\sigma}^2_{2/1} = \frac{1}{n-1} \sum_{i=1}^{n} [y_i - \bar{y} - a - b(x_1 - \bar{x})]^2 \quad (138)$$

8.3. Оценки параметров моделей корреляционных зависимостей

Следующий этап заключается в определении параметров линейных зависимостей a,b и c,d. Следуя желанию максимизировать отношение (135) и (136), необходимо минимизировать в правой части отношения (137) и (138) путём подбора соответствующих значений a,b и c,d.

То есть искомые параметры должны удовлетворять вариационным принципам:

$$\frac{1}{n-1} \sum_{i=1}^{n} [x_i - \bar{x} - c - d(y_1 - \bar{y})]^2 = \min \quad \textbf{(139)}$$

$$\frac{1}{n-1} \sum_{i=1}^{n} [y_i - \bar{y} - a - b(x_1 - \bar{x})]^2 = \min \quad \textbf{(140)}$$

Где минимум ищется по значениям соответствующих параметров. Найденные a и b, c и d называются ***<u>оценками наименьших квадратов</u>***.

Выведем формулы наименьших квадратов оценок в случае соотношения (140).

Необходимым и достаточным условием минимума в этом соотношении (140) является выполнение равенств:

$$\frac{\partial \tilde{\sigma}^2_{2/1}}{\partial a} = 0; \quad \textbf{(141)}$$

$$\frac{\partial \tilde{\sigma}^2_{2/1}}{\partial b} = 0. \quad \textbf{(142)}$$

При подстановке в них левой части соотношения (140) получаем ***систему нормальных уравнений:***

$$- \frac{2}{n-1} \sum_{i=1}^{n} [y_i - \bar{y} - a - b(x_i - \bar{x})] = \mathbf{0},$$

$$- \frac{2}{n-1} \sum_{i=1}^{n} [y_i - \bar{y} - a - b(x_i - \bar{x})] \ (\mathbf{x_i} - \bar{x}) = \mathbf{0}.$$

Так как $\sum(y_i - \bar{y}) = 0$; $\sum(x_i - \bar{x}) = 0$, то эти соотношения дают

$$\mathbf{a = 0,} \tag{143}$$

$$\mathbf{b} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})} = \sqrt{\frac{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \cdot \mathbf{r_{12}} = \frac{\tilde{\sigma}_2}{\sigma_2} \mathbf{r_{12}}, \tag{144}$$

где

$$\mathbf{r_{12}} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{145}$$

Для параметров d и c из (139) аналогично можно получить:

$$\mathbf{c = 0;} \qquad \mathbf{d} = \frac{\tilde{\sigma}_1}{\sigma_2} \mathbf{r_{12}}$$

Аппроксимация (131) и (132):

$$\xi_1 = \mathbf{E}[\xi_2/\xi_1] + \mathbf{E_2} = \bar{x} + \frac{\tilde{\sigma}_1}{\sigma_2} \mathbf{r_{12}} (\xi_2 - \bar{y}) + \varepsilon_1$$

$$\xi_2 = \bar{y} + \frac{\tilde{\sigma}_2}{\sigma_1} \mathbf{r_{12}} (\xi_1 - \bar{x}) + \varepsilon_1$$

Соотношение (145) определяет ***выборочный коэффициент корреляции***, величина которого удовлетворяет неравенству $-1 \le \mathbf{r_{12}} \le 1$, а конкретное значение может быть использовано для проверки гипотезы об отсутствии корреляционной связи:

$$\mathbf{H_0 : \rho_{12} = 0.} \tag{146}$$

В этом случае решающая функция имеет вид:

$$\mathbf{t} = \frac{|r_{12}|\sqrt{n-2}}{\sqrt{1-r^2_{12}}} \tag{147}$$

Эта статистика при выполнении H_0 имеет распределения Стьюдента с «n –2» -степенями свободы. Поэтому необходимо выбрать уровень значимости, и соответствующий ему

порог из таблиц распределения Стьюдента h_{α}; затем вычислить правую часть соотношения (147) и сравнить с порогом. В случае выполнения неравенства

$$\frac{|r_{12}|\sqrt{n-2}}{\sqrt{1-r^2_{12}}} > \mathbf{h}_{\alpha}$$

делается вывод, что гипотеза (146) отвергается на уровне значимости «α», то есть генеральные совокупности признаются зависимыми.
Квадрат коэффициента корреляции используется в качестве показателя тесноты корреляционной связи. Связь считается **<u>тесной</u>**, если $r^2_{12} \geq 0{,}86$.

9. СТАТИСТИЧЕСКИЙ АНАЛИЗ ВРЕМЕННЫХ РЯДОВ

9.1. Понятие временного ряда

Последовательность (набор) данных $x_1, \ldots, x_n$ – называется ***<u>временным рядом</u>***, в тех случаях, когда важным является порядок следования каждого из зарегистрированных значений; так как обычно данные получаются в результате их регистрации в определённые моменты времени.
Основная гипотеза при рассмотрении случайной величины (см. выше) заключается в предположении о независимости отдельных выборочных значений. В этих условиях данные можно перемешивать, меняя их местами, то есть выявить зависимость от времени в этих условиях нельзя. Строго говоря, любые данные представляют собой ***<u>временной ряд</u>***, но в некоторых случаях зависимостью от времени можно пренебречь.
Когда возникает необходимость предсказания будущих значений в регистрируемой последовательности, адекватным является понятие <u>временного ряда</u> (модель).
Предсказание будущих значений ряда может обоснованно осуществить только, если будет выявлено некоторая тенденция их поведения. Именно в силу наличия тенденции, оказывается важным порядок следствия значений временного ряда.
Таким образом <u>важнейшей задачей</u> анализа временного ряда является: определение схемы генерации их значений, которые описывают искомые тенденции.

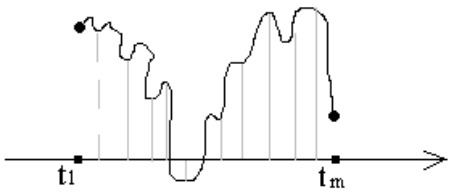
9.2. Основные понятия и модели анализа временных рядов

К числу основных понятий относятся:

1. Тренд, фильтрация и сглаживание.
2. Автоковариация и спектральная плотность.
3. Модели генерации значений.

9.3. Трендовые модели генерации значений временного ряда.

<u>Тренд</u> – неслучайная функция времени, которая описывает основную тенденцию поведения временного ряда.



Обычное математическое представление трендовых моделей имеет вид:

$$\mathbf{x_t = f(t) + \varepsilon\,(t)} \qquad (148)$$

где f(t) – тренд;

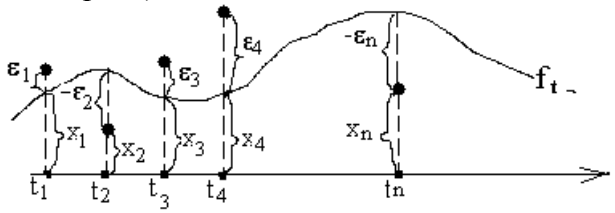
ε (t) – «невязка», характеризующая неточность совпадения значений временного ряда со значениями тренда.

Предполагается выполнение условий несмещённости и некоррелированности:

$$\mathbf{E(\varepsilon_t) = 0}; \qquad (149)$$

$$\mathbf{E(\varepsilon_{t1}\varepsilon_{t2}) =} \begin{cases} \sigma^2{}_\xi, t_1 = t_2 \\ 0, t_1 \neq t_2 \end{cases} \quad . \qquad (150)$$

Более строго предполагаем, что ε_t – имеет независимые значения. При этом, как последовательность она является **стационарной** в том смысле, что соотношение (149) и (150) выполняются при любых значениях t_1 и t_2, независимо от того, где их взяли (неизменной является и дисперсия).



Задача заключается в определении $f(t)$ – как некой функции о времени, так как в этом случае появляется возможность осуществить предсказание значений временного ряда за пределами его интервалов регистрации.

$$\hat{x}_{n+k} = f(t_n + t_k)$$

Возникает проблема определения вида функции f, которая в некотором смысле наилучшим образом позволяет экстраполировать данные за пределы их регистрации. Иногда (чаще всего) в качестве меры отклонения временного анализа от гипотетического тренда используется оценка среднеквадратического отклонения:

$$\widetilde{\sigma}_\varepsilon(f) = \sqrt{\sum_{i=1}^{n} \varepsilon^2{}_i} = \sqrt{\sum_{i=1}^{n} [x_i - f(t_i)]^2} \qquad \mathbf{(151)}$$

Таким образом, основным принципом подбора функций тренда является ***принцип наименьших квадратов:***

$$\widetilde{\sigma}_\varepsilon = \sum_{i=1}^{n} [x_i - f(t_i)]^2 = \min \qquad \mathbf{(152)}$$

Для реализации этого принципа необходимо задавать конкретный вид функциональной зависимости $f(t)$:

для разных временных рядов вид этой функциональной зависимости – разный, поэтому прежде чем переходить к минимизации согласно (152) необходимо провести разведочный анализ данных, которые могут дать ответ о действительном наличии некоторого тренда, и о приблизительном виде соответствующей функциональной зависимости

9.4. Фильтрация и сглаживание временного ряда

9.4.1. Медианная фильтрация (сглаживание)

Фильтрация – это удаление чего-то, что является ненужным или разделение на компоненты.

Сглаживание - это процедура исключения или уменьшения размаха колебаний значений временного ряда.

Медианная фильтрация получила развитие в последние 20 лет.

Говорят, что Ymed – является ***медианой*** подпоследовательности из 2m+1 значений (y_{k-1}, …, y_k, y_{k+1}), если выполняется :

$$Ymed = Z(k), \quad \textbf{(153)}$$

где

$$Z(i) \in (y_{k-m}, \ldots y_{k+m}),\ i=1 \ldots 2m+1, \quad \textbf{(154)}$$

$$Z(1) \le Z(2) \le \ldots \le Z(2m+1). \quad \textbf{(155)}$$

То есть последовательность Z(i) представляет собой упорядоченную по возрастанию последовательность значения “y”

Пример: пусть m=1,2m+1 = 3

0.1, -3, 4

Z_1	Z_2	Z_3
-3	0.1	4

Тогда Ymed = 0.1

В зависимости от значения 2_{m+1}, m=1 – медиана по тройкам, m=2 – медиана по пятеркам и т.д.

В рассмотрение вовлекаются: отрезок исходного временного ряда, содержащий 2m+1 подряд идущих значений.

Процедура фильтрации осуществляется для всех значений временного ряда, кроме первого и последнего.

Достоинства медианой фильтрации заключаются в следующем.

Эта фильтрация позволяет исключить выбросы, то есть значения временного ряда, которые «сильно» отличаются от остальных. При этом эти выбросы заменяются соседними значениями той же последовательности. Количество вовлекаемых в рассмотрение значений исходного временного ряда называется ***апертурой фильтра***. При апертуре 2m+1 можно исключить m – подряд идущих выбросов.

Медианная фильтрация не искажает монотонно возрастающих или монотонно убывающих отрезков временного ряда.

9.4.2. Метод скользящего среднего

В основном методы алгоритмического сглаживания – это различные модификации скользящего среднего.

В основе этих методов лежит следующее соображение: если индивидуальный разброс значений временного ряда x(t) вокруг тренда f(t) характеризуется дисперсией σ^2, то разброс среднего из N членов временного ряда (x(1) + x(2) +…+ x(N))/N около того же значения будет характеризоваться дисперсией σ^2/N.

Алгоритм скользящего среднего заключается в следующем:

$$\hat{f}(t) = \sum_{k=-m}^{m} w_k x(t+k),\ t = m+1, m+2, \ldots, n-m,$$

где w_k – некоторые весовые коэффициенты, в сумме равные 1, т.е.

$$\sum_{k=-m}^{m} w_k = 1.$$

Поскольку, изменяя t от m+1 до n-m диапазон суммирования скользит по временному ряду (при переходе от t к t+1 в составе слагаемых происходит замена только одного слагаемого x(t – m) слагаемым x(t + m + 1)), то эти методы называются методами скользящего среднего (МСС).

Один МСС отличается от другого выбором параметров m и весов w_k.

Определение весов w_k основано на следующей процедуре.

Для 2m+1 элементов временного ряда x(1), x(2),…, x(2m+1) строится полином степени p методом наименьших квадратов.

Значение этого полинома используют для расчёта значений оценки тренда $\hat{f}(t)$ в средней точке этого отрезка ряда m + 1, т.е.

$\hat{f}(m+1) = \hat{x}_1(m+1)$.

Эта же процедура выполняется для отрезка временного ряда x(2),…, x(2m+2).

Оказалось, что такая процедура приводит к процедуре взвешенного скользящего суммирования, по которой веса определяются из следующей таблицы (для m = 1, 2, 3).

m	веса
1	1/3, 1/3, 1/3 (средняя арифметическая)
2	-3/35, 12/35, 17/35, 12/35, –3/35
3	–2/21, 3/21, 6/21, 7/21, 6/21, 3/21, –2/21

9.4.3. Метод экспоненциально взвешенного скользящего среднего (метод Брауна)

Методы скользящего среднего основываются на том, что все значения временного ряда имеют одинаковую информационную ценность. Однако в задачах прогноза, в которых сглаженная функция $\hat{f}(t)$ используется обычно для формирования прогнозов на несколько тактов вперёд, недавние значения x(t) очевидно ценнее, чем значения ряда в далёком прошлом, так как ряд далее будет вести себя так, какова сформировавшаяся тенденция в настоящем и недалёком прошлом.

Эта идея реализована в методе экспоненциально взвешенного скользящего среднего Брауна

$$\hat{f}(t) = \frac{1-\lambda}{1-\lambda^t} \sum_{k=0}^{t-1} \lambda^k x(t-k),\ 0<\lambda<1.$$

$$\hat{f}(t) = \frac{1-\lambda}{1-\lambda^t} \left(x(t) + \lambda x(t-1) + \lambda^2 x(t-2) + \ldots + \lambda^{t-1} x(1)\right)$$

Таким образом, значения временного ряда тоже сглаживаются, однако существуют следующие отличия от обычного МСС

1) скользит только правый край интервала усреднения(левый край закреплён в точке t = 1)

2) веса при x(t – k) экспоненциально уменьшаются по мере «удаления в прошлое»,

3) формула (2) даёт оценку сглаженного значения временного ряда не в средней, а в правой, конечной точке усреднения,

4) нет проблемы крайних значений.

Дисперсия остаточной случайной компоненты после сглаживания

$$D\tilde{\varepsilon}(t) = \sigma^2 \frac{1-\lambda}{1-\lambda^t} \cdot \frac{1+\lambda^t}{1+\lambda},$$

где

$\tilde{\varepsilon}(t)$ – остаточная нерегулярная компонента после сглаживания.

$\sigma^2 = Dx(t)$;

поэтому при значениях λ, не слишком близких к 1, и для достаточно удалённых от прошлого значений t случайные остатки $\tilde{\varepsilon}(t)$ подвержены существенно меньшему разбросу.

Рекуррентный способ вычисления сглаженных значений $\hat{f}(t)$.

$$\hat{f}(t) = \frac{1-\lambda}{1-\lambda^t}(x(t) + \lambda x(t-1) + \lambda^2 x(t-2) + \ldots + \lambda^{t-1} x(1));$$

$$\hat{f}(t+1) = \frac{1-\lambda}{1-\lambda^{t+1}}(x(t+1) + \lambda x(t) + \lambda^2 x(t-1) + \ldots + \lambda^t x(1)) =$$

$$= \frac{1-\lambda}{1-\lambda^{t+1}} x(t+1) + \lambda \frac{1-\lambda^t}{1-\lambda^{t+1}} \frac{1-\lambda}{1-\lambda^t}(x(t) + \lambda x(t-1) + \lambda^2 x(t-2) + \ldots + \lambda^{t-1} x(1)) =$$

$$= \frac{1-\lambda}{1-\lambda^{t+1}} x(t+1) + \lambda \frac{1-\lambda^t}{1-\lambda^{t+1}} \hat{f}(t);$$

$$\hat{f}(t+1) = \frac{1-\lambda}{1-\lambda^{t+1}} x(t+1) + \lambda \frac{1-\lambda^t}{1-\lambda^{t+1}} \hat{f}(t);\ t = 2, \ldots$$

$$\hat{f}(1) = x(1).$$

9.5. Анализ временного ряда с помощью модели авторегрессии

9.5.1. Понятие процесса авторегрессии

Модели авторегрессии – используются для описания схемы генерации значений временных рядов.

Процесс авторегрессии порядка (P) – называется случайная последовательность y_t, которая описывается следующим отношением:

$$y_t = m + \sum_{k=1}^{p} \beta_k (y_{t-k} - m) + \sigma^2_u U_t \qquad \textbf{(156)}$$

здесь y_t процесс авторегрессии порядка (AP(P));

$$E[U_t] = 0:$$

$$E[U_{t1}, U_{t2}] = \begin{cases} 0, t_1 \neq t_2, \\ 1, t_1 = t_2; \end{cases} \qquad \textbf{(157)}$$

где U_t – входной шум (обновляющий процесс);

β_k, к = 1, …, p – вещественные параметры.

Если взять математическое ожидание от левой и правой части, то с учетом свойства (157) получим:

$$E[y_t] = m. \qquad \textbf{(158)}$$

Параметры процесса авторегрессии (AP) не могут быть произвольными и должны удовлетворять условию стационарности (устойчивости), а именно:

$$|Z_i| < 1,\ i = 1, \ldots, p, \qquad \textbf{(159)}$$

где Z_i – корни характеристического уравнения.

$$Z^p - \sum_{k=1}^{p} \beta_k Z^{p-k} = 0. \quad \textbf{(160)}$$

Наиболее часто используются модели авторегрессии первого (АР(1)) и второго (АР(2)) порядков:

$$\text{АР(1)}: y_t = m+\beta(y_{t-1} - m) + \sigma^2{}_u U_t, \quad \textbf{(161)}$$

$$\text{АР(2)}: y_t = m+\beta_1(y_{t-1} - m) +\beta_2(y_{t-2} - m)+ \sigma^2{}_u U_t. \quad \textbf{(162)}$$

*Для процесса АР(1) условия стационарности (158) принимает вид:

$$|\beta| < 1, \quad \textbf{(163)}$$

так как (159) означает $Z - \beta = 0$.
*Для процесса АР(2) характеристическое уравнение имеет вид:

$$Z^2 - \beta_1 Z - \beta_2 = 0 , \quad \textbf{(165)}$$

представление для корней которого следующее

$$Z_{1,2} = \frac{\beta_1}{2} \pm \sqrt{\frac{\beta^2_1}{4} + \beta_2} . \quad \textbf{(166)}$$

Модели авторегрессии часто используются для генерации искусственных временных рядов с целью тестирования методов обработки или некоторых устройств. В этих случаях для выполнения условия стационарности (158) лучше сначала задаться значениями корней, а затем вычислить параметры авторегрессии с использованием формул Виета.

$$\beta_p = (-1)^p Z_1, Z_2, \ldots, Z_p. \quad \textbf{(167)}$$

Для процесса АР(2) могут быть комплексные корни:

$$\beta_1 = -(z_1 + z_2);$$

$$\beta_2 = z_1 z_2.$$

Для вещественности параметров должны выполняться условия:

$$Z_2 = Z_1^*,$$

т.е.

$$d = a, \; c = -b,$$

$$a^2 + b^2 = |Z_k|^2.$$

Следовательно, согласно (159)

$$a^2 + b^2 < 1,$$

$$Z_1 = a + ib , \; Z_2 = d + ic,$$

так как

$$\beta_1 = -(Z_1+Z_2),\ \beta_2 = Z_1Z_2.$$

9.5.2. Свойства автоковариационных функций (АКФ). Уравнение Юла-Уокера

Автоковариационная функция процесса AP(p)

$$R\tau = E[(y_t - m, y_{t+\tau} - m)],\ \tau = 0,1,2,\ldots$$

удовлетворяет так называемому *уравнению Юла-Уокера:*

$$R\tau = \begin{cases} \sum_{k=1}^{p} \beta_k R_{\tau-k}, \tau \geq 1 \\ \sigma^2{}_u + \sum_{k=1}^{h} \beta_k R_{\tau-k}, \tau = 0 \end{cases} \quad \textbf{(168)}$$

Из соотношения (168) следует, что автоковариационную функцию процесса авторегрессии, при любом $\tau \geq 1$ можно вычислить, если известны p – предыдущих значений. В частности первое из равенств (168) - является разностным уравнением, решение которого имеет вид:

$$R\tau = \sum_{k=1}^{p} C_i Z^{\tau}_i \quad \textbf{(169)}$$

если Z_i – простые корни характеристического уравнения.
Для доказательства этого необходимо подставить (169) в (168) имея в виду:

$$R_{\tau-k} = \sum_{k=1}^{p} C_i Z^{\tau-k}_i.$$

В результате в виду (160) получаем:

$$\sum_{k=1}^{p} Ci\ (\ Z^{\tau}_i - \sum_{k=1}^{p} \beta_k\ Z^{\tau}_i) = 0.$$

Примеры:

$$AP(1):\ R\tau = \frac{\sigma^2}{1-\beta^2}\ \beta^{\tau}, \quad \textbf{(170)}$$

$$R_0 = \sigma^2{}_y = \frac{\sigma^2{}_u}{1-\beta^2}, \quad \textbf{(171)}$$

$$AP(2):\ R\tau = \frac{\sigma^2{}_u}{(z_1-z_2)(1-z_1z_2)}\left[\frac{z_1^{\tau+1}}{1-z_1^2} - \frac{z_2^{\tau+1}}{1-z_2^2}\right], \quad \textbf{(172)}$$

Если Z_1 и Z_2 – вещественные, то

$$R\tau = C_1Z_1^{\tau} + C_2Z_2^{\tau}.$$

В зависимости от знака вещественных корней, элемент Z_i^{τ} будет иметь постоянный знак и монотонно убывать, либо иметь колебательный характер, меняя знаки.

Если

$$Z_1 = a + jb,$$

$$Z_2 = a - jb,$$

то эта функция будет иметь вид затухающей косинусоиды:

$$R\tau = A\,\ell^{-\gamma\tau}\cos\eta\tau \qquad \textbf{(173)}$$

где $|Z_1| = |Z_2| = \ell^{-\gamma}$,
где $A = R_0$ – из соотношения (172);
η= arctg b/a.

9.5.3. Условное математическое ожидание

Процессы авторегрессии, как модели реальных временных рядов, используются при решении различных задач анализа их свойств. Существенный интерес представляет возможность решения задачи прогноза, то есть предсказания будущих значений временного ряда на основе зафиксированной предыстории.
Предсказание значений временного ряда на основе предыстории означает вычисление некоторой функции, аргументами которой является предыдущие значения:

$$\hat{y}_t = f[y_{t-1}, y_{t-2}, \ldots, y_{t-N}, \ldots] \qquad \textbf{(174)}$$

Эта функция должна удовлетворять некоторым требованиям оптимальности. В данном случае, речь идёт о минимизации ошибки предсказания, в качестве меры которой используется дисперсия ошибки:

$$\sigma_{пр}^2 = E[(y_t - \hat{y}_t)^2]. \qquad \textbf{(175)}$$

Функция f – должна удовлетворять условию:

$$\sigma_{пр}^2 = \min. \qquad \textbf{(176)}$$

Известно, что этому условию удовлетворяет ***условное математическое ожидание***, то есть должно иметь место:

$$\hat{y}_t = E[y_t / y_{t-1}, y_{t-2}, \ldots, y_{t-N}, \ldots] \qquad \textbf{(177)}$$

Для процесса авторегрессии порядка p имеет место:

$$E[y_t / y_{t-1}, y_{t-2}, \ldots, y_{N-M}, \ldots] = m + \sum_{k=1}^{p}\beta_k(y_{t-k} - m) \qquad \textbf{(178)}$$

То есть наилучший прогноз имеет вид:

$$\hat{y}_t = m + \sum_{k=1}^{p}\beta_k(y_{t-k} - m) \qquad \textbf{(179)}$$

Причём:

$$\sigma_{пр}^2 = \sigma_u^2 \qquad \textbf{(180)}$$

Таким образом, для решения задачи прогноза временных рядов на базе модели авторегрессии необходимо осуществить идентификацию этого ряда в этом классе моделей.
Идентификация (отождествление) – набор модели из определенного класса, который в каком-либо смысле наилучшим образом описывает поведение ряда. Эта процедура

заключается в определении по реальным данным порядка модели авторегрессии, параметров и математического ожидания.
То есть, речь идет об оценке параметров представления (156).

9.5.4. Оценивание параметров модели авторегрессии

Оценивание параметров модели авторегрессии – определение по выборке x_t , t =1, …, N из временного ряда значений оценок математического ожидания ($\tilde{m}$), порядка модели($\tilde{p}$), и параметров $\hat{\beta}_k$, к =1, …, $\tilde{p}$, а так же $\tilde{\sigma}_u^2$.

Существует несколько подходов к оцениванию параметров авторегрессии. В основе их лежат какие-либо принципы, на базе которых можно получить вычислительные формулы. Наиболее свободным от априорных предположений о свойствах временного ряда, является ***принцип наименьших квадратов,*** тогда как метод максимального правдоподобия, например, предполагает вполне определенную функцию плотности вероятности, причём многомерную: $W_N (X_1, …, X_N)$.

Постановка задачи: наблюдается временной ряд, относительно которого выдвигается гипотеза о том, что его значение подчиняется разностному уравнению вида (156), то есть:

$$X_t = m + \sum_{k=1}^{p} \beta_k (X_{t-k} - m) + \sigma_u^2 U_t, \qquad \textbf{(182)}$$

значения которого неизвестны.
Процесс определения этих параметров называется идентификация временных рядов в классе моделей авторегрессии.
Определение параметров вида (182) осуществляется чтобы в дальнейшем осуществить прогноз в виде:

$$\hat{x}t = \tilde{m} + \sum_{k=1}^{\tilde{p}} \tilde{\beta}_k (X_{t-k} - \tilde{m}) \qquad \textbf{(183)}$$

по $\tilde{p}$ - предыдущим значениям, либо организовать решение другой прикладной задачи, которая упрощается, если использовать представление вида (182).
Следует подчеркнуть, что реальный временной ряд может не являться процессом авторегрессии.
Математическое ожидание оценивается обычным образом

$$\tilde{m} = \frac{1}{N} \sum_{k=1}^{N} X_k,$$

можно показать, что эта оценка удовлетворяет *принципу наименьших квадратов*:

$$\sum_{k=1}^{N} [X_k - \tilde{m}]^2 = \min \sum_{k=1}^{N} [X_k - c]^2,$$

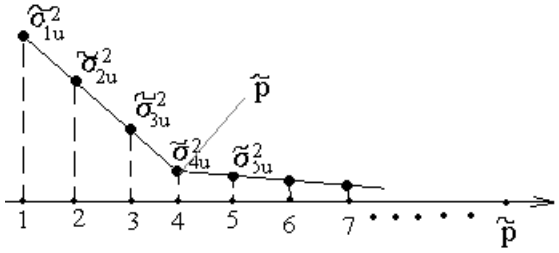
где с – любое вещественное число.
Наиболее трудоёмка процедура определения остальных параметров.
Обозначим $\overset{0}{X}_t = X_t - \tilde{m}$ - центрированная последовательность. Предположим, что параметр β_k, к=1, …,p, где p – известны, тогда оценкой σ_u^2 является

$$\tilde{\sigma}_u^2 = \frac{1}{N} \sum_{k=1}^{N} [\overset{0}{X}_t - \sum_{k=1}^{p} \beta_k \overset{0}{X}_{t-k}]^2$$

Таким образом, чтобы оценить параметр $\tilde{\sigma}_u^2$ необходимо знать параметры β_k и порядок к = 1, …,p.

Так как они на самом деле неизвестны, то процедура оценивания является ***итерационной.*** Последовательно полагают, что $\tilde{p} = 1$, определяют параметры $\tilde{\beta}_{11}$, $\tilde{\sigma}_{1u}{}^2$, $\tilde{p} = 2$, $\tilde{\beta}_{21}$, $\tilde{\beta}_{22}$, $\tilde{\sigma}_{2u}{}^2$, и так далее до $\tilde{p}'$, $\tilde{\beta}_{\tilde{p}1}$, $\tilde{\beta}_{\tilde{p}2}$, …, $\tilde{\beta}_{\tilde{p}\,\tilde{p}}$, $\tilde{\sigma}_{\tilde{p}u}{}^2$.

Правило останова базируется *на критерии Акаике*, который заключается в следующем. Строится график последовательных оценок σ^2_{ku}, л=1,2, …



В начале графика значение оценок $\tilde{\sigma}_u^2$ изменяется (понижается) достаточно быстро, но после некоторого значения ($\tilde{p}$) изменения будут менее резкие. Точка, где угол наклона заметно изменяется, принимается за наиболее подходящую оценку порядка подбираемой модели авторегрессии, а соответствующие оценки параметров принимаются за параметры модели.

Процедура оценивания параметров авторегрессии при фиксированном значении р заключается в минимизации ***квадратичного функционала***:

$$F_{\tilde{p}} = \sum_{t=\tilde{p}+1}^{N} [\overset{0}{X}_t - \sum_{k=1}^{\tilde{p}} \beta_{p\,k} \overset{0}{X}_{t-k}]^2 = \min_{\beta_{\tilde{p}}}. \qquad \textbf{(184)}$$

Необходимо продифференцировать и приравнять нулю:

$$\frac{\partial F_{\tilde{p}}}{\partial \beta_{\tilde{p}i}} = 0,\ I = 1, \ldots, \tilde{p}\ . \qquad \textbf{(185)}$$

Получаем ***систему нормальных уравнений***, которая имеет вид:

$$\sum_{t=\tilde{p}+1}^{N} \overset{0}{X}_t \overset{0}{X}_{t-i} = \sum_{t=\tilde{p}+1}^{N} \sum_{k=1}^{\tilde{p}} \tilde{\beta}_{\tilde{p}\,к} \overset{0}{X}_{t-к} \overset{0}{X}_{t-i}, \quad i = 1, \ldots, \tilde{p} \qquad \textbf{(186)}$$

Если ввести обозначения:

$\tilde{R}_{к\,i} = \dfrac{1}{N-\tilde{p}} \sum_{t=\tilde{p}+1}^{N} \overset{0}{X}_{t-к} \overset{0}{X}_{t-i}$; $\tilde{R} = \{\tilde{R}_{k\,i}\}$, i, k = 1, …, $\tilde{p}$, $\vec{\tilde{r}} = (\tilde{R}_{01}, \tilde{R}_{02}, \tilde{R}_{03}, \ldots, \tilde{R}_{0\,\tilde{p}})$, то

систему уравнений вида (186) можно переписать в матричном виде:

$$\tilde{R}\ \vec{\tilde{\beta}}_p = \vec{r}\ , \qquad \textbf{(187)}$$

где $\vec{\tilde{\beta}}_{\tilde{p}} = (\vec{\tilde{\beta}}_{\tilde{p}\,1}, \ldots, \vec{\tilde{\beta}}_{\tilde{p}\ \tilde{p}})$ – вектор искомых параметров.

Матрица $\tilde{R}$ представляет собой оценку ковариационной матрицы. Если она неособенная, то решение (187) можно записать в виде:

$$\vec{\tilde{\beta}}_{\tilde{p}} = \tilde{R}^{-1}\ \vec{\tilde{r}} \qquad \textbf{(188)}$$

Эта формула вместе с формулами, определяющими элементы матрицы $\tilde{R}$ и вектора $\vec{\tilde{r}}$, применяются столько раз, сколько предписывается критерием Акаике.

ЧАСТЬ II. ОСНОВЫ ЭКОНОМЕТРИКИ

10. ЭКОНОМИЧЕСКАЯ МОДЕЛЬ

Основным элементом экономического исследования является анализ и построение взаимосвязей экономических переменных. Математическое выражение таких взаимосвязей называется экономической моделью.

Пример.

$C = \beta_0 + \beta_1 I,$ **(189)**

I – располагаемый доход семьи;

C – потребление.

Построение экономических моделей осложнено следующими факторами

1. часто эти взаимосвязи не являются строгими, функциональными зависимостями;
2. очень трудно выявлять все факторы, влияющие на данный зависимый экономический показатель;
3. воздействие многих факторов является случайным;
4. экономисты обладают ограниченным набором данных статистических наблюдений, которые к тому же содержат различного рода ошибки.

Если удаётся преодолеть эти трудности, тогда можно построить экономическую модель, выражающую функциональную зависимость некоторой зависимой величины от формирующих её значение факторов. Особенность функциональной зависимости состоит в том, что по значению независимой величины (переменной) можно однозначно, абсолютно точно вычислить, предсказать значение зависимой величины.

11. ЭКОНОМЕТРИЧЕСКАЯ МОДЕЛЬ

Рассмотрим набор реальных статистических данных (C_k, I_k) и изобразим эти данные точками в координатах (C,I).

Таким образом, зависимость между величинами C_k, I_k не функциональная, а стохастическая, случайная. Но эта случайность не такова, что абсолютно невозможно предсказать или объяснить по величине I_k величину C_k, поскольку видна достаточно устойчивая тенденция роста (в среднем). Другими словами, взаимосвязь между величинами C_k, I_k состоит в следующем: точное значение C_k не вычисляется по значению I_k, однако с ростом I_k значение C_k в среднем увеличивается. Такой характер зависимости выражается следующим образом:

$C_k = \beta_0 + \beta_1 I_k + \varepsilon_k,\ k = 1, 2,\ldots,N.$ **(190)**

В общем случае, характер зависимости нелинейный:

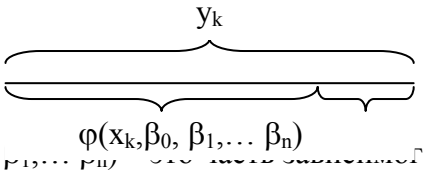
$y_k = \varphi(x_k, \beta_0, \beta_1,\ldots \beta_n) + \varepsilon_k;\ k = 1, 2,\ldots,N.$ **(191)**

Соотношения (190), (191) являются эконометрическими моделями. Таким образом, эконометрическая модель – это выражение *статистической* зависимости между экономическими величинами. Эконометрическая модель строится на основе экономической теории и статистических данных.

11.1. Элементы эконометрической модели и их свойства

1) Вид функции φ называется спецификацией модели. Модель (1) является частным видом эконометрической модели, в которой спецификация линейная. Спецификация φ описывает общий ход экономического процесса, экономическую тенденцию развития, изменения зависимого показателя при изменении независимого.

2) Величина x_k – называется независимой или объясняющей переменной.
3) Величина y_k называется зависимой или объясняемой переменной. Значение величины y_k состоит из двух частей:



величина $\varphi(x_k,\beta_0, \beta_1,\dots \beta_n)$ – это часть зависимого показателя, обусловленная или объяснённая экономическими причинами или просто *объясняемая* часть; ε_k – необъяснённая часть, поскольку невозможно описать все случайные факторы.
4) ε_k – величина, выражающая вклад случайных мелких, незначительных факторов, которые отклоняют реальные статистические данные от значений зависимого показателя, обусловленного экономической тенденцией, однако не изменяют эту экономическую тенденцию.
Основные свойства величин ε_k:
а) эти величины случайные, в противном случае зависимость (2) функциональная;
б) ε_k принимают положительные и отрицательные значения, так как случайные факторы увеличивают или уменьшают величины, обусловленные экономической тенденцией;
в) Абсолютные величины $|\varepsilon_k|$ не должны быть очень большими по сравнению со значениями, вычисляемыми по спецификации. Другими словами, необъяснённая часть показателя у должна быть мала по сравнению с объяснённой. Если же отклонения значительны, тогда спецификация φ неточно описывает экономическую тенденцию, её необходимо уточнять и в число объясняющих факторов вводить факторы, вклад которых приводит к большим значениям ε_k.
5) $\beta_0, \beta_1,\dots, \beta_k$ – параметры спецификации. Для разных видов функции φ количество этих параметров, их смысл и названия разные. Например, для линейной спецификации
$y_k = \beta_0 + \beta_1 x_k + \varepsilon_k$; k = 1, 2,…,N
этих параметров два;
параметр
β_0 – свободный член,
β_1 – угловой коэффициент.

11.2. Классификация переменных эконометрической модели

В зависимости от характера использования переменные эконометрической модели могут быть экзогенными, эндогенными и предопределёнными.
Экзогенные переменные задаются извне, являются планируемыми или управляемыми. Однако не всегда эндогенные переменные находятся в правой части модели.
Эндогенные переменные формируются внутри социально-экономической системы под воздействием экзогенных переменных и во взаимодействии друг с другом. В эконометрической модели они являются объясняемыми переменными.

Предопределённые переменные – переменные-факторы, объясняющие переменные. Предопределённые переменные формируются из всех экзогенных переменных, которые могут относиться к любому моменту времени и *лаговых эндогенных* переменных – эндогенных переменных, значения которых в модели относятся в предыдущие моменты времени, и, следовательно, уже являются известными, заданными.

11.3. Задачи эконометрики

1) По имеющимся статистическим данным подбор спецификации, наиболее точно отражающей экономическую тенденцию. Эта задача может быть уже решена экономической или эконометрической теорией.

2) Оценивание параметров спецификации $\beta_0, \beta_1,\ldots, \beta_k$ и определение качества этих оценок и эконометрической модели в целом. (Насколько хорошо модель объясняет изменение показателя y).

3) Формирование прогнозов на основе построенной модели и выработка рекомендаций для эффективных экономических решений.

12. МОДЕЛИ И МЕТОДЫ РЕГРЕССИОННОГО АНАЛИЗА

12.1. Основные понятия регрессионного анализа

В естественных науках часто речь идет о ***функциональной зависимости*** (связи), когда каждому значению одной переменной соответствует вполне *определенное значение другой* (например, скорость свободного падения в вакууме в зависимости от времени и т.д.).
В экономике в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определенное, а множество возможных значений другой переменной.
Возникновение такой зависимости обусловливается тем, что зависимая переменная подвержена влиянию ряда неконтролируемых или неучтенных факторов, а также тем, что измерение значений переменных неизбежно сопровождается некоторыми случайными ошибками. Примером статистической связи является зависимость урожайности от количества внесенных удобрений, производительности труда на предприятии от его энерговооруженности и т.п.
В силу неоднозначности зависимости между Y и X для исследователя, в частности, представляет интерес усредненная по X схема зависимости, т. е. закономерность в изменении условного математического ожидания $M_X(Y)$ или $M(Y/X = x)$ в зависимости от x.
Если зависимость между двумя переменными такова, что каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой, то такая статистическая зависимость называется **корреляционной или регрессионной.**
Иначе, **регрессионной зависимостью** между двумя переменными называется функциональная зависимость между значениями одной из них и условным математическим ожиданием другой.
Регрессионная зависимость может быть представлена в виде

$$M_x(Y) = \varphi(x)$$

или

$$M_y(X) = \psi(y)$$

где $\varphi(x) \neq \text{const}$, $\psi(y) \neq \text{const}$.
В регрессионном анализе рассматриваются *односторонняя зависимость случайной переменной Y от одной (или нескольких) неслучайной независимой переменной X.* Такая зависимость может возникнуть, например, в случае, когда при каждом фиксированном

значении *X* соответствующие значения *Y* подвержены случайному разбросу за счет действия ряда неконтролируемых факторов.

При этом *зависимую* переменную Y называют также *функцией отклика, объясняемой, выходной, результирующей, эндогенной переменной, результативным признаком;* а независимую переменную *X – объясняющей, входной предсказывающей, предикторной, экзогенной переменной, фактором, регрессором, факторным признаком.*

Уравнение

$$M_x(Y) = \varphi(x, \beta_0, \beta_1, \ldots, \beta_n)$$

называется модельным уравнением регрессии (или просто уравнением регрессии), а функция $\varphi(x)$ – модельной функцией регрессии (или просто функцией регрессии), а ее график — модельной линией регрессии (или просто линией регрессии). $\beta_0, \beta_1, \ldots, \beta_n$ – параметры функциональной зависимости.

Для точного описания уравнения регрессии необходимо знать условный закон распределения зависимой переменной *Y* при условии, что переменная *X* примет значение *x.* В статистической практике такую информацию получить, как правило, не удается, так как обычно исследователь располагает лишь выборкой пар значений (x_i, y_i) ограниченного объема *n.* В этом случае речь может идти только об *оценке (приближенном выражении, аппроксимации)* по выборке функции регрессии. Такой оценкой является *выборочная функция (кривая) регрессии:*

$\hat{y} = \hat{\varphi}(x, b_0, b_1, \ldots, b_n)$

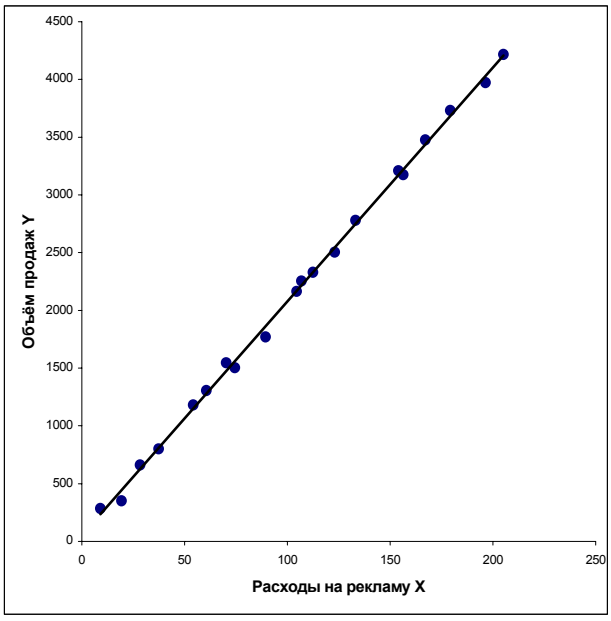
где $\hat{y}$ – *выборочное условное среднее* переменной Y при фиксированном значении переменной $X = x$, $b_0, b_1, \ldots, b_n$ – параметры функции $\hat{\varphi}$.

Это уравнение называется выборочным уравнением регрессии или моделью регрессионной зависимости Y от X.

Вид функции $\hat{\varphi}(x, b_0, b_1, \ldots, b_n)$ называется спецификацией модели выборочной регрессии.

Задачами регрессионного анализа являются следующие:

1) Оценка (выбор) спецификации модели – установление конкретного выражения функции $\hat{\varphi}(x, b_0, b_1, \ldots, b_n)$. Достаточно часто эта задача может быть решена точно в том случае, если заранее известен характер изменения величины Y при изменении X: линейный, экспоненциальный,… Вид зависимости может быть известен теоретически, как результат уже проводившихся исследований или определён визуально при анализе статистических данных (x_i, y_i). Например, для выборки, представленной на следующей диаграмме



линейный характер зависимости очевиден.

2) После выбора спецификации производится оценка параметров спецификации b_0, $b_1,\ldots,b_n$. Для разных спецификаций набор параметров различен. Например, при выборе линейной спецификации

$\hat{y} = b_0 + b_1 x$

следует вычислить два параметра: b_0, b_1. Даже в том случае, если спецификация модели определена точно, значения b_0, b_1 будут являться только оценками истинных параметров уравнения регрессии

$y = \beta_0 + \beta x$.

3) Производится оценка качества полученной регрессии.

Сформулируем основные предпосылки и принципы регрессионного анализа.

1. Объективно существует зависимость одного экономического показателя Y от другого X. Эта зависимость не функциональная, так как на основное течение процесса, экономическую тенденцию накладываются различные случайные факторы. Поэтому для данного значения независимого показателя X = x зависимый показатель может принять значение из некоторого множества с какой-то вероятностью. То есть для каждого значения X величина Y является случайной величиной, распределённой по некоторому закону.

2. Таким образом, каждому значению X соответствует условное математическое ожидание $M_x(Y)$. То есть функциональной является зависимость не самого значения Y от X, а его условного математического ожидания: $M_x(Y) = \varphi(x, \beta_0, \beta_1,\ldots, \beta_n)$. Эта зависимость называется модельной регрессией. В общем случае ни вид функции, ни точные значения параметров $\beta_0, \beta_1,\ldots, \beta_n$ неизвестны, поскольку недоступны генеральные совокупности значений переменной Y при заданных X.

3. Реальным выражением зависимости Y от X является статистическая выборка (x_i,y_i). По этой выборке методами регрессионного анализа получают приближённую функциональную зависимость $\hat{y} = \hat{\varphi}(x,b_0, b_1,\ldots,b_n)$ выборочного условного среднего Y от x.

4. Функциональным зависимостям

$M_x(Y) = \varphi(x, \beta_0, \beta_1,\ldots, \beta_n)$,

$\hat{y} = \hat{\varphi}(x,b_0, b_1,\ldots,b_n)$

соответствуют модели наблюдений – зависимости между реальными статистическими данными (x_i,y_i):

$y_i = \varphi(x_i, \beta_0, \beta_1,\ldots, \beta_n) + \varepsilon_i$,

$y_i = \hat{\varphi}(x_i,b_0, b_1,\ldots,b_n) + e_i$,

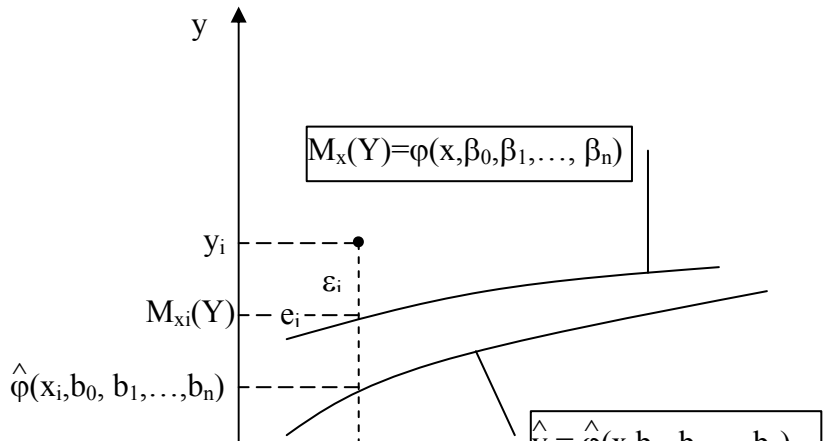
где

$\varepsilon_i = y_i - \varphi(x_i, \beta_0, \beta_1,\ldots, \beta_n) = y_i - M_{xi}(Y)$ – отклонение наблюдаемого значения y_i от своего условного математического ожидания;

ε_i – *ошибка, возмущение* – результат воздействия неучтённых факторов;

$e_i = y_i - \hat{\varphi}(x_i,b_0, b_1,\ldots,b_n)$ – отклонение наблюдаемого значения y_i от вычисленного по теоретической функции регрессии; фактически невязки e_i являются выборочными значениями величин ε_i.

e_i – *невязка*.



5. Методы регрессионного анализа используются для подбора по возможности более точной спецификации $\hat{\varphi}(x,b_0, b_1,\ldots,b_n)$ и оценок параметров $b_0, b_1,\ldots,b_n$ с тем, чтобы выборочная линия регрессии $\hat{\varphi}(x,b_0, b_1,\ldots,b_n)$ приближалась к модельной $\varphi(x,\beta_0,\beta_1,\ldots, \beta_n)$.

12.2. Линейная парная регрессия

12.2.1. Определения

Парной линейной регрессией называется зависимость

$\hat{y} = b_0 + b_1x$

выборочного условного математического ожидания от переменной x. Термин «парная» означает зависимость *двух* переменных Y, X. Термин «линейная» означает их линейную зависимость.

Условное выборочное математическое ожидание – это выборочное среднее значение величины Y при условии, что переменная X приняла значение x.

Модель наблюдения

$y_i = b_0 + b_1x + e_i,$

b_0 – оценка свободного члена β_0,

b_1 – оценка углового коэффициента β_1.

12.2.2. Принцип, метод наименьших квадратов

Согласно принципу наименьших квадратов неизвестные параметры b_0, b_1 выбираются таким образом, чтобы была минимальна сумма квадратов невязок или остатков

$S(b_0,b_1) = \Sigma(\hat{y} - y_i)^2 = \Sigma(b_0 + b_1x_i - y_i)^2 \to \min.$

Следует отметить, что для оценки параметров b_0, b_1 возможны и другие подходы. Например, можно находить эти параметры при минимизации суммы абсолютных величин невязок $\Sigma|\hat{y} - y_i|^2$. Однако вычислительные процедуры, соответствующие принципу наименьших квадратов существенно проще. Эти вычислительные процедуры получили название «Метод наименьших квадратов» (МНК).

Исходя из необходимого условия экстремума функции двух переменных $S(b_0,b_1)$ приравниваем к нулю её частные производные

$$\frac{\partial S}{\partial b_0} = 2\Sigma(b_0 + b_1x_i - y_i) = 0;$$

$$\frac{\partial S}{\partial b_1} = 2\Sigma(b_0 + b_1x_i - y_i)x_i = 0;$$

откуда после преобразований получим систему нормальных уравнений для определения параметров линейной регрессии:

$b_0n + b_1\Sigma x_i = \Sigma y_i;$

$b_0\Sigma x_i + b_1\Sigma x_i^2 = \Sigma x_iy_i.$

Теперь разделим обе части уравнений на n, получим систему нормальных уравнений в виде

$b_0 + b_1\overline{x} = \overline{y}$;

$b_0\overline{x} + b_1\overline{x^2} = \overline{xy}$,

где

$\overline{x} = n^{-1}\Sigma x_i$; $\overline{y} = n^{-1}\Sigma y_i$; $\overline{x^2} = n^{-1}\Sigma x^2_i$; $\overline{xy} = n^{-1}\Sigma x_iy_i.$

Подставляя значение

$b_0 = \bar{y} - b_1\bar{x}$
из первого уравнения системы в уравнение регрессии, получим
$\hat{y} = \bar{y} - b_1\bar{x} + b_1x;$
$\hat{y} - \bar{y} = b_1(\bar{x} - x).$
Коэффициент b_1 называется выборочным угловым коэффициентом регрессии У по Х.
Коэффициент b_1 показывает, на сколько единиц в среднем изменяется выборочное условное среднее $\hat{y}$ при увеличении переменной Х на одну единицу.
Из нормальной системы получаем

$$b_1 = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{\hat{cov}(X,Y)}{s_x^2},$$

где s_x^2 – выборочная дисперсия переменной Х;
$\hat{cov}(X,Y)$ – выборочная ковариация величин Х,Y;
величину b_0 можно найти через коэффициент b_1:

$b_0 = \bar{y} - b_1\bar{x}$.
Через коэффициент b_1 также можно выразить выборочный коэффициент корреляции

$$r = b_1\frac{s_x}{s_y}$$

12.2.3. Свойства оценок параметров парной линейной регрессии

Чтобы полученные МНК оценки *a* и *b* обладали желательными свойствами, необходимо, чтобы выполнялись следующие условия:
1) величина ε_i, является случайной переменной;
2) математическое ожидание ε_i равно нулю: $M(e) = 0$;
3) дисперсия ε_i постоянна: $D(\varepsilon_i) = D(\varepsilon_i) = a^2$ для всех i;
4) значения ε_i независимы между собой.
По теореме Гаусса-Маркова
если условия 1)-4) выполняются, то оценки, сделанные с помощью МНК, обладают следующими свойствами:
1) Оценки являются несмещенными, т.е. математическое ожидание оценки каждого параметра равно его истинному значению: $M(b_0) = \beta_0$; $M(b_1) = \beta_1$. Это вытекает из того, что $M(\varepsilon_i)=0$, и говорит об отсутствии систематической ошибки в определении положения линии регрессии.
2) Оценки состоятельны, так как дисперсия оценок параметров при возрастании числа наблюдений стремится к нулю: $\lim D(b_0) = 0$; $\lim D(b_1) = 0$. Иначе говоря, если *n* достаточно велико, то практически наверняка b_0 близко к β_0, а b_1 близко к β_1: надежность оценки при увеличении выборки растет.
3) Оценки эффективны, они имеют наименьшую дисперсию по сравнению с любыми другими оценками данного параметра, линейными относительно величин y_i.
Перечисленные свойства не зависят от конкретного вида распределения величин ε_i, тем не менее обычно предполагается, что они распределены нормально $N(0;y^2)$. Эта предпосылка необходима для проверки статистической значимости сделанных оценок и определения для них доверительных интервалов. При ее выполнении оценки МНК имеют наименьшую дисперсию не только среди линейных, но среди всех несмещенных оценок.
Если предположения 3) и 4) нарушены, то есть дисперсия возмущений непостоянна и/или значения ε_i связаны друг с другом, то свойства несмещенности и состоятельности сохраняются, но свойство эффективности - нет.

12.2.4. Анализ статистической значимости коэффициентов линейной регрессии

Величины y_i, соответствующие данным x_i, при некоторых теоретических значениях β_0 и β_1, являются случайными. Следовательно, случайными являются и рассчитанные по ним значения коэффициентов b_0 и b_1. Их математические ожидания при выполнении предпосылок об отклонениях ε_i равны, соответственно, β_0 и β_1. При этом оценки тем надежнее, чем меньше их разброс вокруг β_0 и β_1, то есть дисперсия. По определению дисперсии $D(b_1) = M(b_1-\beta_1)^2$; $D(b_0) = M(b_0 - \beta_0)^2$. Надежность получаемых оценок a и b зависит, очевидно, от дисперсии случайных отклонений ε_i, но поскольку по данным выборки эти отклонения (и, соответственно, их дисперсия) оценены быть не могут, они заменяются при анализе надежности оценок коэффициентов регрессии на отклонения переменной y от оцененной линии регрессии $e_i = y - b_0 - b_1 x_i$.

Можно доказать, что

$$D(b_1) = S_{b1}^2 = \frac{S^2}{\Sigma(x_i - \bar{x})^2};$$

$$D(b_0) = S_{b0}^2 = \frac{S^2 \Sigma x_i^2}{n\Sigma(x_i - \bar{x})^2};$$

где $S^2 = \frac{\Sigma e_i^2}{n-2}$ – мера разброса зависимой переменной вокруг линии регрессии *(необъясненная дисперсия)*. S_{b0} и S_{b1} - стандартные отклонения случайных величин b_0 и b_1. Полученный результат можно проинтерпретировать следующим образом.
Коэффициент b_1 есть мера наклона линии регрессии. Очевидно, чем больше разброс значений y вокруг линии регрессии, тем больше (в среднем) ошибка в определении наклона линии регрессии. Если такого разброса нет совсем ($e_i = 0$ и, следовательно, $\sigma^2=0$), то прямая определяется однозначно и ошибки в расчете коэффициентов b_0 и b_1 отсутствуют (а отсюда и значение S^1, "замещающее" σ^2, равно нулю).
В знаменателе величины $D(b_1)$ стоит сумма квадратов отклонений x, от среднего значения $\bar{x}$. Эта сумма велика в том случае, если регрессия оценена на достаточно широком диапазоне значений переменной x, и в этом случае, при данном уровне разброса S^2, очевидно, ошибка в оценке величины наклона прямой будет меньше, чем при малом диапазоне изменения переменной x. Если x_1, и x_2, лежат рядом, то даже небольшое изменение одного из y_i существенно меняет наклон прямой (если x_1 и x_2 далеки друг от друга - ситуация обратная).

Кроме того, чем больше (при прочих равных) число наблюдений n, тем больше $\Sigma(x_i - \bar{x})^2$ и, тем самым, меньше стандартная ошибка оценки. Дисперсия свободного члена уравнения регрессии равна

$D(b_0) = D(b_1)\frac{\Sigma x_i^2}{n}$ - она пропорциональна $D(b)$ и, тем самым, соответствует уже сделанным пояснениям о влиянии разброса y_i вокруг регрессионной прямой и разброса x_i на стандартную ошибку. Чем сильнее меняется наклон прямой, тем больше разброс значений свободного члена. Кроме того, дисперсия и стандартная ошибка свободного члена тем больше, чем больше средняя величина x^2_i. При больших по модулю значениях x даже небольшое изменение наклона регрессионной прямой может вызвать большое изменение оценки свободного члена, поскольку в этом случае в среднем велико расстояние от точек наблюдений до оси y.

12.2.5. Статистика Дарбина-Уотсона

Пусть имеются статистические данные (x_i, y_i).
x_i - независимая(объясняющая) переменная;

y_i - зависимая(объясняемая) переменная, соответствующая x_i.
После применения обычного МНК получено уравнение линейной регресии

$$y = b_0 + b_1 x.$$

Остатки вычисляются следующим образом:

$$e_i = y_i - (ax_i + b).$$

Следует выяснить, являются ли остатки e_i независимыми. Для этого вычисляется статистика Дарбина-Уотсона

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

Если линейная регрессия неадекватна имеющимся статистическим данным, тогда большие серии экспериментальных точек лежат выше либо ниже линии регрессии. В таком случае соседние невязки e_i, e_{i-1} имеют, как правило, одинаковые знаки и примерно равные абсолютные величины. Поэтому, большинство слагаемых $(e_i - e_{i-1})^2$ в числителе величины DW близки к 0. Поэтому статистика Дарбина-Уотсона близка к 0. Таким образом, если $DW \approx 0$, тогда следует сделать вывод о нелинейной зависимости показателя y от показателя x;
Если линейная модель регрессии подходит для описания статистических данных, тогда линия регрессии проходит между экспериментальных точек. В этом случае примерно половина соседних невязок e_{i-1} имеет такой же знак, а половина противоположный невязке e_i. В первом случае $(e_i - e_{i-1})^2 \approx 0$, во втором $(e_i - e_{i-1})^2 \approx 4e_i^2$. Следовательно, если линейная регрессия адекватна, тогда

$$DW = \frac{\sum_{i=2} (e_i - e_{i-1})^2}{\sum_{i=1} e_i^2} \approx \frac{0{,}5 \sum_{i=2} 4e_i^2}{\sum_{i=1} e_i^2} \approx 2.$$

Таким образом, если $DW \approx 2$, тогда зависимость y от x линейна.

12.3. Нелинейная регрессия

Зависимость многих экономических показателей не линейна. Метод наименьших квадратов не предназначен для оценки параметров нелинейных регрессий, кроме полиномиальной. Однако целый класс нелинейных зависимостей можно привести к зависимостям линейным преобразованием независимой и/или зависимой переменной. Такое преобразование называется линеаризацией.
Задача построения нелинейной модели регрессии состоит в следующем
Задана нелинейная спецификация модели
$y = f(x,a,b,\varepsilon)$,
где y - зависимая, объясняемая переменная; x - независимая, объясняющая переменная; a, b - параметры модели, для которых должны быть получены оценки; ε - аддитивный или мультипликативный случайный фактор.
Требуется

1. Преобразовать исходные данные $x \to x^*$, $y \to y^*$ так, чтобы спецификация модифицированной регрессии была линейной:

$y^* = a^* + b^* x^*$

2. Методом наименьших квадратов получить оценки параметров a^*, b^*.

3. По оценкам a^*, b^* вычислить искомые оценки параметров a, b исходной регрессии.

Способы преобразования данных и вычисления параметров a, b по оценкам a^*, b^* приведены в следующей таблице.

Исходная спецификация	Преобразование	Преобразование	Вычисление b по b^*	Вычисление a по a^*

	$x \to x^*$	$y \to y^*$		
$y=a+\frac{b}{x}+\varepsilon$	$x^*=\frac{1}{x}$	$y^*=y$	$b=b^*$	$a=a^*$
$y=\frac{1}{a+bx+\varepsilon}$	$x^*=x$	$y^*=\frac{1}{y}$	$b=b^*$	$a=a^*$
$y=\frac{x}{a+bx+x\varepsilon}$	$x^*=\frac{1}{x}$	$y^*=\frac{1}{y}$	$b=a^*$	$a=b^*$
$y=ae^{bx+\varepsilon}$	$x^*=x$	$y^*=\ln y$	$b=b^*$	$a=e^{a^*}$
$y=ae^{\frac{b}{x}+\varepsilon}$	$x^*=\frac{1}{x}$	$y^*=\ln y$	$b=b^*$	$a=e^{a^*}$
$y=\frac{1}{a+be^{-x}+\varepsilon}$	$x^*=e^{-x}$	$y^*=\frac{1}{y}$	$b=b^*$	$a=a^*$
$y=ax^be^{\varepsilon}$	$x^*=\ln x$	$y^*=\ln y$	$b=b^*$	$a=e^{a^*}$

12.4. Характеристики парной регрессии

Тесноту связи изучаемых явлений оценивает линейный коэффициент корреляции r_{xy} для линейной регрессии

$$r = b_1\frac{s_x}{s_y} = \frac{\hat{cov}\,(X,Y)}{s_x s_y} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{s_x s_y}$$

и индекс корреляции ρ_{xy} – для нелинейной регрессии.

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

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

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

Показатели дисперсии

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

где

$\sum(\bar{y} - y)^2$ – общая дисперсия результативного признака;

$\sum(\bar{y} - \hat{y}_x)^2$ – сумма квадратов отклонений от линии регрессии, обусловленная регрессией, «объяснённая» или «факторная» дисперсия;

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

Доля дисперсии, объясняемая моделью регрессии, в общей дисперсии результативного признака y характеризует коэффициент детерминации R^2:

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

Коэффициент детерминации – это квадрат коэффициента или индекса корреляции.

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

$$F_{факт} = \frac{\sum(y - \hat{y}_x)^2/m}{\sum(y - \hat{y}_x)^2/(n - m - 1)} = \frac{r^2_{xy}}{1 - r^2_{xy}}(n - 2),$$

где

n – объём выборочной совокупности,

m – число параметров при переменной x.

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

Если $F_{табл} < F_{факт}$, то H_0 – гипотеза о случайной природе оцениваемых характеристик токлоняется и признаётся их статистическая значимость и надёжность. Если $F_{табл} > F_{факт}$, то H_0 не отклоняется и признаётся статистическая незначимость, ненадёжность модели регрессии.

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, то есть незначимом их отличии от нуля. Оценка значимости коэффициентов регрессии и корреляции с помощью t-критерия Стьюдента проводится сопоставлением их значений с величиной случайной ошибки:

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

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

$$m_b = \sum(y - \hat{y}_x)^2/(n - 2);\ \sum(\bar{x} - x)^2;$$

$$m_a = \sqrt{\frac{\sum(y - \hat{y}_x)^2}{(n - 2)} \cdot \frac{\sum x^2}{n\sum(\bar{x} - x)^2}},$$

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

Сравнивая фактическое и критическое(табличное) значения t-статистики – $t_{табл}$ и $t_{факт}$ – принимаем или отвергаем гипотезу H_0.

Если $t_{табл} < t_{факт}$, то H_0 отклоняется, т.е. a, b, r_{xy} не случайно отличаются от нуля и сформировались под влиянием систематически действующего фактора x. Если $t_{табл} > t_{факт}$, то H_0 не отклоняется и признаётся случайная природа формирования a, b, r.

Для расчёта доверительных интервалов определяются предельные ошибки каждого показателя:

$\Delta_a = t_{табл}m_a$, $\Delta_b = t_{табл}m_b$.

Доверительные интервалы

$a - t_{табл}m_a < a < a + t_{табл}m_a$, $b - t_{табл}m_b < b < b + t_{табл}m_b$.

12.5. Множественная регрессия

Множественная регрессия – уравнение связи с несколькими независимыми переменными: $y = f(x_1, x_2,..., x_p)$.

Для построения модели множественной регрессии используются следующие функции:

- линейная – $y = a + b_1x_1 + b_2x_2 + ... + b_px_p + \varepsilon$;
- степенная – $y = ax_1^{b1}x_2^{b2}...x_p^{bp}\varepsilon$;
- экспоненциальная – $y = \exp(a + b_1x_1 + b_2x_2 + ... + b_px_p + \varepsilon)$;
- гипербола – $y = \frac{1}{a + b_1x_1 + b_2x_2 + ... + b_px_p + \varepsilon}$.

Последние три функции легко линеаризуются модификацией переменных.
Для оценки параметров линейной регрессии также используется метод наименьших квадратов.
Расчётные формулы для этих параметров

$$a = \frac{\Delta a}{\Delta},\ b_1 = \frac{\Delta b_1}{\Delta},\ ...,\ b_p = \frac{\Delta b_p}{\Delta},$$

где Δ – определитель системы нормальных уравнений

$$\Delta = \begin{vmatrix} n & \sum x_1 & \sum x_2 & ... & \sum x_p \\ \sum x_1 & \sum x_1^2 & \sum x_2 x_1 & ... & \sum x_p x_1 \\ \sum x_2 & \sum x_1 x_2 & \sum x_2^2 & & \sum x_p x_2 \\ ..... & & & & \\ \sum x_p & \sum x_1 x_p & \sum x_2 x_p & & \sum x_p^2 \end{vmatrix}$$

Δa, Δb_1, Δb_2,..., Δb_p – частные определители, которые получаются заменой соответствующего столбца определителя системы на столбец правых частей

$$\begin{vmatrix} \sum y \\ \sum y x_1 \\ ... \\ \sum y x_p \end{vmatrix}$$

Средние коэффициенты эластичности для линейной регрессии рассчитываются по формуле

$$\bar{Э}_j = b_j \frac{\bar{x}_j}{\bar{y}}.$$

Для линейной модели регрессии тесноту совместного влияния факторов на результат оценивает коэффициент множественной корреляции

$$R_{yx_1x_2...x_p} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}},$$

Δr – определитель матрицы парных коэффициентов корреляции, Δr_{11} – определитель матрицы межфакторной корреляции

$$\Delta r = \begin{vmatrix} 1 & r_{yx_1} & r_{yx_2} & ... & r_{yx_p} \\ r_{yx_1} & 1 & r_{x_1x_2} & ... & r_{x_1x_p} \\ r_{yx_2} & r_{x_2x_1} & 1 & & r_{x_2x_p} \\ ..... & & & & \\ r_{yx_p} & r_{x_px_1} & r_{x_px_2} & & 1 \end{vmatrix}$$

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x_1x_2} & r_{x_1x_3} & ... & r_{x_1x_p} \\ r_{x_2x_1} & 1 & r_{x_2x_3} & ... & r_{x_2x_p} \\ ..... & & & & \\ r_{x_px_1} & r_{x_px_2} & r_{x_px_3} & ... & 1 \end{vmatrix}$$

Коэффициент множественной детерминации рассчитывается как квадрат коэффициента множественной корреляции

$$R^2 = R^2_{yx_1x_2...x_p}$$

Значимость уравнения множественной регрессии в целом оценивается с помощью F-критерия Фишера:

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

12.6. Гомо- и гетероскедастичность остатков

Гомоскедастичность остатков означает, что дисперсии возмущающих воздействий постоянны:
$\sigma^2_i = \sigma^2_j$, i, j = 1, 2,..., n.
Гетероскедастичность означает непостоянство этих дисперсий. Непосредственное применение МНК даёт неточные оценки параметров регрессии. Для коррекции гетероскедастичности применяют взвешенный МНК, суть которого в том, чтобы «взвешивать» каждое наблюдение. При этом минимизируется взвешенная сумма квадратов отклонений

$$\sum_{i=1} \frac{1}{\sigma^2_i}(y_i - a - bx_i)^2.$$

Проблема заключается в оценке этих дисперсий.
Для экономических данных достаточно часто величина средней ошибки может быть пропорциональна абсолютному значению независимой величины. В этом случае коррекция гетероскедастичности достаточно проста: следует перейти к модифицированным данным

$$y_i^* = \frac{y_i}{x_i}, \; x_i^* = \frac{1}{x_i}.$$

Соотношения между оценками исходной и модифицированной регрессии:
$a = b^*$, $b = a^*$.

12.7. Системы одновременных уравнений

12.7.1. Модель спроса и предложения

Обозначим
Q^S_t, Q^D_t – соответственно объёмы предложения и спроса в момент времени t;
P_t – цена товара, Y_t – доход в момент времени t.
Предполагается, что на рынке существует равновесие между спросом и предложением $Q^S_t = Q^D_t = Q_t$.
Такое динамическое равновесие может быть представлено следующей моделью
$Q^S_t = \alpha_1 + \alpha_2 P_t + \varepsilon_t$ – предложение;
$Q^D_t = \beta_1 + \beta_2 P_t + \beta_3 Y_t + u_t$ – спрос.
Обозначим отклонения от средних

$q_t = Q_t - \overline{Q}_t$;

$p_t = P_t - \overline{P}_t$;

$y_t = Y_t - \overline{Y}_t$.
Тогда

$q_t + \overline{Q}_t = \alpha_1 + \alpha_2 p_t + \alpha_2 \overline{P}_t + \varepsilon_t$;

$q_t + \overline{Q}_t = \beta_1 + \beta_2 p_t + \beta_2 \overline{P}_t + \beta_3 y_t + \beta_3 \overline{Y}_t + u_t$.
Если усреднить исходные уравнения предложения и спроса, получим

$\overline{Q}_t = \alpha_1 + \alpha_2 \overline{P}_t$;

$\bar{Q}_t = \beta_1 + \beta_2\bar{P}_t + \beta_3\bar{Y}_t$.

С учётом этих соотношений предыдущие два уравнения запишутся как

$q_t = \alpha_2 p_t + \varepsilon_t$;

$q_t = \beta_2 p_t + \beta_3 y_t + u_t$. **(191)**

Эти два соотношения являются моделью предложения и спроса в отклонениях.

В соответствии с этой моделью цена и величина спроса-предложения определяются *одновременно*, поэтому такие системы называются системами одновременных уравнений.

Величины цены и спроса являются эндогенными переменными, которые должны определяться в рамках модели, а величина дохода – экзогенной переменной – её значение задаётся.

Предполагается, что в каждом уравнении экзогенные переменные некоррелированы с ошибкой. Тогда как эндогенные переменные, стоящие в правых частях уравнений, могут иметь отличную от нуля корреляцию с ошибкой в соответствующем уравнении. В этом случае непосредственное применение МНК даёт смещённые и несостоятельные оценки параметров регрессии α_2, β_2, β_3.

12.7.2. Структурная и приведённая форма системы

Система (191) называется структурной формой модели, а коэффициенты α_2, β_2, β_3 – структурными коэффициентами.

Разрешим систему (191) относительно p_t, q_t:

$$q_t = \frac{\alpha_2\beta_3 y_t}{\alpha_2 - \beta_2} + \frac{\alpha_2 u_t - \beta_2\varepsilon_t}{\alpha_2 - \beta_2};$$

$$p_t = \frac{\beta_3 y_t}{\alpha_2 - \beta_2} + \frac{u_t - \varepsilon_t}{\alpha_2 - \beta_2}.$$

Обозначая

$$\pi_1 = \frac{\alpha_2\beta_3}{\alpha_2 - \beta_2};\ \pi_2 = \frac{\beta_3}{\alpha_2 - \beta_2}; \quad \textbf{(192)}$$

$$v_{1t} = \frac{\alpha_2 u_t - \beta_2\varepsilon_t}{\alpha_2 - \beta_2};\ v_{2t} = \frac{u_t - \varepsilon_t}{\alpha_2 - \beta_2},$$

получаем

$q_t = \pi_1 y_t + v_{1t}$;

$p_t = \pi_2 y_t + v_{2t}$. **(193)**

Система (193), в которой все эндогенные переменные явно выражены через экзогенные переменные и случайные остаточные компоненты, называется приведённой формой системы одновременных уравнений.

В этой форме экзогенная переменная y_t некоррелирована с возмущениями v_{1t}, v_{2t}, поэтому МНК даст несмещённые и состоятельные оценки $\hat{\pi}_1$ и $\hat{\pi}_2$ коэффициентов π_1, π_2. Так как

$$\alpha_2 = \frac{\pi_1}{\pi_2}$$

по оценкам $\hat{\pi}_1$, $\hat{\pi}_2$ получаем оценку α_2:

$$\hat{\alpha}_2 = \frac{\hat{\pi}_1}{\hat{\pi}_2},$$

которая по теореме Слуцкого будет состоятельной оценкой структурного коэффициента α_2.

Такой способ оценивания структурных коэффициентов с помощью оценок коэффициентов приведённой формы называется косвенным методом наименьших квадратов.

Очевидно, таким же образом, нельзя воспользоваться соотношениями (192) для получения оценок структурных коэффициентов β_2, β_3. Таким образом, второе уравнение в системе (191) *неидентифицируемо*.

Неидентифицируемость этого уравнения можно также доказать следующим образом. Умножим первое уравнение (191) на λ, а второе – на $(1 - \lambda)$ и сложим их. Получим соотношение

$$q_t = \gamma_2 p_t + \gamma_3 y_t + \eta_t, \qquad \textbf{(194)}$$

где

$\gamma_2 = \lambda\alpha_2 + (1 - \lambda)\beta_3$, $\gamma_3 = (1 - \lambda)\beta_3$, $\eta_t = \lambda\varepsilon_t + (1 - \lambda)u_t$.

Уравнение (194) имеет такой же вид, что и второе уравнение в системе (191). Это значит, что *другие* коэффициенты γ_2, γ_3 также подходят для данных q_t, p_t, y_t. Таким образом, при любом методе оценивания структурных коэффициентов второго уравнения нельзя утверждать, получены оценки именно коэффициентов β_2, β_3.

12.7.3. Идентифицируемость систем одновременных уравнений

Общий вид системы одновременных уравнений

$$\begin{aligned}
&y_{1t} = \beta_{12}y_{2t} + \beta_{13}y_{3t} + \ldots + \beta_{1n}y_{nt} + \alpha_{11}x_{1t} + \ldots + \alpha_{1m}x_{mt} + \varepsilon_{1t};\\
&y_{2t} = \beta_{21}y_{1t} + \beta_{23}y_{3t} + \ldots + \beta_{2n}y_{nt} + \alpha_{21}x_{1t} + \ldots + \alpha_{2m}x_{mt} + \varepsilon_{2t};\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad \textbf{(195)}\\
&y_{nt} = \beta_{n1}y_{1t} + \beta_{n2}y_{2t} + \ldots + \beta_{nn-1}y_{n-1t} + \alpha_{n1}x_{1t} + \ldots + \alpha_{nm}x_{mt} + \varepsilon_{nt}.
\end{aligned}$$

Переменные y_{it} – определяются внутри системы и являются эндогенными;

x_i – внешние по отношению к системе, то есть экзогенные, переменные и лагированные значения эндогенных переменных, которые являются предопределёнными. Индекс t определяет момент наблюдения, t = 1, ..., n, а ε_{1t}, ε_{2t}, ..., ε_{nt} – случайные ошибки. Система (195) – структурная форма модели.

После решения системы (195) относительно эндогенных переменных получаем приведённую форму модели

$$\begin{aligned}
&y_{1t} = \delta_{11}x_1 + \delta_{12}x_2 + \ldots + \delta_{1m}x_m;\\
&y_{1t} = \delta_{21}x_1 + \delta_{22}x_2 + \ldots + \delta_{2m}x_m;\\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\\
&y_{nt} = \delta_{n1}x_1 + \delta_{n2}x_2 + \ldots + \delta_{nm}x_m. \qquad \textbf{(196)}
\end{aligned}$$

Обозначим P – количество эндогенных, Q – количество экзогенных и лаговых переменных, отсутствующих в уравнении. Тогда степень идентифицируемости уравнения определяется из следующих неравенств:

P = Q + 1 – уравнение идентифицируемо;

P > Q + 1 – уравнение неидентифицируемо;

P < Q + 1 – уравнение сверхидентифицируемо.

Эти условия являются лишь необходиыми, так как даже при их выполнении некторые уравнения могут оказаться линейно зависимыми.

Достаточным условием идентификации является следующее.

Определитель матрицы, составленной из коэффициентов при переменных, отсутствующих в исследуемом уравнении, не равен нулю, и ранг этой матрицы не менее числа эндогенных переменных системы минус единица.

Для решения идентифицируемого уравнения применяется косвенный метод наименьших квадратов, для решения сверхидентифицируемых – двухшаговый метод наименьших квадратов.

Косвенный МНК состоит из следующих этапов.

1) Структурную форму модели преобразуют к приведённой. Уравнения приведённой формы независимы друг от друга, поэтому для идентификации приведённых коэффициентов может быть применён обычный МНК.

2) Алгебраически выражают структурные коэффициенты через оценки приведённых. Полученные значения будут оценками структурных коэффициентов.

Двухшаговый МНК состоит из следующих этапов.

1) Вычисляют оценки параметров приведённой формы системы.

2) Определяют эндогенные переменные, находящиеся в правой части сверхидентифицируемого уравнения и находят расчётные значения этой переменной по соответствующим уравенениям приведённой формы.

3) Обычным МНК определяют параметры структурного уравнения, используя в качестве исходных данных фактические значения предопределённых переменных и расчётные значения эндогенных переменных, находящиеся в правой части сверхидентифицируемого уравнения.

13. РЕШЕНИЕ ТИПОВЫХ ЗАДАЧ

13.1. Парная линейная регрессия

Пример 1.

По группе предприятий имеются следующие данные об объёме производства продукции и соответствующем объёме затрат.

№	Объём произв. продукции, X	Объём затрат, Y
1	3,6	330
2	4,6	396
3	5,5	460
4	4,8	430
5	2,7	243
6	2	170
7	7,5	618
8	6,3	540
9	4,1	369
10	4,8	425
11	7,6	646
12	6,5	598
13	11,5	858
14	10,6	820
15	9	810
16	6,9	566
17	5	450
18	11,2	858
19	8,1	656
20	7,8	640
21	4,2	399
22	6,3	518
23	12,1	920
24	9,8	780
25	8,5	696
сумма	171	14196
средние	6,84	567,84

Для изучения тесноты связи между суммой затрат на производство продукции на одно предприятие (результативный признак y) и объёмом произведённой продукции (факторный признак x) рассчитать:

1) уравнение регрессии $\overline{y}_x = a_0 + a_1 x$;

2) парный коэффициент корреляции;

3) коэффициент детерминации;

4) коэффициент эластичности.

Дайте краткий анализ полученных результатов.

Решение

1)

Предварительные вычисления приведены в следующей таблице

№	Объём произв. продукции, X	Объём затрат, Y	X^2	Y^2	XY
1	3,6	330	12,96	108900	1188
2	4,6	396	21,16	156816	1821,6
3	5,5	460	30,25	211600	2530
4	4,8	430	23,04	184900	2064
5	2,7	243	7,29	59049	656,1
6	2	170	4	28900	340
7	7,5	618	56,25	381924	4635
8	6,3	540	39,69	291600	3402
9	4,1	369	16,81	136161	1512,9
10	4,8	425	23,04	180625	2040
11	7,6	646	57,76	417316	4909,6
12	6,5	598	42,25	357604	3887
13	11,5	858	132,25	736164	9867
14	10,6	820	112,36	672400	8692
15	9	810	81	656100	7290
16	6,9	566	47,61	320356	3905,4
17	5	450	25	202500	2250
18	11,2	858	125,44	736164	9609,6
19	8,1	656	65,61	430336	5313,6
20	7,8	640	60,84	409600	4992
21	4,2	399	17,64	159201	1675,8
22	6,3	518	39,69	268324	3263,4
23	12,1	920	146,41	846400	11132
24	9,8	780	96,04	608400	7644
25	8,5	696	72,25	484416	5916
сумма	171	14196	1356,64	9045756	110537
средние	6,84	567,84	54,27	361830,24	4421,48

коэффициенты a_0, a_1 определяем из системы нормальных уравнений

$$a_0 = \frac{(\Sigma Y \cdot \Sigma X^2 - \Sigma X \cdot \Sigma XY)}{n \cdot \Sigma X^2 - (\Sigma X)^2} = (14196 \cdot 135664 - 171 \cdot 110537)/(25 \cdot 1356{,}64 - 171^2) \approx 76{,}4;$$

$$a_1 = \frac{(n \cdot \Sigma XY - \Sigma X \Sigma Y)}{n \cdot \Sigma X^2 - (\Sigma X)^2} = (25 \cdot 110537 - 171 \cdot 14196)/(25 \cdot 1356{,}64 - 171^2) \approx 71{,}9.$$

2)

Коэффициент корреляции

$$r_{xy} = \frac{S_{xy}}{S_x S_y},$$

$$S_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(n-1)} = (25 \cdot 110537 - 171 \cdot 14196)/(25 \cdot 24) \approx 537{,}45;$$

$$S_x = \overline{X^2} - (\overline{X})^2 = 54{,}27 - 6{,}84^2 = 2{,}73;$$

$$S_Y = \overline{Y^2} - (\overline{Y})^2 = 361830{,}24 - 567{,}84^2 = 198{,}46;$$

$$r_{xy} = 537{,}45/(2{,}73 \cdot 198{,}46) \approx 0{,}99.$$

Высокий коэффициент корреляции показывает тесную положительную связь между факторным и результативным признаками.

3)

Коэффициент детерминации является квадратом коэффициента корреляции
$R^2 = 0{,}99^2 = 0{,}98$.
Достаточно высокий коэффициент детерминации характеризует высокое качество полученной модели регрессии.

4)

Коэффициент эластичности

$$Э_X = a_1 \frac{\bar{X}}{\bar{Y}} = 71{,}9 \cdot (6{,}84/567{,}84) \approx 0{,}87.$$

Таким образом, при увеличении объёма производства на 1% затраты увеличиваются в среднем на 0,87%.

Пример 2.

Имеются данные на 12-ти предприятиях объединения.

Предприятие	Капиталовложения, тыс. руб.	Объём производства, 10^4 руб.
1	16,3	52,15
2	16,8	48,15
3	18,5	54,15
4	16,3	50,15
5	17,9	54,15
6	17,4	53,15
7	16,1	53,15
8	16,2	52,15
9	17,0	53,15
10	16,7	52,15
11	17,5	53,15
12	19,1	60,15

Предполагая, что зависимость между переменными имеет линейный характер, провести анализ в следующей последовательности:

1. Построить рабочую таблицу для вычисления статистических характеристик.
2. Построить уравнение регрессии $\hat{y} = b_0 + b_1 x$ (x – капиталовложения), y – объём проивзодства.
3. Вычислить коэффициент корреляции величин y, x.
4. Оценить надёжность полученного уравнения регрессии по критерию Фишера.
5. Оценить надёжность полученного коэффициента корреляции по критерию Стьюдента для уровня значимости 5%.
6. Найти доверительные интервалы для коэффициента b_1 найденной регрессии.

Решение

1)

Предприятие	Капиталовложения, тыс. руб.	Объём производства, 10^4 руб.	Промежуточные результаты					
№	x	y	xy	x^2	y^2	$\hat{y}$	$e_i = y_i - \hat{y}_i$	$e_i^2 = (y_i - \hat{y}_i)^2$
1	16,3	52,15	850,05	265,7	2719,62	51,12	1,03	1,06

Предприятие	Капиталовложения, тыс. руб.	Объём производства, 10^4 руб.	Промежуточные результаты					
2	16,8	48,15	808,92	282,2	2318,42	52,22	-4,07	16,53
3	18,5	54,15	1001,78	342,3	2932,22	55,94	-1,79	3,22
4	16,3	50,15	817,45	265,7	2515,02	51,12	-0,97	0,94
5	17,9	54,15	969,29	320,4	2932,22	54,63	-0,48	0,23
6	17,4	53,15	924,81	302,8	2824,92	53,53	-0,38	0,15
7	16,1	53,15	855,72	259,2	2824,92	50,68	2,47	6,10
8	16,2	52,15	844,83	262,4	2719,62	50,90	1,25	1,56
9	17,0	53,15	903,55	289,0	2824,92	52,65	0,50	0,25
10	16,7	52,15	870,91	278,9	2719,62	52,00	0,15	0,02
11	17,5	53,15	930,13	306,3	2824,92	53,75	-0,60	0,36
12	19,1	60,15	1148,87	364,8	3618,02	57,26	2,89	8,36
Сумма	205,8	635,8	10926,27	3539,6	33774,5	635,8	0,00	38,8
Среднее	17,15	52,98	-	-	-	-	-	-
Дисперсия	0,92	7,97	-	-	-	-	3,53	-
Стандартное отклонение	0,96	2,82	-	-	-	-	1,88	-

2)

$b_1=(n\Sigma xy - \Sigma x\Sigma y)/(n\Sigma x^2 - (\Sigma x)^2) = (12\cdot 10926{,}27 - 205{,}8\cdot 635{,}8)/(12\cdot 3539{,}6 - 205{,}8^2) \approx 2{,}2;$

$b_0 = (\Sigma y - b_1\Sigma x)/n = (635{,}8 - 2{,}19\cdot 205{,}8)/12 \approx 15{,}4;$

$\hat{y} = 15{,}4 + 2{,}2x;$

столбец $\hat{y}$ в рабочей таблице был заполнен по полученному уравнению регрессии с применением абсолютных ссылок на ячейки, содержащие параметры выборочной регрессии.

3)

Коэффициент корреляции

$$r = \frac{S_{xy}}{S_x S_y},$$

$S_{xy} = (n\Sigma xy - \Sigma x\Sigma y)/n(n-1) = (12\cdot 10926{,}27 - 205{,}8\cdot 635{,}8)/12\cdot 11 \approx 2{,}03;$

$r = 2{,}03/(0{,}96\cdot 2{,}82) \approx 0{,}75.$

4)

Определяем

$F_{набл} = S_R^2(n-2)/S_\varepsilon^2 = (7{,}97 - 3{,}53)(12 - 1 - 1)/3{,}53 \approx 12{,}6;$

$n = 12$ – количество данных;

по таблице для 5% уровня значимости и степеней $f_1 = 1$, $f_2 = 10$

$F_{0.05;1;10} = 4{,}96.$

Так как

$F_{набл} > F_{0.05;1;10}$

следует сделать вывод о значимости полученного уравнения регрессии.

5)

$$T_{набл} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = 0{,}75\cdot\sqrt{10}\,/\sqrt{1-0{,}75^2} \approx 3{,}6;$$

критическое табличное значение

$T_{0,05/2;10} = T_{0,025;10} = 2{,}23.$

Так как $T_{набл} > T_{0,025;10}$ делаем вывод о значимости коэффициента корреляции и отклоняем гипотезу об отсутствии линейной корреляционной связи.

6)
Доверительные интервалы для b_1
$b_1 - T_{0,025;10} \cdot S_{b1} \le \beta_1 \le b_1 + T_{0,025;10} \cdot S_{b1}$;
$T_{0,025;10} = 2,23$;
$S_{b1} = S_e/(S_x \cdot \sqrt{n-1}) = 1,88/(0,96 \cdot \sqrt{11}) \approx 0,59$;
$T_{0,025;10} \cdot S_{b1} = 2,23 \cdot 0,59 \approx 1,3$;
теперь получаем доверительные границы
$2,2 - 1,3 \le \beta_1 \le 2,2 + 1,3$;
$0,9 \le \beta_1 \le 3,5$

13.2. Множественная линейная регрессия

Исследовать зависимость между объёмом производства, капитальными вложениями и выполнением норм выработки. Для построения модели собраны данные по исследуемым переменным на 12-ти предприятиях объединения.

Предприятие	Капиталовложения, тыс. руб.	Средний процент выполнения нормы	Объём производства, 10^4 руб.
1	16,334	99,534	52,834
2	16,834	98,934	48,434
3	18,534	99,234	54,234
4	16,334	99,334	50,034
5	17,934	99,834	54,934
6	17,434	99,634	53,934
7	16,134	99,834	53,134
8	16,234	99,734	52,434
9	17,034	99,834	53,034
10	16,734	99,934	52,934
11	17,534	100,034	53,134
12	19,134	100,234	60,134

Предполагая, что зависимость между переменными имеет линейный характер, анализ провести в следующей последовательности:
а) построить уравнение регрессии $\hat{y} = f(x_1)$;
б) построить уравнение регрессии $\hat{y} = f(x_2)$;
в) исследовать модели $y = f(x_1)$; $y = f(x_2)$ и сделать соответствующие выводы;
г) построить уравнение регрессии $\hat{y} = f(x_1, x_2)$ и выполнить исследование множественной модели в полном объёме.

Решение

1. Построение модели $\hat{y} = f(x_1) = b_0 + b_1 x_1$

Предприятие	Капиталовложения, тыс. руб.	Объём производства, 10^4 руб.	Промежуточные результаты					
№	x_1	y	$x_1 y$	x_1^2	y^2	$\hat{y}$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	16,334	52,834	862,991	266,800	2791,432	51,818	1,016	1,032
2	16,834	48,434	815,338	283,384	2345,852	52,431	-3,997	15,979
3	18,534	54,234	1005,173	343,509	2941,327	54,516	-0,282	0,079
4	16,334	50,034	817,255	266,800	2503,401	51,818	-1,784	3,184
5	17,934	54,934	985,186	321,628	3017,744	53,780	1,154	1,332
6	17,434	53,934	940,285	303,944	2908,876	53,167	0,767	0,588
7	16,134	53,134	857,264	260,306	2823,222	51,573	1,561	2,436
8	16,234	52,434	851,214	263,543	2749,324	51,696	0,738	0,545
9	17,034	53,034	903,381	290,157	2812,605	52,677	0,357	0,128
10	16,734	52,934	885,798	280,027	2802,008	52,309	0,625	0,391

11	17,534	53,134	931,652	307,441	2823,222	53,290	-0,156	0,024
Сумма	187,074	579,074	9855,536	3187,538	30519,014	579,074	0,000	25,718
Среднее	17,007	52,643	-	-	-	-	-	-
Дисперсия	0,602	3,477	-	-	-	-	2,572	-
Стандартное отклонение	0,776	1,865	-	-	-	-	1,604	-

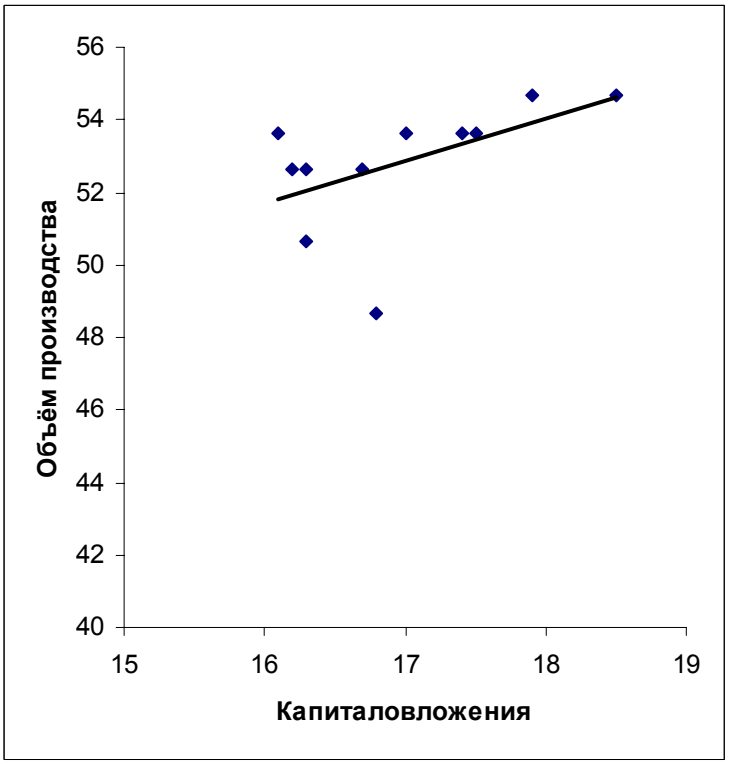
$b_1 = \frac{n\Sigma xy - \Sigma x\Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = (11 \cdot 9855{,}536 - 187{,}074 \cdot 579{,}074)/(11 \cdot 3187{,}538 - 187{,}074^2) \approx 1{,}226;$

$b_0 = (\Sigma y - b_1 \Sigma x)/n = (579{,}074 - 1{,}226 \cdot 187{,}074)/11 \approx 31{,}8;$

$\hat{y} = 31{,}8 + 1{,}226x_1;$

столбец $\hat{y}$ в рабочей таблице был заполнен по полученному уравнению регрессии.

По этим данным строим линию регрессии. На диаграмме также приведены исходные статистические данные.



Коэффициент корреляции

$r_{xy} = \frac{S_{xy}}{S_x S_y}$,

$S_{xy} = (n\Sigma xy - \Sigma x\Sigma y)/n(n-1) = (11 \cdot 9855{,}536 - 187{,}074 \cdot 579{,}074)/(11 \cdot 10) \approx 0{,}738;$

$r_{xy} = 0{,}738/(0{,}776 \cdot 1{,}865) \approx 0{,}510.$

Определяем

$F_{набл} = S_R^2(n - p - 1)/(S_\varepsilon^2 \cdot p) = (3{,}477 - 2{,}572)(11 - 1 - 1)/(2{,}572 \cdot 1) \approx 3{,}167;$

по таблице для 5% уровня значимости и степеней $f_1 = 1$, $f_2 = 9$

$F_{0.05;1;10} = 5{,}12.$

Так как

$F_{набл} < F_{0.05;1;10}$

нельзя сделать вывод о значимости полученного уравнения регрессии.

$T_{набл} = \frac{r\sqrt{n - p - 1}}{\sqrt{1 - r^2}} = 0{,}51 \cdot \sqrt{9} / \sqrt{1 - 0{,}51^2} \approx 1{,}78;$

критическое табличное значение

$T_{0,05;9} = 2{,}26.$

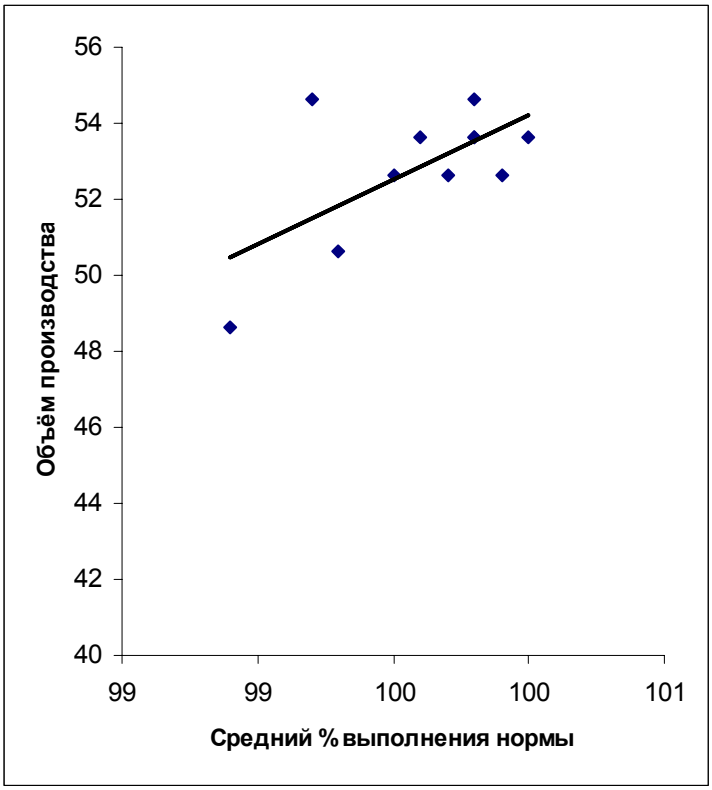
Так как $T_{набл} < T_{0,05;9}$ нельзя сделать вывод о значимости коэффициента корреляции и отклонить гипотезу об отсутствии линейной корреляционной связи.
Доверительные интервалы для b_1
$b_1 - T_{0,05;9} \cdot S_{b1} \le \beta_1 \le b_1 + T_{0,05;9} \cdot S_{b1}$;
$T_{0,05;9} = 2,26$;
$S_{b1} = S_\varepsilon / (S_{x1} \cdot \sqrt{n-1}) = 1,604/(0,776 \cdot \sqrt{10}) \approx 0,654$;
$T_{0,05;9} \cdot S_{b1} = 2,26 \cdot 0,654 \approx 1,477$;
теперь получаем доверительные границы
$1,477 - 1,41 \le \beta_1 \le 1,226 + 1,477$;
$-0,251 \le \beta_1 \le 2,703$.
Так как уравнение регрессии и коэффициент корреляции не являются значимыми, нельзя сделать вывод о существовании линейной зависимости между объёмами капиталовложений и производства для данной группы предприятий.

2. Построение модели $\hat{y} = f(x_2) = b_0 + b_1 x_2$

Предприятие	Средний процент выполнения нормы	Объём производства, 10^4 руб.	Промежуточные результаты					
№	x_2	y	$x_2 y$	x_2^2	y^2	$\hat{y}$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	99,534	52,834	5258,779	9907,017	2791,432	52,323	0,511	0,261
2	98,934	48,434	4791,769	9787,936	2345,852	50,213	-1,779	3,166
3	99,234	54,234	5381,857	9847,387	2941,327	51,268	2,966	8,795
4	99,334	50,034	4970,077	9867,244	2503,401	51,620	-1,586	2,515
5	99,834	54,934	5484,281	9966,828	3017,744	53,378	1,556	2,420
6	99,634	53,934	5373,660	9926,934	2908,876	52,675	1,259	1,585
7	99,834	53,134	5304,580	9966,828	2823,222	53,378	-0,244	0,060
8	99,734	52,434	5229,453	9946,871	2749,324	53,027	-0,593	0,351
9	99,834	53,034	5294,596	9966,828	2812,605	53,378	-0,344	0,119
10	99,934	52,934	5289,906	9986,804	2802,008	53,730	-0,796	0,634
11	100,034	53,134	5315,207	10006,801	2823,222	54,082	-0,948	0,898
Сумма	1095,874	579,074	57694,166	109177,477	30519,014	579,074	0,000	20,804
Среднее	99,625	52,643						
Дисперсия	0,113	3,477					2,080	
Стандартное отклонение	0,336	1,865					1,442	

$$b_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = (11 \cdot 57694,166 - 1095,874 \cdot 579,074)/(11 \cdot 1091177,477 - 1095,874^2) \approx 3,517;$$

$b_0 = (\Sigma y - b_1 \Sigma x)/n = (579,074 - 3,517 \cdot 1095,874)/11 \approx -297,729$;
$\hat{y} = -297,729 + 3,517 x_2$;
столбец $\hat{y}$ в рабочей таблице был заполнен по полученному уравнению регрессии.
По этим данным строим линию регрессии. На диаграмме также приведены исходные статистические данные.



Коэффициент корреляции

$r_{xy} = \frac{S_{xy}}{S_x S_y}$,

$S_{xy} = (n\Sigma xy - \Sigma x \Sigma y)/n(n-1) = (11 \cdot 57694{,}166 - 1095{,}874 \cdot 579{,}074)/(11 \cdot 10) \approx 0{,}397;$

$r_{xy} = 0{,}397/(0{,}336 \cdot 1{,}865) \approx 0{,}634$.

Определяем

$F_{набл} = S_R^2(n-p-1)/(S_\varepsilon^2 \cdot p) = (3{,}477 - 2{,}080)(11-1-1)/(1{,}865 \cdot 1) \approx 6{,}041;$

по таблице для 5% уровня значимости и степеней $f_1 = 1$, $f_2 = 9$

$F_{0.05;1;9} = 5{,}12.$

Так как

$F_{набл} > F_{0.05;1;9}$

следует сделать вывод о значимости полученного уравнения регрессии.

$T_{набл} = \frac{r\sqrt{n-p-1}}{\sqrt{1-r^2}} = 0{,}634 \cdot \sqrt{9} / \sqrt{1 - 0{,}634^2} \approx 2{,}458;$

критическое табличное значение

$T_{0,05;9} = T_{0,05;9} = 2{,}26.$

Так как $T_{набл} > T_{0,025;10}$ делаем вывод о значимости коэффициента корреляции и отклоняем гипотезу об отсутствии линейной корреляционной связи.

Доверительные интервалы для b_1

$b_1 - T_{0,05;9} \cdot S_{b1} \le \beta_1 \le b_1 + T_{0,05;9} \cdot S_{b1};$

$T_{0,05;10} = 2{,}26;$

$S_{b1} = S_\varepsilon/(S_{x1} \cdot \sqrt{n-1}) = 1{,}442/(0{,}336 \cdot \sqrt{10}) \approx 1{,}357;$

$T_{0,05;9} \cdot S_{b1} = 2{,}26 \cdot 1{,}357 \approx 3{,}067;$

теперь получаем доверительные границы

$3{,}517 - 3{,}067 \le \beta_1 \le 3{,}517 + 3{,}067,$

$0{,}449 \le \beta_1 \le 6{,}584$

Из значимости коэффициента корреляции и достаточной степени надёжности уравнения регрессии можно сделать вывод о существовании линейной зависимости между средним процентом выполнения нормы и объёмом производства.

3. Множественная модель регрессии $\hat{y} = f(x_1,x_2) = b_0 + b_1x_1 + b_2x_2$

Из системы нормальных уравнений

$$b_0 = \frac{\Delta_0}{\Delta};\ b_1 = \frac{\Delta_1}{\Delta};\ b_2 = \frac{\Delta_2}{\Delta};$$

$$\Delta = \begin{vmatrix} n & \Sigma x_1 & \Sigma x_2 \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma x_2x_1 \\ \Sigma x_2 & \Sigma x_2x_1 & \Sigma x_2^2 \end{vmatrix}$$

$$\Delta_0 = \begin{vmatrix} \Sigma y & \Sigma x_1 & \Sigma x_2 \\ \Sigma yx_1 & \Sigma x_1^2 & \Sigma x_2x_1 \\ \Sigma yx_2 & \Sigma x_2x_1 & \Sigma x_2^2 \end{vmatrix}$$

$$\Delta_1 = \begin{vmatrix} n & \Sigma y & \Sigma x_2 \\ \Sigma x_1 & \Sigma yx_1 & \Sigma x_2x_1 \\ \Sigma x_2 & \Sigma yx_2 & \Sigma x_2^2 \end{vmatrix}$$

$$\Delta_2 = \begin{vmatrix} n & \Sigma x_1 & \Sigma y \\ \Sigma x_1 & \Sigma x_1^2 & \Sigma yx_1 \\ \Sigma x_2 & \Sigma x_2x_1 & \Sigma yx_2 \end{vmatrix}$$

Все элементы определителей уже подсчитаны ранее, кроме Σx_2x_1:

№	x_1	x_2	x_1x_2
1	16,334	99,534	1625,7884
2	16,834	98,934	1665,455
3	18,534	99,234	1839,203
4	16,334	99,334	1622,5216
5	17,934	99,834	1790,423
6	17,434	99,634	1737,0192
7	16,134	99,834	1610,7218
8	16,234	99,734	1619,0818
9	17,034	99,834	1700,5724
10	16,734	99,934	1672,2956
11	17,534	100,034	1753,9962
Сумма	187,074	1095,874	18637,078

Далее вычисляем

$$\Delta_0 = \begin{vmatrix} 11 & 187,1 & 1095,9 \\ 187,1 & 3187,5 & 18637,1 \\ 1095,874 & 18637,08 & 109177,5 \end{vmatrix} = 74,5344$$

$$\Delta_1 = \begin{vmatrix} 579,07 & 187,074 & 1095,874 \\ 9855,54 & 3187,538 & 18637,08 \\ 57694,17 & 18637,08 & 109177,5 \end{vmatrix} = -25189,425$$

$$\Delta_2 = \begin{vmatrix} 11 & 579,07 & 1095,874 \\ 187,074 & 9855,54 & 18637,08 \\ 1095,874 & 57694,17 & 109177,5 \end{vmatrix} = 98,3646$$

$$\Delta_3 = \begin{vmatrix} 11 & 187,074 & 579,07 \\ 187,074 & 3187,538 & 9855,54 \\ 1095,874 & 18637,08 & 57694,17 \end{vmatrix} = 275,436$$

$b_0 = -25189,425/74,534 \approx -337,96$; $b_1 = 98,365/74,534 \approx 1,320$; $b_2 = 275,436/74,534 \approx 3,695$.

Модель регрессии

$\hat{y} = -337{,}96 + 1{,}32x_1 + 3{,}695x_2$

Коэффициент корреляции

$$r_{x1x2} = \frac{\overline{x_1x_2} - \bar{x}_1\bar{x}_2}{\sigma_{x1}\sigma_{x2}} = \frac{\frac{18637{,}078}{11} - 99{,}62 \cdot 17{,}01}{0{,}776 \cdot 0{,}336} \approx -0{,}053;$$

Определитель матрицы парных коэффициентов корреляции

$$\begin{vmatrix} 1 & 0{,}510 & 0{,}634 \\ 0{,}510 & 1 & -0{,}053 \\ 0{,}634 & -0{,}053 & 1 \end{vmatrix} = 0{,}301$$

Определитель матрицы межфакторной корреляции

$$\begin{vmatrix} 1 & -0{,}053 \\ -0{,}053 & 1 \end{vmatrix} = 0{,}997$$

Коэффициент множественной корреляции

$$R = R_{yx1x2} = \sqrt{1 - \frac{0{,}301}{0{,}997}} \approx 0{,}836.$$

F–критерий

$$F = \frac{R^2}{1 - R^2} \cdot \frac{n - m - 1}{m} = \frac{0{,}836^2}{1 - 0{,}836^2} \cdot \frac{11 - 2 - 1}{2} = 9{,}262;$$

табличное значение для $\alpha = 0{,}05$

$F_{0{,}05;8;2} = 4{,}46.$

Так как $F > F_{0{,}05;8;2}$ модель регрессии является значимой. Тесноту связи можно характеризовать как тесную положительную.

Частные F–критерии

$$F_{x1} = \frac{R^2 - r^2_{yx2}}{1 - R^2} \cdot \frac{n - m - 1}{1} = \frac{0{,}836^2 - 0{,}634^2}{1 - 0{,}836^2} \cdot \frac{11 - 2 - 1}{1} = 7{,}871;$$

табличное значение для $\alpha = 0{,}05$

$F_{0{,}05;8;1} = 5{,}32.$

Так как $F_{x1} > F_{0{,}05;8;1}$ целесообразно включать фактор x_1 после фактора x_2.

$$F_{x2} = \frac{R^2 - r^2_{yx1}}{1 - R^2} \cdot \frac{n - m - 1}{1} = \frac{0{,}836^2 - 0{,}51^2}{1 - 0{,}836^2} \cdot \frac{11 - 2 - 1}{1} = 11{,}620;$$

Так как $F_{x2} > F_{0{,}05;8;1}$ целесообразно включать фактор x_2 после фактора x_1.

13.3. Парная нелинейная регрессия

По 8 продовольственным магазинам имеются следующие данные:

Товарооборот, тыс. руб.	7	10	15	20	30	45	60	120
Уровень издержек обращения по отношению к товарообороту, %	10	9	7,2	6	6,3	5,8	5,4	5

а) Найдите уравнение корреляционной связи товарооборота и уровня издержек обращения.

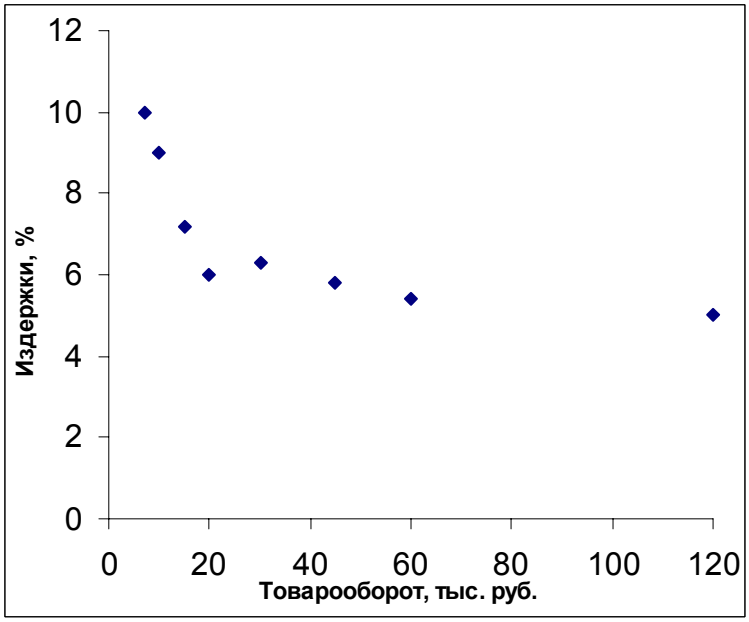
б) Изобразите графически корреляционную связь.

в) Вычислите коэффициенты эластичности, показатели тесноты корреляционной связи

Решение

а)

Изобразим графически приведённые данные



Как видно, зависимость между показателями нелинейная. Будем искать регрессионную зависимость в виде

$y = ax^b$;

переходя к логарифмам, получаем линейную зависимость

$$\ln y = \ln a + b \ln x$$

$y^* = a^* + bx^*$,

где $y = \ln y$, $x^* = \ln x$.

Определяем коэффициенты a*, b для модифицированных данных. Основные вычисления приведены в следующей таблице.

№	$x^*=\ln x$	$y^*=\ln y$	x^2	y^2
1	1,95	2,30	3,79	5,30
2	2,30	2,20	5,30	4,83
3	2,71	1,97	7,33	3,90
4	3,00	1,79	8,97	3,21
5	3,40	1,84	11,57	3,39
6	3,81	1,76	14,49	3,09
7	4,09	1,69	16,76	2,84
8	4,79	1,61	22,92	2,59
Итого	26,04	15,16	91,14	29,15

коэффициенты a*, b определяем из известных соотношений

$$a^* = \frac{\Sigma y_i \Sigma x_i^2 - \Sigma x_i \Sigma x_i y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = (15{,}15 \cdot 91{,}14 - 26{,}04 \cdot 47{,}81)/(8 \cdot 91{,}14 - 26{,}04^2) \approx 2{,}68;$$

$$b = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n\Sigma x_i^2 - (\Sigma x_i)^2} = (8 \cdot 47{,}81 - 26{,}04 \cdot 15{,}16)/(8 \cdot 91{,}14 - 26{,}04^2) \approx -0{,}24.$$

По величине a* = 2,68 потенцированием находим коэффициент a:

$a = e^{2{,}68} \approx 14{,}58$.

Таким образом, корреляционная связь между показателями

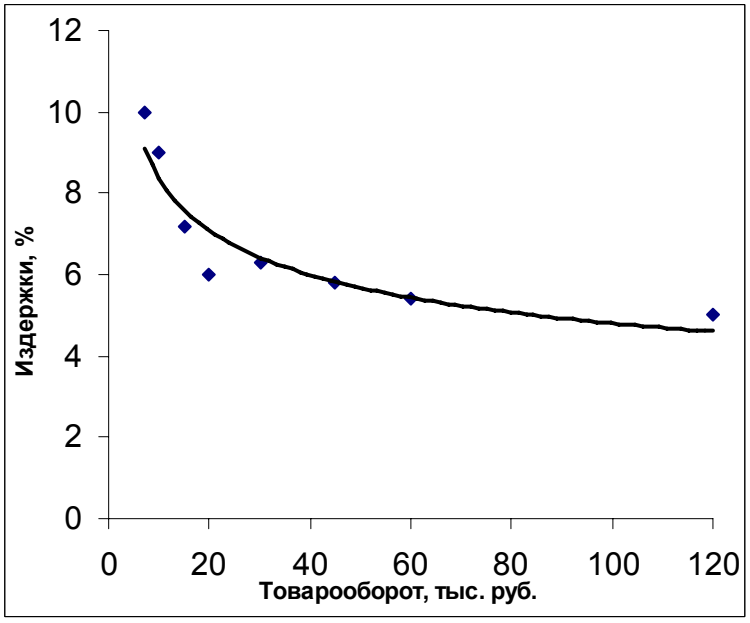
$y = 14{,}58x^{-0{,}24}$.

Строим корреляционную связь по следующей таблице

Товарооборот, X	y = 14,58x-0,24
7	9,12
10	8,37
15	7,59
20	7,08
30	6,42
45	5,82
60	5,43
120	4,60

б)

Вместе с исходными данными получаем следующую диаграмму



в)

Теснота связи характеризуется индексом корреляции

$$\rho = \sqrt{1 - \frac{\sum(y - \hat{y}_x)^2}{\sum(y - \bar{y}_x)^2}}$$

$\hat{y}_x$ – теоретическое значение показателя y;

$\bar{y}_x$ – среднее значение y.

Необходимые вычисления приведены в следующей таблице

	Товарооборот, X	Уровень издержек, y	$(y - \bar{y}_x)^2$	$\hat{y}_x$	$(y - \hat{y}_x)^2$
1	7	10	10,00	9,12	0,77
2	10	9	4,68	8,37	0,40
3	15	7,2	0,13	7,59	0,15
4	20	6	0,70	7,08	1,17
5	30	6,3	0,29	6,42	0,02
6	45	5,8	1,08	5,82	0,00
7	60	5,4	2,07	5,43	0,00
8	120	5	3,38	4,60	0,16
Итого	307	54,7	22,32	-	2,67

средние	38,38	6,84	-	-	-

$\rho = \sqrt{1 - \frac{2,67}{22,32}} \approx 0,94.$

Коэффициент эластичности для степенной зависимости равен показателю степени
Э = –0,24.
Это значит, что при возрастании товарооборота на 1% уровень издержек по отношению к величине товарооборота уменьшится в среднем на 0,24%.

13.4. Система одновременных уравнений

Пример 1.
Рассмотрим следующую модель одновременных уравнений
$y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + \varepsilon_1;$
$y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2;$
$y_3 = b_{31}y_1 + a_{33}x_3 + \varepsilon_3.$

1) Определить степень идентифицируемости каждого уравнения.
2) Составить приведённую форму.

Решение
1)
В первом уравнении две эндогенные переменные y_1, y_2 и отсутствует одна экзогенная, поэтому
P = 2, Q = 1, P = Q + 1,
выполнено необходимое условие идентификации.
Достоточное условие идентификации. Отсутствуют переменные y_3, x_3; матрица коэффициентов при этих переменных

$$\begin{Vmatrix} 0 & a_{23} \\ -1 & a_{33} \end{Vmatrix}$$

определитель этой матрицы не равен нулю, и ранг равен двум, что на единицу меньше количества эндогенных переменных. Поэтому достаточное условие идентификации выполняется.
Во втором уравнении две эндогенные перменные y_1, y_2 и отсутствует одна экзогенная, поэтому
P = 2, Q = 1, P = Q + 1,
выполнено необходимое условие идентификации.
Достоточное условие идентификации. Отсутствуют переменные y_3, x_1; матрица коэффициентов при этих переменных

$$\begin{Vmatrix} 0 & a_{11} \\ -1 & 0 \end{Vmatrix}$$

определитель этой матрицы также не равен нулю, ранг равен количеству эндогенных переменных минус единица. Поэтому достаточное условие идентификации для второго уравнения выполняется.
Третье уравнение сверхидентифицируемо, поскольку
P = 2, Q = 2, P < Q + 1.
2)
Запишем систему как

$y_1 - b_{12}y_2 = a_{11}x_1 + a_{12}x_2 + \varepsilon_1;$
$y_2 - b_{21}y_1 = a_{22}x_2 + a_{23}x_3 + \varepsilon_2;$
$y_3 - b_{31}y_1 = a_{33}x_3 + \varepsilon_3;$
определители системы

$$\begin{vmatrix} 1 & -b_{12} & 0 \end{vmatrix}$$

$$\Delta = \begin{vmatrix} -b_{21} & 1 & 0 \\ -b_{31} & 0 & 1 \end{vmatrix} = 1 - b_{12}b_{21}$$

$$\Delta_1 = \begin{vmatrix} a_{11}x_1 + a_{12}x_2 + \varepsilon_1 & -b_{12} & 0 \\ a_{22}x_2 + a_{23}x_3 + \varepsilon_2 & 1 & 0 \\ a_{33}x_3 + \varepsilon_3 & 0 & 1 \end{vmatrix} = a_{11}x_1 + a_{12}x_2 + \varepsilon_1 + b_{12}(a_{22}x_2 + a_{23}x_3 + \varepsilon_2) =$$
$$= a_{11}x_1 + (a_{12} + b_{12}a_{22})x_2 + b_{12}a_{23}x_3 + \varepsilon_1 + b_{12}\varepsilon_2;$$

$$\Delta_2 = \begin{vmatrix} 1 & a_{11}x_1 + a_{12}x_2 + \varepsilon_1 & 0 \\ -b_{21} & a_{22}x_2 + a_{23}x_3 + \varepsilon_2 & 0 \\ -b_{31} & a_{33}x_3 + \varepsilon_3 & 1 \end{vmatrix} = a_{22}x_2 + a_{23}x_3 + \varepsilon_2 + b_{21}(a_{11}x_1 + a_{12}x_2 + \varepsilon_1) =$$
$$= a_{11}b_{21}x_1 + (a_{22} + b_{21}a_{12})x_2 + a_{23}x_3 + \varepsilon_2 + b_{21}\varepsilon_1;$$

$$\Delta_3 = \begin{vmatrix} 1 & -b_{12} & a_{11}x_1 + a_{12}x_2 + \varepsilon_1 \\ -b_{21} & 1 & a_{22}x_2 + a_{23}x_3 + \varepsilon_2 \\ -b_{31} & 0 & a_{33}x_3 + \varepsilon_3 \end{vmatrix} = a_{33}x_3 + \varepsilon_3 + b_{12}b_{31}(a_{22}x_2 + a_{23}x_3 + \varepsilon_2) +$$
$$+ b_{31}(a_{11}x_1 + a_{12}x_2 + \varepsilon_1) - b_{12}b_{21}(a_{33}x_3 + \varepsilon_3) =$$
$$= a_{11}b_{31}x_1 + (a_{22}b_{12}b_{31} + a_{12}b_{31})x_2 + (a_{33} + a_{23}b_{12}b_{31} - a_{33}b_{12}b_{21})x_3 + b_{31}\varepsilon_1 + b_{12}b_{31}\varepsilon_2 + (1 - b_{12}b_{21})\varepsilon_3;$$

получаем приведённую систему

$$y_1 = \frac{a_{11}}{1 - b_{12}b_{21}}x_1 + \frac{a_{12} + b_{12}a_{22}}{1 - b_{12}b_{21}}x_2 + \frac{b_{12}a_{23}}{1 - b_{12}b_{21}}x_3 + \frac{\varepsilon_1 + b_{12}\varepsilon_2}{1 - b_{12}b_{21}};$$

$$y_2 = \frac{a_{11}b_{21}}{1 - b_{12}b_{21}}x_1 + \frac{a_{22} + b_{21}a_{12}}{1 - b_{12}b_{21}}x_2 + \frac{a_{23}}{1 - b_{12}b_{21}}x_3 + \frac{\varepsilon_2 + b_{21}\varepsilon_1}{1 - b_{12}b_{21}};$$

$$y_3 = \frac{a_{11}b_{31}}{1 - b_{12}b_{21}}x_1 + \frac{a_{22}b_{12}b_{31} + a_{12}b_{31}}{1 - b_{12}b_{21}}x_2 + \frac{a_{33} + a_{23}b_{12}b_{31} - a_{33}b_{12}b_{21}}{1 - b_{12}b_{21}}x_3 + \frac{b_{12}b_{31}\varepsilon_2 + (1 - b_{12}b_{21})\varepsilon_3}{1 - b_{12}b_{21}}.$$

Пример 2.

Рассмотрим следующую модель одновременных уравнений

$$\begin{cases} y_1 = b_{12}y_2 + a_{12}x_2 + \varepsilon_1; \\ y_2 = b_{21}y_1 + a_{21}x_1 + \varepsilon_2; \end{cases}$$

1) Определить степень идентифицируемости каждого уравнения.
2) Составить приведённую форму.
3) Идентифицировать систему.

По этой модели были сгенерированы данные со следующими исходными параметрами
$a_{12} = 0{,}412$; $a_{21} = 0{,}192$; $b_{12} = 0{,}672$; $b_{21} = 0{,}137$;
среднеквадратические отклонения
$\sigma_{\varepsilon 1} = 0{,}003$; $\sigma_{\varepsilon 2} = 0{,}006$.
Данные представлены в следующей таблице

№	y_1	y_2	x_1	x_2
1	1,241	0,187	0,1	2,71
2	1,854	0,301	0,2	4,034
3	2,549	0,401	0,3	5,53
4	3,144	0,516	0,4	6,807
5	4,708	0,734	0,5	10,23
6	5,13	0,82	0,6	11,13
7	5,099	0,844	0,7	11,03

8	5,635	0,927	0,8	12,17
9	4,53	0,798	0,9	9,714
10	7,426	1,217	1	16,06
11	5,269	0,939	1,1	11,28
Средн	4,2	0,7	0,6	9,2

Решение

1)

В первом уравнении две эндогенные переменные y_1, y_2 и отсутствует одна экзогенная, поэтому

$P = 2, Q = 1, P = Q + 1,$

выполнено необходимое условие идентификации. Очевидно выполнено и достаточное условие идентификации, так как отсутствует переменная x_1, коэффициент при которой во втором уравнении не равен нулю. Таким образом, первое уравнение идентифицируемо точно. Анлогично, второе уравнение также точно идентифицируемо.

2)

$$\begin{cases} y_1 = b_{12}y_2 + a_{12}x_2 + \varepsilon_1; \\ y_2 = b_{21}y_1 + a_{21}x_1 + \varepsilon_2; \end{cases}$$

$$y_1 = b_{12}(b_{21}y_1 + a_{21}x_1 + \varepsilon_2) + a_{12}x_2 + \varepsilon_1 = b_{12}b_{21}y_1 + b_{12}a_{21}x_1 + a_{12}x_2 + \varepsilon_1 + b_{12}\varepsilon_2;$$

$$y_1 = \frac{a_{21}b_{12}}{1 - b_{12}b_{21}}x_1 + \frac{a_{12}}{1 - b_{12}b_{21}}x_2 + \frac{\varepsilon_1 + b_{12}\varepsilon_2}{1 - b_{12}b_{21}};$$

$$y_2 = b_{21}\frac{b_{12}a_{21}}{1 - b_{12}b_{21}}x_1 + b_{21}\frac{a_{12}}{1 - b_{12}b_{21}}x_2 + b_{21}\frac{\varepsilon_1 + b_{12}\varepsilon_2}{1 - b_{12}b_{21}} + a_{21}x_1 + \varepsilon_2 =$$

$$= \frac{a_{21}}{1 - b_{12}b_{21}}x_1 + \frac{a_{12}b_{21}}{1 - b_{12}b_{21}}x_2 + \frac{b_{21}\varepsilon_1 + \varepsilon_2}{1 - b_{12}b_{21}}.$$

Получаем приведённую систему

$$\begin{cases} y_1 = \frac{a_{21}b_{12}}{1 - b_{12}b_{21}}x_1 + \frac{a_{12}}{1 - b_{12}b_{21}}x_2 + \frac{\varepsilon_1 + b_{12}\varepsilon_2}{1 - b_{12}b_{21}}; \\ y_2 = \frac{a_{21}}{1 - b_{12}b_{21}}x_1 + \frac{a_{12}b_{21}}{1 - b_{12}b_{21}}x_2 + \frac{b_{21}\varepsilon_1 + \varepsilon_2}{1 - b_{12}b_{21}}; \end{cases}$$

приведённые параметры

$$\delta_{11} = \frac{a_{21}b_{12}}{1 - b_{12}b_{21}}; \delta_{12} = \frac{a_{12}}{1 - b_{12}b_{21}}; u_1 = \frac{\varepsilon_1 + b_{12}\varepsilon_2}{1 - b_{12}b_{21}};$$

$$\delta_{21} = \frac{a_{21}}{1 - b_{12}b_{21}}; \delta_{22} = \frac{a_{12}b_{21}}{1 - b_{12}b_{21}}; u_2 = \frac{b_{21}\varepsilon_1 + \varepsilon_2}{1 - b_{12}b_{21}}.$$

Получаем следующую приведённу модель

$$y_1 = \delta_{11}x_1 + \delta_{12}x_2 + u_1;$$
$$y_2 = \delta_{21}x_1 + \delta_{22}x_2 + u_2;$$

структурные коэффициенты по приведённым вычисляются следующим образом:

$$b_{12} = \frac{\delta_{11}}{\delta_{21}}; b_{21} = \frac{\delta_{22}}{\delta_{12}};$$

$$a_{21} = \delta_{21} - b_{21}\delta_{11}; a_{12} = \delta_{12} - b_{12}\delta_{22}.$$

3)

Первое приведённое уравнение

$$y_1 = \delta_{11}x_1 + \delta_{12}x_2 + u_1.$$

Из системы нормальных уравнений

$$\delta_{11} = \frac{\Delta_{11}}{\Delta}; \delta_{12} = \frac{\Delta_{12}}{\Delta};$$

$$\Delta = \begin{vmatrix} \Sigma\bar{x}_1^2 & \Sigma\bar{x}_2\bar{x}_1 \\ \Sigma\bar{x}_2\bar{x}_1 & \Sigma\bar{x}_2^2 \end{vmatrix}$$

$$\Delta_{11} = \begin{vmatrix} \Sigma\bar{x}_1\bar{y}_1 & \Sigma\bar{x}_2\bar{x}_1 \\ \Sigma\bar{x}_2\bar{y}_1 & \Sigma\bar{x}_2^2 \end{vmatrix}$$

$$\Delta_{12} = \begin{vmatrix} \Sigma\bar{x}_1^2 & \Sigma\bar{x}_1\bar{y}_1 \\ \Sigma\bar{x}_2\bar{x}_1 & \Sigma\bar{x}_2\bar{y}_1 \end{vmatrix}$$

Второе приведённое уравнение

$y_2 = \delta_{21}x_1 + \delta_{22}x_2 + u_2$.

Из системы нормальных уравнений

$$\delta_{21} = \frac{\Delta_{21}}{\Delta};\ \delta_{22} = \frac{\Delta_{22}}{\Delta};$$

$$\Delta = \begin{vmatrix} \Sigma\bar{x}_1^2 & \Sigma\bar{x}_2\bar{x}_1 \\ \Sigma\bar{x}_2\bar{x}_1 & \Sigma\bar{x}_2^2 \end{vmatrix}$$

$$\Delta_{21} = \begin{vmatrix} \Sigma\bar{x}_1\bar{y}_2 & \Sigma\bar{x}_2\bar{x}_1 \\ \Sigma\bar{x}_2\bar{y}_2 & \Sigma\bar{x}_2^2 \end{vmatrix}$$

$$\Delta_{22} = \begin{vmatrix} \Sigma\bar{x}_1^2 & \Sigma\bar{x}_1\bar{y}_2 \\ \Sigma\bar{x}_2\bar{x}_1 & \Sigma\bar{x}_2\bar{y}_2 \end{vmatrix}$$

Далее представлены численные расчёты

№	$\bar{y}_1$	$\bar{y}_2$	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_1^2$	$\bar{x}_2^2$	$\bar{x}_1\bar{x}_2$	$\bar{x}_1\bar{y}_1$	$\bar{x}_2\bar{y}_1$	$\bar{x}_1\bar{y}_2$	$\bar{x}_2\bar{y}_2$
1	-3,0	-0,5	-0,5	-6,4	0,25	41,526	3,222	1,497	19,3	0,256	3,3
2	-2,4	-0,4	-0,4	-5,1	0,16	26,208	2,0478	0,952	12,19	0,159	2
3	-1,7	-0,3	-0,3	-3,6	0,09	13,135	1,0873	0,506	6,11	0,089	1,1
4	-1,1	-0,2	-0,2	-2,3	0,04	5,5089	0,4694	0,218	2,56	0,036	0,4
5	0,5	0,0	-0,1	1,1	0,01	1,1556	-0,107	-0,05	0,508	-0	0
6	0,9	0,1	0,0	2,0	0	3,8937	0	0	1,766	0	0,2
7	0,9	0,1	0,1	1,9	0,01	3,5211	0,1876	0,086	1,621	0,015	0,3
8	1,4	0,2	0,2	3,0	0,04	9,0873	0,6029	0,28	4,219	0,046	0,7
9	0,3	0,1	0,3	0,6	0,09	0,3137	0,168	0,088	0,165	0,03	0,1
10	3,2	0,5	0,4	6,9	0,16	47,723	2,7633	1,276	22,05	0,207	3,6
11	1,0	0,2	0,5	2,1	0,25	4,5257	1,0637	0,517	2,201	0,12	0,5
Всего	0,0	0,0	0,0	0,0	1,1	156,6	11,5	5,4	72,7	1,0	12,2

$$\Delta = \begin{vmatrix} 1,1 & 11,5 \\ 11,50 & 156,6 \end{vmatrix} = 39,9$$

$$\Delta_{11} = \begin{vmatrix} 5,4 & 11,5 \\ 72,68 & 156,6 \end{vmatrix} = 5,5$$

$$\Delta_{12} = \begin{vmatrix} 1,1 & 5,4 \\ 11,5 & 72,7 \end{vmatrix} = 18,1$$

$$\Delta_{21} = \begin{vmatrix} 1,0 & 11,5 \\ 12,22 & 156,6 \end{vmatrix} = 8,8$$

$$\Delta_{22} = \begin{vmatrix} 1,1 & 1,0 \\ 11,5 & 12,2 \end{vmatrix} = 2,5$$

Находим приведённые коэффициенты

$$\hat{\delta}_{11} = \frac{\Delta_{11}}{\Delta} = 5,5/39,9 = 0,138;$$

$$\hat{\delta}_{12} = \frac{\Delta_{12}}{\Delta} = 18,1/39,9 = 0,454;$$

$$\hat{\delta}_{21} = \frac{\Delta_{21}}{\Delta} = 8,8/39,9 = 0,222;$$

$$\hat{\delta}_{22} = \frac{\Delta_{22}}{\Delta} = 2,5/39,9 = 0,062.$$

Теперь находим оценки структурных коэффициентов и их относительные ошибки

$$\hat{b}_{12} = \frac{\hat{\delta}_{11}}{\hat{\delta}_{21}} = 0,138/0,222 = 0,623;\ \delta_{b_{12}} = \frac{|0,623 - 0,672|}{0,672}100\% = 7\%;$$

$$\hat{b}_{21} = \frac{\hat{\delta}_{22}}{\hat{\delta}_{12}} = 0,062/0,454 = 0,136;\ \delta_{b_{21}} = \frac{|0,136 - 0,137|}{0,137}100\% = 0,9\%;$$

$$\hat{a}_{21} = \hat{\delta}_{21} - \hat{b}_{21}\hat{\delta}_{11} = 0,222 - 0,136 \cdot 0,138 = 0,203;\ \delta_{a_{21}} = \frac{|0,203 - 0,192|}{0,192}100\% = 6\%;$$

$$\hat{a}_{12} = \hat{\delta}_{12} - \hat{b}_{12}\hat{\delta}_{22} = 0,454 - 0,623 \cdot 0,062 = 0,415;\ \delta_{a_{21}} = \frac{|0,415 - 0,412|}{0,412}100\% = 0,9\%.$$

ЧАСТЬ III. ЛАБОРАТОРНЫЙ ПРАКТИКУМ

14. ЛАБОРАТОРНАЯ РАБОТА №1. СЛУЧАЙНЫЕ ВЕЛИЧИНЫ, ИХ СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ

14.1. Цель работы

Научиться формировать в *MS Excel* последовательности случайных чисел с заданными статистическими характеристиками. Это необходимо для моделирования экономических объектов, имеющих случайную природу, и последующего эконометрического анализа поведения таких объектов.

14.2. Генерация случайных чисел

Последовательности случайных чисел формируются двумя способами:

1. вызов диалогового окна "Генерация случайных чисел".
2. непосредственное использование функций СЛЧИС(), СЛУЧМЕЖДУ().

Рассмотрим оба этих способа.

14.2.1. Диалоговое окно "Генерация случайных чисел"

Это окно вызывается последовательностью
Сервис/Анализ данных/Генерация случайных чисел
и выглядит следующим образом:

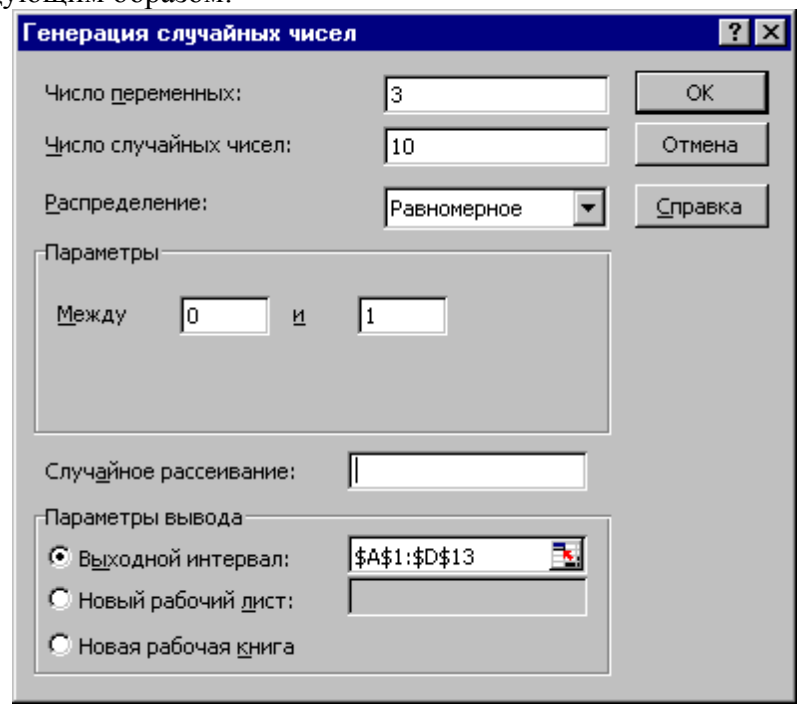


Рис. 1. Диалоговое окно "Генерация случайных чисел"

Число переменных

Здесь следует ввести количество последовательностей случайных чисел (последовательность представляет одну *случайную переменную*). Каждая последовательность располагается в столбце. Если это значение не задать, то каждый столбец в выходном диапазоне (который будет задан в рамке Параметры вывода) будет заполнен последовательностью случайных чисел. Если формируется *одна* последовательность, то следует ввести 1.

Число случайных чисел

Введите число случайных чисел в последовательности. Это число будет определять высоту столбцов выходного диапазона. Если число случайных чисел не будет введено, то все строки выходного диапазона будут заполнены.

Распределение

Выберите распределение, которое необходимо использовать для генерации случайных переменных. Ограничимся только *равномерным* распределением.

Равномерное

Характеризуется верхней и нижней границами интервала. Переменные извлекаются с одной и той же вероятностью для всех значений интервала.

Параметры

Здесь следует задать нижнюю и верхнюю границы интервала. Равномерно распределённые числа принимают значения между границами интервала.

Выходной интервал

Введите ссылку на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные. Может быть задана нижняя правая ячейка интервала.

Новый лист

Установите переключатель, чтобы открыть новый лист в книге и вставить результаты анализа, начиная с ячейки A1. Если в этом есть необходимость, введите имя нового листа в поле, расположенном напротив соответствующего положения переключателя.

Новая книга

Установите переключатель, чтобы открыть новую книгу и вставить результаты анализа в ячейку A1 на первом листе в этой книге.

Пример

После ввода данных, приведённых на рис.1 будут получены следующие результаты

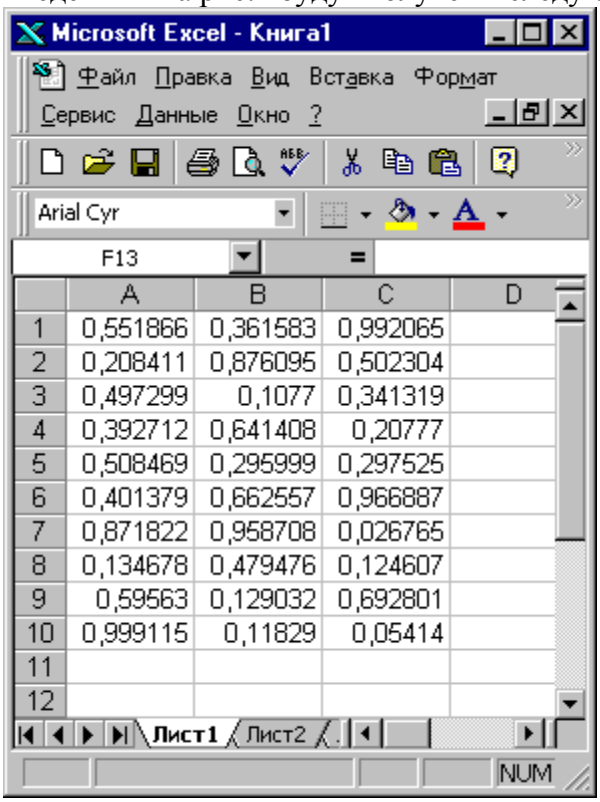


Рис. 2. Последовательности случайных чисел

14.2.2. Использование функций СЛЧИС(), СЛУЧМЕЖДУ()

Функция СЛЧИС() равновероятно формирует вещественное случайное число между 0 и 1. Эта функция не имеет аргументов. Для создания последовательности таких чисел нужно выполнить следующее:

1. ввести в ячейку функцию СЛЧИС(), выделить эту ячейку;
2. подвести курсор к чёрному квадратику в правом нижнем углу ячейки (рис.3), и, когда курсор примет форму чёрного крестика, потянуть его вправо (влево, вверх, вниз) с нажатой левой клавишей мыши (рис.4);

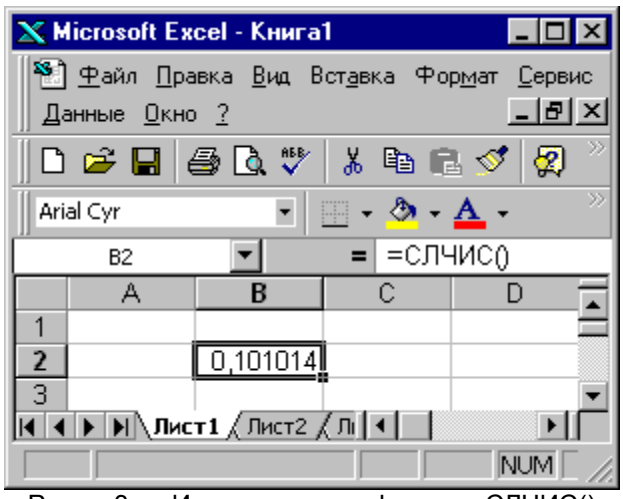


Рис. 3. Использование функции СЛЧИС()

3. потянув чёрный квадрат строки (рис.4) случайных чисел вниз(вверх), можно получить несколько последовательностей случайных чисел.

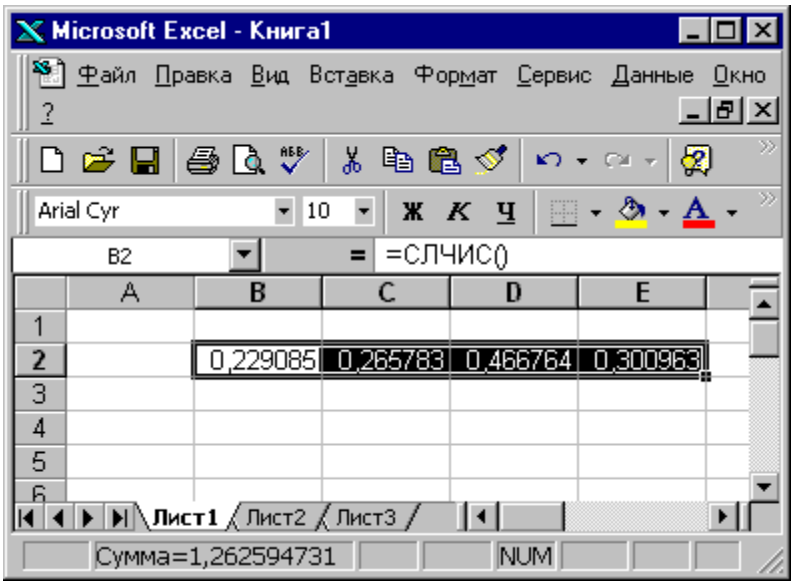


Рис. 4. Случайная последовательность

Функция СЛУЧМЕЖДУ(нижн_граница;верхн_граница) равновероятно возвращает случайное *целое* число между заданными границами.
нижн_граница - это наименьшее *целое* число, которое возвращает функция СЛУЧМЕЖДУ.
верхн_граница - это наибольшее *целое* число, которое возвращает функция СЛУЧМЕЖДУ.
Необходимо, чтобы нижн_граница ≤ верхн_граница.
Каждый раз, когда рабочий лист перевычисляется, возвращается новое случайное число.
Для получения *вещественных* случайных чисел в диапазоне от *0* до *N* следует использовать произведение СЛЧИС()**N*.

14.3. Вычисление среднего значения, дисперсии и стандартного отклонения случайной

Функция СРЗНАЧ(Диапазон) вычисляет среднее значение чисел в указанном диапазоне. Обозначим это значение X.
Функция ДИСП(Диапазон) вычисляет *среднее значение квадрата* отклонения от X.
Функция СТАНДОТКЛОН(Диапазон) вычисляет корень квадратный из *среднего значения квадрата* отклонения от X.

14.4. Выполнение лабораторной работы

1. Разобрать приведённый пример.

2. Используя диалоговое окно "Генерация случайных чисел" и непосредственный ввод в ячейки функций СЛЧИС(), СЛУЧМЕЖДУ(), сформировать случайные последовательности случайных чисел с заданным диапазоном и в заданном месте таблицы.

Вариант	Интервал случайной величины	Количество последовательностей	Количество чисел в последовательности
1.	вещественные (0; 6,5)	3	6
2.	целые [-10;20]	5	7
3.	вещественные (0;16,75)	2	6
4.	целые [-9;2]	4	5
5.	вещественные (0; 5,25)	3	6
6.	вещественные (-10,5;0)	4	5
7.	целые [-8;13]	2	6
8.	вещественные (-5,9;0)	4	5
9.	целые [-7;5]	6	6
10.	вещественные (-25; 0)	5	6

3. Вычислить среднее значение, дисперсию и стандартное отклонение для всех случайных чисел.

14.5. Контрольные вопросы

1. Что такое случайная величина?
2. Что такое непрерывная и дискретная случайные величины?
3. Каковы статистические характеристики случайных величин?

15. ЛАБОРАТОРНАЯ РАБОТА №2. РАЗВЕДОЧНЫЙ АНАЛИЗ ДАННЫХ

15.1. Построение гистограммы и её свойства

Рассмотрим следующие данные

Табл. 1. Среднегодовое количество осадков, мм (Центрально-Черноземный государственный биосферный заповедник, г. Курск)

Год	Кол-во осадков	Год	Кол-во осадков	Год	Кол-во осадков	Год	Кол-во осадков
1947	547	1959	503	1971	495	1983	489
1948	568	1960	688	1972	457	1984	536
1949	558	1961	601	1973	666	1985	655
1950	583	1962	511	1974	517	1986	549
1951	527	1963	494	1975	490	1987	525
1952	734	1964	677	1976	558	1988	678
1953	404	1965	705	1977	718	1989	537
1954	515	1966	561	1978	644	1990	710
1955	458	1967	406	1979	665	1991	672
1956	674	1968	553	1980	681	1992	587
1957	510	1969	551	1981	697	1993	635
1958	602	1970	678	1982	664	1994	602

Этот ряд наблюдений или измерений можно рассматривать как различные значения одной **случайной величины** - «**Среднегодовое количество осадков»**.
Представим значения этой случайной величины следующим образом: диапазон значений от минимального=404 до максимального=734 разобьем на 5 интервалов: [404,470], [470,536], [536,602], [602,668], [668,734]. Ширина интервала **h=66;** частота $\mathbf{n_i}$ – это количество измерений, попадающих в **i-**й интервал; сумма частот, очевидно, всегда равна общему количеству измерений ($\mathbf{n_1+n_2+n_3+n_4+n_5=4+14+12+13+5=48}$).

Относительная частота - отношение частоты к общему количеству измерений $\mathbf{n_i/n}$, т.е. доля попавших в **i-**й интервал измерений среди общего количества измерений; сумма относительных частот всегда равна 1; $\mathbf{(n_1+n_2+n_3+n_4+n_5)/n=1;}$

Подсчитаем частоты и относительные частоты и занесём в табл. 2;

Табл. 2. Распределение измерений по интервалам

№	Интервалы	Количество данных в интервале $\mathbf{n_i}$ (частота)	относительная частота $\mathbf{n_i/n}$ (площадь прямоугольника)	$\mathbf{n_i/(nh)}$ высота прямоугольника
1.	[404,470]	4	0.083	0.0012
2.	[470,536]	14	0.29	0.0044
3.	[536,602]	12	0.25	0.0037
4.	[602,668]	13	0.27	0.0041
5.	[668,734]	5	0.1	0.0015

По данным таблицы строится столбиковая диаграмма - **гистограмма** следующим образом: на оси абсцисс откладываются интервалы измерений; на каждом интервале строится прямоугольник с высотой $\mathbf{n_i/(nh)}$, поэтому площадь каждого прямоугольника равна относительной частоте $\mathbf{n_i/n}$, следовательно, сумма площадей всех прямоугольников гистограммы равна 1.

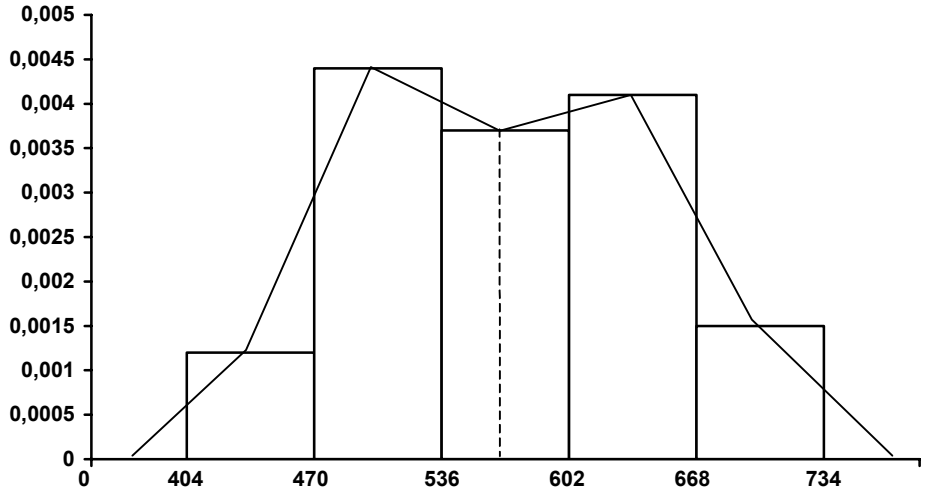


Рис. 1. Гистограмма относительных частот

15.2. Полигон, эмпирическое распределение случайных величин, медиана, мода

Полигон - это кривая, проведённая через середины верхних сторон прямоугольников; площадь, ограниченная **полигоном** также **равна 1**.
Известно, что при больших **n** площадь прямоугольника или относительная частота (которая равна доле попавших в соответствующий интервал измерений) гистограммы приближается к вероятности попадания измерения в соответствующий интервал возможных значений; например, вероятность того, что значение случайной величины **Среднегодовое количество осадков** окажется больше 536 но меньше 602, приблизительно равна 0,25**;** поэтому **полигон** называют **эмпирической плотностью распределения вероятностей** или **эмпирическим распределением случайной величины.**
По гистограмме или полигону можно примерно определить вероятность того, что случайная величина примет значение из некоторого интервала; эта вероятность определяется как часть площади гистограммы, опирающаяся на данный интервал; например, вероятность попадания в интервал [404,536] равна 0,373;
Медиана – это такая величина, которую случайная величина превышает с вероятностью 0,5; таким образом, медиана делит весь диапазон значений на два интервала так, что площади частей гистограммы, опирающихся на эти два интервала, равны по 0,5.
Медиана определяется следующим образом:

$$Me = x_0 + i \cdot \frac{0{,}5\Sigma f_i - S_{Me-1}}{f_{Me}},$$

где x_0 – нижняя граница медианного интервала;
медианным называется первый интервал, накопленная частота которого превышает половину общей суммы частот;
i – величина медианного интервала
S_{Me-1} – накопленная частота интервала, предшествующего медианному;
f_{Me} – частота медианного интервала.
Для приведённого ряда по таблице 3 определяем:
медианным является 3-й интервал;
i = 64;
$S_{Me-1} = 18$;
$\Sigma f_i = 48$;
$f_{Me} = 30$.
Мода – это значение, при котором плотность распределения принимает максимальное значение.
Мода определяется следующим образом:

$$Mo = x_0 + i \cdot \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})},$$

где x_0 – нижняя граница модального интервала;
модальным называется интервал, имеющий наибольшую частоту;
i – величина медианного интервала
f_{Mo} – частота модального интервала;
f_{Mo-1} – частота интервала, предшествующего модальному;
f_{Mo+1} – частота интервала, следующего за модальным.
Для приведённого ряда по таблице 3 определяем:
модальным является 2-й интервал;
i = 64;
$f_{Mo} = 14$;
$f_{Mo-1} = 4$;
$f_{Mo+1} = 12$.
Табл. 3. Накопленные частоты

№	Интервалы	Количество данных в интервале n_i (частота)	Накопленные частоты
1.	[404,470]	4	4
2.	[470,536]	14	18
3.	[536,602]	12	30
4.	[602,668]	13	43
5.	[668,734]	5	48

15.3. Теоретическое распределение случайной величины

Поскольку, при большом количестве измерений **n** площадь прямоугольника гистограммы относительных частот приближается к вероятности попадания измерения в соответствующий интервал измерений, полигон приближается к теоретической плотности распределения вероятностей данной случайной величины.

Во многих случаях теоретическая плотность распределения вероятностей выражается функцией

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

где *a* - среднее значение случайной величины, а σ^2 - её дисперсия. Распределение, определяемое такой функцией, называется **нормальным; график** нормального распределения выглядит следующим образом:

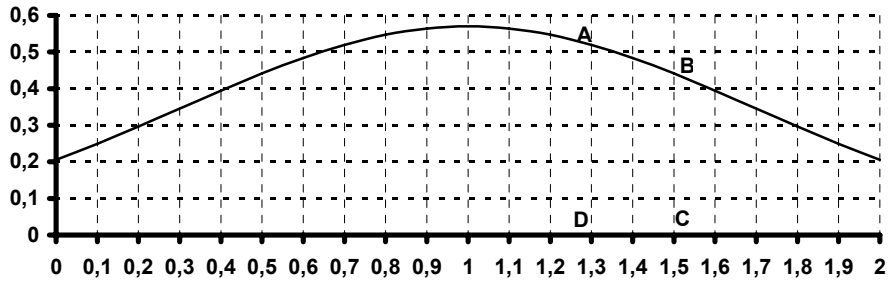


Рис. 2. График нормального распределения с дисперсией 0.49 и средним 1

Вероятность попадания случайной величины в некоторый интервал возможных значений равна площади под графиком функции, опирающейся на этот интервал. Например, вероятность попадания значения случайной величины в интервал **[1.3,1.5]** равна площади криволинейной трапеции **ABCD,** что составляет примерно **0.096.**

15.4. Выполнение лабораторной работы

1. Построить гистограмму по данным своего варианта, разбив весь диапазон на пять интервалов.

2. Анализировать гистограмму:

   а. указать интервал, в который с наибольшей вероятностью попадает значение данного ряда (случайной величины);

b. вычислить среднее значение, моду медиану временного ряда;

c. насколько близка медиана к среднему значению?

3. Вычислить оценки среднего значения a и среднеквадратического отклонения σ, сформировать массив значений функции плотности нормального распределения и построить её график; сравнить с гистограммой;

4. Определить по Рис. 6 и Табл. 2 приблизительно вероятности того, что значение величины «**Среднегодовое количество осадков»** будет принадлежать интервалам [404,536]; [470,602]; [602,734].

15.5. Контрольные вопросы

1. Что такое гистограмма?
2. Что такое функция, плотность распределения случайной величины?
3. Что такое эмпирическая функция, плотность распределения случайной величины?
4. Что такое мода, медиана распределения случайной величины?
5. Что такое нормальное распределения случайной величины?

16. ЛАБОРАТОРНАЯ РАБОТА №3. ЗАВИСИМЫЕ СЛУЧАЙНЫЕ ВЕЛИЧИНЫ

16.1. Цель работы

Предварительный визуальный и корреляционный анализ данных.

16.2. Постановка задачи

Существует зависимость некоторого экономического показателя y от независимого x. Даны массивы показателя x и показателя y. Необходимо провести предварительный разведочный анализ данных.

16.3. Выполнение лабораторной работы

Три набора данных представлены в следующей таблице

1		2		3	
x	y	x	y	x	y
12,0786	113,274	1,43216	144,613	10,3526	99,6214
15,6314	199,546	0,27325	100,151	10,3393	98,4736
28,3647	213,42	2,89057	209,204	9,77984	101,564
41,8782	304,323	9,22231	1017,06	9,98138	95,7555
48,2183	329,273	1,99441	133,718	10,3238	97,1636
64,172	361,869	0,09816	109,011	9,67632	96,2147
74,9673	402,054	2,30822	187,684	10,3477	97,0739
75,8514	386,624	6,30602	553,37	9,93289	104,893
88,3407	464,101	3,51671	235,913	10,1225	100,767
95,6831	468,729	1,96172	124,611	10,2679	100,645
105,703	514,113	3,65577	254,02	10,3265	97,347
116,941	572,861	0,84666	111,463	10,3947	97,3742
127,312	600,06	7,17787	634,812	10,0374	103,819
139,853	659,204	9,95835	1150,23	10,1061	99,8985

Для каждого набора данных

1. Построить точечную диаграмму зависимости y от x.
2. Визуально определить существование тенденции.
3. Для данных с линейной тенденцией вычислить коэффициент корреляции по следующему алгоритму:

$r_{xy} = \frac{S_{xy}}{S_x S_y}$,

S_x, S_y – стандартные отклонения соответственно величин x, y;

$S_{xy} = \frac{n\Sigma xy - \Sigma x \Sigma y}{n(n-1)}$ – выборочная ковариация величин x, y.

Объяснить величину и знак коэффициента корреляции.

4. Для данных, не имеющих тенденции определить центр рассеяния Mx, My и разброс S_x, S_y.

16.4. Контрольные вопросы

a. Какие случайные величины являются независимыми, зависимыми?
b. Какие случайные величины являются коррелированными, некоррелироваными?
c. Что такое коэффициент корреляции?
d. Что означает коэффициент корреляции, близкий к 0, 1, –1?

17. ЛАБОРАТОРНАЯ РАБОТА №4. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

17.1. Линейная парная регрессия

Парной линейной регрессией называется зависимость

$\hat{y} = b_0 + b_1 x$

выборочного условного математического ожидания от переменной x. Термин «парная» означает зависимость *двух* переменных Y, X. Термин «линейная» означает их линейную зависимость.

Условное выборочное математическое ожидание – это выборочное среднее значение величины Y при условии, что переменная X приняла значение x.

Модель наблюдения

$y_i = b_0 + b_1 x + e_i$,

b_0 – оценка свободного члена β_0,

b_1 – оценка углового коэффициента β_1.

17.2. Цель работы

MS Excel обеспечивает эффективную поддержку для проведения регрессионного анализа. Цель работы - освоение инструмента парной линейной регрессии в *MS Excel*.

17.3. Постановка задачи

В некоторой фирме имеются статистические данные (x_i, y_i).

x_i - независимая(объясняющая) переменная - расходы на рекламу продукции фирмы;

y_i - зависимая(объясняемая) переменная - объём продаж, соответствующий расходам x_i.

Следует построить линейную регрессионную модель, объясняющую, как повышение бюджета на рекламу влияет на объём продаж.

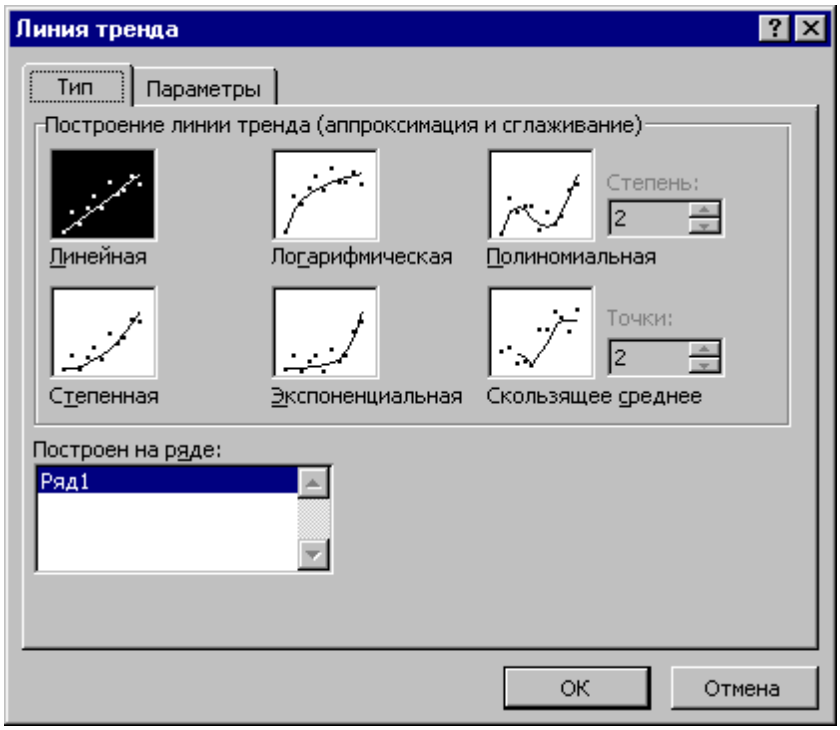
17.4. Выполнение лабораторной работы

Исходные данные приведены в следующей таблице.

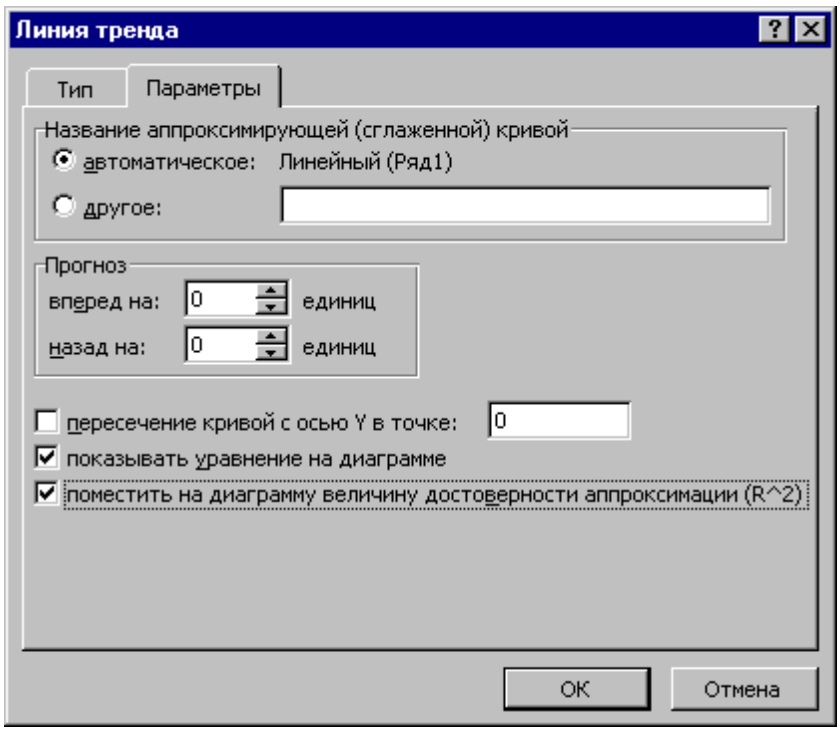
Расходы на рекламу	Объём продаж
48,098754	542,72243
66,48827	720,56027
82,864607	1102,6038
118,45786	1312,86
123,22193	1035,2119
154,35461	1042,4079
170,43262	1177,3656
186,78992	1384,1818
203,46145	1744,9466
220,09022	1607,2908
241,12235	1627,0294
263,92253	1754,7903

Последовательность выполнения работы состоит в следующем.

1. Ввести исходные данные.
2. Построить точеченую даиаграмму.
3. Выделить диаграмму. Выбрать команду Диаграмма/ Добавить линию тренда.
4. Выбрать тип аппроксимации Линейная.



5. Щёлкнуть на вкладке Параметры и установить флажки как на следующем рисунке.



6. Отредактировать положение и величину шрифта уравнения регрессии.
7. Ответить на контрольные вопросы.

17.5. Контрольные вопросы

1. Что такое условное математическое ожидание $M_x(Y)$?
2. Что такое корреляционная и регрессионная зависимости Y от X?
3. Что такое модельное уравнение регрессии?
4. Что такое спецификация модели регрессии, обясняемая и объясняющая переменные, параметры модели?
5. Почему невозможно получить модельное уравнение регрессии?
6. Что такое выборочное уравнение регрессии?
7. Что такое выборочное условное среднее?
8. Каковы задачи регрессионного анализа?
9. Какие модели наблюдения соответствуют модельному и выборочному уравнению регрессии?
10. Что такое парная линейная регрессия, для чего она используется?

18. ЛАБОРАТОРНАЯ РАБОТА №5. МЕТОДЫ ВЫЧИСЛЕНИЯ ПАРАМЕТРОВ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕСИИ

18.1. Диалоговое окно «Линия тренда»

Установка флажка на опции «показывать уравнение на диаграмме» приводит к появлению уравнения регрессии на диаграмме.
Этот метод плох тем, что для дальнейших вычислений, в которых необходимы параметры регрессии, приходится вводить их значения «вручную».

18.2. Расчёт по формулам нормальных уравнений

Формулы для расчёта параметров парной линейной регрессии таковы

$b_0 = \frac{\Sigma y \Sigma x^2 - \Sigma x \Sigma xy}{n\Sigma x^2 - (\Sigma x)^2}$ – свободный член регрессии;

$b_1 = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$ – угловой коэффициент регрессии

где суммирование ведётся по всем выборочным данным.

Расчёт коэффициентов в Excel производится с помощью следующей таблицы

№	X	Y	X^2	XY
1	1,1	1,6	1,21	1,76
2	1,2	1,4	1,44	1,68
3	1,3	1,3	1,69	1,69
4	1,4	1,2	1,96	1,68
5	1,5	1,1	2,25	1,65
Итого	6,5	6,6	8,55	8,46

В столбцы ”X”, “Y” данные вводятся непосредственно. В столбцы “X^2”, “XY” вводятся формулы с соответствующими ссылками на ячейки, содержащие значения X, Y. Затем эти формулы «растягиваются» по столбцу. В последней строке производится суммирование. Далее в некоторые две ячейки вводятся приведённые формулы со ссылками на соответствующие ячейки последней строки таблицы.

18.3. Использование функции «линейн»

Формат функции **линейн**:

линейн(изв_знач_у;изв_знач_х;константа;стат).

Смысл аргументов функции

изв_знач_у – диапазон значений y;

изв_знач_х – диапазон значений x;

константа – устанавливается на 0, если заранее известно, что свободный член равен 0 и на 1 в противном случае;

стат – устанавливается на 0, если не нужен вывод дополнительных сведений регрессионного анализа и на 1 в противном случае.

Использование функции **линейн:**

1. Выделить область пустых ячеек 5×2 (5 строк, 2 столбца) для вывода результатов регрессионной статистики и 1×2 для вывода только коэффициентов b_0, b_1.

2. Ввести функцию **линейн** вручную или через мастера.

3. После корректного ввода функции в левой верхней ячейке выделенной таблицы появится первый итоговый элемент таблицы. Чтобы раскрыть всю таблицу, следует а) нажать клавишу F2, б) затем комбинацию клавиш <CTRL>+<SHIFT>+<ENTER>. Далее появляется регрессионная статистика в следующем порядке

Значение коэффициента b_1	Значение коэффициента b_0
Среднеквадратическое отклонение b_1	Среднеквадратическое отклонение b_0
Коэффициент детерминации R^2	Среднеквадратическое отклонение y
F-статистика	Число степеней свободы

Регрессионная сумма квадратов	Остаточная сумма квадратов

18.4. Выполнение работы

Даны три массива данных x-y. Обработку каждого массива выполнять на отдельном листе.

1		2		3	
x	y	x	y	x	y
12,0786	113,274	11,789	1,07754	6,24128	355,318
15,6314	199,546	24,8361	1,41839	32,5254	689,659
28,3647	213,42	41,5815	1,96181	45,9473	772,645
41,8782	304,323	56,4389	0,72576	72,93	958,635
48,2183	329,273	69,4547	1,84446	90,0468	1263,89
64,172	361,869	85,4163	1,64152	107,361	1541,86
74,9673	402,054	99,4356	1,72272	134,278	1878,93
75,8514	386,624	112,756	2,41526	146,893	1554,35
88,3407	464,101	127,754	2,68287	170,005	1935,97
95,6831	468,729	146,862	3,93581	188,464	1750,06
105,703	514,113	159,52	3,9993	210,064	2257,31
116,941	572,861	174,48	3,99015	229,954	2601,53
127,312	600,06	192,05	4,97495	252,148	3231,12
139,853	659,204	203,777	6,38746	273,043	2639,11

Обработка данных заключается в следующем.

1. Определить параметры линейной регрессии тремя способами.

2. Сформировать массив невязок $e_i = b_1x_i + b_0 – y_i$, сделав абсолютную ссылку на ячейки, содержащие b_1, b_0.

3. Построить диаграмму невязок. Сделать выводы.

4. Исследовать качество модели регрессии как во втором примере раздела 13.2.

19. ЛАБОРАТОРНАЯ РАБОТА №6. АВТОКОРРЕЛЯЦИЯ ОСТАТКОВ. СТАТИСТИКА ДАРБИНА-УОТСОНА

19.1. Цель работы

Объясняемый показатель не всегда линейно зависит от объясняющего. Статистика Дарбина-Уотсона позволяет подтвердить или опровергнуть линейную зависимость показателей. Цель работы - научиться пользоваться статистикой Дарбина-Уотсона.

19.2. Постановка задачи

Имеются статистические данные (x_i, y_i).
После применения обычного МНК получено уравнение линейной регресии

$$y = ax + b.$$

Остатки вычисляются следующим образом:

$$e_i = y_i - (ax_i + b).$$

Следует выяснить, являются ли остатки e_i независимыми. Для этого вычисляется статистика Дарбина-Уотсона

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

19.3. Выполнение лабораторной работы

Два набора данных приведены в следующей таблице

1		2	
X	Y	X	Y
11,788976	1,0775356	12,078558	113,27376
24,836091	1,4183857	15,631428	199,54626
41,581452	1,9618082	28,364705	213,41969
56,438924	0,725762	41,878231	304,32336
69,454714	1,844458	48,218314	329,27322
85,416266	1,6415161	64,171964	361,8687
99,435646	1,7227183	74,967288	402,05448
112,75621	2,4152573	75,851362	386,62402
127,75413	2,6828661	88,340655	464,10133
146,86238	3,9358128	95,683086	468,7287
159,51982	3,9992972	105,70256	514,11295
174,4804	3,9901479	116,94076	572,86126
192,04993	4,9749501	127,31165	600,06004
203,77709	6,3874647	139,85331	659,20398
221,90381	6,8822444	151,50441	671,91049
235,76567	8,6740529	156,8797	682,51971

Обработка каждого набора данных состоит в следующем.

1. Построить точечную даигармму.
2. Выделить диаграмму. Выбрать команду Диаграмма/ Добавить линию тренда.
3. Выбрать тип аппроксимации Линейная.
4. Сформировать массив остатков и массив знаков остатков. Построить точечную диаграмму по этим данным.
5. Вычислить статистику Дарбина-Уотсона через отношение ячеек.
6. По величине статистики Дарбина-Уотсона сделать вывод о зависимости отстаков.

19.4. Контрольные вопросы

1. Дать сравнительные характеристики исходных данных двух разделов, диаграмм остатков и знаков остатков.

2. Как проявляется зависимость остатков относительно линии регрессии?

3. Что такое статистика Дарбина-Уотсона, для чего она предназначена?

4. Как влияет на значение статистики Дарбина-Уотсона зависимость остатков?

5. Как влияет на значение статистики Дарбина-Уотсона независимость остатков?

6. При каких значениях статистики Дарбина-Уотсона можно сделать вывод о зависимости или независимости остатков?

7. Какой вывод следует сделать о характере зависимости между объясняемой и объясняющей переменной, если выяснилось, что остатки зависимы?

20. ЛАБОРАТОРНАЯ РАБОТА №7. НЕЛИНЕЙНЫЕ МОДЕЛИ РЕГРЕССИИ И ИХ ЛИНЕАРИЗАЦИЯ

20.1. Цель работы

Цель работы состоит в освоении способов линеаризации некоторых нелинейных зависимостей. Способы линеаризации описаны в разделе 12.3.

20.2. Постановка задачи

Задана нелинейная спецификация модели
$y = f(x,a,b,\varepsilon)$,
Требуется вычислить искомые оценки параметров *a, b* исходной регрессии.

20.3. Выполнение лабораторной работы

В следующих двух таблицах приведены массивы данных и соответствующие им нелинейные спецификации.

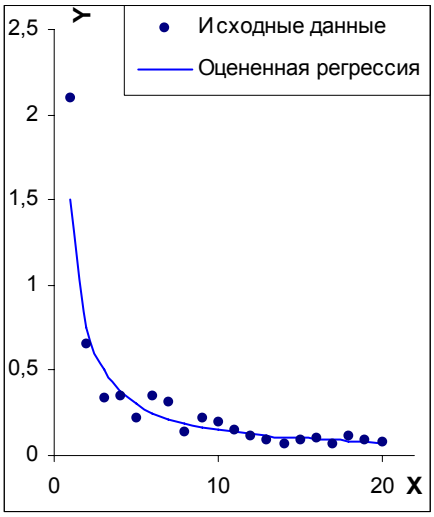
1		2		3		4		5	
$y=b+\frac{a}{x}+\varepsilon$, a>0		$y=b+\frac{a}{x}+\varepsilon$, a<0		$y=\frac{1}{b+ax+\varepsilon}$		$y=\frac{x}{b+ax+x\varepsilon}$, b>0		$y=\frac{x}{b+ax+x\varepsilon}$, b<0	
X	Y	X	Y	X	Y	X	Y	X	Y
0,916	2,343	0,875	-1,26	0,855	0,449	1,001	0,594	0,704	0,64
1,968	1,254	1,715	-0,15	1,927	0,305	2,001	0,734	0,899	0,541
2,928	0,841	3,216	0,424	3,222	0,163	3,001	0,83	1,099	0,468
4,076	0,712	4,291	0,488	4,047	0,133	4	0,825	1,297	0,459
5,079	0,585	4,572	0,522	5,197	0,107	5,002	0,855	1,495	0,421
5,922	0,562	5,526	0,663	6,01	0,119	6,003	0,906	1,698	0,404
7,034	0,556	7,363	0,761	7,089	0,114	6,996	0,915	1,897	0,413
8,034	0,439	8,162	0,786	8,008	0,119	8,003	0,919	2,103	0,382
9,022	0,384	9,217	0,734	8,865	0,098	8,999	0,967	2,3	0,392
9,94	0,357	9,546	0,806	10,19	0,101	9,997	0,978	2,504	0,375
11,05	0,365	11,22	0,842	10,78	0,072	11,00	0,93	2,699	0,375
12,05	0,404	12,15	0,814	11,99	0,064	12	0,988	2,905	0,378
13,09	0,354	13,13	0,87	13,17	0,058	13	0,937	3,1	0,371
14,07	0,317	14,24	0,828	14,25	0,064	14	0,966	3,3	0,367
15,08	0,306	15,36	0,825	15,22	0,054	15	0,994	3,496	0,354
15,92	0,251	16,12	0,881	16,11	0,051	16	0,969	3,698	0,384
16,94	0,375	16,93	0,911	16,86	0,05	17	0,991	3,904	0,354
17,97	0,34	18,33	0,894	17,87	0,049	18	1,007	4,103	0,346
19,09	0,347	18,9	0,888	19,22	0,053	19	0,94	4,302	0,343
20,07	0,372	20,29	0,868	20,08	0,05	20	0,983	4,499	0,366

6		7		8		9		10	
$y=be^{ax+\varepsilon}$, a>0		$y=be^{ax+\varepsilon}$, a<0		$y=be^{a/x+e}$		$y=\frac{1}{b+ae^{-x}+\varepsilon}$		$y=bx^{a}e^{\varepsilon}$	
X	Y	X	Y	X	Y	X	Y	X	Y
11,64	4,198	0,537	2,477	1,038	9,114	-0	0,167	0,01	0,141
21,66	5,322	2,163	2,148	1,49	9,368	0,443	0,25	0,46	1,161
29,19	4,401	5,639	1,485	2,03	9,542	0,877	0,332	0,91	1,618
40,39	5,101	6,615	0,721	2,515	9,641	1,333	0,395	1,36	1,833
50,72	7,306	6,023	0,981	2,972	9,638	1,825	0,592	1,81	1,792
61,01	7,358	8,331	0,848	3,512	9,732	2,25	0,725	2,26	2,117

67,67	7,862	11,94	0,699	4,022	9,723	2,656	0,703	2,71	3,008
77,61	9,336	13,42	0,44	4,513	9,793	3,1	0,837	3,16	2,252
89,6	9,166	14,97	0,417	5,037	9,848	3,648	0,727	3,61	2,586
101,1	10,59	18,86	0,3	5,458	9,836	4,037	0,906	4,06	3,522
108,6	12,63	17,88	0,243	6,009	9,847	4,531	0,904	4,51	3,298
122,3	13,38	20,2	0,215	6,489	9,861	4,919	0,978	4,96	3,249
128,8	16,49	23,73	0,168	7,028	9,896	5,438	0,995	5,41	4,21
139,3	14,29	28,31	0,101	7,548	9,882	5,832	1,007	5,86	4,315
152,4	16,39	27,87	0,11	8,027	9,84	6,254	0,882	6,31	3,979
158,5	16,5	28,82	0,151	8,544	9,932	6,744	1,057	6,76	3,362
168,6	26,37	31,89	0,06	9,045	9,878	7,153	1,251	7,21	4,675
177,9	28,2	33,17	0,093	9,548	9,944	7,637	0,947	7,66	3,733
190	29,45	36,9	0,04	9,987	9,864	8,087	1,113	8,11	5,159
199,1	34,01	38,04	0,041	10,53	9,939	8,502	0,972	8,56	4,023

Последовательность выполнения работы состоит в следующем.

1. Сформировать точеченую диаграмму.
2. Модифицировать данные в соответствии с таблицей, приведённой в разделе 12.3.
3. Построить диаграмму, содержащую облако рассеяния преобразованных данных. Добавить линию тренда.
4. Сформировать массив остатков, их знаков, построить диграмму, содержащую эти данные, вычислить статистику Дарбина-Уотсона.
5. Найти оценки модифицированной линейной регрессии *a*, b**.
6. Вычислить оценки параметров *a, b* исходной регрессии в соответствии с таблицей.
7. Сформировать массив значений нелинейной спецификации с полученными оценками *a, b* и поместить этот массив на диаграмму с исходными данными как на следующем рисунке



8. Сформировать ряд остатков от исходной регрессии и их знаков, вычислить статистику Дарбина-Уотсона.
9. Построить диаграмму остатков и их знаков.

20.4. Контрольные вопросы

1. Что такое линеаризация нелинейной регрессии?

2. Объяснить способы линеаризации, приведённые в таблице.

3. Прокомментировать облако рассеяния, полученное для преобразованных данных и статистику Дарбина-Уотсона для остатков.

4. Прокомментировать взаимное расположение облака рассеяния исходных данных и графика исходной регрессии.

21. ЛАБОРАТОРНАЯ РАБОТА №8. ВЗВЕШЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ

21.1. Цель работы

Освоение применения взвешенного метода наименьших квадратов для коррекции гетероскедастичности остатков

21.2. Постановка задачи

Имеются статистические данные (x_i, y_i), представленные в следующей таблице

X	Y
6,2412802	355,31784
32,525389	689,65859
45,947303	772,64484
72,930043	958,63505
90,046767	1263,894
107,36086	1541,8603
134,27771	1878,934
146,89288	1554,3546
170,0048	1935,9682
188,46383	1750,0638
210,06379	2257,3069
229,95397	2601,5303
252,14755	3231,1214
273,04297	2639,1104
291,62438	2512,6901
305,12028	3108,5159

После применения обычного МНК выясняется гетероскедастичность остатков: стандартное отклонение остатков линейно увеличивается при увеличении независимой переменной. Необходимо применить модификацию взвешенного МНК для коррекции такой гетероскедастичности.

21.3. Выполнение лабораторной работы

1. Построить точечную диаграмму исходных данных, поместить на неё линию тренда, его уравнение и коэффициент детерминации.

2. Сформировать массив остатков.

3. Модифицировать массив независимой переменной следующим образом $x_i^* = 1/x_i$.

4. Модифицировать массив зависимой переменной следующим образом $y_i^* = y_i/x_i$.

5. По полученным данным модифицированной регрессии x_i^*, y_i^* построить диаграмму облака рассеяния, поместить на неё линию тренда с уравнением линии

регрессии и коэффициентом детерминации. Сравнить это уравнение с уравнением исходной регрессии.

21.4. Контрольные вопросы

1. Что такое гомо- и гетероскедастичность остатков?
2. Почему не следует использовать обычный МНК для данных, обладающих свойством гетероскедастичности?
3. В чём состоит суть взвешенного МНК?
4. Стандартные отклонения остатков увеличиваются линейно при увеличении независимой переменной. Как следует модифицировать исходное уравнение регрессии для достижения гомоскедастичности данных?
5. Сформулировать последовательность применения взвешенного МНК для случая гетероскедастичности, сформулированного в п. 4.
6. Каково соотношение оценок углового коэффициента и свободного члена исходной и модифицированной регрессии?

22. ЛАБОРАТОРНАЯ РАБОТА №9. ПРОВЕРКА ГИПОТЕЗЫ О НАЛИЧИИ ТРЕНДА ВО ВРЕМЕННОМ РЯДЕ

22.1. Цель работы

Проверка гипотезы стационарности временного ряда - начальный этап сглаживания. Цель работы - изучение критерия, основанного на выборочной медиане.

22.2. Постановка задачи

Даны значения временного ряда $x(1), x(2),..., x(n)$. Необходимо определить, имеет ли этот ряд неслучайную компоненту, зависящую от времени - тренд.
Пусть x_{med} - выборочная медиана этого временного ряда. Образуем ряд $z(1), z(2),..., z(n)$ следующим образом:
$z(i) = знак(x(i) - x_{med})$.
Серия - это группа подряд идущих +1 или -1.
Обозначим $\nu(n)$ - количество серий; $\tau(n)$ - длина самой протяжённой серии.
Критерий, основанный на выборочной медиане состоит в следующем:
если выполняются оба неравенства
$\nu(n) > 0,5(n + 2 - 1,96\sqrt{n-1}),$
$\tau(n) < 1,43\ln(n+1),$
тогда с вероятностью, заключённой между 0,9025 и 0,95 делается вывод о неизменности среднего значения ряда и об отсутствии тренда. Если хотя бы одно из неравенств не выполняется, тогда с такой же вероятностью следует сделать вывод о наличии тренда.
В лабораторной работе необходимо проверить наличие тренда у двух временных рядов.

22.3. Выполнение лабораторной работы

В следующей таблице приведены два временных ряда. Посредством критерия выборочной медианы определить наличие тренда во временном ряде.

1	-0,385	0,401	0,207	-0,434	-0,343	0,174	-0,277	-0,142	-0,258	-0,194	0,084	0,488	-0,664	0,177	-0,35	0,453	-0,114	0,566	-0,489
2	1,101	1,056	1,071	1,267	1,161	1,366	1,349	1,389	1,415	1,259	1,126	1,413	1,379	1,3	1,171	1,292	1,365	1,375	1,366

Последовательность выполнения работы состоит в следующем.

1. Сформировать массив моментов времени.
2. Ввести значения времнного ряда.
3. Построить изображение временного ряда.
4. Вычислить выборочную медиану по формуле =медиана(*массив*).
5. Используя критерий, определить наличие тренда во временном ряде.

22.4. Контрольные вопросы

1. Что такое временной ряд?
2. Что такое аддитивная и мультипликативная модель временного ряда?
3. Что такое трендовая, циклическая и сезонная компоненты временного ряда?
4. Что такое стационарный временной ряд?

23. ЛАБОРАТОРНАЯ РАБОТА №10. СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ СТАЦИОНАРНОГО ВРЕМЕННОГО РЯДА

23.1. Цель работы

Изучение методов вычисления статистических характеристик стационарных временных рядов в MS Excel

23.2. Постановка задачи

Необходимо оценить статистические характеристики процесса авторегрессии 1-го порядка $x(t) = \alpha x(t-1) + \delta(t)$.
Для таких рядов корреляционная функция равна
$K(\tau) = \alpha^{\tau}$,
поэтому оценка величины α вычисляется как оценка значения корреляционной функции в точке 1:
$\hat{\alpha} = \hat{K}(1)$.

23.3. Выполнение работы

Для выполнения работы необходимо предварительно сформировать достаточно длинный временной ряд: около 2000 значений.

1. В первую строку, начиная с ячейки A11, ввести заголовки t, x(t), τ, $K(\tau)$ – соответственно моменты времени, значения временного ряда, аргумент корреляционной функции, значение корреляционной функции. Все числовые данные формируются начиная со второй строки.

2. В столбец A ввести массив моментов времени от 1 до 2000.

3. В столбец B ввести массив 2000 значений временного ряда с заданными значениями α и дисперсией помехи δ.

4. В столбец C ввести аргументы корреляционной функции от 0 до 30.

5. В ячейку D2 ввести значение 1 (значение корреляционной функции в нуле).

6. В ячейку D3 ввести формулу

=КОРРЕЛ(B$2:СМЕЩ($B$2001;-C2;0;1;1);B2:B$2001)
и растянуть её до ячейки d32. В ячейке D3 будет получена оценка параметра α.

7. Построить диаграммы для двухсот значений временного ряда и 30 значений корреляционной функции.

8. Вычислить среднее значение, стандартное отклонение и дисперсию временного ряда.

23.4. Контрольные вопросы

1. Что такое модель авторегрессии временного ряда?
2. Что такое корреляционная функция временного ряда?

24. ЛАБОРАТОРНАЯ РАБОТА №1. СИСТЕМЫ ОДНОВРЕМЕННЫХ УРАВНЕНИЙ

24.1. Цель работы

Изучение косвенного метода наименьших квадратов для идентификации системы одновременных уравнений.

24.2. Постановка задачи

Задана модель одновременных уравнений

$$\begin{cases} y_1 = b_{12}y_2 + a_{12}x_2 + \varepsilon_1; \\ y_2 = b_{21}y_1 + a_{21}x_1 + \varepsilon_2. \end{cases}$$

Исходные данные приведены в следующей таблице

№	y_1	y_2	x_1	x
1	2,914	-0,12	0,1	7,452
2	0,925	0,098	0,2	2,196
3	1,976	0,108	0,3	4,83
4	5,386	-0,05	0,4	13,56
5	2,683	0,209	0,5	6,451
6	4,779	0,155	0,6	11,79
7	5,413	0,186	0,7	13,35
8	3,973	0,353	0,8	9,497
9	5,015	0,367	0,9	12,1
10	7,807	0,262	1	19,25
11	6,371	0,424	1,1	15,41

1) Определить степень идентифицируемости каждого уравнения.
2) Составить приведённую форму.
3) Найти оценки приведённых параметров модели.
4) Найти оценки структурных параметров модели.

24.3. Контрольные вопросы

1) Что такое система одновременных уравнений?
2) Что такое эндогенные, экзогенные, предопределённые, лаговые переменные?
3) Что такое структурная и приведённая форма системы одновременных уравнений?

4) Что такое идентифицируемость уравнения, каковы необходимые и достаточные условия идентифицируемости?

5) Что такое сверхидентифицируемость?

6) Что такое косвенный и двухшаговый метод наименьших квадратов?

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Тинтнер Г. Введение в эконометрию. М., «Статистика», 1965

2. Чуев Ю.В., Михайлов Ю.Б., Кузьмин В.И. Прогнозирование количественных характеристик процессов. М., «Сов. радио», 1975.- 400 с.

3. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике - М.: Высшая школа, 1979.

4. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Основы моделирования и первичная обработка данных. Справочное изд.- М.: Финансы и статистика, 1983.-471 с.

5. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: Исследование зависимостей. Справочное изд. Под ред. С.А. Айвазяна.- М.: Финансы и статистика, 1985.-487 с.

6. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: Учебник.- М.: Финансы и статистика, 1998.- 352с.

7. Колемаев В.А., Калинина В.Н. Теория вероятностей и математическая статистика:Учебник-М:ИНФРА-М, 2000.

8. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: Дело, 2000.

9. Бородич С.А. Эконометрика: Учеб. пообие Мн.: Новое знание, 2001.- 408с.

10. Айвазян С.А. Основы эконометрики. М.: ЮНИТИ, 2001.

11. Практикум по эконометрике. – Под ред. И.И.Елисеевой. – М.: «Финансы и статистика», 2001.

12. Эконометрика. – Под ред. И.И.Елисеевой. – М.: «Финансы и статистика», 2002.

13. Н.Ш.Кремер, Б.А.Путко. Эконометрика. – М.: ЮНИТИ–ДАНА, 2002.

Учебное издание

Е.Г.Жиляков Ю.М.Перлов
Е.П.Ревтова

ОСНОВЫ ЭКОНОМЕТРИЧЕСКОГО АНАЛИЗА ДАННЫХ

Учебное пособие

В авторской редакции