Григоров М. Ю.
Кульминский Д. Д.

Методы статистического анализа и прогнозирования данных

Учебно-методическое пособие

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Государственное образовательное учреждение
высшего профессионального образования
«Санкт-Петербургский государственный
морской технический университет»

М. Ю. Григоров, Д. Д. Кульминский

МЕТОДЫ СТАТИСТИЧЕСКОГО АНАЛИЗА И ПРОГНОЗИРОВАНИЯ ДАННЫХ

Учебно-методическое пособие

Электронное издание
локального распространения

Санкт-Петербург
Наукоемкие технологии
2026

ISBN 978-5-00271-141-3

УДК 519.22/.25(075.8)
ББК 22.172
Г83

Рецензенты:
В. А. Ненашев, доктор технических наук;
Ю. В. Якубович, кандидат физико-математических наук

Г83 **Григоров М. Ю., Кульминский Д. Д.**
Методы статистического анализа и прогнозирования данных [Электронный ресурс]: учебно-методическое пособие / М. Ю. Григоров, Д. Д. Кульминский. – Электрон, текстовые дан. (2,3 Мб). – СПб.: Наукоемкие технологии, 2026. – 119 с. – 1 электрон., опт. диск (CD-ROM).

ISBN 978-5-00271-141-3

Учебно-методическое пособие посвящено практическому применению статистических методов для анализа, моделирования и прогнозирования данных. Рассматриваются регрессионные модели, системы одновременных уравнений, модели временных рядов, критерии согласия, а также методы оценивания параметров (МНК, ММП). Приводятся алгоритмы решения задач в программных средах MS Excel, MATLAB и STATISTICA.

Учебно-методическое пособие ориентировано на студентов всех форм обучения, проходящих подготовку по направлениям 09.03.01, 17.03.01 и 15.04.06.

Минимальные системные требования:

- процессор: Intel x86, x64, AMD x86, x64 не менее 1 ГГц;
- оперативная память RAM ОЗУ: не менее 512 МБайт;
- свободное место на жестком диске (HDD): не менее 120 МБайт;
- операционная система: Windows XP и выше;
- Adobe Acrobat Reader;
- дисковод CD-ROM;
- мышь.

ISBN 978-5-00271-141-3

Учебное издание

Григоров Максим Юрьевич
Кульминский Данил Дмитриевич

Методы статистического анализа и прогнозирования данных

Учебно-методическое пособие

Электронное издание
локального распространения

Издается в авторской редакции
Верстка *К. В. Михайлов*
Главный редактор *В. М. Коровин*

Издательство «Наукоемкие технологии»
ООО «Корпорация «Интел Групп»
https://publishing.intelgr.com
E-mail: publishing@intelgr.com
Тел.: +7 (812) 945-50-63
Интернет-магазин издательства
https://shop.intelgr.com/

Подписано к использованию 06.06.2026 г.
Объем издания – 2,3 Мб.
Комплектация издания – 1 CD.
Тираж 100 CD.

ISBN 978-5-00271-141-3

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Современный специалист в области информатики, экономики или инженерии ежедневно сталкивается с задачей: есть «сырые» данные – нужно превратить их в обоснованные выводы или прогноз. Будь то анализ продаж, обработка сигналов с датчиков, оценка экономических рисков или тестирование качества продукции – везде требуются одни и те же компетенции: умение выдвинуть гипотезу, применить адекватный статистический метод и корректно интерпретировать результат.

Цель данного учебно-методического пособия – сформировать у обучающегося устойчивые практические навыки статистической обработки данных, построения регрессионных моделей, проверки гипотез и прогнозирования временных рядов с использованием современных программных сред (MS Excel, MATLAB, STATISTICA, R).

В отличие от классических учебников по статистике, здесь мы сознательно смещаем акцент в сторону вычислительной реализации. Теоретические сведения даются ровно в том объеме, который необходим для осознанного выбора метода и понимания получаемых результатов. Основное внимание уделяется пошаговым алгоритмам работы, примерам программного кода и разбору конкретных задач.

Пособие построено как комплекс практических работ. Каждая работа включает:

– краткое, но достаточное теоретическое введение;

– формальную постановку задачи;

– детальное решение в одном или двух пакетах (Excel / MATLAB);

– варианты заданий для самостоятельного выполнения;

– четкие требования к отчету.

После освоения материала студент сможет:

1. Проводить первичную статистическую обработку любых одномерных выборок.

2. Проверять гипотезы о виде распределения (критерии Пирсона, Колмогорова-Смирнова).

3. Моделировать случайные величины с заданным законом распределения.

4. Строить и анализировать модели парной, множественной и нелинейной регрессии.

5. Идентифицировать системы эконометрических уравнений.

6. Оценивать параметры методом максимального правдоподобия.

7. Строить прогнозы на основе моделей временных рядов (ARIMA, SARIMA).

Пособие предназначено для студентов всех форм обучения направлений 09.03.01, 17.03.01 и 15.04.06, а также может быть полезно всем, кто начинает работать с данными в технических и экономических приложениях.

1. ВВОДНЫЕ СВЕДЕНИЯ

1.1. ОСНОВНЫЕ РАЗДЕЛЫ

Рассмотрим основные группы моделей в статистическом моделировании:

Регрессионные модели – данные модели представляют собой уравнение, в котором объясняемая переменная выступает в виде функции от объясняющих переменных;

Системы одновременных уравнений – представляют собой набор тождеств и регрессионных уравнений;

Модели временных рядов – представляют собой случайные процессы, которые необходимо анализировать и прогнозировать.

Оценивание подобных моделей включает два основных этапа теоретический и эмпирический.

Предпосылкой теоретического этапа выступает следующее представление. Считается, что определено все множество реализаций показателей, или, на языке выборочного метода статистики, определена генеральная совокупность. Зная или полагая те или иные статистические свойства генеральной совокупности, можно теоретически определить параметры модели.

На эмпирическом этапе исследователь располагает лишь выборочными значениями показателей. На этом этапе можно оценить, но точно нельзя определить значения параметров модели, поскольку они являются случайными величинами. Оценка производится, чтобы получить как можно более точные и статистически достоверные значения неизвестных параметров модели, которые характеризуют генеральную совокупность всех возможных реализаций показателей.

Характеристики генеральной совокупности, как правило, неизвестны, поэтому их оценивают по выборочным данным. Согласно выборочному методу статистики характеристики генеральной совокупности принято называть параметрами, а характеристики выборочной совокупности принято называть оценками. Выборочная оценка дает удовлетворительное приближение для оцениваемого параметра, если она отвечает ряду требований. Эти требования характеризуются такими терминами, как «несмещенность», «эффективность» и «состоятельность».

Оценка называется несмещенной, если ее математическое ожидание равно оцениваемому параметру при любом объеме выборки. В противном случае оценка называется смещенной.

Несмещенная оценка называется эффективной, если она имеет минимальную дисперсию по сравнению с другими выборочными оценками.

Оценка называется состоятельной, если при увеличении объема выборки она стремится к оцениваемому параметру.

Метод наименьших квадратов (МНК) и его различные модификации – является одним из важнейших методов. Регрессионный анализ, основанный на МНК, дает наилучшие результаты из всех возможных, когда выполняются условия теоремы Гаусса-Маркова. При выполнении этих условий регрессионной модель называется классической нормальной регрессионной моделью. В случае, когда условия теоремы Гаусса-Маркова не выполняются обычно применяются методы, которые могут работать при более слабых предположениях. Одним из таких методов является метод максимального правдоподобия.

Рассмотрим подробнее три основных класса моделей.

Регрессионные модели с одним уравнением

В таких моделях зависимая (объясняемая) переменная y представляется в виде функции $f(x_1,\ldots,x_k, \beta_1,\ldots,\beta_n)$, где $x_1,\ldots,x_k$ – независимые (объясняющие) переменные, а $\beta_1,\ldots,\beta_n$ – некоторые параметры. В зависимости от вида функции $f(x_1,\ldots,x_k, \beta_1,\ldots,\beta_n)$ модели делятся на линейные и нелинейные. Тема регрессионных моделей с одним уравнением является основной в такой дисциплине как эконометрика.

Системы одновременных уравнений

Данные модели описываются системами уравнений. Системы могут состоять из тождеств и регрессионных уравнений, в которые могут входить не только объясняющие, но и объясняемые переменные.

Рассмотрим в качестве примера классическую модель спроса и предложения. Пусть Q_t^D – спрос на товар в момент времени t,

Q_t^S – предложение товара в момент времени t, P_t – цена товара в момент времени t, Y_t – доход в момент времени t. В экономической теории составляется следующая модель спроса и предложения:

$$\begin{cases} Q_t^S = \alpha_1 + \alpha_2 P_t + \alpha_3 P_{t-1} + \varepsilon_t, \\ Q_t^D = \beta_1 + \beta_2 P_t + \beta_3 Y_t + \mu_t, \\ Q_t^S = Q_t^D, \end{cases}$$

где первое уравнение системы представляет собой модель предложения, второе уравнение – модель спроса, а третье уравнение представляет собой модель равновесия спроса и предложения.

Модели временных рядов

Данные модели включают вопросы анализа и синтеза временных рядов в целях изучения динамики изучаемых процессов, а также в целях прогнозирования этих процессов. Модели временных рядов предполагают выделение в них таких компонент, как тренд (основная тенденция) T(t), периодическая (осциллятивная, конъюнктурная, сезонная) компонента S(t), а также случайная (стохастическая) компонента ε_t. Модели принято делить на аддитивные, мультипликативные и смешанные.

Пример аддитивной модели:

$$y(t) = T(t) + S(t) + \varepsilon_t.$$

Пример мультипликативной модели:

$$y(t) = T(t)S(t)\varepsilon_t.$$

Пример смешанной модели:

$$y(t) = T(t) + S(t)\varepsilon_t.$$

К моделям временных рядов относят также модели авторегрессии и скользящего среднего (ARIMA), а также ряд других. Их общей чертой является то, что они описывают поведение временного ряда исходя из его предыдущих значений.

1.2. ОСНОВНЫЕ МАТЕМАТИЧЕСКИЕ ПАКЕТЫ

В настоящее время любое исследование, так или иначе связанное с анализом данных затруднительно проводить без использования каких-либо специализированных пакетов математического моделирования, поскольку расчеты довольно трудны для аналитического решения и требуют использования численных методов, а, следовательно, и эффективных алгоритмов. В качестве основных можно выделить следующие:

– MS Excel
– STATISTICA
– R
– MATLAB

В данном подразделе рассмотрим подробнее каждый из них.

Пакет MS Excel

Для построения базовых моделей регрессионного анализа хорошо подходит пакет Excel входящий в Microsoft Office. Однако в данном пакете не заложено большинство тестов или продвинутых моделей, но при необходимости опытный пользователь может прописать все необходимые процедуры сам, используя VBA.

К преимуществам MS Excel следует отнести следующее:

– Простота работы. Не требуется обладать знаниями в программировании, чтобы работать с данным пакетом;

– Стабильность работы. Поскольку MS Excel является платным, в нем не содержится большого количества ошибок в коде, по сравнению с бесплатными программами;

– Распространенность. В настоящее время на подавляющем большинстве персональных компьютеров стоит это программное обеспечение;

– Кроссплатформенность. Кроме Excel для компьютеров под Windows существуют его почти полные аналоги для операционных систем Mac и Linux.

В качестве недостатков MS Excel можно выделить:

– Платное распространение;

– Ограниченность при анализе временных рядов;
– Малое количество встроенных статистических функций;
– Ограниченные возможности визуализации данных.

Пакет STATISTICA

Пакет STATISTICA производителем которого является фирма StatSoft Inc. включает в себя большое количество методов статистического анализа. Интерфейс данного приложения в целом напоминает интерфейс пакета MS Excel. Пакет предоставляет пользователям возможности одномерного и многомерного статистического анализа, а также позволяет проводить обучение нейронных сетей.

К преимуществам пакета STATISTICA следует отнести:
– Большое количество встроенных статистических функций;
– Хорошие возможности визуализации данных;
– Относительную простоту использования.

В качестве недостатков пакета STATISTICA можно выделить:
– Ограниченность при анализе временных рядов;
– Высокая цена.

Статистический пакет R

Статистический пакет R – свободное программное обеспечение для расчета статистических моделей и построения графиков.

Преимущества пакета R следующие.
– Кроссплатформенность;
– Бесплатное распространение;
– Большое количество пользовательских модулей и моделей;
– Легкость в изучении.

К недостаткам пакета R следует отнести следующее:
– По умолчанию пользовательский интерфейс неудобен. Требует установки какой-либо IDE (среды разработки);
– Ошибки в бесплатно распространяемых пользовательских модулях.

Пакет MATLAB

MATLAB является одним из старейших, тщательно проработанных и проверенных временем математическим пакетом. В настоящее время вышел далеко за пределы специализированной системы для матричных расчетов и стал одним из наиболее мощных универсальных математическим пакетом.

Достоинства MATLAB:

– Быстрота вычислений, относительно других распространенных математических пакетов;

– Широкие возможности визуализации данных;

– Большое количество встроенных функций, как статистических, так и эконометрических.

К недостаткам можно отнести:

– Крайне высокую стоимость;

– Требует навыков в программировании;

– Довольно высокий порог вхождения.

2. КОМПЛЕКС ПРАКТИЧЕСКИХ РАБОТ

2.1. СТАТИСТИЧЕСКАЯ ОБРАБОТКА ОДНОМЕРНОЙ ВЫБОРКИ

Цель работы – получение основных навыков обработки одномерной выборки в пакетах MS Excel и MATLAB.

Пакет MS Excel отлично подходит для простых задач вычисления числовых характеристик выборки. Для вычисления выборочных числовых характеристик средствами MS Excel можно использовать встроенные функции категории «Статистические».

Функция СРЗНАЧ возвращает значение выборочного среднего $\overline{x}$, функция ДИСП позволяет получить значение оценки дисперсии S_x^2, а при помощи функции ДИСПР можно получить значение дисперсии $\tilde{S}_x^2$.

Функция СТАНДОТКЛОН вычисляет выборочное среднеквадратическое отклонение S_x, а функция СТАНДОТКЛОНП дает возможность получить значение среднеквадратического отклонения $\tilde{S}_x$. Значение выборочного момента корреляции (ковариацию) $\hat{V}_{XY}$ можно рассчитать, используя функцию КОВАР, а выборочный коэффициент корреляции r_{xy} можно вычислить, обратившись к функции КОРРЕЛ.

В то же время, при вычислении выборочных числовых характеристик в MS Excel можно воспользоваться возможностями пакета анализа. Процедура действий в этом случае, следующая:

1. Открыть меню Сервис и выбрать Анализ данных.
2. Указать необходимую строку в списке Инструменты анализа.
3. Ввести входной и выходной диапазоны ячеек и установить необходимые параметры.

Так, например, для одновременного вычисления выборочного среднего и дисперсии, а также других характеристик выборки, может быть использована процедура «Описательная статистика».

Эта процедура позволяет получить очень полный статистический отчет. Для выполнения процедуры необходимо:

1. Выполнить команду Сервис – Анализ данных, в появившемся списке «Инструменты анализа» выбрать строку «Описательная статистика» и нажать «Ок».
2. В появившемся диалоговом окне указать входной диапазон анализируемых данных.
3. Указать входной диапазон, т.е. указать адрес ячейки на листе.
4. В разделе Группировка установить переключатель в положение «по столбцам».
5. Установить флажок в поле «Итоговая статистика», нажать ОК.

В результате проведенного анализа в указанном выходном диапазоне для каждого столбца данных выводятся следующие статистические характеристики:

1) среднее (выборочное среднее $\bar{x}$),
2) стандартная ошибка (величина $\frac{S_x}{\sqrt{n}}$),
3) медиана (выборочная квантиль второго порядка),
4) мода (наиболее часто повторяющееся выборочное значение),
5) стандартное отклонение (величина S_x),
6) дисперсия выборки (выборочная дисперсия S_x^2),
7) эксцесс (оценка коэффициентов эксцесса),
8) асимметричность (оценка коэффициента асимметрии),
9) интервал (размах выборки $\Delta = x_{max} - x_{min}$),
10) минимум (наименьшее выборочное значение $x_{\min}$),
11) максимум (наибольшее выборочное значение $x_{\max}$),
12) сумма (сумма всех выборочных значений),
13) счет (объем выборки).

Этапы выполнения работы

1. Получение допуска к работе. Необходимо переписать данные своего варианта N (см. приведенные ниже варианты заданий к работе №1, выборка объемом 50).

2. Выполнить аналитически от руки или в электронном виде:
2.1. Построение вариационного и статистического рядов, найти размах выборки.
2.2. Построение таблицы абсолютных и относительных частот группированной выборки, расчет интервалов провести по формуле Стерджеса.
2.3. Построить эмпирическую функцию распределения, гистограмму, полигон частот.
3. Средствами MS Excel и MATLAB найти оценки математического ожидания, дисперсии (смещенной и несмещенной), медианы и моды. Построить графики эмпирической функции распределения, гистограмму и полигон частот.

Решение задачи в пакете MATLAB

Для начала нам необходима выборка, с которой можно работать. В данном примере мы ее сгенерируем сами. Обратите внимание, что у каждого студента выборка уже задана вариантом задания, и ее не нужно будет генерировать.

```
clear all
close all
clc
% Генерация выборки, для дальнейшей работы
% мат. ожидание генерируемой выборки
mu = 0;
% Среднеквадратическое отклонение
sigma = 1;
% Объем выборки
n = 50;
%  Генерация нормально распределенных
случайных чисел
X = normrnd(mu,sigma,n,1);
% Генерация лог-нормально распределенных
% случайных чисел
% Данная выборка является в нашем случае
% входной
x = exp(X);
```

Далее построим вариационный ряд, определим количество интервалов и найдем абсолютную частоту попадания элемента выборки в каждый из интервалов.

```
% Построение вариационного ряда
x = sort(x);
% Поиск минимального и максимального
% элементов выборки
xmax = max(x);
xmin = min(x);

% Определим количество интервалов
% по формуле Стерджесса
b = 3.332;
r = ceil(1+b*log10(n));

% Длина интервала
stp = (xmax-xmin)/r;

% Определяем середины интервалов
centr = [];
centr(1) = xmin+(stp/2);
for i=2:1:r
    centr(i) = centr(i-1)+stp;
end

% Определяем абсолютную частоту
k1 = xmin;
i = 1;
while i<=r
    k2 = 0;
    for j=1:n
        if (x(j)>=k1) & (x(j)<=k1+stp)
            k2 = k2+1;
        end
    end
    freqn(i) = k2;
    k1 = xmin+stp*i;
    i = i+1;
end
```

Рассчитаем числовые характеристики выборки и выведем их на экран, при помощи следующего программного кода:

```
% Числовые характеристики выборки:
% Выборочное среднее
m = mean(x);
% Дисперсия
D = var(x);
% Ср. кв. отклонение
SKO = std(x);
% Мода
moda = mode(x);
% Медиана
med = median(x);
% Коэффициент эксцесса
kurt = kurtosis(x);
% Коэффициент асимметрии
skew = skewness(x);

% Вывод значений
fprintf('Максимальное значение =
%f\n',xmax);
fprintf('Минимальное значение =
%f\n',xmin);
fprintf('Количество интервалов = %f\n',r);
fprintf('Длина одного интервала =
%f\n',r);
fprintf('Выборочное среднее = %f\n',m);
fprintf('Выборочная дисперсия = %f\n',D);
fprintf('Ср. кв. отклонение = %f\n',SKO);
fprintf('Мода = %f\n',moda);
fprintf('Медиана = %f\n',med);
fprintf('Коэффициент эксцесса =
%f\n',kurt);
fprintf('Коэффициент асимметрии =
%f\n',skew);
```

Далее построим полигон частот, гистограмму и эмпирическую функцию распределения, которые показаны на рис. 1–3 соответственно.

```
% Построение полигона частот
figure()
plot(centr,freqn/n,'r-o')
xlabel('Интервалы');
ylabel('Относительная частота')
grid on
% Построение гистограммы
figure()
histogram(x,r)
xlabel('Интервалы');
ylabel('Частота')
grid on
% Построение эмпирической
% функции распределения
figure()
ecdf(x)
% Подпись оси 0X
xlabel('x')
% Подпись оси 0Y
ylabel('F(x)')
% Добавление сетки на график
grid on
```

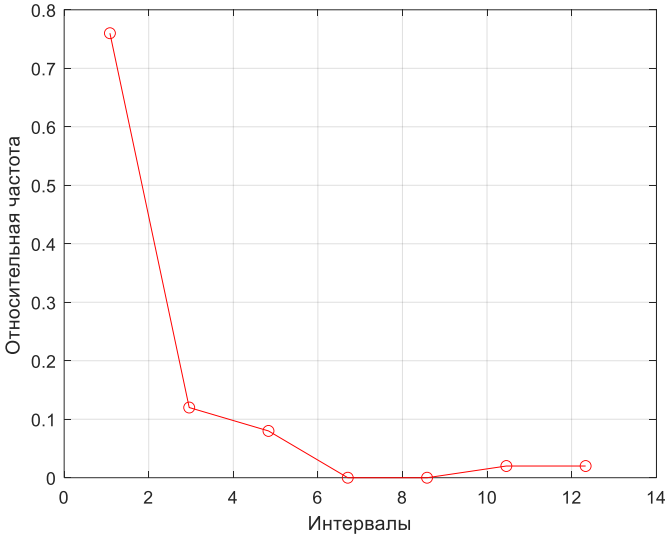


Рис. 1. Полигон частот

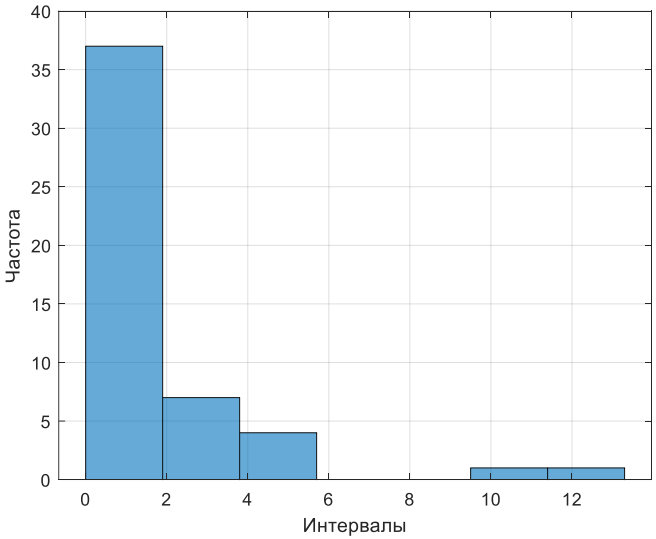


Рис. 2. Гистограмма

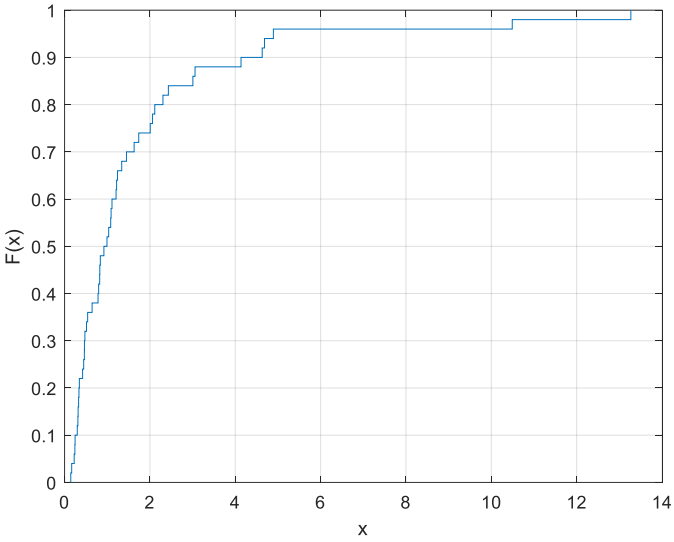


Рис. 3. Эмпирическая функция распределения

Выполнение работы в Excel в данной лабораторной работе мы пропустим, Excel по умолчанию не предоставляет возможности автоматизации процессов поиска интервалов, и выполнение работы в нем напоминает процесс аналитического расчет характеристик.

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. Аналитический расчет необходимых параметров на отдельном листе бумаги.
4. Расчет параметров в пакете MS Excel.
5. Расчет параметров в пакете MATLAB.
6. Графики гистограммы, полигона частот и эмпирической функции распределения.
7. Выводы по проделанной работе.

2.2. КРИТЕРИИ СОГЛАСИЯ

Цель работы – ознакомиться с существующими критериями согласия, получить навыки применения наиболее популярных критериев в современных математических пакетах.

Статистической гипотезой называется определенное предположение о свойствах распределения вероятностей, которым описываются наблюдаемые случайные величины.

Пусть, например, мы имеем выборку $\{x_1, x_2, \ldots, x_n\}$ значений случайной величины X, удовлетворяющей нормальному распределению с математическим ожиданием μ и дисперсией σ^2. Рассматривая выборочные значения x_i, мы можем предположить, например, что $\mu = 0$. Назовем это предположение основной или нулевой гипотезой H_0. Наряду с основной обычно рассматривают альтернативную гипотезу H_1, которая может состоять в отрицании основной ($\mu \neq 0$), но может иметь и другой вид. Если, например, есть веские основания полагать, что $\mu = 1$, это значение принимаем в качестве гипотезы H_1.

Располагая выборочными данными, мы можем сделать правильный выбор между конкурирующими гипотезами H_0 и H_1 только с большей или меньшей вероятностью. Если мы отвергаем основную гипотезу H_0, а она на самом деле верна, то мы совершаем ошибку, которую принято называть ошибкой первого рода. Если же мы принимаем гипотезу H_0, а она не верна, то мы совершаем ошибку второго рода.

Процесс проверки статистической гипотезы приведен на рис. 4.

Вероятность α отвергнуть основную гипотезу в случае, когда она верна, т.е. совершить ошибку 1-го рода, называется уровнем значимости нулевой гипотезы. Обычно уровень значимости полагают равным 0,05 или 0,01. Если нулевая гипотеза не верна, но мы ее принимаем, то совершается ошибка 2-го рода, вероятность которой обозначается β. Число $1 - \beta$ называется мощностью критерия и указывает на вероятность правильного выбора альтернативной гипотезы H_1.

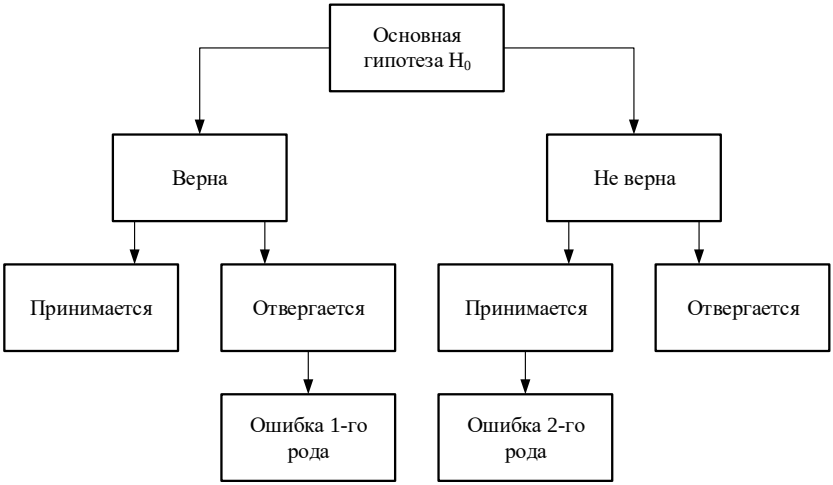


Рис. 4. Процесс проверки статистической гипотезы

Критерии согласия – это критерии проверки гипотез о соответствии эмпирического распределения теоретическому распределению вероятностей.

Существует два основных типа критериев:

1. Основанные на изучении разницы между теоретической плотностью распределения и эмпирической гистограммой.
2. Основанные на расстоянии между теоретической и эмпирической функциями распределения вероятностей.

К первому типу относятся:

– Критерий согласия χ^2;
– Критерий числа пустых интервалов;
– Квартильный критерий Барнетта-Эйсена.

Ко второму типу относятся:

– Критерий Джини;
– Критерий Колмогорова-Смирнова;
– Критерий омега-квадрат;
– Критерий Купера;
– Критерий Ватсона;
– Критерий Реньи и т.д.

В данной работе рассмотрим два наиболее популярных критерия каждого типа, а именно критерий согласия χ^2 (критерий Пирсона) и критерий Колмогорова-Смирнова, начнем с первого.

Во многих практических задачах модель закона распределения заранее не известна и возникает задача выбора модели, согласующейся с результатами наблюдений над случайной величиной. Предположим, что выборка $\{x_1, x_2, \ldots, x_n\}$ произведена из генеральной совокупности с неизвестной функцией распределения (теоретической), относительно которой имеются две гипотезы:

$$H_0 : F(x) = F_0(x),$$
$$H_1 : F(x) \neq F_0(x),$$

где $F_0(x)$ – теоретическая функция распределения.

Критерий χ^2 предполагает, что результаты наблюдений сгруппированы в вариационный ряд. Поскольку при формулировке гипотезы H0 чаще всего необходимо оценивать несколько параметров закона распределения, то последовательность действий, следующая:

1. Сформировать гипотезу о модели закона распределения случайной величины (СВ) и по результатам наблюдений найти оценки неизвестных параметров модели.
2. Подставить в модель закона распределения СВ оценки неизвестных параметров, в результате чего модель станет полностью определенной.

Пусть наблюдаемая СВ X принимает только значения $b_1, b_2, \ldots, b_k$ с неизвестными вероятностями $p_1, p_2, \ldots, p_k$. Основная гипотеза H_0 выделяет среди всех распределений СВ, одно фиксированное распределение, для которого значения вероятностей известны и равны p_i. Обозначим через m_i, число тех элементов выборки $\{x_1, x_2, \ldots, x_k\}$ которые приняли значение b_i. В силу закона больших чисел наблюдаемая частота p_i' с ростом объема выборки будет стремиться к вероятности p_i. Гипотеза принимается если все p_i' мало отличаются от p_i.

Введем статистику

$$\chi^2 = \chi^2\left(x_1, x_2, \ldots, x_n\right) = \sum_{i=1}^{k} \frac{\left(m_i - np_i\right)^2}{np_i}.$$

Эта статистика является мерой равномерной близости $p_i^{|}$ к p_i, m_i – наблюдаемые частоты, np_i – теоретические значения соответствующих частот.

При практической реализации следует следить, чтобы объем выборки был достаточно велик, для этого следует выполнять неравенство $m_i > 5$ при всех k, в противном случае маловероятные значения присоединяют к другим значениям, причем объединенным значениям приписывают суммарную вероятность.

Критерий согласия Колмогорова-Смирнова. В силу теоремы Гливенко-Кантелли эмпирическая функция распределения $F^{|}(x)$ представляет собой состоятельную оценку теоретической функции распределения $F(x)$. Поэтому можно сравнивать $F^{|}(x)$ с гипотетической $F_0(x)$ и, если мера расхождения между ними мала, считать справедливой гипотезу H_0. Наиболее естественная и простая мера – это равномерное расстояние между $F^{|}(x)$ и $F_0(x)$

$$D = \sup_{-\infty<x<+\infty} \left|F^{|}(x) - F_0(x)\right|.$$

заметим, что D случайная величина, потому что ее значения зависят от функции $F^{|}(x)$. Если гипотеза H_0 справедлива и $n \to \infty$, то $F^{|}(x) \to F(x)$ при всяком x.

Рассмотрим статистику

$$D_n = D_n(x_1, x_2, ..., x_n) = \sqrt{n} \sup_{-\infty<x<+\infty} \left|F^{|}(x) - F_0(x)\right|.$$

Критерий Колмогорова указывает принять гипотезу H_0 если $D_n < C$ и отвергнуть ее в противном случае, где C – критическое значение критерия. При $n \to \infty$ значение C совпадает с $(1 - \alpha)$% квантилью распределения Колмогорова.

При практической реализации сначала по выборке составляют вариационный ряд, а затем находят значение статистики D_n. Для этого может использоваться одна из следующих формул:

$$D_n = \sqrt{n} \max_{1\le i\le n} \left[\frac{i}{n} - F_0\left(x_i^*\right), F_0\left(x_i^*\right) - \frac{i-1}{n}\right]$$

или

$$D_n = \sqrt{n}\left[\max_{1\le i\le n}\left|F_0\left(x_i^*\right) - \frac{2i-1}{2n}\right| + \frac{1}{2n}\right],$$

где x_i^* – элемент вариационного ряда. После сравнивают значение статистики с критическим значением для заданного уровня значимости и принимают/отвергают гипотезу.

Этапы выполнения работы

1. Согласовать с преподавателем номер своего варианта.
2. Используя выборку по варианту разложить ее на необходимое количество интервалов.
3. Рассчитать математическое ожидание, среднеквадратическое отклонение и дисперсию выборки.
4. Средствами пакета MS Excel проверить гипотезу H_0 о том, что выборка подчиняется нормальному распределению, с помощью критерия согласия Пирсона.
5. Средствами пакета MATLAB проверить гипотезу H_0 о том, что выборка подчиняется нормальному распределению, с помощью критерия согласия Колмогорова-Смирнова. Построить гистограмму, эмпирическую функцию распределения, а также теоретическую функцию распределения нормального закона на одном графике.

Решение задачи в пакете MS Excel

Пусть имеется выборка объемом 165 элементов поделенная на 10 интервалов, с серединами интервалов и эмпирическими частотами, показанными на рис. 5.

	A	B
1	Середина интервала	Частота эмпирическая
2	-3,2	5
3	-1,6	8
4	0	14
5	1,6	26
6	3,2	33
7	4,8	25
8	6,4	22
9	8	19
10	9,6	7
11	11,2	6

Рис. 5. Параметры выборки

Стоит задача с помощью критерия согласия Пирсона проверить выборку на соответствие нормальному распределению. Для начала рассчитаем необходимые параметры. Количество элементов выборки рассчитывается как сумма эмпирических частот. В нашем случае 165, в пакете MS Excel за это отвечает функция =СУММ(). Среднее значение можно рассчитать используя функцию =СУММПРОИЗВ(A2:A11;B2:B11)/E1. Для возможности вычисления среднеквадратического отклонения (СКО) проведем отдельно вычисление по формуле $\left(x-\bar{x}\right)^2$. Вычислим это в ячейках C2-C11 используя команду =(A2-$E$2)^2*B2, где в ячейке E2 находится среднее значение выборки. Теперь мы можем рассчитать СКО используя формулу =КОРЕНЬ(СУММ(C2:C11)/$E$1), где в ячейке E1 хранится значение объема выборки. И также нам необходимо вычислить значение длины половины интервала, это выполняется аналитически, в нашем случае оно равно 1,6. В результате у нас получится 4 значения, показанные на рис. 6.

D	E
Кол-во элементов выборки=	165
Среднее=	4,033939
СКО=	3,377331
Длина половины интервала=	1,6

Рис. 6. Полученные значения

Далее вычислим теоретические частоты для того, чтобы можно было воспользоваться критерием Пирсона. Они вычисляются с помощью формулы =НОРМ.РАСП(A2:$A$11;$E$2;$E$3;0)*$E$1*$E$4.

Теперь мы можем вычислить значение критерия $\chi^2_{эмп}$. Используя для расчета составляющих формулу =(B2-G2)^2/G2 записанную в ячейку H2. Затем просуммировав ячейки H2-H11 находим искомое значение $\chi^2_{эмп}$. Значение $\chi^2_{теор}$ вычисляем с помощью встроенной функции =ХИ2.ОБР(0,05;7), где 0,05 это

уровень значимости α, а второе число обозначает количество степеней свободы, которое вычисляется по формуле количество интервалов отнять вычисленные параметры (в нашем случае их два – среднее и СКО) и дополнительно отнять единицу. Получили $\chi^2_{теор} = 2{,}167349$, а $\chi^2_{эмп} = 5{,}903442$.

Поскольку $\chi^2_{эмп} > \chi^2_{теор}$ то гипотеза H_0 о нормальном распределении выборки отвергается с уровнем значимости 0,05.

Решение задачи в пакете MATLAB

Пусть имеется выборка, объемом 120 элементов представляющая собой оценки студентов за экзамен по 100 балльной системе. Задача состоит в том, чтобы проверить с помощью критерия согласия Колмогорова-Смирнова гипотезу H_0 о принадлежности выборки к нормальному распределению.

Запишем выборку в MATLAB и построим ее гистограмму, которая показана на рис. 7. Код программы показан далее.

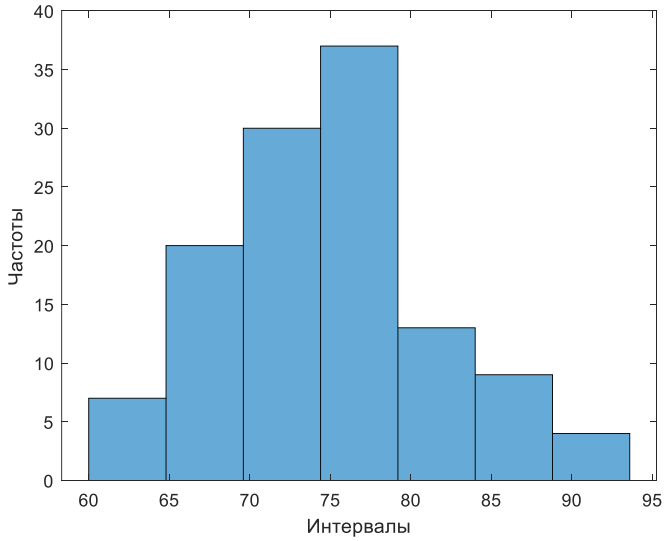


Рис. 7. Гистограмма выборки

```
clear all
close all
clc

x_empirical = [77 74 80 76 77 93 64 83 75
...
    82 70 88 78 77 66 74 87 69 75 68 64 71
...
    68 75 83 80 67 78 84 91 76 73 67 89 67
...
    75 72 69 71 72 72 81 77 71 77 77 67 75
...
    83 74 74 81 67 69 63 73 77 79 86 73 84
...
    71 67 84 71 78 68 68 77 84 70 69 81 81
...
    77 66 76 75 83 77 78 77 74 72 68 75 78
...
    71 74 70 88 70 74 76 71 79 62 77 80 71
...
    72 79 76 77 71 64 74 63 68 78 69 73 79
...
    76 84 89 63 69 79 83];

figure()
histogram(x_empirical,7);
xlabel('Интервалы')
ylabel('Частоты')
```

Далее рассчитаем выборочную дисперсию, выборочное среднее и СКО.

```
sr = mean(x_empirical);
SKO = std(x_empirical);
variance = var(x_empirical);
fprintf('Выборочное среднее = %f\n',sr);
fprintf('СКО = %f\n',SKO);
fprintf('Выборочная дисперсия =
%f\n',variance);
```

Исходя из точечных оценок и гистограммы, принимаем в качестве гипотезы H0 утверждение, что выборка подчиняется нормальному распределению с параметрами математического ожидания 75 и среднеквадратического отклонения 7. Проверим данную гипотезу с помощью критерия согласия Колмогорова-Смирнова.

Для этого выборку сначала необходимо преобразовать к стандартному нормальному распределению.

```
x_empirical= (x_empirical-75)/7;
[h p] = kstest(x_empirical,'alpha',0.01)
```

В нашем случае функция возвращает логический ноль, это значит, что критерий согласия Колмогорова-Смирнова не отклоняет нулевую гипотезу на уровне значимости 0,01.

Построим графики эмпирической и теоретической функции распределения.

```
figure()
cdfplot(x_empirical)
xlabel('x')
ylabel('y')
hold on
x_values                                        =
linspace(min(x_empirical),max(x_empirical
));
plot(x_values,normcdf(x_values,0,1),'r-')
legend('Эмпирическая                  функция
распределения', ...
    'Теоретическая функция распределения',
...
    'Location','best')
```

Результат показан на рис. 8.

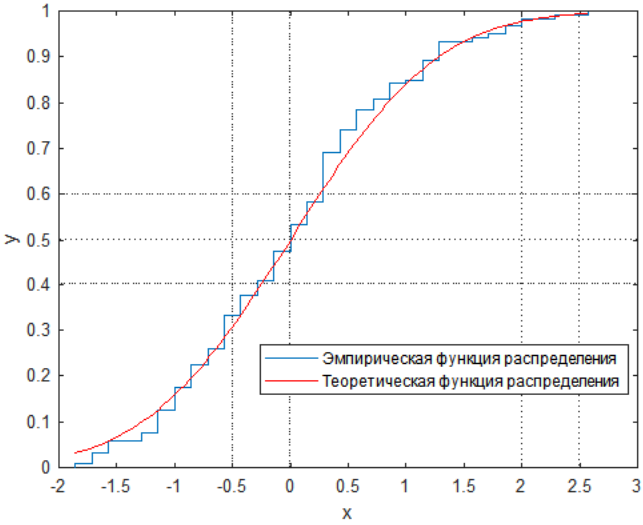


Рис. 8. Эмпирическая и теоретическая функции распределения

Варианты заданий

В качестве исходной выбирается выборка, которая была использована в работе №1, по согласованию с преподавателем, студент может сам предложить интересующую его выборку.

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. Проверка выборки на соответствие нормальному распределению в пакете MS Excel используя критерий согласия Пирсона.
4. Проверка выборки на соответствие нормальному распределению в пакете MATLAB используя критерий согласия Колмогорова-Смирнова.
5. Графики эмпирической и теоретической функции распределения.
6. Выводы по проделанной работе.

2.3. МОДЕЛИРОВАНИЕ СЛУЧАЙНЫХ ВЕЛИЧИН С ЗАДАННЫМ ЗАКОНОМ РАСПРЕДЕЛЕНИЯ

Цель работы – получить навыки генерации случайных величин с заданным законом распределения в современных математических пакетах.

Для начала вспомним несколько основных определений.

Случайная величина (СВ) – величина, которая при испытаниях принимает одно из возможных значений, наперед неизвестно какое. Бывают дискретными и непрерывными.

Законом распределения дискретной случайной величины называют соответствие между возможными значениями случайной величины и вероятностями их появления. Сумма всех вероятностей $\Sigma p_i = 1$. Закон распределения также может быть задан аналитически (формулой) и графически (многоугольником распределения, соединяющим точки (x_i; p_i)).

Функция распределения – функция $F(x)$, характеризующая распределение случайной величины или случайного вектора; вероятность того, что случайная величина X примет значение, меньшее или равное x, где x – произвольное действительное число.

Плотностью распределения вероятностей непрерывной случайной величины X называется функция $f(x)$ – первая производная от функции распределения F(x). Смысл плотности распределения состоит в том, что она показывает, как часто появляется случайная величина X в некоторой окрестности точки x при повторении опытов.

Коэффициент асимметрии – числовая характеризующая степени несимметричности распределения данной случайной величины.

Коэффициент эксцесса – мера остроты пика распределения случайной величины.

В statistics and machine learning toolbox системы MATLAB имеются функции расчета плотностей вероятности и функций распределения для многих известных распределений. Имена функций для расчета плотностей вероятности оканчиваются

буквами pdf (probability density function), а для расчета функций распределения – буквами cdf (cumulative distribution function).

Приведем необходимые для выполнения работы функции:

1. `y=unifpdf(x,a,b)` – расчет значения плотности вероятности в точке x для равномерного распределения на промежутке (a; b).

2. `y=normpdf(x,m,sigma)` – расчет значения плотности вероятности в точке x для нормального распределения, где m – математическое ожидание, sigma – среднее квадратическое отклонение (СКО).

3. `y=exppdf(x,mu)` – расчет значения плотности вероятности в точке x для экспоненциального распределения с параметром mu, равным математическому ожиданию (!) случайной величины.

4. `y=unifcdf(x,a,b)` – расчет значения функции распределения в точке x для равномерного распределения на промежутке (a; b).

5. `y=normcdf(x,m,sigma)` – расчет значения функции распределения в точке x для нормального распределения, где m – математическое ожидание, sigma – СКО.

6. `y=expcdf(x,mu)` – расчет значения функции распределения в точке x для экспоненциального распределения с параметром mu, равным математическому ожиданию (!) случайной величины.

Случайные числа с различными законами распределения обычно моделируются с помощью преобразований одного или нескольких независимых значений базовой случайной величины. Базовая случайная величина α – это случайная величина с распределением $R(0;1)$ (равномерным распределением в интервале (0;1)). Независимые случайные величины с распределением $R(0;1)$ обозначаются символами $\alpha_1, \alpha_2, \ldots, \alpha_n$. В математическом пакете имеется стандартная программа моделирования базовой случайной величины (см. функцию rand).

Рассмотрим алгоритмы моделирования случайных величин, имеющих равномерное, нормальное и экспоненциальное распределение.

Равномерное распределение R(a,b), (a<b)

Если плотность вероятности *f*(x) есть величина постоянная на определенном промежутке [a, b], то закон распределения (ЗР) называется равномерным. На рис. 9 и 10 ниже изображены графики функции распределения вероятностей и плотность вероятности равномерного закона распределения соответственно.

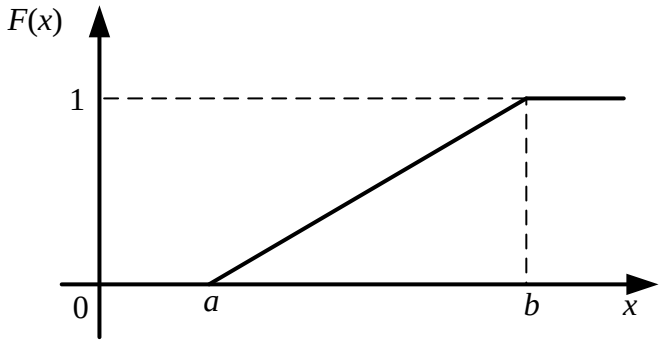


Рис. 9. График функции распределения равномерного закона распределения

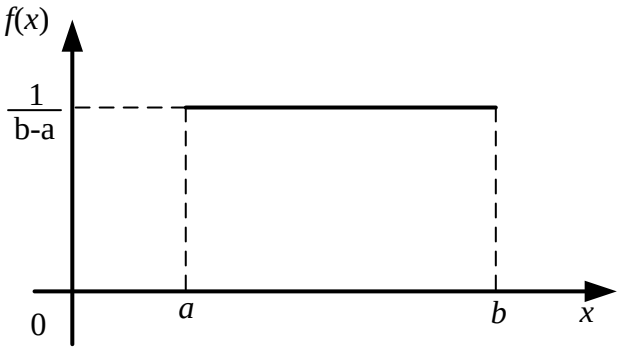


Рис. 10. График плотности вероятности равномерного закона распределения

Аналитическое представление равномерного ЗР имеет вид

$$f(x)=\begin{cases}\dfrac{1}{b-a}, & a\le x\le b\\ 0, & x<a \quad x>b\end{cases}$$

$$F(x)=\begin{cases}0, & x\le a\\ \dfrac{(x-a)}{(b-a)}, & a\le x\le b\\ 1, & x>b\end{cases}$$

Аналитическое выражение для моделирования нормальной СВ имеет вид:

$$X=a+(\mathrm{b}-\mathrm{a})*\alpha,$$

где α – независимые равномерно распределенные случайные величины на интервале [0; 1).

Алгоритм моделирования равномерно распределенной случайной величины, следующий:

1. Задать параметр a = <<номер варианта>>.
2. Вычислить $b = 2a$.
3. Сгенерировать α.
4. Вычислить $X = a + (b - a)*\alpha$.
5. Повторить шаги 1-4 требуемое количество раз.

Нормальное распределение N(m, σ), (σ > 0)

Среди законов распределения СВ наиболее распространённым является нормальный закон распределения. Нормальное распределение задаётся функцией плотности вероятности:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-m)^2}{2\sigma^2}\right)$$

где параметр m – математическое ожидание, а параметр σ – среднеквадратическое отклонение распределения.

График плотности вероятности случайной величины, имеющей нормальный закон распределения, математическое ожидание 0 и среднеквадратическое отклонение 1 показан на рис. 11.

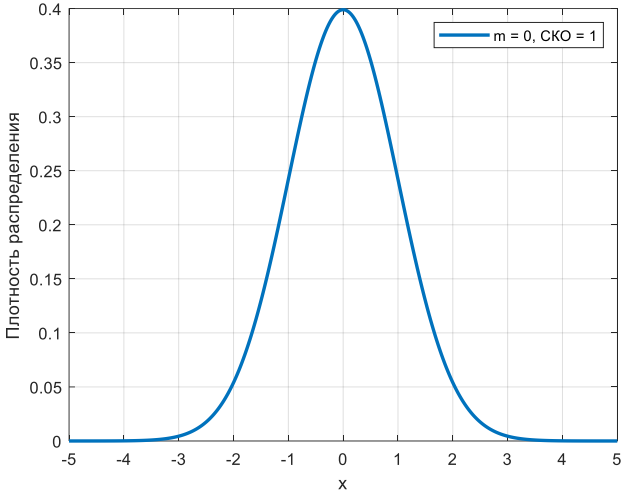


Рис. 11. Плотность вероятности стандартного нормального распределения

Аналитическое выражение для моделирования нормальной СВ имеет вид:

$$Y = m + \sigma\sqrt{-2\ln(\alpha_1)} * \sin(2\pi\alpha_2),$$

где α_1 и α_2 – независимые равномерно распределенные случайные числа на интервале [0;1), m – математическое ожидание СВ, σ – среднеквадратическое отклонение.

Алгоритм моделирования нормальной СВ состоит из следующих шагов:

1. Сгенерировать α_1 и α_2.
2. Зарезервировать константу $c = 2\pi$.
3. Создать переменную $r = \sqrt{-2\ln(\alpha_1)}$.
4. Создать переменную $\varphi = \alpha_2 c$.
5. Получить реализацию нормальной случайной величины $X_i = r_i \sin(\varphi_i)$.
6. Реализовать нормальную СВ с требуемыми параметрами $Y_i = m + \sigma X_i$.
7. Повторить шаги 1–6 требуемое количество раз.

Экспоненциальное распределение E(λ), (λ>0)

Данное распределение моделирует время между двумя одинаковыми событиями, происходящими друг за другом. Получило широкое распространение в теории массового обслуживания.

Аналитическая запись функции распределения экспоненциального закона распределения выглядит следующим образом

$$f(x)=\begin{cases}\lambda\exp(-\lambda x), & x\geq 0;\\ 0, & x<0,\end{cases}$$

а плотность распределения вероятности задается следующим аналитическим выражением

$$F(x)=\begin{cases}1-\exp(-\lambda x), & x\geq 0;\\ 0, & x<0.\end{cases}$$

В качестве примера приведем графики функции распределения и плотности распределения экспоненциальной СВ с параметром $\lambda = 1$, показанные на рис. 12 и 13 соответственно.

Аналитическое выражение для моделирования экспоненциальной СВ имеет вид:

$$X=-\frac{1}{\lambda}\ln(\alpha),$$

где α равномерно распределенное случайное число на интервале [0;1).

Алгоритм моделирования экспоненциально распределенной СВ:

1. Задать параметр λ.
2. Сгенерировать α.
3. Вычислить $X_i = -(1/\lambda)\ln(\alpha_i)$.
4. Повторить шаги 1–3 требуемое количество раз.

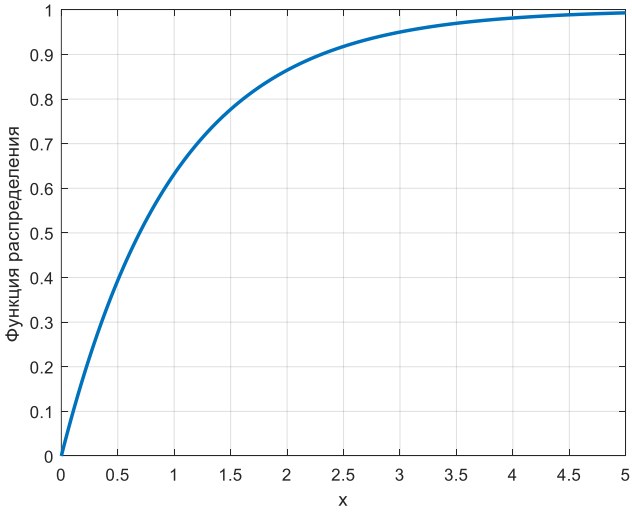


Рис. 12. Функция распределения экспоненциального закона распределения

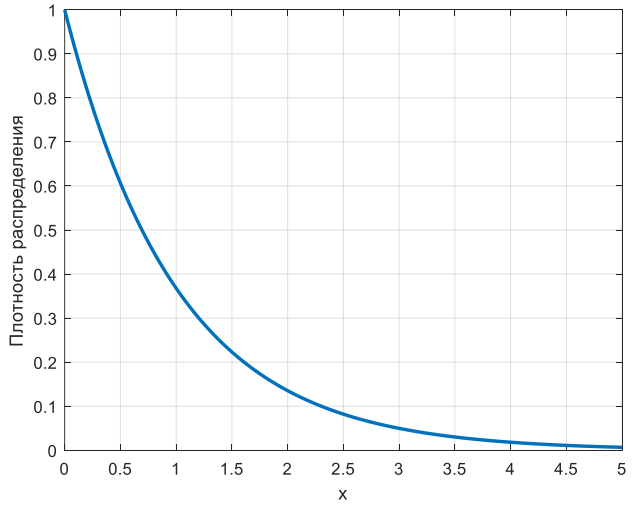


Рис. 13. Плотность распределения экспоненциального закона распределения

Этапы выполнения работы

1. Получить графики функций распределения и плотности распределения в пакетах Excel и MATLAB для следующих трех видов распределений согласно варианту (варианты заданий см. в конце лабораторной работы).
2. Сгенерировать выборку объемом 300 используя встроенную функцию генерации случайных чисел пакетов Excel и MATLAB для каждого из законов распределения – нормального, равномерного и экспоненциального с параметрами, указанными в варианте.
3. Сгенерировать выборку аналогичного объема используя аналитические выражения для каждого из законов распределения.
4. Построить гистограммы полученных выборок. Получить точечные оценки первых четырех моментов выборки.
5. Построить эмпирические функции распределения для каждого распределения. Проверить выборки на соответствие заданному распределению используя критерий согласия Колмогорова-Смирнова.
6. Повторить пункты 2–4 для выборок объемом 700 и 1500 чисел.
7. Построить графики зависимости значений каждого из моментов распределения от объема выборки.

Решение задачи в пакете MS Excel

В качестве примера рассмотрим средства Моделирования СВ, имеющиеся в Excel, при подключении надстройки «Пакет анализа». Здесь имеются генераторы случайных чисел со следующими законами распределения:

1. Равномерное.
2. Нормальное.
3. Дискретное.
4. Бернулли.
5. Биномиальное.
6. Пуассона.
7. Модельное.

Также требуемое распределение можно получить по описанным выше алгоритмам. Используя функцию СЛЧИС. Функция СЛЧИС, генерирующая случайные последовательности, равномерно распределенные на отрезке [0,1]. Ее синтаксис СЛЧИС() и она не имеет аргументов. Выделяется диапазон ячеек, вводится = СЛЧИС() и используется комбинация Ctrl+Shift+Enter.

Равномерное распределение таким образом формируется используя следующий синтаксис =(b-a)*СЛЧИС()+a.

Для генерации других распределений, отсутствующих в Excel, можно использовать, например, метод обратных функций и датчик равномерно распределенных последовательностей. Так, например, методом обратных функций используя алгоритмы моделирования, приведенные в работе можно смоделировать экспоненциальное распределение, отсутствующее в пакете анализа следующим образом:

Используя функцию =LN(СЛЧИС())/A1, записав значение λ в ячейку A1.

После генерации случайных чисел первым делом необходимо построить гистограмму, чтобы визуально оценить полученный результат. Для этого требуется:

1. Выделить мышкой столбцы таблицы, значения которых будут отображены на осях гистограммы.
2. Находясь во вкладке «Вставка» нажимаем по кнопке «Гистограмма», которая расположена на ленте в блоке инструментов «Диаграммы».
3. В открывшемся списке выбрать одну из пяти простых типов диаграмм, в нашем случае – гистограмму.

Решение задачи в пакете MATLAB

Построим функцию плотности вероятности нормального закона f с математическим ожиданием Mx=0 и СКО sigma=1.

```
clear all
close all
clc
Mx=0;                          % мат. ожидание
sigma=1;                       % СКО
```

```
% требуемый диапазон значений СВ
x=Mx-3*sigma:0.1:Mx+3*sigma;
% расчет значений плотности вероятности
f=normpdf(x, Mx, sigma);
% создание нового графического окна для графика
figure()
% команда построение графика
plot(x,f)
xlabel('x')
ylabel('f(x)')
grid on
```

Далее построим функцию плотности распределения нормального закона с такими же параметрами.

```
% расчет значений функции распределения
F=normcdf(x, Mx, sigma);
figure()
plot(x,F)
xlabel('x')
ylabel('F(x)')
```

После выполнения указанных команд мы получим графики, показанные на рис. 14 и 15.

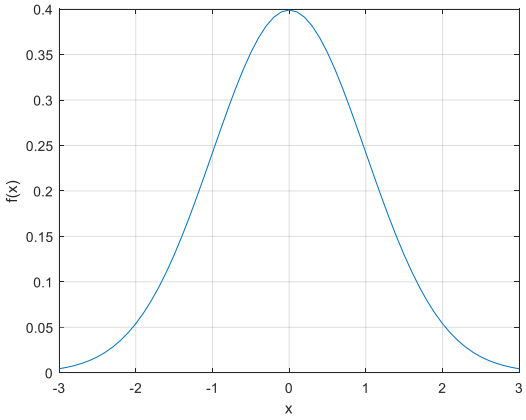


Рис. 14. Результат выполнения первого блока команд

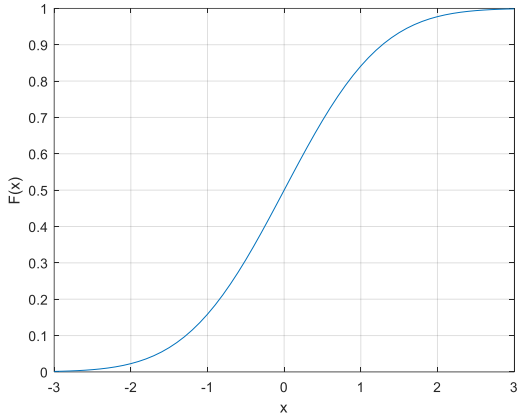


Рис. 15. Результат выполнения второго блока команд

Теперь получим выборку заданной длины псевдослучайной величины, распределенной по нормальному закону с помощью MATLAB. Сгенерируем выборку из N=1000 псевдослучайных чисел, распределенных по нормальному закону с параметрами Mx=0 и sigma=1.

```
N=1000;
n = 1:1:N;
Y=normrnd(Mx, sigma, N, 1);
figure()
plot(n,Y)
grid on
xlabel('Номер элемента в выборке')
ylabel('Значение случайной величины')
```

В результате получим график показанный на рис. 16.

Далее получим реализацию случайной величины, используя аналитическое выражение. Для этого используем следующий код.

```
c=2*pi;
% Генерация первой равномерно распределенной
СВ
a1=rand(1,N);
```

```
% Генерация второй равномерно распределенной
СВ
a2=rand(1,N);
r=sqrt(-2*log(a1));
f=a2*c;
X1=r.*sin(f);
Y1=Mx+sigma*X1;
figure()
plot(Y1)
grid on
xlabel('Номер элемента в выборке')
ylabel('Значение случайной величины')
```

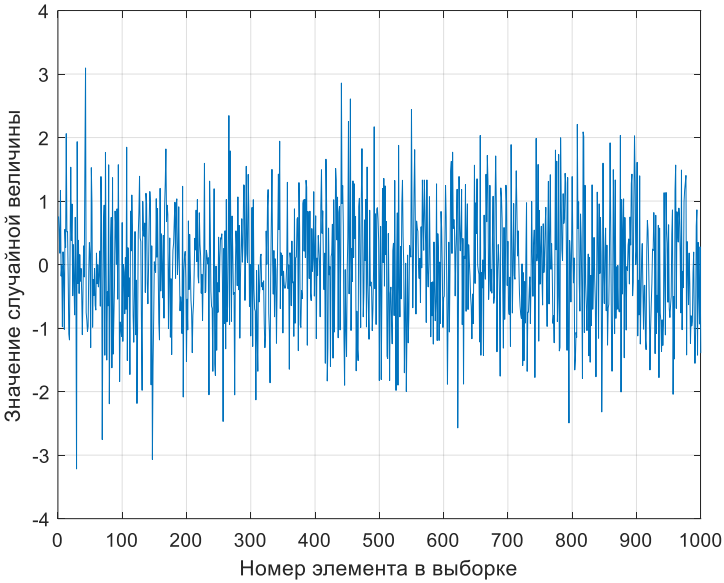


Рис. 16. График значений СВ

Далее необходимо проанализировать полученные результаты. Построим гистограммы выборок, полученных с помощью функции MATLAB и с помощью выражения. Для этого используем следующий код.

```
% Построение гистограммы для СВ
% полученной с помощью функции MATLAB
figure()
bar(YY,p1)
title('Гистограмма СВ полученной с помощью
функции MATLAB')
xlabel('Значение')
ylabel('Относительная частота')
grid on
% Построение гистограммы для СВ
% полученной с помощью выражения
figure()
bar(YY1,p2)
title('Гистограмма СВ полученной с помощью
выражения')
xlabel('Значение')
ylabel('Относительная частота')
grid on
```

Рис. 17 и 18 показывают, что гистограммы случайных величин, полученных функциями MATLAB и аналитическим выражением, а если быть точнее с использованием преобразования Бокса-Мюллера визуально схожи с нормальным распределением.

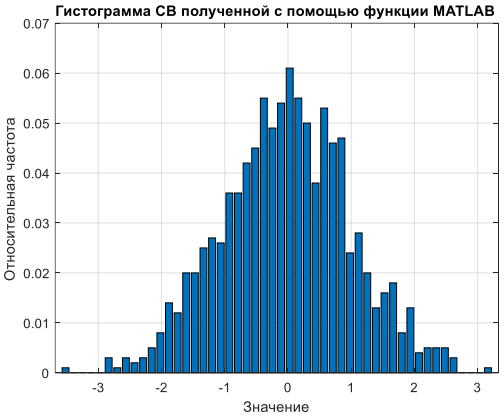


Рис. 17. Гистограмма СВ, полученной с помощью встроенной функции

После визуальной оценки получим выборочные моменты.

```
% Расчет моментов для выборки
% полученной с помощью MATLAB
m_Y = mean (Y) % вычисление мат. ожидания
st_Y = std(Y) % вычисление СКО
sk_Y    =    skewness(Y)    %    вычисление
коэффициента ассиметрии
kur_Y    =    kurtosis(Y)    %    вычисление
коэффициента эксцесса

% Расчет моментов для выборки
% полученной с аналитического выражения
m_Y = mean (Y1) % вычисление мат. ожидания
st_Y = std(Y1) % вычисление СКО
sk_Y    =    skewness(Y1)    %    вычисление
коэффициента ассиметрии
kur_Y    =    kurtosis(Y1)    %    вычисление
коэффициента эксцесса
```

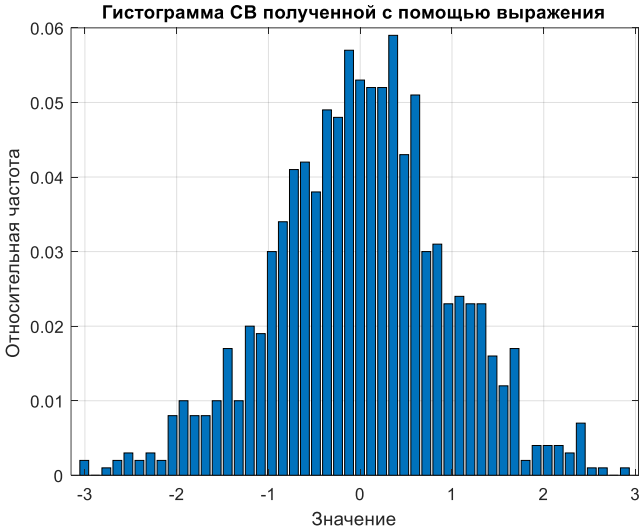


Рис. 18. Гистограмма СВ, полученной с помощью преобразования Бокса-Мюллера

В заключении используя критерий согласия Колмогорова-Смирнова проверим гипотезу о соответствии эмпирических функций распределения выборок, полученных с помощью функции MATLAB и аналитического выражения нормальному распределению.

```
% Построим эмпирические функции
распределения
% MATLAB
[F_emp_m val_m] = ecdf(Y);
% Аналитическое выражение
[F_emp_expr val_expr] = ecdf(Y1);
figure()
plot(val_m,F_emp_m,'b',val_expr,F_emp_exp
r,'--g',x,F,'-.r','LineWidth',2)
axis([-3 3 0 1])
grid on
legend('MATLAB',                'Выражение',
'Теоретическая')
xlabel('x')
ylabel('F(x)')
% Создание тестового распределения
% Нормальное с параметрами m=0; sigma=1
test_cdf                                =
makedist('normal','mu',0,'sigma',1);
% Используем критерий согласия
% Колмогорова-Смирнова
% Получаем вероятность того, что
%   гипотеза   о   соответствии   выборки
нормальному
% распределению подтверждается
[h p] = kstest(Y,'CDF',test_cdf); % MATLAB
[h1   p1]   =   kstest(Y1,'CDF',test_cdf);   %
Выражение
```

Функция kstest возвращает логический 0 если гипотеза подтверждается, и логическую единицу если гипотеза отклоняется. По умолчанию уровень значимости стоит 0,05.

Варианты заданий

Для каждого варианта параметры распределений определяются следующим образом:

1. Для нормально распределения $N(m, \sigma)$, $\sigma > 0$:
   $m =$ <<номер варианта>>, $\sigma =$<< $\sqrt{\text{номер варианта}}$.
2. Для равномерного распределения $R(a,b)$, $a < b$:
   $a =$ <<номер варианта>>, $b = 2a$.
3. Для экспоненциального распределения $E(\lambda)$, $\lambda > 0$:

$$\lambda = \begin{cases} \dfrac{N}{2}, & N < 10, \\ \dfrac{N}{5}, & 10 \le N \le 20, \\ \dfrac{N}{10}, & 20 < N. \end{cases} \text{ где } N = \text{<<номер варианта>>},$$

Номер варианта определяется в соответствии со списком группы, который студенты предоставляют преподавателю на первом занятии.

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. Графики гистограммы выборок, имеющих нормальное распределение объемом 300, 700 и 1500 чисел, сформированные с помощью встроенной функции и с помощью аналитического выражения.
4. Графики гистограммы выборок, имеющих равномерное распределение объемом 300, 700 и 1500 чисел, сформированные с помощью встроенной функции и с помощью аналитического выражения.
5. Графики гистограммы выборок, имеющих экспоненциальное распределение объемом 300, 700 и 1500 чисел, сформированные с помощью встроенной функции и с помощью аналитического выражения.

6. Для каждого объема выборки необходимо вычислить выборочные моменты и занести их в соответствующую таблицу.
7. Графики теоретической функции распределения и плотности распределения вероятности для каждого распределения.
8. Результаты проверки выборки на соответствие заданному закону распределения с помощью критерия согласия.
9. Выводы по проделанной работе.

2.4. МОДЕЛИРОВАНИЕ УРАВНЕНИЯ РЕГРЕССИИ

Цель работы – освоить методы моделирования уравнения регрессии и метод оценки его параметров, а также изучить возможности пакетов Excel и MATLAB по моделированию и анализу параметров уравнения регрессии.

В эконометрических исследованиях часто встречается ситуация, когда каждому значению переменной x соответствует (условное) распределение вероятностей переменной y. Эта зависимость неоднозначна, поэтому в эконометрических исследованиях актуальной является задача поиска закономерностей изменения параметров закона распределения y в зависимости от x. Зависимость между значениями одной из переменных и условным математическим ожиданием другой называется корреляционной зависимостью. В общем случае распределение y может зависеть от $x_1, x_2, \ldots, x_n$.

Зависимую переменную y называют выходной переменной, независимую называют – входной переменной или регрессором. Уравнения связи между ними называют уравнением регрессии. В случае единственной входной переменной регрессию называют парной, в общем случае – множественной.

По условию вхождения переменных и постоянных коэффициентов (параметров) в уравнение регрессии различают линейную по переменным (или параметрам) и нелинейную.

Приведем основные определения необходимые при выполнении данной работы.

Тренд – это долговременная тенденция изменения исследуемого временного ряда. Тренды могут быть описаны различными уравнениями – линейными, логарифмическими, степенными и так далее. Фактический тип тренда устанавливают на основе подбора его функциональной модели статистическими методами либо сглаживанием исходного временного ряда.

Корреляция – это статистическая взаимосвязь двух или нескольких случайных величин. При этом изменения значений одной или нескольких из этих величин сопутствуют систематическому изменению значений другой или других величин. Математической мерой корреляции двух случайных величин служит корреляционное отношение, либо коэффициент корреляции.

Регрессия – в теории вероятностей и математической статистике – математическое выражение, отражающее зависимость зависимой переменной y от независимых переменных x при условии, что это выражение будет иметь статистическую значимость.

Выборочное среднее – это приближение теоретического среднего распределения, основанное на выборке из него.

Выборочная дисперсия – это оценка теоретической дисперсии распределения, рассчитанная на основе данных выборки.

Выборочная ковариация – выборочная ковариация является мерой взаимосвязанности двух переменных и позволяет выразить данную связь одним числом.

При исследовании экономических закономерностей законы распределения значений выходной переменной неизвестны. Поэтому для приближенной оценки (аппроксимации) истинной функции регрессии используется выборочный метод.

В современных условиях вычисление коэффициентов корреляционной зависимости можно производить, используя компьютерные программы, например, в MS Excel существуют опции «Регрессия» и «Корреляция», находящиеся в надстройке «Пакет анализа».

Таким образом, регрессионная модель представляет связь количественных показателей экономики, как некоторую

закономерность в среднем по совокупности наблюдений, в виде аналитической формулы (функции).

Эконометрическое исследование количественного показателя включает формулировку вида модели, соответствующей экономической теории. Прежде всего, определяется круг факторов, влияющих на изучаемый показатель. (В зависимости от количества факторов, включенных в модель, различают парную и множественную регрессии). Парная регрессия достаточна, если используется при моделировании один доминирующий фактор, если такого нет, то для анализа изучаемого показателя предлагается множественная регрессия. Далее рассмотрим парную регрессию.

Пусть имеется n пар чисел (x_i,y_i), $i=1,2,\ldots,n$, относительно которых предполагается, что они отвечают линейной зависимости между величинами x и y: $y=a+bx$, возможно, с некоторой ошибкой ε_i, так что

$$y_i = a + bx_i + \varepsilon_i, i = 1,2,\ldots n. \quad (1)$$

Как можно определить какими должны быть наилучшие значения параметров a и b?

Применяя метод наименьших квадратов, зададимся условием, при котором сумма квадратов ошибок ε_i будет наименьшей:

$$\sum_{i=1}^{n} \varepsilon_i^2 \to \min. \quad (2)$$

Подставляя значения ε_i из (1) в (2), получим функцию

$$\Phi(a,b) = \sum_{i=1}^{n} (a + bx_i - y_i)^2 \to \min. \quad (3)$$

Необходимым условием минимума этой функции, как известно, является равенство нулю ее частных производных по a и b:

$$\frac{\partial \Phi}{\partial a} = 0, \ \frac{\partial \Phi}{\partial b} = 0$$

Вычисляя производные, приходим к системе уравнений

$$\begin{cases} \sum_{i=1}^{n}\left(a+bx_i-y_i\right)=0, \\ \sum_{i=1}^{n}\left(a+bx_i-x_i\right)=0. \end{cases} \tag{4}$$

Заметим, что эту систему можно записать короче в виде

$$\begin{cases} \sum_{i=1}^{n}\varepsilon_i=0, \\ \sum_{i=1}^{n}\varepsilon_i x_i=0. \end{cases}$$

Система (4) равносильна системе

$$\begin{cases} na+b\sum x_i=\sum y_i, \\ a\sum x_i+b\sum x_i^2=\sum x_i y_i. \end{cases} \tag{5}$$

решение которой находится без большого труда:

$$a=\frac{\sum y_i\sum x_i^2-\sum x_i\sum x_i y_i}{n\sum x_i^2-\left(\sum x_i\right)^2}\equiv\hat{a},$$

$$b=\frac{n\sum x_i y_i-\sum x_i\sum y_i}{n\sum x_i^2-\left(\sum x_i\right)^2}\equiv\hat{b}.$$

Условимся далее обозначать вычисленные значения параметров как $\hat{a}$ и $\hat{b}$, чтобы отличать их от неизвестных точных значений a и b.

Введем обозначения:

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i,$$

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i,$$

$$s_x^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\overline{x}^2,$$

$$c_{xy} = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x} * \overline{y}.$$

В курсах математической статистики величины $\overline{x}$, $\overline{y}$ называются выборочными средними, s_x^2 – выборочной дисперсией, c_{xy} – выборочной ковариацией. Теперь формулу для $\hat{b}$ можно переписать в виде

$$\hat{b} = \frac{c_{xy}}{s_x^2},$$

а выражение для $\hat{a}$ получается из первого уравнения системы (5):

$$\hat{a} = \overline{y} - \hat{b}\overline{x}\ .$$

Из данной формулы видно, что точка $\left(\overline{x}, \overline{y}\right)$ лежит на прямой $y = \hat{a} + \hat{b}x.$ Поэтому функцию зависимости между величинами *x* и *y* можно записать также в виде: $y - \overline{y} = \hat{b}\left(x - \overline{x}\right).$

Этапы выполнения работы

1. Сгенерируйте ряд независимых переменных X, представляющий собой массив чисел от 1 до 50.
2. Рассчитайте значения зависимой переменной Y, по уравнению в соответствии с параметрами, заданными по варианту (уравнение и параметры указаны в пункте «варианты заданий»).
3. Используя встроенный в пакеты Excel и MATLAB, генератор случайных чисел сгенерируйте массив, состоящий из 50 случайных чисел. Распределение случайных чисел – нормальное с параметрами математического ожидания 0 и среднеквадратического отклонения 1.
4. Рассчитайте набор значений:

$$y_i = a + bx_i + \varepsilon_i,\ i = 1, 2, ..., 50.$$

где параметры *a* и *b* заданы по варианту, а значения ε получены в пункте 3.

5. Скопируйте (не подряд) любые 10 пар (x_i, y_i) в другой массив (при выполнении в пакете Excel можно скопировать их на

другой лист), и далее для каждого из массивов чисел – 10 пар и 50 пар выполняем следующие пункты.

6. Построить график зависимости показателя y_i от фактора x_i. Обязательно подписать оси на графике.

7. На построенный в пункте 6 график нанести линию тренда.

8. Получить коэффициенты $\hat{a}$ и $\hat{b}$ прямой $y = \hat{a} + \hat{b}x$.

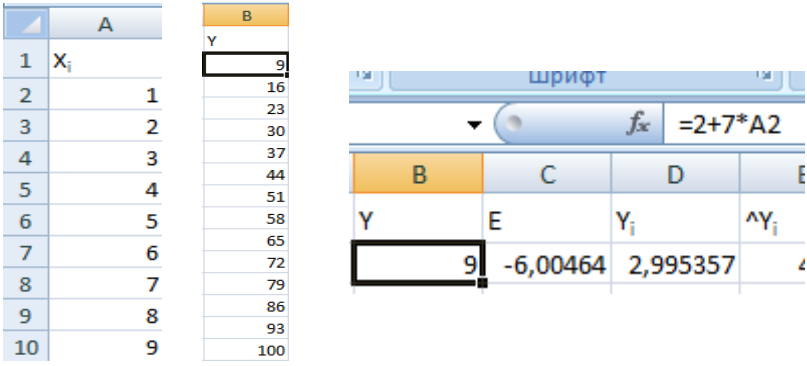
9. Построить набор значений $\hat{y}$ по уравнению $\hat{y}_i = \hat{a} + \hat{b}x_i$. Добавить эту прямую к графикам, полученным в результате выполнения пунктов 6 и 7.

10. Проверить совпала ли полученная в пункте 9 прямая с линией тренда.

Решение задачи в пакете MS Excel

1. Создаем ряд независимых случайных величин X из 50 значений. Берём от 1 до 50 (рис. 19а).

2. Формируем значения зависимой переменной Y* в соответствии с уравнением вашего варианта. Здесь Y=a+b*X. (рис. 16б, 16в).



а) б) в)

Рис. 19. Формирование *X* и *Y*

3. Выполняем генерацию ряда случайных чисел e_i ($i = 1, \ldots, 50$), используя пакет анализа – «Генерация случайных чисел» (рис. 20). Пакет «Анализ данных» находится в графе данные, если он по умолчанию выключен, то, необходимо обратиться к преподавателю. Случайные числа – ошибки – должны быть распределены по нормальному закону распределения $N(m,\sigma)$ величина разброса должна быть сопоставима с выбранными значениями независимой переменной. (Например, если значения фактора X изменяются в пределах от 1 до 50, то величину разброса можно задать 20).

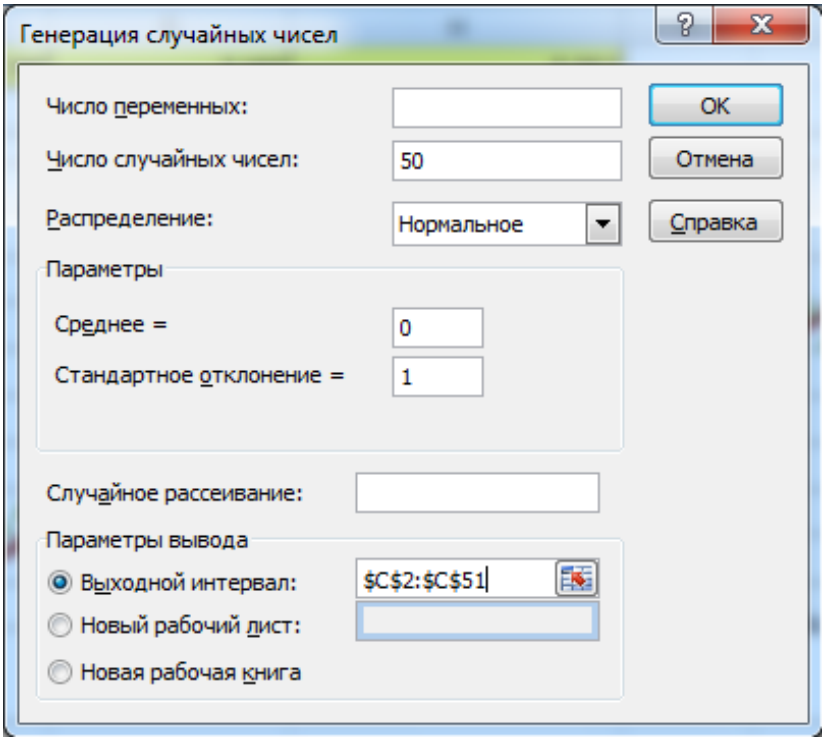


Рис. 20. Генерация случайных чисел

В выходном интервале указываем диапазон ячеек, куда хотим записать сгенерированные случайные числа.

4. Вычисляем набор значений $y_i = a + bx_i + \varepsilon_i$, $i = 1, 2, ..., 50$. как это показано на рис. 21:

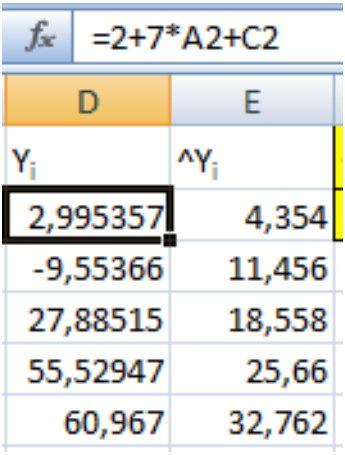


Рис. 21. Вычисление набора значений y_i

Здесь столбец A2 – это значение X и C2 – это значение случайной величины e по нормальному закону распределения полученной в пункте 2.

5. Скопируем любые 10 пар (x_i, y_i) НЕ ПОДРЯД! – на отдельный лист.

6. Построим диаграмму зависимости показателя y_i от фактора x_i. При построении выбираем тип диаграммы «Точечная» (без отрезков, соединяющих точки). Подписываем оси, название диаграммы и названия рядов. Для это кликнем мышкой на любое из значений X и нажмём «вставка» – диаграмма «точечная», выберем без линий соединения.

7. На диаграмму нанесем линию тренда. Для этого следует выделить правой кнопкой мыши получившуюся кривую и выбрать «Добавить линию тренда». В открывшемся меню Параметры линии тренда выбрать линейную аппроксимацию. Далее поставить флажок напротив полей: «Показывать уравнение на диаграмме» и «Поместить на диаграмму величину достоверности аппроксимации R^2».

Кликнем правой кнопкой на любую из точек и выберем «добавить линию тренда» этот процесс показан на рис. 22–24.

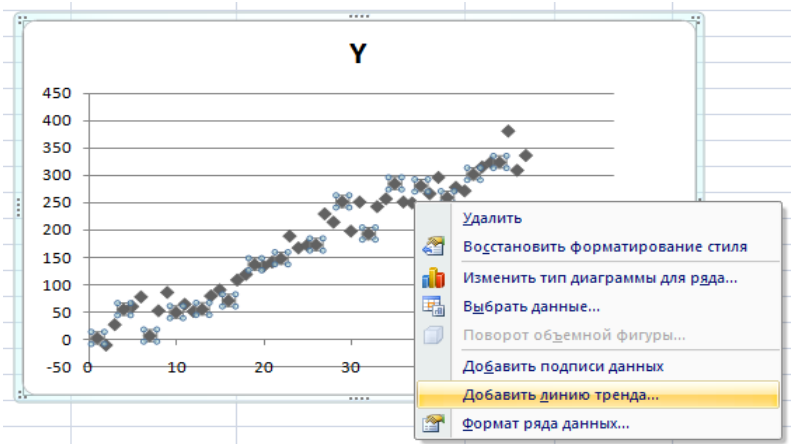


Рис. 22. Добавление линии тренда

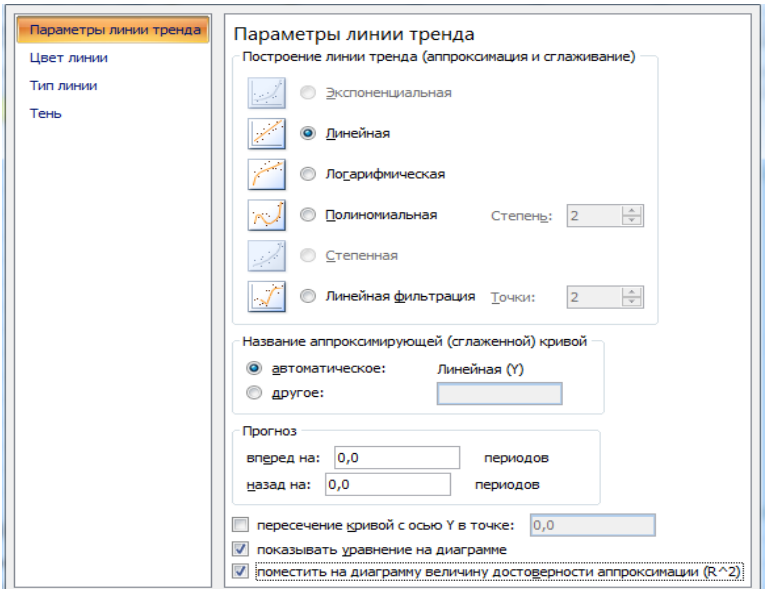


Рис. 23. Выбор параметров линии тренда

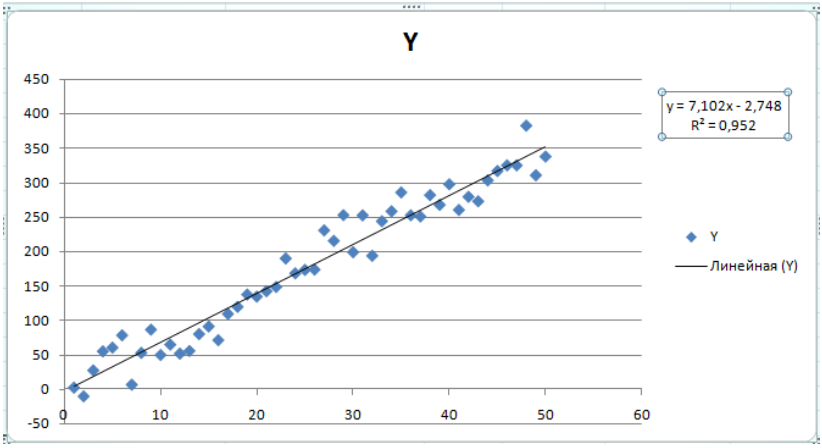


Рис. 24. Полученный результат

В результате должна появиться прямая линия.

9. Получим коэффициенты a и b прямой: $y = \hat{a} + \hat{b}x$ с помощью пакета анализа. Выделите цветом ячейки, содержащие оценки коэффициентов a и b, а также коэффициент детерминации как на рис. 25.

F	G	H
^a	^b	R^2
-2,748	7,102	0,952

Рис. 25. Значения оценок коэффициентов a и b а также коэффициент детерминации полученные по графику

Для получения коэффициентов с помощью пакета анализа выполняем следующую последовательность действий:

Вкладка «Данные» → «Анализ данных» → «Регрессия».

В диалоговом окне этой процедуры поля «Входной интервал» Y (задаём значения y_i), «Входной интервал X» (задаём значения x_i), как на рис. 26, и далее запишите коэффициенты эмпирической прямой $y = \hat{a} + \hat{b}x$, где $\hat{a}$ и $\hat{b}$ – полученны в п. 8

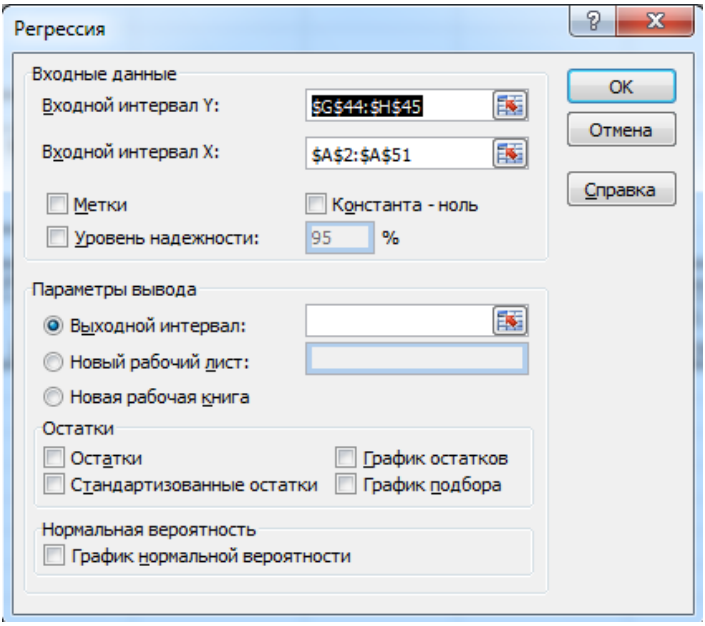


Рис. 26. Работа блока «регрессия»

В графе входной интервал мы указываем место, где хотим получить результат. Результат будет выглядеть так, как показано на рис. 27.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,975760024
R-квадрат	0,952107625
Нормированный R-квадрат	0,951109867
Стандартная ошибка	23,46093613
Наблюдения	50

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	525232,4841	525232,5	954,2472203	2,48698E-33
Остаток	48	26419,94517	550,4155		
Итого	49	551652,4292			

	Коэффициенты	*Стандартная ошибка*	*татисти*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-2,748410271	6,73655686	-0,40798	0,68509929	-16,2931654	10,79634486	-16,2931654	10,7963448
Переменная X 1	7,102288045	0,229915277	30,89089	2,48698E-33	6,640012405	7,564563684	6,640012405	7,56456368

Рис. 27. Результаты работы блока «Регрессия»

Далее строим набор значений $\hat{y}_i$ по уравнению: $\hat{y}_i = \hat{a} + \hat{b}x_i$, i=1,…,n и добавляем график этой прямой на диаграмму. Убеждаемся, что линия тренда и построенная прямая совпадают как на рис. 28. Красная линия – это линия, полученная по оценкам.

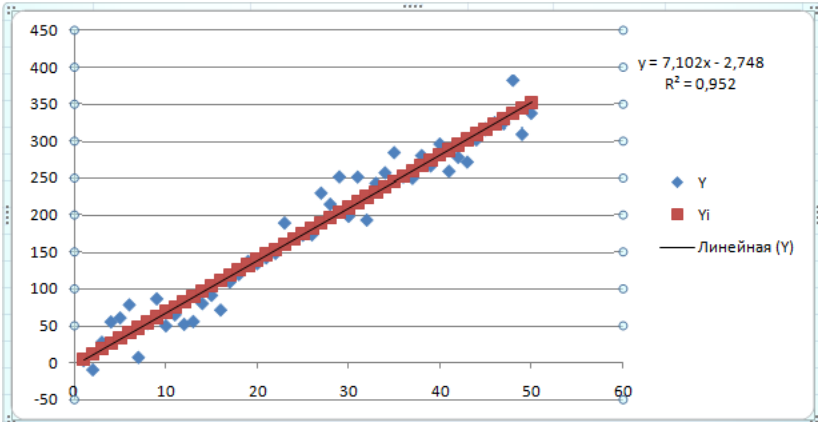


Рис. 28. График экспериментальной прямой и линии тренда

Решение задачи в пакете MATLAB

```
clear all
close all
clc

% Ввод входных параметров
X = 1:1:50;
% Значения a и b берутся из варианта задания
a_ist = 2;
b_ist = 7;
% Вычисление Y без ошибки
Y = a_ist + b_ist.*X;
% Генерация случайной величины
Error = normrnd(0,20,[1 length(X)]);
% Вычисление наблюдаемых значений Y
Y_nabl = a_ist + b_ist.*X + Error;
```

```
% Нанесение линии тренда
n = 1; % 1 - линейная интерполяция
p = polyfit(X,Y_nabl,n);
f = polyval(p,X); % построить линию тренда

% Вычисление оценок a и b
a_est = (sum(Y_nabl)*sum(X.^2) ...
    -sum(X)*sum(X.*Y_nabl))/ ...
    (length(X)*sum(X.^2)-(sum(X))^2);
b_est = (length(X)*sum(X.*Y_nabl) ...
    -sum(X)*sum(Y_nabl))/ ...
    (length(X)*sum(X.^2)-(sum(X))^2);
% Вычисление оценки Y по МНК
Y_est = a_est + b_est.*X;
% Вычисление коэффициента корреляции
r                                          =
b_est*(sqrt(var(X))/sqrt(var(Y_nabl)));
% Вычисление средней ошибки аппроксимации
per1 = zeros(length(X));
for i=1:1:length(X)
    per1(i) = abs((Y(i)-Y_est(i))/Y(i));
end
A = (1/length(X))*sum(per1)*100;
% Отрисовка графики
figure()
% График наблюдаемых значений X и Y
plot(X,Y_nabl,'o')
xlabel('x')
ylabel('y')
grid on
hold on
% Нанесение линии тренда
plot(X,f,'-.g','LineWidth',2)
% Нанесение прямой расчитанной по МНК
plot(X,Y_est,'--r','LineWidth',4)
legend({'Наблюдаемые значения X и Y', ...
    'Тренд','Прямая расчитанная по МНК'},
...
```

```
    'Location','southeast')
% Нанесение коэффициента детерминации на
график
str = {'R^2 = ', num2str(r^2)};
text(10,300,str)
```

Варианты заданий

Таблица 1. Коэффициенты уравнения $y = a + bx$

№ варианта	a	b	№ варианта	a	b
1	9	-2	11	4	3
2	6	3	12	7	4
3	4	5	13	2	-4
4	1	6	14	8	-3
5	2	-4	15	1	6
6	10	-3	16	3	2
7	16	-5	17	13	-7
8	2	4	18	8	-2
9	6	-2	19	11	-6
10	5	2	20	2	7

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. Две выборки объемом 50 и 10 чисел.
4. Для каждой выборки требуется построить графики зависимости $y(x)$ с нанесенной линией тренда и рассчитанной по найденным коэффициентам прямой.
5. Таблица со значениями найденных коэффициентов уравнения регрессии и коэффициентами детерминации.
6. Выводы по работе.

2.5. МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Цель работы – освоить методы моделирования множественной линейной регрессии и метод проверки ее значимости в пакете STATISTICA.

В случае нескольких объясняющих переменных, линейное уравнение регрессии расширяется:

$$y_i = a + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + e_i,\ i = 1, 2..., n,$$

где $b_1, b_2, ..., b_k$ коэффициенты при объясняющих переменных, n – объем выборки, k – количество объясняющих переменных. Теперь вместо прямой у нас имеется линейная модель – связь между каждым коэффициентом и его переменной (признаком) линейная.

Перечислим основные термины:

Среднеквадратическая ошибка (root mean squared error) – квадратный корень из среднеквадратической ошибки регрессии (наиболее широко используемый метрический показатель для сравнения регрессионных моделей).

Стандартная ошибка остатков (residual standard error) – то же самое, что и среднеквадратическая ошибка, но скорректированная для степеней свободы.

t-статистика (t-statistic) – метрический показатель для сравнения важности переменных в модели, получаемый в результате деления регрессионного коэффициента для какого-либо предиктора на стандартную ошибку коэффициента.

Взвешенная регрессия (weighted regression) – регрессия, в которой записям поставлены в соответствие разные веса.

Все другие понятия из простой линейной регрессии, такие как подгонка наименьшими квадратами и определение подогнанных значений и остатков, расширяются на множественную линейную регрессию. Например, подогнанные значений задаются следующей формулой:

$$y_i^* = a^* + b_1^* x_{1i} + b_2^* x_{2i} + ... + b_k^* x_{ki} + e_i,\ i = 1, 2..., n,$$

На практике регрессионные задачи удобно решать, используя готовые компьютерные программы из пакетов MS Excel, Statistica

и др. В частности, функция ЛИНЕЙН из Excel позволяет оценить параметры парной линейной регрессии a и b. Не останавливаясь на методике работы в Excel с функцией ЛИНЕЙН. При этом результаты вычислений представляются следующей таблицей:

Таблица 2. Результаты вычислений функции ЛИНЕЙН

Значение параметра b	Значение параметра a
Стандартное отклонение S_b	Стандартное отклонение S_a
Коэффициент детерминации R^2	Стандартное отклонение S
F-статистика	Число степеней свободы n-2
Регрессионная сумма квадратов RSS	Остаточная сумма квадратов ESS

Таким образом, множественную линейную регрессию строят при изучении связи между исследуемым показателем и несколькими объясняющими переменными. При наличии статистически значимой линейной связи множественной линейной регрессии можно применять для прогнозирования показателя и для оценки влияния возможных изменений факторов на показатель.

Этапы выполнения работы

1. Постройте множественную линейную регрессию на все факторы с помощью функции ЛИНЕЙН.

2. Запишите полученное уравнение, дайте интерпретацию коэффициентам.

3. Интерпретируйте множественный коэффициент детерминации.

4. Вычислите скорректированный коэффициент детерминации по формуле:

$$R^2_{скорр} = 1-(1-R^2)\cdot\frac{n-1}{n-k-1}.$$

Сравните с коэффициентом детерминации R^2.

1. Сделайте предварительный вывод о тесноте линейной зависимости между показателем y и объясняющими факторами $x_1, x_2, \ldots, x_k$ на основании сравнения коэффициентов детерминации.

2. Выберите приемлемый уровень значимости α (вероятность ошибки; обычно – число из отрезка [0,05; 0,1]) для исследования статистической значимости. Проверьте статистическую значимость модели в целом с помощью *F*-теста.

3. Проверьте статистическую значимость коэффициентов $\hat{b}_i$, *i = 1,..,m,* с помощью *t*-теста.

4. Сделайте предварительные выводы о возможности использования построенной модели на основании анализа статистической значимости, проведенного в пунктах 6 и 7.

5. Получите результаты регрессионного анализа с помощью Пакета Анализа для этого нажмите «Сервис», затем «Анализ данных» и выберите «Регрессия». В диалоговом окне этой процедуры отметьте дополнительно поле «Остатки» с тем, чтобы облегчить вычисление средней ошибки аппроксимации.

6. Вычислите среднюю ошибку аппроксимации по формуле:

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

Сделайте выводы о возможности использования построенной модели для прогнозирования.

1. Выполните расчет прогнозного значения показателя $\hat{y}$, предполагая, что прогнозные значения факторов x_i^*, *i = 1,…,k,* составят 30% от их среднего уровня $\overline{x}_i$, *i = 1,…,k,*. Дайте интерпретацию результатам прогнозирования.

2. Импортируйте исходные данные в программу STATISTICA. Для этого необходимо проделать следующее:

2.1. Запустить программу STATISTICA. В меню «Файл» выбрать позицию «Создать». Заполнить поля «Число переменных» и «Число наблюдений». Установить переключатель «Положение» в позицию «В новой рабочей книге» затем нажмите «ОК».

2.2. Скопировать заданный массив значений и поместить его в Таблицу данных в STATISTICA. Обозначить названия столбцов.

3. Восстановите коэффициенты *a* и $b_1, b_2, \ldots, b_k$ в модуле «Множественная регрессия». Для этого необходимо: Во вкладке «Анализ» выбрать «Множественная регрессия» и указать

зависимые и независимые переменные. Нажать «Ок». С помощью кнопки «Итоговая таблица регрессии» получить оценки коэффициентов регрессии $b_1, b_2, \ldots, b_k$ и свободного члена a, расположенные в столбце «В» полученной таблицы. Выделите эти коэффициенты цветом.

4. Постройте прогноз для тех же значений факторов, что и в пункте 11. Для этого необходимо в диалоге «Результаты множественной регрессии» перейти во вкладку «Остатки» и выбрать «предсказанные/наблюдаемые значения». Далее нажать кнопку «Предсказать зависимую переменную», заполнить поля для независимых переменных прогнозными значениями факторов, которые были получены в Excel. Нажать «ОК». Число на пересечении строки «Предсказ.» и столбца «В-веса*знач» в полученной таблице является искомым прогнозом зависимой переменной. Выделите цветом полученное значение. Затем сохраните полученную в пакете STATISTICA информацию для того, чтобы вставить ее в отчет. Для этого необходимо таблицы «Итоговые статистики», «Итоги регрессии для зависимой переменной», «Предсказанные значения» сохранить в одной рабочей книге.

5. Вернитесь на рабочий лист Excel. Постройте корреляционную матрицу. Для этого выберите «Пакет анализа» затем «Корреляция». Выделите цветом столбец, в котором находятся коэффициенты парной корреляции между объясняемой переменной y и объясняющими переменными $x_1, x_2, \ldots, x_k$.

6. Выясните, есть ли в построенной модели малоинформативные факторы? Если есть, то перечислите их и отметьте другим цветом ячейки, содержащие соответствующие коэффициенты корреляции.

Решение задачи в пакете MS Excel

1. Пусть имеется входной набор данных, состоящий из двух объясняющих переменных и одной объясняемой. В нашем случае в качестве объясняемой переменной будет расход топлива автомобиля, измеряемый в литрах на 100 км, в качестве

объясняющих переменных выступают вес автомобиля в килограммах и количество лошадиных сил.

2. Для начала построим корреляционную матрицу и оценим статистическую связь между переменными. Для этого воспользуемся пакетом анализа, и выполним следующую последовательность действий – перейдем во вкладку «Данные» и выберем «Анализ данных». В открывшемся окне выберем «Корреляция». Поскольку входной набор данных в виде таблицы у нас располагается в столбцах A, B и C, то в поле входной интервал заносим ячейки $A$1:$C$101, выбираем группирование по столбцам и метки в первой строке, после указываем ячейку, в которую хотим вывести готовую корреляционную матрицу. В нашем случае это ячейка F2. Вышеописанный процесс показан на рис. 29.

Проанализируем матрицу. Обратим внимание на коэффициенты парной корреляции между переменными «Расход» и «Вес автомобиля», обращаем внимание на наличие сильной прямой связи. Аналогично и между переменными «Расход» и «Количество лошадиных сил». Особое внимание нужно обратить на коэффициент парной корреляции между двумя объясняемыми переменными – «Вес автомобиля» и «Количество лошадиных сил». Наблюдаем между ними наличие линейной зависимости т.е. одна переменная влияет на другую и наоборот, и значение коэффициента корреляции близкое к 0,9, следовательно возможно наличие мультиколлинеарности.

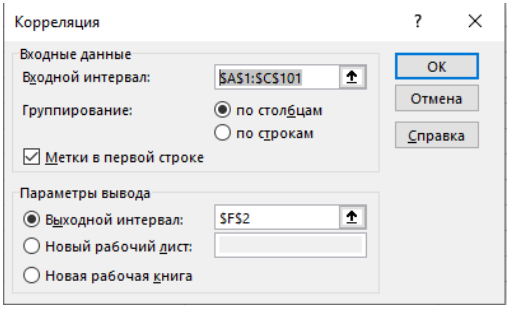


Рис. 29. Ввод исходных данных для построения корреляционной матрицы

В результате получаем корреляционную матрицу, показанную на рис. 30.

F	G	H	I
Корреляционная матрица			
	Расход, л/100 км	*Вес автомобиля, кг*	*Количество лошадиных сил*
Расход, л/100 км	1		
Вес автомобиля, кг	**0,834621467**	1	
Количество лошадиных сил	0,818723813	**0,873321784**	1

Рис. 30. Корреляционная матрица

1. Построим множественную модель регрессии. Для этого опять воспользуемся пакетом анализа. Выполним следующую последовательность действий – перейдем во вкладку «Данные» и выберем «Анализ данных». В открывшемся окне выберем «Регрессия». Заполняем появившееся окно так, как показано на рис. 31. Обязательно ставим галочку в поле «Остатки».

В результате получаем значения коэффициентов регрессии, значения стандартной ошибки для каждого коэффициента регрессии, значение *t*-статистики, которое получается в результате деления коэффициента регрессии на стандартную ошибку, далее выводится *p*-значение т.е. вероятность принятия нулевой гипотезы о том, что коэффициент регрессии в нашем случае не является статистически значимым. Поскольку *p*-значение везде меньше 0,05, следовательно все наши переменные статистически значимы.

Следующим шагом проверим остатки модели регрессии на автокорреляцию. Для этого найдем значение статистики Дарбина-Уотсона.

2. Перейдем к остаткам нашей модели. Критерий Дарбина-Уотсона можно вычислять двумя способами, через суммы остатков и с помощью коэффициента автокорреляции первого порядка:

$$DW = \frac{\sum_{t=2}^{T}\left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{T} e_t^2} \approx 2\left(1 - r\right).$$

Рассчитаем коэффициент автокорреляции. Для этого нам необходимо в два столбца записать значения e_t и e_{t-1}. Запишем их в столбцы K и L ячейки 33–131, как это показано на рис. 32.

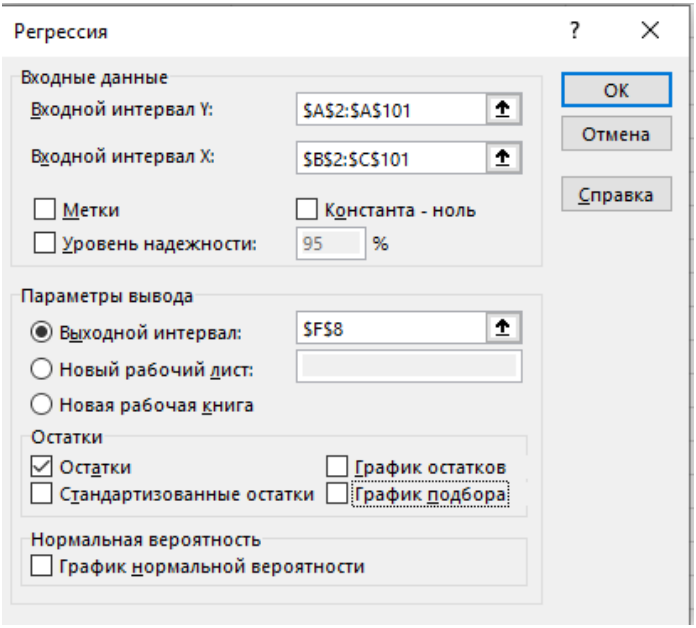


Рис. 31. Ввод исходных данных для построения множественной регрессии

K	L
et-1	et
-0,06991	0,799016
0,799016	-0,59962
-0,59962	1,041679
1,041679	0,489598
0,489598	-2,09255

Рис. 32. Вычисление вспомогательных массивов остатков

Теперь для полученных значений построим матрицу корреляции, способом, аналогичным пункту 2. Значение коэффициента парной корреляции между e_t и e_{t-1} получилось равным 0,55. Рассчитаем критерий Дарбина-Уотсона и получим значение приблизительно равное 0,9.

На практике применение критерия основано на сравнении величины критерия с теоретическими значениями d_L и d_U для заданного числа наблюдений n, числа объясняющих переменных модели и уровня значимости.

В нашем случае количество наблюдений n=100, объясняющих переменных у нас 2, и уровень значимости 5%. Найдем по таблице значения d_L и d_U, они получились равными 1,63 и 1,72 соответственно.

Критерием пользуются следующим образом:

– Если величина критерия меньше d_L, то гипотеза о независимости случайных отклонений отвергается, следовательно присутствует положительная автокорреляция;

– Если величина критерия больше d_U то гипотеза не отвергается;

– Если $d_L < DW < d_U$, то нет достаточных оснований для принятия решения;

– В нашем случае $DW < d_L$, поэтому автокорреляция присутствует.

Проверим данное заключение с помощью другого теста.

1. Поскольку мы уже нашли коэффициент корреляции между остатками, несложно будет провести тест Бройша-Годфри. Особенностью данного теста является то, что его можно использовать практически всегда, в отличии от теста Дарбина-Уотсона, поскольку он позволяет оценить автокорреляцию любого порядка. Тем не менее, следует помнить, что тест Бройша-Годфри является асимптотическим, то есть для достоверности выводов требуется большой объем выборки. С помощью статистики Стьюдента проверим наш коэффициент корреляции, который мы получили в пункте 4, по формуле:

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

где r – это коэффициент парной корреляции, n – объем выборки.

В нашем случае значение $t_{набл}$ получилось равным 6,521, а критическое значение найдем с помощью функции СТЬЮДЕНТ.ОБР(0,95;98), 0,95, потому что уровень значимости у нас 5%, а степеней свободы у нас n-2=98. $t_{крит}$ в нашем случае получилось 1,66, следовательно $t_{набл} > t_{крит}$, таким образом нулевая гипотеза отвергается, коэффициент корреляции значим, и в модели регрессии присутствует автокорреляция остатков случайных отклонений.

Проверим модель на Гетероскедастичность. Используем для этого тест Уайта. Для этого нам необходимо оценить регрессию квадратов остатков на все объясняющие переменные, их квадраты, попарные произведения и константу. Математически это записывается следующим образом:

$$e_t^2 = a + b_1 x_1 + b_{11} x_1^2 + b_2 x_2 + b_{22} x_2^2 + b_{12} x_1 x_2,$$

где x_1 в нашем случае это вес автомобиля, а x_2 это количество лошадиных сил. Вычислим необходимое, а остатки скопируем из оценённых в предыдущих пунктах. Получим таблицу, представленную на рис. 33.

Далее, с помощью пакета анализа и модуля «Регрессия» строим необходимое уравнение регрессии. В качестве «Входного интервала Y» выбираем столбец, где хранятся квадраты остатков. В качестве «Входного интервала X» выбираем столбцы, где хранятся необходимые объясняющие переменные. Указываем ячейку, куда вывести итоги. Нажимаем «Ок».

В последней оцененной регрессии находим коэффициент множественной детерминации R2. Вычисляем тестовую статистику по формуле nR2. При выполнении нулевой гипотезы тестовая статистика имеет распределение «хи – квадрат» с m-1 степенями свободы, где m – число оцениваемых в последней регрессии коэффициентов. В нашем случае тестовая статистика равна 59,55, а критической значение, вычисленное с помощью ХИ2.ОБР равняется 9,48. Таким образом нулевую гипотезу о гомоскедастичности отвергаем. Присутствует гетероскедастичность.

H1

	A	B	C	D	E	F	G
1	Расход, л/100 км	Вес автомобиля, кг	Количество лошадиных сил	x1^2	x2^2	x1*x2	квадраты остатков
2	13,06666667	1589,403974	130	2526204,99	16900	206622,5166	0,004887419
3	15,68	1675,133811	165	2806073,29	27225	276397,0788	0,638426401
4	13,06666667	1558,559376	150	2429107,33	22500	233783,9064	0,359539198
5	14,7	1557,198585	150	2424867,43	22500	233579,7877	1,085095895
6	13,83529412	1564,456137	140	2447523,01	19600	219023,8592	0,239706012
7	15,68	1969,064683	198	3877215,73	39204	389874,8072	4,37874592
8	16,8	1974,961444	220	3900472,71	48400	434491,5177	3,197802676
9	16,8	1955,910369	215	3825585,37	46225	420520,7294	2,247689187

Рис. 33. Вычисленные значения для построения регрессии

Далее рассмотрим пример выполнения в пакете MATLAB.

Решение задачи в пакете MATLAB

Пусть имеется тот же самый набор данных, где переменная, `x1_kg` обозначает вес автомобиля, в килограммах, `y_km` обозначает расход в литрах на 100 километров, а `x2` обозначает количество лошадиных сил. Тогда, для поиска коэффициентов множественной регрессии можно воспользоваться следующим программным кодом.

```
% Матрица объясняющих переменных
X = [ones(size(x1_kg)) x1_kg x2 x1_kg.*x2];
% b - Вычисленные коэффициенты регрессии;
% r - Вектор остатков;
% stats - Выводит значения R^2
%         F-статистики
%         p-значение
%         Оценку дисперсии ошибки
%         ~ - для пропуска некоторых
%         выводимых функцией значений
[b,~,r,~,stats] = regress(y,X);
% Построение 3-D графика
scatter3(x1_kg,x2,y,'filled')
hold on
x1fit = min(x1_kg):100:max(x1_kg);
x2fit = min(x2):10:max(x2);
```

```
[X1FIT,X2FIT] = meshgrid(x1fit,x2fit);
YFIT = b(1) + b(2)*X1FIT + b(3)*X2FIT +
b(4)*X1FIT.*X2FIT;
mesh(X1FIT,X2FIT,YFIT)
xlabel('Вес, Кг')
ylabel('Л.с.')
zlabel('Расход, л/100 Км')
view(50,10)
hold off
```

На экран выведется график, показанный на рис. 34.

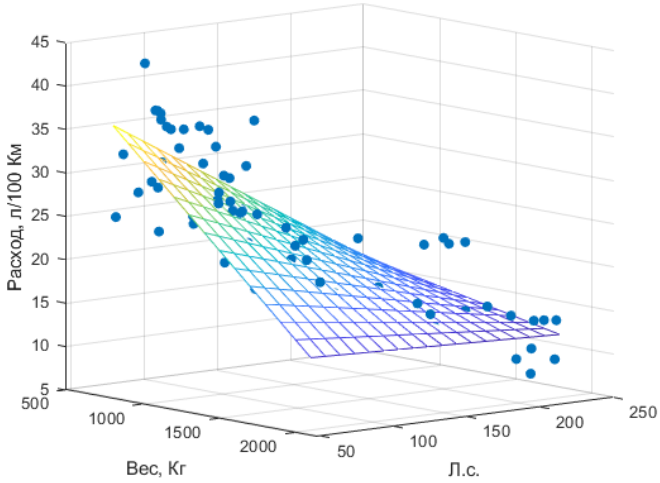


Рис. 34. График построенной модели множественной регрессии

Полученные коэффициенты, и статистические значения в дальнейшем можно проанализировать на гетероскедастичность, мультиколлинеарность и автокорреляцию аналогично тому, как мы сделали это в пакете MS Excel.

Варианты заданий

Исходные данные представляют собой многомерную выборку $(y_i, x_1^i, x_2^i, \ldots, x_k^i)$, $i = 1, \ldots n$. По выборке необходимо построить множественную регрессию на все факторы, проверить ее статистическую значимость и исследовать вопрос мультиколлинеарности.

Требования к содержанию отчета

Отчет по работе должен быть представлен в распечатанном виде и содержать следующую информацию:

1. Титульный лист.
2. Формулировка задачи и таблица исходных данных.
3. Итоговая таблица с результатами обращения к функции ЛИНЕЙН.
4. Построенное уравнение регрессии, экономическая интерпретация его коэффициентов.
5. Интерпретация множественного коэффициента детерминации R^2.
6. Расчет скорректированного коэффициента детерминации, его сравнение с множественным коэффициентом детерминации, предварительный вывод о тесноте линейной зависимости на основании этого сравнения.
7. Выбранный уровень значимости α, значение $F_{крит}$, его сравнение с F-статистикой и вывод о статистической значимости модели в целом.
8. t-статистики коэффициентов $\hat{b}_i$, $i = 1, \ldots, k$, значение $t_{крит}$ и вывод о статистической значимости каждого из коэффициентов.
9. Общий вывод о возможности использования модели на основании анализа статистической значимости.
10. Итоговые таблицы с результатами обращения к «Пакету анализа», включая таблицу остатков.
11. Расчет средней ошибки аппроксимации (таблица «Остатки» со вспомогательными столбцами). Предварительный вывод о качестве построенной модели на основании полученного значения.

12. Расчет прогнозных значений факторов (формулы!) и прогнозного значения показателя в соответствии с заданием. Экономическая интерпретация прогноза. Комментарии о правдоподобности сделанного прогноза.

13. Таблица исходных данных в пакете STATISTICA.

14. Итоговая таблица регрессии (в пакете STATISTICA) с выделенными цветом коэффициентами.

15. Итоговая таблица, содержащая прогноз для тех же значений факторов, что и в пункте 11 (в пакете STATISTICA) с выделенным цветом предсказанным значением.

16. Корреляционная матрица с выделенными цветом ячейками в соответствии с заданием.

17. Анализ модели на наличие малоинформативных факторов и ярко выраженной мультиколлинеарности.

18. Окончательный вывод о качестве построенной модели (как можно использовать построенную модель?) и правдоподобности прогноза.

2.6. НЕЛИНЕЙНАЯ РЕГРЕССИЯ

Цель работы – освоить основные методы моделирования нелинейной регрессии и научиться методике проверки ее значимости в пакете STATISTICA.

Связь между объясняемой и объясняющей переменной не обязательно линейна. Например, в медицине отклик на дозу препарата часто не линеен: удвоение дозы обычно не приводит к удвоенному отклику. Спрос на продукт не является линейной функцией маркетинга расходования денег, поскольку в какой-то момент спрос, вероятно, будет удовлетворен. Существует несколько способов, которыми регрессия может быть расширена для получения этих нелинейных эффектов.

Перечислим основные термины:

Параболическая регрессия (polynomial regression) – добавляет в регрессию полиномиальные члены (квадраты, кубы и т. д.).

Сплайновая регрессия (spline regression) – подгонка гладкой кривой с серией полиномиальных сегментов.

Узлы (knots) – значения, которые отделяют сплайновые сегменты.

Обобщенные аддитивные модели (generalized additive models) -сплайновые модели с автоматизированным выбором узлов.

Когда статистики говорят о нелинейной регрессии, они всегда ссылаются на модели, которые не могут быть подогнаны при помощи наименьших квадратов. Какие модели не являются линейными? По существу, это все модели, где отклик не может быть выражен как линейная комбинация объясняющих переменных или их некая трансформации. Нелинейные регрессионные модели более жесткие и вычислительно более емкие для выполнения подгонки, поскольку они требуют численной оптимизации. По этой причине, если это возможно, предпочтение отдается использованию линейной модели.

Рассмотрим параболическую регрессию, или как ее еще называют полиномиальную, регрессия связана с включением в состав уравнения регрессии полиномиальных членов. Например, квадратичная регрессия между объясняемой переменной *y* и объясняющей *x* примет следующую форму:

$$y_i = a + b_1 x_i + b_2 x_i^2 + b_k x_i^k + e_i .$$

Параболическая регрессия захватывает только определенное количество кривизны в нелинейной связи. Добавление членов более высоких степеней, таких как кубический биквадратный полином, часто приводит к нежелательной "волнистости" в уравнении регрессии. Альтернативный, и часто превосходящий, подход к моделированию нелинейных связей состоит в использовании сплайнов. Сплайны предоставляют способ гладко интерполировать между фиксированными точками. Сплайны первоначально использовались чертежниками для нанесения плавной кривой, в частности, в судостроении и самолетостроении.

Техническое определение сплайна является серией кусочно-непрерывных полиномов. В отличие от линейного члена, для которого коэффициент имеет прямое значение, коэффициенты для

сплайнового члена не интерпретируемы. Вместо этого полезнее использовать визуальное отображение для выявления природы сплайновой подгонки. В отличие от полиномиальной модели, сплайновая модель намного ближе соответствует сглаженной, демонстрируя большую гибкость сплайнов. В этом случае линия подогнана к данным намного ближе. Означает ли это, что сплайновая регрессия является лучшей моделью? Не обязательно. С экономической точки зрения нет никакого смысла в том, чтобы, чтобы очень небольшие дома (площадью менее 100 $м^2$) имели более высокую стоимость, чем дома немного большего размера. Это, возможно, артефакт искажающей переменной.

Таким образом, ключевые идеи для нелинейной регрессии, следующие:

1. Выбросы в регрессии – это записи с большим остатком.
2. Мультиколлинеарность может вызвать числовую нестабильность в подгонке уравнения регрессии.
3. Искажающая переменная – это важная объясняющая переменная, может быть упущена из модели и может привести к уравнению регрессии с мнимыми связями.
4. Член уравнения, характеризующий взаимодействия между двумя переменными, необходим, если эффект одной переменной зависит от уровня другой.
5. Параболическая регрессия может подгонять нелинейные связи между объясняемой и объясняющими переменными.

Возможно, никакой другой статистический метод не видел более широкого применения на протяжении многих лет, чем регрессия – процесс установления связи между многочисленными объясняющими переменными и объясняемой переменной. Ее фундаментальная форма линейна: каждая объясняющая переменная имеет коэффициент, который описывает линейную связь между предиктором и исходом. Более усовершенствованные формы регрессии, такие как параболическая и сплайновая регрессия, допускают нелинейность связи. В классической статистике главный упор делается на отыскании хорошей подгонки к наблюдаемым данным, чтобы объяснить или описать какое-либо явление, и сила этой подгонки зависит от того, как

используются традиционные метрические показатели для диагностики модели. В науке о данных, в отличие от этого, как правило, цель состоит в том, чтобы предсказывать значения для новых данных, поэтому используются метрические показатели, основанные на предсказательной точности для вневыборочных данных. Кроме того, применяются методы отбора переменных с целью уменьшить размерность и создать более компактные модели.

Вспомним систему уравнений, полученную в (5) работы 2.4. посвященной линейной регрессии.

$$\begin{cases} na + b\sum x_i = \sum y_i, \\ a\sum x_i + b\sum x_i^2 = \sum x_i y_i. \end{cases}$$

Соответствующую систему нормальных уравнений можно получить для любого уравнения линейной и нелинейной регрессии. При этом для нелинейной зависимости необходимо произвести линеаризацию переменных.

Нелинейные регрессии бывают трех типов.

1. Регрессии линейные по оцениваемым параметрам:

1.1. Равносторонняя гипербола – $y = a + \frac{b}{x}$;

1.2. Логарифмическая гипербола – y = a + b ln(x); y = $a + b_1x +$ $+b_2x^2 + \ldots + b_nx^n$.

1.3. Полиномы различных степеней – y = $a + b_1x + b_2x^2 + \ldots + b_nx^n$.

2. Регрессии нелинейные по оцениваемым параметрам, но внутренне линейные:

2.1. Степенная – $y = ax^b$;

2.2. Экспоненциальная – $y = ae^{bx}$;

2.3. Показательная – $y = ab^x$.

3. Регрессии нелинейные по оцениваемым параметрам, и внутренне нелинейные:

$$y=a+b_1x^b_2;\ y=a\left(1-\frac{1}{1-x^b}\right).$$

Линеаризацию переменных можно произвести только для первых двух типов нелинейных уравнений регрессии. Первые приводят к линейному виду простой заменой переменных, вторые – с помощью логарифмирования (см. табл. 3). Внутренне нелинейные к линейному виду не приводятся. Оценка параметров регрессий нелинейных по оцениваемым параметрам, но внутренне линейных, основывается, как правило, на минимизации суммы квадратов отклонений в логарифмах. В результате оценки параметров для линеаризуемых уравнений оказываются несколько смещенными, то есть заниженными.

Это будет наглядно видно из дальнейшего анализа. Поэтому для таких регрессий предпочтительнее использовать нелинейные методы наименьших квадратов. В них, как и в классическом методе наименьших квадратов, находят значения параметров уравнения регрессии, при которых сумма квадратов отклонений S фактических данных y от расчетных $\hat{y}_x$ является минимальной.

Таблица 3. Линеаризация нелинейных регрессий

Регрессия	Исходное уравнение	Преобразованное уравнение
Линейные по оцениваемым параметрам		
Равносторонняя гипербола	$y=a+\frac{b}{x}$	$y=a+bx_1,\left(x_1=\frac{1}{x}\right)$
Логарифмическая	$y=a+b\ln(x)$	$y=a+bx_1,\left(x_1=\ln(x)\right)$
Квадратичная (полином 2-й степени)	$y=a+b_1x+b_2x^2$	$y=a+b_1x_1+b_2x_2,$ $\left(x_1=x;\ x_2=x^2\right)$
Кубическая (полином 3-й степени)	$y=a+b_1x+b_2x^2+b_3x^3$	$y=a+b_1x_1+b_2x_2+b_3x_3,$ $\left(x_1=x;x_2=x^2;x_3=x^3\right)$

Регрессия	Исходное уравнение	Преобразованное уравнение
Нелинейные по оцениваемым параметрам, но внутренне линейные		
Степенная	$y = ax^b$	$\ln(y) = \ln(a) + b\ln(x)$
Экспоненциальная	$y = ae^{bx}$	$\ln(y) = \ln(b) + bx$
Показательная	$y = ab^x$	$\ln(y) = \ln(a) + x\ln(b)$

Для решения этой задачи применяют два основных метода:

– прямую минимизацию функции S методами нелинейной оптимизации, которые позволяют находить экстремумы выпуклых линий (сюда можно отнести различные методы наискорейшего спуска (градиентные методы), например, метод обобщенного понижающего градиента, используемый в табличном процессоре Microsoft Excel, и др.);

– решение системы нелинейных уравнений, полученной из необходимого условия экстремума функции – равенства нулю частных производных по каждому из параметров (эта система решается итерационными методами, например, методом Гаусса-Ньютона, который используется в статистическом пакете Statistica, и др.).

Существуют также методы оценивания параметров нелинейной регрессии, которые сочетают в себе два вышеизложенных метода. Сюда можно отнести метод Левенберга-Марквардта, являющийся сочетанием направления Ньютона-Гаусса и метода наискорейшего спуска. Данный метод используется во многих статистических пакетах (Statistica, IBM SPSS Statistics и др.).

Этапы выполнения работы

1. Подготовить данные.

2. Рассчитать оптимальные значения параметров a и b уравнения линейной регрессии.

3. Провести нелинейное оценивание параметров полиномиальной, логарифмической, степенной и экспоненциальной регрессии.

4. Построить графики результатов линейного и нелинейного оценивания степенной и экспоненциальной регрессии.

Решение задачи в пакете MS Excel

1. Подготовим данные. Для этого занесем входные данные в Excel в столбцы A-D занесем соответственно порядковый номер, город, количество высших учебных заведений и численность обучающихся. Подготовленные данные показаны на рис. 35, для наглядности строки с 5 по 17 скрыты.

	A	B	C	D
1	№	Город	Кол-во ВУЗов, ед.	Численность обучающихся, тыс. чел
2	1	Белгород	7	45,4
3	2	Брянск	4	26,3
4	3	Владимир	2	27,3
18	17	Ярославль	7	28,7

Рис. 35. Подготовленные данные

2. Осуществим расчет параметров линейной, полиномиальной, логарифмической, степенной и экспоненциальной регрессий в Microsoft Excel с помощью нелинейного оценивания. Для этого используем надстройку Поиск решения, в которой реализован поиск решения нелинейных задач методом обобщенного понижающего градиента (ОПГ). Подготовим необходимые ячейки в MS Excel в соответствии с рис. 36.

	A	B	C	D	E	F
1	№	Город	Кол-во ВУЗов, ед.	Численность обучающихся, тыс. чел	Оценка y_x	Остатки $y-y_x$
2	1	Белгород	7	45,4		
3	2	Брянск	4	26,3		
4	3	Владимир	2	27,3		
18	17	Ярославль	7	28,7		
19						
20	Расчет параметров регрессии					
21	a					
22	b1					
23	b2					
24	Сумма $(y-y_x)^2 \rightarrow min$					
25	Оценка параметров регрессии					
26	Общая дисперсия					
27	Остаточная дисперсия					
28	Коэф. Детерминации					

Рис. 36. Подготовка ячеек для дальнейших расчетов

Рассчитаем параметры уравнения линейной регрессии $\hat{y}_x = a + bx$. Для этого вначале введем в ячейки E2 и F2 соответственно формулы **=$C$21+$C$22*C2** и **=D2-E2** и скопируем их в ячейки **E3:F18**.

Затем найдем оптимальные значения параметров уравнения регрессии *a* и *b*, минимизируя сумму квадратов остатков (отклонений фактических уровней ряда от предсказанных значений ряда). Для этого введем в ячейку **C24** функцию **=СУММПРОИЗВ(F2:F18;F2:F18)** и выполним команду «Данные», «Поиск решения». В диалоговом окне Параметры поиска решения установим параметры в соответствии с рис. 37.

В результате будут получены оптимальные значения параметров уравнения регрессии *a* и *b* как показано на рис 38.

Уравнение линейной регрессии имеет вид:

$$\hat{y}_x = 5,03 + 5,5x.$$

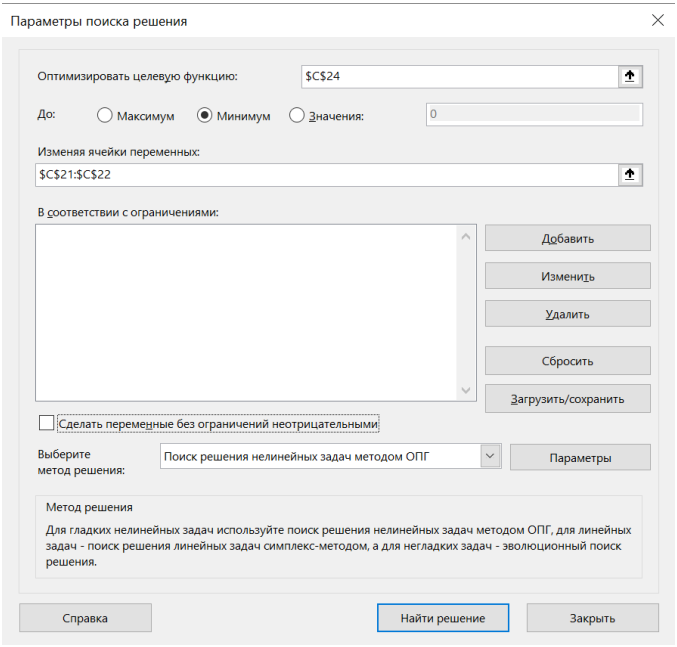


Рис. 37. Ввод параметров для работы модуля «Поиск решения»

Насколько уравнение регрессии соответствует изучаемой совокупности, оценим с помощью индекса детерминации R^2, который рассчитывается по формуле:

$$R^2 = 1 - \frac{\sigma_\varepsilon^2}{\sigma_y^2},$$

где σ^2_y – общая дисперсия объясняемой переменной (результативного признака), σ^2_ε – остаточная дисперсия объясняемой переменной. Которые рассчитываются по формулам:

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y - \bar{y})^2}{n},$$

$$\sigma_\varepsilon^2 = \frac{\sum_{i=1}^{n}(y - \hat{y}_x)^2}{n},$$

где n – объем выборки.

	A	B	C	D	E	F
1	№	Город	Кол-во ВУЗов, ед.	Численность обучающихся, тыс. чел	Оценка y_x	Остатки $y-y_x$
2	1	Белгород	7	45,4	43,53642	1,863577
18	17	Ярославль	7	28,7	43,53642	-14,8364
19						
20	Расчет параметров регрессии					
21	a		5,032732			
22	b1		5,500527			
23	b2					
24	Сумма $(y-y_x)^2 \rightarrow min$		1389,059			
25	Оценка параметров регрессии					
26	Общая дисперсия					
27	Остаточная дисперсия					
28	Коэф. Детерминации					

Рис. 38. Оптимальные значения параметров *a* и *b*

Проведем необходимые вычисления. Для этого введем:

– в ячейку **C26** функцию **=ДИСП.Г(D2:D18)** для расчета общей дисперсии;
– в ячейку **C27** формулу **=C24/A18** для расчета остаточной дисперсии;
– в ячейку **C28** формулу **=1-C27/C26** для расчета индекса детерминации.

Результаты выводятся в следующем виде как показано на рис. 39.

25	Оценка параметров регрессии	
26	Общая дисперсия	316,2187
27	Остаточная дисперсия	81,70934
28	Коэф. Детерминации	0,741605

Рис. 39. Параметры уравнения регрессии

Аналогичный результат можно получить, если рассчитать параметры уравнения регрессии с помощью пакета анализа и модуля «Регрессия».

3. Аналогично проведем нелинейное оценивание параметров полиномиальной, логарифмической, степенной и экспоненциальной регрессий. Для этого для каждого уравнения регрессии создадим копии листа с расчетными данными по линейной регрессии и на скопированных листах проведем соответствующие расчеты: на каждом листе введем в ячейку E2 соответствующие формулы и скопируем их в ячейки E3:E15:

1. **=$C$21+$C$22*C2+$C$23*C2^2** (полиномиальная регрессия).
2. **=$C$21+$C$22*LN(C2)** (логарифмическая регрессия).
3. **=$C$21*C2^$C$22** (степенная регрессия).
4. **=$C$21*EXP($C$22*C2)** (экспоненциальная регрессия).

Получим следующие результаты нелинейного оценивания (см. табл. 4). Для сравнения в таблице приведены также результаты линейного оценивания параметров регрессий.

Нелинейное оценивание степенной и экспоненциальной регрессий показывает лучшие результаты по сравнению с линейным оцениванием, поскольку получен более высокий индекс детерминации, характеризующий достоверность аппроксимации. Это следует из того, что оценки параметров для линеаризуемых уравнений, как правило, оказываются несколько смещенными, то есть заниженными.

Таблица 4. Линейное и нелинейное оценивание параметров регрессии

Регрессия	Линейное оценивание	Нелинейное оценивание
Линейная	$\hat{y}_x = 5,03 + 5,5x$ $R^2 = 0,786$	$\hat{y}_x = 5,03 + 5,5x$ $R^2 = 0,786$
Нелинейные регрессии, линейные по оцениваемым параметрам		
Полино-миальная	$\hat{y}_x = 16,723 + 0,597x + 0,388x^2$ $R^2 = 0,786$	$\hat{y}_x = 16,723 + 0,597x + 0,388x^2$ $R^2 = 0,786$

Регрессия	Линейное оценивание	Нелинейное оценивание
Логарифмическая	$\hat{y}_x = 1{,}814 + 21{,}224\ln(x)$ $R^2 = 0{,}526$	$\hat{y}_x = 1{,}814 + 21{,}224\ln(x)$ $R^2 = 0{,}526$
Нелинейные регрессии, нелинейные по оцениваемым параметрам		
Степенная	$\hat{y}_x = 112{,}156x^{0{,}595}$ $R^2 = 0{,}586$	$\hat{y}_x = 7{,}17x^{0{,}933}$ $R^2 = 0{,}586$
Экспоненциальная	$\hat{y}_x = 14{,}127x^{0{,}142x}$ $R^2 = 0{,}698$	$\hat{y}_x = 14{,}736x^{0{,}141x}$ $R^2 = 0{,}777$

Таким образом, нелинейное оценивание регрессий в Microsoft Excel, нелинейных по оцениваемым параметрам, приводит к лучшим результатам по сравнению с линейным оцениванием. Поэтому его рекомендуется применять на практике для такого рода регрессий, причем как внутренне линейных, так и внутренне нелинейных.

Варианты заданий

Варианты заданий выдает преподаватель на занятиях по расписанию. По согласованию с преподавателем студент может предложить свой набор входных данных для анализа.

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. Расчет параметров линейной регрессии.
4. Расчет параметров нелинейных регрессий.
5. Таблица, где приведены вычисленные параметры и коэффициенты детерминации.
6. Рисунки, показывающие наглядное сравнение линейной и нелинейной модели регрессии для выбранных данных.

2.7. СИСТЕМЫ ЭКОНОМЕТРИЧЕСКИХ УРАВНЕНИЙ

Цель работы – ознакомиться с основами систем эконометрических уравнений, получить навыки идентификации систем.

Довольно часто сложные процессы и явления не могут быть описаны одним уравнением регрессии. Для построения моделей таких процессов необходимо рассматривать системы регрессионных уравнений.

Рассмотрим модель вида:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2, \end{cases} \quad (1)$$

где y_1, y_2, y_3 называются эндогенными или зависимыми переменными, x_1, x_2, x_3, x_4 называются экзогенными или независимыми переменными, а параметры a_{ij} и b_{ij} называются параметрами модели или структурными коэффициентами.

Система вида (1) называется структурной формой модели, и в данной форме поиск параметров a_{ij} и b_{ij} является затруднительным. Поэтому обычно от структурной формы модели переходят к приведенной форме модели вида

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + ... + \delta_{1n}x_n, \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + ... + \delta_{2n}x_n, \\ ....., \\ y_m = \delta_{m1}x_1 + \delta_{m2}x_2 + ... + \delta_{mn}x_m. \end{cases} \quad (2)$$

Параметры приведенной формы модели δ_{ij} могут быть оценены по методу наименьших квадратов. По этим параметрам затем можно рассчитать структурные коэффициенты модели a_{ij} и b_{ij}. Для возможности соответствия между параметрами структурной и приведенной формами необходимо выполнить условия идентификации. Следует отметить, что структурные формы модели бывают трех видов:

1. **Идентифицируемые**. Для того чтобы структурная форма модели была идентифицируема, необходимо чтобы каждое

уравнение системы было идентифицируемо. В этом случае число параметров структурной формы равно числу параметров приведенной формы.

2. **Неидентифицируемые**. Если хотя бы одно уравнение структурной формы неидентифицируемо, то вся модель считается неидентифицируемой. В этом случае число коэффициентов приведенной формы модели меньше, чем число коэффициентов структурной формы.

3. **Сверхидентифицируемые**. Модель сверхидентифицируема, если число приведенных коэффициентов больше числа структурных коэффициентов. В этом случае можно получить два и более значений одного структурного коэффициента на основе коэффициентов приведенной формы модели. В сверхидентифицируемой модели хотя бы одно уравнение сверхидентифицируемо, а остальные уравнения идентифицируемы.

Если обозначить число эндогенных переменных в *i*-том уравнении структурной формы модели через H, а число предопределенных переменных, которые содержатся в системе, но не входят в данное уравнение через D, то условие идентифицируемости модели может быть записано в виде следующего правила:

1. Если D+1 < H – уравнение неидентифицируемо.
2. Если D+1 = H – уравнение идентифицируемо.
3. Если D+1 > H – уравнение сверхидентифицируемо.

Данное правило является необходимым, но недостаточным условием для идентификации. Отметим в системе эндогенные и экзогенные переменные, отсутствующие в рассматриваемом уравнении, но присутствующие в системе. Из коэффициентов при этих переменных в других уравнениях составим матрицу. При этом, если переменная стоит в левой части уравнения, то коэффициент надо брать с обратным знаком. Если определитель полученной матрицы не равен нулю, а ранг не меньше, чем количество эндогенных переменных в системе без одного, то достаточное условие индетификации для данного уравнения выполнено.

Проверим каждое уравнение системы (1) на выполнение неоходимого и достаточного условия идентификации.

В первом уравнении три эндогенных переменных: y_1, y_2, y_3 (H=3). В нем отсутствуют экзогенные переменные x_3 и x_4 (D=2). Необходимое условие идентификации D+1=H выполнено.

Для проверки на достаточное условие составим матрицу из коэффициентов при переменных x_3 и x_4 (см. таблицу 5). В первом столбце таблицы показано, что коэффициенты при экзогенных переменных x_3 и x_4 взяты из уравнений 2 и 3 системы. Во втором уравнении эти переменные присутствуют и коэффициенты при них равны a_{23} и a_{24}, соответственно. В третьем уравнении эти переменные отсутствуют, т.е. коэффициенты при них равны нулю. Так как вторая строка матрицы состоит из нулей, определитель матрицы равен нулю. Значит, достаточное условие не выполнено, и первое уравнение нельзя считать идентифицируемым.

Таблица 5. Матрица, составленная из коэффициентов при переменных x_3 и x_4

Уравнения, из которых взяты коэффициенты при переменных	Переменные	
	x_3	x_4
2	a_{23}	a_{24}
3	0	0

Во втором уравнении две эндогенные переменные: y_1 и y_2 (H=2). В нем отсутствует экзогенная переменная x_1 (D=1). Необходимое условие идентификации D+1=H выполнено.

Для проверки на достаточное условие составим матрицу из коэффициентов при переменных y_3 и x_1, которые отсутствуют во втором уравнении (см. таблицу 6).

Таблица 6. Матрица, составленная из коэффициентов при переменных y_3 и x_1

Уравнения, из которых взяты коэффициенты при переменных	Переменные	
	y_3	x_1
1	b_{13}	a_{11}
3	–1	a_{31}

Определитель представленной в таблице 5 матрицы не равен нулю, а ранг матрицы равен 2. Значит, достаточное условие выполнено, и второе уравнение идентифицируемо.

В третьем уравнении три эндогенные переменные: y_1,y_2,y_3 (H=3). В нем отсутствует экзогенные переменные x_3 и x_4 (D=2). Необходимое условие идентификации D+1=H выполнено.

Проверим на достаточное условие третье уравнение составим матрицу из коэффициентов при переменных x_3 и x_4. Согласно таблице 7, определитель матрицы равен нулю. Значит, достаточное условие не выполнено, уравнение неидентифицируемо.

Таблица 7. Матрица, составленная из коэффициентов при переменных x_3 и x_4

Уравнения, из которых взяты коэффициенты при переменных	Переменные	
	x_3	x_4
1	0	0
2	a_{23}	a_{24}

При оценивании коэффициентов структурной модели используется ряд методов. Рассмотрим косвенный метод наименьших квадратов (КМНК), который применяется в случае точно идентифицируемой структурной модели.

Пусть имеется следующая идентифицируемая модель, содержащая две эндогенные и две экзогенные переменные:

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + \varepsilon_1, \\ y_2 = b_{21} y_1 + a_{22} x_2 + \varepsilon_2. \end{cases}$$

Имеется следующий набор входных данных, представленный в таблице 8.

Таблица 8. Фактические данные для построения модели

n	y_1	y_2	x_1	x_2
1	33,0	37,1	3	11
2	45,9	49,3	7	16
3	42,2	41,6	7	9
4	51,4	45,9	10	9
5	49,0	37,4	10	1
6	49,3	52,3	8	16
Сумма	270,8	263,6	45	62
Среднее значение	45,133	43,930	7,500	10,333

Структурную модель преобразуем в приведенную форму модели.

$$\begin{cases} y_1 = \delta_{11} x_1 + \delta_{12} x_2 + u_1, \\ y_2 = \delta_{21} x_1 + \delta_{22} x_2 + u_2, \end{cases}$$

где u_1 и u_2 – случайные ошибки.

Для каждого уравнения приведенной формы при расчете коэффициентов δ можно применить МНК.

Для упрощения расчетов можно работать с отклонениями от средних уровней $Y=y-\bar{y}$ и $X=x-\bar{x}$ ($\bar{y}$ и $\bar{x}$ – средние значения). Преобразованные таким образом данные таблицы 8 сведены в таблицу 9. Здесь же показаны промежуточные расчеты, необходимые для определения коэффициентов δ_{lk}.

Для нахождения коэффициентов δ_{lk} первого приведенного уравнения можно использовать следующую систему нормальных уравнений:

$$\begin{cases}\sum Y_1 X_1 = \delta_{11}\sum X_1^2 + \delta_{12}\sum X_1 X_2,\\ \sum Y_1 X_2 = \delta_{11}\sum X_1 X_2 + \delta_{12}\sum X_2^2.\end{cases}$$

Таблица 9. Преобразованные данные для построения приведенной формы модели

n	Y_1	Y_2	X_1	X_2	$Y_1 \cdot X_1$	X_1^2	$X_1 \cdot X_2$	$Y_1 \cdot X_2$	$Y_2 \cdot X_1$	$Y_2 \cdot X_2$	X_2^2
1	-12,13	-6,784	-4,500	0,667	54,599	20,250	-3,002	-8,093	30,528	-4,525	0,445
2	0,767	5,329	-0,500	5,667	-0,383	0,250	-2,834	4,347	-2,664	30,198	32,115
3	-2,933	-2,308	-0,500	-1,333	1,467	0,250	0,667	3,910	1,154	3,077	1,777
4	6,267	1,969	2,500	-1,333	15,668	6,250	-3,333	-8,354	4,922	-2,625	1,777
5	3,867	-6,541	2,500	-9,333	9,667	6,250	-23,33	-36,09	-16,35	61,048	87,105
6	4,167	8,337	0,500	5,667	2,084	0,250	2,834	23,614	4,168	47,244	32,115
Сумма	0,002	0,001	0,000	0,002	83,102	33,500	-29,00	-20,67	21,755	134,41	155,33

Подставляя рассчитанные в таблице 5 значения сумм, получим

$$\begin{cases}83{,}102 = 33{,}5\cdot\delta_{11} - 29{,}001\cdot\delta_{12},\\ -20{,}667 = -29{,}001\cdot\delta_{11} + 155{,}33\cdot\delta_{12}.\end{cases}$$

Решение этих уравнений дает значения $\delta_{11} = 2{,}822$ и $\delta_{12} = 0{,}394$. Первое уравнение приведенной формы модели примет вид

$$y_1 = 2{,}822\cdot x_1 + 0{,}394\cdot x_2 + u_1.$$

Для нахождения коэффициентов δ_{2k} второго приведенного уравнения можно использовать следующую систему нормальных уравнений:

$$\begin{cases}\sum Y_2 X_1 = \delta_{21}\sum X_1^2 + \delta_{22}\sum X_1 X_2,\\ \sum Y_2 X_2 = \delta_{21}\sum X_1 X_2 + \delta_{22}\sum X_2^2.\end{cases}$$

Подставляя рассчитанные в таблице 9 значения сумм, получим

$$\begin{cases}21{,}755 = 33{,}5\cdot\delta_{21} - 29{,}001\cdot\delta_{22},\\ 134{,}417 = -29{,}001\cdot\delta_{21} + 155{,}33\cdot\delta_{22}.\end{cases}$$

Решение этих уравнений дает значения $\delta_{21} = 1{,}668$ и $\delta_{22} = 1{,}177$. Второе уравнение приведенной формы модели примет вид

$$y_2 = 1{,}668 \cdot x_1 + 1{,}177 \cdot x_2 + u_2 .$$

Для перехода от приведенной формы к структурной форме модели найдем x_2 из второго уравнения приведенной формы модели

$$x_2 = \left(y_2 - 1{,}668 \cdot x_1\right)/1{,}177 .$$

Подставим это выражение в первое уравнение приведенной модели, найдем структурное уравнение

$$y_1 = 2{,}822x_1 + 0{,}394 \cdot \left(y_2 - 1{,}668x_1\right)/1{,}177 =$$

$$= 2{,}822x_1 + 0{,}335y_2 - 0{,}558x_1 = 0{,}335y_2 + 2{,}264x_1 .$$

Таким образом, $b_{12} = 0{,}335$; $a_{11} = 2{,}264$.

Найдем x_1 из первого уравнения приведенной формы модели

$$x_1 = \left(y_1 - 0{,}394 \cdot x_2\right)/2{,}822 .$$

Подставим это выражение во второе уравнение приведенной модели, найдем структурное уравнение

$$y_2 = 1{,}177x_2 + 1{,}668 \cdot \left(y_1 - 0{,}394x_2\right)/2{,}822 =$$

$$= 1{,}177x_2 + 0{,}591y_1 - 0{,}233x_2 = 0{,}591y_1 + 0{,}944x_2 .$$

Таким образом, $b_{21} = 0{,}591$; $a_{22} = 0{,}944$.

Свободные члены структурной формы находим из уравнений

$$a_{01} = \overline{y}_1 - b_{12}\overline{y}_2 - a_{11}\overline{x}_1 = 45{,}133 - 0{,}335 \cdot 43{,}93 - 2{,}264 \cdot 7{,}5 = 13{,}436$$

$$a_{02} = \overline{y}_2 - b_{21}\overline{y}_1 - a_{22}\overline{x}_2 = 43{,}93 - 0{,}591 \cdot 45{,}133 - 0{,}944 \cdot 10{,}333 = 7{,}502$$

Окончательный вид структурной модели

$$\begin{cases} y_1 = a_{01} + b_{12}y_2 + a_{11}x_1 + \varepsilon_1 = 13{,}436 + 0{,}335y_2 + 2{,}264x_1 + \varepsilon_1, \\ y_2 = a_{02} + b_{21}y_1 + a_{22}x_2 + \varepsilon_2 = 7{,}502 + 0{,}591y_1 + 0{,}944x_2 + \varepsilon_2. \end{cases}$$

Варианты заданий

Проверить необходимые и достаточные условия структурной формы модели:

Таблица 10. Системы одновременных уравнений

Вариант 1	Вариант 2
$\begin{cases} y_1 = b_{13}y_3 + a_{11}x_1 + a_{13}x_3, \\ y_2 = b_{21}y_1 + b_{23}y_2 + a_{22}x_2 + a_{24}x_4, \\ y_3 = b_{32}y_2 + a_{31}x_1 + a_{33}x_3. \end{cases}$	$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 + a_{13}x_3, \\ y_2 = b_{21}y_1 + b_{22}y_2 + a_{23}x_3, \\ y_3 = b_{33}y_3 + b_{32}y_2 + a_{32}x_2. \end{cases}$
Вариант 3	Вариант 4
$\begin{cases} y_1 = b_{11}y_2 + a_{11}x_1 + a_{22}x_2, \\ y_2 = b_{21}y_1 + b_{23}y_2 + a_{13}x_3, \\ y_3 = b_{13}y_3 + a_{31}x_1 + a_{33}x_3 + a_{23}x_4. \end{cases}$	$\begin{cases} y_1 = b_{13}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + b_{23}y_2 + a_{23}x_3, \\ y_3 = b_{32}y_3 + a_{31}x_1 + a_{33}x_3 + a_{34}x_4. \end{cases}$
Вариант 5	Вариант 6
$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + b_{33}y_2 + a_{32}x_2. \end{cases}$	$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \\ y_4 = b_{41}y_4 + b_{42}y_3 + a_{41}x_1 + a_{42}x_2. \end{cases}$
Вариант 7	Вариант 8
$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + b_{22}y_4 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \\ y_4 = b_{42}y_3 + a_{42}x_2. \end{cases}$	$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_4 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \\ y_4 = b_{41}y_4 + b_{42}y_3 + a_{43}x_3. \end{cases}$
Вариант 9	Вариант 10
$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2, \\ y_2 = b_{21}y_1 + b_{22}y_3 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \\ y_4 = b_{42}y_3 + a_{42}x_2. \end{cases}$	$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{13}x_3, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4, \\ y_3 = b_{31}y_4 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2. \\ y_4 = b_{41}y_4 + b_{42}y_3 + a_{43}x_3. \end{cases}$

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. Необходимое и достаточное условие идентификации системы.
4. Результаты проверки системы по варианту.
5. Выводы по работе.

2.8. ОЦЕНКА ПАРАМЕТРОВ МЕТОДОМ МАКСИМАЛЬНОГО ПРАВДОПОДОБИЯ

Цель работы – приобретение практических навыков оценки неизвестных параметров методом максимального правдоподобия при заданном наборе данных.

Метод наименьших квадратов позволяет получить оценки параметров уравнения регрессии, обладающие «хорошими» свойствами, если выполняются достаточно сильные условия теоремы Гаусса-Маркова. Однако на практике, в случае реальных данных эти условия часто не выполняются. В этом случае обычно применяются методы, которые могут работать при более слабых предположениях. Одним из таких методов является метод максимального правдоподобия.

Отправной точкой получения оценок параметров случайной величины является предположение о том, что распределение наблюдаемого явления известно, за исключением конечного числа неизвестных параметров. Эти параметры оцениваются такими значениями, которые придают наблюдаемым значениям наивысшую вероятность, наивысшее правдоподобие. Таким образом, метод максимального правдоподобия обеспечивает способ оценивания совокупности параметров, характеризующих распределение, если мы знаем или предполагаем, что знаем общий вид этого распределения.

Предположим, что имеется функция плотности вероятностей $f\ (y_i \mid x_i;\ \theta)$ где y – зависимая переменная, x – независимая переменная, θ – К-мерный вектор неизвестных параметров и

предположим, что наблюдения взаимно независимы. В данной ситуации функция совместной плотности распределения вероятностей выборки $y_1, y_2 \ldots y_N$ (условная по X = ($x_1, x_2, \ldots x_N$)) задается как:

$$f\left(y_1,...,y_N \mid X;\theta\right)=\prod_{i=1}^{N} f\left(y_i \mid x_i;\theta\right).$$

Тогда функция правдоподобия для имеющейся выборки задается в виде

$$L\left(\theta \mid y, X\right)=\prod_{i=1}^{N} L_i\left(\theta \mid y_i, x_i\right)=\prod_{i=1}^{N} f\left(y_i \mid x_i;\theta\right),$$

и является функцией от вектора неизвестных параметров θ. Для некоторых целей удобно использовать так называемые вклады правдоподобия, обозначаемые как $L_i(\theta \mid y_i,x_i)$, которые отражают, какой вклад в функцию правдоподобия вносит наблюдение i. ММП-оценка $\hat{\theta}$ для вектора неизвестных параметров θ есть решение

$$\max_{\theta} \log L\left(\theta\right)=\max_{\theta} \sum_{i=1}^{N} \log L_i\left(\theta\right),$$

где log $L(\theta)$ – это логарифмическая функция правдоподобия. Следует отметить, что для простоты записи были исключены другие аргументы.

Отметим, что оценку по методу максимального правдоподобия можно определить аналитически только в частных случаях. В общем же, требуется численная оптимизация. Однако для многих стандартных моделей в современных математических пакетах имеются эффективные алгоритмы.

Этапы выполнения работы

1. Сгенерировать 50 случайных чисел с параметрами математического ожидания и среднеквадратического отклонения заданными по варианту. Варианты задания в конце лабораторной работы.

2. Вычислить логарифмические вклады правдоподобия, а затем просуммировав их найти значение логарифмической функции правдоподобия.

3. Используя алгоритмы оптимизации, предлагаемые пакетами MS Excel и MATLAB для первого пакета соответственно максимизировать целевую функцию, а для второго минимизировать.

4. Получить оценки математического ожидания и среднеквадратического отклонения по методу максимального правдоподобия.

5. Сравнить полученные значения с теми, что были заданы по варианту.

6. Повторить пункты 1–5 для выборок размером 150 и 500 чисел. Сравнить полученные значения с теми, что были заданы по варианту.

7. Сгенерировать случайную величину, имеющую равномерное распределение с параметрами a и b равными математическому ожиданию и среднеквадратическому отклонению соответственно заданными по вашему варианту. Провести поиск оценок по методу максимального правдоподобия по пунктам 1–5. Сделать соответствующие выводы.

Решение задачи в пакете MS Excel

Генерируем 50 нормально распределенных случайных чисел с параметрами математического ожидания 12 и среднеквадратического отклонения 3, как это показано на рис. 40. Данные-Анализ данных-Генерация случайных чисел.

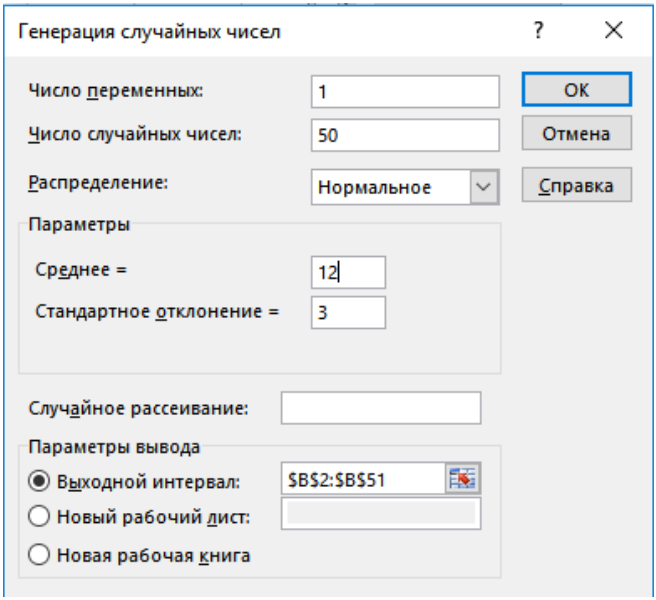


Рис. 40. Генерация случайных чисел

Получаем случайные числа, соответствующие их порядковому номеру как это показано на рис. 41.

	A	B
1	№ числа	Случайные числа
2	1	10,77764253
3	2	8,192250364
4	3	12,80392056
5	4	10,9287528
6	5	16,7775211
7	6	9,635277396
8	7	8,64035613
9	8	10,48832431
10	9	9,198638084

Рис. 41. Соответствие случайного числа порядковому номеру

Теперь, для расчета логарифмической функции правдоподобия нам необходимо задать начальные значения параметров математического ожидания и дисперсии, зададим их равными 0 и 1 соответственно, и занесем в ячейки E2 и E3 соответственно, как это показано на рис. 42.

	D	E
1	Параметры	
	M	0
	D = σ^2	1

Рис. 42. Ввод параметров математического ожидания и дисперсии

Таким образом, у нас есть все необходимое, для нахождения вкладов логарифмической функции правдоподобия:

$$l\left(x_1,...,x_n;\mu,\sigma^2\right)=\sum_{i=1}^{N}\left[-\frac{1}{2}\ln\left(2\pi\right)-\frac{1}{2}\ln\left(\sigma^2\right)-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right].$$

В ячейку C2 записываем формулу как это показано на рис. 43:

=-0,5*LN(2*3,1416)-0,5*LN(2*$E$3)-((B2-$E$2)^2)/(2*$E$3)

Рис. 43. Ввод формулы для расчета логарифмической функции правдоподобия

Здесь $E$2 и $E$3 – указатели на ячейки, где находятся значения параметров математического ожидания и дисперсии соответственно пример рассчитанных вкладов логарифмической функции правдоподобия показан на рис. 44.

Теперь находим искомую логарифмическую функцию правдоподобия суммируя ячейки C2-C51 используя функцию СУММ(C2:C51) и запишем результат в ячейку C52. Далее максимизируем логарифмическую функцию правдоподобия средствами MS EXCEL «Поиск решения». Если эта надстройка не активирована, то необходимо провести следующие действия

пройти по цепочке: Файл-Параметры-Надстройки. Появится окно как на рис. 45.

	A	B	C
1	№ числа	Случайные числа	Логарифмическая функция правдоподобия
2	1	10,77764253	-59,34430256
3	2	8,192250364	-34,82199631
4	3	12,80392056	-83,23570411
5	4	10,9287528	-60,9843322
6	5	16,7775211	-142,0081204
7	6	9,635277396	-47,68479854
8	7	8,64035613	-38,59339032
9	8	10,48832431	-56,2679867
10	9	9,198638084	-43,5729846

Рис. 44. Вычисленные значения вкладов логарифмической функции правдоподобия

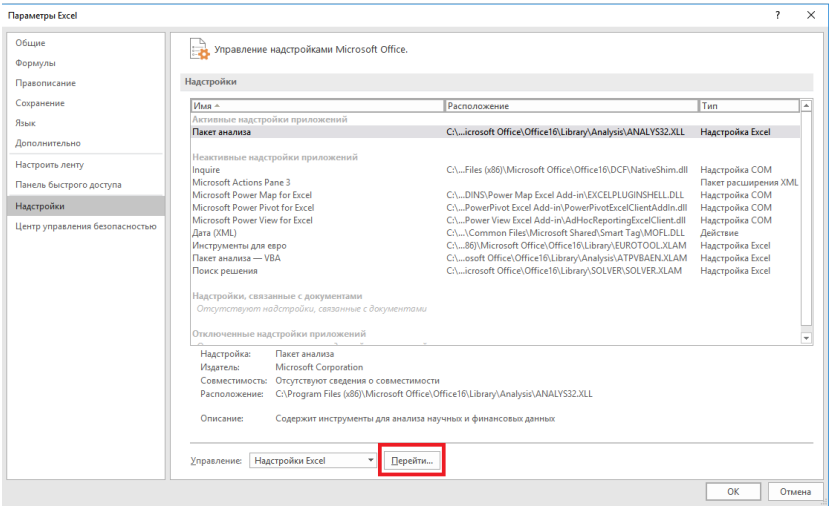


Рис. 45. Окно активации надстроек

В открывшемся окне нажать «Перейти» и поставить галочку в поле поиск решения, как это показано на рис. 46.

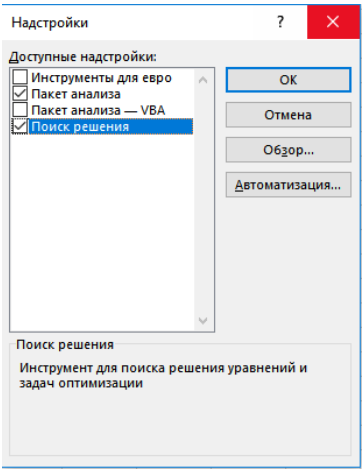


Рис. 46. Активация надстройки «Поиск решения»

Теперь заходим во вкладку «Данные», выбираем поиск решения. Появляется окно, показанное на рис. 47:

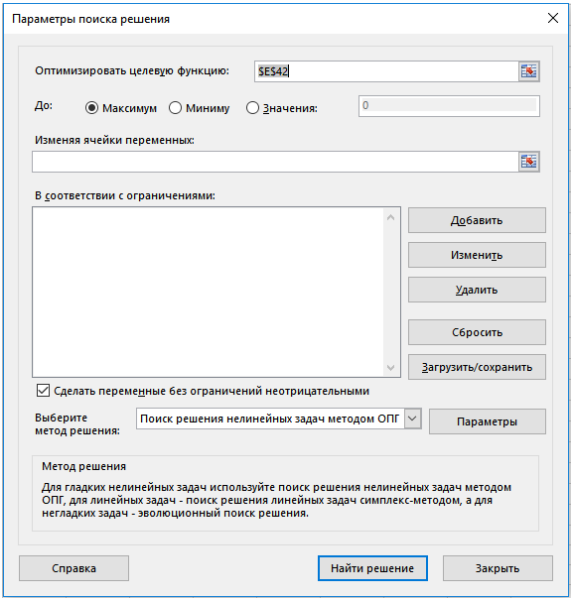


Рис. 47. Работа блока «Поиск решения»

В поле «Оптимизировать целевую функцию» мы выбираем ячейку, где у нас содержится значение логарифмической функции правдоподобия (в нашем случае С52) и задаем параметры следующим образом:

1. В поле «До:» ставим указатель в графе «Максимум».
2. В поле «Изменяя ячейки переменных:» мы выбираем наши начальные значения математического ожидания и дисперсии (Е2 и Е3) соответственно.

3. В поле «В соответствии с ограничениями:» выбираем «Добавить» и указываем ограничение на дисперсию как показано на рис. 48.

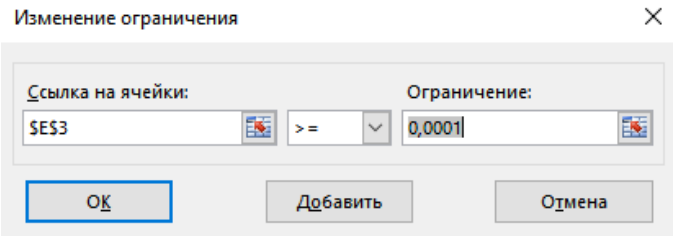


Рис. 48. Ввод ограничений

По завершении ввода параметров нажимаем «Найти решение». В случае если все введено правильно и решение существует EXCEL выдаст следующее окно как на рис. 49. Нажимаем «ОК».

После проведения данных операция, можно заметить, что начальные значения математического ожидания и дисперсии, указанные в ячейках Е2 и Е3 соответственно изменились на значения, близкие к тем, которые мы задавали при генерации случайных чисел, как это показано на рис. 50.

Теперь найдем средствами MS EXCEL выборочное среднее (математическое ожидание) и дисперсию выборки. Используем формулы =СРЗНАЧ (B2:B51) и =ДИСП.Г (B2:B51), получаем значения 11,3 и 9,16 соответственно.

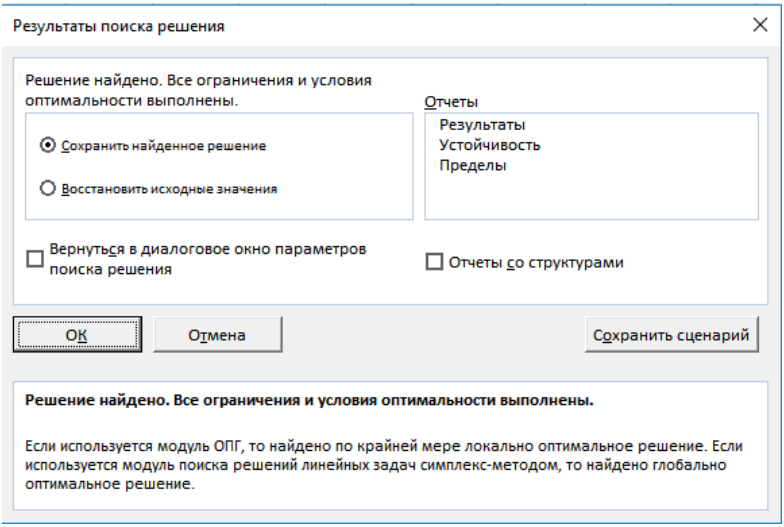


Рис. 49. Окно результатов поиска решения

	D	E
	Параметры	
	M	11,33
	D = σ^2	9,16

Рис. 50. Оценки математического ожидания и дисперсии по методу максимального правдоподобия

Решение задачи в пакете MATLAB

```
clear all
close all
clc

X = 12 + 3 * randn(50);
[Parameters, LLF] = get_norm_theta_ML(X);
```

```
function [Parameters, LLF] = get_norm_theta_ML(X)
% Вычисление отрицательной логарифмической функции правдоподобия
% по выборке X
fun = @(theta) get_Neg_norm_LLF(theta, X);
% Задание параметров для работы оптимизационной функции
x0 = [0, 1]';
A = [];
b = [];
Aeq = [];
beq = [];
lb = [-inf, 10^(-6)]';
ub = [inf, inf]';
nonlcon = [];
options = optimset('Algorithm', 'interior-point', 'Display', 'off');
% Вычисление параметров
[Parameters, NegLLF] = fmincon(fun,x0,A,b,Aeq,beq,lb,ub,nonlcon,options);
LLF = -NegLLF;

function [l] = get_Neg_norm_LLF(theta, X)
%
% Вычисление отрицательной логарифмической функции правдоподобия
%
n = length(X);
l = 0;
for i = 1:n
    l = l + (-0.5 * log(2 * pi) - 0.5 * log(theta(2)) - 0.5 * (X(i) - theta(1))^2 / theta(2));
end
l = -l;
```

Варианты заданий

Таблица 11. Параметры μ и σ

Номер варианта	μ	σ	Номер варианта	μ	σ
1	3	5	11	7	3
2	5	1	12	1	5
3	2	4	13	5	2
4	8	3	14	12	5
5	6	2	15	1	5
6	2	4	16	4	4
7	5	3	17	15	2
8	4	2	18	13	5
9	1	3	19	14	9
10	5	5	20	16	7

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. Параметры своего варианта.
4. Результаты вычисления параметров по методу максимального правдоподобия в пакете MS Excel.
5. Результаты вычисления параметров по методу максимального правдоподобия в пакете MATLAB.
6. Найденная величина ошибки между истинными значениями параметров и параметрами, найденными по методу максимального правдоподобия.
7. Выводы по работе.

2.9. МОДЕЛИРОВАНИЕ И АНАЛИЗ ВРЕМЕННОГО РЯДА

Цель работы – ознакомиться с основными моделями авторегрессии, получить навыки их моделирования в современных математических пакетах.

Временной ряд $y_{t1}, y_{t2}, ... y_T$, $y_{t_i} \in R$ (R – множество действительных чисел) это совокупность значений какого-либо признака измеренных через постоянные временные интервалы. Например, это может быть средняя цена акций компании Intel на бирже или месячный объем производства телевизоров. Пример временного ряда показан на рис. 51.

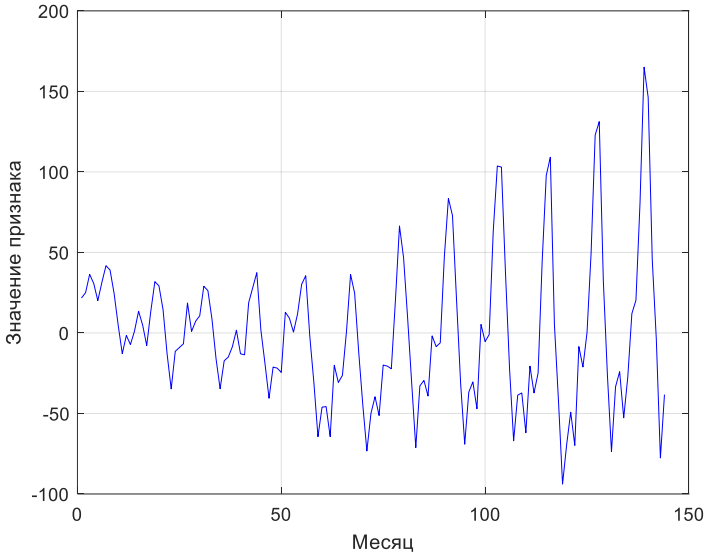


Рис. 51. Пример временного ряда

Задача прогнозирования временного ряда состоит в том, чтобы найти функцию f_T, такую, чтобы

$$y_{T+h} \approx f_T(y_{t_1}, y_{t_2}, ..., y_T, h) \equiv \hat{y}_{T+h|T},$$

где $h \in (1,2,\ldots,H)$, H – горизонт прогнозирования.

Особенностью работы с временными рядами является то, что при их прогнозировании мы надеемся, что значения ряда в предыдущие моменты времени несут информацию о его поведении в будущем, в отличии от классических задач обработки выборок, где полагается, что выборки простые, независимые и одинаково распределенные.

Компоненты временных рядов:

Тренд – плавное долгосрочное изменение уровня ряда.

Сезонность – циклические изменения уровня ряда с постоянным периодом.

Цикл – изменения уровня ряда с переменным периодом (экономические циклы – подъем/спад, циклы, связанные с популяцией животных, периоды солнечной активности).

Ошибка – непрогнозируемая случайная компонента ряда.

В качестве примера на рис. 52 приведены временной ряд с трендом и сезонностью.

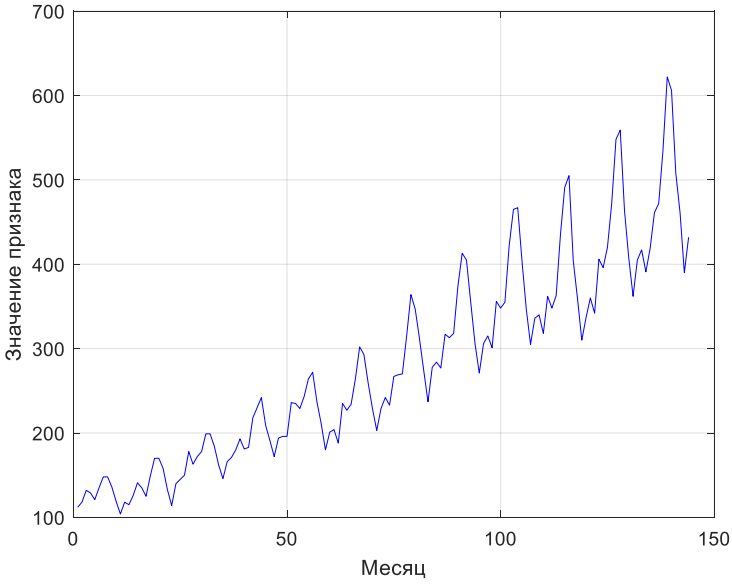


Рис. 52. Пример временного ряда с трендом и сезонностью

Автокорреляция – корреляция ряда с самим собой, но сдвинутая на заданное количество отсчетов. Количество отсчетов, на которое мы сдвигаем ряд, прежде чем вычислить автокорреляцию называется лагом автокорреляции. Формула для коэффициента автокорреляции временного ряда выглядит следующим образом:

$$r_t = \frac{\sum_{t=\tau+1}^{n} \left(y_t - \overline{y}_{1\tau}\right) * \left(y_{t-\tau} - \overline{y}_{2\tau}\right)}{\sqrt{\sum_{t=\tau+1}^{n} \left(y_t - \overline{y}_{1\tau}\right)^2 * \sum_{t=\tau+1}^{n} \left(y_{t-\tau} - \overline{y}_{2\tau}\right)^2}},$$

где τ – величина сдвига – лаг, определяет порядок коэффициента автокорреляции.

График автокорреляции называемый коррелограммой для временного ряда, показанного на рисунке 2 приведен на рис. 53.

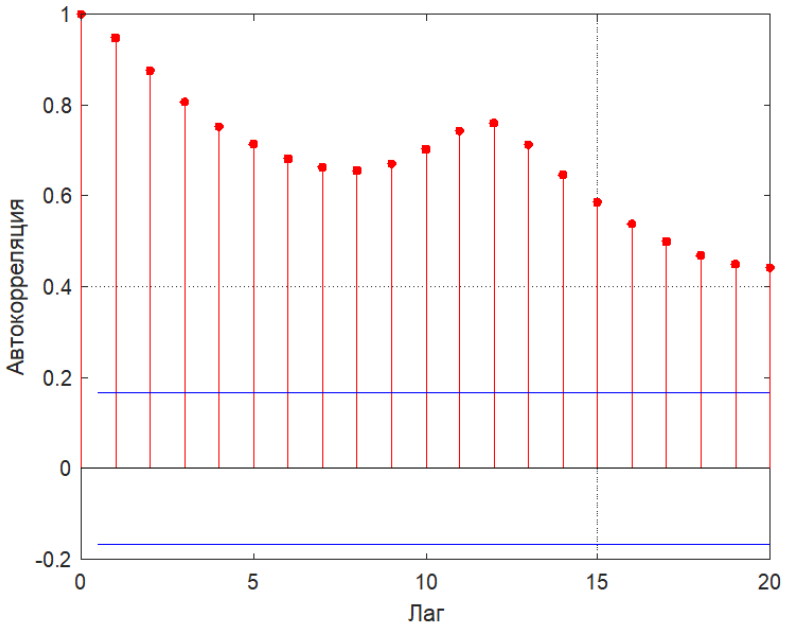


Рис. 53. Коррелограмма временного ряда

Временной ряд $y_{t1}, y_{t2},\ldots,y_T,\ y_{t_i} \in R$ называется стационарным если $\forall s \in \mathbb{Z}$ распределение $y_t,\ldots y_{t+s}$ не зависит от t, т.е. его свойства не зависят от времени. Такой ряд называется стационарным в узком смысле (сильно стационарным).

Следует отметить, что если во временном ряду присутствует тренд или сезонность, то данный ряд будет нестационарным, однако если во временном ряду есть цикл, то он не всегда будет нестационарным.

Если нам приходится работать с нестационарными рядами, существует набор преобразований, проведя которые можно привести ряд к стационарному. К этим преобразованиям относят:

1. Логарифмирование.
2. Дифференцирование ряда т.е. переход к попарным разностям его соседних значений.
3. Сезонное дифференцирование ряда т.е. переход к попарным разностям его значений в соседних сезонах.

Существует большое количество моделей прогнозирования временных рядов это например:

1. Регрессионные модели.
2. Авторегрессионные модели.
3. Модели экспоненциального сглаживания.
4. Модели прогнозирования на нейронных сетях.
5. Модели прогнозирования, основанные на цепях Маркова и т.д.

В рамках данной работы остановимся на авторегрессионных моделях как на наиболее большом и успешно применяемом на практике классе моделей. Для удобства дальнейших выкладок введем разностный оператор D, который будет обозначать:

$$Dy_t = y_{t-1}.$$

Рассмотрим простую модель авторегрессии AR порядка p ($AR(p)$) которая представляет собой линейную комбинацию p предыдущих значений ряда и шумовой компоненты

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t,$$

где y_t – стационарный ряд, ε_t – белый гауссов шум с нулевым средним и постоянной дисперсией. Другой способ записи данной модели выглядит следующим образом:

$$\phi(D)y_t = \left(1 - \phi_1 D - \phi_2 D^2 - \ldots - \phi_p B^p\right) y_t .$$

Рассмотрим простую модель скользящего среднего порядка q (*MA*(q)) которая представляет собой линейную комбинацию q последних значений шумовой компоненты.

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q},$$

где y_t и ε_t – обозначают тоже самое, что и в модели авторегрессии. Другой способ записи данной модели выглядит следующим образом:

$$y_t = \theta(D)\varepsilon_t = \left(1 + \theta_1 D - \theta_2 D^2 - \ldots - \theta_q B^q\right)\varepsilon_t .$$

Рассмотрим модель авторегрессии скользящего среднего *ARMA*(p,q) которой согласно теореме Вольда может быть описан любой стационарный ряд с любой точностью:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q},$$

другая форма записи которого выглядит следующим образом:

$$\phi(D)y_t = \theta(D)\varepsilon_t .$$

Рассмотрим модель ARIMA (Autoregressive integrated moving average). Временной ряд описывается моделью *ARIMA* (p, d, q), если ряд его разностей

$$\nabla^d y_t = (1 - D)^d \, y_t ,$$

описывается моделью *ARMA* (p, q).

$$\phi(D)\nabla^d y_t = \theta(D)\varepsilon_t .$$

Рассмотрим ряд, имеющий сезонную компоненту S.Пусть имеется модель *ARMA* (p, q) и добавим P авторегрессионных компонент:

$$+\phi_S y_{t-S} + \phi_{2S} y_{t-2S} + \ldots + \phi_{PS} y_{t-PS}$$

и Q компонент скользящего среднего:

$$+\theta_S \varepsilon_{t-S} + \theta_{2S} \varepsilon_{t-2S} + \ldots + \theta_{PS} \varepsilon_{t-QS} .$$

Такая модель называется *SARMA* (p,q)x(P,Q).

И рассмотрим модель *SAIRMA* (*p,d,q*)x(*P,D,Q*). это модель *SARMA* (*p,q*)x(*P,Q*) для ряда, к которому *d* раз было применено обычное дифференцирование и D раз – сезонное. Практики чаще всего называют ARIMA как раз эту модель.

Возникает вопрос – как при таком обилии параметров правильно их подобрать? Если все остальные параметры фиксированы, коэффициенты регрессии лучше всего подбирать методом наименьших квадратов, поскольку если шум белый, то МНК дает наилучшую оценку. Для нахождения коэффициентов θ, шум предварительно оценивается с помощью остатков авторегрессии. Параметры *d* и *D* подбирают таким образом, чтобы ряд стал стационарным, т.е. дифференцировать ряд можно любое количество раз. В случае сезонного ряда рекомендуется начать с сезонного дифференцирования, однако следует помнить, что чем меньше раз вы продифференцируете ряд, тем меньше будет итоговая дисперсия прогноза. Для параметров q,Q,p,P существуют различные информационные критерии, например критерий Акаике, чем меньше значение критерия, тем лучше модель.

Рассмотрим средства, которые предлагает пакет MATLAB и его расширение Econometrics Toolbox для возможностей анализа временных рядов и создания моделей авторегрессии.

Пусть имеется некий набор данных. Для начала его необходимо импортировать в MATLAB.

Для начала создадим новый скрипт и сохраним его в какой-либо папке. Далее нам нужно скачать необходимый набор данных и поместить его в папку, где уже находится ваш скрипт. Если все сделано правильно в графе «Current folder» появится ваш набор данных, как на рис. 54. После указанных действий этого необходимо 2 раза нажать на него левой кнопкой мыши, откроется окно, показанное на рис. 55. В этом окне будет предложено выделить необходимые для импорта в MATLAB данные. В нашем случае мы просто нажимаем «Import Selection».

После этого данные появятся в workpsace, появится наш набор данных. Данные готовы к работе, как это показано на рис. 56. После указанных действий для дальнейшей работы потребуется

встроенное в MATLAB приложение под названием Econometric modeler.

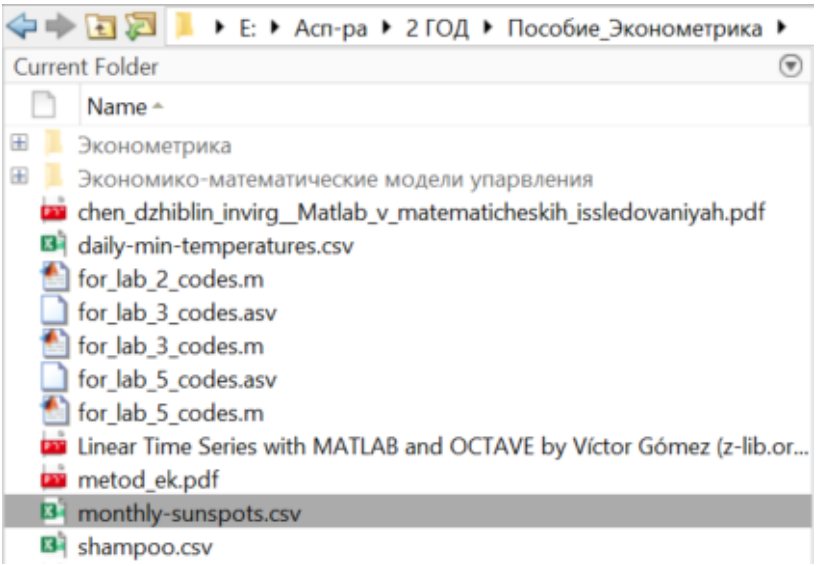


Рис. 54. Пример набора данных

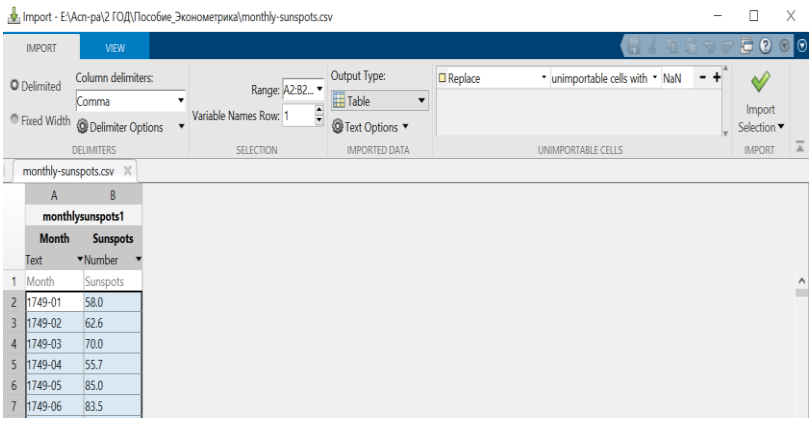


Рис. 55. Окно импорта данных в систему MATLAB

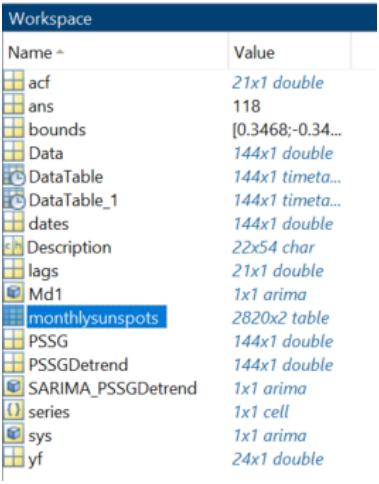


Рис. 56. Успешно импортированные данные

Чтобы запустить Econometric modeler в командной строке MATLAB напишем «econometricModeler». Откроется главное окно приложения, показанное на рис. 57. Красными цифрами обозначены основные элементы интерфейса.

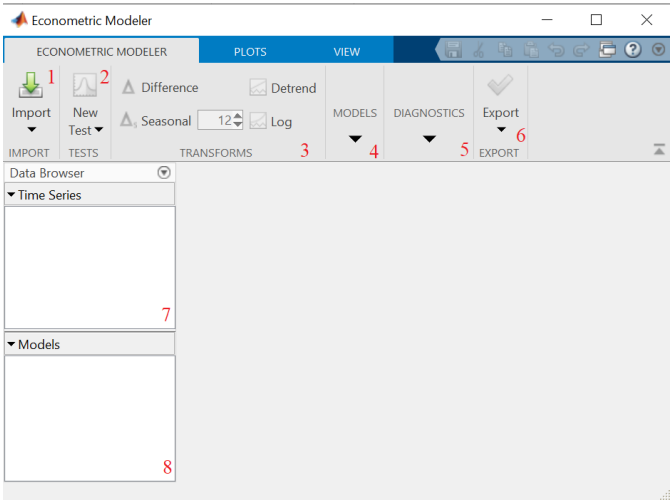


Рис. 57. Интерфейс приложения Econometric modeler

Перечислим элементы интерфейса:

1. Кнопка «Import» нажатие на которую позволяет загрузить в приложение исследуемый временной ряд.

2. Кнопка «New Test» при нажатии на данную кнопку будет предложен выбор основных тестов для временных рядов, например тест на стационарность.

3. Поле «Transforms» позволяет осуществлять преобразования, с помощью которых можно обеспечить стационарность ряда.

4. Поле «Models» позволяет выбрать и создать необходимую модель прогнозирования временного ряда.

5. Поле «Diagnostics» предоставляет возможности диагностирования полученной модели.

6. Кнопка «Export» при нажатии позволяет экспортировать полученную модель в workspace или экспортировать ее в качестве функции MATLAB.

7. Поле «Time Series» показывает загруженные временные ряды, а также ряды после преобразований.

8. Поле «Models» показывает все модели, которые были сгенерированы.

Дальнейшее знакомство с Econometric modeler продолжим в пункте работы «Пример выполнения в пакете MATLAB».

Этапы выполнения работы

1. Получить у преподавателя набор данных, с которым предстоит работать.

2. Загрузить набор данных в используемый пакет.

3. Построить график временного ряда.

4. Визуально попробовать определить является ли ряд стационарным.

5. Проверить гипотезу принятую в пункте 4 с помощью какого-либо теста на стационарность.

6. В случае, если ряд не стационарен провести необходимые преобразования, и обеспечить стационарность (в случае если в пункте 5 вы подтвердили гипотезу о стационарности ряда, этот пункт пропускается).

7. Построить три класса моделей *ARMA*, *ARIMA* и *SARIMA*. С помощью информационных критериев подобрать наилучшую модель.

8. Спрогнозировать с помощью полученной модели дальнейшее поведение временного ряда на длину N/4, где N – общая длина данных по оси OX.

9. Построить график спрогнозированного временного ряда.

Решение задачи в пакете MATLAB

1. Загрузим набор входных данных и построим график временного ряда. Для этого импортируем в MATLAB набор данных и запустим Econometric modeler. После этого импортируем из workspace MATLAB-а временной ряд. Для этого в главном окне Econometric modeler нажимаем кнопку «Import» и выбираем вкладку «Import From Workspace». Поскольку в описываемом здесь случае набор данных назывался – PSSG, в открывшемся меню выбора данных для импорта выбираем соответствующую переменную ставим напротив нее галочку и нажимаем кнопку «Import» как это показано на рис. 58.

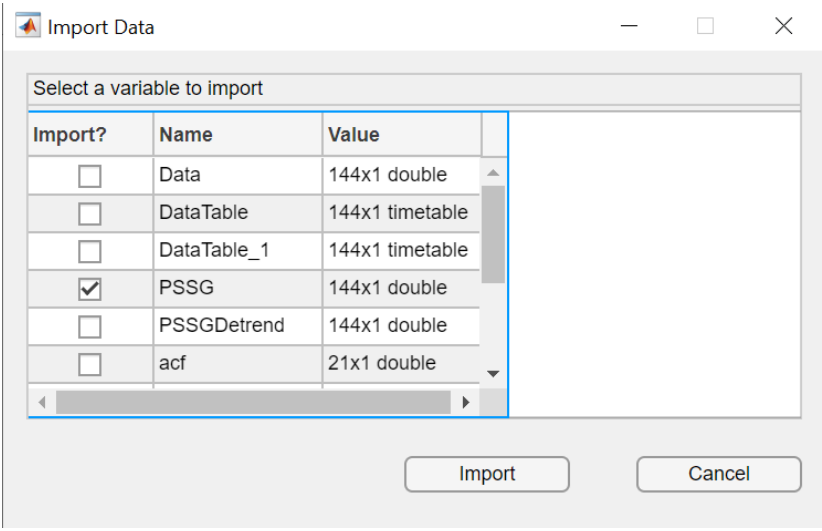


Рис. 58. Выбор необходимой переменной

После выбора переменной программа автоматически построит график временного ряда, как это показано на рис. 59.

2. Визуально оцениваем стационарность. Поскольку ряд имеет явно выраженную трендовую составляющую, делаем вывод о нестационарности ряда. Проверим это встроенным тестом. Для этого нажимаем кнопку «New Test». В появившемся меню выбираем «Leybourne-McCabe Test» в появившемся окне выбираем переменную и нажимаем «Run Test». Данный тест в качестве нулевой гипотезе принимает стационарность ряда, в нашем случае получаем логическое false, т.е. гипотеза о стационарности ряда отвергается.

3. Поскольку ряд нестационарен, имеет явно выраженный тренд и сезонную компоненту, можно сразу построить модель SARIMA.

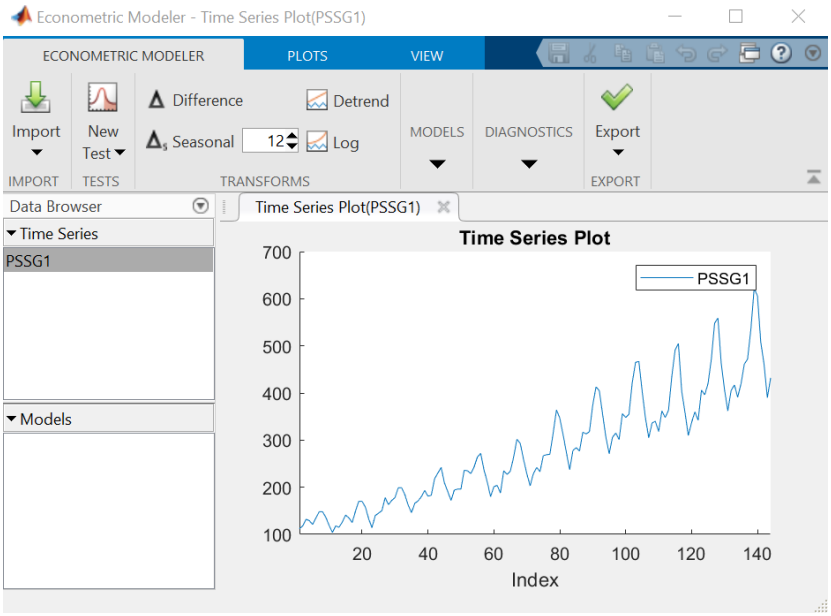


Рис. 59. График временного ряда

Данная модель учитывает тренд и сезонную составляющую, как было показано в теоретической части данной работы. Для того, чтобы построить модель SARIMA мы выбираем необходимую переменную и во вкладке «Models» находим модель SARIMA. Нажимаем на нее. Заметим, что сезонная составляющая ряда имеет период в 12 месяцев. Поэтому в открывшемся окне в поле «Period» запишем цифру 12. В поля «Degree of integration», «Autoregressive Order» и «Moving Average Order» запишем соответственно цифры 1,2 и 2, а затем нажимаем кнопку «Estimate». В появившемся окне появится временной ряд и результаты работы модели SARIMA, как это показано на рис. 60.

Мы также можем посмотреть параметры модели, в графическом окне справа сверху рис. 60, а также посмотреть значения информационных критериев в окне справа снизу на рис. 60. Готовую модель мы можем импортировать обратно в Workspace MATLAB. Для этого нажмем кнопку «Export» и выберем пункт «Export Variables». Выбираем модель которую мы хотим экспортировать и нажимаем «Export». После этого можем закрывать Econometric modeler и продолжить работу непосредственно в MATLAB.

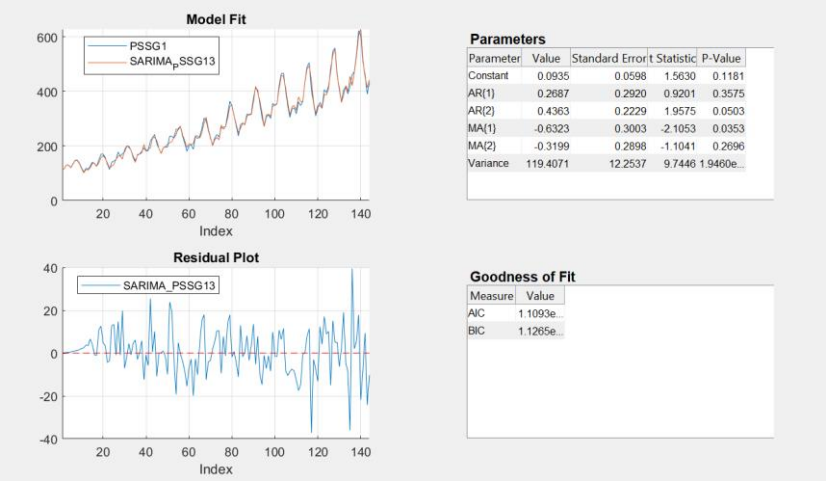


Рис. 60. Результаты работы программы

4. Теперь можно построить график прогноза используя следующий код MATLAB:

```
% Запишем в переменную построенную модель
sys = SARIMA_PSSG13;
% Подгоняем модель к данным
Md1 = estimate(sys,PSSG);
% Прогнозируем значения
yf = forecast(Md1,24,'Y0',PSSG);
% Построение графика прогноза
figure()
plot(1:length(PSSG),PSSG,'b', ...
    length(PSSG):length(PSSG) ...
    +length(yf),[PSSG(end);yf],'r')
legend('Наблюдаемые
значения','Прогнозируемые значения')
grid on
xlabel('Месяц')
ylabel('Значение признака')
```

Требования к содержанию отчета

1. Титульный лист.
2. Цель работы.
3. График временного ряда.
4. Предварительные выводы о стационарности временного ряда.
5. Проверка временного ряда на стационарность при помощи статистического теста.
6. Результаты построения модели прогнозирования.
7. Графики прогноза временного ряда по построенной модели.
8. Выводы по работе.

ЗАКЛЮЧЕНИЕ

В пособии подробно были рассмотрены статистические методы моделирования, обработки и анализа данных для экономических процессов.

Рассмотрены теоретические и практические аспекты применения методов моделирования и статистической обработки данных, а именно регрессия, анализ и прогнозирование временных рядов, проверка статистических гипотез и т.д. Описаны методы и алгоритмы анализа статистических данных.

Инструменты моделирования и обработки данных, рассмотренные в данном учебно-методическом пособии, имеют важное теоретическое и практическое значение при изучении моделей, связанных с эконометрическими процессами, с их прогнозированием и выбором адекватных.

Для дальнейшего изучения вопросов применения статистических методов моделирования, обработки и анализа данных в экономической сфере рекомендуется пользоваться современными ссылками на учебную и научную литературу, приведенную в списке использованных источников.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Доугерти К. Введение в эконометрику: Пер. с англ. – М.: ИНФРА-М, 2003. – 402 с.

2. Кремер Н. Ш., Путко Б. А. Эконометрика: Учебник для вузов / Под ред. проф. Н. Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2010. – 311 с.

3. Гармаш, А. Н. Экономико-математические методы и прикладные модели: учебник для бакалавриата и магистратуры / А. Н. Гармаш, И. В. Орлова, В. В. Федосеев; под ред. В. В. Федосеева. – 4-е изд., перераб. и доп. – М. : Издательство Юрайт, 2017. – 328 с

4. Математика для экономистов: от арифметики до эконометрики: учеб.-справ. пособие для бакалавров / Н. Ш. Кремер, Б. А. Путко, И. М. Тришин, М. Н. Фридман ; под ред. Н. Ш. Кремера. – 3-е изд., перераб. и доп. – М. : Юрайт, 2012. – 685 с.

5. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2-х т. – Т. 1. Айвазян С. А., Мхитарян В. С. Теория вероятностей и прикладная статистика. – М: ЮНИТИ-ДАНА, 2006. – 656 с.

6. Прикладная статистика. Основы эконометрики: Учебник для вузов: В 2-х т. – Т. 2. Айвазян С. А. Основы эконометрики. – М: ЮНИТИ-ДАНА, 2007. – 432 с.

7. Антохонова, И. В. Методы прогнозирования социально-экономических процессов: учеб. пособие для вузов / И. В. Антохонова. – 2-е изд., испр. и доп. – М. : Издательство Юрайт, 2018. – 213 с.

8. Бабешко Л. О. Основы эконометрического моделирования: учеб. пособие / Л. О. Бабешко. – Изд. 4-е. – М.: КомКнига, 2010. – 428 с.

9. Wooldridge J. Introductory Econometrics: A Modern Approach 7th edition. South-Western College Publishers. 2018. P. 816.

10. Stock J., Watson M. Introduction to Econometrics 4th edition. Addison-Wesley. 2018. P. 800.

11. Cunningham S. Causal Inference: The Mixtape. Yale University Press. 2021. P. 512.

12. Wooldridge J. Econometric Analysis of Cross Section and Panel Data 2nd edition. MIT Press. 2010. P. 2010.

13. Adams C. Learning Microeconometrics with R. Chapman and Hall/CRC. 2020. P. 398.

14. Lee M. Matching, Regression Discontinuity, Difference in Differences, and Beyond. Oxford University Press. 2016. P. 280.