Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
"САМАРСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ"

В.К. Семенычев, В.Н. Кожухова,
А.А. Коробецкая

ТЕХНОЛОГИИ И ИНСТРУМЕНТАРИЙ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Практикум

Самара
Издательство
Самарского государственного экономического университета
2019

УДК 004.9(076)
ББК У.в6я7
С30

Рецензенты: кафедра математических методов в экономике
Самарского национального исследовательского университета
имени академика С.П. Королева (Самарский университет)
(зав. кафедрой доктор экономических наук, профессор **М.И. Гераськин**);
кафедра "Менеджмент и логистика на транспорте"
Самарского государственного университета путей сообщения
(зав. кафедрой доктор экономических наук, профессор **В.А. Хайтбаев**)

Издается по решению
редакционно-издательского совета университета

Семенычев, Валерий Константинович.

С30 Технологии и инструментарий анализа больших данных [Электронный ресурс] : практикум / В.К. Семенычев, В.Н. Кожухова, А.А. Коробецкая. - Самара : Изд-во Самар. гос. экон. ун-та, 2019. - 1 электрон. опт. диск. - Систем. требования: процессор Intel с тактовой частотой 1,3 ГГц и выше; 256Мб ОЗУ и более; MS Windows XP/Vista/7/10; Adobe Reader; разрешение экрана 1024×768; привод CD-ROM. - Загл. с титул. экрана. - № госрегистрации: 0322000586.

ISBN 978-5-94622-979-1

Практикум включает в себя шесть лабораторных работ, реализующих последовательное усложнение анализа больших данных. Объект анализа рассматривается в практикуме как слабо структурированный (содержащий стохастическую компоненту в данных) и допускающий, в силу этого, неоднозначное моделирование, предполагающий как обработку больших выборок данных (временных и/или пространственных), так и анализ больших выборок конкурирующих моделей для обоснования наиболее адекватных цели анализа.

Предназначен для студентов магистратуры направления подготовки "Прикладная информатика" по программам "Технологии и инструментарий анализа больших данных" и "Информационная аналитика в цифровой экономике", может быть использован также студентами бакалавриата при изучении курсов "Математическое моделирование" и "Математические и инструментальные методы принятия решений" и аспирантами, студентами других направлений подготовки и слушателями курсов повышения квалификации.

УДК 004.9(076)
ББК У.в6я7

ISBN 978-5-94622-979-1

Содержание

Введение

Практикум направлен на приобретение и последующее развитие навыков поиска и идентификации организационно-экономических резервов в целях повышения эффективности производства продукции и ее конкурентоспособности путем привлечения возможностей цифровых решений, основанных на работе с большими данными (*big data*).

Известно, что цифровые технологии в сравнении с аналоговыми обеспечивают высокую точность исходных данных и в разы - большую скорость их передачи. Однако реализация цифровой экономики, столь актуальной для инновационного развития России в условиях экономических санкций, заключается не только в передаче и хранении цифровых *данных* средствами информационно-коммуникационных технологий, но и в большей мере в обретении навыков формирования *знаний* об объектах анализа.

В условиях конкуренции каждое предприятие, освоившее современные достижения цифровой трансформации, способно на опережение налаженного бизнеса фирмы из любой страны за счет более высокой операционной эффективности, значительного роста продаж, создания новых сбытовых каналов, взрывного роста числа коммерческих сделок, ускорения целевого вывода новой продукции на рынок.

Например, современные цифровые платформы могут совершать до 140 тыс. коммерческих сделок за одну секунду, с ежедневным оборотом свыше 17 млрд долл.; капитализация стартапа Uber (услуги такси) уже стала сравнима со стоимостью нефтяного гиганта АО "Роснефть"[1], превышающей 4 трлн руб[2]. Новейшие стартапы проникают в экономическую среду по всему миру, подрывают баланс влияний фирм-лидеров на традиционно сложившихся рынках коммерческих услуг за

---

[1] Королева А. В России признали новую угрозу экономике // Expert Online. 2016.24.11.

[2] Капитализация «Роснефти» впервые превысила 4 трлн руб. // Инф. Агентство РБК. URL: http://www.rbc.ru/rbcfreenews/584918c99a79476700c16ca5.

счет аккумулирования значительных инвестиционных ресурсов и большой операционной эффективности.

Будем исходить из того, что цифровые данные социально-экономических объектов всегда содержат в себе *погрешности первичных измерений, сбои при их передаче, влияния теневой экономики, действия количественно не измеряемых факторов, преднамеренного использования информации о конкурентах, технологиях и рынках, а также зачастую неучет нестационарности или эволюции объектов* за счет инновационных эндогенных (управленческих или технико-технологических) или экзогенных факторов.

Поэтому реальные экономические объекты анализа следует считать слабоструктурированными, допускающими не единственное решение. Кроме того, реальные выборки показателей объекта, по которым осуществляется анализ (собственно моделирование и прогнозирование как конечная цель моделирования), зачастую из-за эволюции не предполагают длинных выборок, как того требует классическая статистика. В этом случае первичные данные остаются, по сути, большими, так как они должны подвергаться многократной и различной обработке для обоснования выбора, возможно, нелинейной модели объекта с оправданным числом параметров (например, как при эволюции). Результаты анализа представляют собой альтернативные вероятностные (точечные и интервальные) оценки объекта, предполагая порой и дополнительную цифровую обработку первичных данных для повышения достоверности.

При этом в странах мира с рыночной экономикой массово используется аппарат современной математической статистики и эконометрики, в определенной мере превышающий уровень их использования в реформируемой российской экономике. Обработка *big data* на мезоуровне экономики (отрасли и регионы) и на микроуровне экономики (муниципальные образования и предприятия), где достаточно велики возможности лиц, принимающих решения, предполагает массовое обучение альтернативным эконометрическим моделям и практике компьютерных расчетов. Согласно исследованию Всемирного экономического форума (ВЭФ) о готовности промышленности ста стран мира к будущим радикальным изменениям российская промышленность оказывается не готовой конкурировать с мировыми лидерами уже в среднесрочной перспективе из-за низкой конкурентоспособности многих предприятий, высокого инвестиционного риска, слабого использования *big data* и эконометрической аналитики.

Данный практикум формировался по принципу "от простого к более сложному", предлагая в ряде случаев выход за рамки традиционных вузовских курсов для получения дополнительных теоретических сведений

и программных навыков. Лабораторные работы реализуют наиболее известные типовые и актуальные задачи по эконометрическому моделированию и прогнозированию статики и динамики социально-экономических и технических объектов на реальных, зачастую относительно малых выборках. Заметим на перспективу, что для оперативной оценки отклика объекта анализа на разного рода воздействия (как внешние, так и внутренние) при откровенно малых выборках (меньших 20-50 наблюдений), встречающихся на практике, целесообразно обратиться к более сложному методу "бутстреп".

Практикум включает в себя шесть лабораторных работ с указаниями по их выполнению и оформлению, а также содержит контрольные вопросы, проверяющие усвоение теории.

Первая работа лабораторного практикума является, по сути, вводной и посвящена основным понятиям теории вероятностей и математической статистики, методам и свойствам выборочных оценок параметров распределений на выборках разного объема.

Во второй работе требуется построить *парные* (линейную и нелинейные) регрессионные модели по имеющимся статистическим данным, оценить качество полученных моделей, выбрать лучшую из них (по критериям точности и адекватности).

В третьей работе необходимо построить модель *множественной линейной* регрессии, произведя отбор (редукцию) наиболее значимых факторов. Основным методом идентификации моделей во второй и в третьей лабораторных работах является традиционный метод наименьших квадратов (МНК), приведены необходимые условия, налагаемые на стохастическую компоненту, для его применения на практике.

В четвертой работе рассматриваются *нелинейные по параметрам и факторам* регрессионные модели, оцениваются их параметры. Здесь же иллюстрируются большие возможности для идентификации многопараметрических моделей более современным методом - генетическим алгоритмом (ГА), реализуемым с помощью программной среды *R*, затрагивая разделы эконометрики, получившие широкое применение в современной цифровой экономике.

Пятая работа посвящена моделированию финансовых временных рядов *при помощи ARIMA-моделей различного порядка*. Здесь же рассмотрены и вопросы декомпозиции рядов на компоненты, что демонстрирует реализацию системного подхода к слабо структурируемым объектам анализа с присутствием стохастической компоненты.

В шестой лабораторной работе анализируются *логистические модели трендов*, отражающие возможность переменной динамики (эволюционного развития) социально-экономических и технических объектов как с фиксированной, так и с произвольной асимметрией, адаптивно

настраиваемой на возможно большую область применения. Реализация наиболее известных в отечественной эконометрике логистических трендов Ферхюльста и Гомпертца с фиксированными асимметриями дополнена и другими более сложными, в том числе авторскими, адаптивными моделями и методами их идентификации.

Все предложенные для обучения лабораторные работы выполняются по индивидуальным вариантам для обучающихся средствами *MS Excel* 2016 или языка программирования *R*.

В отдельном разделе рассматриваются функции *R*, необходимые для загрузки и выгрузки больших данных для их последующей обработки. Все изложенные в лабораторных работах методы могут быть применены для анализа полученных таким образом больших данных.

Определим общую характеристику и место *R* среди десятков известных программ по обработке данных. Классическая статистика в практикуме основывается на проверке статистических гипотез - априорных предположений о свойствах исследуемых данных. Когда компьютеры были еще слабы в области отображения информации, особенно графической, практика ориентировалась на построение моделей и на проверку гипотез. Подобный подход реализован в известных программных средствах (например, *SPSS*), базирующихся на электронных таблицах и управляющихся с помощью меню. Фактически первые версии программных продуктов *SPSS*, а также и *SAS* Analytics состояли из подпрограмм, которые можно было вызвать из основной программы (на *Fortran* или на другом языке) с целью подгонки и проверки модели из имеющегося набора моделей. Альтернативой этому подходу была концепция разведочного анализа данных *EDA*, который применяется для нахождения связей между переменными в ситуациях, когда отсутствуют (или недостаточны) априорные представления о природе этих связей. Как правило, при разведочном анализе учитывается и сравнивается большое число переменных, а для поиска закономерностей применяются разные методы.

R соединяет в себе обе концепции, позволяя использовать как основные статистические методы, так и более сложные, специально созданные методы многомерного анализа, предназначенные для отыскания закономерностей в многомерных данных. *R* в отличие от *SAS* и *SPSS* является **универсальным языком программирования**, разработанным для применения в таких областях, как разведочный анализ данных, классические статистические тесты и высокоуровневая графика.

R обладает рядом существенных достоинств. *R* - свободная программная среда вычислений **с открытым исходным кодом**. Это реализация языка *S* с дополнительными моделями, разработанными в языке ***S-Plus***.

R доступен в соответствии с лицензией *GNU*. На этом фундаменте *R* продолжает развиваться в значительной степени посредством добавления пакетов, представляя собой коллекцию наборов данных, функций языка *R*, документации и динамически загружаемых элементов на языке *C* или *Fortran*. Посредством этих пакетов исследователи могут с легкостью обмениваться вычислительными методами со своими коллегами.

R **обладает собственной обширной и непрерывно расширяющейся библиотекой пакетов**, содержащей большое количество готовых решений различных задач. Некоторые пакеты имеют ограниченную область применения, какие-то представляют целые области статистики, а другие отражают новейшие разработки. Многие новые разработки в области статистики, эконометрики, биологии, химии, медицины, географии и других прикладных областей сначала появляются как *R*-пакеты и только потом реализуются в коммерческих программных продуктах.

***R* - это мощный скриптовый язык.** Скриптом называется программный сценарий, любая исполняемая процедура, которая запускается автоматически или же с помощью команды пользователя. Скрипты используют не только в программировании, но и для повторяемых и сложных для запоминания пользователем операций. Для обработки неупорядоченных данных требуются возможности языка программирования: продукты *SAS* и *SPSS* также имеют скриптовые языки для решения отдельных задач, однако *R* был создан именно как язык программирования и поэтому является более подходящим средством для этой цели, полезным инструментом в области анализа больших массивов данных. Он уже **интегрирован в ряд коммерческих пакетов**, таких как *IBM SPSS* и *InfoSphere*, а также *Mathematica*.

R - язык, ориентированный на статистику, который можно рассматривать в качестве конкурента для таких аналитических систем, как *SAS Analytics*, не говоря уже о более простых пакетах *StatSoft STATISTICA* или *Minitab*.

R **органично интегрируется и с системами публикации** документов, что позволяет встраивать статистические результаты и графику из среды *R* в документы публикационного качества.

Как язык программирования *R* подобен многим другим языкам. Любой человек, который когда-либо писал программный код, найдет в *R* множество знакомых моментов. Отличительные особенности *R* лежат в статистической философии, которую он исповедует. Язык *R* имеет легкий синтаксис - это универсальный инструмент, созданный специально для работы с данными. *R* также включает в себя чрезвычайно мощные графические возможности.

Авторы практикума не один десяток лет занимаются рассмотренными вопросами, а В.Н. Кожуховой получены соответствующие сертификаты по применению *R* от НИУ ВШЭ и Института биоинформатики СПбАУ РАН. Предложенные лабораторные работы в течение нескольких лет прошли апробацию в вузах.

Время выполнения и номенклатура отдельных лабораторных работ могут быть различными в зависимости от подготовленности контингента обучающихся и их предпочтений по выбору тем.

Общие рекомендации по выполнению лабораторных работ

В данном практикуме рассматривается работа с *R* из-под операционной системы *Windows*. Для выполнения практикума читателю необходимо установить:

1) язык *R*:

классический https://cran.rstudio.com/bin/windows/base/

или PRO https://mran.revolutionanalytics.com/download/

2) графический интерфейс пользователя *RStudio*:

https://www.rstudio.com/products/rstudio/download/ - набор разработанных интегрированных инструментов, чтобы наиболее продуктивно использовать *R*.

RStudio включает в себя консоль, редактор с подсветкой синтаксиса, поддерживающий как прямое выполнение кода, так и инструменты для построения графиков, сохранение истории команд, отладку и управление рабочим пространством.

При установке *R* будет лучше, если имя пользователя *Windows* будет написано латиницей. Если имя пользователя набрано кириллицей (например, "Иван"), то необходимо создать нового пользователя с именем без кириллицы (например, "*Ivan*"), а установку программ и дальнейшую работу рекомендуется проводить, войдя в систему под "английским" пользователем.

На время установки *R* и *RStudio* также рекомендуется отключить антивирусную программу. Другим вариантом избежания ошибок, связанных с кириллицей, является и установка *R*, и дальнейшая работа от имени администратора. При сохранении *.*R* файлов не следует использовать названия файлов или папок, содержащих русские буквы или пробелы. Для работы в *R* потребуется стабильное интернет-соединение, так как может понадобиться онлайн-загрузка дополнительных пакетов для обработки данных.

Англоязычных источников по *R* чрезвычайно много. При возникновении вопросов, ответов на которые нет в справке *R*, можно использовать следующие ресурсы:

http://rseek.org/ - *Google*, модифицированный под поиск по источникам, связанным с *R*;

http://stats.stackexchange.com/ - обращаться с вопросами по поводу статистических методов и их реализации в *R*;

http://stackoverflow.com/ - обращаться с вопросами по поводу программирования в *R*;

http://vk.com/rstatistics - группа в социальной сети "ВКонтакте", участники которой могут ответить на возникшие вопросы;

http://R-analytics.blogspot.ru/ - один из самых масштабных русскоязычных проектов по *R*.

Данный курс предполагает знание читателями базового курса эконометрики и основ программирования на любом языке высокого уровня. Список литературы [1, 2, 5-7, 13] ориентирует на наиболее известные публикации, посвященные разным сторонам решения задачи использования среды *R*.

Общие требования к оформлению отчетов по лабораторным работам

Отчет оформляется в текстовом процессоре *MS Word* или аналогичном и сдается в печатном виде. Отчет по лабораторным работам должен содержать: титульный лист; индивидуальное задание; основную часть, которая должна содержать результаты выполнения каждого пункта задания с краткими комментариями и указанием расчетных формул, а при необходимости следует приводить основные промежуточные расчеты; выводы, в которых следует привести и сопоставить основные численные результаты работы.

Рекомендуется использовать шрифт *Times New Roman*, 14 пт, абзацный отступ 1,25 см, межстрочный интервал - 1 или 1,5, поля: верхнее 1,5-2 см, нижнее 2 см, правое 1-1,5 см, левое 2,5-3 см. Страницы нумеруются снизу по центру, кроме титульного листа.

Лабораторная работа 1
Элементы теории вероятностей. Основные виды распределений. Генерация случайных величин

Цель работы: напомнить основные понятия плотности (дифференциального закона распределения) и интегральной функции распределения дискретных и непрерывных случайных величин, рассчитать их теоретические числовые характеристики. Научиться разыгрывать случайные величины, рассчитывать выборочные числовые характеристики.

Задание. Лабораторная работа выполняется в *Microsoft Excel* 2016. Варианты работы находятся в приложении.

1. Для заданных по вариантам значений параметров рассчитать теоретическую плотность распределения и интегральную функцию распределения:
   - для биномиального закона распределения;
   - распределения Пуассона;
   - нормального распределения;
   - равномерного распределения;
   - распределения хи-квадрат;
   - распределения Фишера;
   - распределения Стьюдента.

Построить графики указанных распределений. Значения аргументов функций подобрать самостоятельно. Рассчитать **математическое ожидание, дисперсию, среднее квадратическое отклонение, моду и медиану.**

2. Сгенерировать выборку из $n = 100$ значений случайной величины:
   - с равномерным законом распределения;
   - с нормальным законом распределения.

Параметры распределений берутся в соответствии со своим номером варианта из таблицы исходных данных к лабораторной работе.

Для каждой выборки следует:

а) построить вариационный ряд;

б) разбить полученный ряд на m интервалов равной ширины h и подсчитать количество значений, попавших в каждый интервал (эмпирические абсолютные частоты);

в) рассчитать высоту столбцов и построить гистограмму эмпирического распределения;

г) рассчитать накопленные относительные частоты попадания в интервалы и построить интегральную функцию эмпирического распределения.

3. Рассчитать оценки среднего выборочного значения, выборочную дисперсию, выборочное среднее квадратическое отклонение (СКО); моду и медиану по интервальному (группированному) ряду; квартили по исходным (негруппированным) выборкам. Построить **"ящик с усами"**, а также гистограмму распределения с помощью встроенных диаграмм *Excel*, сравнить их с гистограммой, полученной в п. 2.

4. Проделать пп. 2, 3 данной лабораторной работы для новой выборки объемом из 1000 значений. Как изменилась эмпирическая плотность (гистограмма) распределения и выборочные числовые характеристики? Сравнить их с теоретическими (теми, которые были заданы в п. 1), что можно о них сказать?

Указания к выполнению

1. Построение графиков плотностей и функций распределения

Диапазон изменения аргументов функций выбирайте самостоятельно для получения графиков функций, на которых будет понятна принципиальная форма изображаемой кривой.

Биномиальный закон распределения описывает вероятность количества "успехов" в последовательности из n испытаний. Функция плотности распределения устанавливается двумя параметрами: вероятностью успеха в каждом испытании p и числом испытаний n:

$$P_n(k) = C_n^k p^k q^{n-k}.$$

Используйте функцию =БИНОМ.РАСП(число_успехов; число_испытаний; вероятность_успеха; интегральная). Число успехов - аргумент функции, интегральная - логический признак, может принимать значения 0 - ложь (плотность распределения) и 1 - истина (интегральная функция распределения).

Краткое обозначение: B ($\boldsymbol{n}$, $\boldsymbol{p}$).

Распределение Пуассона (один параметр $\lambda = np$) имеет вид

$$P_n(k) = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Используйте функцию = ПУАССОН.РАСП(*x*; среднее; интегральная). Среднее - это параметр λ, *x* - это аргумент функции. Интегральная - аналогично предыдущему примеру.

Краткое обозначение: $Pois(\lambda)$.

Нормальный закон распределения задается плотностью:

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

где M, σ - параметры распределения (математическое ожидание и стандартное отклонение, соответственно).

Можно воспользоваться функцией НОРМ.РАСП (*x*; среднее; стандартное_откл; интегральная), где *x* - аргумент; среднее - *M*; стандартное_откл - σ; интегральная- логический признак, принять равным 0 для плотности распределения (и 1 - для интегральной функции распределения).

Краткое обозначение: $N(M,\sigma^2)$.

Равномерный закон распределения задается плотностью:

$$f(x)=\begin{cases}\dfrac{1}{b-a}, a\le x\le b,\\ 0, \text{иначе}\end{cases}$$

Функция равномерного распределения:

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

В *Excel* нет встроенной функции равномерного распределения. Постройте самостоятельно, воспользовавшись функцией ЕСЛИ.

Краткое обозначение: $U(a,b)$.

Хи-квадрат распределение с *k* степенями свободы - это распределение **суммы квадратов** *k* независимых стандартных **нормальных** случайных величин. Нормальная СВ называется стандартной, если имеет $M = 0$, $\sigma = 1$.

Плотность распределения хи-квадрат:

$$f_{x^2(k)}(x)=\frac{(1/2)^{\frac{k}{2}}}{\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2}-1}e^{-\frac{x}{2}},$$

где $\Gamma(k/2)$ - гамма-функция.

Используйте функцию = ХИ2.РАСП(*x*; степени_свободы; интегральная). Степени_свободы - это параметр *k*, *x* - это аргумент функции, интегральная - аналогично предыдущим примерам.

Краткое обозначение: $x^2\left(k\right)$

Распределение Фишера (F-распределение). Пусть Y_1, Y_2 - две независимые случайные величины, имеющие распределение хи-квадрат: $Y_i \sim x^2\left(k_i\right)$, где $k_i \in \mathbb{N}, i=1,2$. Тогда распределение случайной величины $F=\frac{Y_1/k_1}{Y_2/k_2}$ называется распределением Фишера со степенями свободы k_1 и k_2.

Используйте функцию = F.РАСП(*x*; степени_свободы1; степени_свободы2; интегральная). Степени_свободы - это параметры k_1 и k_2, *x* - это аргумент функции, интегральная - аналогично предыдущим примерам.

Краткое обозначение: $F\left(k_1,k_2\right)$.

Распределение Стьюдента (t-распределение). Пусть $Y_0, Y_1, \ldots, Y_k$ - независимые стандартные нормальные случайные величины, такие что $Y_i \sim N(0,1), i=0,\ldots,k.$. Тогда распределение случайной величины *t*, где $t=\frac{Y_0}{\sqrt{\frac{1}{k}\sum_{i=1}^{k}Y_i^2}}$, называется распределением Стьюдента с *k* степенями свободы. Ее распределение абсолютно непрерывно и имеет плотность

$$f_t(y)=\frac{\Gamma\left(\frac{k+1}{2}\right)}{\sqrt{\pi k}\,\Gamma\left(\frac{k}{2}\right)}\left(1+\frac{y^2}{k}\right)^{-\frac{k+1}{2}},$$

где Г - гамма-функция Эйлера.

Используйте функцию = СТЬЮДЕНТ.РАСП(*x*; степени_ свободы; интегральная). Степени свободы - это параметр k, x - это аргумент функции, интегральная - аналогично предыдущим примерам.

Краткое обозначение: $T(k)$.

Для каждого закона распределения нужно получить график, как на рис. 1.1. Всего должно быть семь графиков плотностей (для дискретных распределений - рядов) распределения и семь графиков функций распределения.

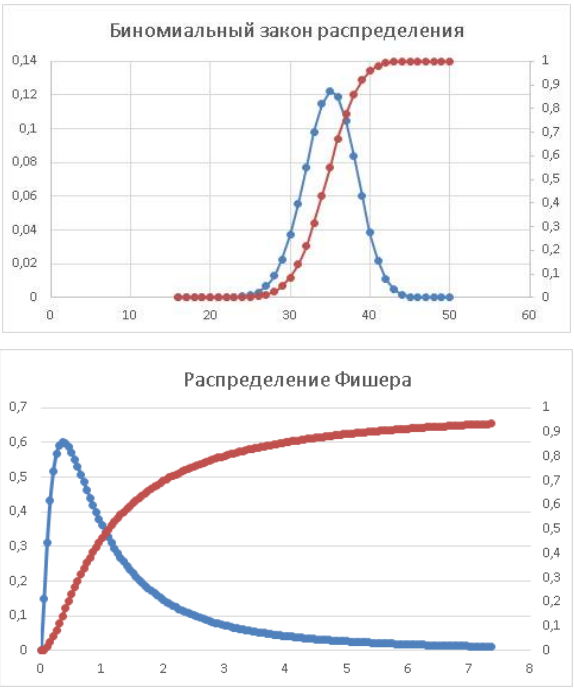


Рис. 1.1. **Примеры графиков распределений**

Для распределений Фишера и Стьюдента шаг аргумента лучше брать меньше (см. рис. 1.2 как пример подбора аргумента для исходных параметров).

Замечание: первые два закона дискретные, а гладкий график ряда распределения строится исключительно из соображений наглядности.

	B	C	D
2	**Вариант примера**		
3			
4	Бином. распр.	n	50
5		p	0,7
6	Распр. Пуассона	λ	16
7	Норм. распр.	M	243
8		σ	25
9	Равном. распр.	a	14
10		b	17
11	Хи-квадрат	k	20
12	Фишера	k1	5
13		k2	3
14	Стьюдента	k	3

	F	G	H	I	J	K	L	M	N	O	P	Q
3		Бином			Пуассон			Норм			Равном	
4	аргумент X	плотность	функция	аргумент X	плотность	функция	аргумент X	плотность	функция	аргумент X	плотность	функция
5	16	2,7289E-08	3,3771E-08	0	1,1254E-07	1,1254E-07	140	3,2887E-06	1,8944E-05	10	0	0
6	17	1,2735E-07	1,6112E-07	1	1,8006E-06	1,9131E-06	145	7,3485E-06	4,4274E-05	11	0	0
7	18	5,4476E-07	7,0588E-07	2	1,4405E-05	1,6318E-05	150	1,5776E-05	9,9611E-05	12	0	0
8	19	2,1408E-06	2,8467E-06	3	7,6824E-05	9,3142E-05	155	3,2541E-05	0,00021577	13	0	0
9	20	7,7426E-06	1,0589E-05	4	0,0003073	0,00040044	160	6,4489E-05	0,00045009	14	0,333333	0
10	21	2,5809E-05	3,6398E-05	5	0,00098335	0,00138379	165	0,00012279	0,00090426	15	0,333333	0,3333
11	22	7,9382E-05	0,00011578	6	0,00262226	0,00400604	170	0,00022464	0,00175016	16	0,333333	0,6667
12	23	0,00022549	0,00034127	7	0,00599374	0,00999978	175	0,00039485	0,0032641	17	0,333333	1
13	24	0,00059191	0,00093318	8	0,01198747	0,02198725	180	0,0006668	0,00586774	18	0	1
14	25	0,00143637	0,00236955	9	0,02131106	0,04329832	185	0,00108192	0,01017044	19	0	1
15	26	0,00322262	0,00559217	10	0,0340977	0,07739602	190	0,00168664	0,01700302	20	0	1
16	27	0,00668396	0,01227613	11	0,04959665	0,12699267	195	0,00252626	0,02742895	21	0	1
17	28	0,01281092	0,02508704	12	0,06612887	0,19312154	200	0,00363548	0,04271622			
18	29	0,02267679	0,04776384	13	0,08138938	0,27451092	205	0,00502659	0,06425549			
19	30	0,03703876	0,0848026	14	0,09301644	0,36752736	210	0,00667748	0,09341751			
20	31	0,05575728	0,14055988	15	0,09921753	0,46674489	215	0,00852277	0,13135688			
21	32	0,07724706	0,21780694	16	0,09921753	0,56596242	220	0,01045145	0,17878638			
22	33	0,09831444	0,31612138	17	0,09338121	0,65934363	225	0,01231405	0,2357625			
23	34	0,11470018	0,43082156	18	0,08300552	0,74234915	230	0,0139397	0,30153179			
24	35	0,12234686	0,55316843	19	0,06989938	0,81224853	235	0,01516122	0,37448417			
25	36	0,11894834	0,67211676	20	0,05591951	0,86816803	240	0,01584321	0,45224157			
26	37	0,10501745	0,77713422	21	0,04260534	0,91077337	245	0,01590671	0,53188137			
27	38	0,08382972	0,86096394	22	0,0309857	0,94175907	250	0,01534425	0,61026125			
28	39	0,06018544	0,92114938	23	0,02155527	0,96331434	255	0,0142213	0,6843863			
29	40	0,03861899	0,95976837	24	0,01437018	0,97768452	260	0,01266372	0,75174777			
30	41	0,02197829	0,98174665	25	0,00919692	0,98688144	265	0,01083456	0,81057035			
31	42	0,01098914	0,9927358	26	0,00565964	0,99254108	270	0,00890614	0,85992891			
32	43	0,00477048	0,99750628	27	0,00335386	0,99589494	275	0,0070339	0,89972743			
33	44	0,00177086	0,99927714	28	0,00191649	0,99781143	280	0,00533741	0,93056338			
34	45	0,00055093	0,99982807	29	0,00105737	0,9988688	285	0,00389129	0,95352134			
35	46	0,00013973	0,9999678	30	0,00056393	0,99943274	290	0,00272574	0,96994596			

Рис. 1.2. **Примеры расчета значений функций**

Квантиль распределения - значение, которое анализируемая случайная величина не превышает с фиксированной вероятностью. Квантиль порядка 0,5 называется **медианой**: $F\left(x_{0,5}\right)=0,5,$ где $x_{0,5}$ - значение медианы распределения.

Медианой M_e случайной величины X называют то ее значение, для которого выполняется равенство вероятностей событий, т.е. плотность вероятностей справа и слева одинакова и равна половине (0,5).

$$P(-\infty < X < M_e) = P(M_e < X < \infty)$$
$$F(M_e) - F(-\infty) = F(+\infty) - F(M_e)$$
$$2F(M_e) = 1 \Rightarrow F(M_e) = 0,5$$

Модой M_o дискретной случайной величины X называют те ее возможные значения, которые соответствуют наибольшей вероятности появления, т.е. такое значение величины X, которое случается чаще всего при анализе наблюдений.

В случае непрерывной случайной величины модой называют то ее возможное значение, которому соответствует максимальное значение плотности вероятностей:

$$f\left(M_o\right) = \max_x f\left(x\right).$$

В зависимости от вида функции *f(x)* случайная величина X может иметь различное количество мод. Если случайная величина имеет одну моду, то такое распределение вероятностей называют одномодальным; если распределение имеет две моды - двухмодальным, более двух - мультимодальным. Существуют и такие распределения, которые не имеют моды, их называют амодальными (например, равномерное).

Графически мода и медиана изображены на рис. 1.3.

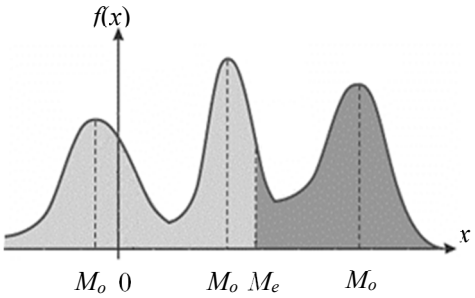


Рис. 1.3. **Мода и медиана распределения**

В табл. 1.1 приведены значения основных характеристик распределений.

Таблица 1.1

Основные числовые характеристики распределений

Характеристика	$B(n,p)$	$Pois(\lambda)$	$N(M,\sigma^2)$	$U(a,b)$	$x^2(k)$	$F(k_1,k_2)$	$T(k)$
$M(X)$	np	λ	M	$\frac{a+b}{2}$	k	$\frac{k_2}{k_2-2}$, если $k_2>2$	0, если $k>1$
$D(X)$	npq	λ	σ^2	$\frac{(b-a)^2}{12}$	$2k$	$\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}$, если $k_2>4$	$\frac{k}{k-2}$, если $k>2$
$\sigma(X)$	$\sqrt{npq}$	$\sqrt{\lambda}$	σ	$\frac{(b-a)}{2\sqrt{3}}$	$\sqrt{2k}$	$\sqrt{\frac{2k_2^2(k_1+k_2-2)}{k_1(k_2-2)^2(k_2-4)}}$, если $k_2>4$	$\sqrt{\frac{k}{k-2}}$, если $k>2$
M_o	$\lfloor (n+1)p \rfloor$	$\lfloor \lambda \rfloor$	M	-	$k-2$, если $k \geq 2$	$\frac{k_1-2}{k_1}\frac{k_2}{k_2+2}$, если $k_1>2$	0
M_e	$\lfloor np \rfloor$	$\approx\lfloor \lambda+1/3-0{,}02/\lambda \rfloor$	M	$\frac{a+b}{2}$	$\approx k-2/3$	-	0

Здесь $\lfloor x \rfloor$ означает округление до ближайшего целого в меньшую сторону, "пол". Например, $\lfloor 25,9 \rfloor = 25$

ОКРВНИЗ.МАТ(число).

Цель выполнения пп. 2, 3, 4 данной работы - сравнить характеристики теоретического и эмпирического распределений.

Эмпирическое распределение будет создано с помощью искусственной выборки процедурой генерации случайных чисел.

Теоретические значения числовых характеристик распределения были частично рассчитаны в табл. 1.1. Приведем теоретические значения первого и третьего квартилей для равномерного и нормального законов (табл. 1.2). Подумайте, как получились эти формулы.

Таблица 1.2

Основные числовые характеристики распределений

Квартиль	N (M, σ^2)	U (a, b)
Q_1	$M + F^{-1}(0,25)\sigma$	$\frac{3}{4}a + \frac{1}{4}b$
Q_3	$M + F^{-1}(0,75)\sigma$	$\frac{1}{4}a + \frac{3}{4}b$

где $F^{-1}_{(p)}$ - квантиль стандартного нормального распределения порядка p.

$F^{-1}_{(p)}$ можно рассчитать по формуле = НОРМ.СТ.ОБР(p).

2. Генерация случайных величин

Генерация случайных величин может быть необходима для реализации численного моделирования (симуляции) по методу Монте-Карло для точечной и интервальной оценок точности методов идентификации моделей (подробнее показано на конкретных примерах моделирования, например в [11].

Генерировать случайные числа необходимо с теми параметрами, которые ранее использовались при построении графиков равномерного и нормального распределения.

Покажем выполнение п. 2 на примере равномерного распределения с параметрами 14 и 17, взятыми из п. 1.

Генерация случайных чисел вызывается командой: *Данные - Анализ данных - Генерация случайных чисел*. Если *Анализ данных* на вкладке *Данные* отсутствует, его необходимо установить через *Параметры Excel - Надстройки - Перейти - Пакет анализа* (рис. 1.4).

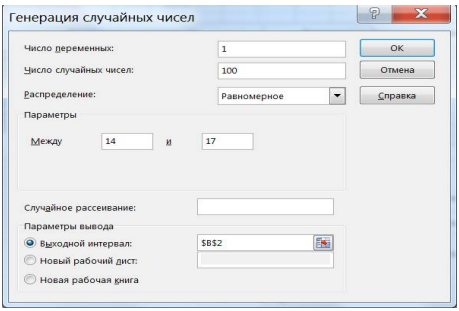


Рис. 1.4. **Генерация 100 значений равномерно распределенной случайной величины с параметрами 14 и 17**

Построить вариационный ряд - значит упорядочить полученные значения по возрастанию.

Опишем процедуру разбиения полученного ряда на m интервалов равной ширины h и подсчета количества значений, попавших в каждый интервал (эмпирические абсолютные частоты).

Минимальное число столбцов (интервалов) гистограммы распределения рассчитывается по формуле Стерджесса: $m = 1 + 3{,}322\, lg\, n$.

Ширина интервала $h = \dfrac{x_{\max} - x_{\min}}{m-1}$, начальное значение первого интервала $x_{\min} - \dfrac{h}{2}$

Упорядочить выборку по возрастанию: выделить диапазон значений; вкладка *Главная - Сортировка и фильтр - Сортировка по возрастанию.*
Минимальное и максимальное значения в выборке: МИН (число1; число2; ...) и МАКС (число1; число2; ...).
Количество наблюдений, попавших в *i*-й интервал, определяется с помощью функции ЧАСТОТА (массив_данных; массив_интервалов).
Массив_интервалов - правые границы интервалов. Формулу необходимо ввести как формулу массива.
После ввода формулы в первую ячейку (например, как на рис. 6, H7) выделите весь нужный диапазон (например, H7:H14 - для 8 интервалов), нажмите клавишу F2, а затем нажмите одновременно клавиши CTRL+SHIFT+ENTER. Если формула не будет введена как формула массива, отобразится только одно ее значение в ячейке H7.
Не вводите фигурные скобки с клавиатуры (рис. 1.5 и 1.6).

	A	B	C	D	E	F	G	H
1	№	Вариац. ряд						
2	1	14,05511643		n=	100	x_{min}=	14,05512	
3	2	14,0737022		m=	8	x_{max}=	16,98297	
4	3	14,12762841		h=	0,418265			
5	4	14,14365062						
6	5	14,15994751		№	начало	конец	середина	частоты
7	6	14,16489151		1	13,84598	14,26425	14,05512	7
8	7	14,17368084		2	14,26425	14,68251	14,47338	9
9	8	14,2773217		3	14,68251	15,10078	14,89165	9
10	9	14,35065767		4	15,10078	15,51904	15,30991	15
11	10	14,35551012		5	15,51904	15,93731	15,72818	18
12	11	14,36338389		6	15,93731	16,35557	16,14644	19
13	12	14,41740165		7	16,35557	16,77384	16,56471	12
14	13	14,53614917		8	16,77384	17,1921	16,98297	11
15	14	14,57423627		Сумма				**100**

Рис. 1.5. **Разбиение выборки на интервалы**

	A	B	C	D	E	F	G	H
1	№	Вариац. ряд						
2	1	14,0551164281137		n=	100	x_{min}=	=МИН(B2:B101)	
3	2	14,0737022003845		m=	=ОКРУГЛ(1+3,322*LOG10(E2);0)	x_{max}=	=МАКС(B2:B101)	
4	3	14,1276284066286		h=	=(G3-G2)/(E3-1)			
5	4	14,1436506241035						
6	5	14,1599475081637		№	начало	конец	середина	частоты
7	6	14,1648915066988		1	=G2-E4/2	=E7+$E$4	=(E7+F7)/2	=ЧАСТОТА(B2:B101;F7:F14)
8	7	14,1736808374279		2	=E7+$E$4	=E8+$E$4	=(E8+F8)/2	=ЧАСТОТА(B2:B101;F7:F14)
9	8	14,2773216956084		3	=E8+$E$4	=E9+$E$4	=(E9+F9)/2	=ЧАСТОТА(B2:B101;F7:F14)
10	9	14,3506576738792		4	=E9+$E$4	=E10+$E$4	=(E10+F10)/2	=ЧАСТОТА(B2:B101;F7:F14)
11	10	14,3555101168859		5	=E10+$E$4	=E11+$E$4	=(E11+F11)/2	=ЧАСТОТА(B2:B101;F7:F14)
12	11	14,3633838923307		6	=E11+$E$4	=E12+$E$4	=(E12+F12)/2	=ЧАСТОТА(B2:B101;F7:F14)
13	12	14,4174016541032		7	=E12+$E$4	=E13+$E$4	=(E13+F13)/2	=ЧАСТОТА(B2:B101;F7:F14)
14	13	14,5361491744743		8	=E13+$E$4	=E14+$E$4	=(E14+F14)/2	=ЧАСТОТА(B2:B101;F7:F14)
15	14	14,5742362743004		Сумма				**=СУММ(H7:H14)**

Рис. 1.6. **Разбиение выборки на интервалы в режиме отображения формул**

Высота *i*-го столбца гистограммы эмпирического распределения (рис. 1.7) рассчитывается по формуле

$$h_i = \frac{n_i}{n \cdot h},$$

где n_i - количество наблюдений, попавших в *i*-й интервал (столбец частоты);

h - ширина интервала;

n - число наблюдений в выборке.

n=	100		x_{min}=	14,05511643				
m=	8		x_{max}=	16,98297067				
h=	0,418264892							

№	начало	конец	середина	частоты	отн. частоты	высоты	э.ф.р.
1	13,84598398	14,26425	14,05511643	7	0,07	0,167358058	0,07
2	14,26424887	14,68251	14,47338132	9	0,09	0,215174646	0,16
3	14,68251377	15,10078	14,89164621	9	0,09	0,215174646	0,25
4	15,10077866	15,51904	15,3099111	15	0,15	0,35862441	0,4
5	15,51904355	15,93731	15,728176	18	0,18	0,430349292	0,58
6	15,93730844	16,35557	16,14644089	19	0,19	0,454257586	0,77
7	16,35557333	16,77384	16,56470578	12	0,12	0,286899528	0,89
8	16,77383823	17,1921	16,98297067	11	0,11	0,262991234	1
Сумма				100	1		

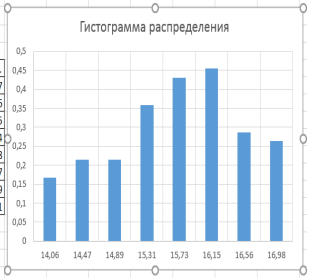
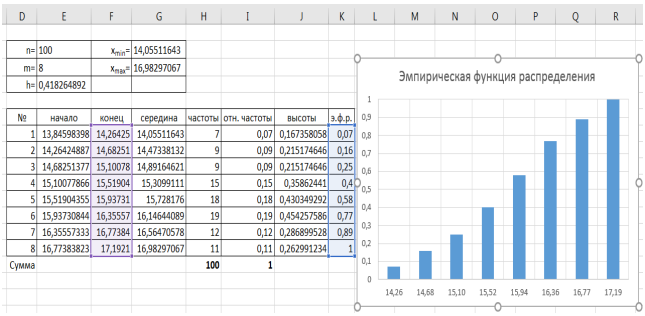


Рис. 1.7. **Гистограмма по выборке равномерного распределения**

Гистограмму распределения необходимо строить по серединам интервалов (ось *X*), интегральную функцию эмпирического распределения (э. ф. р. - рис. 1.8) - по правым границам интервалов.

Эмпирическая функция распределения представляет собой диапазон накопленных частот n_i/n.



n=	100		x_{min}=	14,05511643				
m=	8		x_{max}=	16,98297067				
h=	0,418264892							

№	начало	конец	середина	частоты	отн. частоты	высоты	э.ф.р.
1	13,84598398	14,26425	14,05511643	7	0,07	0,167358058	0,07
2	14,26424887	14,68251	14,47338132	9	0,09	0,215174646	0,16
3	14,68251377	15,10078	14,89164621	9	0,09	0,215174646	0,25
4	15,10077866	15,51904	15,3099111	15	0,15	0,35862441	0,4
5	15,51904355	15,93731	15,728176	18	0,18	0,430349292	0,58
6	15,93730844	16,35557	16,14644089	19	0,19	0,454257586	0,77
7	16,35557333	16,77384	16,56470578	12	0,12	0,286899528	0,89
8	16,77383823	17,1921	16,98297067	11	0,11	0,262991234	1
Сумма				100	1		

Рис. 1.8. **Эмпирическая функция распределения**

Таким образом, должна получиться табл. 1.3 со следующими столбцами.

Таблица 1.3

Вспомогательная таблица при разбиении выборки

№ интервала	**Начало интервала**	**Конец интервала**	**Середина интервала**	**Частота n_i**	**Относ. частота n_i / n**	**Высота h_i**	**э.ф.р.**
1	$x_{\min} - \frac{h}{2}$	$\left(x_{\min} - \frac{h}{2}\right) + h$	$\frac{\text{Начало+конец}}{2}$	…	…	…	…
2	$\left(x_{\min} - \frac{h}{2}\right) + h$	$\left(x_{\min} - \frac{h}{2}\right) + 2h$	…	…	…	…	…
…	…	…	…	…	…	…	…

Вариант выполнения расчетов на листе Excel в режиме отображения формул представлен на рис. 1.9.

	A	B	C	D	E	F	G	H	I	J	K
1	№	Вариац. ряд									
2	1	14,0551164281137		n=	100	x_{min}=	=МИН(B2:B101)				
3	2	14,0737022003845		m=	=ОКРУГЛ(1+3,322*LOG10(E2);0)	x_{max}=	=МАКС(B2:B101)				
4	3	14,1276284066286		h=	=(G3-G2)/(E3-1)						
5	4	14,1436506241035									
6	5	14,1599475081637		№	начало	конец	середина	частоты	отн. частоты	высоты	э.ф.р.
7	6	14,1648915066988		1	=G2-E4/2	=E7+$E$4	=(E7+F7)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H7/$E$2	=I7/$E$4	=I7
8	7	14,1736808374279		2	=E7+$E$4	=E8+$E$4	=(E8+F8)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H8/$E$2	=I8/$E$4	=K7+I8
9	8	14,2773216956084		3	=E8+$E$4	=E9+$E$4	=(E9+F9)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H9/$E$2	=I9/$E$4	=K8+I9
10	9	14,3506576738792		4	=E9+$E$4	=E10+$E$4	=(E10+F10)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H10/$E$2	=I10/$E$4	=K9+I10
11	10	14,3555101168859		5	=E10+$E$4	=E11+$E$4	=(E11+F11)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H11/$E$2	=I11/$E$4	=K10+I11
12	11	14,3633838923307		6	=E11+$E$4	=E12+$E$4	=(E12+F12)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H12/$E$2	=I12/$E$4	=K11+I12
13	12	14,4174016541032		7	=E12+$E$4	=E13+$E$4	=(E13+F13)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H13/$E$2	=I13/$E$4	=K12+I13
14	13	14,5361491744743		8	=E13+$E$4	=E14+$E$4	=(E14+F14)/2	=ЧАСТОТА(B2:B101;F7:F14)	=H14/$E$2	=I14/$E$4	=K13+I14
15	14	14,5742362743004		Сумма				**=СУММ(H7:H14)**	**=СУММ(I7:I14)**		

Рис. 1.9. **Таблица разбиения выборки на интервалы в режиме отображения формул**

3. Расчет выборочных числовых характеристик

Расчет выборочных описательных статистик производится по следующим формулам:

$\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$ - выборочное среднее;

$S_x^2=\frac{1}{n}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2$ - выборочная дисперсия;

$S_x=\sqrt{S_x^2}$ - выборочное СКО,

где x_i - сгенерированный ряд наблюдений в выборке.

выборочное среднее: СРЗНАЧ (число1; число2; ...)
выборочная дисперсия: ДИСП.Г (число1; число2; ...)
выборочное СКО: СТАНДОТКЛОН.Г (число1; число2; ...) или КОРЕНЬ (ДИСП.Г (число1; число2; ...))

Мода для выборки из *дискретного* распределения - значение во множестве наблюдений, которое встречается наиболее часто. Иногда в совокупности встречается более чем одна мода (например: 6, 2, 6, 6, 8, 9, 9, 9, 10; мода = 6 и 9).

Немного сложнее с *интервальными* данными, когда вместо конкретных значений в выборке имеются интервалы. В этом случае говорят о **модальном интервале**, т.е. интервале, частота которого максимальна относительно других интервалов. Однако и здесь можно отыскать конкретное модальное значение, хотя оно будет условным и примерным, так как нет точных исходных данных. Есть общее правило, по которому рассчитывается мода в интервальных данных. Представим, что есть маленький набор данных, как в табл. 1.4. Для наглядности изобразим соответствующую гистограмму (рис. 1.10). Требуется найти модальное значение.

Таблица 1.4

Пример интервального ряда для расчета моды

Интервал	Частота
101-200	20
201-300	50
301-400	60
401-500	40
501-600	25

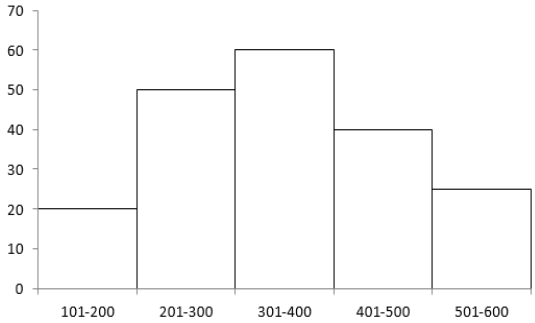


Рис. 1.10. **Гистограмма распределения для примера интервального ряда табл. 1.4**

Вначале нужно определить модальный интервал, который соответствует интервалу с наибольшей частотой. В этом небольшом примере это третий интервал от 301 до 400. На графике это самый высокий столбец.

Теперь нужно определить конкретное значение в выборке, которое соответствует максимальному количеству.

Делается допущение о том, что интервалы выше и ниже модального в зависимости от своей частоты имеют разный вес и "перетягивают" моду в свою сторону.

Если частота интервала, следующего за модальным, больше, чем частота интервала перед модальным, то мода будет правее середины модального интервала, и наоборот (рис. 1.11).

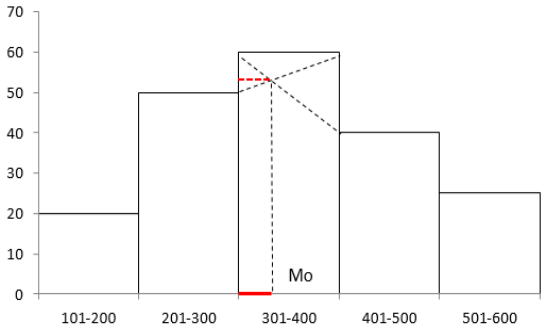


Рис. 1.11. **Модальный интервал**

На рис. 1.11 отчетливо видно, что соотношение высоты столбцов, расположенных слева и справа от модального, определяет близость моды

к левому или правому краю модального интервала. Задача по расчету модального значения состоит в том, чтобы найти точку пересечения линий, соединяющих модальный столбец с соседними (как показано на рис. 1.11 пунктирными линиями), и соответствующего значения признака. По данному рисунку нетрудно вывести формулу расчета моды в интервальном ряду.

Формула моды имеет следующий вид:

$$MO = x_{MO} + h_{MO} \frac{n_{MO} - n_{MO-1}}{\left(n_{MO} - n_{MO-1}\right) + \left(n_{MO} - n_{MO+1}\right)},$$

где MO - мода;

x_{MO} - значение начала модального интервала;

h_{MO} - ширина модального интервала;

n_{MO} - частота модального интервала;

n_{MO}^{-1} - частота интервала, находящегося перед модальным;

n_{MO}^{+1} - частота интервала, находящегося после модального.

Второе слагаемое формулы моды соответствует длине красной линии на рисунке выше. Рассчитаем моду для этого примера:

$$MO = 301 + 100 \frac{60 - 50}{(60 - 50) + (60 - 40)} = 334,3.$$

Таким образом, мода интервального ряда представляет собой сумму, состоящую из значения начального уровня модального интервала и отрезка, который определяется соотношением частоты ближайших интервалов от модального.

Если в ряду есть наблюдения с одинаковыми значениями, то все просто:
= МОДА.ОДН - рассчитывает моду по заданным значениям.
= МОДА.НСК - позволяет рассчитать сразу несколько модальных значений (одинаковых максимальных частот) для одного ряда данных, если они есть. Функцию нужно вводить как формулу массива, перед этим выделив количество ячеек, равное количеству требуемых модальных значений. Иногда действительно модальных значений может быть несколько. Однако для этих целей предварительно лучше посмотреть на диаграмму распределения.

В задании же лабораторной работы полученный при генерации вариационный ряд не содержит повторяющихся значений.

Моду для интервальных данных одной функцией в *Excel* рассчитать **нельзя** - нужно набирать формулу вручную.

Для выборки из **равномерного** закона моду рассчитывать **не нужно**.

Медиана для выборки из дискретного распределения - срединное значение для ранжированного ряда, половина чисел имеет значения бóльшие, чем медиана, а другая половина чисел - меньшие (например: 1, 2, 3, 4, 5, то медиана =3; если число значений в ряду четно, например: 1, 2, 3, 4, 5, 6, то медиана равна (3+4) / 2).


МЕДИАНА (число1; число2;...).

Так происходит поиск или расчет медианы в дискретных данных.

Однако данные, как в нашем случае, могут быть еще и **интервальными**, где выбрать конкретное значение не представляется возможным, так как конкретных значений просто нет.

Как и в моде, медиану в таком случае рассчитывают по некоторому общепринятому правилу.

Для начала находят **медианный интервал**.

Это такой интервал, через который проходит искомое медианное значение. Определяется он с помощью накопленной доли ранжированных интервалов. Где накопленная доля впервые перевалила через 50 % всех значений, там и медианный интервал со скрытой внутри медианой. Исходим из предположения, что распределение данных внутри медианного интервала равномерное (т. е. 30 % ширины интервала - это 30 % значений, 80 % ширины - 80 % значений и т. д.). Отсюда, зная количество значений от начала медианного интервала до 50 % всех значений совокупности (разница между половиной количества всех значений и накопленной частотой предмедианного интервала), можно найти, какую долю они занимают во всем медианном интервале.

Эта доля переносится на ширину медианного интервала, указывая на конкретное значение, именуемое медианой. Для небольшого примера рассчитаем медиану по следующим данным (табл. 1.5, рис. 1.12).

Таблица 1.5

Пример интервального ряда для расчета медианы

Интервал	Частота
100-200	20
200-300	50
300-400	60
400-500	40
500-600	30

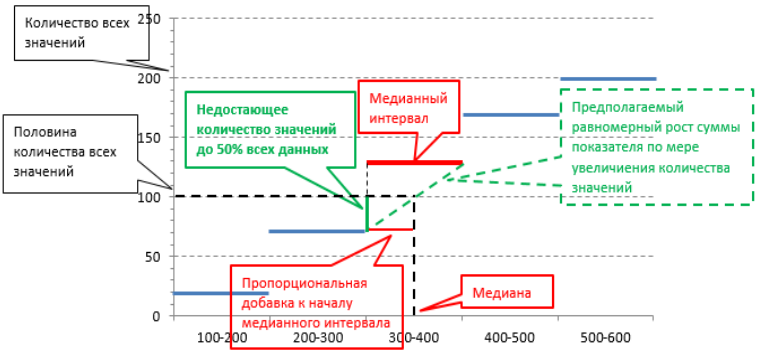


Рис. 1.12. **Расчет медианы по интервальным данным**

Формула медианы имеет следующий вид:

$$ME = x_{ME} + h_{ME} \frac{n/2 - S_{ME-1}}{n_{ME}},$$

где x_{ME} - начало медианного интервала;

h_{ME} - ширина медианного интервала;

$n/2$ - количество всех значений, деленное на два;

$S_{(ME-1)}$ - суммарное количество наблюдений, которое было накоплено до начала медианного интервала, т.е. накопленная частота предмедианного интервала;

n_{ME} - частота медианного интервала.

Требуется найти медиану, т.е. то значение, меньше и больше которого по половине количества всех наблюдений. Для начала произведем расчет эмпирической функции распределения (табл. 1.6).

Таблица 1.6

Пример интервального ряда для расчета медианы

№ п/п	Интервал	Частота	Накопленная частота	Э.ф.р
1	100-200	20	20	0,1
2	200-300	50	**70**	0,35
3	**300**-400	**60**	130	0,65
4	400-500	40	170	0,85
5	500-600	30	200	1
Сумма		200		

По последней колонке определяем медианный интервал - 300-400 (накопленная доля впервые более 50 %, в нашем случае - 0,65, т. е. 65 %).

Ширина интервала - 100. Теперь остается подставить данные в приведенную выше формулу и рассчитать медиану:

$$ME = 300 + 100\frac{\frac{200}{2} - 70}{60} = 350.$$

Медиану по интервальным данным одной функцией в *Excel* рассчитать **нельзя** - нужно набирать формулу вручную.

Расчет первого и третьего квартилей распределения Q_1, Q_3 по исходным данным осуществляется двумя способами.

Обозначим p - порядок квартиля (в нашем случае - 0,25 и 0,75 для Q_1, Q_3, соответственно).

$$Q_{4p} = x_{\lfloor v \rfloor} + (v - \lfloor v \rfloor)(x_{\lfloor v \rfloor + 1} - x_{\lfloor v \rfloor}),$$

где v - вспомогательное число, рассчитывающееся двумя разными способами.

I способ - основной (название в *Excel* 2016 - **инклюзивная** медиана):

$$v = (n-1)p + 1$$

II способ (название в *Excel* 2016 - **эксклюзивная** медиана):

$$v = (n+1)p$$

С помощью данных формул можно также посчитать по выборке квантиль любого порядка p.

Заметим, что формулы медианы по интервальному ряду и по исходным данным дают приблизительно одинаковые (но не равные) результаты.

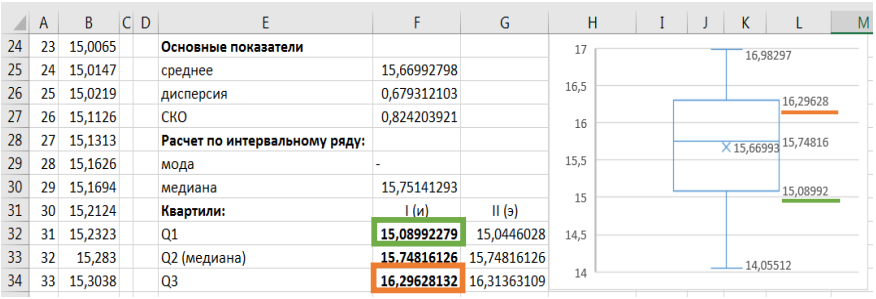
Чтобы построить график "ящик с усами", нужно выделить диапазон исходных данных и выбрать *Вставка - Статистическая гистограмма - Ящик с усами.*

В настройках графика выставите *Инклюзивная медиана*, чтобы Q_1, Q_3 были рассчитаны первым способом.

Отобразите подписи данных. **Убедитесь, что ваши расчеты совпадают с теми, что рассчитал *Excel* для графика**.

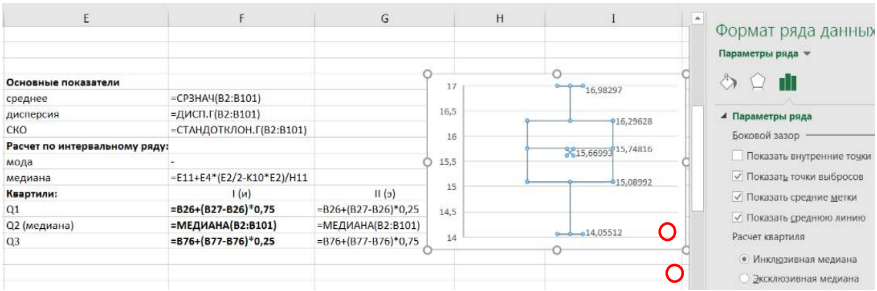
В качестве длин "усов" *Excel* отображает минимум и максимум по выборке.

На рис. 1.13 и 1.14 представлены расчеты для нашей выборки из примера.



	A	B	C	D	E	F	G
24	23	15,0065			**Основные показатели**		
25	24	15,0147			среднее	15,66992798	
26	25	15,0219			дисперсия	0,679312103	
27	26	15,1126			СКО	0,824203921	
28	27	15,1313			**Расчет по интервальному ряду:**		
29	28	15,1626			мода	-	
30	29	15,1694			медиана	15,75141293	
31	30	15,2124			**Квартили:**	I (и)	II (э)
32	31	15,2323			Q1	**15,08992279**	15,0446028
33	32	15,283			Q2 (медиана)	**15,74816126**	15,74816126
34	33	15,3038			Q3	**16,29628132**	16,31363109

Рис. 1.13. **Расчет числовых характеристик выборочного распределения и "ящик с усами"**



	E	F	G
	Основные показатели		
	среднее	=СРЗНАЧ(B2:B101)	
	дисперсия	=ДИСП.Г(B2:B101)	
	СКО	=СТАНДОТКЛОН.Г(B2:B101)	
	Расчет по интервальному ряду:		
	мода	-	
	медиана	=E11+E4*(E2/2-K10*E2)/H11	
	Квартили:	I (и)	II (э)
	Q1	**=B26+(B27-B26)*0,75**	=B26+(B27-B26)*0,25
	Q2 (медиана)	**=МЕДИАНА(B2:B101)**	=МЕДИАНА(B2:B101)
	Q3	**=B76+(B77-B76)*0,75**	=B76+(B77-B76)*0,75

Рис. 1.14. **Расчет числовых характеристик выборочного распределения в режиме формул**

Заметим, что расчет моды не выполнен, так как не имеет смысла, ибо выборка в примере извлечена из равномерного закона распределения. В случае же нормального закона необходимо самостоятельно рассчитать моду.

Построим теперь **гистограмму распределения с помощью встроенной гистограммы** *Excel* (рис. 1.15).

В данном случае *Excel* разбивает выборку также на равноотстоящие интервалы, только берет за начало первого интервала минимальное значение в выборке, а за конец последнего - максимальное.

Выделить диапазон исходных данных. Выбрать *Вставка - Статистическая гистограмма - Гистограмма*.
Выделить на получившейся гистограмме ось *X*, вызвав меню правой кнопкой мыши, выбрать *Формат оси* и выставить число интервалов такое же, как в гистограмме из п. 2в (в примере - 8 интервалов). Указать число знаков после запятой, равное трем.

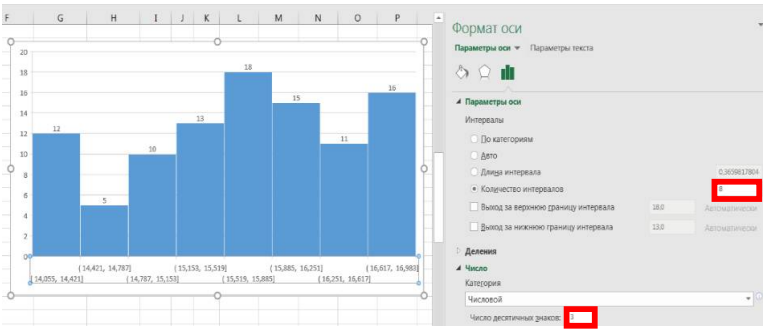


Рис. 1.15. **Встроенная гистограмма распределения**

Таким образом, пп. 2 и 3 приводят к следующему результату (пример выполнения на листе в *Excel* на рис. 1.16):

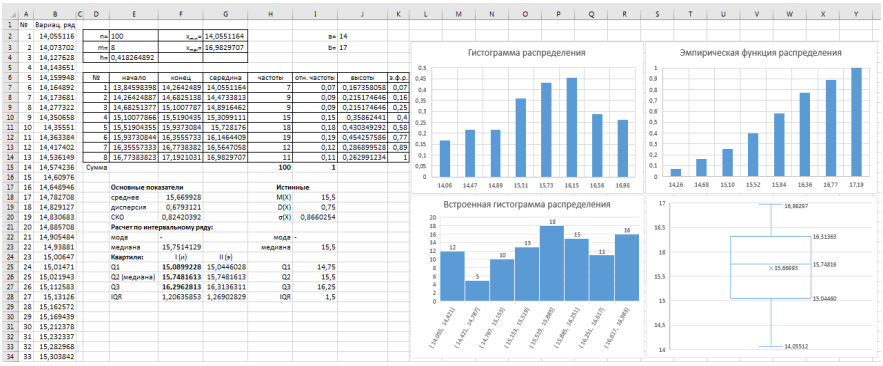


Рис. 1.16. **Пример выполнения п. 2**

Выполните самостоятельно п. 4 задания. Далее нужно проделать пп. 2, 3, 4 для нормального закона распределения, а также сгенерировать 1000 значений выборки, проделать аналогичные пункты и сравнить результаты расчетов.

Отчет должен содержать следующие результаты. В п. 1 нужно привести заполненную табл. 1.1 и построенные 14 графиков функций. Не следует приводить данные для построения графиков. На графике должны быть указаны: название (какой закон распределения, какие у него параметры), подписаны оси абсцисс и ординат. Линии сетки обязательны. Легенда нужна, если на одной диаграмме несколько графиков.

Главный критерий качества графиков: они должны быть легко читаемы и понятны человеку, **не читавшему задание этой лабораторной работы**. В пп. 2-4 для каждого из 4 случаев (100 значений, равномерное; 1000 значений, равномерное; 100 значений, нормальное; 1000 значений,

нормальное) нужно привести заполненную табл. 1.3, четыре графика: два - гистограммы (своей и встроенной), эмпирической функции распределения, "ящик с усами", а также рассчитанные выборочные характеристики.

Графики должны быть оформлены аналогично п. 2а, т.е. с подписями названий, осей и легендой (если требуется для читаемости). В п. 4 нужно указать, как изменились выборочные характеристики выборки в 1000 наблюдений от выборки в 100 наблюдений, похожи ли они на теоретические, рассчитанные в табл. 1.1, стали ли они точнее? Постараться понять, что и почему с ними произошло, и пояснить. Не нужно вставлять сгенерированные выборки.

Генерация случайных выборок с заданными статистическими характеристиками широко используется при проведении **компьютерных экспериментов, заменяя натурные, по методу Монте-Карло.**

Контрольные вопросы

1. Чем отличаются теоретические и выборочные числовые характеристики?
2. Какие способы разбиения выборки на интервалы вам известны?
3. Объясните два способа расчета выборочной моды и медианы.
4. Какие свойства оценок вам известны? Что можно сказать об оценках параметров распределений при увеличении объемов выборок?
5. Как построить "ящик с усами"?
6. Рассчитайте для своей выборки в 100 значений из нормального закона распределения выборочный 77 % квантиль.

Лабораторная работа 2
Парная линейная регрессия

Цель работы: научиться оценивать значения параметров линейной (и нелинейной, но сводимой к линейной) парной регрессии; качество полученной регрессии и параметров; сравнивать модели между собой и осуществлять их выбор.

Задание. Лабораторная работа выполняется в *Microsoft Excel* 2016. Варианты работы находятся в дополнительном файле "Варианты_ЛР2.xlsx".

Имеются две выборки равного объема для показателей Y и X. Предполагается наличие зависимости уровней Y от уровней X.

1. Визуальный анализ исходных данных.

Построить график зависимости Y от X. Выдвинуть предположение о наличии и функциональном виде зависимости.

2. Корреляционный анализ.

Проверить наличие корреляционной зависимости между Y и X, Y и X_2, Y и $1/X$, Y и lnX, lnY и X, lnY и lnX.

3. Идентификация параметров моделей регрессии:

1) линейной;
2) параболической;
3) логарифмической;
4) обратной (гиперболической);
5) показательной (экспоненциальной);
6) степенной.

4. Проверка общего качества моделей.

Проверить общее качество полученных уравнений регрессии. Сравнить с результатами, полученными в п. 2. Отсеять неудовлетворительные по точности модели.

5. Выбор наилучшей модели.

Из оставшихся моделей выбрать наилучшую, учитывая количественные критерии и качественный анализ.

6. Оценка качества идентификации параметров.

Проверить статистическую значимость оценок параметров для выбранной модели и построить их доверительные интервалы.

7. Прогнозирование по модели.

Построить точечную и интервальную оценку Y при заданном значении $X=x$. Если в п. 4 были отсеяны все модели, п/п. 6-7 рассчитать для линейной модели. Если в п. 5 была выбрана параболическая модель, п. 6 рассчитать для линейной модели.

Указания к выполнению

Модели парной регрессии

Общий вид модели парной линейной регрессии

Модели **пространственной** динамики описывают взаимосвязи между различными показателями. Модели **временно́й** динамики описывают развитие одного фактора во времени. **Смешанные** пространственно-временные модели описывают взаимодействие нескольких показателей с учетом их развития во времени.

Простейшая пространственная модель - модель **парной линейной регрессии:**

$$Y_k = \beta_0 + \beta_1 X_k + \varepsilon_k, \quad M[Y \mid X] = \beta_0 + \beta_1 X,$$

где Y_k - зависимая переменная (регрессор); X_k - независимая переменная (фактор); β_0, β_1 - параметры модели; ε_k - стохастическая компонента (случайные остатки, ошибки, невязки, погрешности); $k = \overline{1,n}$ - номер наблюдения; n - общее число наблюдений.

Это *теоретическое* уравнение регрессии, для которого необходимо оценить значения ее параметров по конкретным выборкам X и Y.

В результате получим *эмпирическое* уравнение регрессии, модельные (ожидаемые) значения Y_k^* и оценки ошибок e_k:

$$Y_k^* = b_0 + b_1 X_k = m_{Y|X}, \quad e_k = Y_k - Y_k^*.$$

Этапы построения модели

1. Постановка задачи, анализ предметной области - выбор исследуемых показателей Y, X

2. Сбор исходных данных - формирование выборок (рядов), по которым будет строиться модель

$$Y_k, X_k, n$$

3. *Спецификация* модели (структурная идентификация) - определение общего вида модели

$$Y = f(\Theta, X, \varepsilon)$$

$$Y_k = b_0 + b_1 X_k + e_k$$

4. *Идентификация* модели (параметрическая идентификация) - оценка значений параметров модели

$$\Theta^* \Rightarrow Y^* = f(\Theta^*, X)$$

$$Y_k^* = b_0 + b_1 X_k$$

5. *Верификация* модели - проверка качества модели и ее параметров

6. *Интерпретация* модели - анализ результатов моделирования

Все этапы взаимосвязаны, зачастую после выполнения какого-то этапа возможно возвращение к предыдущим. Например, после спецификации модели могут потребоваться дополнительные статистические данные или после верификации модели она оказывается неудовлетворительной и приходится рассматривать другую модель.

Как правило, выдвигается несколько гипотез о возможном виде модели, из которых затем выбирается наилучшая. Критерии выбора могут быть различными.

Метод наименьших квадратов (МНК, method of Least Squares, LS)

Суть метода заключается в минимизации отклонений модели от исходных данных по всем точкам выборки. Чтобы отклонения модели (их часто называют **невязками**) с противоположными знаками не компенсировали друг друга при суммировании, они возводятся в квадрат.

Общий вид метода иллюстрируют выражения:

$$SS = \sum_{\kappa=1}^{\nu} \varepsilon_\kappa{}^2 \xrightarrow[\Theta]{} \min \text{ или}$$

$$\sum_{k=1}^{n} \left(Y_k - f(\Theta, X_k)\right)^2 \xrightarrow[\Theta]{} \min \text{ или}$$

$$\Theta^* = \underset{\Theta}{\operatorname{argmin}} \sum_{k=1}^{n} \left(Y_k - f(\Theta, X_k)\right)^2.$$

Если параметры Θ входят в модель линейно, то существует единственный минимум. Чтобы его найти, необходимо приравнять частные производные функции потерь *SS* по всем параметрам к нулю:

$$\frac{\partial SS}{\partial \Theta} = 0.$$

Для парной линейной регрессии:

$$SS = \sum_{k=1}^{n} \left(Y_k - b_0 - b_1 X_k \right)^2 \xrightarrow[b_0, b_1]{} \min$$

$$\begin{cases} \dfrac{SS}{b_0} = -2 \sum\limits_{k=1}^{n} \left(Y_k - b_0 - b_1 X_k \right) = 0 \\ \dfrac{SS}{b_1} = -2 \sum\limits_{k=1}^{n} \left(Y_k - b_0 - b_1 X_k \right) X_k = 0 \end{cases}$$

$$\begin{cases} \sum\limits_{k=1}^{n} Y_k - \sum\limits_{k=1}^{n} b_0 - \sum\limits_{k=1}^{n} b_1 X_k = 0 \\ \sum\limits_{k=1}^{n} Y_k X_k - \sum\limits_{k=1}^{n} b_0 X_k - \sum\limits_{k=1}^{n} b_1 X_k^2 = 0 \end{cases}$$

$$\begin{cases} n b_0 + b_1 \sum\limits_{k=1}^{n} X_k = \sum\limits_{k=1}^{n} Y_k \\ b_0 \sum\limits_{k=1}^{n} X_k + b_1 \sum\limits_{k=1}^{n} X_k^2 = \sum\limits_{k=1}^{n} Y_k X_k \end{cases}$$

$$\begin{cases} b_0 + b_1 \dfrac{1}{n} \sum\limits_{k=1}^{n} X_k = \dfrac{1}{n} \sum\limits_{k=1}^{n} Y_k \\ b_0 \dfrac{1}{n} \sum\limits_{k=1}^{n} X_k + b_1 \dfrac{1}{n} \sum\limits_{k=1}^{n} X_k^2 = \dfrac{1}{n} \sum\limits_{k=1}^{n} Y_k X_k \end{cases}$$

$$\begin{cases} b_0 + b_1 m_X = m_Y \\ b_0 m_X + b_1 m_{X^2} = m_{YX} \end{cases}$$

$$b_0 = m_Y - b_1 m_X$$

$$\left(m_Y - b_1 m_X \right) m_X + b_1 m_{X^2} = m_{YX}$$

$$m_Y m_X - b_1 m_X^2 + b_1 m_{X^2} = m_{YX}$$

$$b_1 = \frac{m_{YX} - m_Y m_X}{m_{X^2} - m_X^2} = \frac{\operatorname{cov}(Y, X)}{s_X^2}$$

Таким образом получим

$$b_1 = \frac{m_{YX} - m_Y m_X}{m_{X^2} - m_X^2} = \frac{\mathrm{cov}(Y,X)}{s_X^2} = r_{YX}\frac{s_Y}{s_X}, \quad b_0 = m_Y - b_1 m_X.$$

b_1 - [выборочный] коэффициент регрессии Y по X показывает, на сколько единиц изменится Y при изменении X на 1.

Отметим, что b_0 и b_1 - оценки, зависящие от конкретной выборки, т.е. *случайные величины*.

Оптимальными считаются *состоятельные*, *несмещенные* и *эффективные* оценки (BLUE-оценки - best linear unbiased estimates).

Состоятельность - при увеличении объема выборки оценка приближается к своему истинному значению (сходится по вероятности):

$$\lim_{n\to\infty} P\left(\left|\theta - \theta^*\right| \le \lambda\right) = 1, \lambda > 0\,.$$

При объеме выборки n, стремящемся к бесконечности, вероятность P того, что отклонение оценки θ^* от истинного значения θ будет меньше некоторого малого положительного числа λ, стремится к 1.

Несмещенность - среднее значение оценки равно ее истинному значению:

$$M[\theta^*] = \theta.$$

Иногда говорят об асимптотической несмещенности:

$$\lim_{n\to\infty} M[\theta^*] = \theta,$$

т.е. **несмещенность достигается при увеличении объема выборки**.

Эффективность - оценка обладает наименьшей дисперсией (среди всех методов ее идентификации):

$$D[\theta^*] \to \min.$$

Поскольку все методы учесть невозможно, говорят об эффективности оценок в некотором классе применяемых методов идентификации.

Теорема Гаусса - Маркова. Оценки параметров линейной регрессии являются эффективными в классе линейных несмещенных оценок, если выполняются следующие условия (**условия Гаусса - Маркова**) для применения МНК:

1) Математическое ожидание стохастической компоненты ε_k равно нулю:

$$M\left[\varepsilon_k\right] = 0\,.$$

Если это условие не выполняется, то модель содержит систематическую ошибку. Когда в модели присутствует свободный член (константа b_0), по результатам МНК всегда выполняется $M[e_k] = 0$. Если в исходных данных присутствовала систематическая ошибка, то она просто прибавится к b_0, поэтому обычно считают, что данное условие выполняется автоматически. Проверить его выполнение практически невозможно.

2) Дисперсия ε_k постоянна для всех наблюдений (*гомоскедастичность*):

$$\forall k \neq i \; D[\varepsilon_k] = D[\varepsilon_i] = \sigma_\varepsilon^2 = const.$$

Нарушение этого условия называется *гетероскедастичностью*, и она нередко возникает на практике. Существуют специальные тесты для обнаружения гетероскедастичности (Голдфелда - Квандта, Глейзера, ранговой корреляции Спирмена и др.). Вместо МНК тогда применяется МВНК - метод взвешенных наименьших квадратов.

3) Случайные отклонения для различных наблюдений ε_k и ε_i являются некоррелированными (*отсутствие автокорреляции в невязке*):

$$\forall k \neq i \; \mathrm{cov}[\varepsilon_k; \varepsilon_i] = 0.$$

Это условие также зачастую нарушается на практике, особенно для временных моделей. Для обнаружения автокорреляции используется тест Дарбина -Уотсона, для идентификации таких моделей применяют обобщенный метод наименьших квадратов (ОМНК). Если выполняются условия 3 и 5, то ε_k и ε_i являются независимыми.

4) Случайные отклонения ε_k должны быть некоррелированы с объясняющей переменной X:

$$\mathrm{cov}[\varepsilon_k; X_k] = 0.$$

Нарушение данного условия означает, что в случайных остатках содержится один или несколько неучтенных факторов. Решением в этом случае может быть переход от линейной модели к нелинейной или от парной регрессии к множественной. Также возможно, что неправильно определены зависимая и независимая переменные. Однако, как и для условия 1, обнаружить нарушение этого условия нельзя. Если в ε_k содержится линейная зависимость от X_k, то она будет включена в b_1 и для оценок $cov[e_k; X_k] = 0$ будет выполняться всегда.

Для временных моделей данное условие выполняется автоматически, поскольку в них в качестве независимой переменной выступает время - неслучайная величина, которая не может коррелировать со случайной.

5) Стохастическая компонента ε_k имеет нормальное распределение:

$$\varepsilon_k \sim N(0,\sigma_\varepsilon).$$

Данное условие не является обязательным и необходимо лишь для проверки качества модели и оценок параметров. На малых выборках проверить его выполнение практически невозможно. Иногда его заменяют более мягким требованием симметричности закона распределения. Обычно считают, что оно выполняется по закону больших чисел, т.е. ε_k аккумулирует все множество факторов, не учтенных в модели. При выполнении всех пяти условий Гаусса - Маркова модель парной линейной регрессии называется *классической нормальной линейной регрессионной моделью.*

Нелинейная парная регрессия

Нелинейные модели регрессии позволяют описывать более сложную форму зависимости *Y* от *X*. Различают *нелинейные по переменным* и *нелинейные по параметрам* (или *существенно нелинейные*) модели.

Наиболее распространенные нелинейные по переменным, но линейные по параметрам модели:

1) параболическая $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$;

2) полиномиальная $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_p X^p + \varepsilon$;

3) логарифмическая $Y = \beta_0 + \beta_1 In\ X + \varepsilon$;

4) обратная (гиперболическая) $Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon$.

Такие модели могут быть напрямую идентифицированы с помощью МНК. Например, для гиперболической модели можно принять:

$Q = \frac{1}{X} \Rightarrow Y = \beta_0 + \beta_1 Q + \varepsilon$ - линейная модель.

Для параболической модели необходимо по МНК оценить значения уже трех параметров:

$$SS = \sum_{k=1}^{n} \left(Y_k - b_0 - b_1 X_k - b_2 X_k^2 \right)^2 \xrightarrow[b_0, b_1, b_2]{} \min$$

$$\begin{cases} \dfrac{\partial SS}{\partial b_0} = -2\sum\limits_{k=1}^{n}\left(Y_k - b_0 - b_1 X_k - b_2 X_k^2\right) = 0 \\ \dfrac{\partial SS}{\partial b_1} = -2\sum\limits_{k=1}^{n}\left(Y_k - b_0 - b_1 X_k - b_2 X_k^2\right) X_k = 0 \\ \dfrac{\partial SS}{\partial b_2} = -2\sum\limits_{k=1}^{n}\left(Y_k - b_0 - b_1 X_k - b_2 X_k^2\right) X_k^2 = 0 \end{cases}$$

$$\begin{cases} nb_0 + b_1\sum\limits_{k=1}^{n} X_k + b_2\sum\limits_{k=1}^{n} X_k^2 = \sum\limits_{k=1}^{n} Y_k \\ b_0\sum\limits_{k=1}^{n} X_k + b_1\sum\limits_{k=1}^{n} X_k^2 + b_2\sum\limits_{k=1}^{n} X_k^3 = \sum\limits_{k=1}^{n} Y_k X_k \\ b_0\sum\limits_{k=1}^{n} X_k^2 + b_1\sum\limits_{k=1}^{n} X_k^3 + b_2\sum\limits_{k=1}^{n} X_k^4 = \sum\limits_{k=1}^{n} Y_k X_k^2 \end{cases}$$

$$\begin{cases} b_0 + b_1 m_X + b_2 m_{X^2} = m_Y \\ b_0 m_X + b_1 m_{X^2} + b_2 m_{X^3} = m_{YX} \\ b_0 m_{X^2} + b_1 m_{X^3} + b_2 m_{X^4} = m_{YX^2} \end{cases}$$

Данную систему проще решать в матричном виде:

$$B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \quad A = \begin{pmatrix} 1 & m_X & m_{X^2} \\ m_X & m_{X^2} & m_{X^3} \\ m_{X^2} & m_{X^3} & m_{X^4} \end{pmatrix} \quad C = \begin{pmatrix} m_Y \\ m_{YX} \\ m_{YX^2} \end{pmatrix}$$

$$AB = C$$

$$B = A^{-1}C$$

Модели, нелинейные по параметрам, зачастую могут быть **линеаризованы** с помощью различных преобразований исходного выражения.

Наиболее распространены на практике следующие *нелинейные по параметрам* модели и преобразования:

5) показательная (экспоненциальная) $Y = a\gamma^X \cdot e = e^{\beta_0 + \beta_1 X} \cdot e$:

$$Z = \ln Y = \ln\left(e^{\beta_0 + \beta_1 X} \cdot \varepsilon\right) = \beta_0 + \beta_1 X + \ln\varepsilon = \beta_0 + \beta_1 X + \xi;$$

6) степенная $Y = aX^{\beta_1} \cdot e$:

$$Z = \ln Y = \ln\left(\alpha X^{\beta_1} \cdot \varepsilon\right) = \ln\alpha + \beta_1 \ln X + \ln\varepsilon = \beta_0 + \beta_1 U + \xi;$$

где $\beta_0 = \ln\alpha$, $U = \ln X$, $\xi = \ln\varepsilon$.

Чтобы такая реализация была возможна, стохастическая компонента ε включается в модель как **мультипликативная**. После линеаризации новая стохастическая компонента ξ уже будет входить в модель **аддитивно.**

Применяя МНК к линеаризованным моделям, необходимо учитывать, что условия Гаусса - Маркова должны выполняться уже для ξ, а не для ε . Несмещенность, эффективность и состоятельность также гарантируются для β_0, β_1, а не для α, γ.

Тогда про ε можно сказать, что она должна иметь несимметричное логнормальное распределение, а $M[\varepsilon] = 1$, $D[\varepsilon] = const$, автокорреляция и корреляция с X отсутствуют.

Визуально модель будет выглядеть как гетероскедастическая, поскольку ε - это относительная погрешность и условие $D[\varepsilon] = const$ означает постоянство доли ошибки, а не ее абсолютного значения (рис. 2.1).

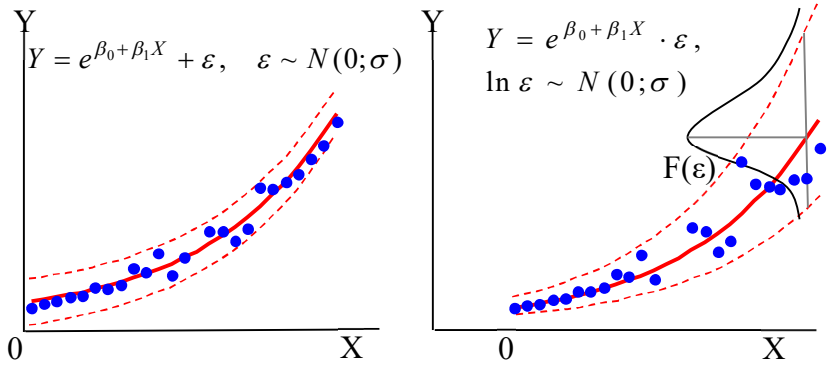


Рис. 2.1. **Аддитивная и мультипликативная модели**

Из-за этого обстоятельства мультипликативные и аддитивные модели трудно сравнивать между собой. Формально для мультипликативной модели можно вычислить погрешности $e_k = Y_k - Y_k^*$, но характеристики их будут другими. Налагаемые условия на структуру вхождения стохастической компоненты и на закон распределения являются существенными ограничениями.

Если модель линеаризовать не удается, то применяют **нелинейный МНК** (НМНК), т.е. численное решение задачи минимизации и системы

нелинейных алгебраических уравнений. Для НМНК доказана лишь состоятельность получаемых оценок параметров при выполнении условий Гаусса - Маркова.

Проверка качества уравнения регрессии

Как правило, выдвигается несколько гипотез о возможном виде модели, из которых затем выбирается наиболее адекватная. **Адекватность** модели означает ее соответствие исходным данным с точки зрения цели моделирования. Следует отличать *адекватность* модели от ее **точности**, т.е. близости модельных данных к исходным. Адекватность является более общим понятием, включающим в себя точность. Кроме того, точность - это количественная характеристика, а адекватность - качественная:

$$\text{низкая точность} \Rightarrow \text{низкая адекватность}$$

$$\text{высокая точность} \not\Rightarrow \text{высокая адекватность}$$

Таким образом, проверка точности модели необходима, но не достаточна для обоснования ее адекватности. Дополнительно требуется качественное обоснование модели. Проверка качества уравнения регрессии включает проверку ее общей точности, а также точности оценок параметров. Следует проверить: а) является ли модель и ее параметры статистически значимыми (позволяет ли имеющаяся выборка судить о генеральной совокупности, т.е. являются ли они **репрезентативными**); б) какова практическая ценность модели (насколько информация, полученная по модели, снимает неопределенность в отношении регрессора Y).

Критерии общей точности модели

Существует большое количество критериев точности, используемых как в экономике, так и в технике. Наиболее простыми можно считать **средние абсолютное и относительное отклонения** (mean absolute error, mean absolute percentage error):

$$MAE = \frac{1}{n}\sum_{k=1}^{n}\left|Y_k - Y_k^*\right|,$$

$$MAPE = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{Y_k - Y_k^*}{Y_k}\right| \cdot 100\%.$$

Первый критерий выражается в тех же единицах измерения, что и моделируемый ряд динамики, поэтому он может быть использован только для сравнения моделей одного и того же ряда динамики. С другой

стороны, он позволяет оценить "физическое содержимое" погрешности моделирования.

Второй критерий (среднее относительное отклонение) является безразмерной величиной и позволяет судить о точности модели, сравнивать их между собой. Однако ее значение также во многом зависит от уровней ряда динамики. Если значения Y_k велики по сравнению со своим разбросом, то MAPE-оценка будет малой, а если Y_k близки к нулю, то MAPE-оценка будет большой вне зависимости от точности модели. Если же имеются наблюдения, строго равные нулю, то использовать относительные величины вообще невозможно.

Коэффициент корреляции (выборочный) определяет силу (степень тесноты) линейной зависимости (связи) между показателями:

$$r_{YX} = \frac{cov(Y;X)}{s_Y s_X} = \frac{m_{YX} - m_Y m_X}{\sqrt{\left(m_{Y^2} - m_Y^2\right)\left(m_{X^2} - m_X^2\right)}} = r_{XY}, \quad -1 \le r_{YX} \le 1.$$

Чем ближе модуль $|r_{YX}|$ к 1, тем точнее линейная модель. При $|r_{YX}| = 1$ между X и Y существует функциональная зависимость, при $|r_{YX}| = 0$ линейная связь полностью отсутствует. Знак r_{XY} определяет направление зависимости (положительная или отрицательная).

Для нелинейных моделей коэффициент корреляции рассчитывается **для их линеаризованной формы**, например, для логарифмической модели рассчитывается так:

$$r_{YlnX} = \frac{cov(Y;lnX)}{s_Y s_{lnX}}.$$

Как и оценки параметров моделей, выборочный коэффициент корреляции является также случайной оценкой, зависящей от выборки. Поэтому его значение следует оценивать с позиций статистики.

Проверка *гипотезы о статистической значимости коэффициента корреляции:*

$H_0: \quad r_{XY} = 0$ - линейная зависимость отсутствует,

$H_1^{(1)}: \quad r_{XY} \neq 0$ - линейная зависимость присутствует.

Для проверки гипотезы рассчитывается специальный показатель *t*-статистика:

$$t_r = \frac{r_{XY}\sqrt{n-2}}{\sqrt{1 - r_{XY}^2}},$$

которая сравнивается с критической точкой распределения Стьюдента $t_{\frac{\alpha}{2},n-2}$, где α - уровень значимости, $n-2$ - число степеней свободы, количество наблюдений минус число параметров.

Если $|t_r| < t_{1-\frac{\alpha}{2},n-2}$, то нет оснований для отклонения H_0. Если $|t_r| \geq t_{1-\frac{\alpha}{2},n-2}$, то H_0 отклоняется в пользу $H_1^{(1)}$.

Обычно считают, что если $|r_{YX}| \geq 0,5$, то линейная зависимость отсутствует, если $|r_{YX}| \geq 0,7$ - линейная зависимость слабая, если $|r_{YX}| \geq 0,9$ - присутствует сильная линейная зависимость.

Коэффициент детерминации (coefficient of determination) R^2:

$$R^2 = 1 - \frac{\sum_{k=1}^{n}\left(Y_k^* - Y_k\right)^2}{\sum_{k=1}^{n}\left(Y_k - M\left[Y_k\right]\right)^2} = 1 - \frac{\sum_{k=1}^{n} e_k^2}{\sum_{k=1}^{n}\left(Y_k - M\left[Y_k\right]\right)^2}.$$

Здесь $ESS = \sum_{k=1}^{n}\left(Y_k^* - Y_k\right)^2$ (сумма квадратов отклонений, unexplained sum of squares) - мера остаточного, не объясненного моделью разброса исходных данных. $TSS = \sum_{k=1}^{n}\left(Y_k^* - M\left[Y_k\right]\right)^2$ (общая сумма квадратов, total sum of squares) - мера общего рассеивания Y_k относительно линии математического ожидания. $M\left[Y_k\right]$.

Смысл R^2: показать, какая доля зависимого показателя не является случайной, т.е. описывается моделью. Другая интерпретация: насколько лучше найденная модель объясняет исследуемую зависимость, чем горизонтальная прямая $M\left[Y_k\right]$.

Для линейных моделей $0 \leq R^2 \leq 1$. Чем ближе коэффициент детерминации к единице, тем точнее модель. При $R^2 = 1$ модель проходит точно через исходные данные. Для нелинейных моделей коэффициент детерминации может быть отрицательным $R^2 \in -\infty;1$, если модель совершенно не объясняет показатель (даже хуже, чем просто горизонтальная прямая). Причинами могут быть, например, вычислительная ошибка, неверный выбор вида модели.

Для линейных моделей можно использовать и другую формулу:

$$R^2 = \frac{\sum_{k=1}^{n}\left(Y_k^* - M\left[Y_k\right]\right)^2}{\sum_{k=1}^{n}\left(Y_k - M\left[Y_k\right]\right)^2},$$

где $ESS = \sum_{k=1}^{n}\left(Y_k^* - M\left[Y_k\right]\right)^2 = TSS - USS$ (объясненная сумма квадратов, explained sum of squares) - мера объясненного моделью разброса.

Также для линейных моделей $R^2 = r_{XY}^2$. Поэтому для нелинейных моделей иногда рассчитывают индекс корреляции $\rho_{XY} = \sqrt{R^2}$.

F-тест используется для проверки качества уравнения регрессии, т.е. статистической значимости самого уравнения и коэффициента детерминации.

$H_0: \quad R^2 = 0$ - модель статистически **НЕ значима**,

$H_1^{(1)}: \quad R^2 > 0$ - модель статистически значима.

F-критерий Фишера (F-статистика):

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

При выполнении указанных выше пяти условий Гаусса - Маркова случайная величина F имеет распределение Фишера $F_{1-\alpha;m-1;n-m}$, где m - число параметров модели; α - уровень значимости (вероятность отвергнуть модель, когда она статистически значима).

Если $F > F_{1-\alpha;m-1;n-m}$, то нулевую гипотезу H_0 следует отклонить, т. е. принять модель и R^2 статистически значимыми и надежными.

Критерии выбора модели

Основным недостатком R^2 является то, что при усложнении модели он возрастает и поэтому не может служить достоверным критерием выбора одной модели из нескольких. Рекомендацией может быть использование **скорректированного коэффициента детерминации**, который учитывает число степеней свободы модели (усложнение модели):

$$\overline{R}^2 = 1 - \frac{USS/(n-m-1)}{TSS/(n-1)} = 1 - (1-R^2)\cdot\frac{n-1}{n-m-1} \le R^2,$$

но однако не учитывает линейность/нелинейность модели.

Можно применить и *информационные критерии* выбора модели с учетом числа параметров.

Информационный критерий Акаике (*Akaike information criterion*):

$$AIC = n\ln USS + 2m,$$

где m - количество параметров в модели.

Абсолютное значение AIC не имеет смысла, данный критерий может служить лишь для сравнения моделей. Чем меньше величина AIC, тем лучше модель соответствует исходным данным.

Критерий Шварца (Schwarz Criterion), или Байесовский информационный критерий (Bayesian information criterion):

$$SC = n\ln USS + m\ln n.$$

Чем меньше величина SC, тем модель лучше, причем критерий Шварца более "чутко" относится к увеличению числа параметров модели, чем Акаике. Оба критерия SC и AIC рекомендуется использовать для задач большой размерности ($n/m > 40$).

Известен и **скорректированный критерий Акаике**:

$$AIC_c = AIC + \frac{2m(m+1)}{n-m-1}.$$

Какой из перечисленных критериев использовать и как учитывать нелинейность модели, остается на усмотрение эксперта.

Проверка качества оценок параметров

При выполнении условий Гаусса - Маркова можно проверить гипотезу о статистической значимости оценок параметров, т.е. их отличие от нуля:

$H_0: \ \theta^* = 0$ - параметр статистически НЕ значим,

$H_1^{(1)}: \ \theta^* \neq 0$ - параметр статистически значим.

Для проверки этой гипотезы строится двухсторонняя **t-статистика**:

$$t_\theta = \frac{\theta^*}{S_\theta},$$

где S_θ - стандартные ошибки параметров θ.

Для парной линейной регрессии:

$$S_{b_1}^2 = \frac{\sum e_k^2}{(n-2)\sum (X_k - m_X)^2};$$

$$S_{b_0}^2 = \frac{\sum e_k^2 \sum X_k^2}{n(n-2)\sum\left(X_k - m_X\right)^2} = m_{X^2} S_{b_1}^2.$$

Для нелинейных моделей вместо X необходимо подставлять соответствующие факторы ($1/X$, $\ln X$). Значение t-статистики сравнивается с критическим значением распределения Стьюдента $t_{kp} = t_1 - \frac{\alpha}{2}, n-2$.

Если $|t_\theta| \geq t_{kp}$, то параметр считается статистически значимым.

Используя t-статистику, можно также построить **доверительный интервал** значений параметра θ:

$$\left(\theta^* - t_{кр} S_\theta ; \theta^* + t_{кр} S_\theta\right)$$

и доверительный интервал прогнозного значения регрессии при заданном $X{=}x$ (M $[Y|X{=}x]$):

$$(Y^* - t_{kp}\triangle; Y^* + t_{kp}\triangle), \triangle = \sqrt{\frac{\sum e^2}{n-2}\left(1 + \frac{1}{n} + \frac{(m_x - x)^2}{\sum\left(X_k - m_X\right)^2}\right)}.$$

1. **Визуальный анализ исходных данных**

Рассмотрим пример. Исследуется зависимость распространенности сети Интернет от общего уровня потребления населения по России. Требуется определить ожидаемый уровень доступа в Интернет по Самарской области при увеличении потребления на 2 %.

Для исследования были выбраны следующие статистические показатели (www.gks.ru, www.fedstat.ru):

– доля домохозяйств, имеющих доступ к сети Интернет, на конец года, %;

– совокупные денежные расходы домохозяйств за год в текущих ценах, руб.

Значения показателей за 2011 г. доступны для большинства субъектов Российской Федерации ($n = 79$). Однако значения совокупных денежных расходов не являются сопоставимыми для различных регионов из-за разного уровня цен.

Чтобы перевести показатель из текущих цен в сопоставимые, воспользуемся дополнительными сведениями о текущей стоимости фиксированного набора потребительских товаров и услуг, руб.

Таким образом, хотя модель строится для двух показателей, исходные статистические данные содержат три выборки: X_0 и X_1, на основе которых рассчитывается X и Y (рис. 2.2, 2.3).

	A	B	C	D	E	F	G	H	I
1		Y	%	Доля лиц (домохозяйств), имеющих доступ к сети Инте					
2		X0	руб.	Денежные расходы домохозяйств в текущих ценах					
3		X1	руб.	Стоимость фиксированного набора потребительских то					
4		X	млн. руб.	Денежные расходы домохозяйств в сопоставимых цена					
5									
6		**Y**	**X0**	**X1**	**X**				
7	Российская Федерация	37,0	15 225 580,00	9 072,71	15,23				
9	г.Москва	71,9	29 186 272,00	12 728,09	29,19				
10	г.Санкт-Петербург	57,9	24 076 267,00	9 557,75	24,08				
11	Алтайский край	28,6	11 631 089,00	7 859,78	11,63				
12	Амурская область	34,3	12 307 217,00	10 142,35	12,31				
13	Архангельская область	44,9	16 241 259,00	10 347,76	16,24				
14	Астраханская область	39,0	12 896 105,00	8 072,68	12,90				
15	Белгородская область	31,3	12 792 408,00	7 862,71	12,79				

Рис. 2.2. **Расчет значений фактора X**

E7 =C7/10^6*$D$7/D7

	A	B	C	D	E	F
6		**Y**	**X0**	**X1**	**X**	
7	Российская Федерация	37	15225580	9072,71	=C7/10^6*$D$7/D7	
9	г.Москва	71,9	29186272	12728,09	=C9/10^6*$D$7/D9	
10	г.Санкт-Петербург	57,9	24076267	9557,75	=C10/10^6*$D$7/D10	
11	Алтайский край	28,6	11631089	7859,78	=C11/10^6*$D$7/D11	
12	Амурская область	34,3	12307217	10142,35	=C12/10^6*$D$7/D12	
13	Архангельская область	44,9	16241259	10347,76	=C13/10^6*$D$7/D13	

Рис. 2.3. **Расчет значений фактора X в режиме отображения формул**

Y_k - зависимая (эндогенная) переменная, регрессор;

$X_k = \frac{X_k^0}{1\,000\,000}\frac{X_{PФ}^1}{X_k^1}$ - независимая (экзогенная) переменная, фактор.

В качестве эталонного уровня цен выбрано значение $X_{PФ}^1$ в среднем по России.

В других расчетах данные по РФ не используются, поскольку являются усреднением всех остальных значений.

Построим график зависимости Y от X (рис. 2.4).

Используйте диаграмму типа "Точечная", без линий. Проверьте правильность выбора осей, выберите удобный масштаб. Включите вертикальные линии сетки, отметьте названия и единицы измерения осей.

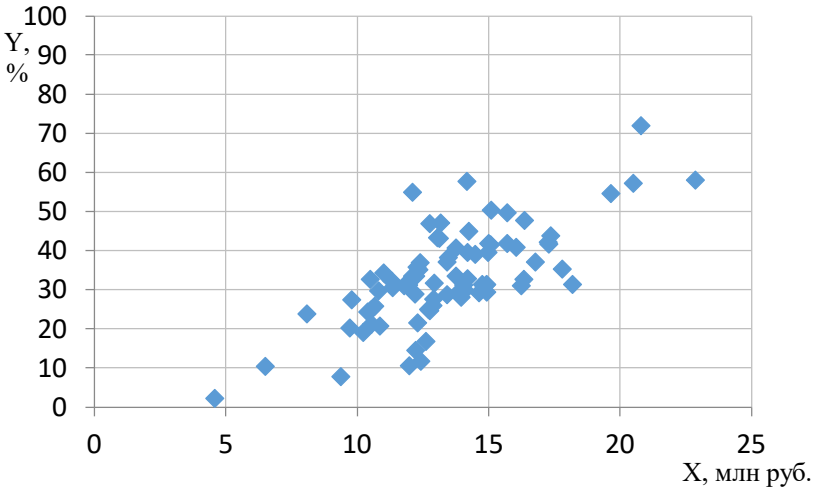


Рис. 2.4. **График зависимости *Y* от *X***

График позволяет *предположить* наличие положительной линейной взаимосвязи между исследуемыми показателями, однако видим, что разброс значений вокруг предполагаемой линии регрессии достаточно велик. Тем не менее выдвинуть какие-либо предположения о характере нелинейной зависимости в данном случае по графику довольно сложно: отсутствуют видимые минимумы или максимумы, нельзя сказать, выпуклая или вогнутая должна быть функция.

2. Корреляционный анализ

Чтобы проверить возможные виды зависимостей, рассчитаем парные коэффициенты корреляции между *Y* и *X*, *Y* и X^2, *Y* и *1/X*, *Y* и *lnX*, *lnY* и *X*, *lnY* и *lnX*. Каждый из них соответствует определенной двухпараметрической модели (табл. 2.1):

Таблица 2.1

Двухпараметрические модели парной регрессии

Переменные	Вид модели
$Y\|X$	$Y = \beta_0 + \beta_1 X + \varepsilon$
$Y\|X^2$	$Y = \beta_0 + \beta_1 X^2 + \varepsilon$
$Y\|1/X$	$Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon$
$Y\|\ln X$	$Y = \beta_0 + \beta_1 \ln X + \varepsilon$

Окончание табл. 2.1

Переменные	Вид модели
$\ln Y \mid X$	$Y = e^{\beta_0 + \beta_1 X} \cdot \varepsilon$
$\ln Y \mid \ln X$	$Y = \alpha X^{\beta_1} \cdot \varepsilon$

Обратите внимание, что $Y|X^2$ соответствует квадратичной, а не полной параболической зависимости $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$.

Воспользуемся следующей формулой коэффициента корреляции:

$$r_{XY} = \frac{m_{XY} - m_X m_Y}{s_X s_Y}, \quad s_X = \sqrt{m_{X^2} - m_X^2} = \sqrt{\frac{1}{n}\sum (X_k - m_X)^2},$$

где s_X, s_Y - смещенные оценки среднеквадратического отклонения.

m_X=СРЗНАЧ (диапазон X)
s_{0X} =СТАНДОТКЛОН.Г(диапазон X) или СТАНДОТКЛОНП (диапазон X) в более ранних версиях
r_{XY} =КОРРЕЛ (диапазон X; диапазон Y)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	k	Y	X	X^2	1/X	lnX	lnY		XY	X^2Y	Y/X	YlnX	XlnY	lnXlnY
77	76	31,7	12,93	167,30	0,077	2,560	3,456		410,02	5303,31	2,45	81,15	44,71	8,85
78	77	47,0	13,18	173,75	0,076	2,579	3,850		619,53	8166,39	3,57	121,20	50,75	9,93
79	78	57,2	20,51	420,62	0,049	3,021	4,047		1173,11	24059,35	2,79	172,79	82,99	12,22
80	79	31,0	11,78	138,79	0,085	2,466	3,434		365,21	4302,56	2,63	76,46	40,46	8,47
81														
82	∑	2647,4	1067,8	15132,7	6,2	203,6	270,2		37848,6	564399,6	193,1	6986,8	3732,4	703,5
83	m	33,5	13,5	191,6	0,1	2,6	3,4		479,1	7144,3	2,4	88,4	47,2	8,9
84	s	12,20	2,98	84,60	0,02	0,24	0,50		258,55	5597,47	0,69	37,20	14,72	1,79
85	r								0,720	0,703	-0,651	0,710	0,688	0,760

Рис. 2.5. **Расчет выборочных характеристик и коэффициентов корреляции**

	A	B	C	D	E	F	G	H	I	J
1	k	Y	X	X^2	1/X	lnX	lnY		XY	X^2Y
2	1	2,1	4,58588112689496	=C2^2	=1/C2	=LN(C2)	=LN(B2)		=B2*C2	=D2*B2
79	=A78+1	71,9	20,80426692749003	=C79^2	=1/C79	=LN(C79	=LN(B79		=B79*C79	=D79*B79
80	=A79+1	57,9	22,8544362819251	=C80^2	=1/C80	=LN(C80	=LN(B80		=B80*C80	=D80*B80
81										
82	∑	=СУМ	=СУММ(C$2:C$80)	=СУММ	=СУММ	=СУММ(	=СУММ(		=СУММ(I$2:I$80)	=СУММ(J$2:J$80)
83	m	=СРЗ	=C82/$A$80	=СРЗН	=СРЗН	=СРЗНАЧ	=СРЗНАЧ		=СРЗНАЧ(I$2:I$80)	=СРЗНАЧ(J$2:J$80)
84	s	=СТА	=КОРЕНЬ(D83-C83^2)	=СТАН	=СТАН	=СТАНД	=СТАНД		=СТАНДОТКЛОНП(I$2:I$80)	=СТАНДОТКЛОНП(J$2:J$80)
85	r								=(I$83-$B$83*C$83)/($B$84*C$84)	=(J$83-$B$83*D$83)/($B$84*D$84)
86									=КОРРЕЛ($B$2:$B$80;C$2:C$80)	=КОРРЕЛ($B$2:$B$80;D$2:D$80)

Рис. 2.6. **Расчет выборочных характеристик и коэффициентов корреляции в режиме отображения формул**

Таким образом, для всех предполагаемых зависимостей (рис. 2.5, 2.6) коэффициент корреляции по модулю близок к 0,7 (заметная линейная

связь). Наибольшее значение коэффициента корреляции у показательной зависимости (0,760), наименьшее - у обратной (гиперболической) (0,651), но отличия невелики.

Следовательно, все эти зависимости потенциально могут дать неплохую модель.

3. **Идентификация моделей**

Рассмотрим возможные варианты вида регрессии:

1) линейная $Y = \beta_0 + \beta_1 X + \varepsilon$;

2) параболическая $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$;

3) обратная (гиперболическая) $Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon$;

4) логарифмическая $Y = \beta_0 + \beta_1 \ln X + \varepsilon$;

5) показательная (экспоненциальная) $Y = \alpha\gamma^X \cdot \varepsilon = e^{\beta_0 + \beta_1 X} \cdot \varepsilon$;

6) степенная $Y = \alpha X^{\beta_1} \cdot \varepsilon$.

Выполним идентификацию предложенных моделей.

Параболическая, гиперболическая и логарифмическая модели линейны по параметрам и могут быть идентифицированы с помощью МНК.

Чтобы идентифицировать нелинейные по параметрам модели, выполним их линеаризацию путем логарифмирования:

5) $\ln Y = \beta_0 + \beta_1 X + \ln \varepsilon = \beta_0 + \beta_1 X + \xi$;

6) $\ln Y = \ln \alpha + \beta_1 \ln X + \ln \varepsilon = \beta_0 + \beta_1 \ln X + \xi$.

Рассчитаем оценки параметров моделей b_0, b_1, b_2 по формулам коэффициентов регрессии.

Для регрессий с двумя параметрами:

$$b_1 = r_{YX} \frac{s_Y}{s_X}, \quad b_0 = m_Y - b_1 m_X$$

Проверьте себя!
Линейная регрессия =ЛИНЕЙН (диапазон *Y*; диапазон *X*)
Показательная регрессия =ЛГРФПРИБЛ (диапазон *Y*; диапазон *X*)
Эти функции можно использовать в одной ячейке, тогда они возвращают значение b_1, либо как формулы массива, чтобы получить оба параметра. Для этого выделите формулу с ячейкой и соседнюю справа от нее, нажмите F_2, а затем Ctrl+Shift+Enter. В ячейке с формулой отобразится b_1, а в соседней - b_0.
Другие модели (кроме гиперболической) можно получить, добавив на график исходных данных линии тренда.

Через b_0, b_1 рассчитаем оценки a, g параметров a, γ для показательной и степенной регрессий:

$$a = e^{b_0}; \quad g = e^{b_1}.$$

Экспонента x вычисляется с помощью функции =EXP(x)

Для параболической регрессии с тремя параметрами (рис. 2.7):

$$B = A^{-1}C$$

$$B = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \quad A = \begin{pmatrix} 1 & m_X & m_{X^2} \\ m_X & m_{X^2} & m_{X^3} \\ m_{X^2} & m_{X^3} & m_{X^4} \end{pmatrix} \quad C = \begin{pmatrix} m_Y \\ m_{YX} \\ m_{YX^2} \end{pmatrix}$$

Или, по методу определителей (Крамера, рис. 2.8):

$$b_0 = \frac{\det A_0}{\det A}, b_1 = \frac{\det A_1}{\det A}, b_2 = \frac{\det A_2}{\det A},$$

где $\det A$ - определитель матрицы A; $\det A_0, \det A_1, \det A_2$, - частные определители (в матрице A соответствующий столбец заменен на C).

Обратную матрицу можно рассчитать с помощью функции МОБР (матрица). Результат этой функции является массивом (выделить область результата, начиная с ячейки с формулой, нажать F_2, нажать Ctrl+Shift+Enter).
Умножение векторов и матриц осуществляется функцией МУМНОЖ (вектор, матрица). Ее результатом также является массив.
Для расчета определителя матрицы используйте функцию МОПРЕД (матрица).

	A	H	I	J	K	L	M	N	O	P	Q
1	k		XY	X^2Y	Y/X	YlnX	XlnY	lnXlnY	X^3	X^4	YX^2
2	1		1495,83	31119,58	3,46	218,23	88,94	12,98	9004,45	187331,01	31119,58
86			0,720	0,703	-0,651	0,710	0,688	0,760			
87											
88	b0		-6,36516	14,10383	60,59409	-60,1132	1,861403	-0,68208			
89	b1		2,950246	0,101317	-345,459	36,32257	0,115301	1,591384			
90			2,950246	0,101317	-345,459	36,32257	0,115301	1,591384			
91											
92			A			C	det A				
93			1	13,5	191,6	33,5					
94			13,5	191,6	2836,3	479,1					
95			191,6	2836,3	43849,2	7144,3	2297,024				
96			A0			b0	det A0				
97			33,5	13,51635	191,5533						
98			479,1	191,5533	2836,329						
99			7144,3	2836,329	43849,23	-9,44522	-21695,9				
100			A1			b1	det A1				
101			1	33,5	191,5533						
102			13,51635	479,1	2836,329						
103			191,5533	7144,3	43849,23	3,41339	7840,64				
104			A2			b2	det A2				
105			1	13,51635	33,5						
106			13,51635	191,5533	479,1						
107			191,5533	2836,329	7144,3	-0,0166	-38,1326				
108			A^{-1}			B					
109			154,418	-21,494	0,716	-9,44522					
110			-21,494	3,116	-0,108	3,41339					
111			0,716	-0,108	0,004	-0,0166					

Рис. 2.7. **Расчет оценок параметров параболической модели матричным методом и методом Крамера**

	A	I	J	K	L	M
1	k	XY	X^2Y	Y/X	YlnX	XlnY
2	1	=B2*C2	=D2*B2	=E2*B2	=B2*F2	=C2*G2
88	b0	=$BS83-C83*I89	=$BS83-D83*J89	=$BS83-E83*K89	=$BS83-F83*L89	=$GS83-C83*M89
89	b1	=I85*$BS84/C84	=J85*$BS84/D84	=K85*$BS84/E84	=L85*$BS84/F84	=M85*$GS84/C84
90		=ЛИНЕЙН($BS2:$BS80;C2:C80;1)	=ЛИНЕЙН($BS2:	=ЛИНЕЙН($BS2	=ЛИНЕЙН($BS2:$BS80;F2:F80;1)	=ЛИНЕЙН($GS2:$GS8
91						
92		A			C	det A
93		1	=C83	=D83	=B83	
94		=C83	=D83	=O83	=I83	
95		=D83	=O83	=P83	=Q83	=МОПРЕД(I93:K95)
96		A0			b0	det A0
97		=L93	=J93	=K93		
98		=L94	=J94	=K94		
99		=L95	=J95	=K95	=M97/$M$93	=МОПРЕД(I97:K99)
100		A1			b1	det A1
101		=I93	=L93	=K93		
102		=I94	=L94	=K94		
103		=I95	=L95	=K95	=M101/$M$93	=МОПРЕД(I101:K103)
104		A2			b2	det A2
105		=I93	=J93	=L93		
106		=I94	=J94	=L94		
107		=I95	=J95	=L95	=M105/$M$93	=МОПРЕД(I105:K107)
108		A^{-1}			B	
109		=МОБР(I93:K95)	=МОБР(I93:K95)	=МОБР(I93:K95)	=МУМНОЖ(I109:K111;I93:I95)	
110		=МОБР(I93:K95)	=МОБР(I93:K95)	=МОБР(I93:K95)	=МУМНОЖ(I109:K111;I93:I95)	
111		=МОБР(I93:K95)	=МОБР(I93:K95)	=МОБР(I93:K95)	=МУМНОЖ(I109:K111;I93:I95)	

Рис. 2.8. **Расчет оценок параметров параболической модели матричным методом и методом Крамера в режиме отображения формул**

Рассчитаем модельные уровни регрессии и покажем их на графике (рис. 2.9):

Чтобы построить гладкие графики моделей, отсортируйте исходные данные по *X*. Выделите массивы для *X* и *Y* (с заголовками), на вкладке "Главная" → "Сортировка и фильтр" → "Настраиваемая сортировка". В окне выберите сортировку по столбцу *X*. Остальные данные должны рассчитаться автоматически.

В данном случае порядок появления пар *X* и *Y* не важен (все наблюдения являются равнозначными), хотя в некоторых случаях он имеет значение (наблюдения не равнозначны).

Таким образом, получены следующие модели регрессии:

1) $Y_k^{*1} = -6{,}37 + 2{,}95X_k$;

2) $Y_k^{*2} = -9{,}45 + 3{,}41X_k - 0{,}02X_k^2$;

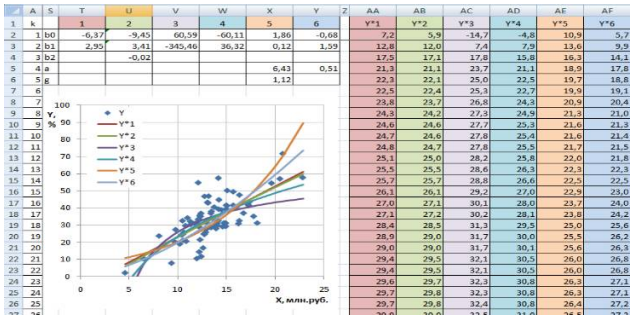


Рис. 2.9. **График корреляционного поля с наложенными шестью моделями регрессии**

3) $Y_k^{*3} = 60,6 - \dfrac{345,5}{X_k}$;

4) $Y_k^{*4} = -60,1 + 36,3 \ln X_k$;

5) $Y_k^{*5} = 6,43 \cdot 1,12^{X_k} = e^{1,86+0,12X_k}$;

6) $Y_k^{*6} = 0,51 X_k^{1,59}$.

4. Оценка общего качества моделей

Для оценки качества модели необходимо вычислить случайные остатки (рис. 2.10)

$$e_k = Y_k - Y_k^*$$

для каждой модели (в том числе и для мультипликативных).

Если вычисления выполнены правильно, то для моделей с аддитивной помехой $m_e = 0$. Для мультипликативных моделей $m_e > 0$ из-за асимметричности логнормального закона распределения.

	A	AA	AB	AC	AD	AE	AF	AG	AH	AI	AJ	AK	AL	AM
1	k	Y*1	Y*2	Y*3	Y*4	Y*5	Y*6		e1	e2	e3	e4	e5	e6
2	1	7,2	5,9	-14,7	-4,8	10,9	5,7		-5,1	-3,8	16,8	6,9	-8,8	-3,6
75	74	46,2	46,1	41,2	44,5	50,1	49,4		-11,0	-10,9	-6,0	-9,3	-14,9	-14,2
76	75	47,3	47,1	41,6	45,2	52,4	51,1		-16,0	-15,8	-10,3	-13,9	-21,1	-19,8
77	76	51,6	51,2	43,0	48,1	62,0	57,8		3,0	3,4	11,6	6,5	-7,4	-3,2
78	77	54,1	53,6	43,7	49,6	68,5	61,9		3,1	3,6	13,5	7,6	-11,3	-4,7
79	78	55,0	54,4	44,0	50,1	70,8	63,3		16,9	17,5	27,9	21,8	1,1	8,6
80	79	61,1	59,9	45,5	53,5	89,7	73,5		-3,2	-2,0	12,4	4,4	-31,8	-15,6
81														
82	∑	2647	2647	2647	2647	2568	2575		0,00	0,00	0,00	0,00	79,30	71,95
83	m	33,51	33,51	33,51	33,51	32,51	32,6		0,00	0,00	0,00	0,00	1,00	0,91
84	s	8,78	8,787	7,943	8,654	12,55	11,36		8,46	8,46	9,25	8,59	9,80	8,97

Рис. 2.10. **Вычисление ошибок моделей**

Вычислим коэффициент детерминации по формуле

$$R^2 = 1 - \frac{\sum_{k=1}^{n} e_k^2}{\sum_{k=1}^{n} (Y_k - m_Y)^2}.$$

Значение R^2, которое можно отобразить на диаграмме для линии тренда, вычисляется по другой формуле, как квадрат коэффициента корреляции r^2.

Для моделей, линейных по параметрам, эти величины совпадут, а для нелинейных $R^2 < r^2$.

Проверим его статистическую значимость с помощью F-критерия:

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

где m - число параметров в модели, в нашем случае 2 или 3.

Чем больше значений в выборке и чем выше R^2, тем больше значение F-критерия и тем надежнее модель.

F-критерий должен превышать критическое значение F-распределения Фишера $F_{kp} = F_{1-\alpha;m-1;n-m}$ с уровнем значимости α и степенями свободы $m-1$ и $n-m$:

$$F > F_{кр}.$$

Такую проверку можно осуществлять только при выполнении всех условий Гаусса - Маркова и для моделей, линейных по параметрам.

Также можно проверить статистическую значимость линеаризованных форм нелинейных моделей, используя вместо R^2 квадрат коэффициента корреляции, рассчитанного ранее, r^2.

Обычно α выбирают равным 0,1; 0,05; 0,01; 0;001, учитывая объемы выборок и уровень зашумленности модели. В данном случае объем выборки приближается к 100, поэтому, хотя шум достаточно велик, назначим $\alpha = 0,01$.

Имея в своем распоряжении компьютер, можно вычислить $\alpha_{\text{кр}}$ из условия $F = F_{кр}$, т.е. найти вероятность, с которой R^2 не является статистически значимым.

$F_{кр}$	=F.РАСП.ОБР(1-α ; $m-1$; $n-m$)
$\alpha_{кр}$	=F.РАСП (F; $m-1$; $n-m$)

Кроме того, вычислим MAPE-оценку, которая показывает среднее процентное отклонение реальных данных от модельных:

$$MAPE = \frac{1}{n}\sum_{k=1}^{n}\left|\frac{Y_k - Y_k^*}{Y_k}\right| \cdot 100\%.$$

Чем меньше MAPE-оценка, тем точнее модель. Допустимой будем считать MAPE <30 %, удовлетворительной MAPE <20 %, хорошей MAPE <10 %.

Модуль числа x вычисляется с помощью функции = ABS(x).

MAPE-оценка не имеет статистического смысла, но зачастую бывает более наглядной. Кроме того, стоит обратить внимание на MAPE в мультипликативных моделях, так как в них ошибка задается как доля от значений показателя.

	A	AO	AP	AQ	AR	AS	AT	AU	AV	AW	AX	AY	AZ	BA	BB
1	k	(Y-mY)^2	e1^2	e2^2	e3^2	e4^2	e5^2	e6^2		|e1/Y|	|e2/Y|	|e3/Y|	|e4/Y|	|e5/Y|	|e6/Y|
2	1	986,7	25,6	14,1	283,5	47,5	77,7	13,0		2,4	1,8	8,0	3,3	4,2	1,7
78	77	561,2	9,4	13,1	180,9	57,6	126,6	21,9		0,1	0,1	0,2	0,1	0,2	0,1
79	78	1473,7	285,2	306,9	779,0	473,9	1,2	73,8		0,2	0,2	0,4	0,3	0,0	0,1
80	79	594,8	10,0	4,0	154,3	19,0	1011,6	244,1		0,1	0,0	0,2	0,1	0,5	0,3
81															
82	∑	11748,8	5655,4	5649,8	6764,1	5831,8	7659,7	6426,1		21,8	21,2	29,1	23,1	24,2	21,3
83	m	148,7	71,6	71,5	85,6	73,8	97,0	81,3		0,276	0,268	0,368	0,293	0,306	0,269
84	s	245,91	113,88	113,84	138,42	116,60	169,38	125,43		0,409	0,373	0,947	0,483	0,518	0,318

	A	S	T	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG	AH
1	k		1	2	3	4	5	6		Y*1	Y*2	Y*3	Y*4	Y*5	Y*6		e1
2	1	b0	-6,37	-9,45	60,59	-60,11	1,86	-0,68		7,2	5,9	-14,7	-4,8	10,9	5,7		-5,1
3	2	b1	2,95	3,41	-345,46	36,32	0,12	1,59		12,8	12,0	7,4	7,9	13,6	9,9		-2,4
4	3	b2		-0,02						17,5	17,1	17,8	15,8	16,3	14,1		6,2
5	4	a					6,43	0,51		21,3	21,1	23,7	21,1	18,9	17,8		-13,5
6	5	g					1,12			22,3	22,1	25,0	22,5	19,7	18,8		-2,1
7	6	R^2	0,519	0,519	0,424	0,504	0,348	0,453		22,5	22,4	25,3	22,7	19,9	19,1		4,9
8	7	r^2	0,519		0,424	0,504	0,474	0,578		23,8	23,7	26,8	24,3	20,9	20,4		-4,8
9	8	m	2	3	2	2	2	2		24,3	24,2	27,3	24,9	21,3	21,0		0,0
10	9	F	26,2967	12,9855	22,9372	25,7904	24,7493	28,2054		24,6	24,6	27,7	25,3	21,6	21,3		8,0
11	10	Fкр.	6,97592	4,89584	6,97592	6,97592	6,97592	6,97592		24,7	24,6	27,8	25,4	21,6	21,4		0,0
12	11	αкр.	0,00000	0,00003	0,00002	0,00001	0,00001	0,00000		24,8	24,7	27,8	25,5	21,7	21,5		-3,6
13	12	MAPE	27,6%	26,8%	36,8%	29,3%	30,6%	26,9%		25,1	25,0	28,2	25,8	22,0	21,8		0,7

Рис. 2.11. **Вычисление характеристик качества моделей**

Проанализируем результаты, показанные на рис. 2.11. По примерному правилу R^2>0,5 и с использованием общей формулы для расчета R^2 получаем, что только линейная (1) и логарифмическая (4) модели обладают приемлемой точностью. Однако в п. 2 для всех линеаризованных моделей коэффициент корреляции был в районе 0,7 (для линейных по параметрам моделей R^2 равен квадрату R, а для нелинейных - нет).

В то же время F-критерий показывает, что в данном случае все модели можно признать статистически значимыми с вероятностью ошибки (первого рода) менее 0,001. Причиной тому служит большой объем исходной выборки.

MAPE-оценка для всех моделей находится в районе 25-35 %, т.е. довольно велика. Отметим, насколько различны соотношения MAPE и R^2: наименьшая MAPE у 2 и 6 моделей, а наибольший R^2 у 1 и 4. Какой из критериев считать более важным, зависит от целей исследования и характера исходных данных (что важнее - абсолютная ошибка или относительная, наблюдается ли гетероскедастичность помехи и т. д.).

В данном случае можно однозначно отметить низкое качество гиперболической модели (3) и экспоненциальной модели (5).

Для выбора наилучшей модели рассчитаем скорректированный коэффициент детерминации, информационный критерий Акаике и критерий Шварца:

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

$$AIC = n \ln USS + 2m,$$

$$SC = n \ln USS + m \ln n.$$

Если число наблюдений мало (n/m <40), критерий Шварца использовать не рекомендуется, а критерий Акаике необходимо скорректировать:

$$AIC_c = AIC + \frac{2m(m+1)}{n-m-1}.$$

В данном случае это соотношение нарушается для параболы $79/3 \approx 26$, но с небольшой натяжкой можно использовать основные критерии и для нее. Для наглядности в данном примере рассчитаем все критерии (рис. 2.12).

В этих формулах часто приходится ссылаться на n - число наблюдений. Чтобы не вводить его вручную, но и не искать нужную ячейку на листе, можно задать для нее имя. Выделите ячейку, содержащую n, и в окне "Имя" (рядом со строкой формулы) впишите "n".
Теперь вместо абсолютной ссылки на ячейку можно использовать назначенное имя (рис. 2.13).

	A	S	T	U	V	W	X	Y
1	k		1	2	3	4	5	6
2	1	b0	-6,37	-9,45	60,59	-60,11	1,86	-0,68
7	6	R^2	0,519	0,519	0,424	0,504	0,348	0,453
8	7	r^2	0,519		0,424	0,504	0,474	0,578
9	8	m	2	3	2	2	2	2
10	9	F	26,2967	12,9855	22,9372	25,7904	24,7493	28,2054
11	10	Fкр.	6,97592	4,89584	6,97592	6,97592	6,97592	6,97592
12	11	αкр.	0,00000	0,00003	0,00002	0,00001	0,00001	0,00000
13	12	MAPE	27,6%	26,8%	36,8%	29,3%	30,6%	26,9%
14	13	R^2_c	0,506	0,500	0,409	0,491	0,331	0,439
15	14	AIC	686,6	688,5	700,7	689,0	710,6	696,7
16	15	AIC_c	686,6	688,7	700,8	689,1	710,6	696,7
17	16	SC	691,3	695,6	705,5	693,8	715,3	701,4

Рис. 2.12. **Расчет критериев качества моделей**

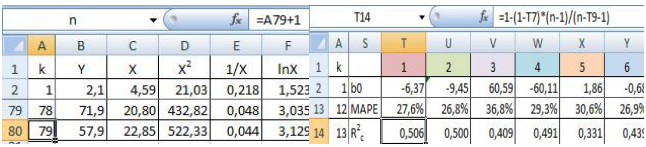


Рис. 2.13. **Именование ячейки**

Скорректированный R^2 всегда меньше обычного. Разница между скорректированным и обычным критерием Акаике в данном случае мала. Критерий Шварца, который всегда (для $n > 7$) превышает критерий Акаике и более чувствителен к включению дополнительных параметров в модель, в данном случае не дает ощутимо отличающихся результатов для параболической модели (2), которая содержит 3 параметра.

Формально все критерии указывают на наилучшее качество линейной модели (1), а наихудшим качеством обладают гиперболическая модель (3) и экспоненциальная модель (5). Формальность такого выбора опять же связана с тем, что две модели из шести являются нелинейными по параметрам и сумма квадратов отклонений USS для них будет завышенной. Из 4 линейных по параметрам моделей наилучшим качеством однозначно обладает линейная модель (1).

Таким образом, выбор осуществляется между линейной (1) и степенной (6) моделью. Несмотря на неудовлетворительные значения некоторых критериев для последней, соответствующая ей линеаризованная модель является наилучшей (r^2 =0,58). Чтобы сделать правильный выбор, обратимся к сути рассматриваемых параметров.

Регрессор Y представляет собой долю населения (домохозяйств), имеющих доступ к сети Интернет в различных субъектах РФ. Очевидно, что он изменяется в пределах от 0 до 100 %. Фактически на данный момент он варьируется от 2,1 % (Республика Ингушетия) до 71,9 % (Москва) и скорее всего в настоящее время 100 %-ный охват населения недостижим по техническим причинам, т. е. показатель ограничен сверху уровнем порядка 90-95 %.

Фактор X - общий объем расходов домохозяйств, для которого устранена разница в ценах между субъектами РФ, характеризует совокупный спрос домохозяйств. Данный показатель не может быть отрицательным, но установить его верхнюю границу, пожалуй, невозможно.

Очевидно, что при увеличении совокупного спроса все большее число домохозяйств получает доступ к сети Интернет. Но при приближении к 100 %-ному уровню темпы роста будут замедляться, поскольку остается все меньше людей, которые нуждаются в Интернете и имеют техническую возможность доступа.

При очень низком объеме расходов (ниже определенного уровня, например, прожиточного минимума) доля пользователей сети будет близка к 0 и практически не будет изменяться (небольшой процент, скорее всего, все же останется - те, кому Интернет необходим для работы, учебы). Таким образом, имеем картину: почти постоянный уровень Y при низких значениях X, рост Y при росте X, замедление роста при приближении Y к максимуму.

Такая динамика называется **логистической**. Один из вариантов (не единственный!) логистической кривой (логисты, *S*-образной кривой) в сравнении с моделями (1) и (6) показан на графике (рис. 2.14).

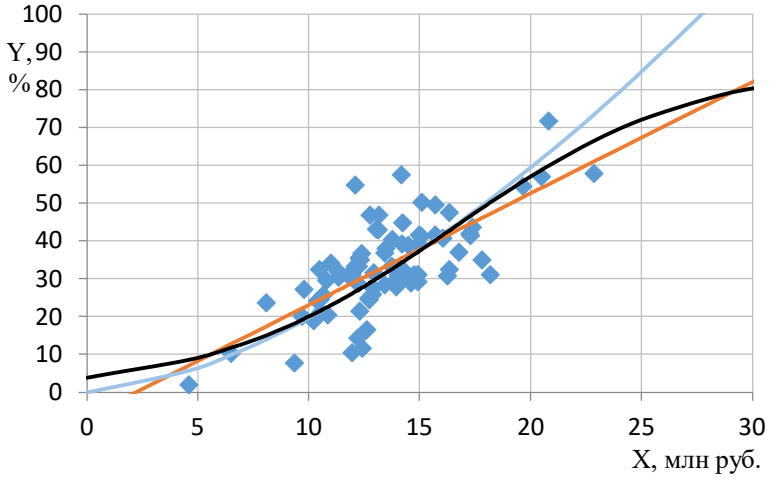


Рис. 2.14. **Логистическая модель**

Построение логистических кривых выходит за рамки данной лабораторной работы, поскольку они являются существенно нелинейными по параметрам, и будет впоследствии продемонстрировано в лабораторной работе № 6.

Однако можно заметить, что середина логисты близка к прямой линии, начало (нижняя половина) представляет собой выпуклую вниз кривую (ветвь параболы, экспонента, степенная функция), а верхняя половина - вогнутую кривую (другая ветвь параболы, гипербола, сумма константы и экспоненты с отрицательным показателем).

Следовательно, модель (6) в большей степени соответствует малым значениям X и Y, а модель (1) - близким к средним значениям. При больших значениях X (более 20 млн руб.) более грубой является степенная функция, поскольку не слишком быстро возрастает. Поэтому решение об адекватности модели зависит от того, какой диапазон X нас интересует, т.е. от цели моделирования.

В задании сказано, что требуется определить ожидаемый уровень доступа в Интернет по Самарской области при увеличении общего объема расходов населения на 2 %, т.е. для уровня расходов населения, равного 17,6 млн руб., что существенно превышает средний уровень (на 30 % m_X,

или 1,4 s_X), поэтому предпочтение стоит отдать линейной модели. Но если бы нас интересовала, к примеру, Республика Мордовия (X=10,2 млн руб.), следовало бы выбрать степенную зависимость.

Также можно отметить, что степенная функция проходит через начало координат, т.е. при нулевых расходах ожидаемый уровень доступа к сети Интернет равен нулю, но при любых малых расходах Y будет положительным. Прямая линия, напротив, пересекает ось X в точке 2,15 млн руб., т.е. при расходах менее этой суммы ожидаемый уровень доступа к сети Интернет будет отрицательным.

Также по линейной модели 100 %-ный уровень доступа будет достигнут при X=36,1млн руб., а по степенной - 27,7 млн руб.

5. Оценка качества идентификации параметров

Выполним оценку качества и построение доверительных интервалов оценок параметров линейной модели (1), выбранной в качестве наилучшей в п. 5.

Заметим, что в реальности эта проверка должна осуществляться для каждой построенной модели вместе с общей оценкой качества модели, но в лабораторной работе она была вынесена в отдельный пункт для сокращения объема расчетов. Для нелинейных моделей все расчеты проводятся для линеаризованной формы модели, а доверительный интервал исходных параметров можно рассчитать по связывающим эти параметры формулам.

В первую очередь необходимо рассчитать стандартные ошибки параметров модели:

$$S_{b1} = \sqrt{\frac{\sum e_k^2}{(n-2)\sum (X_k - m_X)^2}};\ \ S_{b0} = \sqrt{m_{X^2}}\, S_{b_1}.$$

Затем для каждого параметра рассчитывается t-статистика:

$$t_{b0} = \frac{b_0}{S_{b0}},\quad t_{b1} = \frac{b_1}{S_{b1}}.$$

Для нахождения критического уровня $t_{кр} = t_{1-\alpha/2,n-2}$ зададимся уровнем значимости $\alpha = 0{,}01$. Если

$$|t_{b0}| > t_{кр},\ |t_{b1}| > t_{кр},$$

то соответствующий параметр признается статистически значимым.

Как и для *F*-критерия, можно рассчитать $\alpha_{кр}$ для полученных *t*-статистик (вероятность принять параметр статистически значимым, тогда как он таковым не является).

$t_{кр}$ =СТЬЮДЕНТ.ОБР(1- α /2; $n-2$)

$\alpha_{крb0}$ =СТЬЮДЕНТ.РАСП (abs (t_{b0}); $n-2$)

Наконец, определим полуразмах доверительного интервала для каждого параметра:

$$\delta_{b0} = t_{кр} S_{b0},\ \delta_{b1} = t_{кр} S_{b1}.$$

Можно показать на графике, насколько изменится модель, если параметры будут равны границам своих интервалов (рис. 2.15).

	A	BD	BE	BF	BG	BH	BI	BJ	BK	BL	BM	BN	BO	BP	BQ	BR
1	k	Y* = b0 + b1 * X												(X-mX)^2	Y*1(-δ)	Y*1(+δ)
2	1			S	t	$\alpha_{кр.}$	δ		(b*- δ ; b*+ δ)					79,75	-8,60	22,93
3	2	b0	-6,37	4,48	-1,42	0,15968	11,84		(-18,21 ; 5,48)					49,26	-4,59	30,20
4	3	b1	2,95	0,3239	9,11	0,00000	0,8555		(2,09 ; 3,81)					29,53	-1,28	36,23
5	4	$t_{кр.}$			2,64									17,22	1,41	41,12
6	5													14,46	2,14	42,44
7	6													13,91	2,29	42,72
8	7													10,89	3,20	44,36
9	8													9,78	3,56	45,01
10	9													9,12	3,78	45,42
11	10													8,98	3,83	45,51
12	11													8,81	3,89	45,62
13	12													8,17	4,12	46,03
14	13													7,37	4,42	46,58
15	14													7,01	4,56	46,84
16	15													6,29	4,86	47,37
17	16													4,84	5,50	48,54
18	17													4,68	5,58	48,69
19	18													3,01	6,47	50,31
20	19													2,42	6,85	50,99
21	20													2,39	6,87	51,03
22	21													1,97	7,17	51,57

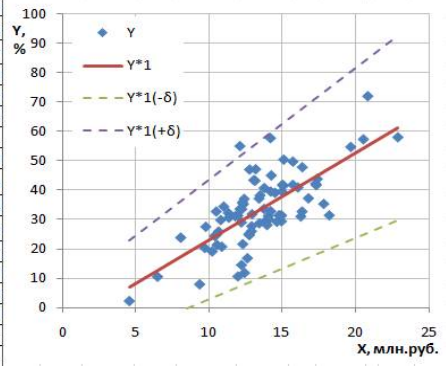


Рис. 2.15. **Изменение модели при равенстве параметров границам своих доверительных интервалов**

В данном случае с доверительной вероятностью $\alpha = 0{,}01$ параметр b_1 является статистически значимым, а параметр b_0 - нет. Вероятность ошибочно принять параметр b_0 статистически значимым составляет 0,160 - достаточно большое значение, а для b_1 эта вероятность равна нулю с точностью до пятого знака. Это означает, что наклон прямой определен достаточно надежно, а ее точка пересечения с осью *Y* - нет.

Как следствие, доверительный интервал b_0 накрывает 0. По графику видно, что наклон прямой в пределах доверительного интервала изменяется незначительно, а вот возможный сдвиг по вертикали - очень сильно. Это обстоятельство могло бы служить причиной для отклонения линейной модели.

6. **Прогнозирование по модели**

Рассчитаем значение фактора *x*, для которого необходимо выполнить прогнозирование.

$$x = 1,02 X_{Сам.обл} = 1,02 \cdot 17,30 \approx 17,64.$$

Точечная оценка прогноза рассчитывается простой подстановкой *x* в уравнение регрессии:

$$m_{Y|x} = Y^{*1}(x) = -6,37 + 2,95x \approx 45,69.$$

На данный момент уровень распространенности сети Интернет для Самарской области составляет 41,6, т.е. увеличение объема расходов населения на 2 % приведет к среднему увеличению распространенности сети Интернет на 9,8 %.

Для построения доверительного интервала прогноза рассчитаем (уровень значимости также равен 0,01):

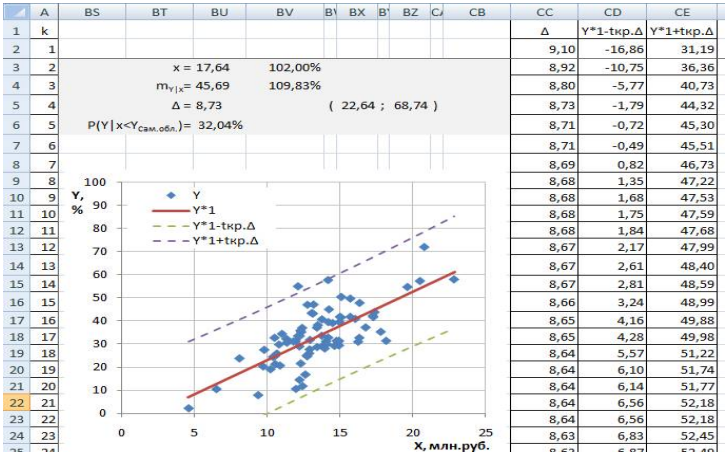
$$\left(Y^{*1}(x) - t_{кр}\Delta; Y^{*1}(x) + t_{кр}\Delta\right),$$

$$t_{кр} = t_{1-\frac{\alpha}{2}, n-2}, \ \Delta = \sqrt{\frac{\sum e^2}{n-2}\left(1 + \frac{1}{n} + \frac{\left(m_X - x\right)^2}{\sum\left(X_k - m_X\right)^2}\right)}.$$

Обратите внимание, что значение Δ зависит от *x*, для которого рассчитывается интервал, точнее от его удаленности от среднего значения. Поэтому, если построить границы доверительных интервалов для всей прямой, они будут расширяться по краям (чем дальше от среднего *X*, тем менее надежен прогноз).

По той же причине нельзя построить границы доверительного интервала, просто подставив в модель границы доверительных интервалов параметров, как это было показано на графике в п. 6.

Из графика можно видеть, что с вероятностью 99 % при увеличении объема расходов населения на 2 % уровень распространенности сети Интернет будет составлять от 22,6 до 68,7, т.е. существует довольно большой шанс, что он не изменится или уменьшится (рис. 2.16).



k	Δ	Y*1-tкр.Δ	Y*1+tкр.Δ
1	9,10	-16,86	31,19
2	8,92	-10,75	36,36
3	8,80	-5,77	40,73
4	8,73	-1,79	44,32
5	8,71	-0,72	45,30
6	8,71	-0,49	45,51
7	8,69	0,82	46,73
8	8,68	1,35	47,22
9	8,68	1,68	47,53
10	8,68	1,75	47,59
11	8,68	1,84	47,68
12	8,67	2,17	47,99
13	8,67	2,61	48,40
14	8,67	2,81	48,59
15	8,66	3,24	48,99
16	8,65	4,16	49,88
17	8,65	4,28	49,98
18	8,64	5,57	51,22
19	8,64	6,10	51,74
20	8,64	6,14	51,77
21	8,64	6,56	52,18
22	8,64	6,56	52,18
23	8,63	6,83	52,45

Рис. 2.16. **Построение доверительного интервала прогноза**

Вероятность этого события составляет 32,0 %:

$$P(Y \mid x < Y_{Сам.обл.}) = F_t\left(\frac{Y^{*1}(x) - Y_{Сам.обл.}}{\Delta}; n-2\right) = F_t\left(0{,}469; n-2\right) = 0{,}3204\,,$$

где $F_t(\ldots; n-2)$ - значение закона распределения Стьюдента (двухстороннего) с числом степеней свободы $n-2$.

Отчет по лабораторной работе должен содержать:

1. Общее и индивидуальное задание.
2. Основная часть. Результаты выполнения каждого пункта задания с краткими комментариями и указанием расчетных формул. При необходимости приводить основные промежуточные расчеты. Весь качественный анализ можно перенести в вывод.
3. Вывод. Привести и сопоставить основные численные результаты работы. Попытаться дать качественное обоснование полученной модели, пояснить смысл ее параметров и функционального вида.

Контрольные вопросы

1. Каковы основные виды моделей, для каких целей они строятся?
2. Перечислите этапы построения модели. Насколько строгим является их порядок?
3. В чем отличие теоретического и эмпирического уравнений регрессии?
4. В чем заключается суть МНК?

5. Какие оценки считаются "синими"?

6. Перечислите условия Гаусса - Маркова. Какие из них являются необязательными, а какие следует проверять?

7. Какие виды нелинейных моделей регрессии принято выделять? Приведите примеры.

8. На графике зависимости Y от X наблюдается наличие минимума. Какую модель регрессии можно рекомендовать?

9. Какие проблемы возникают при линеаризации нелинейных по параметрам моделей?

10. В чем отличие адекватности и точности модели? Что обычно подразумевают под проверкой качества уравнения регрессии?

11. Перечислите известные вам показатели общего качества уравнения регрессии.

12. Охарактеризуйте качество линейной модели с R^2=0,554, если исходная выборка содержит 20 значений? 10 значений? 100 значений?

13. Зачем нужны интервальные оценки параметров и прогнозных значений?

14. От чего зависит ширина доверительного интервала параметров?

Лабораторная работа 3
Множественная линейная регрессия

Цель работы: научиться осуществлять отбор факторов и оценивать значения параметров множественной линейной регрессии; выявлять мультиколлинеарность факторов; оценивать качество модели множественной регрессии.

Задание. Лабораторная работа выполняется в Microsoft Excel 2016. Варианты работы находятся в дополнительном файле "Варианты_ЛР3.xlsx".

Имеются выборки равного объема для показателей Y и X_1, X_2, X_3, X_4. Предполагается наличие линейной зависимости уровней Y от X_1, X_2, X_3, X_4. Необходимо выполнить:

1) корреляционный анализ, построив корреляционную матрицу, проанализировать ее и отобрать неколлинеарные факторы;

2) идентификацию модели со всеми включенными факторами и модели с выбранными факторами;

3) проверку качества полученной модели - расчет классического и скорректированного коэффициента детерминации;

4) расчет коэффициентов эластичности, рассчитав средние коэффициенты эластичности для каждого фактора, включенного в модель, пояснить их смысл.

Указания к выполнению

Модели множественной регрессии

Уравнение множественной регрессии описывает зависимость результативного признака y_k от нескольких факторов $x_{ik}, i = \overline{1,m}$.

Уравнение **множественной линейной регрессии** имеет вид

$$y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_m x_{mk} + \varepsilon_k .$$

Как и в случае парной регрессии, для построения модели необходимо решить задачи ее спецификации, идентификации и верификации.

В общем случае спецификация модели множественной регрессии включает:

- отбор факторов модели;
- выбор функционального вида модели.

Модель множественной регрессии может быть и нелинейной как по переменным, так и по параметрам, например:

– логарифмическая $y_k = \beta_0 + \beta_1 \ln x_{1k} + \beta_2 \ln x_{2k} + ... + \beta_m \ln x_{mk} + \varepsilon_k$;

– степенная $y_k = \alpha \cdot x_{1k}^{\beta_1} \cdot x_{2k}^{\beta_2} \cdot ... \cdot x_{mk}^{\beta_m} \cdot \varepsilon_k$;

– экспоненциальная $y_k = e^{\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_m x_{mk}} \cdot \varepsilon_k$;

– гиперболические $y_k = b_0 + b_1 \frac{1}{x_{1k}} + b_2 \frac{1}{x_{2k}} + ... + b_m \frac{1}{x_{mk}} + e_k$, или

$$y_k = \frac{1}{\beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_m x_{mk} + \varepsilon_k};$$

– смешанные $y_k = \beta_0 + \beta_1 x_{1k} + \beta_2 \frac{1}{x_{2k}} + \beta_3 x_{3k}^2 + ... + \beta_m \ln x_{mk} + \varepsilon_k$ и др.

В данной лабораторной работе рассматривается только линейная модель. Что касается нелинейных, то для них справедливо все то же, что и в случае парной регрессии - нелинейные по параметрам модели необходимо линеаризовать, учитывая вхождение стохастической компоненты.

Широко используется модель множественной регрессии в стандартизованном масштабе:

$$\hat{y}_k = \hat{\beta}_1 \hat{x}_{1k} + \hat{\beta}_2 \hat{x}_{2k} + ... + \hat{\beta}_m \hat{x}_{mk} + \hat{\varepsilon}_k,$$

$\hat{y}_k = \frac{y_k - \overline{y}}{S_y}$, $\hat{x}_{ik} = \frac{x_{ik} - \overline{x_i}}{S_{x_i}}$ - нормированные и центрированные величины:

$m_{\hat{y}} = 0, S_{\hat{y}} = 1, m_{\hat{x}_i} = 0, S_{\hat{x}_i} = 1.$

Уравнение регрессии в стандартизованном масштабе не содержит свободного члена, а переменные $\hat{y}_k$, $\hat{x}_{ik}$ выражаются в долях от своих СКО.

Параметры (или их оценки) стандартизованного уравнения регрессии связаны с параметрами в *естественном* масштабе следующими соотношениями:

$$\beta_i = \hat{\beta}_i \frac{S_y}{S_{x_i}} \quad \beta_0 = \overline{y} - \beta_1 \overline{x_1} - \beta_2 \overline{x_2} - ... - \beta_m \overline{x_m}.$$

Значения параметров $\hat{\beta}_i$ сравнимы между собой, а параметров β_i - нет. Но β_i имеют качественную интерпретацию: каждый коэффициент линейной регрессии показывает, на сколько изменится зависимая переменная при увеличении соответствующей независимой переменной на 1 (и неизменности других переменных).

Для качественного анализа влияния факторов на результативный признак часто используют коэффициенты эластичности. *Средний коэффициент эластичности* рассчитывается по формуле

$$Э_{yxi} = b_i \frac{\overline{x_i}}{\overline{y}} \cdot 100\% .$$

Средний коэффициент эластичности показывает, на сколько процентов от своего среднего значения изменится регрессор при увеличении фактора на 1 %.

Идентификация моделей множественной линейной регрессии (МНК)

Идентификация модели множественной линейной регрессии осуществляется с помощью МНК.

$$\sum_{k=1}^{n} \varepsilon_k^2 \xrightarrow[B]{} \min .$$

Идентификацию можно выполнить несколькими способами; в том числе через уравнение регрессии в естественном масштабе или в стандартизованном масштабе.

1-й способ. Решение СЛАУ, являющейся реализацией МНК:

$$\begin{cases} \sum y_k = nb_0 + b_1 \sum x_{1k} + b_2 \sum x_{2k} + \ldots + b_m \sum x_{mk} \\ \sum x_{1k} y_k = b_0 \sum x_{1k} + b_1 \sum x_{1k}^2 + b_2 \sum x_{1k} x_{2k} + \ldots + b_m \sum x_{1k} x_{mk} \\ \ldots \\ \sum x_{mk} y_k = b_0 \sum x_{mk} + b_1 \sum x_{1k} x_{mk} + b_2 \sum x_{2k} x_{mk} + \ldots + b_m \sum x_{mk}^2 \end{cases}$$

2-й способ. В матричном виде уравнение множественной регрессии имеет вид

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \quad B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_m \end{pmatrix} \quad X = \begin{pmatrix} 1 & x_{11} & \ldots & x_{m1} \\ 1 & x_{12} & & x_{m2} \\ \vdots & & \ddots & \\ 1 & x_{1n} & & x_{mn} \end{pmatrix} \quad E = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$Y^* = XB, \;\; E = Y - Y^* = Y - XB.$$

Тогда МНК записывается в виде

$$E^T E = \left(Y - XB\right)^T \left(Y - XB\right) \underset{B}{\rightarrow} \min .$$

Его решение:

$$-2X^T\left(Y-XB\right)=0$$

$$\left(X^TX\right)B=X^TY$$

$$B=\left(X^TX\right)^{-1}X^TY.$$

Замечание: под МНК нередко понимают именно эту формулу.

3-й способ. Уравнению регрессии в стандартизованном масштабе соответствует следующая система:

$$\begin{cases}\sum\hat{x}_{1k}\hat{y}_k=\hat{b}_1\sum\hat{x}_{1k}^2+\hat{b}_2\sum\hat{x}_{1k}\hat{x}_{2k}+\ldots+\hat{b}_m\sum\hat{x}_{1k}\hat{x}_{mk}\\ \sum\hat{x}_{2k}\hat{y}_k=\hat{b}_1\sum\hat{x}_{1k}\hat{x}_{2k}+\hat{b}_2\sum\hat{x}_{2k}^2+\ldots+\hat{b}_m\sum\hat{x}_{2k}\hat{x}_{mk}\\ \ldots\\ \sum\hat{x}_{mk}\hat{y}_k=\hat{b}_1\sum\hat{x}_{1k}\hat{x}_{mk}+\hat{b}_2\sum\hat{x}_{2k}\hat{x}_{mk}+\ldots+\hat{b}_m\sum\hat{x}_{mk}^2\end{cases}$$

Каждая сумма в данной системе преобразуется к соответствующему коэффициенту корреляции, например:

$$\sum\hat{x}_{1k}\hat{x}_{2k}=\frac{\sum\left(x_{1k}-\overline{x_1}\right)\left(x_{2k}-\overline{x_2}\right)}{S_{x_1}S_{x_2}}=\frac{n\cdot\mathrm{cov}(x_1;x_2)}{S_{x_1}S_{x_2}}=n\cdot r_{x_1x_2}.$$

Таким образом:

$$\begin{cases}r_{x_1y}=\hat{b}_1+r_{x_1x_2}\hat{b}_2+\ldots+r_{x_1x_m}\hat{b}_m\\ r_{x_2y}=r_{x_1x_2}\hat{b}_1+\hat{b}_2+\ldots+\hat{b}_mr_{x_2x_m}\\ \ldots\\ r_{x_my}=r_{x_1x_m}\hat{b}_1+r_{x_2x_m}\hat{b}_2+\ldots+\hat{b}_m\end{cases}$$

Например, для $m=2$: $\hat{b}_1=\dfrac{r_{yx_1}-r_{x_1x_2}r_{yx_2}}{1-r_{x_1x_2}^2}$ $\hat{b}_2=\dfrac{r_{yx_2}-r_{yx_1}r_{x_1x_2}}{1-r_{x_1x_2}^2}$.

При любом способе решения для множественной линейной регрессии справедлива **теорема Гаусса - Маркова**. Однако к пяти условиям Гаусса - Маркова, накладываемым на уравнение парной регрессии, добавляется еще одно, шестое: отсутствие *мультиколлинеарности,* т.е. линейной зависимости между объясняющими переменными:

$$r_{x_ix_j}=r_{ij}=0,\quad i\neq j.$$

При нарушении этого условия оценки параметров перестают быть эффективными, а решение системы уравнений в целом становится неустойчивым. Можно сказать, что наличие линейной зависимости между факторами не позволяет "разделить" их влияние на эндогенную переменную и правильно рассчитать стоящие при них коэффициенты.

При *совершенной (строгой)* мультиколлинеарности между факторами существует явная функциональная зависимость:

$$x_{ik} = a_0 + a_1 x_{jk}, \quad \left| r_{ij} \right| = 1.$$

На практике наиболее распространена *несовершенная* мультиколлинеарность, т.е. корреляционная зависимость между факторами:

$$x_{ik} = a_0 + a_1 x_{jk} + \xi_k, \quad r_{ij} \neq 0.$$

Для устранения мультиколлинеарности используются следующие приемы:

1. Исключение факторов из модели. Если между двумя факторами существует мультиколлинеарность, один из них, менее информативный, следует исключить.
2. Замена переменных - переход от исходных данных к их разностям, темпам роста и т. п.
3. Изменение формы модели - переход от линейной зависимости к нелинейной. Это возможно, только если для нелинейной модели сохраняется зависимость между факторами и регрессором.
4. Получение новой выборки. На практике это не всегда возможно, но на другой выборке показатели могут оказаться некоррелированными.

Мультиколлинеарность. Корреляционная матрица

Проверка наличия мультиколлинеарности осуществляется путем анализа матрицы *парных* коэффициентов корреляции:

$$R = \begin{pmatrix} 1 & r_{12} & \ldots & r_{1m} \\ r_{21} & 1 & \ldots & r_{2m} \\ \ldots & & \ddots & \\ r_{m1} & r_{m2} & \ldots & 1 \end{pmatrix}.$$

При отсутствии мультиколлинеарности корреляционная матрица должна иметь вид

$$R = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & & \ddots & \\ 0 & 0 & \ldots & 1 \end{pmatrix}, \quad \left| R \right| = 1.$$

При совершенной мультиколлинеарности:

$$R = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 1 & \dots & 1 \\ \dots & & \ddots & \\ 1 & 1 & \dots & 1 \end{pmatrix}, \quad |R| = 0.$$

Таким образом, чем ближе определитель матрицы R к 0, тем выше мультиколлинеарность. Если $|R|$ близок к 1, то мультиколлинеарность отсутствует.

Более строго проверку можно выполнить с помощью критерия x^2.

$H_0 : |R| = 1$ - отсутствие мультиколлинеарности;

$H_1 : |R| < 1$ - наличие мультиколлинеарности.

Рассчитывается критерий, имеющий распределение x^2 с $\frac{m}{2}(m-1)$ степенями свободы: $\chi^2 = -\left(n - 1 - \frac{2m+5}{6}\right) lg|R|$.

Если $x^2 > x^2_{кр} = x^2_{\alpha}; \frac{m}{2}(m-1)$, то гипотеза H_0 отклоняется и в модели присутствуют коррелирующие факторы.

Факторы, оказывающие наибольшее влияние друг на друга и наименьшее на результативный признак, необходимо исключить из модели. Можно по одному исключать "наихудшие" факторы до тех пор, пока мультиколлинеарность не исчезнет.

Недостаток парных коэффициентов корреляции заключается в том, что они не учитывают косвенное влияние факторов друг на друга. Необходимо рассчитывать *частные* коэффициенты корреляции, которые очищены от влияния других факторов.

Частный коэффициент корреляции, очищенный от влияния *одного* фактора x_k, рассчитывается по формуле

$$r_{ij.k} = \frac{r_{ij} - r_{ik} r_{jk}}{\sqrt{\left(1 - r^2_{ik}\right)\left(1 - r^2_{jk}\right)}}.$$

Частные коэффициенты корреляции, очищенные от влияния всех факторов, рассчитываются через обратную матрицу $C = R^{-1}$:

$$r_{ij} = -\frac{c_{ij}}{\sqrt{c_{ii} c_{jj}}}.$$

Таким образом, получают матрицу частных коэффициентов корреляции:

$$\overset{\circ}{R} = \begin{pmatrix} -1 & r_{12} & \dots & r_{1m} \\ r_{21} & -1 & \dots & r_{2m} \\ \dots & & \ddots & \\ r_{m1} & r_{m2} & \dots & -1 \end{pmatrix}.$$

Можно вычислить и частные коэффициенты корреляции между результативным признаком и факторами, например:

$$r_{yx_1x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}}.$$

Частные коэффициенты корреляции, очищенные ото всех факторов, также можно получить через обратную матрицу *C*, но тогда в корреляционную матрицу нужно добавить строку и столбец для *Y*:

$$R_y = \begin{pmatrix} 1 & r_{yx_1} & \dots & r_{yx_m} \\ r_{yx_1} & 1 & \dots & r_{1m} \\ \dots & & \ddots & \\ r_{yx_m} & r_{m1} & \dots & 1 \end{pmatrix} \quad C_y = R_y^{-1} = \begin{pmatrix} c_{yy} & c_{yx_1} & \dots & c_{yx_m} \\ c_{yx_1} & c_{x_1x_1} & \dots & c_{1m} \\ \dots & & \ddots & \\ c_{yx_m} & c_{m1} & \dots & c_{x_mx_m} \end{pmatrix} \quad r_{yx_i} = -\frac{c_{yx_i}}{\sqrt{c_{yy} \cdot c_{x_ix_i}}}.$$

Частные коэффициенты корреляции позволяют судить о взаимосвязи между двумя переменными при фиксированных значениях других переменных.

Проверка качества уравнения множественной регрессии. Отбор факторов

Качество уравнения регрессии может быть проверено с помощью тех же показателей, что и для парной регрессии - MAE, MAPE-оценки, коэффициента детерминации и др.

В данной работе используются следующие критерии.

Коэффициент множественной детерминации:

$$R^2 = 1 - \frac{\sum_{k=1}^{n} \left(y_k^* - y_k\right)^2}{\sum_{k=1}^{n} \left(y_k - \bar{y}\right)^2} = 1 - \frac{S_e^2}{S_y^2}.$$

Скорректированный коэффициент детерминации:

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

***F*-критерий Фишера**:

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

Если $F > F_{\alpha;m;n-m-1}$, то нулевую гипотезу H_0 следует отклонить и принять модель и R^2 статистически значимыми и надежными.

***t*-статистика**: $t_i = \frac{b_i}{S_{b_i}}$,

$$S_{b_i}^2 = S_e^2 Z_{ii}, \quad S_e^2 = \frac{1}{n-m-1} \sum e_k^2, \quad Z = \left(X^T X \right)^{-1}.$$

Если $|t_i| \ge t_{кр} = t_{\alpha/2;n-m-1}$, то соответствующий параметр можно считать статистически значимым и надежным.

Отбор факторов, включаемых в уравнение регрессии, можно выполнить двумя путями:

1) последовательным *включением* факторов в модель (**пошаговый отбор факторов**) - сначала в модель включается один наиболее значимый фактор, затем второй и т.д, пока добавление новых факторов в модель повышает ее качество (скорректированный R^2);

2) последовательным *исключением* факторов - сначала в модель включаются все факторы, затем наименее информативные по одному исключаются из модели, пока не начнет уменьшаться ее качество.

Какой путь выбрать - зависит от конкретной задачи. Если большинство из рассматриваемых факторов достаточно сильно коррелированы с регрессором, то проще идти методом исключения.

Если факторов с достаточно сильной корреляцией немного, то удобнее применить пошаговый отбор.

В целом, при выборе факторов, которые могут быть потенциально включены в модель, необходимо, чтобы они обладали двумя *свойствами*:

1) были количественно измеримы;
2) не были коррелированы между собой.

1. Корреляционный анализ

Рассмотрим следующий пример. Исследуется взаимосвязь показателей качества жизни населения по выборке для 25 регионов (табл. 3.1):

Y - средняя ожидаемая продолжительность жизни при рождении, лет;

X_1 - уровень рождаемости, чел. на 1000 чел. населения;

X_2 - доля населения с денежными доходами ниже величины прожиточного минимума, % от всего населения;

X_3 - среднедушевые доходы населения, у.е.;

X_4 - объем социальных выплат, млрд у.е.

Таблица 3.1

Пример исходных данных для модели множественной регрессии

k	Y	X_1	X_2	X_3	X_4
1	68,1	10,2	11,2	14,04	6,09
2	68,2	10,5	14,0	16,27	6,79
3	69,0	11,7	11,9	23,41	4,50
4	68,2	11,3	12,0	16,41	4,71
5	66,6	8,8	14,3	11,25	5,72
6	68,6	11,9	11,0	21,22	4,69
7	68,3	11,4	11,3	14,72	6,11
8	67,3	9,0	14,3	11,31	6,65
9	68,6	11,4	12,6	23,04	5,18
10	68,4	12,0	12,5	21,67	5,41
11	69,1	11,1	10,5	20,80	5,83
12	69,1	12,3	11,2	21,55	4,85
13	68,8	12,0	12,5	18,08	5,57
14	68,7	12,5	13,0	19,81	5,58
15	68,6	11,2	15,1	16,16	6,52
16	68,6	12,5	12,8	18,87	5,70
17	69,0	12,2	12,2	22,43	5,72
18	68,5	10,5	13,9	17,06	6,84
19	67,9	10,9	12,9	20,53	5,43
20	69,7	13,1	11,8	23,49	6,02
21	68,5	10,4	11,6	21,98	5,11
22	68,6	11,9	13,1	19,48	5,34
23	68,3	12,5	12,1	21,30	4,95
24	67,0	8,1	15,2	11,22	7,43
25	68,0	10,1	12,3	20,33	6,06

Необходимо исследовать корреляционные зависимости между переменными. Составим матрицу парных коэффициентов корреляции:

$$R = \begin{pmatrix} 1 & r_{12} & \dots & r_{1m} \\ r_{21} & 1 & \dots & r_{2m} \\ \dots & & \ddots & \\ r_{m1} & r_{m2} & \dots & 1 \end{pmatrix}.$$

Рассчитаем также парные коэффициенты корреляции между Y и X_i.

Корреляционную матрицу можно получить так (рис. 3.1, 3.2). Используем пакет "Анализ данных" → "Корреляция".

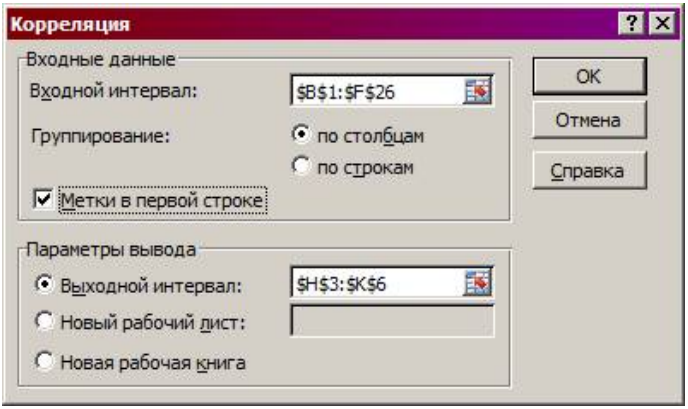


Рис. 3.1. **Окно команды "Корреляция"**

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	k	Y	X1	X2	X3	X4							
2	1	68,1	10,2	11,2	14,04	6,09							
3	2	68,2	10,5	14,0	16,27	6,79			*Y*	*X1*	*X2*	*X3*	*X4*
4	3	69,0	11,7	11,9	23,41	4,50		Y	1				
5	4	68,2	11,3	12,0	16,41	4,71		X1	0,8436	1			
6	5	66,6	8,8	14,3	11,25	5,72		X2	-0,563	-0,522	1		
7	6	68,6	11,9	11,0	21,22	4,69		X3	0,7874	0,7532	-0,579	1	
8	7	68,3	11,4	11,3	14,72	6,11		X4	-0,382	-0,57	0,635	-0,639	1

Рис. 3.2. **Результат расчета корреляционной матрицы**

Таблица 3.2

Корреляционная матрица

	Y	X_1	X_2	X_3	X_4
Y	1	0,8436	-0,563	0,7874	-0,382
X_1	0,8436	1	-0,522	0,7532	-0,57
X_2	-0,563	-0,522	1	-0,579	0,635
X_3	0,7874	0,7532	-0,579	1	-0,639
X_4	-0,382	-0,57	0,635	-0,639	1

Из табл. 3. видим, что между всеми факторами существует корреляционная связь различной силы (от 0,5 до 0,75). Проверим наличие мультиколлинеарности с помощью определителя $|R|$.

$$|R| = 0{,}1342.$$

Определитель матрицы достаточно близок к нулю.

Рассчитаем частные коэффициенты корреляции (табл. 3.3) с помощью обратной матрицы $C = R^{-1}$:

Таблица 3.3

Расчет частных коэффициентов корреляции

C					***R*⁰**				
	X_1	X_2	X_3	X_4	i\j	1	2	3	4
X_1	2,407	0,205	-1,522	0,269	1	-1,000	-0,097	**0,585**	-0,120
X_2	0,205	1,849	0,407	-0,797	2	-0,097	-1,000	-0,179	0,406
X_3	-1,522	0,407	2,810	0,669	3	**0,585**	-0,179	-1,000	-0,276
X_4	0,269	-0,797	0,669	2,087	4	-0,120	0,406	-0,276	-1,000

Cy				
6,614	-4,059	1,605	-2,934	-2,681
-4,059	4,898	-0,780	0,279	1,915
1,605	-0,780	2,239	-0,305	-1,448
-2,934	0,279	-0,305	4,111	1,858
-2,681	1,915	-1,448	1,858	3,174

i	***ryx*ᵢ**
1	0,713
2	-0,417
3	0,563
4	0,585

Наибольшее влияние на результативный признак Y оказывает фактор X_1, наименьшее - X_2.

Мультиколлинеарность в наибольшей степени вызвана корреляцией между X_1 и X_3. Следовательно, один из них необходимо исключить из модели. В данном случае лучше исключить X_3, поскольку он оказывает меньшее влияние на Y.

Высокое значение парного коэффициента корреляции между Y и X_3 обусловлено именно мультиколлинеарностью.

2. **Отбор факторов и идентификация модели**

Выполним идентификацию модели с включением переменных X_1, X_2 и X_4.

$$y_k^* = b_0 + b_1 x_{1k} + b_2 x_{2k} + b_4 x_{4k}.$$

Параметры вычислим через матричное представление исходных данных (табл. 3.4).

$$B = \left(X^T X \right)^{-1} X^T Y.$$

Таблица 3.4

Расчет частных коэффициентов корреляции

X^TX

25	**279,392**	**315,425**	**142,831**
279,392	3160,13	3505,13	1583,48
315,425	3505,13	4018,38	1816,48
142,831	1583,48	1816,48	829,291

$Z = (X^TX)^{-1}$

18,4455	**-0,77039**	**-0,48326**	**-0,64738**
-0,77039	0,04192	0,01115	0,02822
-0,48326	0,01115	0,04631	-0,03948
-0,64738	0,02822	-0,03948	0,14531

X^TY

1709,8
19125,1
21561,2
9763,96

$B = Z(X^TY)$

63,8283
0,45287
-0,16085
0,26815

X^T =ТРАНСП (матрица_X)
X^TX =МУМНОЖ (ТРАНСП (матрица_X); матрица_X)
$(X^TX)^{-1}$ =МОБР (матрица_X^TX)
Все формулы необходимо использовать как формулы массивов (выделить диапазон, F2, Ctrl+Shift+Enter).

Таким образом, модель имеет вид

$$y_k^* = 63{,}8 + 0{,}453x_{1k} - 0{,}161x_{2k} + 0{,}268x_{4k}.$$

Коэффициент детерминации:

$$R^2 = 1 - \frac{S_e^2}{S_y^2} = 0{,}779.$$

$$\overline{R}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-4} = 0{,}747.$$

Окончательно модель имеет вид

$$y_k^* = 63,8 + 0,453x_{1k} - 0,161x_{2k} + 0,268x_{4k}, \quad R^2 = 0,779.$$

Полученная модель описывает 77 % исходных данных, 23 % приходятся на случайные отклонения.

Самостоятельно реализуйте идентификацию модели со всеми включенными факторами.

3. **Проверка качества полученной модели**

Проверим общее качество модели с помощью F-теста:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-4}{3} = 24,6 > F_{кр} = F_{0,01;3;21} = 4,8 .$$

$F > F_{kp}$, следовательно, R^2 и полученная модель статистически значимы и надежны.

Таким образом, получена зависимость средней ожидаемой продолжительности жизни при рождении от уровня рождаемости, доли малоимущего населения (с доходом менее прожиточного минимума) и объема социальных выплат:

$$y_k^* = 63,8 + 0,453x_{1k} - 0,161x_{2k} + 0,268x_{4k}, \quad R^2 = 0,779.$$

Фактор среднедушевого дохода был исключен из модели из-за сильной коррелированности с другими факторами.

4. **Расчет коэффициентов эластичности**

Вычислим средние коэффициенты эластичности для каждого фактора, включенного в модель, по формуле

$$Э_{yx_i} = b_i \frac{\overset{m}{x_i}}{m_y} \cdot 100\%$$

$$Э_{yx1} = \ 7,47\ \%$$

$$Э_{yx2} = -3,21\ \%$$

$$Э_{yx4} = \ 2,32\ \%$$

При увеличении рождаемости X_1 на 1 % средняя ожидаемая продолжительность жизни при рождении Y увеличится на 7,47 %. При увеличении доли населения с доходом ниже прожиточного минимума X_2 на 1 % Y снизится на 3,21 %. При увеличении объема социальных выплат X_4 на 1 % Y увеличится на 2,32 %.

Контрольные вопросы

1. Каков смысл коэффициентов множественной регрессии в естественном масштабе?

2. В каких величинах выражаются переменные в модели множественной регрессии в стандартизованном масштабе?

3. По результатам идентификации получены значения $\hat{\beta}_1 = -1{,}78$, $\hat{\beta}_2 = 0{,}31$. Какой из факторов оказывает большее влияние на результативный признак?

4. Что показывают средние коэффициенты эластичности?

5. Запишите решение МНК в матричном виде.

6. Запишите систему для решения МНК через коэффициенты корреляции.

7. Какое условие Гаусса - Маркова добавляется к условиям для парной линейной регрессии? Каковы последствия его нарушения?

8. Что такое корреляционная матрица?

9. В каком диапазоне изменяется определитель корреляционной матрицы $|R|$?

10. Что можно сказать о модели множественной регрессии, для которой $|R| = 0{,}05$? $|R| = 0{,}78$?

11. Чем частные коэффициенты корреляции отличаются от парных?

12. Как рассчитывается скорректированный коэффициент детерминации? Как он соотносится с обычным R^2?

13. Какими основными свойствами должны обладать факторы модели множественной регрессии?

Лабораторная работа 4
Парная нелинейная регрессия в *R*

Цель работы: научиться применять генетические алгоритмы (ГА) для поиска оценок параметров нелинейных моделей

Задание. Лабораторная работа выполняется на языке программирования *R* в среде *RStudio*. Варианты лабораторной работы находятся в приложении.

1. Выполнить генерацию значений *x* и *y*.
2. Построить график зависимости *y* от *x*.
3. Задать функцию зависимости *y* от *x* в *R* и соответствующую ей функцию МНК.
4. Определить минимальные и максимальные значения параметров.
5. Запустить поиск решения с помощью ГА. Размер популяции, предельное число итераций и условие останова алгоритма заданы в варианте.
6. Построить график значений fitness-функции и график с исходными данными и полученной моделью.
7. Повторить пункты 5 и 6 еще 2 раза и сравнить полученные результаты.
8. Выбрать наилучший из трех результатов по значению коэффициента детерминации.

Указания к выполнению работы

1. Генетические алгоритмы

Для идентификации сложных нелинейных по параметрам моделей можно применять генетический алгоритм, который решает задачу моделирования путем случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую эволюцию.

Постановка задачи выглядит следующим образом: имеется целевая функция от многих переменных, у которой необходимо найти глобальный экстремум: $f(x_1, x_2, ..., x_n)$.

Например, это функция МНК для логистической модели Гомпертца:

$$Q(C, A_0, \alpha, K_0) = \sum_{k=0}^{n} \left(Y_k - C - A_0 e^{-e^{-\alpha \triangle (k - k_0)}} \right)^2,$$

которую необходимо минимизировать.

Независимые переменные следует представить в виде хромосом. Преобразование независимых переменных в хромосомы осуществляют с помощью кодирования, которое задается в двоичном формате. Используются N бит для каждого параметра, причем N может быть различным для каждого параметра.

Хромосома представляется вектором длины $L = N_m$, где m - число независимых переменных (параметров) функции. Каждая особь состоит из массива $X(0...L\text{-}1)$ и значения функции f на переменных, извлеченных из этого массива.

В общем случае генетический алгоритм состоит из следующих шагов:

1. Генерация начальной популяции - заполнение популяции особями, в которых независимые переменные (закодированные битами) заполнены случайным образом.

2. Выбор родительской пары: берутся K особей с минимальными значениями целевой функции, и из них составляются все возможные пары $K(K-1)/2$.

3. Кроссинговер: случайным образом выбирается точка t на массиве $X(0...L\text{-}1)$.

4. Все элементы массива с индексами 0-t новой особи (потомка) заполняются элементами с теми же индексами, но из массива X первой родительской особи. Остальные элементы заполняются из массива второй родительской особи. Для второго потомка действия производятся наоборот: элементы 0-t берутся от второй родительской особи, а остальные - от первой.

5. Новые особи с некоторой вероятностью мутируют: инвертируется случайный бит массива X этой особи. Вероятность мутации обычно полагают порядка 1 %.

6. Полученные особи-потомки добавляются в популяцию после переоценки. Обычно новую особь добавляют взамен самой неудачной старой особи при условии, что значение функции на новой особи меньше значения функции на старой особи.

7. Если самое лучшее решение в популяции не удовлетворяет, то осуществляется переход на шаг 2.

Критерием окончания процесса могут служить заданное количество поколений или схождение популяции.

Более сложный генетический алгоритм содержит такие шаги, как отбор особей для размножения и генерация пар из отобранных особей. При этом каждая особь может быть задействована в одной и более паре, в зависимости от используемого алгоритма.

Рассмотрим решение задачи нелинейного МНК с помощью генетического алгоритма на встроенном наборе данных *trees* в среде *R*.

Для работы подключим следующие пакеты:

```
library("GA") # генетические алгоритмы
library("dplyr") # работа с наборами данных
library("ggplot2") # графики
library("spuRs")# содержит набор данных trees
```

Получим описание набора данных по деревьям `trees` из пакета `spuRs` командой (рис. 4.1):

```
help("trees", package = "spuRs")
```

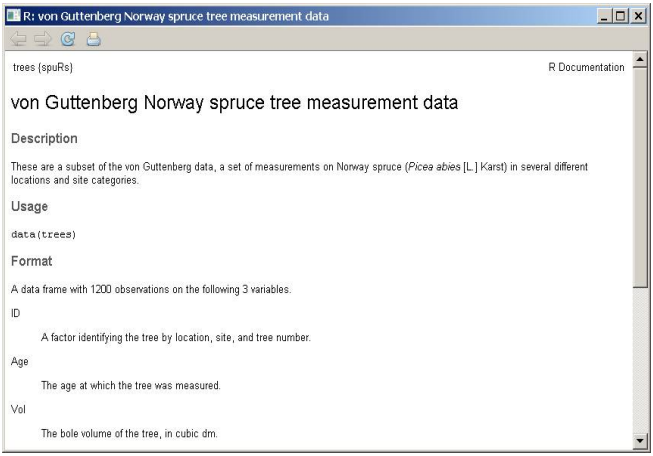


Рис. 4.1. **Справка по набору данных trees из пакета spuRs**

В этом наборе данных содержится 1200 наблюдений и три переменных (`ID` дерева, который включает номер местоположения, площадки и номер дерева; возраст дерева; объем в куб. дм). Будем рассматривать зависимость объема дерева (переменная `Vol`) от возраста (переменная `Age`).

Активируем набор данных командой:

```
data("trees", package = "spuRs")
```

Посмотрим на этот набор данных:

```
> glimpse(trees)
Observations: 1200
Variables:
$ ID  (fctr)  1.1.1,  1.1.1,  1.1.1,  1.1.1,  1.1.1,  1.1.1,  1.1.1,
1.1.1,  1.1.1,  1.1.1,  1....
$ Age (dbl) 9.67, 19.67, 29.67, 39.67, 49.67, 59.67, 69.67, 79.67,
89.67, 99.67, 109....
```

```
$ Vol (dbl) 5.0, 38.0, 123.0, 263.0, 400.0, 555.0, 688.0, 820.0,
928.0, 1023.0, 1104....
```

Выберем для дальнейшей работы только данные для деревьев с определенным местоположением, например, *ID* = 1.3.11 (подробнее см.: Zeide B. Analysis of growth equations // Forest Science. 1993. № 39 (3). С. 549-616). Поместим в переменную `tree` выбранную часть исходной выборки:

```
tree<- trees[trees$ID == "1.3.11", 2:3]
```

Командой `trees[]` можно выбрать нужные столбцы из набора данных. Таким образом, предыдущая команда выбирает из набора данных только второй и третий столбцы, но так, чтобы в первом столбце ID был при этом равен 1.3.11.

Посмотрим теперь на `tree`:

```
> glimpse(tree)
Observations: 12
Variables:
$ Age (dbl) 2.44, 12.44, 22.44, 32.44, 42.44, 52.44, 62.44, 72.44,
82.44, 92.44, 102....
$ Vol (dbl) 2.2, 20.0, 93.0, 262.0, 476.0, 705.0, 967.0, 1203.0,
1409.0, 1659.0, 1898...
```

Получен набор данных из 12 наблюдений с двумя переменными.

Попробуем описать зависимость объема дерева (`y`) от его возраста (`x`) с помощью логистической функции Ричардса:

$$y = a\left(1 - e^{-bx}\right)^c.$$

Затем нужно будет применить метод наименьших квадратов, причем параметры входят в модель нелинейно:

$$a,b,c = \arg\min_{a,b,c} \sum_{i=1}^{12} \left(Vol_i - a\left(1 - e^{-b \cdot Age_i}\right)^c \right)^2.$$

Минимизировать сумму квадратов ошибок будем численно с помощью генетического алгоритма. ГА реализован в *R* функцией ga из пакета GA.

Функция `ga` максимизирует функцию `fitness`, которую следует задать. Поскольку стоит задача минимизации, воспользуемся следующим фактом:

$$\arg\max_{\Theta}(\Theta) = \arg\min_{\Theta}(-\Theta).$$

Для этого необходимо задать в *R* функцию Ричардса - она будет использована для функции `fitness`, указанной в аргументах функции `ga`. Пусть вектор параметров функции обозначен `theta`. Тогда `a = theta[1], b = theta[2], c = theta[3]`. Задаем функцию `richards`:

```
richards<- function(x, theta)
theta[1] * (1 - exp(-theta[2] * x))^theta[3]
```

В первой строке указан аргумент функции *x* и вектор параметров `theta`. Через пробел в следующей строке описывается сама функция Ричардса, порядок параметров `a, b, c` в векторе `theta` должен сохраняться.

Задаем функцию `fit` суммы квадратов ошибок и после пробела ставим перед функцией суммы квадратов ошибок знак минус:

```
fit<- function(theta, x, y) -sum((y - richards(x, theta))^2).
```

В функции `ga` задаются следующие параметры:

`fit` - нужная нам функция для аргумента `fitness`, которую максимизирует генетический алгоритм;

`type` выберем `real-valued`, поскольку возраст деревьев и объем являются действительными числами;

`x` и `y` - это, соответственно, возраст дерева (`tree$Age`) и объем дерева (`tree$Vol`) - аргументы функции `fitness`;

`min` - это вектор минимальных значений параметров `a, b, c` функции Ричардса;

`max` - вектор максимальных значений параметров `a, b, c`;

`crossover` - функция в *R*, выполняющая кроссовер, т.е. функция, которая образует потомков, объединив часть генетической информации от родителей;

`popsize` - размер популяции;

`maxiter` - максимальное число итераций, после которого работа генетического алгоритма прекращается;

`run` - число последовательных поколений без какого-либо улучшения в значении `fitness`-функции перед остановом алгоритма и т.д.

Запишем результат выполнения функции `ga` в переменную `myGA`, задав значения описанным аргументам:

```
    myGA<- ga(type = "real-valued", fitness = fit,x = tree$Age,
    y = tree$Vol, min = c(3000, 0, 2), max = c(4000, 1, 4),
    popSize = 500, crossover = gareal_blxCrossover, maxiter = 5000,
run = 200, names = c("a", "b", "c"))
```

Функция при работе будет выводить в консоль значения fit на каждой итерации алгоритма:

```
Iter = 1 | Mean = -74580573 | Best = -6766.369
Iter = 2 | Mean = -64840448 | Best = -6766.369
Iter = 3 | Mean = -56793683 | Best = -6766.369
Iter = 4 | Mean = -48988112 | Best = -6766.369
Iter = 5 | Mean = -38083470 | Best = -6766.369
Iter = 6 | Mean = -31671411 | Best = -4661.229
Iter = 7 | Mean = -21836845 | Best = -4661.229
...
```

Посмотрим на краткие результаты подбора функции Ричардса на исходные данные:

```
>summary(myGA)
+----------------------------------+
| Genetic Algorithm |
+----------------------------------+
GA settings:
Type = real-valued
Population size = 500
Number of generations = 5000
Elitism = 25
Crossover probability = 0.8
Mutation probability = 0.1
Search domain
a b c
Min 3000 0 2
Max 4000 1 4
GA results:
Iterations = 792
Fitness function value = -2774.113
Solution =
 a b c
[1,] 3589.788 0.01546055 2.786145
```

Таким образом, получено следующее уравнение регрессии:

$$\hat{y} = 3589.788\left(1 - e^{-0.015x}\right)^{2.786}.$$

Командой `plot(myGA)` можно посмотреть, как изменялись значения `fitness`-функции на протяжении всех 792 итераций алгоритма (рис. 4.2).

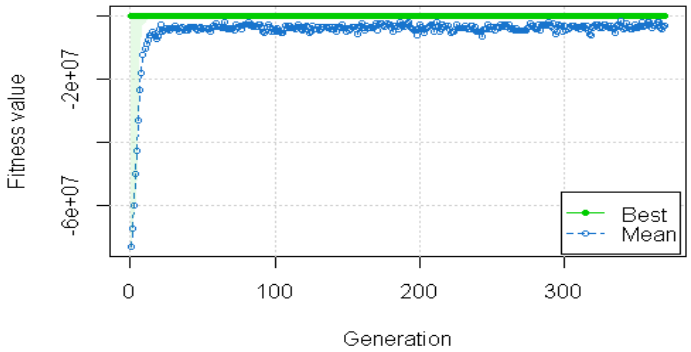


Рис. 4.2. **Изменение значений fitness-функции в зависимости от числа итераций**

Модельные значения y можно получить командой:

```
richards(tree$Age, myGA@solution)
```

Построим на графике исходную выборку и результат подбора функции Ричардса (рис. 4.):

```
ggplot() +
 geom_point(aes(x=tree$Age, y=tree$Vol)) +
 geom_line(aes(x=tree$Age, y=richards(tree$Age, myGA@solution)))
```

В данной команде указано:

`ggplot()` активирует построение графика с помощью функций пакета ggplot2;

`geom_point(aes(x=tree$Age, y=tree$Vol))` - строит точками зависимость объема дерева от возраста по исходной выборке;

`geom_line(aes(x=tree$Age, y=richards(tree$Age, myGA@solution)))` - строит линией модельные значения объема деревьев в зависимости от возраста, рассчитанные по функции Ричардса с параметрами, найденными генетическим алгоритмом.

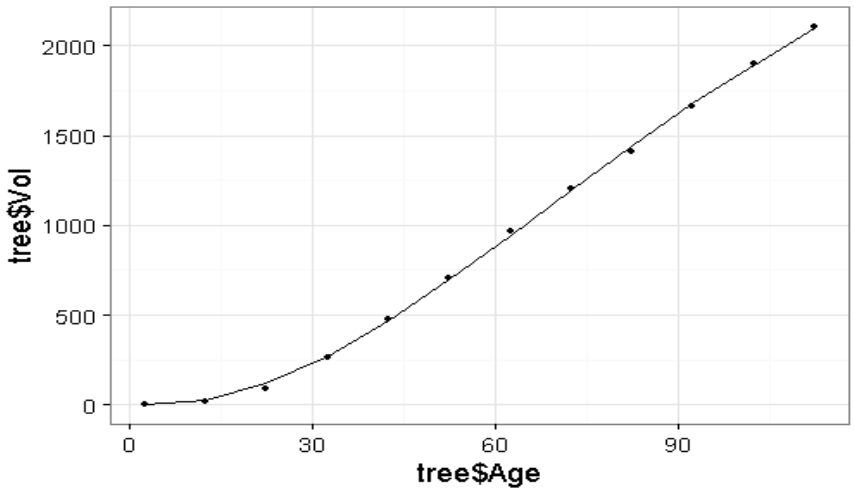


Рис. 4.3. **График исходных данных и полученной функции Ричардса**

Для оценки точности модели рассчитаем ее коэффициент детерминации:

```
RSS <- sum((tree$Vol-richards(tree$Age, myGA@solution))^2)
TSS <- sum((tree$Vol-mean(tree$Vol))^2)
R2 <- 1-RSS/TSS
> R2
[1] 0.9995553
```

Построим по полученной модели прогноз объема дерева для возраста 150 и 200 лет:

```
Age_forecast <- c(150,200)
richards(Age_forecast, myGA@solution)
```

```
[1] 2691.205 3156.228
```

Для построения графика прогноза в виде гладкой линии потребуется большое число точек, которые удобнее задать в виде последовательности от 115 до 200 с шагом 5:

```
Age_forecast <- seq(115,200, by = 5)

ggplot() +
geom_point(aes(x=tree$Age, y=tree$Vol)) +
geom_line(aes(x=tree$Age, y=richards(tree$Age, myGA@solution))) +
geom_line(aes(x=Age_forecast, y=richards(Age_forecast, myGA@solu
tion), color = "maroon")) +
theme_bw(base_size = 18)
```

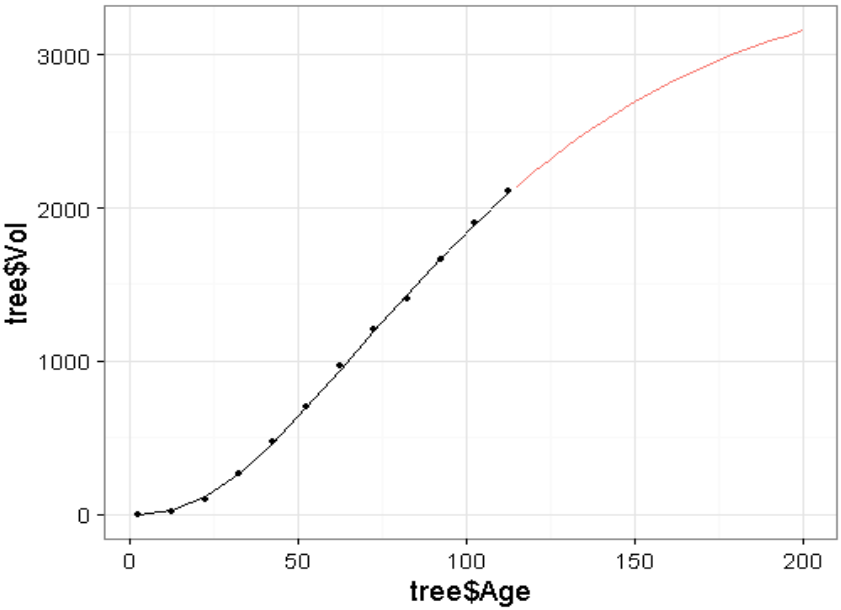


Рис. 4.4. **График исходных данных и полученной логистической функции Ричардса с прогнозом до 200 лет**

Результат показан на рис. 4.4.

2. Генерация данных

Выполним генерацию значений *x* с нормальным законом распределения, математическим ожиданием 45 и среднеквадратическим отклонением (СКО) 7. Объем выборки зададим равным 100:

```
x <- rnorm(n = 100, mean = 45, sd = 7)
```

Для генерации *y* необходимо задать формулу зависимости, например, линейную, и наложить на нее случайную помеху *ε*:

```
a <- 20
b <- 5
y <- a + b*x
```

Генерацию *ε* также выполним по нормальному закону распределения, с нулевым математическим ожиданием и единичной дисперсией.

Однако, поскольку генераторы случайных чисел неидеальны, необходимо дополнительно выполнить нормировку и центрирование значений:

```
er <- rnorm(n = 100, mean = 0, sd = 1)
er <- (er- mean(er))/sd(er)
```

После этого можно наложить помеху на формулу зависимости с заданным СКО, равным 2:

```
Ser <- 5
y <- y + er*Ser
```

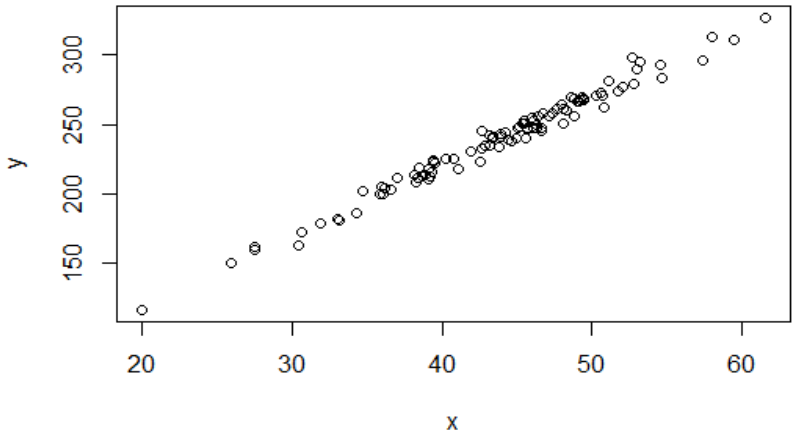


Рис. 4.5. **Сгенерированная зависимость *y* от *x***

В результате получим зависимость, показанную на рис. 4.5.

Контрольные вопросы

1. Максимизирует или минимизирует генетический алгоритм в *R* fitness-функцию?
2. Как задать в *R* функцию $y = a(x - b)^3 + c$?
3. Каким образом в *R* можно сгенерировать выборку со стандартным нормальным законом распределения?

Лабораторная работа 5
Временные ряды в *R*

Цель работы: изучить базовые способы анализа, моделирования и прогнозирования временных рядов с помощью *ARIMA*-моделей на примере биржевой статистики акций различных компаний.

Задание. Лабораторная работа выполняется на языке программирования *R* в среде *RStudio*. Варианты лабораторной работы находятся в приложении.

1. Загрузить данные из временного ряда для своего варианта за последние полгода.
2. Построить график временного ряда с графиками автокорреляционной и частной автокорреляционной функций.
3. Подобрать вручную порядок *ARIMA*-модели. Для подбора использовать визуальный анализ графика.
4. Подобрать *ARIMA*-модель автоматически.
5. Построить прогноз на указанное в варианте число шагов.Показать прогноз на графике.

Указания к выполнению работы

1. Анализ временных рядов

ARIMA-модель временного ряда расшифровывается как интегрированная модель авторегрессии - скользящего среднего (англ. *auto-regressive integrated moving average*). *ARIMA*-модель определяется тремя параметрами порядка: (*p, d, q*). Процесс подгонки модели *ARIMA* иногда называют методологией Бокса - Дженкинса (Box -Jenkins).

Авторегрессионная компонента $AR(p)$ использует прошлые значения временного ряда *Y* в уравнении регрессии. Параметр *p* определяет число лагов, используемых в модели. Например, $AR(2)$, или, что то же самое, $ARIMA(2,0,0)$, представляется следующей формулой:

$$Y_t = C + \varphi_1 Y_{t-1} + \varphi_2 Y_{t-2} + e_t,$$

где C, φ_1, φ_2 - параметры модели.

Параметр *d* представляет собой порядок дифференцирования в интегрированной компоненте $I(d)$ *ARIMA*-модели. Дифференцирование временного ряда здесь означает лишь взятие разностей между текущим и

предыдущим значениями d раз. Обычно дифференцирование используется для стабилизации временного ряда, при нарушении предпосылки его стационарности (определение будет дано ниже).

Компонента скользящего среднего $MA(q)$ представляет собой невязку модели, представленную линейной комбинацией предыдущих значений невязок. Порядок компоненты q определяет число этих предыдущих значений невязки, включаемое в модель:

$$Y_t = C + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \ldots + \theta_2 e_{t-q} + e_t.$$

Компонента скользящего среднего, а также авторегрессионная и интегрированная компоненты образуют *ARIMA*-модель **без учета сезонности**, которая записывается следующим линейным уравнением:

$$\Delta^d Y_t = c + \sum_{i=1}^{p} \varphi_i \Delta^d Y_{t-i} + \sum_{j=1}^{q} \theta_j e_{t-j} + e_t,$$

где Δ^d - оператор взятия разности временного ряда порядка d.

Заметим, что эта модель предполагает отсутствие в ряду сезонной компоненты, что означает, что перед построением *ARIMA*-модели необходимо бывает провести процесс десезонализации (это будет показано в примере ниже).

ARIMA-модель также может предполагать структуру временного ряда с учетом сезонности. В таком случае модель включает два набора параметров: (p, d, q), как в предыдущем случае, и $(P, D, Q)_m$, описывающие сезонную компоненту из m периодов.

Методология *ARIMA* имеет свои ограничения. Эти модели напрямую полагаются на прошлые значения и поэтому лучше всего работают с длинными и стабильными рядами. Также обратите внимание, что *ARIMA* просто аппроксимирует исторические закономерности и поэтому не стремится объяснить внутренние механизмы формирования структуры данных.

Загрузка пакетов и данных

Для работы в *R* с временными рядами, и *ARIMA* - моделями в частности, подключим следующие пакеты:

```
library("lubridate") # работа с датами
library("zoo") # работа с временными рядами
library("xts") # дополнительные функции для работы с временными
рядами
library("dplyr") # работа с наборами данных
library("ggplot2") # графики
library("forecast") # прогнозы
library("lmtest")# тестирование гипотез в линейных моделях
library('tseries') # работа с временными рядами
```

```
library("quantmod") # загрузка данных с различных источников
```

Для начала посмотрим, как сгенерировать тестовые выборки. Для этого воспользуемся функцией `arima.sim` из базового пакета `stats`.

Выполним симуляцию *ARIMA* (2,0,2)-процесса в 100 наблюдений по модели: $y_t = 0.9y_{t-1} - 0.5y_{t-2} + \varepsilon_t - 0.2\varepsilon_{t-1} + 0.3\varepsilon_{t-2}$:

```
arima.sim(n = 100, list(ar = c(0.9, -0.5), ma = c(-0.2, 0.3))
```

В этой функции необходимо указать объем выборки, а также в параметре *list* коэффициенты модели.

Модель вида $y_t = y_{t-1} + \varepsilon_t$ называется процессом случайного блуждания:

```
arima.sim(n=100, list(order=c(0,1,0)).
```

Модель вида $y_t = \varepsilon_t$ называется белым шумом:

```
arima.sim(n=100, list(order=c(0,0,0)))
```

В качестве примера временного ряда для последующего анализа используем набор данных о количестве арендованных велосипедов из службы проката . Данные собраны по дням.

```
daily_data = read.csv('day.csv', header=TRUE,
stringsAsFactors=FALSE)
```

Визуальный анализ данных

Построим временной ряд и визуально оценим его на предмет выбросов, волатильности или нерегулярных наблюдений. В данном случае число арендованных велосипедов меняется день ото дня. Однако даже при такой волатильности видны некоторые закономерности. Например, в зимние месяцы велосипеды используются реже, а в летние месяцы чаще (рис. 5.1):

```
daily_data$Date = as.Date(daily_data$dteday)
ggplot(daily_data, aes(Date, cnt)) + geom_line() +
scale_x_date('month') + ylab("Daily Bike Checkouts") + xlab("").
```

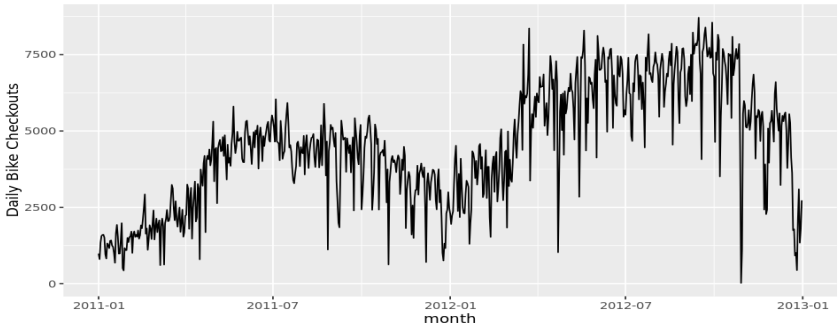


Рис. 5.1. **График временного ряда**

В некоторых случаях количество взятых напрокат велосипедов упало ниже 100 в день и увеличилось до более чем 4000 на следующий день. Это "выбросы", которые могут повлиять на модель, исказив значения параметров. *R* предоставляет удобный способ удаления "выбросов" временных рядов с помощью `tsclean()` из пакета `forecast`, которая определяет и заменяет выбросы, используя сглаживание и декомпозицию рядов. Этот метод также может вводить значения в ряд на место пропущенных, если таковые имеются. Обратите внимание, что используется команда `ts()` для создания объекта временного ряда для передачи в `tsclean()`:

```
count_ts = ts(daily_data[, c('cnt')])
daily_data$clean_cnt = tsclean(count_ts)
ggplot() + geom_line(data = daily_data, aes(x = Date,
y = clean_cnt)) + ylab('Cleaned Bicycle Count')
```

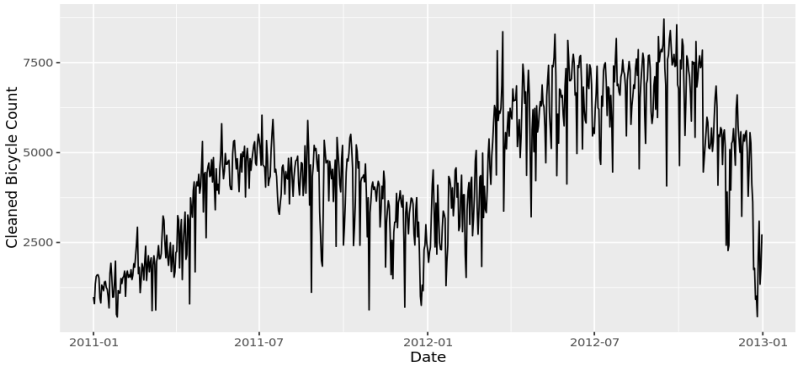


Рис. 5.2. **График временного ряда после удаления выбросов**

Даже после удаления выбросов (рис. 5.2) ежедневные данные остаются довольно волатильными. Визуально можно провести линию тренда через временной ряд, сглаживая шумные колебания. Эта линия может быть описана одним из самых простых, но очень полезных понятий в анализе временных рядов, известном как скользящее среднее. Это интуитивная концепция, которая усредняет точки в течение нескольких периодов времени, тем самым сглаживая наблюдаемые данные в более стабильные предсказуемые ряды.

Формально скользящее среднее (*MA*) порядка *m* может быть рассчитано взятием средних временного ряда *Y*, по *k* значений около каждой точки:

$$MA = \frac{1}{m} \sum_{j=-k}^{k} y_{t+j},$$

где $m = 2k + 1$. Вышеуказанная величина также называется симметричной скользящей средней, так как в вычислении участвуют данные с каждой стороны от точки.

Обратите внимание, что скользящее среднее в этом контексте отличается от компонента $MA(q)$ в приведенном выше определении *ARIMA*. Скользящее среднее $MA(q)$ как часть структуры *ARIMA* относится к лагам невязок и их линейным комбинациям, в то время как суммарная статистика скользящего среднего относится к методам сглаживания данных.

Чем шире окно скользящей средней, тем более гладким становится исходный ряд. В нашем примере с велосипедами мы можем взять еженедельную или ежемесячную скользящую среднюю, сглаживая ряд во что-то более стабильное и, следовательно, предсказуемое (рис. 5.3):

```
daily_data$cnt_ma = ma(daily_data$clean_cnt, order=7) # using the
clean count with no outliers
daily_data$cnt_ma30 = ma(daily_data$clean_cnt, order=30)

ggplot() +
geom_line(data = daily_data, aes(x = Date, y = clean_cnt, colour
= "Counts")) +
geom_line(data = daily_data, aes(x = Date, y = cnt_ma,   colour
= "Weekly Moving Average"))  +
geom_line(data = daily_data, aes(x = Date, y = cnt_ma30, colour
= "Monthly Moving Average"))  +
ylab('Bicycle Count')
```

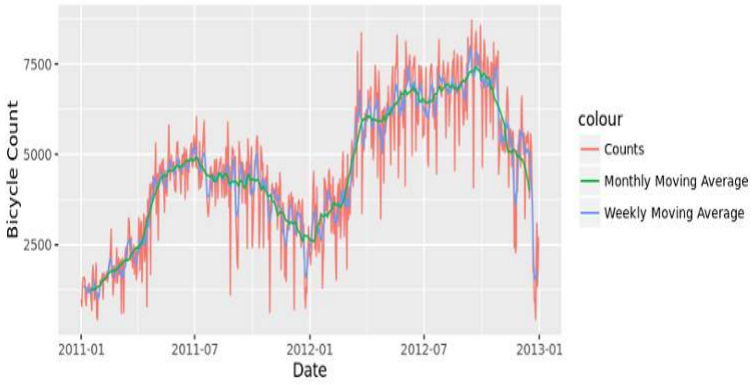


Рис. 5.3. **График временного ряда и сглаженные значения**

В дополнение к волатильности для моделирования ежедневных данных x может потребоваться определение нескольких уровней сезонности, таких как день недели, месяц, год, праздники и т.д. Для простоты смоделируем сглаженный ряд недельной скользящей средней (синяя линия на рисунке).

Декомпозиция

Основными структурными элементами при анализе временных рядов являются сезонность, тренд и цикл. Не каждый ряд будет иметь все три (или любую) из этих компонент, но, если они присутствуют, декомпозиция ряда может помочь понять его поведение и подготовить основу для построения прогноза.

Сезонная составляющая относится к колебаниям данных, связанных с календарными циклами. Например, летом и в теплую погоду на велосипедах может кататься больше людей, а в холодные месяцы меньше. Обычно сезонность фиксируется в некотором периоде (например, квартал или месяц).

Трендовая составляющая - это общая картина ряда: увеличивается или уменьшается количество арендуемых велосипедов с течением времени.

Циклическая составляющая включает уменьшающиеся или увеличивающиеся паттерны, которые не являются сезонными. Обычно компоненты тренда и цикла группируются вместе. Компонента тренд-цикличности оценивается с помощью скользящих средних.

Наконец, часть ряда, которая не может быть отнесена к сезонным, циклическим или трендовым компонентам, называется остатком, или ошибкой.

Процесс извлечения этих компонент называется декомпозицией.

Формально, если Y - количество арендованных велосипедов, мы можем разложить ряд двумя способами: используя аддитивную или мультипликативную модель:

$$Y = S_t + T_t + E_t$$
$$Y = S_t T_t E_t,$$

где S_t - сезонная составляющая, T_t - тренд и цикл, а E_t - оставшаяся ошибка.

Аддитивная модель обычно более уместна, когда сезонная или трендовая составляющая непропорциональна уровню ряда, так как мы можем просто сложить компоненты вместе, чтобы реконструировать ряд. С другой стороны, если компонента сезонности изменяется с изменением уровня или тренда ряда, простого "наложения" компонент будет недостаточно для восстановления ряда. В этом случае более подходящей может оказаться мультипликативная модель.

Как упоминалось выше, модели *ARIMA* могут быть адаптированы как к сезонным, так и к несезонным данным. Сезонная *ARIMA* требует более сложной спецификации структуры модели, хотя процесс определения (P, D, Q) аналогичен процессу выбора определения параметров порядка. Поэтому мы рассмотрим, как десезонализировать ряд и использовать модель *ARIMA* без сезонности.

Сначала вычислим сезонную составляющую данных с помощью функции `stl()`. STL вычисляет сезонную составляющую ряда с помощью сглаживания и корректирует исходный ряд путем вычитания сезонности:

```
count_ma = ts(na.omit(daily_data$cnt_ma), frequency=30)
decomp = stl(count_ma, s.window="periodic")
deseasonal_cnt <- seasadj(decomp)
plot(decomp)
```

Обратите внимание, что `stl()` по умолчанию предполагает аддитивную структуру модели. Подставьте параметр `allow.multiplicative.trend = TRUE` для использования мультипликативной модели.

В случае аддитивной модельной структуры та же задача разложения ряда и устранения сезонности может быть выполнена простым вычитанием сезонной составляющей из исходного ряда. `seasadj()` - удобный метод внутри пакета `forecast`.

Что касается параметра частоты в объекте временного ряда ts(), то мы задаем периодичность данных, т.е. количество наблюдений за период. Поскольку мы используем сглаженные ежедневные данные, у нас есть 30 наблюдений в месяц (рис. 5.4).

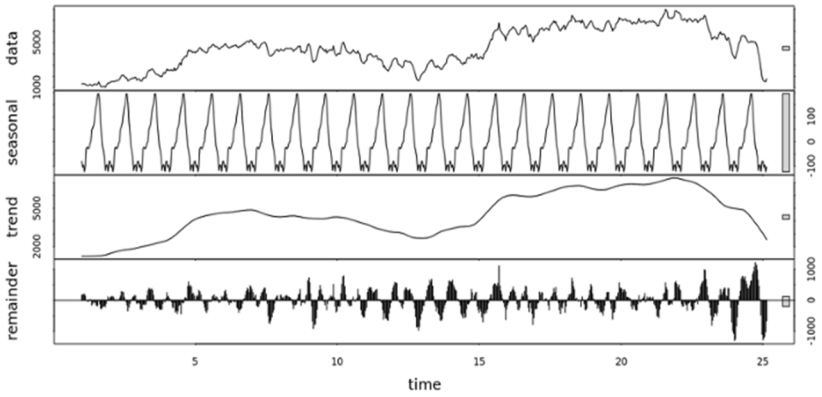


Рис. 5.4. **Десезонализация временного ряда**

Стационарность

Применение ARIMA-модели требует, чтобы временной ряд был стационарен. Ряд называется стационарным, когда его среднее, дисперсия и

автоковариация инвариантны по времени. Это предположение имеет интуитивный смысл, поскольку ARIMA использует предыдущие лаги рядов для моделирования, а моделирование стабильных рядов с согласованными свойствами предполагает меньшую неопределенность. На рис. 5.5 слева показан пример стационарного ряда, где значения данных колеблются с постоянной дисперсией вокруг среднего значения 1; справа показан нестационарный ряд; среднее значение этого ряда будет отличаться в разных временных отрезках.

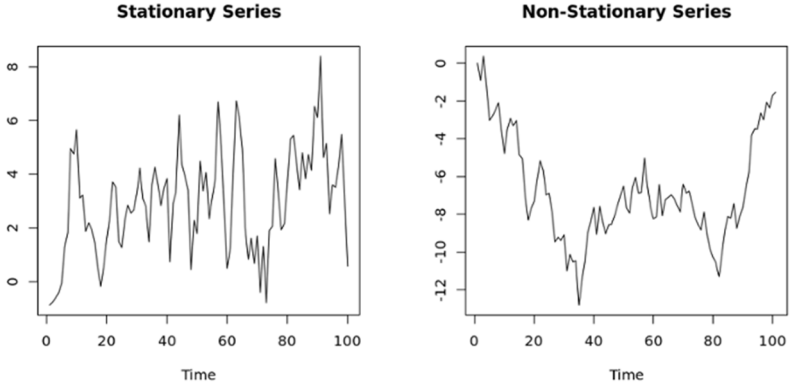


Рис. 5.5. **Стационарный и нестационарный временной ряд**

Расширенный тест Дики - Фуллера (ADF) является формальным статистическим тестом на стационарность. Нулевая гипотеза предполагает, что ряд нестационарен. Процедура ADF проверяет, можно ли объяснить изменение *Y* запаздывающим значением и линейным трендом. Если вклад запаздывающего значения в изменение *Y* несущественен и присутствует трендовая составляющая, то ряд является нестационарным и нулевая гипотеза не будет отвергнута.

Наш ряд данных о прокате велосипедов является нестационарным; среднее количество взятых напрокат велосипедов изменчиво во времени. Формальный тест ADF не отвергает нулевую гипотезу о нестационарности:

```
adf.test(count_ma, alternative = "stationary")
Augmented Dickey-Fuller Test
data: count_ma
Dickey-Fuller = -0.2557, Lag order = 8, p-value = 0.99
alternative hypothesis: stationary
```

Обычно нестационарные ряды можно корректировать простым преобразованием, таким как взятие разности. Дифференцирование ряда может помочь выявить тенденцию или циклы. Идея дифференцирования

состоит в том, что если исходный ряд данных не имеет постоянных свойств с течением времени, то переход от одного периода к другому имеет. Разница рассчитывается путем вычитания значений одного периода из значений предыдущего периода:

$$\Delta^1 Y_t = Y_t - Y_{t-1}.$$

Различия более высокого порядка вычисляются аналогичным образом. Например, разность второго порядка ($d = 2$) просто выражается через первые разности:

$$\Delta^2 Y_t = \Delta^1 Y_t - \Delta^1 Y_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}).$$

Аналогичным образом дифференцирование может использоваться, если существует сезонность при определенных лагах. В таком случае вычитание значения для "сезонности" предыдущего периода представляет собой изменение от одного периода к другому, а также от одного сезона к другому:

$$\Delta^d Y_t = (Y_t - Y_{t-s}) - (Y_{t-1} - Y_{t-s-1}).$$

Число рассчитанных разностей представляет *d*-компоненту *ARIMA*. Теперь перейдем к диагностике, которая поможет определить порядок разностей.

Автокорреляции и выбор порядка модели

Автокорреляционные функции (АКФ, англ. ACF) являются полезным визуальным инструментом для определения стационарности ряда. Эти графики также могут помочь выбрать параметры для модели *ARIMA*. Если ряд коррелирован с его лагами, то, как правило, существуют некоторые трендовые или сезонные компоненты, и поэтому его статистические свойства не являются постоянными во времени.

АКФ показывает корреляцию между рядом и его лагами. В дополнение к определению порядка дифференцирования графики АКФ могут помочь в определении порядка модели $MA(q)$. Графики частной АКФ (ЧАКФ, англ. PACF), как следует из названия, отображают корреляцию между переменной и ее лагами, которая не объясняется предыдущими лагами. Графики ЧАКФ полезны при определении порядка модели $AR\ (p)$.

R показывает 95 %-ные границы значимости синими пунктирными линиями. Существует значительная автокорреляция со многими лагами в нашем наборе данных по прокату велосипедов, как показано на рис. 5.6. Однако это может быть связано с переносом корреляции с первого или раннего лага, так как график ЧАКФ (рис. 5.7) показывает только всплеск при лагах 1 и 7:

```
Acf(count_ma, main='')
Pacf(count_ma, main='')
```

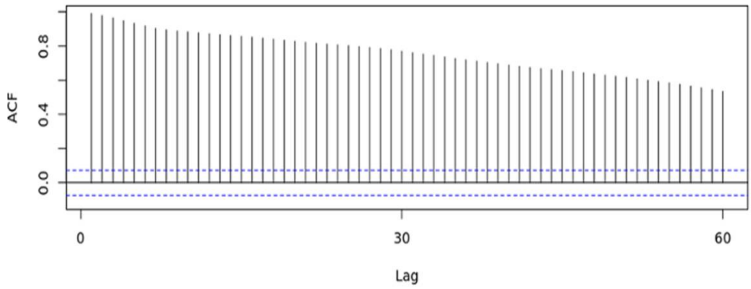


Рис. 5.1. **График АКФ**

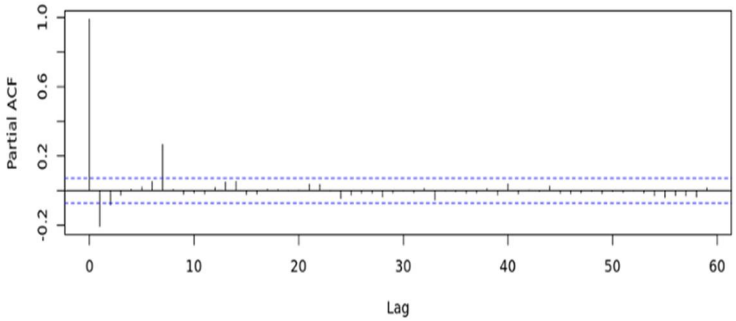


Рис. 5.7. **График ЧАКФ**

Мы можем начать с порядка $d = 1$ и повторно оценить, требуется ли дальнейшее дифференцирование.

Расширенный тест Дики - Фуллера на разностных данных (рис. 5.8) отвергает нулевые гипотезы нестационарности. При построении разностного ряда мы видим колебания около 0 без видимого сильного тренда. Это говорит о том, что разности первого порядка являются достаточными и должны быть включены в модель.

```
count_d1 = diff(deseasonal_cnt, differences = 1)
plot(count_d1)
adf.test(count_d1, alternative = "stationary")

Augmented Dickey-Fuller Test

data:  count_d1
Dickey-Fuller = -9.9255, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary
```

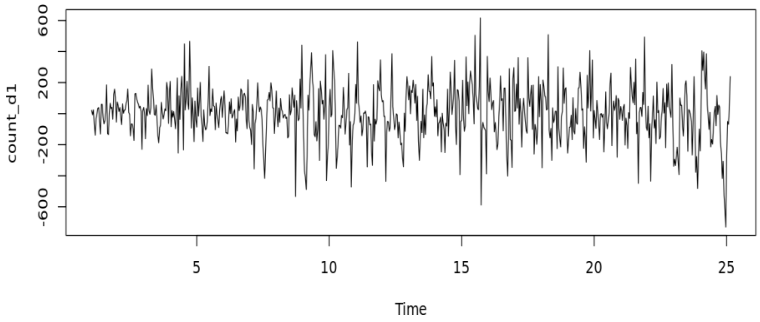


Рис. 5.8. **Первые разности временного ряда**

Определим параметры p и q, рассчитав АКФ и ЧАКФ для ряда первого порядка разностей:

```
Acf(count_d1, main='ACF for Differenced Series')
Pacf(count_d1, main='PACF for Differenced Series')
```

Существуют значительные автокорреляции при лагах 1 и 2 и выше (рис. 5.9). Графики частной автокорреляции показывают значительный всплеск при лагах 1 и 7 (рис. 5.10). Это предполагает, что можно протестировать модели с компонентами *AR* или *MA* порядка 1, 2 или 7. Всплеск в лаге 7 дает возможность предположить, что существует сезонная составляющая с периодом, возможно, в день недели.

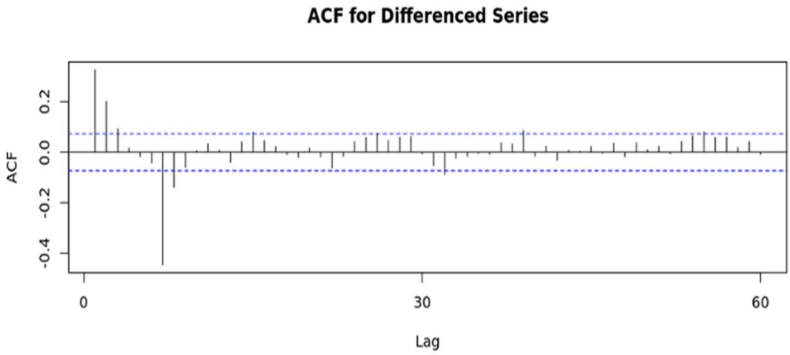


Рис. 5.9. **АКФ для первых разностей временного ряда**

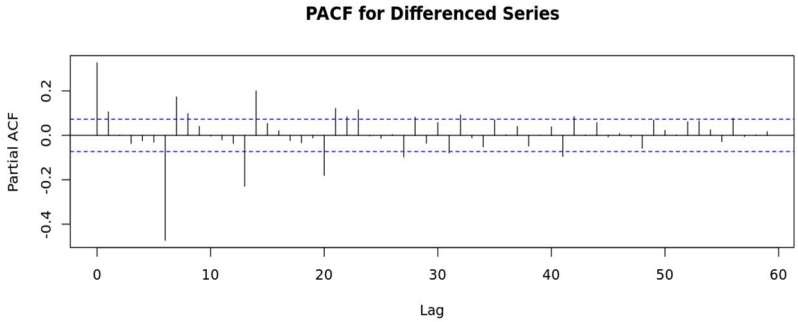


Рис. 5.10. **ЧАКФ для первых разностей временного ряда**

Подгонка ARIMA-модели

Пакет forecast позволяет пользователю явно указать порядок модели с помощью функции `arima()` или автоматически сгенерировать набор оптимальных (*p, d, q*) с помощью `auto.arima()`. Эта функция выполняет поиск комбинаций параметров модели и выбирает тот набор, который оптимизирует критерии качества модели.

Существует ряд таких критериев для сравнения качества подгонки различных моделей. Два из наиболее широко используемых - это информационный критерий Акаике (*AIC*) и Байесовский информационный критерий (*BIC*). Эти критерии тесно взаимосвязаны и могут быть истолкованы как оценка того, сколько информации будет потеряно при выборе данной модели. При сравнении моделей необходимо минимизировать *AIC* и *BIC*.

Пока `auto.arima()` может быть очень полезна, по-прежнему важно выполнить предыдущие пункты, чтобы понять методологию Бокса - Дженкинса и интерпретировать результаты модели. Обратите внимание, что `авто.arima()` также позволяет пользователю задать максимальный порядок для *(p, d, q)*, который по умолчанию равен 5.

Мы можем указать структуру ARIMA без сезонности и подогнать модель к данным, полученным после десезонализации. Параметры (1,1,1), предложенные автоматизированной процедурой, соответствуют нашим ожиданиям, основанным на шагах выше; модель включает в себя разность степени 1 и использует авторегрессионный член первого лага и модель скользящего среднего порядка 1:

```
auto.arima(deseasonal_cnt, seasonal=FALSE)
```

```
Series: deseasonal_cnt
ARIMA(1,1,1)

Coefficients:
         ar1      ma1
      0.5510  -0.2496
s.e.  0.0751   0.0849

sigma^2 estimated as 26180:  log likelihood=-4708.91
AIC=9423.82   AICc=9423.85   BIC=9437.57
```

Модель может быть записана в виде:

$$\Delta^d \hat{Y}_t = 0,551 Y_{t-1} - 0,2496 e_{t-1} + E,$$

где E - некоторая ошибка, и порядок разности равен единице.

Оценка модели

Мы можем начать с изучения графиков АКФ и ЧАКФ для остатков модели (рис. 5.11). Если параметры и структура модели заданы правильно, то значительных автокорреляций не будет.

```
fit<-auto.arima(deseasonal_cnt, seasonal=FALSE)
tsdisplay(residuals(fit),   lag.max=45,   main='(1,1,1)   Model
Residuals')
```

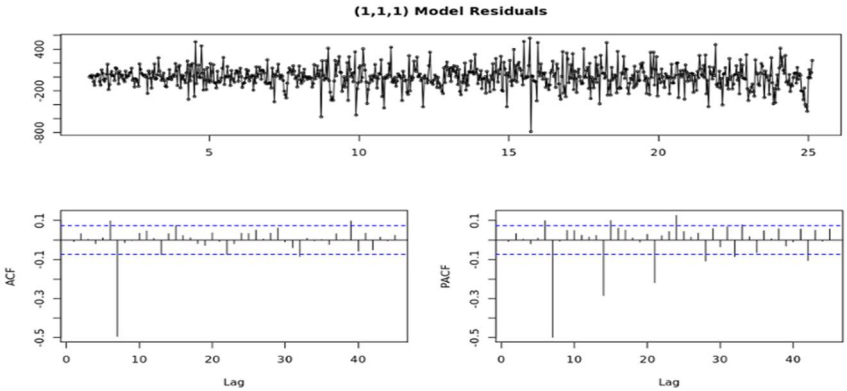


Рис. 5.11. **АКФ и ЧАКФ для остатков модели без сезонности**

В графиках АКФ и ЧАКФ, а также остатков модели присутствует четкая закономерность, повторяющаяся с лагом 7. Это говорит о том, что наша модель может быть лучше с другой спецификацией, такой как $p = 7$ или $q = 7$.

Мы можем повторить процесс подгонки с учетом компонента *MA* (7) и снова изучить диагностические графики. На этот раз значимых автокорреляций нет. Если модель указана неправильно, это обычно отражается в остатках в виде трендов, асимметрии или любых других закономерностей, не захваченных моделью. В идеале остатки должны выглядеть как белый шум, т.е. быть нормально распределенными. Для построения этих диагностических моделей можно использовать удобную функцию `tsdisplay()`. Графики остатков показывают меньший диапазон ошибок, более или менее центрированный вокруг нуля (рис. 5.12). Мы можем наблюдать, что *AIC* меньше для модели порядка (1, 1, 7):

```
fit2 = arima(deseasonal_cnt, order=c(1,1,7))

fit2

tsdisplay(residuals(fit2), lag.max=15, main='Seasonal Model
Residuals')

Call:
arima(x = deseasonal_cnt, order = c(1, 1, 7))

Coefficients:
      ar1     ma1     ma2     ma3     ma4     ma5     ma6     ma7
   0.2803  0.1465  0.1524  0.1263  0.1225  0.1291  0.1471  -
0.8353
s.e.  0.0478  0.0289  0.0266  0.0261  0.0263  0.0257  0.0265
0.0285

sigma^2 estimated as 14392:  log likelihood = -4503.28,  aic =
9024.56
```

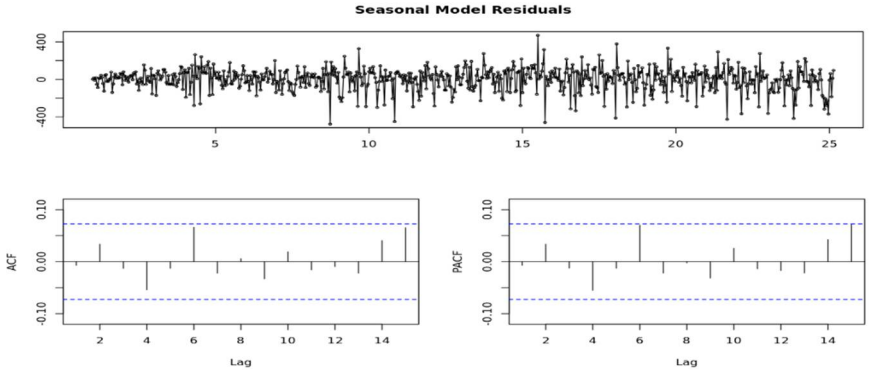


Рис. 5.12. **АКФ и ЧАКФ для остатков модели с сезонностью**

Прогнозировать с использованием подогнанной модели в *R* очень просто. Мы можем указать горизонт прогноза *h* периодов вперед для прогнозов, которые будут сделаны, и использовать подогнанную модель для расчета этих прогнозов (рис. 5.13):

```
fcast <- forecast(fit2, h=30)
plot(fcast)
```

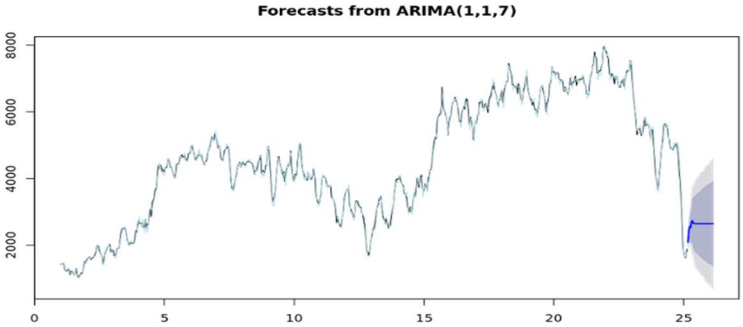


Рис. 5.13. **Прогноз по модели ARIMA (1, 1, 7) без сезонности**

Тем не менее, синяя линия, представляющая прогноз, кажется очень наивной: она приближается к прямой довольно скоро, что кажется маловероятным, учитывая прошлое поведение ряда. Напомним, что модель предполагает ряд без сезонности и отличается от исходных нестационарных данных. Иными словами, построенные на графике прогнозы основаны на предположении, что других сезонных колебаний в данных не будет, а изменение количества взятых напрокат велосипедов от одного дня к другому будет более или менее постоянным в среднем и дисперсии. Этот прогноз может быть наивной моделью, но он иллюстрирует процесс выбора модели *ARIMA*, а также может служить в качестве ориентира для оценки по мере построения более сложных моделей.

Как можно улучшить прогноз? Одним из простых изменений является добавление сезонной компоненты, которую мы извлекли ранее. Другой подход заключается в том, чтобы включить в модель компоненты (P, D, Q), которые по умолчанию используются в `auto.arima()`. Перенастраивая модель на те же данные, мы видим, что все еще может быть некоторая сезонность в ряду с сезонной составляющей, описанной *AR* (1):

```
fit_w_seasonality = auto.arima(deseasonal_cnt, seasonal=TRUE)
fit_w_seasonality

Series: deseasonal_cnt
```

```
ARIMA(2,1,2)(1,0,0)[30]

Coefficients:
         ar1      ar2      ma1     ma2     sar1
      1.3644  -0.8027  -1.2903  0.9146  0.0100
s.e.  0.0372   0.0347   0.0255  0.0202  0.0388

sigma^2 estimated as 24810:  log likelihood=-4688.59
AIC=9389.17   AICc=9389.29   BIC=9416.68

seas_fcast <- forecast(fit_w_seasonality, h=30)
plot(seas_fcast)
```

Обратите внимание, что параметры (*p, d, q*) также изменились после включения сезонной составляющей. Мы можем пройти тот же процесс оценки остатков модели и графиков АКФ/ЧАКФ и корректировки структуры, если это необходимо. Например, мы замечаем, что та же картина присутствует в автокорреляциях с лагом 7, что предполагает, что может потребоваться компонента более высокого порядка (рис. 5.14, 5.15).

Обе вышеприведенные прогнозные оценки имеют доверительные границы: 80 %-ные доверительные интервалы выделены темно-синим цветом, а 95 %-ные - светло-синим. Долгосрочные прогнозы обычно имеют большую неопределенность, так как модель использует для будущих значений ранее предсказанные значения. В таком случае это отражается в форме доверительных границ, поскольку они начинают расширяться с увеличением горизонта.

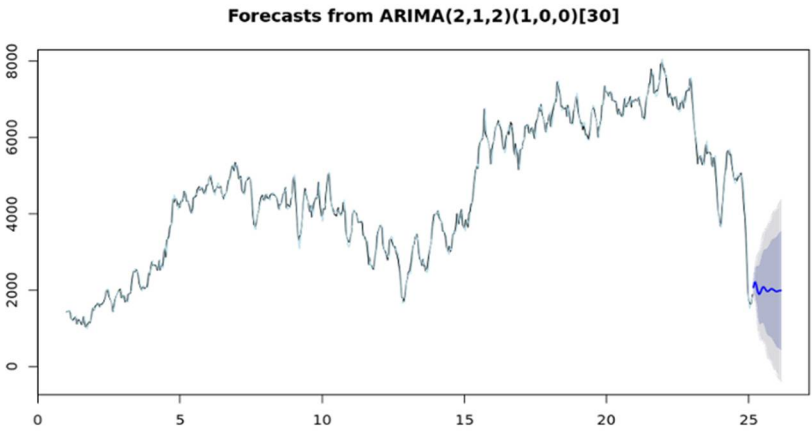


Рис. 5.14. **Прогноз по модели ARIMA (2, 1, 2) (1,0,0) с сезонностью**

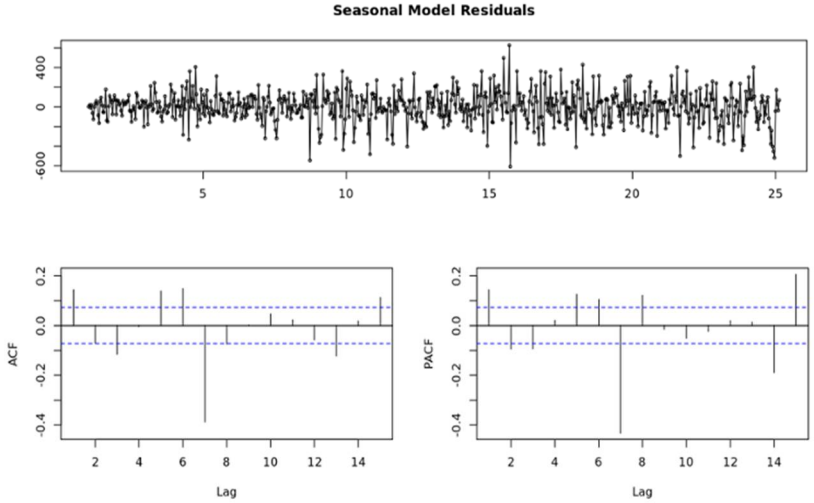


Рис. 5.15. **Остатки модели ARIMA (2, 1, 2) (1,0,0) с сезонностью**

Поведение границ в доверительных интервалах может сигнализировать о необходимости использования более стабильной модели. Очень важно смотреть на границы прогноза и иметь в виду ожидаемую ошибку, связанную с точечными оценками.

2. Загрузка данных из различных источников в *R*

Для корректного перевода дат на английский язык в русскоязычной *Windows* необходимо сначала выполнить следующую команду:

```
Sys.setlocale("LC_TIME","C")
```

Пакет `quantmod` позволяет загружать данные из нескольких источников, а именно:

– Yahoo! Finance (OHLC data) http://finance.yahoo.com/;

– Federal Reserve Bank of St. Louis FRED® (11,000 экономических временных рядов) http://research.stlouisfed.org/fred2/;

– Oanda, The Currency Site (FX and Metals) http://www.oanda.com/.

Для загрузки данных используется одна и та же функция `getSymbols`, например:

```
    #данные о стоимости акций Гугл с finance.yahoo.com за период с 1 января 2015 г. по 1 декабря 2015 г.
    getSymbols("GOOG", src="yahoo", from="2015-01-01", to="2015-12-01")
```

GOOG - это краткое название акций компании на бирже (ти́кер). Теперь эти данные загружены в переменную с таким же названием GOOG.

Если не указывать начальную или конечную дату, то будут загружены все доступные данные. Можно посмотреть шесть первых значений и шесть последних значений командами `head` и `tail`, соответственно.

```
head(GOOG)
```

```
 GOOG.Open    GOOG.High    GOOG.Low    GOOG.Close    GOOG.Volume
GOOG.Adjusted
2015-01-02 529.0124 531.2724 524.1023 524.8124 1447600 524.8124
2015-01-05 523.2624 524.3324 513.0623 513.8723 2059800 513.8723
2015-01-06 515.0024 516.1773 501.0523 501.9623 2899900 501.9623
2015-01-07 507.0023 507.2463 499.6522 501.1023 2065100 501.1023
2015-01-08 497.9922 503.4823 491.0022 502.6823 3353600 502.6823
2015-01-09 504.7623 504.9223 494.7922 496.1723 2069400 496.1723
```

```
tail(GOOG)
```

```
 GOOG.Open    GOOG.High    GOOG.Low    GOOG.Close    GOOG.Volume
GOOG.Adjusted
2015-11-23 757.45 762.708 751.82 755.98 1414500 755.98
2015-11-24 752.00 755.279 737.63 748.28 2333100 748.28
2015-11-25 748.14 752.000 746.06 748.15 1122100 748.15
2015-11-27 748.46 753.410 747.49 750.26 838500 750.26
2015-11-30 748.81 754.930 741.27 742.60 2035300 742.60
2015-12-01 747.11 768.950 746.70 767.04 2129900 767.04
```

Посмотрим на график (рис. 5.16):

```
barChart(GOOG, theme = "white")
```

Рис. 5.16. **График стоимости акций Гугл GOOG**

Приведем еще несколько примеров:

```
#данные о стоимости акций Yahoo с finance. yahoo.com
```

```
getSymbols("YHOO",src="yahoo")
#данные о стоимости акций Apple с finance.yahoo.com
getSymbols("AAPL",src="yahoo",  from="2015-01-01",  to="2015-12-
01")
```

Можно построить график японских свечей (рис. 5.17):

```
candleChart(AAPL, multi.col=TRUE, theme="white")
```

Рис. 5.17. **График стоимости акций *Apple AAPL***

Для более детальной информации можно обратиться на сайт разработчиков пакета (http://www.quantmod.com/).

Контрольные вопросы

1. В чем заключается методология Бокса - Дженкинса?
2. В чем отличие автокорреляционной и частной автокорреляционной функций?
3. Что такое случайное блуждание и белый шум?
4. Как можно сравнить модели временного ряда по точности?

Лабораторная работа 6
Прогнозирование временных рядов с помощью логистических моделей тренда

Цель работы: смоделировать и спрогнозировать динамику численности пользователей Интернета в различных странах с помощью логистических моделей, осуществить обоснованный выбор наилучшей из них.

Задание. Лабораторная работа выполняется с использованием оригинального программного комплекса, разработанного авторами данного практикума: "Эконометрическое моделирование рядов динамики". Варианты лабораторной работы находятся в приложении.

Имеются статистические данные для одной из стран - участниц Организации экономического сотрудничества и развития (ОЭСР) о числе абонентов высокоскоростного доступа в Интернет на 100 человек населения по полугодиям, начиная с 1-го полугодия 2002 г. по 1-е полугодие 2010 г.

1. Смоделируйте и спрогнозируйте временной ряд каждой из 4 моделей последовательно на 1, 2 и 3 шага прогноза.

2. Для каждой модели и глубины прогноза определите значение точки перегиба $\left(k*, Y(k*)\right)$.

3. Выберите наилучшую модель на основании показателей R^2 и *MAPE*.

Для каждого шага прогноза составьте таблицу следующего вида (табл. 6.1):

Таблица 6.1

Вид итоговой таблицы для отчета по лабораторной работе

k	Y_k	Y_k^o			
		Модель Ферхюльста	**Модель Рамсея**	**Модель Гомпертца**	***GRM***
…	…	…	…	…	…
R^2					
MAPE					
Уровень насыщения					
k^*					
$Y(k*)$					

Как изменяются значения R^2, оценки уровня насыщения и точки перегиба в зависимости от объема выборки для каждой модели?

Как изменяются значения MAPE-оценки в зависимости от глубины прогноза для каждой модели? Постройте графики.

Задание для всей группы

Сравните полученные результаты для различных вариантов лабораторной работы:

– какая страна из участниц ОЭСР раньше всех начала внедрять высокоскоростной доступ в Интернет (значение абсциссы точки перегиба минимально)?

– в какой стране самая высокая численность абонентов, пользующихся высокоскоростным Интернетом?

Указания к выполнению

Модели логистической динамики и методы их идентификации

Среди моделей эконометрики большую и практически важную группу образуют логистические модели (логисты), или, как их еще называют, сигмоидальные модели (сигмоиды), или пользуются и таким названием, как S-образные кривые роста.

Логистическая зависимость может отражать тренд сложной, эволюционирующей динамики зависимости одного экономического показателя (определяемого) от другого (определяющего) экономического показателя в случае "пространственной динамики".

Логистическая зависимость характеризует чаще и "временную динамику" определяемого показателя от времени, как бы "интегрируя" через время действие всех факторов.

Наиболее часто в приложениях рассматривается динамика логистического роста определяемого показателя. При этом эволюция динамики определяемого показателя отражается в том, что скорость его роста изменяется с течением времени (первая производная логистической функции неотрицательна, вторая производная меняет свой знак с “+” на “-“, проходя через *точку перегиба*), а его рост является ограниченным: стремится к некоторому пределу. Логистическая динамика уменьшения определяемого показателя встречается реже: имеет место отрицательная первая производная, а в точке перегиба вторая производная меняет свой знак.

На рис. 6.1 представлен вид растущей логистической модели. Она подходит для описания такого процесса, при котором определяемый показатель проходит полный цикл развития. Можно, конечно, логистическую тенденцию считать объединением трех разных по типу

трендов (тенденций): параболического с ускоряющимся ростом на первом этапе, линейного - на втором этапе и гиперболического с замедляющимся ростом на втором этапе.

Но предпочтительнее рассмотрение всего цикла развития как единого цикла тенденций со сложными переменными (эволюционирующими) свойствами, но с постоянным направлением изменений в сторону увеличения (или уменьшения) уровней определяемого показателя.

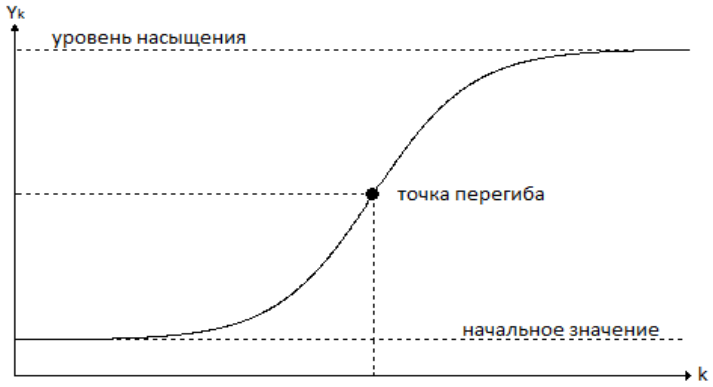


Рис. 6.1. **Общий вид логистической функции роста**

Таким образом, основными характеристиками логистической функции являются:

– нижняя горизонтальная асимптота, или начальное значение функции;

– точка перегиба, в которой *значение второй производной функции равно нулю;*

– верхняя горизонтальная асимптота, или уровень насыщения.

В общем случае положение точки перегиба не является фиксированным, а кривая, изображенная на рис. 6.1, не обязательно будет симметричной: для нее значение ординаты точки перегиба всегда равно половине уровня насыщения.

На рис. 6.2 представлены примеры асимметричных логистических моделей. Используются следующие обозначения: 0 - точка, соответствующая половине уровня насыщения; 1 - точка перегиба находится левее половины уровня насыщения; 2 - точка перегиба находится правее половины уровня насыщения.

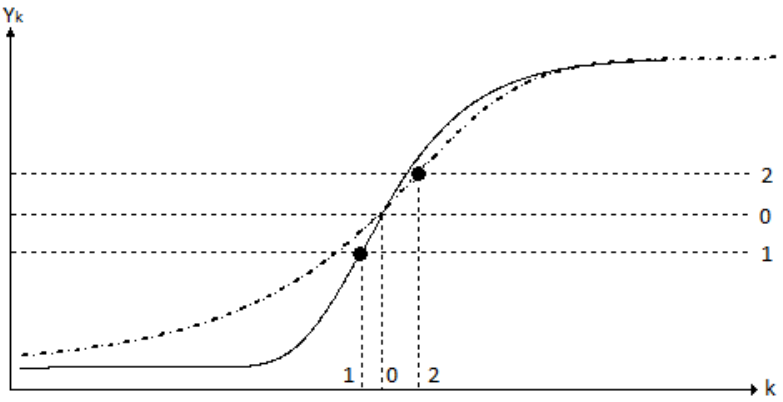


Рис. 6.2. **Асимметричные логистические функции роста**

Широко распространенными областями применения логистических функций в моделировании являются:

– жизненные циклы товаров, в частности, изменение спроса на товары, обладающие способностью достигать некоторого уровня насыщения;

– доля насыщения рынка новыми товарами и услугами, в том числе описание числа пользователей Интернета и сотовой связи;

– оценка изменения числа семей, имеющих радио и телевидение;

– рост населения страны в страховых исследованиях;

– развитие биологических популяций;

– развитие тех или иных показателей технологических нововведений, смена технологий;

– доля неграмотных жителей среди населения;

– динамика антисоциального поведения и др.

В данной лабораторной работе в качестве примера будет рассматриваться временная динамика роста числа абонентов высокоскоростного доступа в Интернет в ряде стран, входящих в ОЭСР. Модели логистической динамики наблюдений уровней определяемого показателя Y_k содержат обязательно логистический тренд D_k и стохастическую компоненту ε_k. Возможно присутствие в модели и сезонных, и циклических компонент.

Обратимся к наиболее простой аддитивной структуре модели временного ряда:

$$Y_k = D_k + \varepsilon_k,$$

а для стохастической компоненты ε_k будем считать справедливыми условия Гаусса - Маркова, что позволит, применяя метод наименьших квадратов (МНК) для идентификации параметров D_k, получить их оптимальные оценки.

Известно более двадцати моделей логистической динамики различных по сложности и, соответственно, по числу использованных в них параметров, по области применения [4, 10, 11]. Три аналитических выражения широко распространенных на практике логистических моделей, которые будут использоваться при выполнении данной лабораторной работы, сведены в табл. 6.2.

Одна из них модель (GRM - generalized rational innovation diffusion model) задана не в привычном виде аналитической функции, а пересчитывается на основе своего предыдущего значения [3]:

$$f_1 = f_0 + \alpha \frac{f_0(A_0 - f_0)}{A_0 - (1-\sigma) f_0}; \; f_2 = f_1 + \alpha \frac{f_1(A_0 - f_1)}{A_0 - (1-\sigma) f_1} \text{ и т. д.}$$

Таким образом, модель содержит 4 параметра, т.е. начальное значение детерминированной компоненты f_0 также является параметром модели, как и α, A_0, σ. Задача оценки параметров (идентификация) логистической функции в общем случае нетривиальна, поскольку применение МНК непосредственно к самой модели требует минимизации нелинейной функции ошибки.

Так, модель Ферхюльста идентифицируется с помощью численного решения МНК

$$A_0^0, A_1^0, \alpha^0 = \arg \min_{A_0, A_1, \alpha} \sum_{k=1}^{N} \left(Y_k - \frac{A_0}{1 + A_1 e^{-\alpha k}} \right)$$

методом Левенберга - Марквардта, являющегося комбинацией градиентного метода и метода Гаусса - Ньютона, пояснения на примерах приведены и в [10].

Идентификация модели Рамсея осуществляется на основе конструирования обобщенной параметрической модели авторегрессии - скользящего среднего [9]:

$$Y_k = (2\lambda + 1) Y_{k-1} - (\lambda^2 + 2\lambda) Y_{k-2} + \lambda^2 Y_{k-3} + \xi_k,$$

$$\xi_k = \varepsilon_k - (2\lambda + 1)\varepsilon_{k-1} + (\lambda^2 + 2\lambda)\varepsilon_{k-2} - \lambda^2 \varepsilon_{k-3},$$

где ξ_k - гомоскедастическая стохастическая компонента.

Таблица 6.2

Логистические модели и их характеристики

Название модели	**Вид модели**	**Начальное значение**	**Уровень насыщения**	**Точка перегиба** $(k^*, Y(k^*))$		**Симметричность**
				абсцисса	**ордината**	
Модель Ферхюльста (Перла -Рида)	$Y_k = \frac{A_0}{1 + A_1 e^{-\alpha k}} + \varepsilon_k$	0	A_0	$k^* = -\frac{1}{\alpha}\ln\left(\frac{1}{A_1}\right)$	$Y(k^*) = \frac{A_0}{2}$	Симметричная
Модель Рамсея	$Y_k = C(1-(1+ +\alpha k)e^{-\alpha k}) + B_0 + \varepsilon_k$	B_0	$C + B_0$	$k^* = \frac{1}{\alpha}$	$Y(k^*) = B_0 + C - \frac{2C}{e}$	Асимметричная
Модель Гомпертца	$Y_k = C + A_0 e^{-e^{-\alpha(k-k_0)}} + \varepsilon_k$	C	$C + A_0$	$k^* = k_0$	$Y(k^*) = C + \frac{A_0}{e}$	Асимметричная
GRM (generalized rational innovation diffusion model)	$Y_k = f_{k-1} + +\alpha\frac{f_{k-1}(A_0 - f_{k-1})}{A_0 - (1-\sigma)f_{k-1}} + \varepsilon_k$	f_0	A_0	-	$Y(k^*) = \frac{A_0\left(\sqrt{\sigma} - 1\right)}{\sigma - 1}$	Асимметричная

$$\lambda^o = \arg\min_{\lambda} \sum_{k=3}^{n} \left(Y_k - (2\lambda+1)Y_{k-1} + (\lambda^2+2\lambda)Y_{k-2} - \lambda^2 Y_{k-3} \right)^2, \quad \alpha^o = -\ln \lambda^o;$$

$$C^o, B_0^o = \arg\min_{C, B_0} \sum_{k=1}^{N} \left(Y_k - C(1-(1+\alpha^o k)e^{-\alpha^o k}) - B_0 \right)^2.$$

Для идентификации модели Гомпертца используется метод Гаусса - Ньютона, который сводит задачу минимизации нелинейной функции МНК к итерационной минимизации линейных функций [11]. При идентификации модели *GRM* используется эвристический алгоритм *RPROP*, разработанный в теории нейронных сетей [8].

Выбор модели, в большей мере "подходящей" статистическим данным, осуществляется, в зависимости от содержания задачи, по большей точности моделирования, или по большей точности прогнозирования, или с учетом обеих характеристик.

Для характеристики качества моделирования будем использовать коэффициент детерминации:

$$R^2 = \frac{\sum_{k=1}^{N} \left(Y_k^o - M[Y_k] \right)^2}{\sum_{k=1}^{N} \left(Y_k - M[Y_k] \right)^2} = 1 - \frac{\sum_{k=1}^{N} \left(Y_k^o - Y_k \right)^2}{\sum_{k=1}^{N} \left(Y_k - M[Y_k] \right)^2},$$

где Y_k^o - модельные значения ряда динамики.

Обычно считают удовлетворительным качество моделирования при $0{,}7 \leq R^2 \leq 1$.

Качество прогнозирования будем определять с помощью MAPE-оценки:

$$MAPE = \frac{1}{l} \sum_{k=1}^{l} \left| \frac{Y_k - Y_k^o}{Y_k} \right| \cdot 100\%,$$

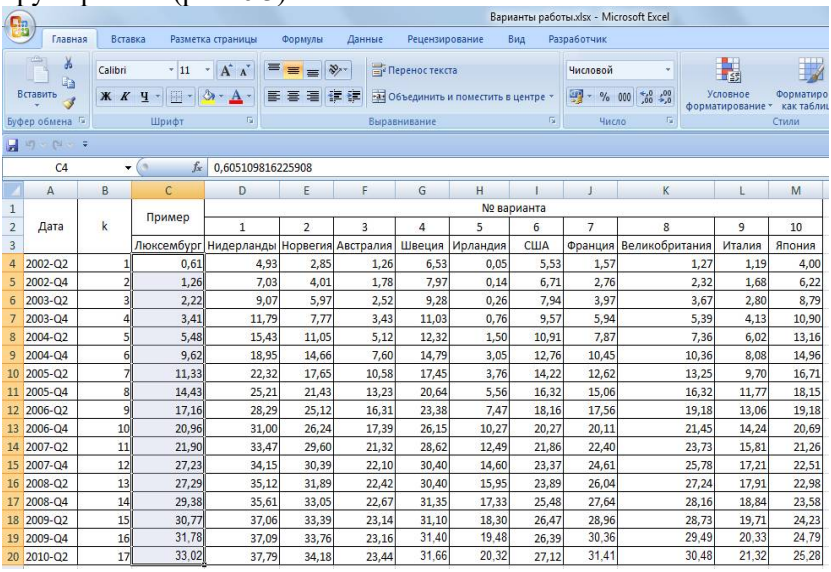
где l - глубина (горизонт) прогноза, который обычно не превышает одной трети от объема анализируемой выборки.

Высокой точностью прогнозирования считают обычно MAPE$\leq$ 10 %.

1. **Моделирование и прогнозирование временного ряда**

В качестве примера рассмотрим данные о числе абонентов высокоскоростного доступа в Интернет на 100 человек населения в Люксембурге.

Скопируйте в буфер обмена ряд данных, соответствующих вашему номеру варианта (рис. 6.3).



Варианты работы.xlsx - Microsoft Excel

Главная | Вставка | Разметка страницы | Формулы | Данные | Рецензирование | Вид | Разработчик

C4: 0,605109816225908

	A	B	C	D	E	F	G	H	I	J	K	L	M
1				№ варианта									
2	Дата	k	Пример	1	2	3	4	5	6	7	8	9	10
3			Люксембург	Нидерланды	Норвегия	Австралия	Швеция	Ирландия	США	Франция	Великобритания	Италия	Япония
4	2002-Q2	1	0,61	4,93	2,85	1,26	6,53	0,05	5,53	1,57	1,27	1,19	4,00
5	2002-Q4	2	1,26	7,03	4,01	1,78	7,97	0,14	6,71	2,76	2,32	1,68	6,22
6	2003-Q2	3	2,22	9,07	5,97	2,52	9,28	0,26	7,94	3,97	3,67	2,80	8,79
7	2003-Q4	4	3,41	11,79	7,77	3,43	11,03	0,76	9,57	5,94	5,39	4,13	10,90
8	2004-Q2	5	5,48	15,43	11,05	5,12	12,32	1,50	10,91	7,87	7,36	6,02	13,16
9	2004-Q4	6	9,62	18,95	14,66	7,60	14,79	3,05	12,76	10,45	10,36	8,08	14,96
10	2005-Q2	7	11,33	22,32	17,65	10,58	17,45	3,76	14,22	12,62	13,25	9,70	16,71
11	2005-Q4	8	14,43	25,21	21,43	13,23	20,64	5,56	16,32	15,06	16,32	11,77	18,15
12	2006-Q2	9	17,16	28,29	25,12	16,31	23,38	7,47	18,16	17,56	19,18	13,06	19,18
13	2006-Q4	10	20,96	31,00	26,24	17,39	26,15	10,27	20,27	20,11	21,45	14,24	20,69
14	2007-Q2	11	21,90	33,47	29,60	21,32	28,62	12,49	21,86	22,40	23,73	15,81	21,26
15	2007-Q4	12	27,23	34,15	30,39	22,10	30,40	14,60	23,37	24,61	25,78	17,21	22,51
16	2008-Q2	13	27,29	35,12	31,89	22,42	30,40	15,95	23,89	26,04	27,24	17,91	22,98
17	2008-Q4	14	29,38	35,61	33,05	22,67	31,35	17,33	25,48	27,64	28,16	18,84	23,58
18	2009-Q2	15	30,77	37,06	33,39	23,14	31,10	18,30	26,47	28,96	28,73	19,71	24,23
19	2009-Q4	16	31,78	37,09	33,76	23,16	31,40	19,48	26,39	30,36	29,49	20,33	24,79
20	2010-Q2	17	33,02	37,79	34,18	23,44	31,66	20,32	27,12	31,41	30,48	21,32	25,28

Рис. 6.3. **Варианты лабораторной работы**

Запустите *EconoModel.exe*.

Выделите первую ячейку таблицы, расположенной в левой части окна, и выберите пункт меню "Правка" - "Вставить" (рис. 6.4). Установите в поле "Объем выборки" значение, равное 16, и в поле "Глубина прогноза" значение, равное 2.

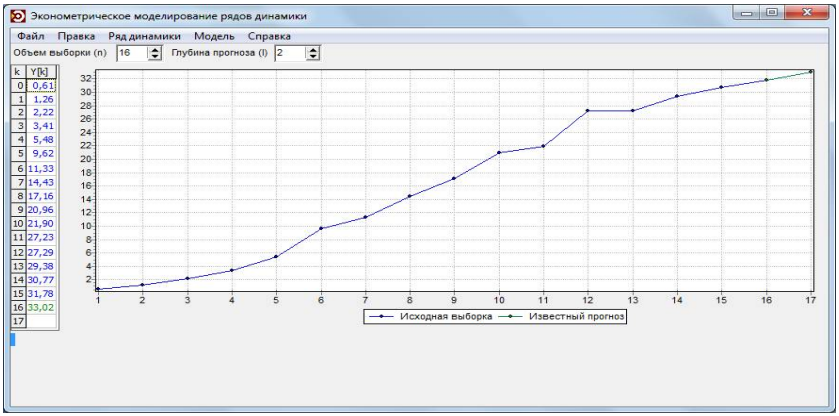


Рис. 6.4. **Главное окно программы с исходным рядом динамики**

Выберите пункт меню "Модель" - "Добавить". В открывшемся окне в списке выбора моделей выберите модель Ферхюльста

$$Y_k = \frac{A_0}{1 + A_1 e^{-\alpha k}} + \varepsilon_k$$

и нажмите кнопку "*OK*" (рис. 6.5). Аналогичным образом добавляются оставшиеся три модели.

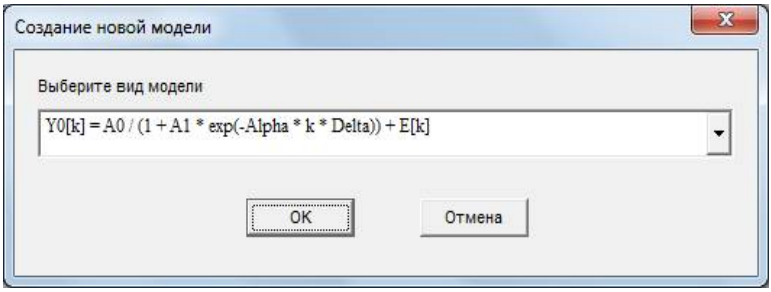


Рис. 6.5. **Окно добавления модели**

Окно программы с добавленными четырьмя логистическими моделями представлено на рис. 6.6.

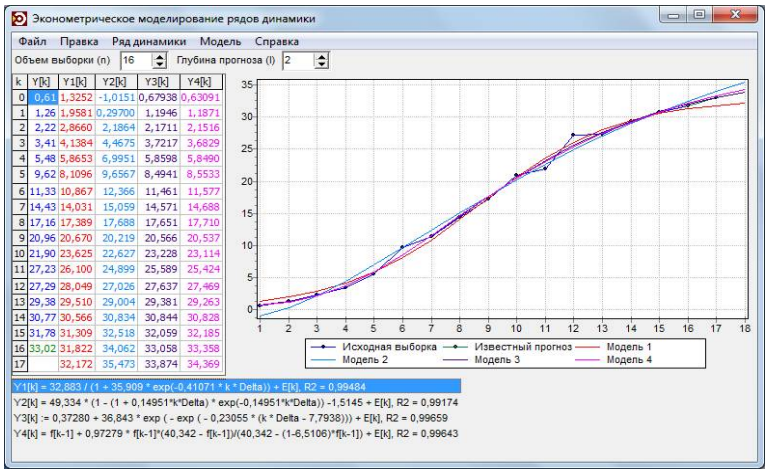


Рис. 6.6. **Окно программы с добавленными четырьмя логистическими моделями**

Для того чтобы скопировать таблицу, щелкните по ней правой кнопкой мыши и выберите пункт меню "Копировать всю таблицу". Чтобы посмотреть и скопировать полученные оценки параметров модели, сделайте двойной щелчок левой кнопкой мыши по выбранной модели из списка, расположенного в нижней части окна программы.

В открывшемся окне (рис. 6.7) выделите таблицу, содержащую оценки параметров и характеристики точности, и нажмите *Ctrl-C*.

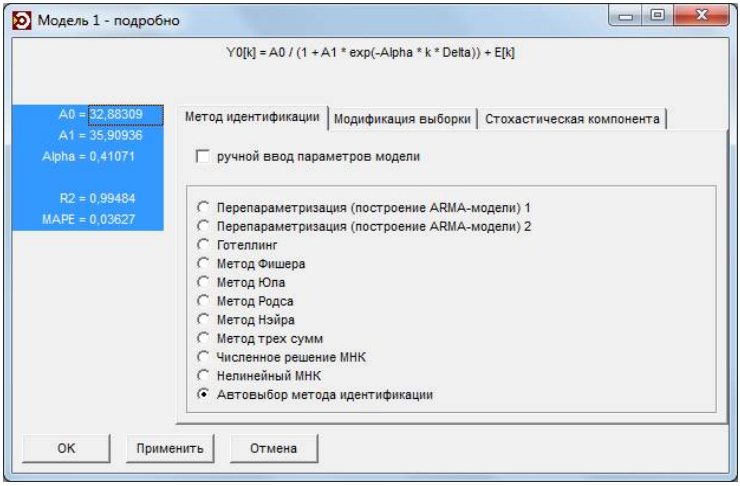


Рис. 6.7. **Окно отображения результатов идентификации для выбранной модели**

Скопируйте все полученные результаты в *MS Excel* и составьте итоговую таблицу по форме, указанной в задании.

2. Расчет значений точек перегиба

Самостоятельно рассчитайте значения оценки уровня насыщения и точки перегиба по формулам, приведенным в табл. 6.2.

Для модели *GRM*, рассчитав значение ординаты, укажите примерный диапазон для абсциссы точки перегиба, опираясь на полученные значения модельного ряда (например, для Люксембурга $Y(k*) = 11{,}36$, $k*$= 6-7).

Пример результатов моделирования для Люксембурга при длине выборки в 16 наблюдений и горизонте прогноза в 1 наблюдение представлен в табл. 6.3.

Таблица 6.3

Итоговая таблица, прогноз 1 шаг вперед, длина рабочей части выборки - 16 наблюдений

k	Y_k	Модель Ферхюльста	Модель Рамсея	Модель Гомпертца	*GRM*
1	0,6051	1,3252	0,4186	0,6785	0,6308
2	1,2620	1,9581	0,9243	1,1942	1,1870
3	2,2158	2,8660	2,2530	2,1711	2,1515
4	3,4113	4,1384	4,1664	3,7220	3,6830
5	5,4844	5,8653	6,4764	5,8603	5,8491
6	9,6226	8,1096	9,0360	8,4944	8,5535
7	11,3285	10,8670	11,7310	11,4610	11,5770
8	14,4319	14,0310	14,4750	14,5710	14,6880
9	17,1568	17,3890	17,2020	17,6510	17,7100
10	20,9606	20,6700	19,8650	20,5660	20,5370
11	21,9029	23,6250	22,4270	23,2280	23,1140
12	27,2266	26,1000	24,8660	25,5890	25,4240
13	27,2906	28,0490	27,1670	27,6370	27,4690
14	29,3774	29,5100	29,3210	29,3810	29,2630
15	30,7697	30,5660	31,3240	30,8450	30,8280
16	31,7781	31,3090	33,1770	32,0600	32,1860
17	33,0161	31,8220	34,8830	33,0600	33,3590
18		32,1720	36,4460	33,8760	34,3700
R^2		0,9948	0,9940	0,9966	0,9964
MAPE		0,62 %	8,94 %	3,25 %	4,18 %
Уровень насыщения		32,8831	50,3775	37,2201	40,3472
k^*		7,7190	6,6885	6,7940	6-7
$Y(k^*)$		16,4415	13,7304	13,9273	11,3583

Для того чтобы построить прогноз на 2 и 3 шага вперед, установите объем выборки в 15 наблюдений, глубину прогноза в 4 наблюдения и объем выборки в 14 наблюдений, глубину прогноза в 6 наблюдений, соответственно.

Аналогично составьте еще две итоговые таблицы.

В табл. 6.4 и 6.5 приведены результаты моделирования для Люксембурга при длине выборки в 15 и 14 наблюдений и горизонте прогноза в 2 и 3 наблюдения, соответственно.

Таблица 6.4

Итоговая таблица, прогноз 2 шага вперед, длина рабочей части выборки - 15 наблюдений

k	Y_k	Модель Ферхюльста	Модель Рамсея	Модель Гомпертца	*GRM*
1	0,61	1,2799	0,25789	0,65673	0,61635
2	1,26	1,9061	0,77532	1,1879	1,1741
3	2,22	2,8109	2,1335	2,179	2,1478
4	3,41	4,0876	4,0873	3,7375	3,695
5	5,48	5,83	6,4437	5,8743	5,8723
6	9,62	8,1025	9,0519	8,4997	8,5707
7	11,33	10,897	11,795	11,454	11,572
8	14,43	14,097	14,586	14,555	14,655
9	17,16	17,473	17,356	17,631	17,66
10	20,96	20,741	20,057	20,551	20,489
11	21,90	23,651	22,654	23,225	23,09
12	27,23	26,057	25,124	25,605	25,445
13	27,29	27,926	27,451	27,676	27,555
14	29,38	29,31	29,627	29,446	29,429
15	30,77	30,297	31,649	30,938	31,085
16	31,78	30,984	33,517	32,18	32,541
17	33,02	31,453	35,235	33,205	33,817
18		31,77	36,809	34,045	34,933
19		31,982	38,244	34,73	35,905
R^2		0,99424	0,99459	0,99609	0,99599
MAPE		2,14 %	7,75 %	2,48 %	4,00 %
Уровень насыщения		32,39721	50,6913	37,5429	42,23975
k^*		7,6233	6,6414	6,82808	6-7
$Y(k^*)$		16,1986	13,6526	14,0193	11,3960

Таблица 6.5

Итоговая таблица, прогноз 3 шага вперед, длина рабочей части выборки - 14 наблюдений

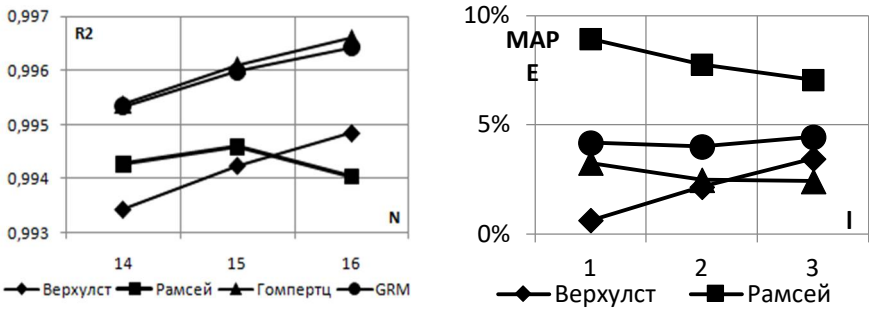
k	Y_k	Модель Ферхюльста	Модель Рамсея	Модель Гомпертца	*GRM*
1	0,61	1,2259	0,16477	0,63351	0,6038
2	1,26	1,8442	0,68821	1,1826	1,1635
3	2,22	2,7461	2,0622	2,1894	2,1463
4	3,41	4,0298	4,0387	3,7551	3,7075
5	5,48	5,7934	6,4225	5,8884	5,8922
6	9,62	8,1031	9,061	8,5023	8,581
7	11,33	10,945	11,836	11,444	11,559

Окончание табл. 6.5

k	Y_k	Модель Ферхюльста	Модель Рамсея	Модель Гомпертца	GRM
8	14,43	14,185	14,659	14,535	14,619
9	17,16	17,573	17,461	17,61	17,613
10	20,96	20,81	20,194	20,539	20,452
11	21,90	23,646	22,822	23,233	23,088
12	27,23	25,95	25,32	25,641	25,502
13	27,29	27,708	27,674	27,746	27,689
14	29,38	28,987	29,876	29,554	29,658
15	30,77	29,886	31,921	31,084	31,42
16	31,78	30,501	33,811	32,364	32,992
17	33,02	30,916	35,549	33,426	34,388
18		31,193	37,14	34,299	35,627
19		31,375	38,593	35,014	36,723
20		31,495	39,913	35,596	37,691
R^2		0,99343	0,99427	0,99537	0,99533
MAPE		3,45%	7,05%	2,42%	4,44%
Уровень насыщения		31,72196	51,1844	38,0135	44,73193
k^*		7,4948	6,6414	6,87994	6-7
$Y(k^*)$		15,8610	13,6898	14,1613	11,5143

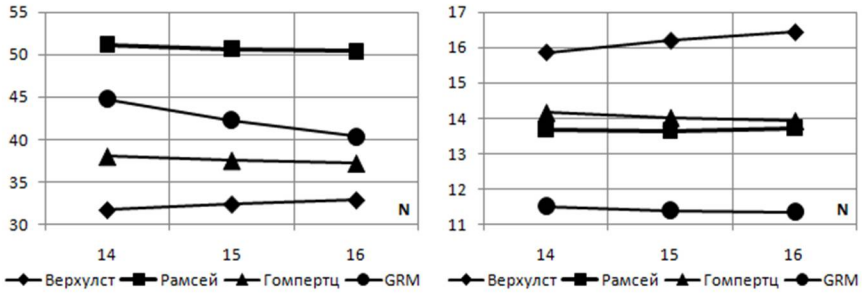
3. Выбор наилучшей модели по критериям точности моделирования и прогнозирования

На основе полученных результатов построим графики зависимости R^2 от объема выборки и *MAPE*-оценки от глубины прогноза (рис. 6.8).



а) б)

Рис. 6.8. **Зависимость (а) R^2 от объема выборки и (б) *MAPE*-оценки от глубины прогноза**



а) б)

Рис. 6.9. **Зависимость оценки уровня насыщения и точки перегиба от объема выборки**

Наилучшее качество моделирования по критерию R^2 в данном примере показали модели Гомпертца и *GRM*. По качеству прогнозирования модель Гомпертца оказалась более устойчивой при увеличении горизонта прогноза. Данная модель также дает стабильную оценку уровня насыщения (рис. 6.9).

Заметим, что модель Рамсея дала самую высокую оценку уровня насыщения, в то время как для модели Ферхюльста оценка уровня насыщения явно занижена и продолжает снижаться с уменьшением объема выборки, а оценка точки перегиба увеличивается.

Отчет по лабораторной работе должен включать:

– задание на лабораторную работу;
– графики исходного ряда и сглаженного ряда по предложенным и идентифицированным моделям (с прогнозными значениями);
– три итоговые таблицы;
– графики зависимости R^2, оценки уровня насыщения и точки перегиба от объема выборки и *MAPE*-оценки от глубины прогноза;
– выводы о выборе наилучшей модели для каждого из трех случаев.

Контрольные вопросы

1. Что такое логистическая функция? Назовите основные характеристики логистической кривой и сферы применения логистических моделей.
2. Как изменяются знаки первой и второй производной логистической функции? Где находится точка перегиба?

3. Назовите основные компоненты временного ряда.

4. Перечислите условия Гаусса - Маркова и условия оптимальности получаемых МНК-оценок идентификации.

5. Что показывает коэффициент детерминации? В каких пределах он изменяется, что характеризует?

6. Как вычисляется *MAPE*-оценка прогноза?

Обработка больших данных в *R*

Существуют три основных проблемы работы с большими данными.

1. Проблемы, требующие извлечения подмножества, выборки или сводки из источника больших данных. Может понадобиться дополнительная аналитика по подмножеству, при этом подмножество само по себе может быть довольно большим.

2. Проблемы, требующие повторного вычисления для многих подгрупп данных, например, нужно приспособить одну модель для тысяч отдельных объектов. Результаты могут быть объединены после завершения.

3. Проблемы, требующие одновременного использования всех данных. Эти проблемы необратимо велики; они должны решаться в масштабе хранилища данных.

Задачи анализа больших данных в *R*:

1. Передача подмножества данных из хранилища в *R*.
2. Обработка данных в *R* и передача результата в хранилище данных.

Рассмотрим следующий пример. Имеются статистические данные по авиалиниям: информация о прибытии и отправлении всех коммерческих рейсов в США между октябрем 1987-го и апрелем 2008 г. Всего 120 000 000 записей, т.е. около 12 ГБ данных (источник данных: http://stat-computing.org/dataexpo/2009/).

В примере требуется решить следующие задачи:

1. Извлечение случайной выборки.
2. Подгонка модели к выборке (в *R*).
3. Оценка по тестовым данным (в БД).

dplyr - пакет, предоставляющий синтаксис обработки данных для *R*. Поставляется со встроенной серверной частью *SQL* и позволяет решить следующие задачи:

1) подключение к СУБД;
2) преобразование кода *R* в SQL, отправка в СУБД для запуска;
3) сбор результатов в *R*.

Подключиться к базе данных, различающихся для разных СУБД, можно с помощью команды *src_function: src_postgres (PostgreSQL), src_sqlite (SQLite), src_mysql (MySQL, MariaDB), src_bigquery (Google BigQuery*, требуется пакет *bigrquery*). Например, для *PostgreSQL* команда выглядит следующим образом:

```
db <- src_postgres (
dbname = 'DATABASE_NAME',
host = 'HOST',
port = 5432,
user = 'USERNAME',
```

```
password = 'PASSWORD')
```

Установим предварительно пакеты *dbplyr, RPostgreSQL* и осуществим подключение к базе данных из примера.

```
library(dplyr)
# Создаем подключение к БД
air <- src_postgres(
dbname = 'airontime',
host = 'sol-eng.cjku7otn8uia.us-west-2.redshift.amazonaws.com',
port = '5439',
user = 'redshift_user',
password = 'ABCd4321')
# Выведем имена таблиц БД
src_tbls(air)
# Создадим ссылки на таблицы БД функцией tbl
flights <- tbl (air, "flights")
carriers <- tbl (air, "carriers")
# Дальше будем обращаться к ссылкам на таблицы так, как к обычным
таблицам данных
clean <- flights %>%
filter (! is.na (arrdelay), ! is.na(depdelay)) %>%
filter (depdelay > 15, depdelay < 240) %>%
filter (year >= 2002 & year <= 2007) %>%
select (year, arrdelay, depdelay, distance, uniquecarrier)
```

Команда *show_query* показывает соответствующий коду в *R* код на SQL:

```
show_query(clean)
## <SQL>
## SELECT "year" AS "year",
## "arrdelay" AS "arrdelay",
## "depdelay" AS "depdelay",
## "distance" AS "distance",
## "uniquecarrier" AS "uniquecarrier"
## FROM "flights"
## WHERE NOT ("arrdelay"IS NULL) AND NOT ("depdelay"IS NULL)
## AND "depdelay" > 15.0 AND "depdelay" < 240.0
## AND "year" >= 2002.0 AND "year" <= 2007.0O
```

Реализуем следующие запросы к данным:

```
q1 <- filter (flights, year < 2007)
q2 <- filter (q1, depdelay > 15)
q3 <- filter (q2, depdelay < 240)
q4 <- select (q3, arrdelay, depdelay, year)
q4
```

Отметим, что пакет *dplyr* не будет получать данные из БД до последнего возможного момента. Пакет совмещает все необходимые для работы запросы в один запрос.

```
show_query(q4)
```

```
## <SQL>
## SELECT "arrdelay" AS "arrdelay", "depdelay" AS "depdelay",
## "year" AS "year"
## FROM "flights"
## WHERE "year" > 2007.0 AND "depdelay" > 15.0 AND "depdelay" <
240.0
```

dplyr выведет только первые 10 строк запроса.

```
clean
## Source: postgres 8.0.2 […]
## From: flights [6,517,621 x 4]
## Filter: ! is.na(arrdelay), !is.na(depdelay), …
## year arrdelay depdelay uniquecarrier
## (int) (int) (int) (chr)
## 1 2007 42 40 9E
## 2 2007 90 94 9E
## 3 2007 19 20 9E
## 4 2007 184 167 9E
## 5 2007 21 30 9E
## 6 2007 178 179 9E
## 7 2007 56 59 9E
## 8 2007 21 21 9E
## 9 2007 50 57 9E
## 10 2007 56 23 9E
## .. ... ... ... ...
```

Команда *collect* () указывает пакету *dplyr* запросить результаты из СУБД целиком в *R*. Результат выполнения команды возвращается в формате *tbl_df*.

```
q5 <- flights %>%
filter (year > 2007, depdelay > 15) %>%
filter (depdelay == 240) %>%
collect ()
```

Команда *collapse ()* запускает принудительное выполнение запроса в СУБД.

```
q6 <- flights %>%
mutate (adjdelay = depdelay - 15) %>%
collapse () %>%
filter (adjdelay > 0)
```

Команда *collapse ()* преобразовывает предшествующие запросы в табличное выражение. Остальные запросы применяются к полученной таблице.

Выберем случайно 1 % от данных из учебной БД, поместив результат в переменную *random*:

```
random <- clean %>%
mutate (x = random ()) %>%
collapse () %>%
filter (x <= 0.01) %>%
select(-x) %>%
```

```
collect ()
```

Fit a model (in R)

Выясним, действительно ли некоторые перевозчики компенсируют потерянное время лучше других, идентифицировав модель в *R*. Рассчитаем прирост во времени (*gain*) как разницу в минутах между задержкой отправления рейса и задержкой прибытия.

```
# gain
random$gain <- random$depdelay - random$arrdelay
# построим модель
mod <- lm(gain ~ depdelay + distance + uniquecarrier,
data = random)
# таблица коэффициентов
coef(mod)
# поместим таблицу коэффициентов в data. frame
coefs <- dummy.coef(mod)
coefs_table <- data.frame(
uniquecarrier = names(coefs$uniquecarrier),
carrier_score = coefs$uniquecarrier,
int_score = coefs$`(Intercept)`,
dist_score = coefs$distance,
delay_score = coefs$depdelay,
row.names = NULL,
stringsAsFactors = FALSE
)
coefs_table
## uniquecarrier carrier_score int_score dist_score delay_score
## 1 9E 0.0000000 -1.540312 0.003083624 -0.01359926
## 2 AA -1.7131012 -1.540312 0.003083624 -0.01359926
## 3 AQ 0.6153050 -1.540312 0.003083624 -0.01359926
## 4 AS 1.4143664 -1.540312 0.003083624 -0.01359926
## 5 B6 -1.9714287 -1.540312 0.003083624 -0.01359926
## 6 CO -1.5865993 -1.540312 0.003083624 -0.01359926
## 7 DH 3.1367039 -1.540312 0.003083624 -0.01359926
## 8 DL -2.6404154 -1.540312 0.003083624 -0.01359926
## 9 EV 2.3434536 -1.540312 0.003083624 -0.01359926
## 10 F9 0.5341914 -1.540312 0.003083624 -0.01359926
## 11 FL -0.8888280 -1.540312 0.003083624 -0.01359926
## 12 HA 1.6712540 -1.540312 0.003083624 -0.01359926
## 13 HP 3.3742529 -1.540312 0.003083624 -0.01359926
## 14 MQ -1.3632398 -1.540312 0.003083624 -0.01359926
## 15 NW -2.0416490 -1.540312 0.003083624 -0.01359926
```

Построим оценки полученной модели для каждого перевозчика.

```
    score <- flights %>%
    filter (year == 2008) %>%
    filter (! is.na (arrdelay)&! is.na(depdelay)&! is.na(dis-
tance)) %>%
    filter (depdelay > 15 & depdelay < 240) %>%
```

```
    filter (arrdelay > -60 & arrdelay < 360) %>%
    select (arrdelay, depdelay, distance, uniquecarrier) %>%
    left_join(carriers, by = c('uniquecarrier' = 'code')) %>%
    left_join (coefs_table, copy = TRUE) %>%
    mutate (gain = depdelay - arrdelay) %>%
    mutate (pred = int_score + carrier_score + dist_score * dis-
tance +
    delay_score * depdelay) %>%
    group_by(description) %>%
    summarize (gain = mean (1.0 * gain), pred = mean(pred))
    scores <- collect(score)
```

Выполним визуализацию полученных оценок.

```
library(ggplot2)
ggplot(scores, aes(gain, pred)) +
geom_point (alpha = 0.75, color = 'red', shape = 3) +
geom_abline (intercept = 0, slope = 1, alpha = 0.15, color =
'blue') +
geom_text (aes (label = substr (description, 1, 20)),
size = 4, alpha = 0.75, vjust = -1) +
labs (title='Average Gains Forecast', x = 'Actual', y = 'Predict-
ed')
```

Команда *copy_to ()* создает таблицу в базе данных из локального data. frame, например:

```
copy_to(air, query5, name = "прибыль")
```

Закроем соединение с БД:

```
rm(air)
gc()
```

dplyr автоматически закрывает соединения, когда удаляется объект подключения и запускается сборщик мусора, *gc*().

Существуют и другие пакеты для работы с большими данными и соединения с БД: DBI, sparkR RHadoop, RevoScaleR, PivotalR, RODBC, RJDBC и др.

Список литературы

1. Code School - Try R [Электронный ресурс]. - Режим доступа: http://tryr.codeschool.com/ (дата обращения: 29.11.2018).
2. Coghlan, A. A Little Book of R for Time Series [Электронный ресурс] / A. Coghlan. - 2015. - Режим доступа: https://media.readthedocs.org/pdf/a-little-book-of-R-for-time-series/latest/a-little-book-of-R-for-time-series.pdf (дата обращения: 29.11.2018).
3. Giovanis, A.N. A Stochastic Logistic Innovation Diffusion Model Studying the Electricity Consumption in Greece and USA [Text] / A.N. Ciovanis, C.H. Skiadas // Technological Forecasting and Social Change. - 1999. - № 61. - P. 235-246.
4. Nakicenovic, A. Diffusion of technologies and social behavior - Springer Verlag and International Institute for Applied Systems Analysis [Text] / N. Nakicenovic, A. Grubler. - Berlin ; New York, 1991. - 605 p.
5. Анализ данных в *R* [Электронный ресурс]. - Режим доступа: https://stepik.org/course/Анализ-данных-в-R-129 (дата обращения: 29.11.2018).
6. Зорин, А.В. Введение в прикладной статистический анализ в пакете *R* [Текст] : учеб.-метод. пособие /А.В. Зорин, М.А. Федоткин. - Нижний Новгород : ННГУ, 2010. - 50 с.
7. Мастицкий, С.Э. Статистический анализ и визуализации данных с помощью *R* [Текст] / С.Э. Мастицкий, В.К. Шитиков. - Москва : ДМК Пресс, 2015. - 496 с.
8. Осовский, С. Нейронные сети для обработки информации [Текст] / пер. с польск. И.Д. Рудинского / С. Осовский. - Москва : Финансы и статистика, 2002. - 344 с.
9. Семенычев, В.К. Идентификация экономической динамики на основе моделей авторегрессии [Текст] / В.К. Семенычев. - Самара: Изд-во СамНЦ РАН, 2004. - 243 с.
10. Семенычев, В.К. Анализ и предложения моделей экономической динамики с кумулятивным логистическим трендом [Текст] : монография / В.К. Семенычев, В.Н. Кожухова. - Самара : Изд-во СамНЦ РАН, 2013. - 152 с.
11. Семенычев, В.К. Предложения эконометрического инструментария моделирования и прогнозирования эволюционных процессов [Текст] / В.К. Семенычев, А.А. Коробецкая, В.Н. Кожухова. - Самара: Изд-во САГМУ, 2015. - 384 с.
12. Шитиков, В.К. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. [Текст] / В.К. Шитиков, Г.С. Розенберг. - Тольятти : Кассандра, 2013. - 314 с.
13. Эконометрика [Электронный ресурс] // Coursera. - Режим доступа: https://www.coursera.org/learn/ekonometrika (дата обращения: 29.11.2018).

Приложение

Индивидуальные варианты заданий для лабораторной работы № 1

Варианты		**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
Бином. распр.	n	70	70	55	40	60	40	80	50	30	70	20	40	90
	p	0,4	0,5	0,8	0,8	0,5	0,1	0,5	0,2	0,3	0,3	0,7	0,3	0,7
Распр. Пуассона	λ	17	11	10	16	14	13	5	7	12	8	15	6	9
Норм. распр.	M	430	135	200	125	220	370	270	185	125	150	105	180	320
	σ	42	15	30	19	22	37	27	15	20	26	15	30	32
Равном. распр.	a	1	6	0	0	13	10	0	1	20	6	15	20	4
	b	4	10	1	9	15	20	2	4	50	9	20	30	8
Хи-квадрат	k	18	12	11	17	15	14	6	8	13	9	16	7	10
Фишера	k_1	7	50	100	30	10	50	4	100	8	5	20	100	10
	k_2	18	8	7	15	17	16	10	100	9	13	14	11	6
Стьюдента	k	46	23	20	50	40	35	5	11	29	15	39	6	19

Индивидуальные варианты заданий для лабораторных работ № 4

Варианты	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Среднее x	7	30	100	20	50	10	25	40	60	33
Ст. отклон. x	2	3	15	4	6	1	4	7	10	8
Объем выборки	50	100	150	70	50	30	60	90	110	80
Формула y	$y=a+b(x-c)^2+\varepsilon$			$y=a+\frac{b}{x-c}+\varepsilon$		$y=a+e^{bx+c}+\varepsilon$		$y=\frac{a}{1+e^{-b(x-c)}}+\varepsilon$		
a	10	100	–5	10	150	20	1000	50	80	20
b	–2	5	–0,1	3	–10	0,8	–0,25	0,25	0,3	0,22
c	5	32	90	10	30	0,5	5	35	52	30
Ст. отклон. помехи	4	28	14	0,05	0,1	3000	0,25	3	6	1,5

Индивидуальные варианты заданий для лабораторной работы № 5

Варианты	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Компания	Yandex	Twitter	Facebook	ВКонтакте	Amazon	EBay	Microsoft	Сбербанк	Вымпелком	МТС
Тикер	*YNDX*	*TWTR*	*FB*	*VK*	*AMZN*	*EBAY*	*MSFT*	*SBNC*	*VIP*	МВТ
Глубина прогноза	10	15	20	25	30	25	20	15	10	20

Индивидуальные варианты заданий для лабораторной работы № 6

Дата	k	Пример	№ варианта									
			1	2	3	4	5	6	7	8	9	10
		Люксем-бург	Нидер-ланды	Норве-гия	Австра-лия	Шве-ция	Ирлан-дия	США	Фран-ция	Великобрита-ния	Ита-лия	Япо-ния
2002-Q2	1	0,61	4,93	2,85	1,26	6,53	0,05	5,53	1,57	1,27	1,19	4,00
2002-Q4	2	1,26	7,03	4,01	1,78	7,97	0,14	6,71	2,76	2,32	1,68	6,22
2003-Q2	3	2,22	9,07	5,97	2,52	9,28	0,26	7,94	3,97	3,67	2,80	8,79
2003-Q4	4	3,41	11,79	7,77	3,43	11,03	0,76	9,57	5,94	5,39	4,13	10,90
2004-Q2	5	5,48	15,43	11,05	5,12	12,32	1,50	10,91	7,87	7,36	6,02	13,16
2004-Q4	6	9,62	18,95	14,66	7,60	14,79	3,05	12,76	10,45	10,36	8,08	14,96
2005-Q2	7	11,33	22,32	17,65	10,58	17,45	3,76	14,22	12,62	13,25	9,70	16,71
2005-Q4	8	14,43	25,21	21,43	13,23	20,64	5,56	16,32	15,06	16,32	11,77	18,15
2006-Q2	9	17,16	28,29	25,12	16,31	23,38	7,47	18,16	17,56	19,18	13,06	19,18
2006-Q4	10	20,96	31,00	26,24	17,39	26,15	10,27	20,27	20,11	21,45	14,24	20,69
2007-Q2	11	21,90	33,47	29,60	21,32	28,62	12,49	21,86	22,40	23,73	15,81	21,26
2007-Q4	12	27,23	34,15	30,39	22,10	30,40	14,60	23,37	24,61	25,78	17,21	22,51
2008-Q2	13	27,29	35,12	31,89	22,42	30,40	15,95	23,89	26,04	27,24	17,91	22,98
2008-Q4	14	29,38	35,61	33,05	22,67	31,35	17,33	25,48	27,64	28,16	18,84	23,58
2009-Q2	15	30,77	37,06	33,39	23,14	31,10	18,30	26,47	28,96	28,73	19,71	24,23
2009-Q4	16	31,78	37,09	33,76	23,16	31,40	19,48	26,39	30,36	29,49	20,33	24,79
2010-Q2	17	33,02	37,79	34,18	23,44	31,66	20,32	27,12	31,41	30,48	21,32	25,28

Учебное издание

Семенычев Валерий Константинович
Кожухова Варвара Николаевна
Коробецкая Анастасия Александровна

ТЕХНОЛОГИИ И ИНСТРУМЕНТАРИЙ АНАЛИЗА БОЛЬШИХ ДАННЫХ

Практикум

Руководитель издательской группы О.В. Егорова
Редактор Т.В. Федулова
Корректор Л.И. Трофимова
Компьютерная верстка - М.А. Куприянова

Подписано к изданию 10.12.2019. Печ. л. 8,19.
ФГБОУ ВО "Самарский государственный экономический университет".
443090, Самара, ул. Советской Армии, 141.