С. Ф. Каморников, С.С. Каморников

ЭКОНОМЕТРИКА
(учебное пособие)

Гомель, 2012

Рецензенты:
заведующий кафедрой экономических теорий УО «ГГУ им. Ф.Скорины», доктор экономических наук, профессор *Б.В. Сорвиров*; заведующий кафедрой экономической кибернетики и теории вероятностей УО «ГГУ им. Ф.Скорины», доктор физико-математических наук, профессор *Ю.В. Малинковский*

К185 Каморников С. Ф. , Каморников С. С. Эконометрика : учеб. пособие. – М. : Интеграция, 2012. – 262 с.

Пособие представляет собой начальный курс эконометрики, включающий базовые методы построения и анализа регрессионных моделей. Главное внимание уделяется основам моделирования, парным и множественным моделям линейной и нелинейной регрессии, анализу временных рядов и систем одновременных уравнений. Теоретический материал сопровождается большим числом решённых задач. Приведены индивидуальные задания, контрольные вопросы и тесты. Продемонстрированы возможности табличного процессора *MS Excel* для решения задач эконометрического моделирования и прогнозирования.

Пособие рекомендуется при изучении дисциплины «Эконометрика и экономико-математические методы и модели» для студентов экономических специальностей всех форм обучения, а также лиц, проводящих самостоятельные эконометрические расчёты. Может быть использовано преподавателями для методической поддержки практических и лабораторных занятий по эконометрике.

ОГЛАВЛЕНИЕ

Глава 4. Модели множественной регрессии

Глава 5. Эконометрический анализ классических модельных предположений

Глава 6. Моделирование временных рядов

Глава 7. Системы эконометрических уравнений

Введение

Во всех программах подготовки экономических кадров учебный курс «Эконометрика и экономико-математические методы и модели» является одной из базовых дисциплин. Объясняется это тем, что современные требования к уровню экономического образования предполагают наличие у выпускников широких математических и статистических знаний, необходимых при решении задач, связанных с моделированием и количественным анализом экономических явлений и процессов.

Настоящее пособие ориентировано на студентов экономических специальностей, **впервые приступающих к изучению эконометрики.** Поэтому при изложении материала делается упор не на математическую полноту и строгость утверждений и доказательств, а на понимание постановок задач и методов их решения. Мы полагаем, что если студент хорошо понял идею метода, то при необходимости он самостоятельно сможет более детально и глубоко изучить его, воспользовавшись дополнительной литературой.

Доказательства же всех приведенных в пособии теоретических положений читатель может найти в источниках, указанных в списках литературы, которые даны в каждой главе. Каждая глава учебника – это только введение в большую область эконометрики, и если появится желание выйти за пределы пособия, то приведенные литературные списки помогут это желание удовлетворить. Этому будут способствовать также небольшие тематические обзоры вопросов и проблем современной эконометрики, предложенные в конце каждой главы.

Обучение эконометрике опирается на знание курсов высшей математики, информатики, экономической теории, теории вероятностей и математической статистики, а также общей теории статистики. При изложении учебного материала предполагается, что читатель владеет основами указанных курсов в объеме, необходимом для экономических специальностей.

Уяснение сущности эконометрического моделирования невозможно без рассмотрения общих теоретических положений моделирования в целом и экономико-математического моделирования в частности. Поэтому в пособие включен соответствующий раздел 1 «Математическое моделирование в экономике», состоящий из двух глав.

Второй раздел «Эконометрические модели» является основным в пособии. Каждая из пяти глав этого раздела содержит основные теоретические выводы и методики решения задач. К сожалению, небольшой объем времени, отводимый учебным планом на изучение эконометрики, предполагает рассмотрение только ключевых вопросов. К таковым мы относим вопросы, связанные с:

- построением парных и множественных линейных и нелинейных регрессионных моделей;
- рассмотрением основных ошибок, возникающих при нарушении классических модельных предположений, методикой их диагностики и устранения;
- изучением эконометрических моделей, выраженных системой одновременных уравнений;
- анализом структуры временных рядов.

Эконометрика в значительной степени обеспечивает формирование необходимых умений моделирования и прогнозирования. Такая подготовка не может обойтись без отработки практических навыков построения и анализа эконометрических моделей. С этой целью в пособие включен комплекс задач, описывающих ситуации экономической реальности (в примерах выполнения типовых задач и контрольных заданиях для самостоятельной работы). Кроме того, в пособие включены интегрированные задачи, которые могут быть использованы преподавателями для организации лабораторных занятий. Отметим, что некоторые задачи составлены с использованием уже опубликованных материалов, приведенных в списке литературы.

Контроль над усвоением теоретических положений может быть осуществлен с помощью системы вопросов для самоконтроля, сопровождающих каждую главу пособия, и тестовых заданий, помещенных в конце каждого раздела.

В пособие включены компьютерные задания по базовым темам. При этом сами задания предусматривают не только оценку параметров модели, но и содержательную интерпретацию результатов. Мы исходим из того, что при эконометрическом анализе важны не только сами расчеты и обоснования показателей, но, прежде всего, те пояснения и выводы, которые показывают, что и как эти показатели характеризуют. Студент должен уметь не только строить корреляционные и регрессионные таблицы, но и интерпретировать их. Для этого интегрированные задания в пособии дополнены специальными отчетами, призванными развивать соответствующие аналитические качества.

Компьютерные задания предусматривают выполнение задач в среде табличного процессора Microsoft Excel, являющегося тем программным продуктом, с которым современный экономист проводит основную массу своих расчетов. Для описания порядка выполнения компьютерных заданий в пособии используется версия Excel 2003 (в последних версиях Excel порядок применения инструментов, необходимых для решения рассматриваемых нами задач, отличается незначительно).

Учебное пособие прошло апробацию в Гомельском филиале Международного университета «МИТСО» и отражает определенный опыт преподавания эконометрики в этом учебном заведении.

РАЗДЕЛ 1
Математическое моделирование в экономике

«Экономико-математическая модель представляет собой концентрированное выражение общих взаимосвязей и закономерностей экономического явления в математической форме».

Академик Немчинов В.С.

Глава 1
Теоретические основы экономико-математического моделирования

Основные понятия: *модель, моделирование, знаковая модель, математическая модель, экономическая модель, экономико-математическая модель, экономико-математическое моделирование, оптимизационная модель, балансовая модель, модели микроуровня и макроуровня, статические и динамические модели, детерминированные и вероятностные модели*.

Литература: [1], [6], [17].

1.1. Понятие о модели и моделировании

Одним из основных методов исследования экономических явлений и процессов является основанный на принципе аналогии метод моделирования.

Необходимость использования этого метода в целом определяется тем, что многие объекты (или проблемы, относящиеся к этим объектам) непосредственно исследовать или вовсе невозможно, или же это исследование требует слишком высоких затрат времени и средств.

Такая ситуация проявляется и в экономической практике, где результаты деятельности во многом зависят от качества принимаемых решений. Конечно, в привычных, повторяющихся условиях управляющие решения могут приниматься на основе накопленного опыта, интуиции или здравого смысла. Но наилучшие решения не всегда лежат на поверхности, аналогов в прошлом опыте может и не быть, а цена ошибки при современных масштабах производства очень велика. Самым надежным способом выбора хорошего решения была бы постановка экспериментов непосредственно на объекте. Однако такие эксперименты зачастую невозможны или затруднительны ввиду их дороговизны, длительных сроков проведения, опасности нежелательных последствий.

В тех случаях, когда нельзя провести управляемый эксперимент, остается одна возможность – построить модель рассматриваемой ситуации и провести необходимые эксперименты с ней.

Модель (от латинского слова *modelus* – мера, образец) – это искусственно созданный объект произвольной природы, который строится с целью получения новых знаний об объекте-оригинале и отражает существенные для рассматриваемой задачи свойства оригинала.

Из определения следует, что модель должна воспроизводить оригинал в его основных чертах так, чтобы ее изучение давало новую информацию об объекте. При этом изучение одних сторон моделируемого объекта может осуществляться ценой отказа от отражения других. Поэтому любая модель замещает оригинал лишь в строго ограниченном смысле. Для получения более полной информации об изучаемом объекте может быть построено несколько моделей, характеризующих его с разных сторон.

Моделирование – это процесс построения, изучения и применения моделей.

Главная особенность моделирования состоит в том, что модель выступает как инструмент, который исследователь ставит между собой и объектом с целью изучения последнего. Таким образом, процесс моделирования включает в себя три элемента: субъект исследования (исследователь), объект исследования (оригинал), модель. Ситуацию иллюстрирует рисунок 1.1.

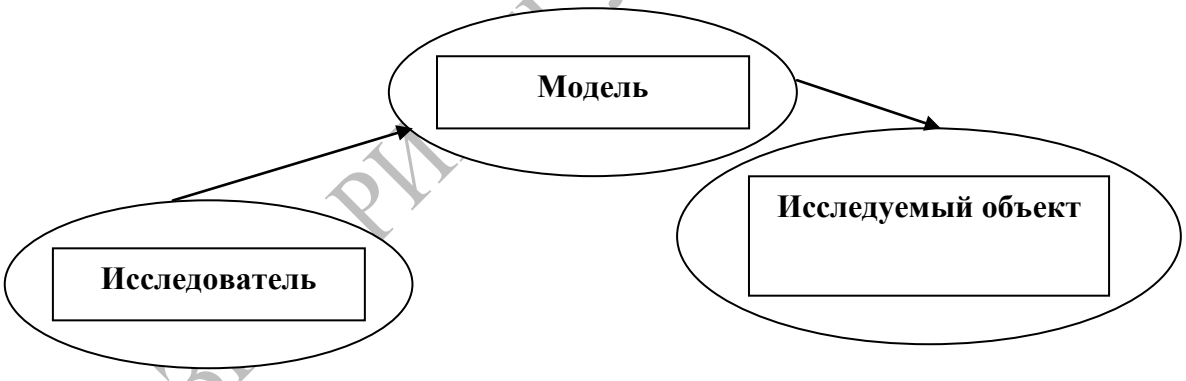


Рис. 1.1. Роль модели в процессе исследования

Сущность процесса моделирования поясняет схема, представленная на рисунке 1.2.

Пусть необходимо изучить некоторый объект **А**. Первоначально исследователь с помощью некоторых средств конструирует модель **В** – условный образ объекта **А**, приближенно воссоздающий объект **А**. При этом важнейшим является вопрос о степени сходства оригинала и модели. Этот вопрос требует детального анализа и решения в зависимости от конкретной ситуации. Очевидно, что модель утрачивает свой смысл как в случае

тождества с оригиналом (тогда она перестает быть моделью), так и в случае чрезмерного отличия от оригинала.

После построения модель **В** исследуется. Конечным результатом такого исследования является получение совокупности знаний о модели **В.**

Далее осуществляется интерпретация полученных знаний о модели, т.е. перенос их с модели на оригинал и формирование множества знаний об объекте **А** на основе знаний, полученных по модели **В.**

Наконец, заканчивается моделирование практической проверкой полученных знаний, их использованием для управления объектом **А.**

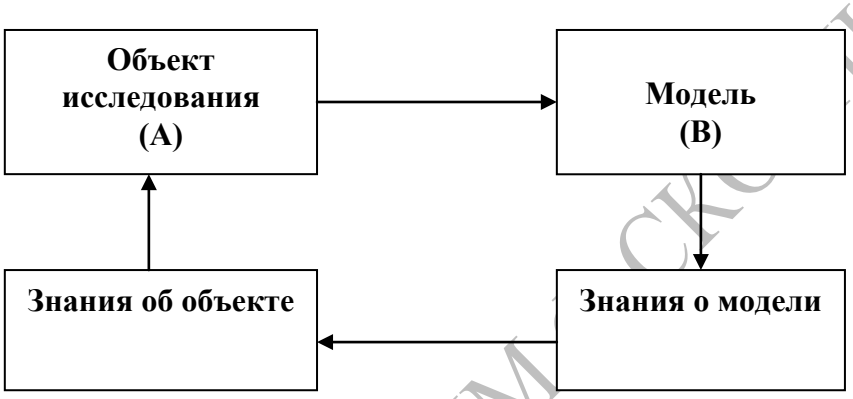


Рис. 1.2. Сущность процесса моделирования

Отметим, что метод моделирования широко использует не только экономическая практика, но и экономическая наука. Значительный опыт построения моделей накоплен учеными, применявшими их для анализа экономических процессов и явлений. В частности, с использованием моделирования связаны практически все работы, удостоенные Нобелевской премии по экономике (Д. Хикс, Р. Солоу, В.В. Леонтьев, Л.В. Канторович, П. Самуэльсон и др.).

1.2. Классификация моделей

В зависимости от выбранных средств построения можно выделить три основных типа моделей: концептуальные, материальные и знаковые (рисунок 1.3):

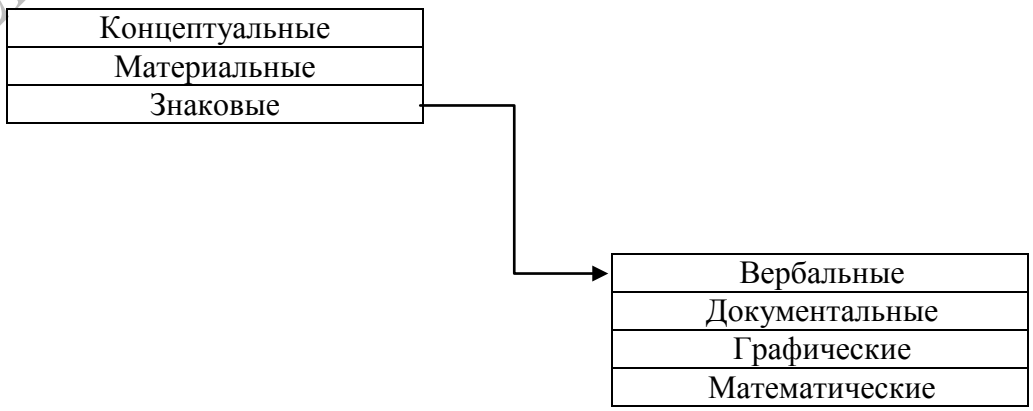


Рис. 1.3. Классификация моделей по способу представления

Концептуальная модель – это некоторое идеальное представление объекта, которое формируется в сознании человека в виде некоторого мыслимого образа.

Детализация концептуальной модели, приведение ее к виду, позволяющему экспериментировать с моделью для получения информации об объекте, может осуществляться в двух формах: материальной и знаковой.

Материальные модели существуют объективно и реализуются с помощью некоторых материальных средств (их можно "потрогать руками").

Материальные модели в экономике практически не используются. Редкие попытки имели лишь демонстрационное применение, а не служили средством изучения закономерностей экономики. Примером материальной модели, применявшейся в экономике, является *гидравлическая модель*. В ней потоки воды имитировали потоки денег и товаров, а резервуары отождествлялись с такими экономическими категориями как объем промышленного производства, объем личного потребления и т.д.

В отличие от многих естественных и особенно технических наук в экономике преобладает знаковое моделирование.

Знаковые модели отображают свойства оригинала либо на естественном языке, либо с помощью различных символов. Условно они разделяются на *вербальные* (выраженные в разговорной форме, например, на русском языке), *документальные* (выраженные с помощью некоторого текста), *графические* и *математические*.

Примерами вербальных и документальных моделей являются словесные или текстовые описания экономических процессов или явлений, технико-экономические задания, пояснительные записки к проектам и т.д.

Графические модели служат для отображения свойств объекта в виде некоторого графического образа. Это могут быть статистические таблицы, графики некоторых функциональных зависимостей между экономическими показателями, диаграммы динамики процессов, гистограммы, схемы и т.д.

Графические модели, как наиболее удобный инструмент научного экономического анализа, были введены автором теории рыночного ценообразования Альфредом Маршаллом, который в 1890 году первым графически изобразил задачу нахождения равновесия между спросом и предложением. Позже в экономической теории и практике (как на микроуровне, так и на макроуровне) графические модели получили широкое распространение.

Математическая модель отображает свойства объекта на языке математики, т.е. с помощью математических символов и соотношений: функциональных и логических зависимостей, неравенств, систем алгебраических, дифференциальных или других уравнений и т.п.

Основное преимущество математической модели перед другими видами знаковых моделей заключается в том, что она позволяет провести детальный количественный анализ исследуемого объекта, помогает предсказать, как поведет себя этот объект в различных условиях, и дает рекомендации для выбора наилучших вариантов решения проблемы. Тем самым исследователь избавляется от необходимости проведения реальных экспериментов.

Математические модели использовались в иллюстративных и исследовательских целях еще в XVIII веке (Ф. Кенэ, экономическая таблица; А. Смит, классическая макроэкономическая модель; Д. Рикардо, модель международной торговли). В настоящее время решение практически любой экономической задачи в той или иной мере связано с математическим моделированием.

Кроме представленной классификации моделей, различают и многие другие:

по отрасли знаний (экономические, социальные, биологические и др.);

по фактору времени (статические и динамические);

по принадлежности к иерархическому уровню (модели микроуровня и модели макроуровня);

по степени причинной обусловленности (детерминированные и вероятностные).

Под *экономической моделью* понимается упрощенное формальное описание некоторого экономического явления. Примерами экономических моделей являются модель потребительского выбора, модель фирмы, модель экономического роста, модель равновесия на товарном рынке.

Обычно экономическая модель строится в вербальной или документальной форме по следующему алгоритму: 1) формулируется предмет и цель исследования; 2) выделяются структурные или функциональные элементы экономической системы, выявляются наиболее важные качественные характеристики этих элементов; 3) качественно описываются взаимосвязи между элементами.

В *статических моделях* описывается состояние объекта в конкретный момент или период времени. *Динамические модели* описывают взаимосвязи переменных во времени.

Детерминированные модели предполагают жесткие функциональные связи между переменными модели. Вероятностные (или *стохастические*) модели допускают наличие случайных воздействий на исследуемые показатели.

1.3. Принципы моделирования

Моделирование объектов может проводиться различными способами и приводить в итоге к разным моделям. Главное, чтобы при построении моделей учитывались следующие общие принципы:

1. Модель должна реализовываться простыми средствами, легко оцениваться и проверяться. Результаты, полученные из модели, должны быть

ясны как её создателю, так и лицу, принимающему решение (*принцип простоты*).

2. Модель должна достаточно широко охватывать моделируемый объект, учитывать существенные черты оригинала и при этом не должна его сильно упрощать (*принцип полноты*).

3. Модель должна иметь поведение, структуру и функции, подобные таковым у моделируемого объекта (*принцип подобия*).

4. Отклонения параметров модели от соответствующих параметров моделируемого объекта не должно выходить за рамки допустимой точности (*принцип точности*).

5. На основании исследования модели должно быть возможным обнаружение новых свойств у моделируемого объекта (*принцип новизны*).

6. Проведение исследований и экспериментов на модели должно быть более удобным по сравнению с моделируемым объектом (*принцип удобности*).

1.4. Экономико-математическая модель

Под *экономико-математической моделью* понимается экономическая модель, в которой структурные элементы и их взаимосвязи выражены в математической форме (т.е. в виде математической модели). Естественно, что под *экономико-математическим моделированием* понимают процесс построения, изучения и применения математических моделей в экономике.

Среди экономико-математических моделей особое место занимают эконометрические, балансовые и оптимизационные модели.

Эконометрические модели – это модели, в которых описываются корреляционно-регрессионные зависимости. Эти модели широко используются для построения производственных функций и прогнозирования экономических явлений.

Балансовые модели представляют собой систему балансов производства и распределения. Они опираются на аппарат матричной алгебры и применяются при планировании деятельности различных отраслей экономики.

Оптимизационные модели представляют собой систему математических уравнений, подчиненных определенной целевой функции и служащих для отыскания оптимальных решений конкретной экономической задачи. Эти модели применяются для описания условий функционирования экономических систем.

Экономико-математические модели классифицируются и по другим признакам. В частности, как и для любых моделей, для них характерно разделение по *фактору времени, принадлежности к иерархическому уровню* и *степени причинной обусловленности*.

Кроме того, экономико-математические модели различают по *целевому назначению* (теоретические и прикладные), по *уровням исследуемых экономических процессов* (производственно-технологические и социально-

экономические), по *форме математических зависимостей* (линейные и нелинейные).

1.5. Этапы экономико-математического моделирования

Процесс математического моделирования можно условно разделить на три этапа:

1. *Этап формализации*, т.е. перевод рассматриваемой задачи с естественного языка на язык математических терминов и обозначений. При этом осуществляется переход от реальной ситуации к математической модели. Результатом этого этапа является построенная математическая модель.

Формализация любой экономико-математической модели начинается с описания переменных и параметров. При этом различают *экзогенные* и *эндогенные* переменные. Первые из них задаются вне модели, а вторые определяются в ходе расчетов по модели.

2. *Этап исследования построенной математической модели*, т.е. выбор наиболее подходящего метода решения и решение поставленной математической задачи.

3. *Этап интерпретации*, т.е. анализ полученных математических результатов и объяснение их в терминах исходной экономической задачи.

Отметим, что выделяют и так называемый *подготовительный* (или *постановочный*) этап. На этом этапе осуществляется сбор и анализ информации по моделируемому объекту, формируются цели и задачи, выбираются средства решения.

Контрольные вопросы

1. Сформулируйте понятия «модель» и «моделирование».
2. С какими целями осуществляется моделирование в экономике?
3. В чем заключается сущность метода моделирования?
4. По каким признакам классифицируются модели?
5. Как классифицируются модели по способу представления?
6. К какому классу моделей относятся математические модели?
7. Как определяется математическая модель?
8. Приведите примеры применения знаковых моделей в экономике.
9. Что понимается под экономико-математической моделью?
10. Охарактеризуйте эконометрические, балансовые и оптимизационные экономико-математические модели.
11. Как классифицируются модели по фактору времени и степени причинной обусловленности?
12. В чем отличие статических моделей от динамических?
13. В чем отличие детерминированных моделей от вероятностных?
14. Сформулируйте основные принципы моделирования.
15. В чем заключаются принципы простоты и полноты?
16. Назовите этапы экономико-математического моделирования.

17. Какие задачи решаются на этапе формализации модели?
18. Какие задачи решаются на этапе интерпретации модели?

Глава 2
Теоретические основы эконометрики

Основные понятия: *эконометрика, эконометрическая модель, спецификация модели, параметризация модели, верификация модели, функциональные и статистические зависимости, корреляционная зависимость, уравнение регрессии, регрессионная модель.*

Литература: [2-4], [9-11], [15].

2.1. Эконометрика как наука

Основные результаты экономической теории носят качественный характер. Например, экономическая теория, формулируя закон спроса, утверждает, что снижение цены товара (при неизменности всех прочих факторов) приводит к увеличению спроса на данный товар. Однако экономическая теория не указывает количественных оценок такого увеличения, т.е. не отвечает на вопрос: насколько увеличится спрос при уменьшении цены товара на определенную величину. Расчет количественных характеристик и есть главная задача эконометрики.

Эконометрические методы позволяют построить модели взаимосвязей в экономике, количественно оценить зависимости, отражающие эти взаимосвязи, и использовать полученные оценки либо для прогнозирования, либо для объяснения внутренних механизмов исследуемых экономических явлений.

Термин «эконометрика» введен в 1930 году лауреатом Нобелевской премии по экономике норвежским ученым Рагнаром Фришем и отражает содержание эконометрики как науки. Этот термин представляет собой комбинацию терминов «экономика» и «метрика» и означает «измерения в экономике».

Эконометрика – это наука, в которой на базе экономической теории и реальных статистических данных строятся математические модели массовых экономических явлений с целью количественного подтверждения или опровержения определенных экономических гипотез и прогнозирования соответствующих экономических показателей.

Таким образом, объектом исследования эконометрики являются реальные экономические процессы, а предметом ее исследования – количественные характеристики взаимосвязей в экономике.

2.2. Эконометрика и другие науки

Эконометрика занимает свою нишу научных и прикладных исследований на стыке экономической теории, экономической статистики, математического моделирования, теории вероятностей и математической статистики. Связь эконометрики с отмеченными научными дисциплинами проявляется по-разному:

1) экономическая теория с помощью качественного анализа устанавливает совокупность факторов и показателей, влияющих на изучаемое экономическое явление, их роль и теоретические взаимосвязи;

2) экономическая статистика обеспечивает информационную основу экономических исследований, представляя эмпирические данные выбранных экономических показателей в виде таблиц, диаграмм, графиков и обеспечивая их первичную обработку; при этом различают *пространственные* данные (взятые по разным объектам за один и тот же период или момент времени) и *временные* данные (рассматриваемые для одного экономического объекта в последовательные моменты времени);

3) математическое моделирование формализует экономическую задачу на языке математики;

4) для исследования построенных экономико-математических моделей в эконометрике используется специфический математический аппарат, опирающийся на теорию вероятностей и математическую статистику.

Исследование эконометрической модели даже с небольшим числом факторов в «ручном» исполнении весьма трудоемко. В последнее время на помощь исследователям пришли компьютерные технологии, позволяющие значительно упростить расчет регрессионных моделей. Для этих целей используются специализированные пакеты прикладных программ, среди которых в среде экономистов наибольшее распространение получил табличный процессор Excel. Таким образом, кроме указанных выше научных дисциплин, эконометрика тесно связана с компьютерными технологиями, используя их для автоматизации вычислений.

Предлагаемая ниже схема (рисунок 2.1) наглядно отражает существующие связи эконометрики с другими дисциплинами и технологиями.

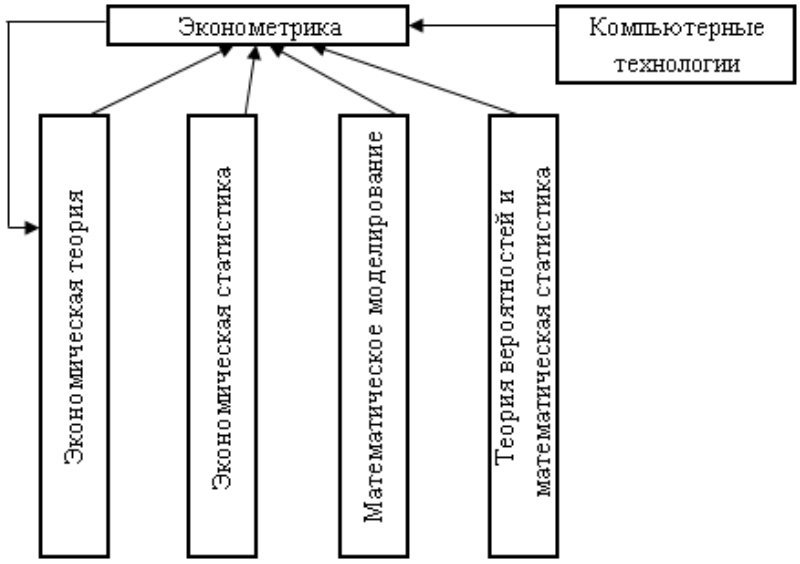


Рис. 2.1. Схема связи эконометрики с другими дисциплинами

Опираясь на экономическую теорию, моделирование, теорию вероятностей и статистику, используя вычислительные возможности компьютерных технологий, эконометрика количественно обосновывает экономические гипотезы, углубляет и совершенствует тем самым экономическую теорию и дает рекомендации для проведения продуманной и целенаправленной экономической политики.

2.3. Эконометрические модели и их типы

Одной из основных задач эконометрики является построение и анализ эконометрической модели. При этом под *эконометрической моделью* понимается такая форма представления исследуемой экономической задачи с помощью математических терминов и соотношений, которая удобна для проведения количественного анализа на основе имеющихся статистических данных.

Существуют различные классификации эконометрических моделей. Различают:

1) *эконометрические модели микроуровня* (уровень отдельного предприятия);

2) *эконометрические модели мезоуровня* (уровень отрасли или региона);

3) *эконометрические модели макроуровня* (уровень национальной или мировой экономики).

По фактору времени различают *статические* и *динамические эконометрические модели*. Первые из них исследуют состояние экономической системы в определенный момент времени. Вторые (их еще называют *моделями временных рядов*) строятся по данным, характеризующим изучаемые объекты за ряд последовательных периодов времени.

В зависимости от формы математического представления эконометрические модели подразделяют на *модели с одним уравнением* и *модели системы одновременных уравнений*. В первом случае объясняемый фактор *y* выражается через объясняющие переменные $x_1, x_2, ..., x_m$ с помощью одного уравнения. При этом модель называется *парной*, если уравнение связывает только две переменные x, y. Если же речь идет о зависимости величины *y* от нескольких факторов $x_1, x_2, ..., x_m$ $(m \geq 2)$, то модель с одним уравнением называется *множественной моделью*.

В зависимости от вида функции, выражающей *объясняемый* (*результативный*) *фактор y* через *объясняющие переменные* $x_1, x_2, ..., x_m$, эконометрические модели разделяются на *линейные* и *нелинейные*.

При изучении достаточно сложных экономических явлений взаимосвязи между исследуемыми показателями описываются, как правило, не одним, а несколькими уравнениями. Это и приводит к необходимости использования *модели системы одновременных уравнений*. Наиболее известный пример такой модели дает модель равновесия спроса и предложения в рыночной экономике.

2.4. Этапы эконометрического моделирования

Выделяют следующие этапы решения эконометрической задачи.

1. *Постановочный этап*. Он предполагает:

а) определение целей и задач исследования;

б) выделение факторов и показателей, определяющих изучаемые экономические процессы;

в) установление на базе экономической теории роли выбранных показателей.

2. *Этап спецификации*. На этом этапе осуществляется выбор формулы связи между переменными, обозначающими выделенные факторы. Как правило, эта формула имеет весьма общий вид и содержит группу параметров, требующих статистической оценки.

3. *Этап параметризации*. На этом этапе решается задача оценки значений параметров выбранной функции связи, т.е. задача подбора коэффициентов функции таким образом, чтобы эта функция в некотором смысле наилучшим образом отражала зависимость между объясняемым фактором и независимыми переменными.

4. *Этап верификации*. Он предполагает проверку адекватности модели, т.е. проверку того, в какой степени построенная модель соответствует реальному экономическому явлению или процессу. Кроме того, здесь выясняется, насколько удачно решены проблемы спецификации и параметризации, совершенствуется форма модели, уточняется состав объясняющих переменных, устанавливается точность расчетов по данной модели, общее качество уравнения, статистическая значимость найденных параметров.

Как видим, методика эконометрического моделирования вполне согласуется с общей схемой построения и анализа математической модели, которая предполагает прохождение этапов формализации, исследования и интерпретации модели. Отличие проявляется лишь в специфичной для эконометрики терминологии.

Предложенная периодизация решения эконометрической задачи весьма условна. Это связано, с одной стороны, с тем, что сами этапы эконометрического моделирования могут пересекаться и взаимно дополнять друг друга, а с другой стороны, с тем, что связь между этапами имеет возвратный характер: полное эконометрическое исследование порой предполагает прохождение нескольких циклов до тех пор, пока модель не будет признана качественной. Сущность эконометрических исследований, их последовательность и цикличность наглядно демонстрирует схема на рисунке 2.2.

Указанная схема подчеркивает одну из главных особенностей эконометрической модели, заключающуюся в том, что модель органично связана со статистическими данными, на основе которых построена. Эта связь существенно проявляется на всех этапах решения: при выборе общего типа модели, исходя из вида корреляционного поля, построенного по эмпирическим данным; при нахождении параметров модели, опираясь на метод наименьших квадратов; при анализе качества модели по схеме проверки статистических гипотез на основе выборочных данных.

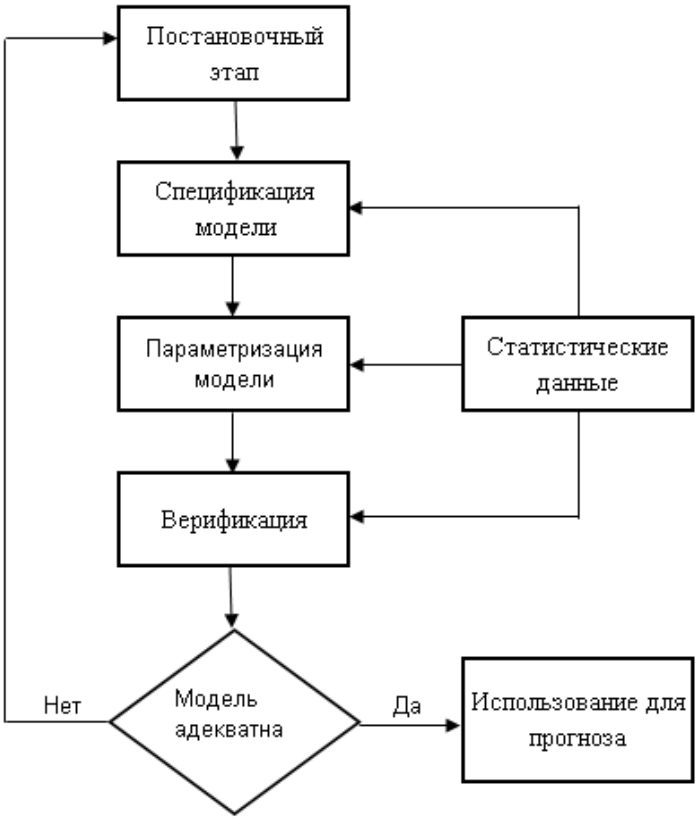


Рис. 2.2. Общая схема решения эконометрической задачи

2.5. Пример эконометрического исследования

Предположим, что некоторая риэлтерская фирма желает составить для себя точное представление об ожидаемой цене на квартиры.

Первый шаг такого исследования состоит в том, чтобы установить факторы, определяющие цену *p*, которая выступает в данном примере в качестве результирующего фактора и является зависимой (или, иначе, объясняемой) переменной. Конечно, значение цены квартиры зависит практически от бесконечного количества факторов. К ним относятся, например, площадь квартиры, количество ее комнат, площадь кухни, совмещенность или несовмещенность санузла, этажность дома, номер этажа, на котором расположена квартира, удаленность квартиры от центра города, наличие квартир на вторичном рынке жилья и новых квартир и многие-многие другие. Выберем лишь те, которые оказывают наиболее существенное влияние на цену *p*. Отнесем к ним площадь квартиры *s* и удаленность ее от центра города *l*. Эти величины называются *независимыми* (или *объясняющими*) *факторами* (их еще называют *регрессорами*). Доля влияния остальных факторов незначительна, их игнорирование в среднем не приведет к существенным отклонениям цены *p*. Поэтому все они рассматриваются как одна случайная переменная ε (она называется

возмущением или *ошибкой*). В результате зависимость переменной p разбивается на две части – зависимость *объясненную* (связанную с переменными s и l) и *случайную* ε.

Отметим, что фирма располагает данными по n квартирам города, причем для каждой из них известны площадь $s_1, s_2, \ldots, s_n$, удаленность от центра города $l_1, l_2, \ldots, l_n$ и цена $p_1, p_2, \ldots, p_n$.

Второй шаг исследования состоит в построении эконометрической модели. Речь идет о том, чтобы на основании имеющихся статистических данных определить объясненную часть переменной p, рассматривая случайную составляющую как случайную величину, подобрать функцию $f(s,l)$ так, чтобы она наиболее точно соответствовала статистическим данным. При такой постановке задачи эконометрическая модель имеет вид, схематически изображенный на рисунке 2.3.

Наблюдаемое значение цены p	=	Объясненная часть цены $f(s, l)$	+	Случайная составляющая часть цены ε

Рис. 2.3. Эконометрическая модель задачи

Очевидно, что стоимость квартиры тем больше, чем больше ее площадь, и тем меньше, чем дальше она расположена от центра города. Это означает, что между переменными p и s существует прямая зависимость, а между p и l – обратная. Поэтому в качестве общей формулы можно взять, например, формулу

$$f(s,\ l) = as + \frac{b}{1+l}. \qquad (2.1)$$

Тогда эконометрическая модель задается уравнением

$$p = as + \frac{b}{1+l} + \varepsilon\ , \qquad (2.2)$$

где p – цена квартиры; s – площадь квартиры; l – расстояние до центра города.

После спецификации модели решение задачи вступает в этап параметризации. На этом этапе необходимо подобрать параметры a и b таким образом, чтобы при подстановке в формулу (2.1) значений $s_i, l_i \quad (i = 1, 2, ..., n)$ получались значения, расположенные в среднем как можно ближе к значениям p_i (i = 1, 2, …, n).

Отметим, что успех математического моделирования во многом зависит от спецификации модели. Поэтому эконометрист, кроме модели (2.2), построит и другие, а потом из них выберет наилучшую.

Поясним еще один момент, касающийся практических целей эконометрического моделирования. Предположим, что параметризованная на основании равенства (2.2) эконометрическая модель имеет вид:

$$p = 800s + \frac{6300}{1+l} + \varepsilon . \qquad (2.3)$$

Тогда, опираясь на модель (2.3), в частности, можно:

а) спрогнозировать цену квартиры (например, квартира площадью в 40 м^2, расположенная в двух км от центра города, предположительно будет стоить $p \approx 800 \cdot 40 + \frac{6300}{1+2} = 34100$ денежных единиц);

б) оценить целесообразность (существенность) факторов *s* и *l*;

в) выявить влияние на цену квартиры каждого фактора;

г) получить статистические доказательства надежности выводов а)-в).

Теперь риэлтор может легко определить ожидаемую цену любой квартиры, даже если ее аналогов нет в базе данных фирмы. В этом и состоит главное практическое приложение полученного результата.

2.6. Функциональные и статистические зависимости

Различают два вида зависимостей между величинами (факторами): *функциональную* и *статистическую*.

При *функциональной* зависимости двух величин значению одной из них соответствует единственное значение другой. Функциональные зависимости широко представлены в естественных науках, но между экономическими показателями они проявляются достаточно редко. Это связано с тем, что в экономике влияние между исследуемыми величинами, как правило, проявляется через ряд других факторов, а порой и по ряду других причин.

В экономике взаимосвязи между величинами чаще носят статистический характер. *Статистической* называют зависимость, при которой изменение одной из величин влечет изменение распределения другой, и эта величина принимает некоторые значения с определенной вероятностью. Конечно, функциональную зависимость можно считать частным случаем статистической: значению одного фактора соответствуют значение другого фактора с вероятностью, равной единице. Однако на практике такое рассмотрение функциональной связи применения не нашло.

Наиболее важным частным случаем статистической зависимости является *корреляционная* зависимость, при которой значению одной из величин ставится в соответствия среднее значение (математическое ожидание) другой.

2.7. Эконометрическое моделирование

Как показывает пример 2.5, в большинстве случаев между экономическими показателями проявляется не однозначная (функциональная) зависимость, а такая, при которой каждому конкретному набору независимых факторов $x_1, x_2, ..., x_m$ соответствует не одно, а множество значений зависимой переменной *y* из некоторой области. Поэтому зависимый фактор *y* является случайной величиной. Например, цена на квартиры одинаковой площади, расположенной на одинаковом расстоянии от центра города, не является однозначной, она колеблется в некотором интервале. В таких случаях каждому конкретному набору объясняющих факторов $x_1, x_2, ..., x_m$ соответствует некоторое вероятностное значение зависимой переменной *y*. Чаще всего в практике экономических исследований используется корреляционная зависимость, когда в качестве этого вероятностного значения выступает математическое ожидание $M(y \mid x_1, x_2, ..., x_m)$.

Зависимость, заданная соотношением

$$M(y \mid x_1, x_2, ..., x_m) = f(x_1, x_2, ..., x_m),$$

называется *уравнением множественной регрессии*. Соответственно соотношение вида $M(y \mid x) = f(x)$ называется *уравнением парной регрессии*.

Чтобы избежать громоздкого обозначения $M(y \mid x_1, x_2, ..., x_m)$ для математического ожидания переменной *y*, используется запись $\tilde{y}$. При таких соглашениях уравнение множественной регрессии имеет вид $\tilde{y} = f(x_1, x_2, ..., x_m)$ (соответственно $\tilde{y} = f(x)$ в случае парной регрессии).

Понятно, что реальное значение зависимой переменной *y* не совпадает с условным математическим ожиданием $M(y \mid x_1, x_2, ..., x_m)$. Мы можем говорить только о том, что

$$y \approx M(y \mid x_1, x_2, ..., x_m) = f(x_1, x_2, ..., x_m).$$

Для отражения того факта, что реальные значения переменной *y* могут быть различными при одном и том же наборе объясняющих переменных, фактическая зависимость *y* от $x_1, x_2, ..., x_m$ должна быть дополнена некоторым слагаемым ε, которое и указывает на случайный характер величины *y*. Таким образом, сама величина *y* разбивается на две части (как это, в частности, указано на рисунке 2.3): одна из них (объясняемая) имеет вид $M(y \mid x_1, x_2, ..., x_m)$ и задает ту часть *y*, которая объясняется факторами $x_1, x_2, ..., x_m$, вторая часть ε является случайной величиной и определяет влияние на *y* неучтенных уравнением $M(y \mid x_1, x_2, ..., x_m) = f(x_1, x_2, ..., x_m)$ других факторов.

При таком естественном разделении связь фактора y с факторами $x_1, x_2, ..., x_m$ задается соотношением

$$y = M(y \mid x_1, x_2, ..., x_m) + \varepsilon \qquad (2.4)$$

или

$$y = f(x_1, x_2, ..., x_m) + \varepsilon . \qquad (2.5)$$

В курсе математической статистики уравнения (2.4), (2.5) называются *регрессионными моделями* (или *уравнениями регрессионной модели*).

Таким образом, *эконометрическая модель* – это форма представления взаимосвязи экономических показателей в виде суммы двух слагаемых, первое из которых отражает влияние на результативный признак выбранных факторов, а второе – влияние случайных величин.

Теперь мы можем сформулировать общую постановку задачи эконометрического моделирования. Она заключается в следующем: по имеющимся данным n наблюдений за изменением признака y в зависимости от наборов значений факторов $x_1, x_2, ..., x_m$ выбрать эконометрическую модель $y = f(x_1, x_2, ..., x_m) + \varepsilon$, оценить ее параметры и статистически обосновать, что факторы $x_1, x_2, ..., x_m$ существенны, а построенная функция $f(x_1, x_2, ..., x_m)$ такова, что наиболее точно соответствует данным наблюдений.

2.8. Методологические аспекты эконометрического моделирования

Для построения и анализа эконометрических моделей используется специфический статистический и математический аппарат. В частности, для установления тесноты связи между переменными регрессионной модели используется *корреляционный анализ*.

Корреляция (от латинского слова c*orrelatio* – взаимозависимость) в широком смысле слова означает связь между явлениями и процессами, а корреляционный анализ позволяет оценить силу, или тесноту, этой связи, используя понятия ковариации и корреляции.

Поэтому основными задачами корреляционного анализа являются:

– количественное измерение степени зависимости переменных;

– отбор факторов, оказывающих наибольшее влияние на результативный признак;

– обнаружение неизвестных причинных связей.

Что касается последнего, то с помощью только корреляционного анализа нельзя указать, какую переменную следует принимать в качестве причины, а какую – в качестве следствия. Для выявления причинной взаимообусловленности, количественные оценки взаимосвязей, полученные с помощью корреляционного анализа, должны быть обязательно дополнены

глубоким анализом сущности изучаемого явления на базе экономической теории.

Корреляционный анализ существенно связан с методами *регрессионного анализа*, которые направлены на установление формы зависимости между переменными (т.е. формы функции $f(x_1, x_2, ..., x_m)$) и оценку параметров функции регрессии (т.е. на выделение из некоторого множества функций той функции, которая дает наилучшее приближение к исходным данным).

Таким образом, основными задачами регрессионного анализа являются:

– определение вида уравнения регрессии по имеющимся данным наблюдений (спецификация модели);

– оценка параметров уравнения по реальным данным (параметризация модели);

– анализ качества уравнения, проверка адекватности уравнения эмпирическим данным, улучшение качества уравнения (верификация модели).

Термин «*регрессия*» (от латинского слова *regressio* – движение назад, возврат в прежнее состояние) ввел английский статистик Френсис Гальтон в конце XIX века при анализе влияния роста родителей и более отдаленных предков на рост детей. Гальтон заметил, в частности, что рост детей у высоких родителей в среднем меньше среднего роста родителей, а у низких родителей наблюдается обратная закономерность. Таким образом, осуществляется возврат среднего роста детей аномальных родителей к среднему росту людей в данном регионе. Кроме того, по его модели рост ребенка определяется наполовину родителями, на четверть – родителями родителей и т.д. Другими словами, модель Гальтона характеризует движение назад по генеалогическому дереву. Эти наблюдения и были положены в основу выбора терминологии. В настоящее время термин «*регрессия*», конечно, не отражает всей сущности регрессионного метода, но продолжает использоваться для описания статистической связи между случайными величинами.

Отметим еще, что какой бы хорошо подогнанной и математически обоснованной не была модель, ее главное содержание определяется экономической теорией, а результат моделирования представляет интерес лишь в том случае, когда он имеет экономическую интерпретацию.

Контрольные вопросы

1. Дайте определение эконометрики.
2. Что является объектом исследования эконометрики?
3. Каков предмет исследования эконометрики?
4. С какими науками связана эконометрика?
5. Как эконометрика связана с экономической теорией?
6. В чем проявляется связь эконометрики и экономической статистики?

7. Как эконометрика связана с математическим моделированием?

8. Что понимается под эконометрической моделью?

9. Как проявляется связь эконометрики с теорией вероятностей и математической статистикой?

10. Какова роль компьютерных технологий в эконометрике?

11. Изложите классификацию эконометрических моделей в зависимости от иерархического уровня решаемых задач.

12. Как классифицируются эконометрические модели по фактору времени?

13. Изложите классификацию эконометрических моделей в зависимости от формы математического представления.

14. Дайте определение парной и множественной эконометрических моделей.

15. Как определяются линейные и нелинейные эконометрические модели?

16. Назовите основные этапы эконометрического моделирования.

17. Какие задачи решаются на постановочном этапе эконометрического моделирования?

18. Что понимается под спецификацией модели?

19. Какие задачи решаются на этапе параметризации эконометрической модели?

20. Как осуществляется верификация эконометрической модели?

21. Какова последовательность прохождения этапов эконометрического моделирования?

22. В чем проявляется цикличность эконометрического моделирования?

23. Как в эконометрической модели учитываются объясняемая и случайная доли зависимой переменной?

24. Какая переменная называются объясняемой, а какая – регрессором?

25. В чем заключаются практические приложения эконометрических моделей?

26. Как осуществляется прогнозирование на основе эконометрических моделей?

27. Сформулируйте общую постановку эконометрической задачи.

28. Что такое уравнение регрессии?

29. Чем регрессионная модель отличается от уравнения регрессии?

30. Какие методы применяются для построения и анализа эконометрических моделей?

31. Какая зависимость называется корреляционной?

32. Перечислите основные задачи корреляционного анализа.

33. Перечислите основные задачи регрессионного анализа.

Тестовые задания

Выберите правильные ответы из предложенных вариантов:

1. Моделирование применяется в тех случаях, когда:

а) реальный объект нельзя исследовать непосредственно;

б) исследование реального объекта требует слишком высоких затрат времени и средств;

в) необходимо объяснить и спрогнозировать поведение сложного реального объекта.

2. Модель – это объект произвольной природы, который:

а) создается с целью получения новых знаний об объекте-оригинале;

б) отражает существенные для рассматриваемой задачи свойства оригинала;

в) строится для упрощения исследования объекта-оригинала.

3. Процесс моделирования включает:

а) построение модели;

б) изучение модели;

в) применение модели.

4. В зависимости от выбранных средств представления модели делятся на:

а) концептуальные, материальные и знаковые;

б) статические и динамические;

в) парные и множественные.

5. Математическая модель является:

а) концептуальной;

б) материальной;

в) знаковой.

6. По степени причинной обусловленности модели делятся на:

а) учебные, опытные, игровые;

б) экономические, социальные, биологические;

в) статические и динамические;

г) детерминированные и вероятностные.

7. Перевод рассматриваемой экономической задачи на язык математических терминов и соотношений производится на этапе:

а) исследования;

б) подготовительном;

в) интерпретации;

г) формализации.

8. Выбор наиболее подходящего метода решения и решение поставленной математической задачи производится на этапе:

а) исследования;

б) подготовительном;

в) интерпретации;

г) формализации.

9. Сбор и анализ информации по моделируемому объекту, формирование целей и задач, выбор средств моделирования производится на этапе:

а) исследования;

б) подготовительном;

в) интерпретации;

г) формализации.

10. Анализ полученных математических результатов и объяснение их в терминах исходной экономической задачи производится на этапе:

а) исследования;

б) подготовительном;

в) интерпретации;

г) формализации.

11. Экономико-математическое моделирование осуществляется в следующей последовательности:

а) исследование модели – формализация – подготовительный этап – интерпретация;

б) подготовительный этап – формализация – исследование модели – интерпретация;

в) подготовительный этап – формализация – интерпретация – исследование модели;

г) подготовительный этап – интерпретация – формализация – исследование модели.

12. Эконометрика – это:

а) раздел экономической теории, связанный с анализом статистической информации;

б) специальный раздел математики, посвященный анализу экономической информации;

в) наука, которая осуществляет качественный анализ взаимосвязей экономических явлений и процессов;

г) наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов.

13. Предметом эконометрики является:

а) определение наблюдаемых в экономике количественных закономерностей;

б) сбор и обработка статистических данных;

в) изучение экономических законов.

14. Совокупность факторов и показателей, влияющих на изучаемое экономическое явление, их роль и теоретические взаимосвязи при построении эконометрической модели устанавливает:

а) экономическая теория;

б) экономическая статистика;

в) математическое моделирование;

г) теория вероятностей и математическая статистика.

15. Формализацию экономической задачи и перевод ее на язык математики при построении эконометрической модели обеспечивает:

а) экономическая теория;

б) экономическая статистика;

в) математическое моделирование;

г) теория вероятностей и математическая статистика.

16. Исследование построенной эконометрической модели обеспечивает:

а) экономическая теория;

б) экономическая статистика;

в) математическое моделирование;

г) теория вероятностей и математическая статистика.

17. Информационную базу при построении эконометрической модели обеспечивает:

а) экономическая теория;

б) экономическая статистика;

в) математическое моделирование;

г) теория вероятностей и математическая статистика.

18. Эконометрическая модель микроуровня описывает:

а) уровень национальной или мировой экономики;

б) уровень отдельного предприятия;

в) уровень отрасли или региона.

19. По фактору времени различают:

а) парные и множественные эконометрические модели;

б) статические и динамические эконометрические модели;

в) модели с одним уравнением и модели системы одновременных уравнений;

г) линейные и нелинейные модели.

20. Парной эконометрической моделью является:

а) модель $y = f(x) + \varepsilon$;

б) модель $y = f(x_1, x_2) + \varepsilon$;

в) модель $y = \frac{3z}{z+1} + \varepsilon$;
г) модель $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$.

21. Множественной эконометрической моделью является:
а) модель $y = f(x, z) + \varepsilon$;
б) модель $y = f(x_1) + \varepsilon$;
в) модель $y = \frac{3z}{z+1} + \varepsilon$;
г) модель $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$.

22. Линейной эконометрической моделью является:
а) модель $y = \frac{x}{x-4} + \varepsilon$;
б) модель $y = 2x^5 \cdot \varepsilon$;
в) модель $y = \frac{3z}{z+1} + \varepsilon$;
г) модель $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$.

23. Нелинейной эконометрической моделью является:
а) модель $y = \frac{4x^3}{3x-4} + \varepsilon$;
б) модель $y = 2x + \varepsilon$;
в) модель $y = \frac{2x+3z}{z+1} + \varepsilon$;
г) модель $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$.

24. На этапе спецификации эконометрической модели осуществляется:
а) проверка адекватности модели;
б) оценка значений параметров выбранной функции связи;
в) определение целей и задач исследования;
г) выбор формы связи между переменными.

25. На этапе верификации эконометрической модели осуществляется:
а) проверка адекватности модели;
б) оценка значений параметров выбранной функции связи;
в) определение целей и задач исследования;
г) выбор формулы связи между переменными.

26. Регрессионная модель имеет вид:
а) $y \approx M(y \mid x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m)$;

б) $y = f(x_1, x_2, \ldots, x_m) + \varepsilon$;

в) $\tilde{y} = f(x_1, x_2, \ldots, x_m)$;

г) $M(y \mid x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m)$.

27. В регрессионной модели $y = f(x) + \varepsilon$ регрессором является:

а) переменная y;

б) переменная x;

в) переменная ε.

28. В регрессионной модели $y = f(x) + \varepsilon$ результативный признак величина обозначается:

а) переменной y;

б) переменной x;

в) переменной ε.

29. В регрессионной модели $y = f(x) + \varepsilon$ влияние неучтенных факторов отражается:

а) переменной y;

б) переменной x;

в) переменной ε.

30. Зависимость, при которой изменение одной величин влечет изменение среднего значения (математического ожидания) другой, называется:

а) функциональной;

б) корреляционной;

в) логической.

Ответы тестовых заданий

Номер задания	Ответы	Номер задания	Ответы	Номер задания	Ответы
1	а), б), в)	11	б)	21	а), г)
2	а), б), в)	12	г)	22	г)
3	а), б), в)	13	а)	23	а), в)
4	а)	14	а)	24	г)
5	в)	15	в)	25	а)
6	г)	16	г)	26	б)
7	г)	17	б)	27	б)
8	а)	18	б)	28	а)
9	б)	19	б)	29	в)
10	в)	20	а), в)	30	б)

РАЗДЕЛ 2
Эконометрические модели

«Эконометрика – это не то же самое, что экономическая статистика. Она не идентична и тому, что мы называем экономической теорией. Эконометрика не является синонимом применения математики в экономике. Это единство всех трех составляющих. И это единство образует эконометрику».

Лауреат Нобелевской премии Фриш Р.

Глава 3
Модели парной регрессии

Основные понятия: *парная регрессия, линейная парная модель, нелинейная регрессия, классическая линейная регрессионная модель, корреляционное поле* (*диаграмма рассеивания*)*, метод наименьших квадратов, линеаризация модели, выборочная ковариация, линейный коэффициент корреляции, общая, факторная и остаточная дисперсии, индекс корреляции, коэффициент детерминации, коэффициент эластичности, стандартная ошибка регрессии, средняя ошибка аппроксимации, точечный и интервальный прогнозы.*

Литература: [2-4], [7], [9], [15-16].

3.1. Постановочный этап

Хотя поведение экономического показателя зависит практически от бесконечного множества факторов, экономическая теория выделила и исследовала значительное число устоявшихся связей между парами показателей. В частности, хорошо изучены зависимость спроса и предложения от цены товара, зависимость уровня безработицы от инфляции, зависимость объема производства от величины основных фондов, зависимость между производительностью труда и уровнем механизации и многие другие.

Поэтому парная регрессионная модель является достаточно распространенной эконометрической моделью, описывающей корреляционную взаимосвязь двух экономических показателей. Ее преимущества перед другими моделями заключаются в относительной простоте построения и исследования, возможности представления графическими средствами и ясной экономической интерпретации

параметров. Кроме того, парная модель всегда служит начальной точкой более глубокого эконометрического анализа.

Парная регрессия представляет собой модель, где среднее значение *зависимой переменной y* рассматривается как функция одной *независимой переменной* (*регрессора*) *x*; *уравнение парной регрессионной модели* имеет вид

$$y = f(x) + \varepsilon. \qquad (3.1)$$

В уравнении (3.1) величина ε является случайной и указывает на случайный характер величины *y*. Сама величина *y* разбивается на две части: одна из них имеет вид $f(x)$ и оценивает объясняемую часть *y*, а вторая часть ε определяет влияние на *y* неучтенных *уравнением парной регрессии* $\tilde{y} = f(x)$ других факторов.

Случайная величина ε называется также *возмущением*. Она включает влияние случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

Наибольшую опасность в практическом использовании регрессионных моделей представляют ошибки измерения. Если ошибки спецификации можно уменьшить, изменяя форму модели, а ошибки выборки – увеличивая объем выборки, то ошибки измерения практически сводят на нет все усилия по количественной оценке связи между признаками.

Общая постановка задачи парного эконометрического моделирования заключается в следующем: по имеющимся данным *n* наблюдений за изменением признака *y* в зависимости от наборов значений фактора *x* выбрать эконометрическую модель $y = f(x) + \varepsilon$, оценить ее параметры и статистически обосновать, что построенная функция $f(x)$ наиболее точно соответствует данным наблюдений.

3.2. Классификация парных моделей

Относительно формы зависимости выделяются линейные и нелинейные парные регрессионные модели. *Линейная парная регрессионная модель* имеет вид $y = a + bx + \varepsilon$. *Нелинейная регрессия* выражается нелинейной функцией $f(x)$ в уравнении (3.1).

Примером модели парной линейной регрессии является предложенная Кейнсом модель $C = C_0 + bI + \varepsilon$ зависимости частного потребления C от располагаемого дохода I, где $C_0 > 0$ – величина автономного потребления, $0 < b \le 1$ – предельная склонность к потреблению.

Классическим примером парной нелинейной модели является модель Филлипса $y = a + \frac{b}{x} + \varepsilon$, характеризующая соотношение между уровнем

безработицы (x) в процентах и процентным изменением заработной платы (y). Кроме того, часто оказываются нелинейными производственные функции, функции спроса и др.

Наиболее часто используется линейная регрессия. Внимание к ней объясняется четкой экономической интерпретацией ее параметров и тем, что в большинстве случаев нелинейные формы связи для выполнения расчетов преобразуют в линейную форму с помощью специальных процедур линеаризации.

Линейная парная регрессионная модель называется *классической линейной регрессионной моделью*, если она удовлетворяет следующим модельным предположениям:

1. Математическое ожидание случайной переменной ε равно нулю.

2. Дисперсия случайной переменной ε постоянна для всех наблюдений (это условие называется *условием гомоскедастичности*).

3. Отсутствует систематическая связь между значениями случайной переменной для различных наблюдений.

4. Объясняющая и случайная переменные независимы.

Предположения 1–4 называются *условиями Гаусса–Маркова*.

Классическая линейная регрессионная модель называется *нормальной*, если в дополнение к условиям Гаусса-Маркова выполняется следующее условие.

5. Случайная переменная имеет нормальный закон распределения вероятностей с нулевым математическим ожиданием и постоянной дисперсией.

Нелинейная парная регрессия делится на два типа:

1) *регрессия, нелинейная относительно включенной в уравнение объясняющей переменной, но линейная по оцениваемым параметрам* (например, полином $\tilde{y} = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n$, гипербола $\tilde{y} = a + \frac{b}{x}$);

2) *регрессия, нелинейная по оцениваемым параметрам* (например, степенная $\tilde{y} = a \cdot x^b$, показательная $\tilde{y} = a \cdot b^x$, экспоненциальная $\tilde{y} = e^{ax+b}$ регрессии).

В зависимости от характера парной регрессии различают:

– *прямую регрессию* (увеличение объясняющей переменной вызывает увеличение зависимой переменной);

– *обратную регрессию* (увеличение объясняющей переменной вызывает уменьшение зависимой переменной).

Относительно типа соединения переменных различают:

– *непосредственную регрессию* (зависимая и объясняющая переменные связаны непосредственно друг с другом);

– *косвенную регрессию* (объясняющая переменная действует на результативную переменную через какую-то третью или ряд других переменных);

– *нонсенс-регрессию* (ложная регрессия, при которой отсутствует причинная обусловленность связи переменных).

3.3. Спецификация модели

Любое эконометрическое исследование начинается со спецификации модели, т.е. с выбора формы модели. При этом, как правило, спецификация опирается на имеющиеся экономические теории, специальные знания и интуитивные представления об анализируемой экономической системе.

В случае парной регрессии выделяется доминирующий фактор (если такой имеется), который и используется в качестве объясняющей переменной x.

Таким образом, для парной модели спецификация – это определение вида f аналитической зависимости $\tilde{y} = f(x)$. Она проводится одним из трех методов.

1. *Графический метод* заключается в построении *корреляционного поля* (или *диаграммы рассеивания*), которое позволяет произвести визуальный анализ эмпирических данных.

Пусть имеется по n наблюдений $x_1, x_2, ..., x_n$ и $y_1, y_2, ..., y_n$ за поведением переменной x и переменной y. В прямоугольной системе координат по оси абсцисс отмечаются значения независимой переменной x, по оси ординат – значения зависимой переменной y. Расположение построенных точек (x_i, y_i) определяет картину зависимости двух переменных.

По ширине разброса точек можно сделать вывод о степени тесноты связи. Если точки расположены близко друг к другу в виде узкой полосы, то можно утверждать о наличии относительно тесной связи. Если точки разбросаны широко по полю, то имеется слабая связь. Например, на рисунке 3.1а) связь между x и y близка к линейной $\tilde{y} = a + bx$, так как точки сгруппированы относительно прямой (причем отличной от $\tilde{y} = \bar{y}$, когда зависимость отсутствует (рис. 3.1б)). На рисунке 3.1в) связь между x и y может быть экспоненциальной $\tilde{y} = e^{ax+b}$. На рисунке 3.1г) связь между переменными логарифмическая и может быть описана уравнением $\tilde{y} = a + b\ln x$. Параболическая зависимость на рисунке 3.1д) может быть задана уравнением $\tilde{y} = a_0 + a_1 x + a_2 x^2$. Корреляционное поле, по которому определяется отсутствие зависимости, изображено на рисунке 3.1д).

2. *Экспериментальный метод* состоит в том, что зависимость между переменными описывается несколькими моделями, а затем выбирается наиболее качественная путем сравнения величины остаточной дисперсии. Метод удобен при обработке информации с помощью компьютерных технологий.

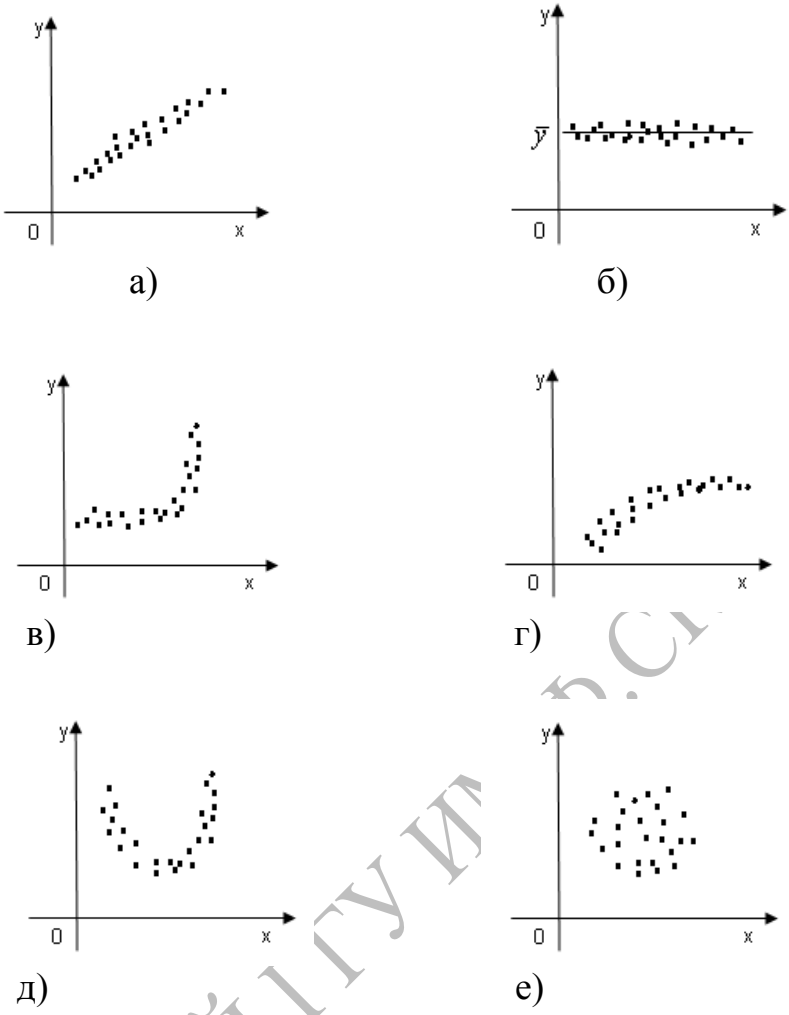


Рис. 3.1. Корреляционное поле: а) линейной, в) экспоненциальной, г) логарифмической, д) параболической зависимостей; б), е) отсутствие зависимостей

3. *Аналитический метод* основан на анализе природы изучаемой взаимосвязи, исходя либо из некоторых теоретических соображений, либо из опыта практических исследований (как в примере 2.5 с оценкой стоимости квартиры).

От правильно выбранной спецификации модели во многом зависит величина случайных ошибок: они тем меньше, чем в большей мере теоретические значения результативного признака ($\tilde{y}_i$) соответствуют наблюдаемым значениям (y_i).

3.4. Параметризация линейной модели

После того, как регрессионная модель специфицирована, производится ее параметризация. В случае парной линейной регрессионной модели $y = a + bx + \varepsilon$ речь идет о количественной оценке коэффициентов a и b. Суть параметризации парной линейной модели состоит в выборе из бесконечного

множества всех прямых плоскости такой прямой, к которой точки корреляционного поля прилегают «наиболее тесным образом» (рис. 3.2). При этом в качестве меры тесноты такого прилегания могут выступать различные критерии.

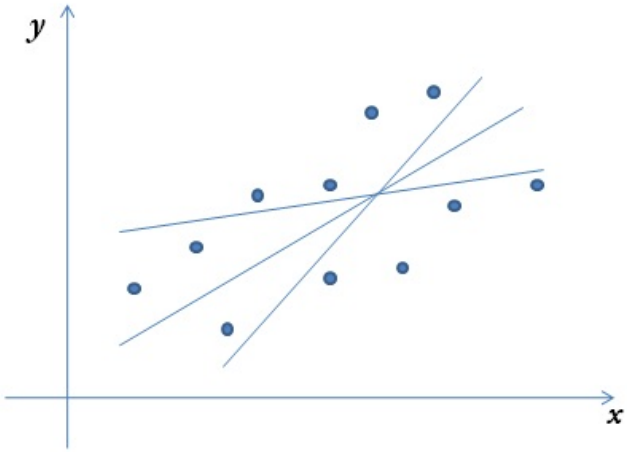


Рис. 3.2. Графическая иллюстрация параметризации
парной линейной модели

Наиболее распространенным методом оценки параметров регрессии является *метод наименьших квадратов* (МНК). Этот метод является наиболее простым с вычислительной точки зрения. Здесь в качестве меры общей близости точек наблюдений от искомой прямой выбирается сумма квадратов отклонений.

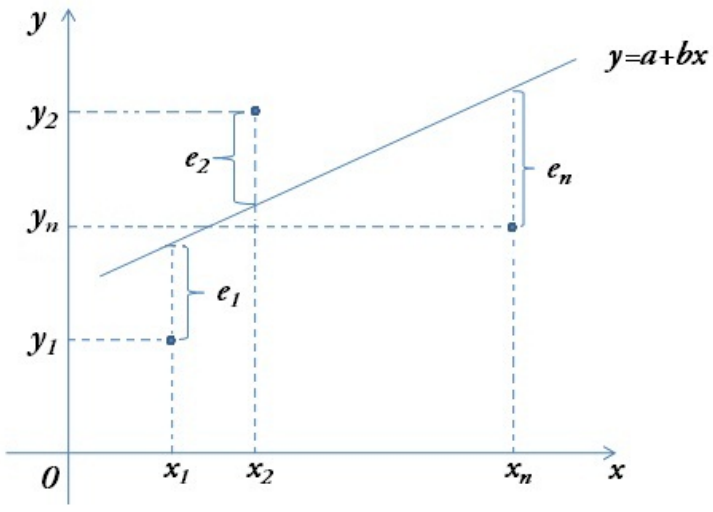


Рис. 3.3. Разброс результатов наблюдений около графика
парной линейной регрессии

Суть МНК (отсюда и название метода) заключается в нахождении параметров модели, при которых минимизируется сумма квадратов отклонений эмпирических (наблюдаемых) значений y_i, $i=1,2,...,n$, зависимой переменной от теоретических значений $\tilde{y}_i = a + bx_i$, $i=1,2,...,n$, полученных по уравнению регрессии: $S = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \to \min$ (при выборе линии регрессии такой критерий позволяет учитывать величину всех остатков в совокупности).

На рисунке 3.3 изображен разброс точек корреляционного поля около линии регрессии. При этом через $e_i = y_i - \tilde{y}_i$ обозначается разность между фактическим и теоретическим значениями зависимой переменной.

Аргументами функции

$$S = e_1^2 + e_2^2 + ... + e_n^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2 = S(a,b)$$

являются неизвестные параметры *a* и *b* уравнения регрессии. Исследование на экстремум данной функции проводится методами дифференциального исчисления. Приравнивая к нулю частные производные функции $S(a,b)$, приходим к системе двух линейных уравнений, из которой находятся оценки неизвестных коэффициентов *a* и *b* уравнения регрессии:

$$\begin{cases} na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \\ a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i. \end{cases} \qquad (3.2)$$

Решая систему, получим

$$b = \frac{\overline{xy} - \bar{x}\cdot\bar{y}}{\overline{x^2} - (\bar{x})^2},\ a = \bar{y} - b\cdot\bar{x}, \qquad (3.3)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

В уравнении $\tilde{y} = a + bx$ коэффициент *b* при переменной *x* указывает, на какую величину изменится в среднем значение *y* при изменении фактора *x* на одну единицу измерения. В этом и заключается экономическая интерпретация коэффициента *b*. Коэффициент *a* формально показывает

прогнозируемый уровень показателя y при значении x, равном нулю. Правда, если $x = 0$ находится достаточно далеко от выборочных значений $x_1, x_2, \ldots, x_n$, то буквальная интерпретация коэффициента a может и не иметь ясного смысла.

Свойства коэффициентов регрессии a и b в линейной парной модели $y = a + bx + \varepsilon$ существенным образом зависят от свойств случайной составляющей ε. Поэтому метод наименьших квадратов не всегда обеспечивает оптимальные свойства оценкам параметров a и b. Однако это имеет место, если линейная модель является классической.

Теорема Гаусса–Маркова. *Если линейная парная модель является классической* (*т.е. для нее выполнены модельные предположения 1–4 пункта 3.2 данной главы*)*, то оценки коэффициентов парной линейной регрессии, полученные с помощью МНК, являются несмещенными, состоятельными и эффективными*.

Теорема Гаусса–Маркова гарантирует, что:

а) в определении линии регрессии отсутствует систематическая ошибка;

б) при возрастании числа n наблюдений дисперсия оценок параметров регрессии стремится к нулю;

в) оценки имеют наименьшую дисперсию по сравнению с любыми другими оценками.

Выполнимость модельного предположения 1 означает, что ошибки различных наблюдений поступают с разными знаками и компенсируют друг друга. Таким образом, исключается ситуация, когда ошибки систематически накапливаются с одним и тем же знаком. Проверка условия 1 отчасти может быть проведена путем анализа вида корреляционного поля. Если точки наблюдений разбросаны хаотично по отношению к линии регрессии, то условие 1 выполняется. Если же такой разброс имеет системный характер (например, в некоторой части корреляционного поля имеет место систематическое смещение точек в одну сторону от линии регрессии), то это дает основание для того, чтобы усомниться в точности вычислений либо правильности спецификации.

Требование постоянства дисперсии случайной переменной говорит о том, что все наблюдения производятся с одинаковой точностью. Поэтому в русскоязычной литературе говорят, что имеет место равноточная схема наблюдений. Выполнимость модельного предположения 2 также может быть установлена с помощью графического анализа корреляционного поля: условие 2 выполняется, если точки наблюдений расположены внутри полосы постоянной ширины, окаймляющей линию регрессии.

Отметим, что невыполнение условия 1 приводит к смещению оценок коэффициентов регрессии, а в случае невыполнения условия 2 оценки параметров могут быть неэффективными.

Последнее имеет место и при невыполнении модельных предположений 3-4, так как в этом случае коррелированность случайных членов между собой или с независимой переменной может определять систематическую тенденцию к тому, что предыдущие наблюдения будут влиять на случайный член последующих наблюдений. Условие 3 часто нарушается в динамических моделях, но практически всегда выполняется в статических.

Что касается условия 5 (предположения о нормальности), то выполнимость или невыполнимость его не оказывает существенного влияния на качество оценок параметров регрессии, но в то же время является достаточно важным. Условие 5 необходимо для проверки статистических гипотез значимости уравнения регрессии и его коэффициентов, а также для установления их доверительных интервалов. Связано это с тем, что сам аппарат проверки статистических гипотез базируется на свойствах нормально распределенных случайных величин.

3.5. Параметризация нелинейной модели

Параметризация нелинейных парных моделей имеет свою специфику. Конечно, можно воспользоваться нелинейным МНК. Идея такого подхода такова же, как и в линейном случае: минимизируется сумма квадратов отклонений эмпирических (наблюдаемых) значений y_i, $i = 1,2,...,n$, зависимой переменной от теоретических значений $\tilde{y}_i = f(x_i)$, $i = 1,2,...,n$, полученных по уравнению регрессии:

$$S = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \to \min .$$

Главный недостаток такого подхода заключается в том, что для нахождения значений параметров приходится решать весьма сложные системы нелинейных уравнений.

В эконометрических исследованиях большее распространение получил метод определения параметров нелинейной парной регрессии, основанный на применении процедур линеаризации. Он состоит в том, что с помощью соответствующих преобразований исходных переменных исследуемая зависимость представляется в виде линейного соотношения между преобразованными переменными.

Для оценки параметров регрессии, нелинейной относительно включенной в уравнение объясняющей переменной, но линейной по оцениваемым параметрам, используется подход, называемый «замена переменных». Суть его состоит в замене «нелинейных» объясняющих переменных новыми «линейными» переменными. После этого к новой регрессии применяется обычный МНК. Например, модель гиперболической регрессии $y = a + \frac{b}{x} + \varepsilon$

параметризуется после замены $x' = \frac{1}{x}$ и $y' = y$ как линейная модель $y' = a + bx' + \varepsilon$.

Для оценки параметров регрессии, нелинейной по оцениваемым параметрам, часто применяется метод логарифмирования с последующей заменой переменных. Например, уравнение экспоненциальной регрессии $\tilde{y} = e^{a+bx}$ параметризуется как линейная модель $y' = a + bx' + \varepsilon$ после логарифмирования и введения новых переменных $y' = \ln y$ и $x' = x$.

В таблице 3.1 приведены виды нелинейных парных регрессий и формулы замен переменных.

Метод линеаризации для оценки параметров парной регрессии удобен тем, что он может быть легко реализован в стандартных пакетах прикладных компьютерных программ (например, Excel). Недостаток подхода проявляется в том, что оценки параметров регрессии в таком случае получаются несколько смещенными. Это связано с тем, что сами оценки находятся не из условия минимизации суммы квадратов отклонений исходных переменных, а из условия минимизации суммы квадратов отклонений для преобразованных переменных, что не одно и то же.

Таблица 3.1. Оценки параметров нелинейных моделей регрессии

Вид регрессии	Линеаризующее преобразование
Экспоненциальная регрессия $\tilde{y} = e^{a+bx}$	$x' = x,\ y' = \ln y$ (после логарифмирования)
Логарифмическая регрессия $\tilde{y} = a + b\ln x$	$x' = \ln x,\ y' = y$
Степенная регрессия $\tilde{y} = a \cdot x^b$	$x' = \ln x,\ y' = \ln y$ (после логарифмирования)
Показательная регрессия $\tilde{y} = a \cdot b^x$	$x' = x,\ y' = \ln y$ (после логарифмирования)
Гиперболическая регрессия $\tilde{y} = a + \frac{b}{x}$	$x' = \frac{1}{x},\ y' = y$

3.6. Оценка тесноты линейной связи между переменными

Уравнение парной регрессии всегда дополняется показателями тесноты связи между переменными. При использовании линейной регрессии такими показателями являются выборочная ковариация и линейный коэффициент корреляции.

Выборочной ковариацией $\mathrm{cov}(x, y)$ называется среднее произведений отклонений значений переменных x и y от своих средних величин $\bar{x}$, $\bar{y}$: $\mathrm{cov}(x, y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x} \cdot \bar{y}.$

Пусть точка $(\bar{x}, \bar{y})$ является центром корреляционного поля эмпирической зависимости между переменными x и y (рис. 3.4).

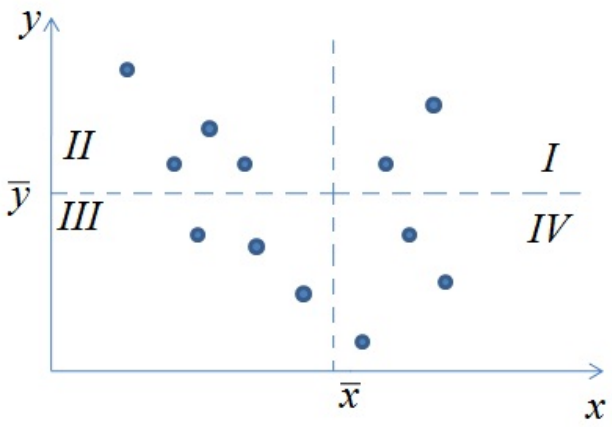


Рис. 3.4. Разброс точек корреляционного поля около средних значений

Тогда вертикальная и горизонтальная прямые, проведенные через нее, разделяют плоскость поле на четыре области. Положительный вклад в ковариацию формируется в областях I и III, а отрицательный – в областях II и IV. Если положительные вклады преобладают над отрицательными, то ковариация будет положительной, а большинство точек поля будет сосредоточено в областях I и III, группируясь возле возрастающей прямой.

Поэтому ковариация характеризует не только величину рассеивания значений факторов x и y, но и линейную связь между ними:

– при $\mathrm{cov}(x, y) > 0$ связь между факторами x и y прямая, т.е. большим значениям x соответствуют большие значения y;

– при $\mathrm{cov}(x, y) < 0$ связь между факторами x и y обратная, т.е. большим значениям x соответствуют меньшие значения y;

– при $\mathrm{cov}(x, y) \to 0$ линейная связь между x и y отсутствует.

Более подходящим измерителем взаимосвязи переменных x и y, чем выборочная ковариация, является линейный коэффициент корреляции r_{xy}. Основная причина этого заключается в том, что ковариация зависит от единиц, в которых измеряются переменные x и y, в то время как коэффициент корреляции есть величина безразмерная.

Линейным коэффициентом корреляции r_{xy} факторов x и y называется величина, определяемая по формуле

$$r_{xy} = \frac{\mathrm{cov}(x;y)}{\sigma_x \sigma_y} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\sigma_x \sigma_y}.$$

Линейный коэффициент корреляции логически связан с коэффициентом b линейной регрессии. Эта связь, в частности, проявляется через формулу

$$r_{xy} = b \cdot \frac{\sigma_x}{\sigma_y} = b \cdot \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Качественная оценка тесноты линейной связи между переменными x и y в зависимости от величины линейного коэффициента корреляции выявляется по шкале английского статистика Чеддока (таблица 3.2). В соответствии с этой шкалой выделяются пять качественных уровней связи между двумя переменными: слабая, умеренная, заметная, высокая и весьма высокая.

Таблица 3.2. Шкала Чеддока

Теснота связи	Значение линейного коэффициента корреляции при наличии	
	прямой связи	обратной связи
Слабая	0,1–0,3	(-0,1)–(-0,3)
Умеренная	0,3–0,5	(-0,3)–(-0,5)
Заметная	0,5–0,7	(-0,5)–(-0,7)
Высокая	0,7–0,9	(-0,7)–(-0,9)
Весьма высокая	0,9–0,99	(-0,9)–(-0,99)

Выборочный коэффициент корреляции устанавливает также напрвление линейной связи (прямая или обратная):

1) при $0 < r_{xy} < 1$ большим значениям x соответствуют большие значения y (рисунок 3.5 а));

2) при $-1 < r_{xy} < 0$ большим значениям x соответствуют меньшие значения y (рисунок 3.5 б)).

Кроме того, он указывает, что:

– при $r_{xy} = 0$ величины x и y являются некоррелированными: величина коэффициента корреляции, близкая к нулю, говорит об отсутствии линейной связи между величинами, но не об отсутствии связи между ними вообще (рисунок 3.5 в));

– при $|r_{xy}|=1$ существует линейная функциональная зависимость между выборочными значениями x и y (прямая или обратная); точки лежат точно на прямой.

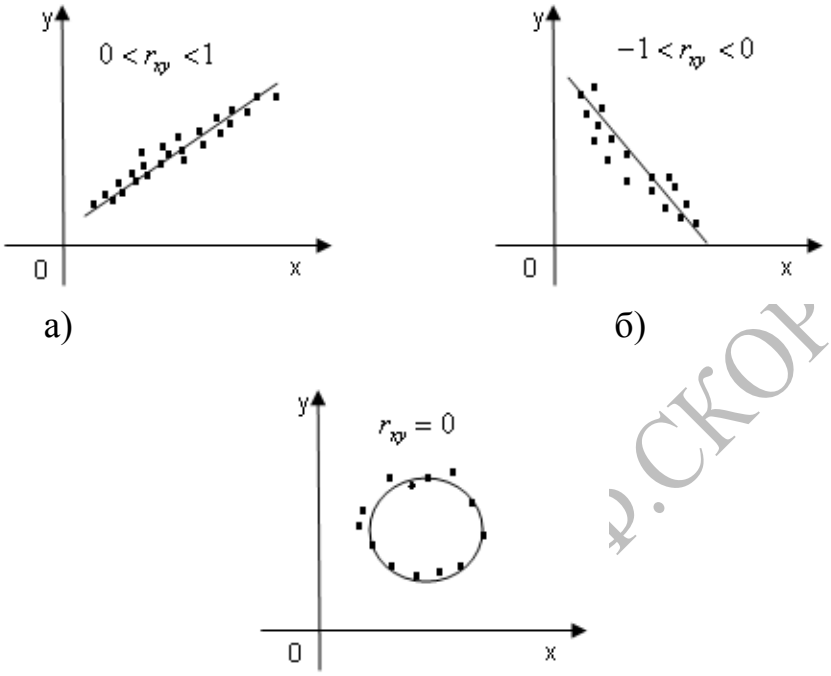


Рис. 3.5. Значение коэффициента корреляции
в зависимости от вида корреляционного поля

3.7. Оценка тесноты нелинейной связи между переменными

Для оценки тесноты взаимосвязи результативного признака с фактором в случае нелинейной парной регрессии анализируются следующие дисперсии:

1) *общая дисперсия* результативного признака y, отражающая влияние на y как основного фактора x, так и неучтенных случайных факторов: $\sigma_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$, где $\bar{y}$ – выборочное среднее значение результативного признака по выборке $y_1, y_2, ..., y_n$;

2) *факторная дисперсия* результативного признака y, отражающая влияние на y только основного фактора x: $\sigma_\phi^2 = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2$, где $\tilde{y}_i = f(x_i)$, $i = 1, 2, ..., n$, – значения результативного признака y, полученные по уравнению регрессии;

3) *остаточная дисперсия* результативного признака y, отражающая влияние на y неучтенных факторов и характеризующая меру разброса зависимой переменной возле линии регрессии: $\sigma^2_{\text{ост}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2$, где $e_i = y_i - \tilde{y}_i$.

В соответствии с теоремой о разложении дисперсии справедливо равенство $\sigma_y^2 = \sigma_\phi^2 + \sigma^2_{\text{ост}}$.

Отсюда следует, что чем меньше в формуле $\sigma_y^2 = \sigma_\phi^2 + \sigma^2_{\text{ост}}$ величина $\sigma^2_{\text{ост}}$, тем меньше точки наблюдений рассеяны относительно линии регрессии, а значит, тем теснее взаимосвязь результативного признака y и основного фактора x.

Поэтому теснота взаимосвязи результативного признака y с фактором x в случае парной нелинейной регрессии оценивается с помощью *индекса корреляции* ρ_{xy}:

$$\rho_{xy} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma_y^2}} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Величина данного показателя удовлетворяет соотношению $0 \le \rho_{xy} \le 1$. Чем ближе ρ_{xy} к единице, тем теснее связь рассматриваемых признаков.

В случае линейной регрессии справедливо равенство $\rho_{xy} = |r_{xy}|$. Отметим, что для нелинейной регрессии линейный коэффициент корреляции r_{xy} дает лишь приближенную оценку связи и в общем случае не совпадает с индексом корреляции ρ_{xy}.

Для относительной (в процентах) характеристики силы связи фактора с результативным признаком может быть использован коэффициент эластичности. При этом различают обобщающие (средние) и точечные коэффициенты эластичности.

Средний коэффициент эластичности $\overline{Э} = f'(\bar{x})\frac{\bar{x}}{\bar{y}}$ на основании вида уравнения $\tilde{y} = f(x)$ парной регрессии позволяет определить, на сколько процентов в среднем по совокупности изменится результирующий признак y

при изменении фактора x на 1% от своего среднего значения $\bar{x}$. В случае линейной парной регрессии $\overline{Э} = b\dfrac{\bar{x}}{\bar{y}} = \dfrac{b \cdot \bar{x}}{a + b \cdot \bar{x}}$.

В эконометрических исследованиях широкое применение получила степенная регрессия $\tilde{y} = a \cdot x^b$. Во многом это связано с тем, что коэффициент b в ней имеет четкую экономическую интерпретацию – он совпадает с коэффициентом эластичности.

Точечный коэффициент эластичности рассчитывается для конкретного значения $x = x_0$ и показывает, на сколько процентов изменится y относительно уровня $f(x_0)$ при изменении x на 1% от уровня x_0. Формула расчета имеет вид $Э_{x_0} = f'(x_0)\dfrac{x_0}{y_0}$.

3.8. Верификация модели: проверка адекватности

Вопрос о возможности практического применения построенной эконометрической модели для прогнозирования экономического показателя может быть решен только после проверки ее общего качества или, иначе, ее адекватности. При этом оценивается степень подгонки теоретических значений $\tilde{y}_i = f(x_i)$ к статистическим данным y_i. Другими словами, проверяется, насколько широко рассеяны точки корреляционного поля относительно линии регрессии $\tilde{y} = f(x)$.

Для анализа общего качества уравнения линейной парной регрессии обычно используется *коэффициент детерминации* $R^2 = r_{xy}^2$. В случае нелинейной регрессии используется *индекс детерминации*

$$R^2 = \rho_{xy}^2 = \frac{\textit{факторная регрессия}}{\textit{общая дисперсия}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

Коэффициент и индекс детерминации определяют долю дисперсии результативного признака, обусловленную изменением факторного признака. Тем самым коэффициент и индекс детерминации дают относительную меру влияния фактора на результат, фиксируя одновременно и роль ошибок.

Так как $0 \le R^2 \le 1$, то величина $R^2 \cdot 100\,\%$ в процентах показывает, какая часть изменения зависимой переменной y определяется объясняющей переменной x. Чем выше показатель детерминации, тем лучше модель описывает исходные данные. Соответственно величина $(1 - R^2) \cdot 100\%$ характеризует долю дисперсии переменной y, вызванную влиянием прочих неучтенных в модели факторов.

При значениях показателей тесноты связи меньше 0,7 величина коэффициента детерминации всегда будет ниже 50 %. Это означает, что на долю факторных признаков приходится меньшая часть по сравнению с остальными неучтенными в модели факторами, влияющими на изменение результативного показателя. Построенные при таких условиях регрессионные, конечно, модели имеют низкое практическое значение.

Индекс детерминации нелинейной модели ρ_{xy}^2 можно сравнивать с коэффициентом детерминации r_{xy}^2 для обоснования возможности применения линейной формы. Чем больше кривизна линии регрессии, тем меньше величина r_{xy}^2 по сравнению ρ_{xy}^2. А близость этих показателей указывает на то, что нет необходимости усложнять форму уравнения регрессии и можно использовать линейную модель.

Общее качество модели может быть оценено с помощью *стандартной ошибки регрессии* $s = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}$, которая является несмещенной оценкой среднего квадратического отклонения наблюдаемых значений результативного признака от теоретических значений, рассчитанных по модели. Величина стандартной ошибки регрессии характеризует среднюю величину рассеивания наблюдаемых значений переменной y возле линии регрессии.

Для оценки адекватности уравнения регрессии также используется показатель *средней ошибки аппроксимации*:

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

– среднее отклонение расчетных значений от наблюдаемых. Ошибка аппроксимации не более 8–12 % свидетельствует о хорошем качестве модели.

3.9. Верификация модели: проверка статистической значимости

После того, как найдены показатели регрессионной модели, возникает вопрос. Можно ли считать, что их оценки, полученные на основе выборочных данных, будут такими же и для всей генеральной совокупности? Ведь сами оценки, естественно, изменяются и при добавлении в исходную выборку новых данных, и при переходе к другой выборке. Таким образом, возникает проблема обоснованности распространения выводов, полученных на основе конкретной выборки, на всю генеральную совокупность. Эта проблема получила название проблемы оценки статистической значимости. Решение ее осуществляется с помощью аппарата проверки статистических гипотез.

Оценка значимости уравнения регрессии производится для того, чтобы узнать, пригодно ли уравнение регрессии для практического использования (для прогнозирования) или нет.

Обычно оценка статистической значимости уравнения парной регрессии проводится по двум направлениям:

а) оценивается значимость коэффициента R^2;

б) оценивается значимость коэффициентов регрессии.

Оценка значимости всего уравнения регрессии в целом осуществляется с помощью *F*-критерия Фишера.

F-критерий Фишера заключается в проверке нулевой гипотезы H_0 о статистической незначимости уравнения регрессии (т.е. гипотезы H_0 о том, что $R^2 = 0$).

Для этого выполняется сравнение фактического $F_{\text{набл}}$ и критического (табличного) $F_{\text{кр}}$ значений *F*-критерия Фишера.

Наблюдаемое значение статистики $F_{\text{набл}}$ вычисляется по выборочным данным на основании формулы $F_{\text{набл}} = \frac{R^2}{1-R^2} \cdot (n-2)$. При этом для линейной регрессии вместо R^2 используется значение r_{xy}^2.

По таблицам критических точек F-распределения находится критическое значение статистики $F_{\text{кр}}$ при заданном уровне значимости α. Для линейной регрессии число степеней свободы определяется значениями $k_1 = 1$ и $k_2 = n-2$, где n – число наблюдений. Уровень значимости α – вероятность отвергнуть гипотезу H_0 при условии, что она верна. Обычно величина α принимается равной 0,05 или 0,01.

Если $F_{\text{кр}} < F_{\text{набл}}$, то нулевая гипотеза отвергается, что говорит о соответствии теоретического уравнения регрессии выборочным данным. Если $F_{\text{кр}} > F_{\text{набл}}$, то признается ненадежность уравнения регрессии.

Возможна ситуация, когда некоторые из вычисленных коэффициентов линейной парной регрессии a и b не обладают необходимой степенью значимости. В этом случае такие коэффициенты должны быть исключены из уравнения регрессии. Поэтому проверка статистической значимости построенного уравнения парной линейной регрессии включает в себя также и проверку значимости каждого коэффициента регрессии.

При этом выдвигаются нулевые гипотезы о незначимом отличии от нуля коэффициентов регрессии *a* и *b*, т.е. $H_0 : a = 0$, $H_0 : b = 0$ при альтернативных гипотезах $H_1 : a \neq 0$, $H_1 : b \neq 0$. Проверка данных гипотез осуществляется с помощью t-статистики, имеющей распределение Стьюдента с числом

степеней свободы $n-2$. Для этого рассчитываются стандартные ошибки коэффициентов регрессии

$$m_a = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{(n-2)} \cdot \frac{\sum_{i=1}^{n} x_i^{\,2}}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}}\,,\ m_b = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2}{(n-2)\sum_{i=1}^{n}(x_i-\overline{x})^2}}\,. \qquad (3.4)$$

По выборочным данным для каждого из коэффициентов вычисляются наблюдаемые значения t-статистики $t_{\text{набл}}$ как отношения значений коэффициентов к величине их стандартной ошибки: $t_a = \frac{a}{m_a}$, $t_b = \frac{b}{m_b}$, которые затем сравнивается с табличным значением t-статистики $t_{\text{кр}}$. Если $|t_{\text{набл}}| > t_{\text{кр}}$, то нулевые гипотезы $H_0: a = 0$, $H_0: b = 0$ отвергаются и признается, что коэффициенты регрессии не случайно отличаются от нуля, а значит, они статистически значимы. Если же $|t_{\text{набл}}| < t_{\text{кр}}$, то коэффициенты регрессии статистически не значимы и природа их формирования случайна.

Если незначимым окажется коэффициент a, а выбранная форма модели по некоторым причинам должна быть линейной, то проводится пересчет уравнения регрессии в предположении, что $a = 0$, т.е. строится линейная модель $y = bx + \varepsilon$, не содержащая свободного члена. Если же незначимым окажется коэффициент b, то нужно изменить спецификацию модели с линейной формы на нелинейную.

Так как рассчитанные по выборке значения показателей регрессии являются приближенными, то для оценки того, насколько точные значения показателей могут отличаться от рассчитанных, осуществляется построение доверительных интервалов.

Доверительные интервалы для каждого коэффициента регрессии имеют вид:

$$(a - t_{\text{кр}} m_a; a + t_{\text{кр}} m_a)\,,\ (b - t_{\text{кр}} m_b; b + t_{\text{кр}} m_b)\,.$$

Они определяют пределы, в которых находятся точные значения коэффициентов регрессии с заданным уровнем значимости α.

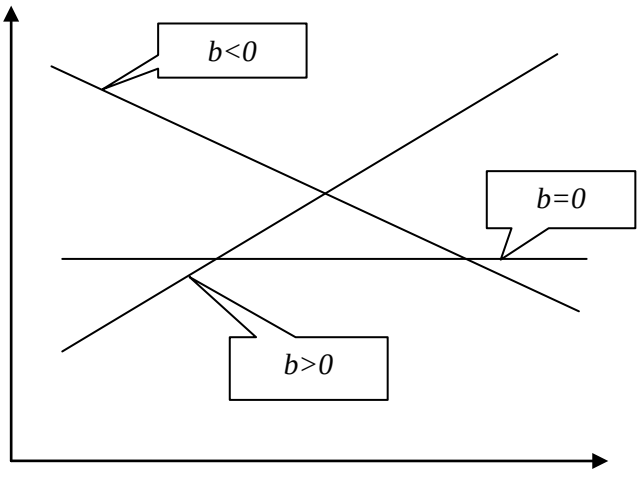


Рис. 3.6. Наклон линии регрессии в зависимости от коэффициента b

Поскольку знак коэффициента регрессии b указывает (рисунок 3.6) либо на рост результативного признака y при увеличении фактора x ($b > 0$), либо на уменьшение результативного признака при увеличении фактора x ($b < 0$), либо на его независимость от независимой переменной x ($b = 0$), то границы доверительного интервала для коэффициента регрессии b не должны иметь различные знаки. В противном случае доверительный интервал предполагает, что истинное значение коэффициента регрессии b одновременно может быть положительным, отрицательным и даже нолем, а это, конечно, недопустимо.

Проверка статистической значимости линейного коэффициента корреляции осуществляется по следующей схеме:

1. Рассчитываются линейный коэффициент корреляции r_{xy} и его стандартная ошибка $m_r = \sqrt{\dfrac{1 - r_{xy}^2}{n - 2}}$.

2. Выдвигается нулевая гипотеза о равенстве нулю коэффициента корреляции $H_0 : r_{xy} = 0$ при альтернативной гипотезе $H_1 : r_{xy} \neq 0$. При проверке нулевой гипотезы используется t-статистика, имеющая распределение Стьюдента с $n - 2$ степенями свободы, где n – объем выборки. По выборке находится наблюдаемое значение статистики $t_{\text{набл}} = \dfrac{|r_{xy}|}{m_r}$, где $m_r = \sqrt{\dfrac{1 - r_{xy}^2}{n - 2}}$ – стандартная ошибка коэффициента корреляции. Для заданного уровня значимости α по таблице критических точек Стьюдента определяется критическая точка $t_{\text{кр}}$. Если $t_{\text{набл}} \geq t_{\text{кр}}$, то нулевая гипотеза об отсутствии корреляционной зависимости величин отвергается, т.е. линейный коэффициент корреляции значим и статистическая зависимость между величинами существует. Если $t_{\text{набл}} < t_{\text{кр}}$, то нулевая гипотеза принимается.

3. Для значимого коэффициента корреляции r_{xy} устанавливается доверительный интервал при уровне значимости α, который имеет вид:

$$(r_{xy} - t_{\text{кр}} \cdot \frac{1 - r_{xy}}{\sqrt{n}}; r_{xy} + t_{\text{кр}} \cdot \frac{1 - r_{xy}}{\sqrt{n}}).$$

Отметим, что проверка статистической значимости линейного коэффициента корреляции с помощью t-статистики Стьюдента может не проводиться, если уже проведена проверка статистической значимости коэффициента детерминации с помощью F-критерия Фишера. Это обусловлено тем, что статистики критериев взаимосвязаны друг с другом.

3.10. Прогнозирование по парной регрессионной модели

Под прогнозированием по парной регрессионной модели понимается нахождение неизвестных значений зависимой переменной y для тех значений независимой переменной x, которых нет в исходных наблюдениях. Различают точечное и интервальное прогнозирование. В первом случае оценка – некоторое число, во втором – интервал, в котором находится истинное значение зависимой переменной с заданной вероятностью.

Прогностические способности модели определяются величиной индекса детерминации. О достаточном качестве прогноза можно говорить, как правило, лишь при значении коэффициента (индекса) детерминации, большем 0,75.

Точечный прогноз y_p результирующего признака y определяется путем подстановки в уравнение регрессии $\tilde{y} = f(x)$ значения x_p независимого фактора: $y_p = f(x_p)$. В случае линейной модели $y_p = a + b \cdot x_p$.

Точечный прогноз явно не реален, поэтому он всегда дополняется расчетом доверительного интервала прогноза. В случае интервального прогноза по парной линейной модели предварительно рассчитывается *стандартная ошибка прогноза*:

$$m_p = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \qquad (3.5)$$

где $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}}$.

Затем строится *доверительный интервал прогноза*

$$(y_p - t_{\text{кр}} \cdot m_p; y_p + t_{\text{кр}} \cdot m_p),$$

т.е. определяются нижняя и верхняя границы интервала прогноза.

Из формулы (3.5) следует, что ширина доверительного интервала прогноза зависит от стандартной ошибки регрессии s (т.е. от качества модели), а также от значения x_p независимого фактора x, как это видно на рисунке 3.7: при $x_p = \overline{x}$ она минимальна, а по мере удаления x_p от $\overline{x}$ она увеличивается.

Отсюда следует, что интервальный прогноз реалистичен в пределах диапазона исходных данных. Экстраполяция кривой регрессии, т.е. ее использование вне пределов наблюдаемого диапазона значений объясняющей переменной, может привести к значительным погрешностям. Поэтому, в частности, долгосрочное прогнозирование по трендовым моделям, где в качестве независимой переменной выступает время, как правило, не оправдывает себя.

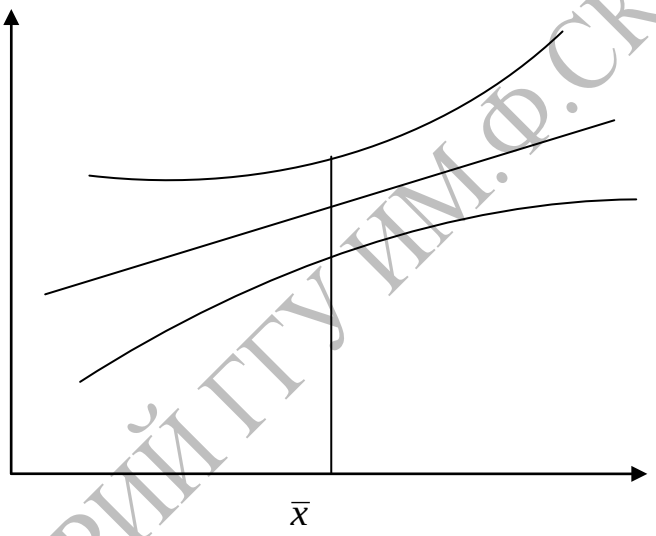


Рис. 3.7. Доверительная полоса линии регрессии

3.11. Обзор некоторых вопросов и проблем парной регрессии

Базовые понятия теории вероятностей и математической статистики в объеме, необходимом для анализа парных регрессионных моделей, представлены в [3-5].

Все оценки основных показателей парной регрессии приведены нами без доказательств. Строгие выводы их можно найти в [2-4].

Базовые нелинейные модели, используемые в парном эконометрическом моделировании, а также примеры их практического использования описываются в [3]. Здесь же охарактеризованы основные критерии «хорошей модели», которые следует учитывать при построении работоспособной модели и сравнении ее с другими моделями.

Считается, что при построении парной регрессионной модели число наблюдений должно в 7-8 раз превышать число рассчитываемых параметров

при переменной *x*. Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла. Если же вид функции усложняется, то соответственно требуется увеличение числа наблюдений, ибо каждый параметр при *x* должен рассчитываться хотя бы по 7 наблюдениям. Если, например, мы выбираем параболическую модель $y = a_0 + a_1 x + a_2 x^2 + \varepsilon$, то требуется объем информации, предполагающий уже не менее 14 наблюдений.

Кроме оценки общего качества и статистической значимости, верификация парной линейной регрессионной модели предусматривает проверку модельных предположений (условий 1-5 классической нормальной линейной модели). Некоторые аспекты такой проверки рассматриваются в главе 5. Подробно проблемы тестирования и устранения негативных явлений в парных моделях (гетероскедастичности, автокорреляции и др.) излагаются в [2,3].

Мы рассматриваем оценку тесноты связи между переменными *x* и *y* как один из элементов верификации модели, хотя в ряде случаев ее полезно провести на этапе спецификации. Это связано с тем, что если линейный коэффициент корреляции r_{xy} оказывается недостаточно высоким (линейная связь слабая, умеренная или даже заметная), то уже на ранней стадии эконометрического моделирования следует задуматься о целесообразности выбора парной модели.

Вскрывая взаимосвязи изучаемых процессов, эконометрические модели не решают вопроса о причине этих взаимосвязей. Может оказаться, что совместные изменения переменных вовсе не означают наличия причинных связей между ними. И если это так, то одна из главных целей эконометрики, состоящая в выработке рекомендаций для принятия эффективного решения, не будет достигнута, так как будет неизвестно, на какой фактор надо воздействовать.

Именно потребность в причинном объяснении корреляции и привела американского генетика С.Райта к созданию метода путевого анализа. Путевой анализ основан на изучении всей структуры причинных связей между переменными и заключается в разложении коэффициента парной корреляции на четыре компоненты:

- компоненту прямого влияния (причина *u*, вызывающая *y*, задается действием переменной *x*);
- компоненту косвенного влияния (причина *u* воздействует на промежуточное звено *x* и тем самым вызывает *y*);
- непричинную компоненту, объясняемую наличием общих причин, воздействующих на *x* и на *y;*
- непричинную компоненту, зависящую от неанализируемой в модели корреляции.

Методика путевого анализа описана в [2].

О значимости коэффициентов линейной регрессии можно судить не только на основании критерия Стьюдента, но также и по значениям показателя *P-значение* из таблицы «Дисперсионный анализ», рассчитанной в режиме работы инструмента "Регрессия" в MS Excel. Коэффициенты признаются значимыми, если *P-значение* меньше заданного уровня значимости α=0,05. Это же касается и оценки значимости коэффициента детерминации.

Примеры решения типовых заданий

Пример 3.1. По статистическим данным, приведенным в таблице 3.3, построить корреляционное поле зависимости спроса y на товар от его цены x и определить форму связи между результирующим признаком y и фактором x.

Таблица 3.3. Статистические данные примера 3.1

Цена товара	Спрос на товар
99	100
82	115
77	210
69	270
52	323
44	478
31	544
29	564
28	570
27,5	574

Решение:

Из вида корреляционного поля (рисунок 3.8) можно сделать вывод о том, что между результирующим признаком y и фактором x существует обратная зависимость, т.е. с ростом цены спрос на товар уменьшается. Такое поведение спроса в зависимости от цены согласуется с выводами экономической теории. Можно предположить также, что форма связи между результирующим признаком y и фактором x является линейной.

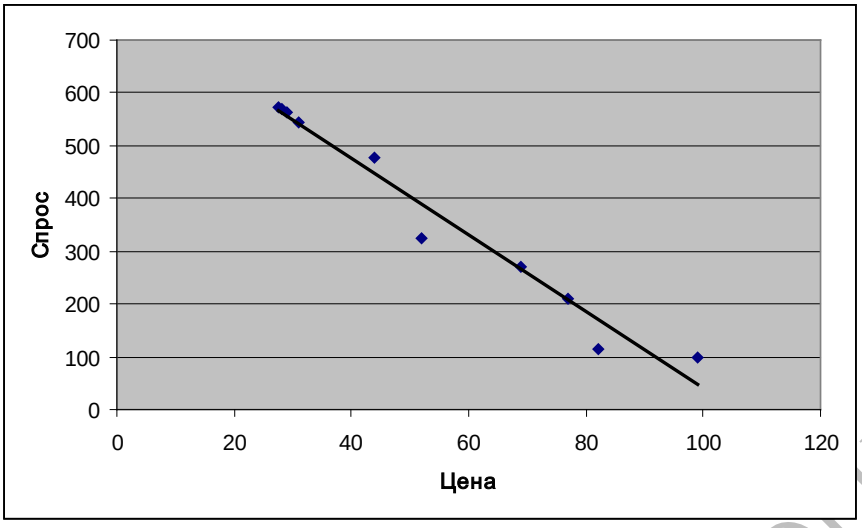


Рис. 3.8. Корреляционное поле статистической зависимости между спросом и ценой

Пример 3.2. В таблице 3.4 представлена информация по семи территориям некоторого региона о значениях процентной доли расходов на покупку продовольственных товаров в общих расходах (y) и значениях среднедневной заработной платы одного работающего (x).

Таблица 3.4. Статистические данные примера 3.2

Расходы на покупку продовольственных товаров в общих расходах (в процентах, y)	Среднедневная заработная плата одного работающего (в денежных единицах, x)
68,8	45,1
61,2	59,0
59,9	57,2
56,7	61,8
55,0	58,8
54,3	47,2
49,3	55,2

1. Для характеристики зависимости y от x рассчитать параметры следующих регрессий:

а) линейной;

б) степенной;

в) показательной;

г) гиперболической.

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

Решение:

1а) Для расчета параметров a и b линейной модели $y = a + bx + \varepsilon$ решаем систему уравнений (3.2) относительно a и b.

Соответствующие вычисления сведем в таблицу 3.5.

Таблица 3.5. Вычисление параметров a и b

	y	x	xy	x^2	y^2	$\tilde{y}$	$y-\tilde{y}$	$\left\|\frac{y-\tilde{y}}{y}\right\|$
1	68,8	45,1	3102,88	2034,01	4733,44	61,3	7,5	10,9
2	61,2	59,0	3610,80	3481,00	3745,44	56,5	4,7	7,7
3	59,9	57,2	3426,28	3271,84	3588,01	57,1	2,8	4,7
4	56,7	61,8	3504,06	3819,24	3214,89	55,5	1,2	2,1
5	55,0	58,8	3234,00	3457,44	3025,00	56,5	-1,5	2,7
6	54,3	47,2	2562,96	2227,84	2948,49	60,5	-6,2	11,4
7	49,3	55,2	2721,36	3047,04	2430,49	57,8	-8,5	17,2
Сумма	405,2	384,3	22162,34	21338,4	23685,76	405,2	0,0	56,7
Среднее	57,89	54,90	3166,05	3048,34	3383,68	-	-	8,1
σ	5,74	5,86	-	-	-	-	-	-
σ^2	32,92	34,34	-	-	-	-	-	-

Теперь по формулам (3.3) находим:

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \approx -0{,}35,$$

$$a = \bar{y} - b\bar{x} = 57{,}89 + 0{,}35 \cdot 54{,}9 = 76{,}88.$$

Уравнение линейной регрессии имеет вид $\tilde{y} = 76{,}88 - 0{,}35 \cdot x$. С увеличением среднедневной заработной платы на одну денежную единицу доля расхода на покупку продовольственных товаров снижается в среднем на 0,35 %-ных пункта. Рассчитаем линейный коэффициент парной корреляции:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = -0{,}35 \cdot \frac{5{,}86}{5{,}74} = -0{,}357.$$

Связь умеренная, обратная.

Определим коэффициент детерминации: $r_{xy}^2 = (-0{,}35)^2 = 0{,}127$.

Изменения результата на 12,7% объясняются влиянием фактора x.

Подставляя в уравнение регрессии фактические значения x, определим теоретические (расчетные) значения $\tilde{y}$. Найдем величину средней ошибки аппроксимации $\overline{A}$:

$$\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \tilde{y}}{y_i}\right| \cdot 100\,\% = 8{,}1\%.$$

В среднем расчетные значения отклоняются от фактических на 8,1%.

Рассчитываем наблюдаемое значение F-критерия: $F_{\text{набл}} = \frac{0{,}127}{0{,}873} \cdot 5 = 0{,}7$. По таблице значений F-критерия при уровне значимости $\alpha = 0{,}05$ находим $F_{\text{кр}} = 6{,}61$. Так как $F_{\text{кр}} > F_{\text{набл}}$, то следует принять гипотезу H_0 о случайной природе выявленной зависимости и статистической незначимости параметров уравнения и показателя тесноты связи.

1б) Построению уравнения $\tilde{y} = a \cdot x^b$ степенной регрессии предшествует процедура линеаризации переменных. В данном примере линеаризация производится путем логарифмирования обеих частей уравнения:

$$\lg y = \lg a + b \lg x,$$
$$\tilde{Y} = C + b \cdot X,$$

где $Y = \lg y, X = \lg x, C = \lg a.$

Все расчеты сведем в таблицу 3.6.

Рассчитаем C и b:

$$b = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sigma_X^2} = \frac{3{,}0572 - 1{,}7605 \cdot 1{,}7370}{0{,}0023} \approx -0{,}298,$$
$$C = \overline{Y} - b\overline{X} = 1{,}7605 + 0{,}298 \cdot 1{,}7370 = 2{,}278.$$

Получим линейное уравнение $\tilde{Y} = 2{,}278 - 0{,}298 \cdot X$.

Выполнив его потенцирование, получим уравнение регрессии: $\tilde{y} = 10^{2{,}278} \cdot x^{-0{,}298} = 189{,}7 \cdot x^{-0{,}298}$.

Таблица 3.6. Вычисление параметров C и b

	Y	X	XY	X^2	$\tilde{y}$	$y - \tilde{y}$	$(y - \tilde{y})^2$	$\left\|\frac{y - \tilde{y}}{y}\right\|$
1	1,8376	1,6542	3,0398	2,7364	61,0	7,8	60,8	11,3
2	1,7868	1,7709	3,1642	3,1361	56,3	4,9	24,0	8,0
3	1,7774	1,7574	3,1236	3,0885	56,8	3,1	9,6	5,2
4	1,7536	1,7910	3,1407	3,2077	55,5	1,2	1,4	2,1
5	1,7404	1,7694	3,0795	3,1308	56,3	-1,3	1,7	2,4
6	1,7348	1,6739	2,9039	2,8019	60,2	-5,9	34,8	10,9
7	1,6928	1,7419	2,9487	3,0342	57,4	-8,1	65,6	16,4

Сумма	12,3234	12,1587	21,4003	21,1355	403,5	1,7	197,9	56,3
Среднее	1,7605	1,7370	3,0572	3,0194	-	-	28,27	8,0
σ	0,0425	0,0484	-	-	-	-	-	-
σ^2	0,0018	0,0023	-	-	-	-	-	-

Подставляя в данное уравнение фактические значения x, получаем теоретические значения результата $\tilde{y}$. По ним рассчитываем показатели тесноты связи: индекс корреляции ρ_{xy} и среднюю ошибку аппроксимации $\overline{A}$:

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{1 - \frac{28,27}{32,92}} = 0,3758, \qquad \overline{A} = 8,0\%.$$

Значения показателей ρ_{xy} и $\overline{A}$ степенной модели указывают, что она несколько лучше линейной регрессии описывает взаимосвязь.

1в) Построению уравнения показательной регрессии $\tilde{y} = a \cdot b^x$ предшествует процедура линеаризации путем логарифмирования обеих частей уравнения:

$$\lg y = \lg a + x \lg b;$$

$$\tilde{Y} = C + B \cdot x,$$

где $Y = \lg y, C = \lg a, B = \lg b$.

Все расчеты сведем в таблицу 3.7.

Вычислим значения параметров регрессии A и B:

$$B = \frac{\overline{xY} - \bar{x} \cdot \bar{Y}}{\sigma_x^2} = \frac{96,5711 - 1,7605 \cdot 54,9}{34,33} \approx -0,0023,$$

$$C = \bar{Y} - B \cdot \bar{x} = 1,7605 + 0,0023 \cdot 54,9 = 1,887.$$

Полученное линейное уравнение принимает вид $\tilde{Y} = 1,887 - 0,0023 \cdot x$.

Произведем потенцирование полученного уравнения, получим уравнение регрессии: $\tilde{y} = 10^{1,887} \cdot 10^{-0,0023 \cdot x} = 77,1 \cdot 0,9947^x$.

Таблица 3.7. Вычисление параметров C и B

	Y	x	xY	x^2	$\tilde{y}$	$y - \tilde{y}$	$(y - \tilde{y})^2$	$\left\lvert\frac{y - \tilde{y}}{y}\right\rvert$
1	1,8376	45,1	82,8758	2034,01	60,7	8,1	65,61	11,8
2	1,7868	59,0	105,4212	3481,00	56,4	4,8	23,04	7,8

3	1,7774	57,2	101,6673	3271,84	56,9	3,0	9,00	5,0
4	1,7536	61,8	108,3725	3819,24	55,5	1,2	1,44	2,1
5	1,7404	58,8	102,3355	3457,44	56,4	-1,4	1,96	2,5
6	1,7348	47,2	81,8826	2227,84	60,0	-5,7	32,49	10,5
7	1,6928	55,2	93,4426	3047,04	57,5	-8,2	67,24	16,6
Сумма	12,3234	384,3	675,9974	21338,4	403,0	-1,8	200,78	56,3
Среднее	1,7605	54,9	96,5711	3048,34	-	-	28,68	8,0
σ	0,0425	5,86	-	-	-	-	-	-
σ^2	0,0018	34,33	-	-	-	-	-	-

Тесноту связи оценим через индекс корреляции ρ_{xy}:

$$\rho_{xy} = \sqrt{1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \sqrt{1 - \frac{28,68}{32,92}} = 0,3589.$$

Связь умеренная. Средняя ошибка аппроксимации в допустимых пределах: $\overline{A} = 8,0\%$. Показательная функция чуть хуже, чем степенная, описывает изучаемую зависимость.

1г) Уравнение гиперболической регрессии $\tilde{y} = a + b \cdot \frac{1}{x}$ линеаризуется заменой $z = \frac{1}{x}$. Тогда $\tilde{y} = a + b \cdot z$.

Все расчеты сведем в таблицу 3.8.

Вычислим значения параметров регрессии a и b:

$$b = \frac{\overline{yz} - \bar{y} \cdot \bar{z}}{\sigma_z^2} = \frac{1,0723 - 57,9 \cdot 0,0184}{0,000005} \approx 1051,4,$$

$$a = \bar{y} - b \cdot \bar{z} = 57,89 - 1051,4 \cdot 0,0184 = 38,5.$$

Получим уравнение регрессии $\tilde{y} = 38,5 + 1051,4 \cdot \frac{1}{x}$. Вычислим индекс корреляции $\rho_{xy} = \sqrt{1 - \frac{27,84}{32,92}} = 0,3944.$

Таблица 3.8. Вычисление параметров a и b

	y	z	yz	z^2	$\tilde{y}$	$y-\tilde{y}$	$(y-\tilde{y})^2$	$\left\|\frac{y-\tilde{y}}{y}\right\|$
1	68,8	0,0222	1,5255	0.000492	61,8	7,0	49,00	10,2
2	61,2	0,0169	1,0373	0.000287	56,3	4,9	24,01	8,0
3	59,9	0,0175	1,0472	0,000306	56,9	3,0	9,00	5,0
4	56,7	0,0162	0,9175	0,000262	55,5	1,2	1,44	2,1
5	55	0,0170	0,9354	0,000289	56,4	-1,4	1,96	2,5
6	54,3	0,0212	1,1504	0,000449	60,8	-6,5	42,25	12,0
7	49,3	0,0181	0,8931	0,000328	57,5	-8,2	67,24	16,6
Сумма	405,2	0,1291	7,5064	0,002413	405,2	0,0	194,90	56,5
Среднее	57,9	0,0184	1,0723	0,000345	-	-	27,84	8,1
σ	5,74	0,002145	-	-	-	-	-	-
σ^2	32,9476	0,000005	-	-	-	-	-	-

Средняя ошибка аппроксимации в допустимых пределах: $\overline{A} = 8{,}1\%$. Для гиперболического уравнения регрессии получена наибольшая оценка тесноты связи $\rho_{xy} = 0{,}3944$ (по сравнению с линейной, степенной и показательной регрессиями).

2. Для гиперболической регрессии вычислим наблюдаемое значения критерия Фишера:

$$F_{\text{набл}} = \frac{\rho_{xy}^2}{1-\rho_{xy}^2} \cdot (n-2) = \frac{0{,}1555}{0{,}8445} \cdot 5 = 0{,}92.$$

Тогда при $\alpha = 0{,}05$ имеем, что $F_{\text{кр}} = 6{,}61 > F_{\text{набл}}$. Следовательно, принимается гипотеза H_0 о статистически незначимых параметрах уравнения гиперболической регрессии. Поэтому общее качество и этой модели следует признать невысоким.

Полученные результаты можно объяснить сравнительно невысокой теснотой зависимости между результирующим признаком y и фактором x, а также небольшим числом наблюдений ($n = 7$).

Пример 3.3. По данным таблицы 3.9 построить линейное уравнение регрессии, отражающее зависимость стоимости квартиры от ее жилой площади.

Для построенного уравнения вычислить:

1) коэффициент корреляции,
2) коэффициент детерминации;
3) наблюдаемое значение критерия Фишера;
4) стандартные ошибки коэффициентов регрессии;
5) доверительные интервалы коэффициентов регрессии.

Осуществить точечный и интервальный прогнозы по построенной модели в случае, когда площадь квартиры составляет 41 кв. м.

Таблица 3.9. Статистические данные примера 3.3

№ п/п	Стоимость (доллары)	Жилая площадь (кв. м.)
1	5000	30,2
2	5200	32
3	5350	32
4	5880	37
5	5430	30
6	5430	30
7	5430	30
8	5350	29
9	5740	33
10	5570	31
11	5530	30
12	6020	34
13	7010	38
14	6420	31
15	7150	39
16	7190	39,5

Решение:

По формулам $b = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2}$, $a = \bar{y} - b \cdot \bar{x}$ находим коэффициенты регрессии: $b = 170{,}239$, $a = 262{,}847$. Поэтому построенная линейная модель имеет вид $y = 262{,}847 + 170{,}239x + \varepsilon$.

Коэффициент регрессии b модели показывает, что в среднем увеличение жилой площади квартиры на 1 кв. метр приводит к увеличению ее стоимости на 170,24 доллара.

Расчет линейного коэффициента корреляции и коэффициента детерминации дает $r_{xy} = 0{,}853$, $R^2 = r_{xy}^2 = 0{,}7281$. Связь между факторами является высокой, поэтому стоимость квартиры существенно зависит от ее жилой площади. Величина $R^2 \cdot 100\% = 72{,}81\%$ показывает, что изменения стоимости квартиры на 72,81% объясняется размером жилой площади.

Расчет дисперсионного отношения Фишера дает значение $F_{\text{набл}} = \dfrac{R^2}{1 - R^2} \cdot 14 = 37{,}504$. Сравнивая наблюдаемое значение F-критерия Фишера с критическим $F_{\text{кр}} = 4{,}60$ при $\alpha = 0{,}05$, получаем, что $F_{\text{кр}} < F_{\text{набл}}$. Таким образом, уравнение регрессии значимо и построенная модель адекватна выборочным данным.

Расчет стандартных ошибок коэффициентов регрессии осуществляется по формулам (3.4). В нашем случае имеем $m_a = 918{,}35$, $m_b = 27{,}79$.

Доверительные интервалы для каждого коэффициента регрессии имеют вид: $(a - t_{\text{кр}} m_a; a + t_{\text{кр}} m_a)$, $(b - t_{\text{кр}} m_b; b + t_{\text{кр}} m_b)$. Зная точечные оценки коэффициентов регрессии, их стандартные ошибки и критическое значение t-статистики Стьюдента $t_{\text{кр}} = 2{,}1448$, находим, что $-1706{,}691 < a < 2232{,}538$, $110{,}616 < b < 229{,}861$.

Для вычисления точечного прогноза подставим значение x_p=41 в полученное уравнение линейной регрессии $\tilde{y} = 262{,}847 + 170{,}239x$. Получим: $y_p = 262{,}847 + 170{,}239 \cdot 41 \approx 7242{,}65$. Таким образом, прогнозируемая по построенной модели стоимость квартиры площадью 41 квадратный метр составляет 7242,65 доллара.

Для построения доверительного интервала прогноза вычислим стандартную ошибку прогноза m_p по формуле $m_p = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, где $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}}$ – стандартная ошибка регрессии. В нашем случае $s = 382{,}9$. Поэтому $m_p = 382{,}9 \cdot \sqrt{1 + \frac{1}{16} + \frac{(41 - 32{,}86)^2}{189{,}76}} \approx 454{,}99$. Теперь нижняя и верхняя границы интервала прогноза при $\alpha = 0{,}05$ и $t_{\text{кр}} = 2{,}1448$ определяются по формулам $y_p - t_{\text{кр}} \cdot m_p$ и $y_p + t_{\text{кр}} \cdot m_p$. Окончательно находим, что доверительный интервал прогноза стоимости квартиры имеет вид $(6266{,}78; 8218{,}51)$. Таким образом, значение цены квартиры площадью 41 квадратный метр с вероятностью 0,95 находится в пределах от 6266,78 до 8218,51 доллара.

Пример 3.4. По данным проведенного опроса восьми групп семей (таблица 3.10) о расходах на питание y и душевом доходе x построить и исследовать линейную регрессионную модель. Найти прогнозное значение результативного фактора y при значении фактора, составляющем 110% от среднего уровня душевого дохода семьи.

Таблица 3.10. Статистические данные примера 3.4

Расходы на продукты питания (тыс. ден. ед.)	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
Душевой доход семьи (тыс. ден. ед.)	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Решение:

Предположим, что связь между доходами семьи и расходами на продукты питания линейная. Для подтверждения нашего предположения построим корреляционное поле (рисунок 3.9) и проанализируем расположение точек наблюдений.

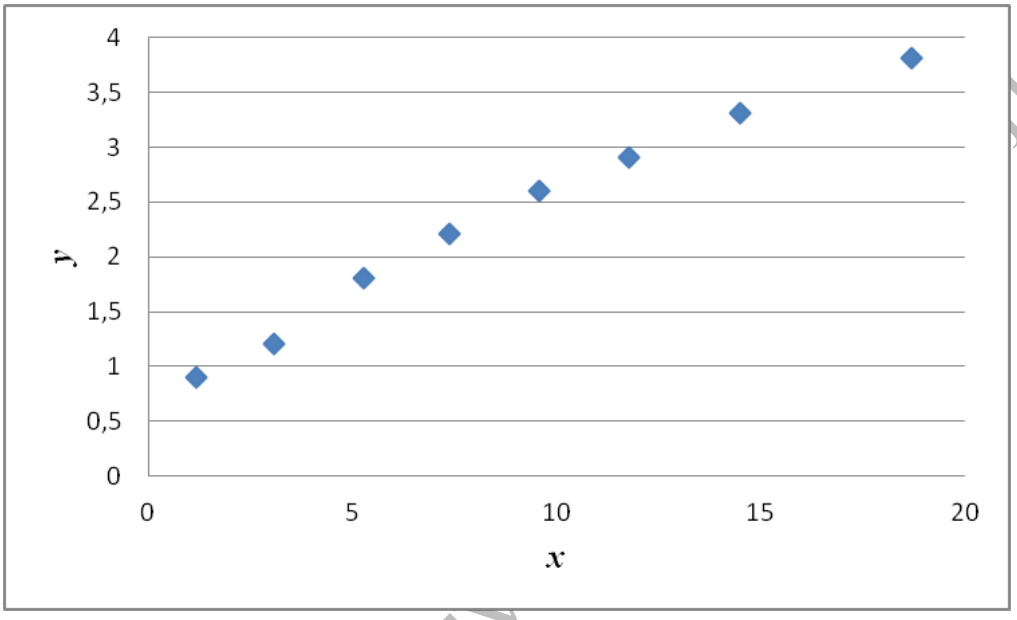


Рис. 3.9. Корреляционное поле примера 3.4

Из графика видно, что точки группируются возле некоторой прямой линии.

Рассчитаем параметры линейного уравнения $\tilde{y} = a + bx$ парной регрессии. Для этого воспользуемся формулами (3.3):

$$b = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} \approx 0{,}169,$$

$$a = \bar{y} - b\bar{x} = 0{,}824.$$

Получили уравнение $\tilde{y} = 0{,}824 + 0{,}169x$. Таким образом, с увеличением душевого дохода семьи на 1000 денежных единиц расходы на питание увеличиваются на 169 денежных единиц.

Как было указано выше, уравнение линейной регрессии всегда дополняется показателем тесноты связи – линейным коэффициентом корреляции r_{xy}. В нашем случае $r_{xy} = 0{,}991$. Близость коэффициента корреляции к 1 указывает на тесную линейную связь между признаками. По шкале Чеддока теснота этой связи характеризуется как весьма высокая.

Коэффициент детерминации $R^2 = r_{xy}^2 = 0{,}982$ показывает, что уравнением регрессии объясняется 98,2% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,8%.

Оценим качество уравнения регрессии в целом с помощью F -критерия Фишера. Рассчитаем наблюдаемое значение F -критерия:

$$F_{\text{набл}} = \frac{R^2}{1-R^2} \cdot (n-2) = \frac{0{,}982}{1-0{,}982} \cdot 6 = 335{,}66.$$

Критическое значение F -статистики $F_{\text{кр}}$ при уровне значимости $\alpha = 0{,}05$ равно 4,6.

Так как $F_{\text{кр}} > F_{\text{набл}}$, то признается статистическая значимость уравнения в целом.

Для оценки статистической значимости коэффициентов регрессии и доверительных интервалов по формулам (3.4) вычислим стандартные ошибки коэффициентов регрессии. В нашем случае имеем $m_a = 0{,}0971$, $m_b = 0{,}0092$.

Вычислим наблюдаемые значения t-статистики Стьюдента:

$$t_a = \frac{a}{m_a} = \frac{0{,}824}{0{,}0971} = 8{,}48,\ t_b = \frac{b}{m_b} = \frac{0{,}169}{0{,}0092} = 18{,}3.$$

Критическое значение t-статистики Стьюдента при уровне значимости $\alpha = 0{,}05$ равно 2,447. Так как неравенство $|t_{\text{набл}}| > t_{\text{кр}}$ выполняется для обоих коэффициентов регрессии, то признается их статистическая значимость.

Доверительные интервалы для каждого коэффициента регрессии имеют вид:

$$(a - t_{\text{кр}} m_a ; a + t_{\text{кр}} m_a),\ (b - t_{\text{кр}} m_b ; b + t_{\text{кр}} m_b).$$

Зная точечные оценки коэффициентов регрессии, их стандартные ошибки и критическое значение t -статистики Стьюдента $t_{\text{кр}} = 2{,}447$, находим, что

$$0{,}586 < a < 1{,}061,\ 0{,}147 < b < 0{,}192.$$

Средняя ошибка аппроксимации $\overline{A} = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \tilde{y}}{y_i} \right| \cdot 100\,\% = 6{,}35\%$ говорит о хорошей точности уравнения регрессии, т.е. свидетельствует о хорошей подгонке модельных (теоретических) данных к исходным данным.

Итак, качество модели является высоким, а значит, уравнение регрессии пригодно для практического использования (для прогнозирования).

Найдем прогнозное значение результативного фактора при значении признака-фактора, составляющем 110% от среднего уровня дохода семьи. Так как $\bar{x} = 8{,}95$, то $x_p = 1{,}1 \cdot \bar{x} = 9{,}845$. Следовательно, необходимо определить прогноз расходов на питание, если душевой доход семьи составляет 9,845 тысяч денежных единиц.

Для вычисления точечного прогноза подставим значение $x_p = 9{,}845$ в полученное уравнение линейной регрессии $\tilde{y} = 0{,}824 + 0{,}169x$. Получим: $y_p = 0{,}824 + 0{,}169 \cdot 9{,}845 \approx 2{,}489$ (тысяч денежных единиц). Значит, если душевой доход семьи составляет 9845 денежных единиц, то расходы на питание будут 2489 денежных единиц.

Для построения доверительного интервала прогноза вычислим стандартную ошибку прогноза m_p по формуле $m_p = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$,

где $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}}$ – стандартная ошибка регрессии. В нашем случае $s = 0{,}1443$. Поэтому $m_p = 0{,}1443 \cdot \sqrt{1 + \frac{1}{8} + \frac{(9{,}845 - 8{,}95)^2}{244{,}42}} \approx 0{,}153$. Теперь нижняя и верхняя границы интервала прогноза при $\alpha = 0{,}05$ и $t_{\text{кр}} = 2{,}447$ определяются по формулам $y_p - t_{\text{кр}} \cdot m_p$ и $y_p + t_{\text{кр}} \cdot m_p$. Окончательно находим, что доверительный интервал прогноза имеет вид $(2{,}114; 2{,}864)$. Таким образом, при душевом доходе, равном 9,845 тысяч денежных единиц, значение расходов на питание с вероятностью 0,95 находится в пределах от 2,114 до 2,864 тысяч денежных единиц.

В заключение на одном графике (рисунок 3.10) изобразим исходные данные и линию регрессии. Из графика видно, что построенная линейная модель достаточно качественно описывает взаимосвязь исследуемых показателей.

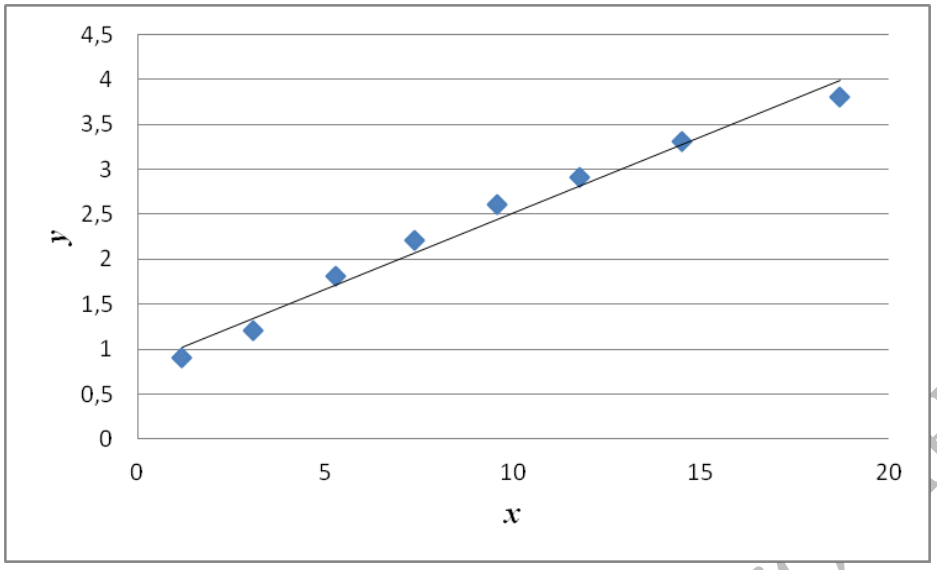


Рис. 3.10. Исходные данные и линия регрессии примера 3.4

Пример 3.5. Применив экспериментальный метод, на основе статистических данных (таблица 3.11) о розничной торговле в Республике Беларусь исследовать зависимость розничного товарооборота y от числа объектов розничной торговой сети x.

Требуется:

1) С помощью вкладки «Мастер диаграмм» по заданным статистическим данным построить точечный график (корреляционное поле) и произвести визуальный анализ эмпирических данных.

2) Среди экспоненциальной, логарифмической, степенной, полиномиальной и линейной регрессионных парных моделей с помощью инструмента «Линия тренда» выбрать уравнение регрессии, которое наилучшим образом соответствует расположению точек корреляционного поля.

3) По выбранному уравнению определить точечный коэффициент эластичности, предполагая, что число объектов сети равно 42 тысячи.

4) По построенной модели выполнить точечный прогноз розничного товарооборота при прогнозном значении числа объектов розничной торговой сети, равном 44 тысячам.

Таблица 3.11. Статистические данные примера 3.5

Год	Число объектов розничной торговой сети, тыс.	Розничный товарооборот в фактически действовавших ценах, млрд руб.
2000	30,8	4197
2001	29,7	8171
2002	29,6	11910

2003	31,1	15170
2004	32,8	19452
2005	34,2	25230
2006	35,4	31062
2007	36,1	38168
2008	41,0	50651
2009	43,4	54736

Решение:

1) С помощью вкладки «Мастер диаграмм» по заданным статистическим данным построим точечный график (рисунок 3.11). Из вида корреляционного поля можно сделать вывод о том, что между результирующим признаком y и фактором x существует прямая зависимость, т.е. с ростом числа объектов розничной торговой сети розничный товарооборот увеличивается.

2) С помощью инструмента «Линия тренда» (методика применения этого инструмента описана в главе 6) построим различные выдыуравнений регрессии, указав индекс детерминации.

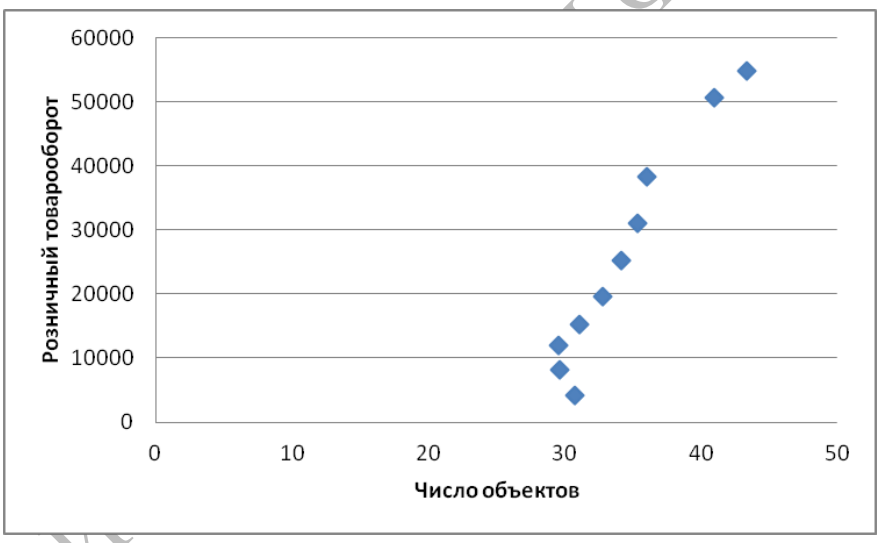


Рис. 3.11. Корреляционное поле примера 3.5

Уравнение экспоненциальной регрессии отображено на рисунке 3.12. Оно имеет вид $\tilde{y} = 115{,}27e^{0{,}1498x}$, при этом $R^2 = 0{,}7323$.

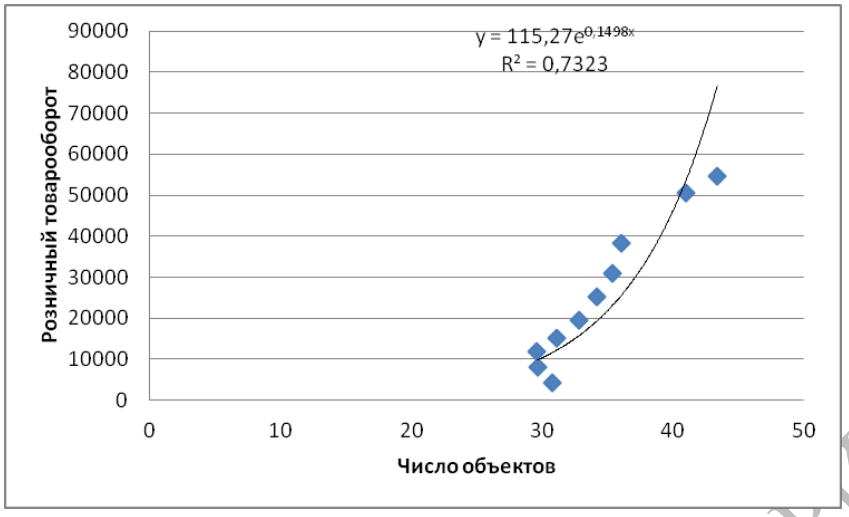


Рис. 3.12. Точечный график с экспоненциальным трендом

Уравнение логарифмической регрессии отображено на рисунке 3.13. Оно имеет вид $\tilde{y} = 129828 \ln x - 432461$, при этом $R^2 = 0{,}9543$.

Уравнение степенной регрессии отображено на рисунке 3.14. Оно имеет вид $\tilde{y} = 0{,}00009x^{5{,}4441}$, при этом $R^2 = 0{,}7557$.

Уравнение полиномиальной регрессии отображено на рисунке 3.15. Оно имеет вид $\tilde{y} = -84{,}675x^2 + 9746{,}4x - 207547$, при этом $R^2 = 0{,}9563$.

Уравнение линейной регрессии отображено на рисунке 3.16. Оно имеет вид $\tilde{y} = 3616{,}3x - 98563$, при этом $R^2 = 0{,}948$.

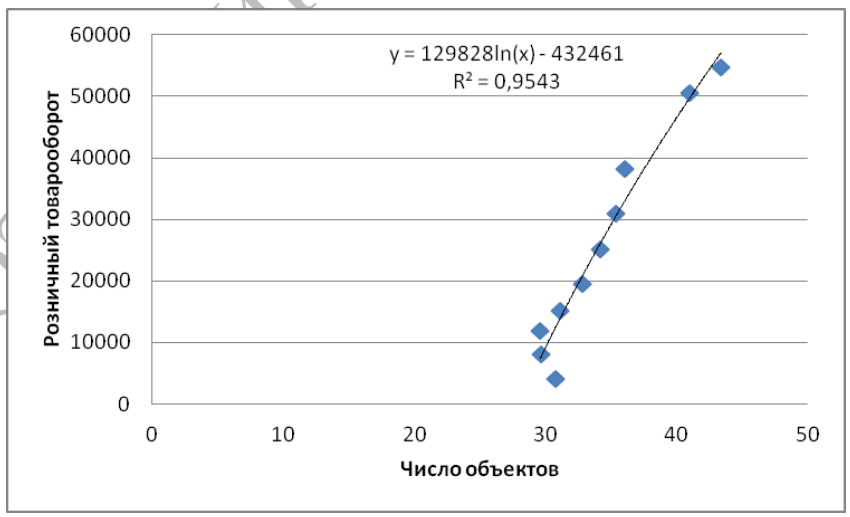


Рис. 3.13. Точечный график с логарифмическим трендом

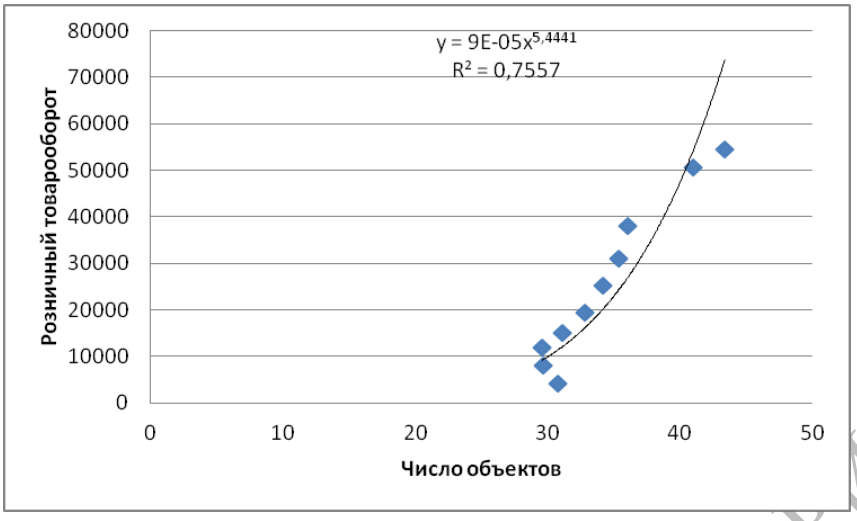


Рис. 3.14. Точечный график со степенным трендом

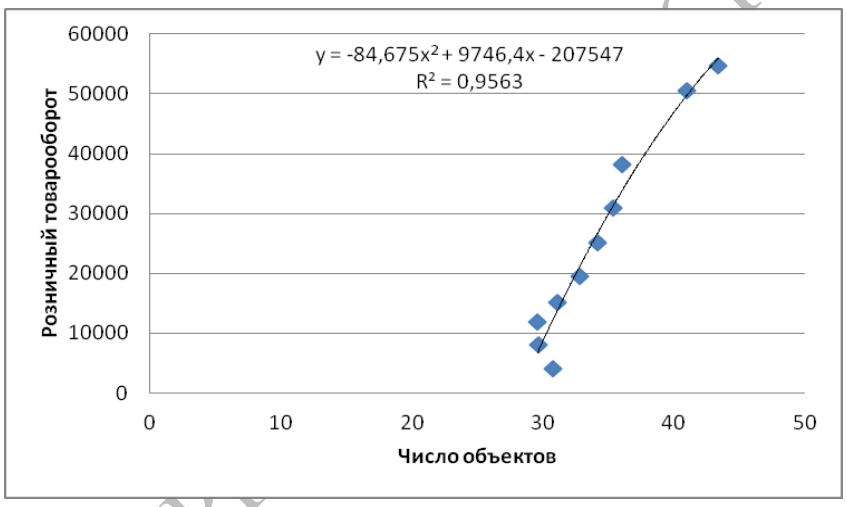


Рис. 3.15. Точечный график с полиномиальным трендом

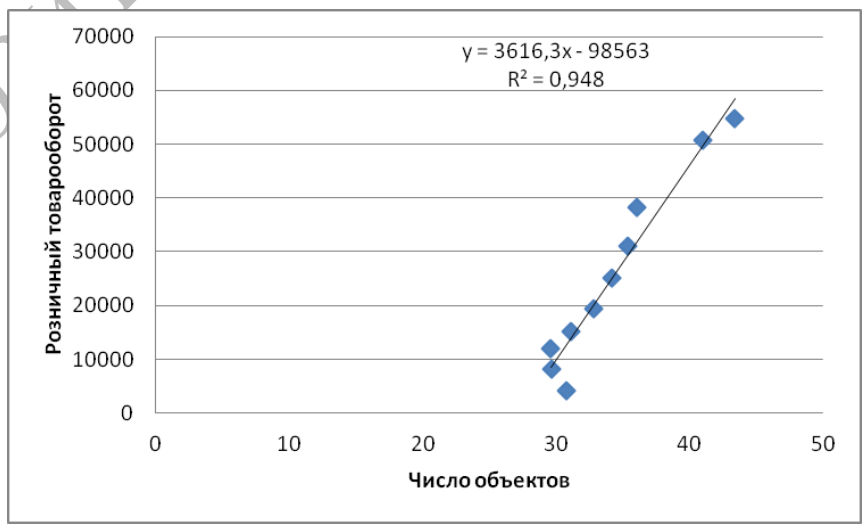


Рис. 3.16. Точечный график с линейным трендом

Сравнение индексов детерминации для всех построенных моделей позволяет сделать вывод о том, что наиболее качественной является полиномиальное уравнение регрессии $\tilde{y} = -84{,}675x^2 + 9746{,}4x - 207547$.

3) Для вычисления точечного коэффициента эластичности воспользуемся формулой $Э_{x_0} = f'(x_0)\dfrac{x_0}{y_0}$, где $f'(x)$ производная функции $f(x)$, $y_0 = f(x_0)$. В нашем случае $f(x) = -84{,}675x^2 + 9746{,}4x - 207547$, $x_0 = 42$. Значит, $f'(x) = -169{,}35x + 9746{,}4$, $y_0 = f(42) = 52435{,}1$, $f'(42) = 2633{,}7$. Поэтому $Э_{42} = 2{,}1$. Следовательно, при изменении x на 1% от уровня $x_0 = 42$ значение y изменится от уровня $f(42) = 52435{,}1$ на 2,1%.

4) Для нахождения по выбранному уравнению регрессии $\tilde{y} = -84{,}675x^2 + 9746{,}4x - 207547$ точечного прогноза розничного товарооборота при прогнозном значении числа объектов розничной торговой сети $x_p = 44$ подставим в уравнение $\tilde{y} = -84{,}675x^2 + 9746{,}4x - 207547$ значение $x_p = 44$. Получим значение, равное 57363,8 млрд руб.

Реализация с помощью ППП *Excel*

Расчет и анализ показателей парной линейной регрессии может быть осуществлен с помощью «Пакета анализа» табличного процессора *Excel*. Основную информацию о линейной модели дает программа «Регрессия». Эта информация содержится в четырех таблицах: «Регрессионная статистика», «Дисперсионный анализ» (две таблицы) и «Вывод остатка».

В таблице «Регрессионная статистика» приводятся значения:

1. **Множественный** R – линейный коэффициент корреляции r_{xy}.
2. R**-квадрат** – коэффициент детерминации R^2.
3. **Нормированный** R^2 – скорректированный R^2 с поправкой на число степеней свободы.
4. **Стандартная ошибка** – стандартная ошибка регрессии s.
5. **Наблюдения** – число наблюдений n.

В первой таблице «Дисперсионный анализ» приведены:

1. Столбец ***df*** – число степеней свободы, равное:

$df = 1$ для строки **Регрессия**;

$df = n - 1$ для строки **Остаток**;

$df = n - 1$ для строки **Итого**.

2. Столбец ***SS*** – сумма квадратов отклонений, равная:

$\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2$ для строки **Регрессия**;

$\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$ для строки **Остаток**;

$\sum_{i=1}^{n}(y_i - \bar{y})^2$ для строки **Итого**.

3. Столбец ***MS*** – дисперсии, определяемые по формуле $\frac{SS}{df}$:

факторная для строки **Регрессия**;
остаточная для строки **Остаток**.

4. Столбец ***F*** **–** наблюдаемое значение *F*-критерия Фишера $F_{\text{набл}}$.

5. Столбец **Значимость *F*** – значение уровня значимости, соответствующее вычисленной *F*-статистике.

Если значимость *F* меньше заданного уровня значимости α, то R^2 статистически значим.

Во второй таблице «Дисперсионный анализ» указаны:

1. **Коэффициенты** – значения коэффициентов *a* и *b*.

2. **Стандартная ошибка** – стандартные ошибки коэффициентов регрессии *a* и *b*.

3. ***t*-статистика** – наблюдаемые значения t-статистики $t_{\text{набл}}$ для коэффициентов регрессии *a* и *b*.

4. ***P*-Значение** – значение уровня значимости, соответствующее вычисленной t-статистике.

Если *P*-значение меньше заданного уровня значимости α, то соответствующий коэффициент регрессии статистически значим.

5. **Нижние 95%** и **Верхние 95%** – нижние и верхние границы 95%-ных доверительных интервалов для коэффициентов уравнения линейной регрессии.

В таблице «Вывод остатка» указаны:

1. **Наблюдение** – номер наблюдения.

2. **Предсказанное** – расчетные (теоретические) значения $\tilde{y}_i$.

3. **Остатки** – разность $y_i - \tilde{y}_i$ между наблюдаемыми и расчетными значениями зависимой переменной.

Методику вычисления ключевых показателей парной линейной регрессии проиллюстрируем на примере следующей задачи (при этом будем опираться на статистические данные, находящиеся в таблице 3.12, и исходить из того, что объем выборки n равен 20).

Задача. Для прогноза возможного объема экспорта на основе ВНП построить и исследовать парную линейную регрессионную модель $\tilde{y} = a + bx + \varepsilon$ зависимости объема экспорта (*y*, усл. ед.) от ВНП (*x*, усл. ед.) Использовать построенную модель для прогноза при x_p=2500.

Требуется:

1) ввести данные;

2) построить корреляционное поле зависимости экспорта y от ВНП x;

3) установить тесноту и вид связи между указанными показателями, т.е. рассчитать ковариацию и корреляцию и проанализировать их;

4) найти точечные и интервальные оценки для коэффициентов регрессии *a* и *b*;

5) оценить коэффициент детерминации и провести анализ общего качества уравнения регрессии;

6) указать стандартную ошибку регрессии;

7) оценить статистическую значимость коэффициентов регрессии *a* и *b* при уровне значимости $\alpha = 0{,}05$, при необходимости получить новое уравнение регрессии со значимыми коэффициентами;

8) выяснить, выполняются ли условия теоремы Гаусса-Маркова; для этого оценить разброс точек на графике остатков, построить гистограмму остатков, проанализировать их числовые характеристики;

9) дать точечный и интервальный прогнозы объема экспорта по заданному значению ВНП.

Результаты вычислений и анализа оформить в виде отчета (*форма отчета прилагается ниже*).

Порядок выполнения:

1) В ячейку A1 введите название ВНП, в ячейку B1 – название Экспорт. В ячейки A2, A3, …, A21 введите данные первого столбца выбранного варианта задания, в ячейки B2, B3, …, B21 – данные второго столбца выбранного варианта.

Присвойте листу 1 название «Исходные данные».

2) На листе «Исходные данные» выполните следующие действия:

– на панели инструментов активизируется кнопка *Мастер диаграмм (шаг 1 из 4)*, в одноименном диалоговом окне (рисунок 3.17) среди стандартных типов выбирается *Точечная* и верхний левый вид диаграммы и нажимается кнопка *Далее>*;

– открывается диалоговое окно *Мастер диаграмм (шаг 2 из 4)*, в котором во вкладке *Диапазон данных* в поле *Диапазон* вводится ссылка на диапазон ячеек A2:B21; нажимается кнопка *Далее>*;

– открывается диалоговое окно *Мастер диаграмм (шаг 3 из 4)*, в котором во вкладке *Заголовки* в поле *Ось X(категорий)* вводится название «ВНП», в поле *Ось Y(значений)* – название «Экспорт»; во вкладке *Легенда* снимается флажок *Добавить легенду* и нажимается кнопка *Далее>*;

– открывается диалоговое окно *Мастер диаграмм (шаг 4 из 4)* в поле *имеющемся* устанавливается флажок и нажимается кнопка *Готово*.

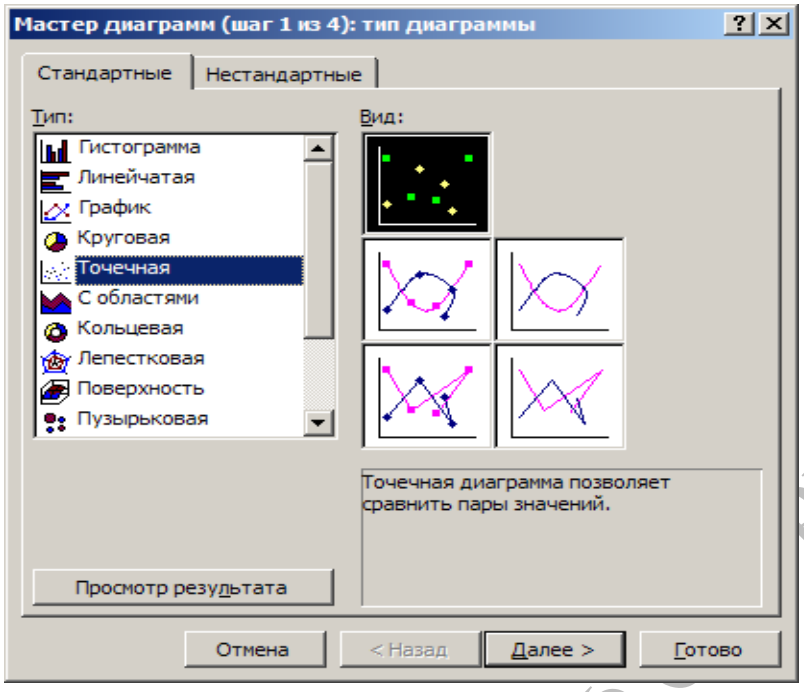


Рис. 3.17. Диалоговое окно «Мастер диаграмм (шаг 1 из 4)»

3) В меню *Сервис* выберите дополнение *Анализ данных*, в предложенных инструментах анализа выделите *Ковариация*, нажмите кнопку *ОК*. Установите значения параметров в появившемся диалоговом окне (рисунок 3.18) следующим образом:

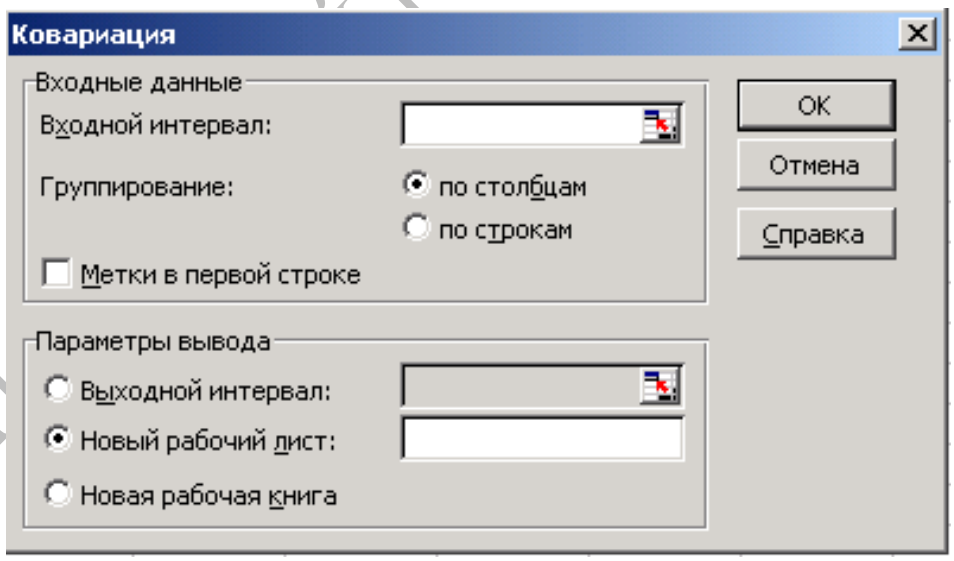


Рис. 3.18. Диалоговое окно «Ковариация»

- *Входной интервал* – введите ссылки на ячейки A1:B21 (курсор установите в поле «Входной интервал», указатель мыши поместите в ячейку

A1, удерживая нажатой левую клавишу, протяните указатель мыши до ячейки B21);

- *Группирование* – флажок *по столбцам* устанавливается автоматически;
- *Метки в первой строке* – установите флажок щелчком левой кнопки мышки;
- *Параметры вывода* – установите флажок на *Новый рабочий лист,* поставив курсор в поле напротив, введите название «Ковариация».

Нажмите *ОК.*

Вернитесь на лист «Исходные данные». В меню *Сервис* выберите опцию *Анализ данных* и выделите *Корреляция*. Установите в диалоговом окне (рисунок 3.19) следующим образом значения параметров:

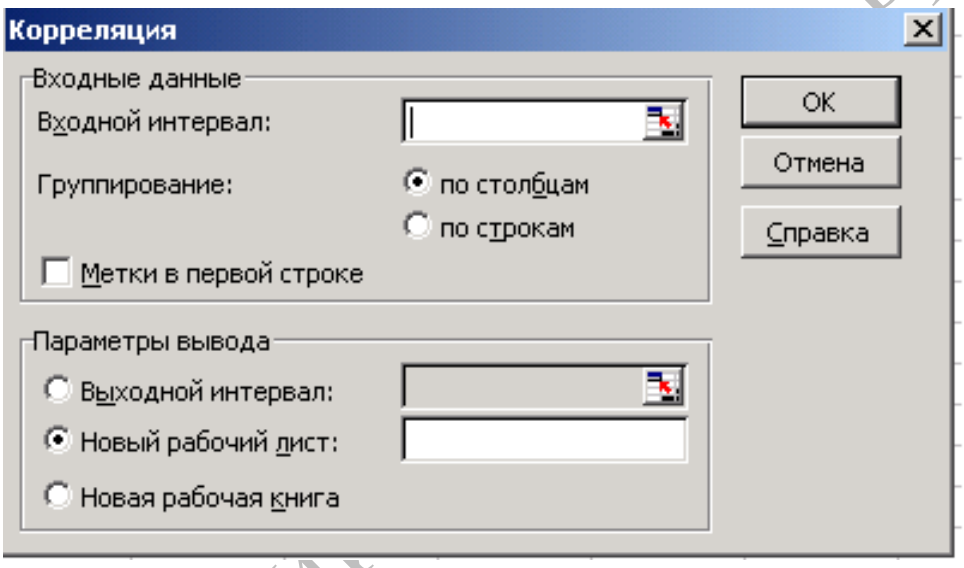


Рис. 3.19. Диалоговое окно «Корреляция»

- *Входной интервал* – введите ссылки на ячейки, содержащие исходные данные A1:B21 (курсор установите в поле «Входной интервал», указатель мыши поместите в ячейку A1, удерживая нажатой левую клавишу, протяните указатель мыши до ячейки B21);
- *Группирование* – установите флажок *по столбцам*;
- *Метки в первой строке* – установите флажок;
- *Параметры вывода* – установите флажок на *Новый рабочий лист,* введите название «Корреляция».

Нажмите *ОК.*

Значение линейного коэффициента корреляции находится на листе «Корреляция» в ячейке B3.

4) Вернитесь на лист «Исходные данные». В меню *Сервис* выберите дополнение *Анализ данных* укажите *Регрессия*. Нажмите кнопку *ОК.*

Установите в диалоговом окне (рисунок 3.20) следующим образом значения параметров:

- *Входной интервал Y* – введите ссылки на ячейки B1:B21;
- *Входной интервал X* – введите ссылки на ячейки A1:A21;
- *Метки* – установите флажок;
- *Уровень надежности* – установите флажок;
- *Константа ноль* – не активизируйте;
- *Параметры вывода* – установите флажок на *Новый рабочий лист* и в поле напротив введите имя «Регрессия»;
- *Остатки* – установите флажок;
- *Стандартизованные остатки* – оставьте пустым;
- *График остатков* – установите флажок;
- *График подбора* – установите флажок;
- *График нормальной вероятности* – оставьте пустым.

Нажмите *ОК*.

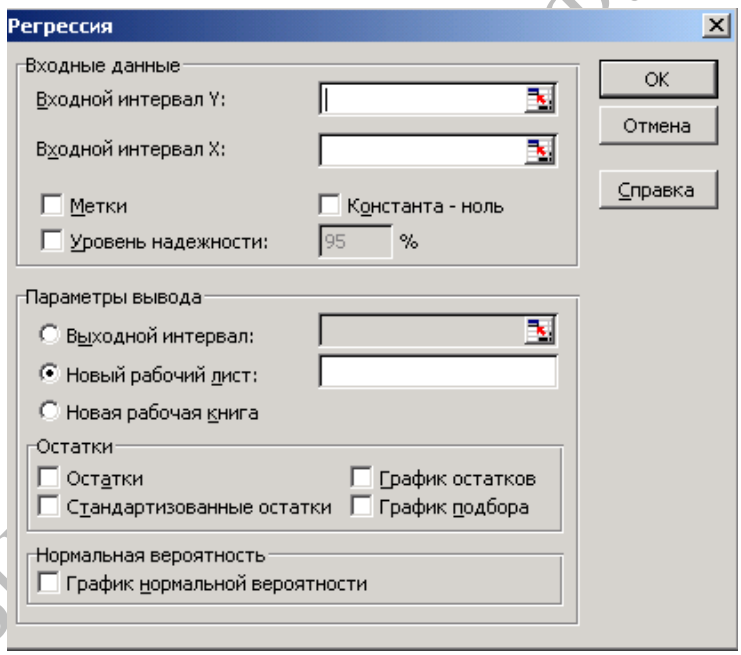


Рис. 3.20. Диалоговое окно «Регрессия»

Расположите диаграммы рядом (на поле диаграммы нажмите левую кнопку мыши, затем поместите курсор на белое поле и при нажатой левой кнопке передвигайте диаграмму вниз) и растяните их (на поле диаграммы нажмите левую кнопку мыши, нижнюю линию границы диаграммы при нажатой левой клавише протяните вниз).

Точечные оценки коэффициентов регрессии *a* и *b* находятся на листе «Регрессия» в ячейках B17 и B18 соответственно. Нижняя и верхняя границы

доверительного интервала вычислены на листе «Регрессия» в ячейках F17 и G17 для коэффициента a и в ячейках F18 и G18 для коэффициента b.

5) Значение коэффициента детерминации R^2 находится на листе «Регрессия» в ячейке B5. Наблюдаемое значение F-критерия Фишера $F_{\text{набл}}$ находится на листе «Регрессия» в ячейке E12.

Вычислите критическое значение $F_{\text{кр}}$ в свободной ячейке E15 следующим образом:

– нажмите на f_x (вставка функций);

– в поле *Категория* окна *Мастер функций* выберите *статистические*, из предложенных ниже функций выделите FРАСПОБР и нажмите *ОК*.

Откроется окно *Аргументы функций*. Заполните поля так:

- *Вероятность* – наберите значение 0,05;
- *Степени свободы 1* – установите курсор в поле и выделите ячейку B12 столбца *df* таблицы «Дисперсионный анализ»;
- *Степени свободы 2* – установите курсор в поле и выделите ячейку B13 столбца *df* таблицы «Дисперсионный анализ».

Нажмите *ОК*.

6) Значение стандартной ошибки регрессии s находятся на листе «Регрессия» в ячейке B7.

7) Наблюдаемые значения t-статистики $t_{\text{набл}}$ коэффициентов регрессии a и b находятся на листе «Регрессия» в ячейках D17 и D18 соответственно.

Вычислите критическое значение $t_{\text{кр}}$ в свободной ячейке D19 следующим образом:

– нажмите на f_x (вставка функций);

– в поле «Категория» окна *Мастер функций* выберите *статистические*, из предложенных ниже функций выделите СТЬЮДРАСПОБР и нажмите *ОК*. Откроется окно «Аргументы функций». Заполните поля:

- *Вероятность* – наберите значение 0,05;
- *Степени свободы* – введите 20–1–1, где 20 – число наблюдений, 1 – число факторов (x) в уравнении регрессии, 1 – число свободных членов (a) в уравнении регрессии.

Нажмите *ОК*.

8) На листе регрессия в меню *Сервис* выберите *Анализ данных*, укажите *Гистограмма*. Нажмите кнопку *ОК*. Значения параметров в появившемся диалоговом окне (рисунок 3.21) установите следующим образом:

- *Входной интервал* – введите ссылки на ячейки C24:C44 (столбец *Остатки* с названием);
- *Интервал карманов* – не заполняйте;
- *Метки* – установите флажок;
- *Выходной диапазон* – введите ссылку на новый рабочий лист «Гистограмма»;

- *Парето* – оставьте пустым;
- *Интегральный процент* – оставьте пустым;
- *Вывод графика* – установите флажок.

Нажмите *ОК*. Растяните диаграмму вниз.

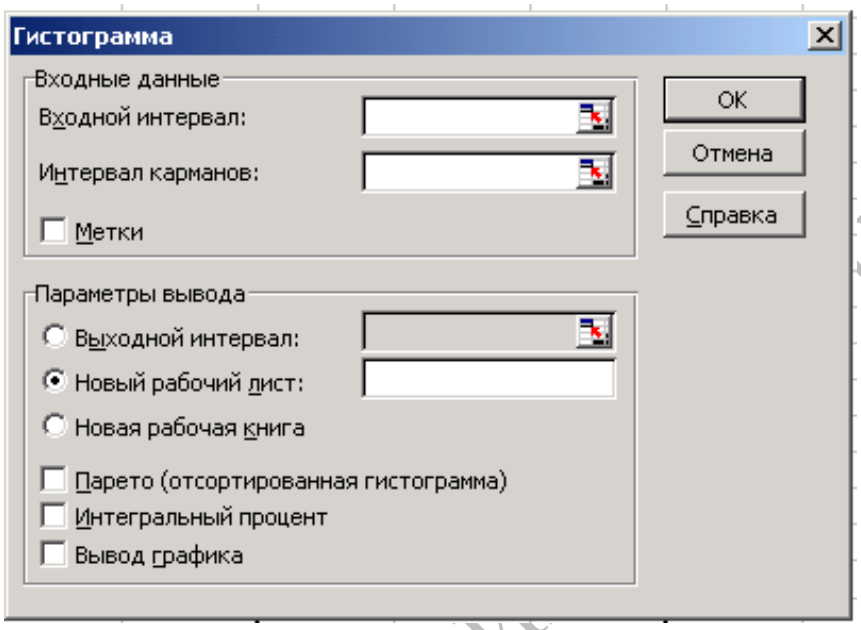


Рис. 3.21. Диалоговое окно «Гистограмма»

Вернитесь на лист «Регрессия». Выберите в опциях меню *Сервис* → *Анализ данных* → *Описательная статистика*, нажмите *ОК*. Значения параметров в диалоговом окне (рисунок 3.22) установите следующим образом:

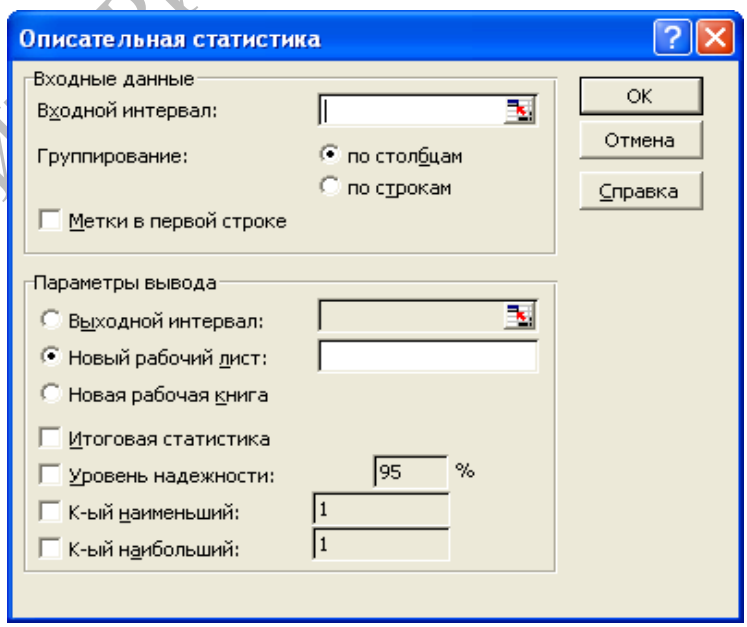


Рис. 3.22. Диалоговое окно «Описательная статистика»

- *Входной интервал* – введите ссылки на ячейки C24:C44 (столбец *Остатки* с названием);
- *Группирование* – установите флажок *по столбцам*;
- *Метки* – установите флажок *в первой строке*;
- *Выходной диапазон* – установите флажок на *Новый рабочий лист* и в поле напротив введите название «Статистика остатков»;
- установите флажки *Итоговая статистика*, уровень надежности (95%).

Нажмите *ОК*.

9) Вернитесь на лист «Регрессия» и в пустой ячейке E22 листа введите формулу

= B17+B18*2500 – точечный прогноз.

На листе «Регрессия» в пустой ячейке E23 вычислите значение m_p по формуле $m_p = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$, где s – стандартная ошибка регрессии.

В пустых ячейках E24 и F24 введите формулы

=E22–D19*E23 – левый конец интервала прогноза;

=E22+D19*E23 – правый конец интервала прогноза.

Приложение: Отчет о результатах вычислений и анализа

1. *Постановочный этап*

Из экономической теории известно, что *Экспорт* зависит от *ВНП* и от многих других факторов. Выделим один фактор – *ВНП*, считая его наиболее существенным. Он является объясняющим фактором для результативного (объясняемого) фактора *Экспорт*. Возникает задача количественного описания зависимости указанных экономических показателей уравнением парной регрессии.

2. *Спецификация модели*

Вид регрессии визуально определяется по корреляционному полю, которое изображено на графике, построенном с помощью вкладки «Мастер диаграмм» по данным 20 наблюдений. Так как точки сгруппированы вдоль прямой (не горизонтальной), то можно предположить, что зависимость экспорта y от ВНП x описывается парной линейной регрессионной моделью $y = a + bx + \varepsilon$, где a, b – неизвестные параметры модели, ε – случайный член, который включает в себя суммарное влияние всех неучтенных в модели факторов.

Выборочная ковариация равна ....., поэтому зависимость (*прямая или обратная*).

Линейный коэффициент корреляции равен ..... Так как он (*больше, меньше*) нуля, то зависимость (*прямая, обратная*). Вывод о силе линейной зависимости определяется по шкале Чеддока. Линейная зависимость (*слабая, умеренная, заметная, высокая, весьма высокая*).

3. *Параметризация модели*

Для оценки параметров уравнения парной регрессии применяется метод наименьших квадратов (МНК). В результате проведения регрессионного анализа получены точечные и интервальные оценки неизвестных параметров.

Точечная оценка параметра a равна ..... Доверительный интервал для параметра a имеет вид (..... , ...).

Точечная оценка параметра b равна ..... Доверительный интервал для параметра b имеет вид (..... , ...).

Таким образом, уравнение регрессии имеет вид (*записать уравнение линейной регрессии*).

4. *Верификация модели*

4.1. Значимость коэффициентов регрессии оценивается с помощью t-статистики.

Для коэффициента a наблюдаемое значение статистики $t_{набл}$ равно ..... Критическое значение $t_{кр}$ равно ..... Так как $|t_{набл}|$ (*больше, меньше*) $t_{кр}$, то коэффициент a (*значим или незначим*).

Для коэффициента b наблюдаемое значение статистики $t_{набл}$ равно ..... Критическое значение $t_{кр}$ равно ..... Так как $|t_{набл}|$ (*больше, меньше*) $t_{кр}$, то коэффициент b (*значим или незначим*).

4.2. Качество построенной модели в целом оценивает коэффициент детерминации. В таблице «Регрессионная статистика» листа «Регрессия» коэффициент детерминации R-квадрат равен ..... *Сделать вывод об общем качестве уравнения.*

Значимость коэффициента детерминации R-квадрат устанавливается с помощью критерия Фишера в таблице «Дисперсионный анализ» листа «Регрессия». Наблюдаемое значение $F_{набл}$ равно ..... Критическое значение $F_{кр}$ равно ..... Так как наблюдаемое значение $F_{набл}$ (*больше, меньше*) $F_{кр}$, то R-квадрат (*значим или незначим). Сделать вывод об общем качестве уравнения.*

4.3. Для того, чтобы оценки параметров линейного уравнения регрессии были несмещенными, состоятельными и эффективными, необходимо выполнение условий Гаусса–Маркова.

4.3.1. Значение «Среднее» из таблицы на листе «Статистика остатков» равно ..... Оно является несмещенной оценкой математического ожидания случайного члена. *Сделать вывод о выполнении предпосылки* 1.

4.3.2. Если на графике остатков листа «Регрессия» точки разбросаны в полосе, то условие 2 Гаусса-Маркова выполняется. *Сделать вывод о выполнении предпосылки.*

4.3.3. Если на графике остатков листа «Регрессия» точки разбросаны возле прямой $y = 0$ хаотично без видимой закономерности, то зависимости между остатками не наблюдается. В этом случае условие 3 Гаусса-Маркова выполняется. *Сделать вывод о выполнении предпосылки*.

4.4.4. *Сделать вывод о нормальности распределения остатков по визуальному анализу гистограммы*.

5. *Прогнозирование*

Если выполняются все условия верификации, то модель является качественной. В противном случае ее надо усовершенствовать: либо на этапе спецификации, либо варьировать выборку. По качественной модели можно прогнозировать объем экспорта по объему ВНП. *Сделать вывод о возможности прогнозирования*.

Точечный прогноз экспорта равен ....., доверительный интервал прогноза имеет вид (....., ...), где центр интервала равен точечному прогнозу, концы интервала получены прибавлением и вычитанием произведения стандартной ошибки прогноза на критическое значение *t*-статистики. *Сделать вывод о качестве прогноза*.

Интегрированные задачи

Задача 3.1. С помощью «Пакета анализа» табличного процессора Excel построить линейную модель и проанализировать зависимость объема экспорта (y, усл. ед.) от ВНП (x, усл. ед.), опираясь на данные, находящиеся в таблице 3.12.

Таблица 3.12. Статистические данные задачи 3.1

1 вариант		2 вариант		3 вариант	
ВНП	экспорт	ВНП	экспорт	ВНП	экспорт
1000	190	1030	120	1450	120
1090	220	1090	150	1570	150
1150	240	1120	170	1630	170
1230	240	1250	180	1850	180
1300	260	1300	210	2034	210
1360	250	1340	210	2170	220
1400	280	1380	220	2250	200
1470	290	1400	250	2310	230
1500	310	1450	290	2810	250
1580	350	1500	310	3000	280
1600	340	1560	300	3064	290
1630	360	1600	330	3200	300
1700	380	1620	310	3300	310
1780	400	1700	350	3500	310
1800	420	1710	340	3800	320
1850	400	1820	380	4000	380
1910	400	1890	400	4100	390

1990	440	1900	400	4080	400
2010	450	1980	420	4120	400
2100	470	2000	430	4200	450

4 вариант		5 вариант		6 вариант	
ВНП	экспорт	ВНП	экспорт	ВНП	экспорт
1010	180	900	80	1100	200
1080	200	980	105	1190	230
1100	230	1050	120	1250	250
1220	230	1140	130	1330	250
1290	250	1200	135	1400	270
1350	230	1250	150	1460	260
1390	260	1300	180	1500	290
1400	280	1360	190	1570	300
1450	290	1400	210	1600	310
1500	300	1480	260	1660	360
1590	340	1500	265	1700	380
1600	330	1590	280	1740	400
1650	350	1610	290	1800	400
1710	370	1690	310	1860	410
1790	390	1710	330	1900	420
1800	400	1790	360	1960	440
1890	410	1800	370	2020	430
1900	400	1840	380	2100	450
1950	410	1900	400	2200	460
1990	420	1910	420	2300	450

7 вариант		8 вариант		9 вариант	
ВНП	экспорт	ВНП	экспорт	ВНП	экспорт
800	100	1000	130	920	200
890	130	1100	150	980	240
910	160	1190	170	1000	280
950	170	1210	180	1190	290
1000	175	1290	185	1200	310
1100	180	1320	200	1240	350
1180	190	1360	220	1290	350
1220	200	1400	240	1320	380
1290	220	1420	250	1380	410
1310	230	1480	240	1410	430
1380	250	1510	260	1420	470
1420	240	1590	260	1500	460
1490	290	1610	280	1520	490
1520	300	1680	300	1590	500
1590	330	1720	310	1610	520
1600	350	1780	330	1630	540

1640	380	1810	350	1700	550
1680	390	1890	350	1740	590
1720	400	1920	380	1800	600
1800	410	2000	400	1820	610

Расчеты провести в соответствии с требованиями и описанием задачи, изложенной в разделе «Реализация с помощью ППП *Excel*». Рассчитать точечный прогноз при x_p=2500. Определить доверительный интервал прогноза для уровня значимости $\alpha = 0{,}05$.

Результаты вычислений и анализа представить в виде отчета по форме, предложенной выше.

Задача 3.2. Для прогноза возможного объема экспорта на основе ВНП построить нелинейную парную регрессионную модель. При этом использовать данные, находящиеся в таблице 3.13 с вариантами заданий.

С помощью вкладки «Мастер диаграмм» по заданным статистическим данным построить точечный график (корреляционное поле) и произвести визуальный анализ эмпирических данных. Среди экспоненциальной, логарифмической, степенной и полиномиальной регрессионных парных моделей с помощью инструмента «Линия тренда» выбрать уравнение регрессии, которое наилучшим образом соответствует расположению точек корреляционного поля.

Таблица 3.13. Статистические данные задачи 3.2

1 вариант $V = 2000 \;\; P = 1900$		2 вариант $V = 1950 \;\; P = 2100$		3 вариант $V = 4150 \;\; P = 1800$	
ВНП	экспорт	ВНП	экспорт	ВНП	экспорт
1000	190	1030	120	1450	120
1090	220	1090	150	1570	150
1150	240	1120	170	1630	170
1230	240	1250	180	1850	180
1300	260	1300	210	2034	210
1360	250	1340	210	2170	220
1400	280	1380	220	2250	200
1470	290	1400	250	2310	230
1500	310	1450	290	2810	250
1580	350	1500	310	3000	280
1600	340	1560	300	3064	290
1630	360	1600	330	3200	300
1700	380	1620	310	3300	310
1780	400	1700	350	3500	310
1800	420	1710	340	3800	320
1850	400	1820	380	4000	380

1910	400	1890	400	4100	390
1990	440	1900	400	4080	400
2010	450	1980	420	4120	400
2100	470	2000	430	4200	450

4 вариант $V = 2000$ $P = 2050$		5 вариант $V = 1950$ $P = 2000$		6 вариант $V = 2250$ $P = 2000$	
ВНП	экспорт	ВНП	экспорт	ВНП	экспорт
1010	180	900	80	1100	200
1080	200	980	105	1190	230
1100	230	1050	120	1250	250
1220	230	1140	130	1330	250
1290	250	1200	135	1400	270
1350	230	1250	150	1460	260
1390	260	1300	180	1500	290
1400	280	1360	190	1570	300
1450	290	1400	210	1600	310
1500	300	1480	260	1660	360
1590	340	1500	265	1700	380
1600	330	1590	280	1740	400
1650	350	1610	290	1800	400
1710	370	1690	310	1860	410
1790	390	1710	330	1900	420
1800	400	1790	360	1960	440
1890	410	1800	370	2020	430
1900	400	1840	380	2100	450
1950	410	1900	400	2200	460
1990	420	1910	420	2300	450

7 вариант $V = 1750$ $P = 1850$		8 вариант $V = 1950$ $P = 1900$		9 вариант $V = 1750$ $P = 1780$	
ВНП	экспорт	ВНП	экспорт	ВНП	экспорт
800	100	1000	130	920	200
890	130	1100	150	980	240
910	160	1190	170	1000	280
950	170	1210	180	1190	290
1000	175	1290	185	1200	310
1100	180	1320	200	1240	350
1180	190	1360	220	1290	350
1220	200	1400	240	1320	380
1290	220	1420	250	1380	410
1310	230	1480	240	1410	430
1380	250	1510	260	1420	470
1420	240	1590	260	1500	460
1490	290	1610	280	1520	490
1520	300	1680	300	1590	500
1590	330	1720	310	1610	520

1600	350	1780	330	1630	540
1640	380	1810	350	1700	550
1680	390	1890	350	1740	590
1720	400	1920	380	1800	600
1800	410	2000	400	1820	610

Для выбранной формы требуется:

1) провести процедуру линеаризации и оценить ее параметры, построить парную регрессионную модель;

2) вычислить индекс корреляции, оценить характер регрессии и тесноту связи между факторами;

3) вычислить индекс детерминации и оценить общее качество уравнения парной регрессии;

4) вычислить среднюю ошибку аппроксимации и оценить точность уравнения регрессии;

5) определить точечный коэффициент эластичности, предполагая, что ВНП равно V (задано в таблице 3.13).

6) выполнить точечный прогноз экспорта при прогнозном значении ВНП, равном P (задано в таблице 3.13).

Контрольные задания

Задание 3.1. Построены парные модели:

1) $y = a + bx^3 + \varepsilon$,

2) $y^a = b + cx^2 + \varepsilon$,

3) $y = a + b\ln x + \varepsilon$,

4) $y = 1 + a\left(1 - x^b\right) + \varepsilon$,

5) $\ln y = a + b\ln x + \varepsilon$,

6) $y = a + b\dfrac{x}{10} + \varepsilon$,

7) $y = a + bx^c + \varepsilon$.

Определить, какие из представленных выше моделей линейны по переменным, линейны по параметрам, не линейны ни по переменным, ни по параметрам.

Задание 3.2. Зависимость среднемесячной производительности труда от возраста рабочих характеризуется моделью $y = a + bx + cx^2 + \varepsilon$. Её использование привело к результатам, представленным в таблице 3.14.

Таблица 3.14. Статистические данные задания 3.2

№ п/п	Производительность труда, y	№ п/п	Производительность труда, y

	фактическая	расчетная		фактическая	расчетная
1	12	10	6	11	12
2	8	10	7	12	13
3	13	13	8	9	10
4	15	14	9	11	10
5	16	15	10	9	9

Оценить качество модели, определив ошибку аппроксимации и индекс корреляции.

Задание 3.3. В таблице 3.15 представлена информация о материалоемкости y (потребление материалов на единицу продукции, кг) и объеме x выпуска продукции (тыс. ед.) по 10 однородным заводам. Построить парную гиперболическую модель зависимости y от x.

Таблица 3.15. Статистические данные задания 3.3

	Значение показателей заводам									
	1	2	3	4	5	6	7	8	8	10
y	9	6	5	4	3,7	3,6	3,5	6	7	3,5
x	100	200	300	400	500	600	700	150	120	250

1) Найти параметры модели $y = a + \frac{b}{x} + \varepsilon$.

2) Оценить тесноту связи с помощью индекса корреляции.

3) Сделать вывод об общем качестве уравнения регрессии.

Задание 3.4. В таблице 3.16 приведены данные о среднемесячной начисленной заработной плате x и доле y денежных доходов, направленных на прирост сбережений во вкладах, займах, сертификатах и на покупку валюты, в общей сумме среднедушевого денежного дохода.

Таблица 3.16. Статистические данные задания 3.4

Доля денежных доходов, направленных на прирост сбережений во вкладах, займах, сертификатах и на покупку валюты, в общей сумме среднедушевого денежного дохода, % , y	Среднемесячная начисленная заработная плата, x
6,9	289
8,7	334
6,4	300
8,4	343
6,1	356
9,4	289

11,0	341
6,4	327
9,3	357
8,2	352
8,6	381

1) Построить корреляционное поле и визуально оценить форму связи между переменными.

2) Рассчитать параметры уравнений линейной, степенной, экспоненциальной, гиперболической парной регрессии.

3) Оценить тесноту связи с помощью показателей корреляции и детерминации.

4) Дать с помощью среднего коэффициента эластичности сравнительную оценку силы связи фактора с результатом.

5) Оценить с помощью средней ошибки аппроксимации качество уравнений.

6) Оценить с помощью F-критерия Фишера статистическую надежность результатов моделирования. По значениям вычисленных характеристик выбрать лучшее уравнение регрессии и дать его обоснование.

7) Рассчитать прогнозное значение результата, если прогнозное значение фактора увеличится на 10% от его среднего уровня. Определить доверительный интервал прогноза для уровня значимости $\alpha = 0{,}05$.

Задание 3.5. В таблице 3.17 приведены статистические данные о среднем размере назначенных пенсий и прожиточном минимуме.

Таблица 3.17. Статистические данные задания 3.5

Средний размер назначенных ежемесячных пенсий, y	Прожиточный минимум в среднем на одного пенсионера в месяц, x
240	178
223	202
221	197
226	201
220	189
250	302
237	215
232	166
215	199
220	180
222	181
231	186
229	250

1) Построить корреляционное поле и визуально оценить форму связи между переменными.

2) Рассчитать параметры уравнений линейной, степенной, экспоненциальной, гиперболической парной регрессии.

3) Оценить тесноту связи с помощью показателей корреляции и детерминации.

4) Оценить с помощью средней ошибки аппроксимации качество уравнений.

5) С помощью F-критерий Фишера оценить статистическую надежность результатов моделирования. По значениям вычисленных характеристик выбрать лучшее уравнение регрессии и дать его обоснование.

6) Рассчитать прогнозное значение результата, если прогнозное значение фактора увеличится на 20% от его среднего уровня. Определить доверительный интервал прогноза для уровня значимости $\alpha = 0{,}05$.

Задание 3.6. В таблице 3.18 приведены данные об уровне механизации работ x и производительности труда y для 14 однотипных предприятий.

Таблица 3.18. Статистические данные задания 3.6

x_i	32	30	36	40	41	47	56	54	60	55	61	67	69	76
y_i	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Необходимо:

1) оценить тесноту и направление связи между переменными с помощью линейного коэффициента корреляции;

2) найти параметры уравнения линейной регрессии.

Задание 3.7. По статистическим данным задания 3.6:

1) определить параметры уравнения линейной регрессии;

2) найти коэффициент детерминации R^2 и пояснить его смысл;

3) проверить значимость уравнения с помощью F-критерия;

4) оценить точечный и интервальный прогнозы на предприятиях с уровнем механизации работ 60%.

Задание 3.8. По статистическим данным таблицы 3.19 построить парную линейную модель, отражающую зависимость удельного веса y бракованной продукции от доли x рабочих со специальной подготовкой. Оценить статистическую значимость коэффициентов уравнения.

Таблица 3.19. Статистические данные задания 3.8

Удельная доля рабочих со специальной подготовкой, %, x	15,1	20,2	30,4	40,3	45,4	55,1	60,6	70,8
Удельный вес бракованной продукции, %, y	18,6	14,7	11,3	9,5	8,4	6,3	5,5	3,6

Задание 3.9. По статистическим данным (таблица 3.20), описывающим зависимость уровня рентабельности на предприятии от скорости товарооборота, построить уравнение парной линейной регрессии. Определить общее качество и статистическую значимость уравнения.

Таблица 3.20. Статистические данные задания 3.9

Число оборотов, x	Уровень рентабельности, %, y
5,49	0,78
4,68	0,38
4,66	0,21
4,53	0,51
4,56	0,95
6,02	1,05
5,72	0,83
5,43	0,99

Задание 3.10. Имеются данные (таблица 3.21) о расходах на питание y и душевом доходе x для девяти групп семей.

Таблица 3.21. Статистические данные задания 3.10

x	y
64	42
159	60
262	89
365	109
470	125
590	145
725	163
930	194
1850	239

Необходимо:

1) построить корреляционное поле и визуально оценить форму связи между переменными;

2) построить уравнение парной линейной регрессии;

3) оценить значимость коэффициентов полученной модели;

4) оценить общее качество модели;

5) осуществить точечный и интервальный прогнозы при условии, что $x = 1050$.

Задание 3.11. Имеются данные (таблица 3.22) по 18 сельскохозяйственным предприятиям.

Таблица 3.22. Статистические данные задания 3.11

Номер хозяйства	Качество земли, балл	Урожайность, ц/га
1	32	19,5
2	33	19
3	35	20,5
4	37	21
5	38	20,8
6	39	21,4
7	40	23
8	41	23,3
9	42	24
10	44	24,5
11	45	24,2
12	46	25
13	47	27
14	49	26,8
15	50	27,2
16	52	28
17	54	30
18	55	30,2

Необходимо:

1) найти коэффициент корреляции между урожайностью зерновых культур и качеством земли;

2) построить уравнение линейной регрессии, которое характеризует зависимость между качеством земли и урожайностью;

3) оценить качество построенной модели;

4) осуществить точечный и интервальный прогнозы урожайности зерновых культур, если качество земли 48 баллов.

Задание 3.12. По статистическим данным, представленным в таблице 3.23, были построены следующие регрессионные модели:

1) линейная $\tilde{y} = 6{,}07 - 0{,}085x$;

2) параболическая $\tilde{y} = -2{,}017 + 3{,}957x - 0{,}367x^2$;

3) экспоненциальная $\tilde{y} = 5{,}918e^{-0{,}043x}$.

Таблица 3.23. Статистические данные задания 3.12

x	3	8	5	10	7	6	4	9	1	2
y	6	5	9	1	8	9	8	4	2	4

Оценить каждую модель, определив для нее индекс детерминации и коэффициент аппроксимации. Дать интерпретацию рассчитанных характеристик и выбрать лучшую модель. Рассчитать точечный прогноз результативного признака y по лучшей модели, если $x = 4$.

Задание 3.13. По статистическим данным, представленным в таблице 3.24, построить линейную модель зависимости объема выпуска продукции y от величины основных фондов x. Оценить качество построенной модели и осуществить точечный и интервальный прогнозы при $x = 24$.

Таблица 3.24. Статистические данные задания 3.13

x	10	12	15	18	20	22	25	28	30
y	2	5	8	12	14	16	20	24	28

Задание 3.14. На основе статистических данных, приведенных в таблице 3.25, необходимо:

1) Построить уравнение линейной парной регрессии между жилой площадью квартиры x и ее ценой y. Вычислить линейный коэффициент корреляции и коэффициент детерминации. Сделать выводы.

2) Вычислить коэффициенты регрессии и оценить их статистическую значимость (на уровне 0,05). Изложить экономическую интерпретацию коэффициентов регрессии.

3) Осуществить точечный и интервальный прогнозы цены квартиры, если ее площадь составляет 65 квадратных метров.

Таблица 3.25. Статистические данные задания 3.14

x, кв.м	39,0	68,4	34,8	39,0	54,7	74,7	71,7	74,5
y, тыс. долларов	15,9	27,0	13,5	15,1	21,1	28,7	27,2	28,3
x, кв.м	40,0	53,0	86,0	98,0	62,6	45,3	56,4	37,0
y, тыс. долларов	22,0	28,0	45,0	51,0	34,4	24,7	30,8	15,9
x, кв.м	67,5	37,0	69,0	40,0	69,1	68,1	75,3	83,7
y, тыс. долларов	29,0	15,4	28,6	15,6	27,7	34,1	37,7	41,9

Контрольные вопросы

1. Что понимается под парной регрессией?

2. Сформулируйте общую постановку задачи парного эконометрического моделирования.

3. Перечислите основные классификационные признаки парных моделей.

4. Какой вид имеет уравнение парной линейной регрессии?

5. Приведите примеры использования линейных и нелинейных парных регрессионных моделей в экономике.

6. Как определяется классическая линейная нормальная парная модель?

7. Сформулируйте и поясните условия Гаусса-Маркова.

8. Как классифицируются нелинейные парные модели?

10. В чем заключается спецификация парной регрессионной модели?

11. Какие методы применяются для выбора вида модели парной регрессии?

12. В чем заключается графический метод спецификации парной модели?

13. Как строится корреляционное поле?

14. В чем заключается экспериментальный метод спецификации парной модели?

15. В чем заключается суть параметризации парной линейной модели?

16. В чем сущность метода наименьших квадратов?

17. Как вычисляются коэффициенты парной линейной регрессии?

18. В чем заключается экономическая интерпретация коэффициентов парной линейной регрессии?

19. Какими свойствами обладают оценки параметров парной линейной модели, полученные с помощью МНК, если модель является классической?

20. Как с помощью МНК найти параметры нелинейной парной регрессии?

21. В чем состоит процедура линеаризации парной модели?

22. Какая величина называется выборочной ковариацией? Что она характеризует?

23. По какой формуле вычисляется линейный коэффициент корреляции?

24. Как по линейному коэффициенту корреляции оценить тесноту линейной связи между факторами?

25. Как линейный коэффициент корреляции связан с коэффициентом b линейной регрессии?

26. Какие показатели применяются для оценки тесноты взаимосвязи результативного признака с фактором в случае нелинейной парной регрессии?

27. Как связаны между собой общая, факторная и остаточная дисперсии?

28. Как вычисляется индекс корреляции?

29. Как вычисляется и что оценивает коэффициент эластичности?

30. С помощью каких показателей осуществляется оценка общего качества уравнения парной регрессии?

31. Как вычисляется и что оценивает показатель средней ошибки аппроксимации?

32. Как вычисляется и что оценивает стандартная ошибка регрессии?

33. С помощью какого критерия и как проверяется значимость уравнения регрессии в целом?

34. С помощью какого критерия и как осуществляется проверка значимости коэффициентов линейной регрессии?

35. Как строятся доверительные интервалы коэффициентов линейной регрессии?

36. Как по парной модели осуществляется точечный прогноз?

37. В чем преимущества интервального прогноза перед точечным?

38. Как по парной линейной модели осуществляется интервальный прогноз?

Тестовые задания

Выберите правильные ответы из предложенных вариантов:

1. В зависимости от характера парной регрессии различают:

а) прямую регрессию и обратную регрессию;

б) линейную регрессию и нелинейную регрессию;

в) непосредственную регрессию и косвенную регрессию.

2. Графический метод спецификации парной модели заключается:

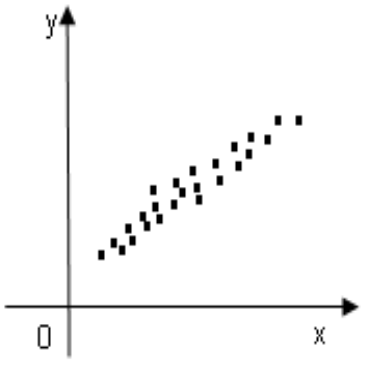
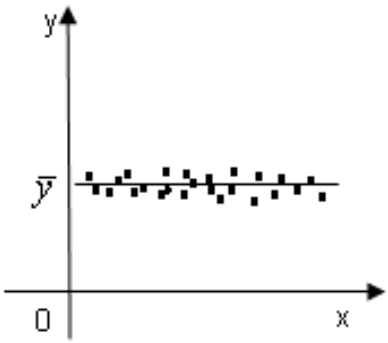
а) в построении и анализе корреляционного поля;

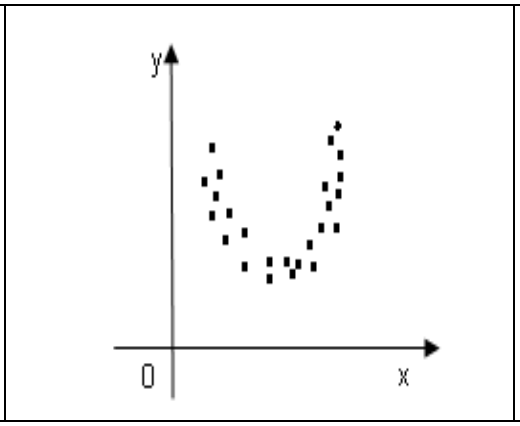
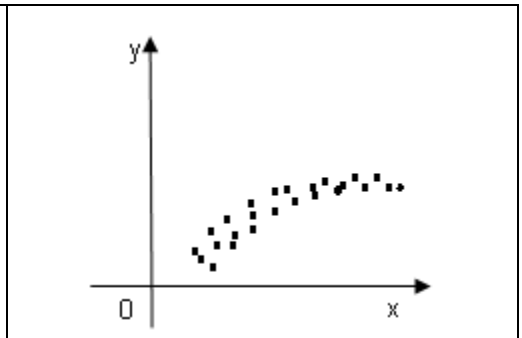
б) в построении нескольких моделей и выборе наиболее качественной;

в) в анализе изучаемой взаимосвязи между признаками;

г) в обращении к экономической теории.

3. Какая из диаграмм рассеивания отражает прямую зависимость между факторами x и y:

а)		в)	

б)		г)	

4. Какое из уравнений задает обратную регрессию:

а) $\tilde{y} = -2x$;

б) $\tilde{y} = \frac{1}{x}$;

в) $\tilde{y} = x^2 - 6x + 9$;

г) $\tilde{y} = \log_2 x$.

5. Уравнение парной линейной регрессии имеет вид:

а) $\tilde{y} = a + \frac{b}{x}$;

б) $\tilde{y} = a + bx$;

в) $y = a + bx + \varepsilon$;

г) $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$.

6. Среди предложенных моделей выделить линейные парные регрессионные модели:

а) $y = 1 - 2x + \varepsilon$;

б) $y = 2x + 3z + \varepsilon$;

в) $y = 2x$;

г) $y = 2x^2 + \varepsilon$.

7. Экспериментальный метод спецификации парной модели заключается:

а) в построении корреляционного поля;

б) в построении нескольких моделей и выборе наиболее качественной;

в) в анализе изучаемой взаимосвязи;

г) в обращении к экономической теории.

8. Метод наименьших квадратов используется для оценивания:

а) величины коэффициента детерминации;

б) параметров линейной регрессии;

в) величины коэффициента корреляции;

г) средней ошибки аппроксимации.

9. Основная идея МНК для построения уравнения регрессии заключается в том, что:

а) сумма квадратов остатков минимизируется;

б) сумма остатков минимизируется;

в) сумма квадратов остатков максимизируется;

г) сумма остатков максимизируется.

10. Рассчитывать параметры парной линейной регрессии можно, если имеется:

а) не менее 5 наблюдений;

б) не менее 7 наблюдений;

в) не менее 10 наблюдений.

11. На основании наблюдений за 50 семьями построено уравнение регрессии $\tilde{y} = 136{,}9 + 0{,}753x$, где y – потребление, x – доход. Соответствуют ли знаки и значения коэффициентов регрессии теоретическим представлениям?

а) да;

б) нет;

в) ничего определенного сказать нельзя.

12. Связь между факторами x и y обратная:

а) при $\mathrm{cov}(x, y) > 0$;

б) при $\mathrm{cov}(x, y) < 0$;

в) при $\mathrm{cov}(x, y) \to 0$.

13. Пусть по статистическим данным получены следующие значения средних: $\bar{x} = 10$, $\bar{y} = 30$, $\overline{xy} = 100$. Тогда:

а) $\mathrm{cov}(x, y) = -200$;

б) $\mathrm{cov}(x, y) = 100$;

в) $\mathrm{cov}(x, y) = 200$.

14. Линейный коэффициент корреляции r_{xy} изменяется в пределах:

а) от 0 до 1;

б) от -1 до 0;

в) от $-\infty$ до $+\infty$;

г) от -1 до 1.

15. Индекс корреляции для нелинейной регрессии находится в пределах:

а) $-1 \le \rho_{xy} \le 0$;

б) $-1 \le \rho_{xy} \le 1$;

в) $0 \le \rho_{xy} \le 1$.

16. Если $r_{xy} = -0{,}6$, то:

а) между *x* и *y* отсутствует какая-либо связь;
б) между *x* и *y* имеется слабая линейная связь;
в) между *x* и *y* имеется высокая линейная связь;
г) между *x* и *y* имеется заметная линейная связь.

17. Вариацию результативного признака y, обусловленную вариацией фактора x, оценивает:

а) коэффициент детерминации *R*-квадрат;
б) коэффициент эластичности;
в) коэффициент корреляции;
г) коэффициент регрессии b.

18. Параметр b степенной регрессии $\tilde{y} = a \cdot x^b$ совпадает с:

а) индексом детерминации;
б) коэффициентом эластичности;
в) индексом корреляции.

19. Величину изменения результативного признака при изменении фактора на одну единицу измерения в случае парной линейной модели оценивает:

а) коэффициент детерминации *R*-квадрат;
б) коэффициент эластичности;
в) коэффициент корреляции;
г) коэффициент регрессии b.

20. В модели линейной парной регрессии $y = 5 + 2x + \varepsilon$ изменение x на 2 единицы вызывает изменение y на:

а) 4 единицы;
б) 2 единицы;
в) 5 единиц;
г) 10 единиц.

21. Коэффициент b в уравнении $\tilde{y} = a + bx$ линейной регрессии совокупного спроса y на велосипеды (тыс. руб.) по цене x (руб.) оказался равным –1. Это означает:

а) увеличение цены на 1% снижает спрос на мобильные телефоны на 1%;
б) увеличение цена на 1 рубль снижает спрос на мобильные телефоны на 1%;

в) увеличение цены на 1% снижает спрос на мобильные телефоны на 1 тысячу рублей;

г) увеличение цены на 1 рубль снижает спрос на мобильные телефоны на 1 тысячу рублей.

22. Если при построении уравнения парной регрессии получен коэффициент детерминации *R*-квадрат, равный 0,98, то:

а) зависимость слабая, незначительная, изменения результативного признака большей частью обусловлены случайными факторами;

б) изменения результативного признака на 0,98% обусловлены изменениями фактора;

в) изменения результативного признака на 98% обусловлены изменениями фактора;

г) изменения результативного признака на 98% обусловлены случайными факторами.

23. Случайными воздействиями обусловлено 12% дисперсии результативного признака. Значит, значение индекса детерминации составило:

а) 88; б) 0,12; в) 0,88; г) 12.

24. Значение коэффициента корреляции равно 0,9. Следовательно, значение коэффициента детерминации составит:

а) 0,3; б) 0,81; в) 0,95; г) 0,1.

25. Если $0{,}1 \le r_{xy} \le 0{,}3$, то о связи между фактором и признаком можно сказать, что она:

а) умеренная;

б) сильная;

в) отсутствует;

г) слабая.

26. О сильной линейной связи между фактором и признаком говорит то, что:

а) $0{,}1 \le r_{xy} \le 0{,}3$;

б) $0{,}3 \le r_{xy} \le 0{,}7$;

в) $0{,}7 \le r_{xy} \le 1$;

г) $0{,}1 \le r_{xy} \le 0{,}5$.

27. Коэффициент корреляции больше нуля, это означает, что

а) связь между переменными тесная;

б) связь между переменными прямая;

в) связь между переменными обратная;

г) связь между переменными заметная.

28. Зависимость спроса от цены характеризуется уравнением регрессии $\tilde{y} = 10{,}34 \cdot x^{1{,}05}$. Следовательно:

а) с увеличением цен на одну единицу спрос увеличивается на 10,34%;
б) с уменьшением цен на 1% спрос снижается на 10,34%;
в) с увеличением цен на 1% спрос снижается на 1,05%.

29. В парной линейной регрессионной модели коэффициент детерминации *R*-квадрат равен:

а) квадрату коэффициента корреляции;
б) квадрату свободного члена;
в) свободному члену;
г) коэффициенту корреляции.

30. Допустимый предел средней ошибки аппроксимации составляет:

а) 8-12%;
б) 15-20%;
в) 20-25%;
г) 30%.

31. Значимость коэффициентов регрессии оценивается с помощью:

а) критерия Стьюдента;
б) критерия Фишера;
в) критерия Пирсона;
г) коэффициента детерминации.

32. Построена линейная парная регрессионная модель $y = 1 - 2x + \varepsilon$. Точечный прогноз по этой модели при $x = 5$ составляет:

а) -8; б) 9; в) -9; г) -10.

33. Замена переменных может быть применена при линеаризации следующей парной модели:

а) $y = ax^b \cdot \varepsilon$;

б) $y = a + \frac{b}{x} + \varepsilon$;

в) $y = \frac{a}{b + x} + \varepsilon$.

Ответы тестовых заданий

Номер задания	Ответы	Номер задания	Ответы	Номер задания	Ответы

1	а)	12	б)	23	в)
2	а)	13	а)	24	б)
3	а), г)	14	г)	25	г)
4	а), б)	15	в)	26	в)
5	б)	16	г)	27	б)
6	а)	17	а)	28	в)
7	б)	18	б)	29	а)
8	б)	19	г)	30	а)
9	а)	20	а)	31	а)
10	б)	21	г)	33	в)
11	а)	22	в)	33	б)

Глава 4
Модели множественной регрессии

Основные понятия: *множественная регрессионная модель, множественная линейная регрессионная модель, матрица парных коэффициентов корреляции, классическая нормальная линейная модель множественной регрессии, коэффициент множественной корреляции, уравнение множественной линейной регрессии в стандартизованном масштабе, коэффициент* (*индекс*) *детерминации, условия Гаусса-Маркова для множественной регрессионной модели, фиктивные переменные, дихотомические переменные, тест Чоу.*

Литература: [2-4], [7], [9-11].

4.1. Постановочный этап

Экономические явления и процессы, как правило, определяются большим числом одновременно и совокупно действующих факторов. При этом не всегда удается выделить один доминирующий фактор, влияющий на результативный признак. В таких случаях, когда необходимо учитывать влияние двух или более факторов, вместо парной регрессии применяется множественная регрессия.

Множественная регрессионная модель имеет вид

$$y = f(x_1, x_2, ..., x_m) + \varepsilon, \qquad (4.1)$$

где y – зависимая переменная (результативный признак), $x_1, x_2, ..., x_m$ – независимые переменные (факторы), ε – случайная ошибка.

Множественные регрессионные модели широко используются при решении проблем спроса и предложения, при описании потребительских и производственных функций, для исследования многих макро- и микроэкономических проблем.

Например, производственная функция представляет собой математическую модель, характеризующую зависимость объема выпускаемой продукции от факторов производства. Наиболее известна модель Кобба-Дугласа $y = A \cdot K^{\alpha} \cdot L^{\beta} \cdot \varepsilon$, где y – объем производства, K – затраты капитала, L – затраты труда, A, α, β – параметры модели, ε – случайная ошибка.

В 30-е годы Кейнс сформулировал гипотезу потребительской модели, которая сегодня чаще всего рассматривается как модель вида $C = f(x, P, M, Z)$, где C – потребление, x – доход, P – цена, индекс стоимости жизни, M – наличные деньги, Z – ликвидные активы. В частности, достаточно реалистична упрощенная модель спроса, имеющая вид $y = a \cdot x^{b} \cdot P^{c} \cdot \varepsilon$, где y – объем потребления товара, x – доход, P – цена, a, b, c – параметры модели, ε – случайная ошибка. Основная цель множественной регрессии – построить модель с большим числом факторов и определить при этом их индивидуальное и совокупное влияние на результативный фактор.

Общая постановка задачи множественной регрессии заключается в следующем: по имеющимся данным *n* наблюдений за изменением признака *y* в зависимости от наборов значений факторов $x_1, x_2, ..., x_m$ выбрать эконометрическую модель $y = f(x_1, x_2, ..., x_m) + \varepsilon$, оценить ее параметры и статистически обосновать, что факторы $x_1, x_2, ..., x_m$ существенны, а построенная функция $f(x_1, x_2, ..., x_m)$ наиболее точно соответствует данным наблюдений.

4.2. Спецификация модели множественной регрессии

Спецификация модели множественной регрессии включает решение двух задач. Первая задача заключается в выборе независимых переменных $x_1, x_2, ..., x_m$. Вторая задача состоит в выборе формы $f(x_1, x_2, ..., x_m)$ зависимости y от переменных $x_1, x_2, ..., x_m$.

При этом сами факторы $x_1, x_2, ..., x_m$, включаемые в модель, должны отвечать следующим требованиям:

1) быть количественно измеримыми;

2) быть тесно связанными с результативным признаком;

2) не должны быть коррелированными между собой.

При нарушении требования 3) невозможно определить индивидуальное влияние отдельных регрессоров $x_1, x_2, ..., x_m$ на результат y, что является весьма актуальным для осуществления прогнозов и принятия управляющих решений.

Отбор факторов $x_1, x_2, ..., x_m$, как правило, осуществляется в несколько этапов. Сначала отбираются факторы, связанные с изучаемым явлением на основе данных теоретического исследования (т.е. на основе экономической

теории, заключений специалистов и т.д.). Далее отобранные факторы подвергаются проверке существенности их влияния на изучаемый показатель с использованием методов математической статистики, малозначимые факторы при этом из модели исключаются.

Один из методов отбора факторов базируется на анализе *матрицы* (*таблицы*) *парных коэффициентов корреляции* (таблица 4.1). Элементами ее являются линейные коэффициенты парной корреляции факторов $x_1, x_2, ..., x_m$ как с зависимой переменной y, так и между собой. Отметим, что в таблице 4.1 коэффициенты $r_{x_i x_j}$ и $r_{x_j x_i}$, а также $r_{x_i y}$ и $r_{y x_i}$ совпадают, так как теснота связи между x_i и x_j такая же, как между x_j и x_i (аналогично, для y и x_i).

По данным такой матрицы можно примерно оценить, какие факторы существенно влияют на переменную y, а какие – несущественно, а также определить взаимосвязь между факторами, т.е. корреляцию между объясняющими переменными. Считается, что две переменные x_i и x_j *явно коллинеарны*, т.е. находятся между собой в высокой линейной зависимости, если $r_{x_i x_j} \geq 0{,}7$. В таком случае одна из них исключается из модели. Предпочтение отдается тому фактору, который достаточно тесно связан с результативным фактором, но имеет при этом наименьшую тесноту связи с другими объясняющими факторами.

Таблица 4.1. Матрица (таблица) парных коэффициентов корреляции

	y	x_1	x_2	…	x_m
y	1	r_{yx_1}	r_{yx_2}	…	r_{yx_m}
x_1	r_{x_1y}	1	$r_{x_1x_2}$	…	$r_{x_1x_m}$
x_2	r_{x_2y}	$r_{x_2x_1}$	1	…	$r_{x_2x_m}$
…	…	…	…	…	…
x_m	r_{x_my}	$r_{x_mx_1}$	$r_{x_mx_2}$	…	1

В процедуре построения множественной регрессионной модели правильный отбор факторов весьма важен. Результаты неправильной спецификации переменных в уравнении отражаются на модели следующим образом:

1) если опущена переменная, которая должна быть включена, то оценки регрессии часто оказываются смещенными;

2) если включена переменная, которая не должна присутствовать в уравнении, то оценки регрессии могут быть несмещенными, но неэффективными.

Подходы к отбору факторов на основе показателей корреляции могут быть разными. Наиболее широкое применение получили следующие два метода, дающие в целом близкие результаты:

– *метод исключения* (на первом шаге строится уравнение регрессии с полным набором факторов, а затем после исключения коллинеарных факторов отбираются факторы, имеющие наибольшее влияние на изменение результативного признака, менее значимые факторы при этом исключаются);

– *метод включения* (заключается в поэтапном введении новых факторов в регрессионную модель).

Относительно формы зависимости различают линейные и нелинейные множественные модели.

Модель множественной линейной регрессии описывается уравнением

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon, \qquad (4.2)$$

где коэффициенты a_j, j = 1, 2,…, m, характеризуют среднее изменение результата с изменением фактора x_j на единицу при неизменном значении других факторов.

Ввиду отмеченной четкой интерпретации коэффициентов a_j уравнения (4.2) линейные множественные модели широко представлены в эконометрическом анализе. Однако реальное соотношение между социально-экономическими явлениями и процессами далеко не всегда можно выразить линейными функциями.

В таком случае прибегают к нелинейным моделям, среди которых наиболее часто применяются:

– *степенная* $y = a_0 \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot ... \cdot x_m^{a_m} \cdot \varepsilon$ (тогда a_j, где j = 1, 2,…, m, – коэффициенты эластичности, которые показывают, на сколько процентов изменится в среднем результат y с изменением фактора x_j на 1% при неизменности действия других факторов);

– *экспоненциальная* $y = e^{a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m} \cdot \varepsilon$;

– *гиперболическая* $y = a_0 + \frac{a_1}{x_1} + \frac{a_2}{x_2} + ... + \frac{a_m}{x_m} + \varepsilon$.

Используются и другие нелинейные функции. Правильность выбора формы модели определяется на этапе верификации.

4.3. Параметризация модели

Оценки неизвестных параметров a_0, a_1, …, a_m линейной модели $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$ множественной регрессии находятся, как и в случае парной регрессии, с помощью метода наименьших квадратов из условия оптимизации функции

$$S = S(a_0; a_1; a_2,..., a_m) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 ,$$

т.е. из условия

$$\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \to \min .$$

Для нахождения параметров a_0, a_1, …, a_m на основании необходимого условия экстремума приравниваются к нулю частные производные функции $S(a_0; a_1; a_2,..., a_m)$ по переменным a_0, a_1, …, a_m. В итоге получается система, содержащая $m+1$ линейных уравнений (по числу параметров) с $m+1$ переменными:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} (x_1)_i + a_2 \sum_{i=1}^{n} (x_2)_i + ... + a_m \sum_{i=1}^{n} (x_m)_i = \sum_{i=1}^{n} y_i , \\ a_0 \sum_{i=1}^{n} (x_1)_i + a_1 \sum_{i=1}^{n} (x_1)_i^{\,2} + a_2 \sum_{i=1}^{n} (x_1)_i (x_2)_i + ... + a_m \sum_{i=1}^{n} (x_1)_i (x_m)_i = \\ \qquad = \sum_{i=1}^{n} (x_1)_i y_i , \\ ..... , \\ a_0 \sum_{i=1}^{n} (x_m)_i + a_1 \sum_{i=1}^{n} (x_m)_i (x_1)_i + a_2 \sum_{i=1}^{n} (x_m)_i (x_2)_i + ... + a_m \sum_{i=1}^{n} (x_m)_i^{\,2} = \\ \qquad = \sum_{i=1}^{n} (x_m)_i y_i . \end{cases} \qquad (4.3)$$

Здесь n – число наблюдений $y_1, y_2,..., y_n$ зависимой переменной y, $(x_j)_1, (x_j)_2,..., (x_j)_n$ – наблюдаемые значения j-го фактора, $j = 1,2,..., m$.

Для решения системы (4.3) может быть применен метод Крамера, метод Гаусса, матричный метод или любой другой метод решения систем линейных уравнений.

Метод наименьших квадратов применительно к множественной линейной регрессионной модели дает хорошие результаты (несмещенные, эффективные и состоятельные оценки параметров регрессии) при выполнении определенных требований к случайной переменной ε.

Теорема Гаусса–Маркова. *Пусть выполняются условия*:

1) *математическое ожидание случайной переменной равно нулю*;

2) *дисперсия случайной переменной одинакова для всех наблюдений* (*постоянство дисперсии называется гомоскедастичностью*, *непостоянство – гетероскедастичностью*);

3) *отсутствует систематическая связь между значениями случайной переменной для различных наблюдений* (*это условие называется условием отсутствия автокорреляции*);

4) *случайная переменная независима от объясняющих переменных.*

Тогда оценки параметров регрессии, полученные по МНК, являются несмещенными, состоятельными и эффективными.

Линейная модель, удовлетворяющая условиям 1-4, называется *классической линейной моделью* множественной регрессии. Если же в дополнение к условиям 1-4 выполняется предположение о нормальном распределении случайной величины ε, то классическая линейная модель называется *нормальной*.

При построении классических линейных множественных регрессионных моделей необходимо выполнение и таких предположений, как:

– *отсутствует мультиколлинеарность* (*нет зависимости между факторами*);

– *число наблюдений существенно больше числа объясняющих переменных* (*по крайней мере, в три раза*);

– *отсутствуют ошибки спецификации.*

Для параметризации нелинейных моделей множественной регрессии часто используется метод линеаризации. Например, в случае модели Кобба-Дугласа $y = A \cdot K^{\alpha} \cdot L^{\beta} \cdot \varepsilon$, где y – объем производства, K – затраты капитала, L – затраты труда, A, α, β – параметры модели, ε – случайная ошибка, осуществляется логарифмирование:

$$\ln y = \ln A + \alpha \cdot \ln K + \beta \cdot \ln L + \ln \varepsilon .$$

Далее по заданным рядам статистических данных рассчитываются ряды их логарифмов, а затем для них с помощью метода наименьших квадратов оцениваются параметры A, α, β модели Кобба-Дугласа.

Экономическая интерпретация коэффициентов α и β в модели Кобба-Дугласа заключается в следующем: при увеличении капиталовложений на 1% объем производства увеличится на α%, а при увеличении затрат труда на 1% объем производства увеличится на β%.

4.4. Верификация модели

Проверка качества оцененной множественной регрессионной модели проводится по следующим направлениям:

– оценка тесноты связи рассматриваемого набора факторов с исследуемым признаком;

– проверка общего качества уравнения регрессии;

– проверка статистической значимости коэффициентов регрессии;

– проверка выполнимости предпосылок МНК.

Независимо от формы связи $\tilde{y} = f(x_1, x_2, ..., x_m)$ (линейной или нелинейной) тесноту совместного влияния факторов на результат оценивает *коэффициент* (*индекс*) *множественной корреляции*:

$$R = R_{yx_1x_2...x_m} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{\frac{\sigma^2_\phi}{\sigma^2_y}},$$

где $\sigma^2_y = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$ – общая дисперсия результативного признака, $\sigma^2_\phi = \frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2$ – факторная дисперсия результативного признака, $\sigma^2_{\text{ост}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2$ – остаточная дисперсия результативного признака. Так как $\sigma^2_y = \sigma^2_\phi + \sigma^2_{\text{ост}}$, то $R_{yx_1x_2...x_m} \in [0;1]$. При этом, чем ближе к 1 индекс множественной корреляции, тем теснее связь результативного признака со всем набором исследуемых факторов.

Величина индекса множественной корреляции больше или равна максимального парного индекса корреляции: $R_{yx_1x_2...x_m} \geq r_{yx_i}$ для всех $i = 1,2,...,m$. При этом при правильном включении факторов в модель величина индекса множественной корреляции будет существенно отличаться от парных индексов корреляции. Если же дополнительно включенные в уравнение множественной регрессии факторы второстепенны, то индекс множественной корреляции может практически совпадать с индексом парной корреляции (различия в третьем, четвертом знаках). Отсюда следует, что сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

Низкое значение индекса множественной корреляции означает, что либо в регрессионную модель не включены существенные факторы, либо рассматриваемая форма связи $\tilde{y} = f(x_1, x_2, ..., x_m)$ не отражает реальные соотношения между переменными, включенными в модель. В обоих случаях требуется дополнительная работа по спецификации модели.

Для линейной модели работа по определению существенных факторов может быть связана с определением стандартизованных коэффициентов регрессии и средних коэффициентов эластичности.

Если коэффициенты множественной линейной регрессии рассматривать в качестве показателей влияния факторов, то следует иметь в виду, что коэффициенты регрессии в линейной модели $y = a_0 + a_1x_1 + a_2x_2 + ... + a_mx_m + \varepsilon$ между собой прямо несравнимы. Их

численные значения зависят от выбранных единиц измерения каждого фактора. Чтобы коэффициенты регрессии стали сопоставимы, их приводят к *стандартизованному масштабу*.

Уравнение множественной регрессии в стандартизованном масштабе имеет вид

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + ... + \beta_m t_{x_m} + \varepsilon ,$$

где $t_y = \dfrac{y - \bar{y}}{\sigma_y}$, $t_{x_j} = \dfrac{x_j - \bar{x}_j}{\sigma_{x_j}}$, j = 1, 2, …, m, – стандартизованные переменные. Связь между стандартизованными коэффициентами β_j и коэффициентами множественной регрессии a_j описывается соотношениями $a_j = \beta_j \dfrac{\sigma_y}{\sigma_{x_j}}$, j = 1, 2, …, m, $a_0 = \bar{y} - a_1\bar{x}_1 - a_2\bar{x}_2 - ... - a_m\bar{x}_m$. Стандартизованные коэффициенты сравнимы между собой, поэтому с их помощью можно ранжировать факторы $x_1, x_2, ..., x_m$ по силе воздействия на результат y.

Средние коэффициенты эластичности для линейной множественной регрессии рассчитываются по формуле $\bar{Э}_{yx_j} = a_j \dfrac{\bar{x}_j}{\bar{y}}$ и показывают, на сколько процентов в среднем изменяется зависимая переменная с изменением на 1% фактора x_j при фиксированном значении других факторов. Сравнение показателей эластичности друг с другом позволяет также ранжировать факторы модели по силе их влияния на результирующий фактор y.

Как правило, выводы о ранжировании влияния факторов на результат на основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности дополняются выводами, полученными на основе анализа матрицы парных коэффициентов регрессии.

Одной из наиболее эффективных оценок общего качества множественной модели и характеристикой ее прогностической силы является коэффициент детерминации R^2. Он рассчитывается как квадрат индекса множественной корреляции, т.е. $R^2 = R^2_{yx_1x_2...x_m} = 1 - \dfrac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$.

Величина $R^2 \cdot 100\%$ показывает, на сколько процентов изменения результативного признака объясняются изменением факторных признаков, включенных в модель.

Недостатком коэффициента детерминации R^2 является то, что он не уменьшается при добавлении новых объясняющих переменных. Ввиду этого при сравнении двух моделей не всегда ясно, за счет чего возрос R^2: за счет простого увеличения числа факторов, либо за счет реального влияния новых введенных факторов. Это, в свою очередь, может привести к ошибочному выводу о значимости влияния факторов на результативный признак. Для того чтобы компенсировать влияние такого эффекта при включении в модель нового фактора, вместо показателя R^2 рассматривают скорректированный коэффициент детерминации $\overline{R}^2 = 1 - \frac{n-1}{n-m-1} \cdot (1 - R^2)$, где m – число объясняющих переменных в модели, а n – число наблюдений.

В отличие от R^2 скорректированный коэффициент детерминации $\overline{R}^2$ может уменьшаться при введении в модель новых объясняющих переменных, не оказывающих существенного влияния на зависимую переменную. В то же время увеличение $\overline{R}^2$ может не означать улучшения качества регрессионной модели.

Как и в случае парной регрессии, общее качество множественной модели может быть оценено с помощью *стандартной ошибки регрессии* $s = \sqrt{\frac{1}{n-(m+1)} \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}$. Величина стандартной ошибки регрессии характеризует среднюю величину рассеивания наблюдаемых значений переменной y относительно теоретических.

Для оценки адекватности уравнения регрессии может быть применена *средняя ошибка аппроксимации*:

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

Ошибка аппроксимации не более 8–12% свидетельствует о хорошем качестве модели.

Оценка статистической значимости уравнения множественной регрессии в целом осуществляется с помощью *F*-критерия Фишера.

F-критерий Фишера заключается в проверке нулевой гипотезы $H_0 : R^2 = 0$ о статистической незначимости уравнения регрессии. Для этого выполняется сравнение фактического $F_{\text{набл}}$ и критического (табличного) $F_{\text{кр}}$ значений *F*-критерия Фишера.

Наблюдаемое значение статистики $F_{\text{набл}}$ вычисляется по выборочным данным на основании формулы $F_{\text{набл}} = \frac{R^2}{1 - R^2} \cdot \frac{n-m-1}{m}$, где m – число

объясняющих переменных в модели, а n – число наблюдений.По таблицам критических точек F-распределения находится критическое значение статистики $F_{кр}$ при заданном уровне значимости α. При этом число степеней свободы определяется значениями $k_1 = m$ и $k_2 = n - m - 1$. Уровень значимости α – вероятность отвергнуть гипотезу H_0 при условии, что она верна.

Если $F_{кр} < F_{набл}$, то нулевая гипотеза отвергается, что говорит о соответствии теоретического уравнения регрессии выборочным данным. Если $F_{кр} > F_{набл}$, то признается ненадежность уравнения регрессии.

Гипотеза о статистической значимости коэффициентов линейной множественной регрессии $H_0 : a_j = 0$, где j = 1, 2, …, m, при альтернативной гипотезе $H_1 : a_j \neq 0$ проверяется с помощью t-статистики, имеющей распределение Стьюдента с числом степеней свободы, равным $n - m - 1$. По выборочным данным вычисляется наблюдаемое значение t-статистики $t_{набл}$ (для каждого коэффициента) как отношение значения коэффициента к величине его стандартной ошибки: $t_{набл} = t_{a_j} = \dfrac{a_j}{s_{a_j}}$. Стандартная ошибка коэффициента регрессии может быть определена по следующей формуле: $s_{a_j} = \dfrac{\sigma_y \sqrt{1 - R^2_{yx_1 \ldots x_m}}}{\sigma_{x_j} \sqrt{1 - R^2_{x_j x_1 \ldots x_m}}} \cdot \dfrac{1}{\sqrt{n - m - 1}}$, где σ_y – среднее квадратическое отклонение для признака y, σ_{x_j} – среднее квадратическое отклонение для фактора x_j, $R^2_{yx_1 \ldots x_m}$ – коэффициент детерминации для уравнения множественной регрессии, $R^2_{x_j x_1 \ldots x_m}$ – коэффициент детерминации зависимости фактора x_j со всеми другими факторами уравнения множественной регрессии.

Наблюдаемые значения t-статистики для каждого коэффициента регрессии затем сравнивается с табличным значением t-статистики $t_{кр}$. Если $|t_{набл}| > t_{кр}$, то нулевая гипотеза $H_0 : a_j = 0$ отвергается и признается, что коэффициент a_j регрессии не случайно отличаются от нуля, а значит, он статистически значим. Если же $|t_{набл}| < t_{кр}$, то коэффициент регрессии статистически не значим и природа его формирования случайна. В таком случае считается, что фактор x_j линейно не связан с зависимой переменной и его рекомендуется исключить из уравнения регрессии. Это не приведет к существенной потере качества модели, но сделает ее более простой и конкретной.

Следует отметить, что в экономических исследованиях исключению переменных из регрессионной модели должен предшествовать тщательный качественный анализ. Иногда может оказаться, что целесообразнее все же оставить в модели одну или несколько объясняющих переменных, хотя они и не оказывают существенного влияния на зависимую переменную.

4.5. Прогнозирование по множественной регрессионной модели

Под прогнозом по множественной регрессионной модели понимается оценка значения зависимой переменной y для значений объясняющих переменных, которых нет в исходных наблюдениях. Как и в случае парной модели, различают *точечный и интервальный прогнозы*.

Точечный прогноз y_p по уравнению регрессии осуществляется путем подстановки значений регрессоров $x_1^0, x_2^0, ..., x_m^0$ в уравнение $\tilde{y} = f(x_1, x_2, ..., x_m)$ регрессии. В случае линейной модели $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$ имеем значение $y_p = a_0 + a_1 x_1^0 + a_2 x_2^0 + ... + a_m x_m^0$. Для классической линейной модели полученный точечный прогноз является несмещенным.

В дополнение к точечному прогнозу можно определить (по аналогии с парным случаем) границы возможного изменения прогнозируемого показателя, т.е. с заданным уровнем значимости вычислить доверительный интервал для прогнозируемого значения y_p зависимой переменной y.

Для компактного описания стандартной ошибки прогноза в случае линейной множественной регрессии введем следующие матричные обозначения:

$$X = \begin{pmatrix} 1 & (x_1)_1 & ... & (x_m)_1 \\ 1 & (x_1)_2 & ... & (x_m)_2 \\ ... & ... & ... & ... \\ 1 & (x_1)_n & ... & (x_m)_n \end{pmatrix}$$

– матрица наблюдаемых значений факторов x_1, x_2, ..., x_m (в матрицу X дополнительно вводится столбец, все элементы которого равны 1; матрица имеет n строк и $m+1$ столбец);

$$X_0 = \begin{pmatrix} 1 \\ x_1^0 \\ ... \\ x_m^0 \end{pmatrix}$$

– вектор-столбец значений факторов x_1, x_2, ..., x_m, для которых необходимо найти интервальный прогноз (вектор-столбец дополняется в первой строке элементом 1).

Пусть X' – транспонированная матрица для матрицы X, а $X_0' = \begin{pmatrix} 1 & x_1^0 & \dots & x_m^0 \end{pmatrix}$.

Тогда стандартная ошибка прогноза определяется по формуле

$$m_p = s \cdot \sqrt{1 + X_0'(X'X)^{-1}X_0}, \qquad (4.4)$$

где $s = \sqrt{\frac{1}{n-(m+1)}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}$ – стандартная ошибка регрессии, $(X'X)^{-1}$ – матрица, обратная к матрице $X'X$.

Затем строится *доверительный интервал прогноза*

$$(y_p - t_{\text{кр}} \cdot m_p; y_p + t_{\text{кр}} \cdot m_p),$$

т.е. определяются нижняя и верхняя границы интервала прогноза (за середину доверительного интервала выбирается точечная оценка y_p, а отступ от нее пропорционален критическому значению $t_{\text{кр}}$ и стандартной ошибке регрессии s).

4.6. Фиктивные переменные

При постановке ряда регрессионных задач приходится рассматривать зависимость некоторого показателя не только от количественных переменных, принимающих значения из определенных числовых интервалов, но также зависимость его от ряда факторов, имеющих два и более качественных уровня. Такая ситуация имеет место, в частности, в следующих примерах.

1. Исследуется зависимость заработной платы работников предприятия от стажа и уровня образования. При такой постановке задачи объясняющими факторами будут стаж работы x_1 и уровень образования x_2. Но если фактор x_1 имеет явно выраженный количественный характер (он может принимать любые значения в интервале, например, от 0 до 70), то фактор x_2 характеризуется только тремя уровнями: «начальное образование», «среднее образование», «высшее образование».

2. Строится модель взаимодействия цены и спроса на некоторый товар с учетом сезонности продаж. В этой модели качественный характер имеет фактор времени года. Он принимает только два значения: «зимний период», «летний период».

Существуют два принципиально различных подхода в решении приведенных задач. Первый заключается в том, чтобы для каждого уровня

качественного признака построить и оценить свою регрессионную модель (в первом примере таких моделей будет три, а во втором – две), а затем изучить различия между ними. Другой подход состоит в том, чтобы качественные факторы некоторым образом ввести в одно уравнение регрессии, а затем исследовать это уравнение.

Качественные факторы, рассматриваемые как переменные регрессионной модели, называются в эконометрике *фиктивными* (или *манекенными*) *переменными*. В противоположность значащим переменным, отражающим количественную сторону показателя, фиктивные переменные играют роль индикаторов, сигнализирующих об уровне рассмотрения задачи. Поэтому фиктивные переменные часто еще называют *индикаторами*.

В качестве фиктивных переменных обычно используются так называемые *дихотомические переменные*, которые имеют только два уровня (например, «фактор действует» – «фактор не действует», «сезон летний» – «сезон зимний», «пол мужской» – «пол женский»).

Главная задача фиктивных переменных – отражение в модели значения качественных факторов, которые порой существенно влияют на структуру связей между значащими переменными и приводят к скачкообразному изменению параметров регрессионной модели.

Коэффициент регрессии при фиктивной переменной интерпретируется как среднее изменение зависимой переменной при переходе от одного уровня к другому при неизменных значениях других факторов. На основе *t*-критерия Стьюдента можно сделать вывод о значимости влияния фиктивной переменной, существенности расхождения на разных уровнях рассмотрения задачи.

4.7. Введение фиктивных переменных в модель

Для того, чтобы ввести фиктивную переменную в регрессионную модель, ей необходимо присвоить некоторые числовые значения, придав тем самым фиктивным переменным количественное содержание. В случае дихотомической переменной это делается следующим образом. Фиктивной переменной придается значение 1, если признак присутствует в наблюдении, и 0 – при его отсутствии. Таким образом, если z – дихотомическая переменная, то в описанном выше двоичном виде она формализуется равенством

$$z = \begin{cases} 1, \textit{ если фактор действует,} \\ 0, \textit{ если фактор не действует.} \end{cases}$$

Что касается фиктивной переменой, имеющей k уровней качества ($k > 2$), то при построении регрессионной модели она заменяется на $k-1$ дихотомическую фиктивную переменную.

Например, при исследовании зависимости заработной платы от стажа работника и его образования модель может быть представлена в виде:

$$y = f(x) + a_1 z_1 + a_2 z_2 + \varepsilon,$$

где $f(x)$ – часть заработной платы, объясняемая стажем,

$$z_1 = \begin{cases} 1, \textit{ если у работника высшее образование}, \\ 0, \textit{ во все остальных случаях}, \end{cases}$$

$$z_2 = \begin{cases} 1, \textit{ если у работника среднее образование}, \\ 0, \textit{ во всех остальных случаях}. \end{cases}$$

Третьей дихотомической переменной z_3 и не требуется, так как если работник имеет начальное образование, то это уже учтено при $z_1 = z_2 = 0$. Более того, с точки зрения требований к качеству модели ее вводить нельзя, так как тогда для любого работника

$$z_1 + z_2 + z_3 = 1,$$

то есть переменные z_1, z_2, z_3 становятся линейно зависимыми, а это приводит к появлению мультиколлинеарности. Такая ситуация совершенной мультиколлинеарности получила название «*ловушка фиктивной переменной*». Чтобы избежать ее, необходимо руководствоваться следующим простым правилом.

Если фиктивная переменная z имеет k качественных уровней, то при моделировании вместо нее используются $k-1$ *дихотомическая переменная* $z_1, z_2, \ldots, z_{k-1}$.

4.8. Тест Чоу

Иногда выборка наблюдений состоит из двух или более подвыборок, и трудно установить, следует ли оценивать одну объединенную регрессию или отдельные регрессии для каждой подвыборки.

Предположим, что ставится задача не только построить модель зависимости цены p квартиры от факторов $x_1, x_2, \ldots, x_m$, но и решить вопрос существенности (или несущественности) влияния фактора: «квартира в панельном или кирпичном доме». Другими словами, необходимо выяснить, можно ли считать одним и тем же уравнение регрессии $\tilde{p} = f(x_1, x_2, ..., x_m)$ для панельных и кирпичных домов или необходимо всю имеющуюся выборку разбить на две части (одну для панельных домов, а другую для кирпичных) и построить для каждой из них свое уравнение регрессии.

Формальный статистический тест для оценки объединенной регрессии в сравнении с регрессиями для подвыборок был предложен Грегори Чоу.

Суть теста Чоу заключается в следующем:

1) полная выборка объема n разбивается на две подвыборки **A** и **B** объемами n_1 и n_2 соответственно ($n = n_1 + n_2$);

2) для полной выборки, а также для подвыборок **A** и **B** оцениваются параметры линейных уравнений регрессии:

$$y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon_0, \quad (0)$$

$$y = a_{01} + a_{11} x_1 + a_{21} x_2 + ... + a_{m1} x_m + \varepsilon_1, \quad (1)$$

$$y = a_{02} + a_{12} x_1 + a_{22} x_2 + ... + a_{m2} x_m + \varepsilon_2; \quad (2)$$

3) выдвигается и проверяется с помощью F-статистики гипотеза о равенстве друг другу соответствующих коэффициентов регрессии, а именно гипотеза $H_0 : a_{i1} = a_{i2}$, $i = 0,1,2,...,m$.

Наблюдаемое значение статистики $F_{\text{набл}}$ вычисляется по выборочным данным на основании формулы

$$F_{\text{набл}} = \frac{S_0 - S_1 - S_2}{S_1 + S_2} \cdot \frac{n - 2m - 2}{m + 1},$$

где S_j – сумма квадратов отклонений выборочных значений y_i от соответствующих значений, рассчитанных по уравнению регрессии (j), $j = 0,1,2$, $i = 1,2,...,n$, n – объем выборки.

Построенная F-статистика имеет распределение Фишера с числами степеней свободы $k_1 = m + 1$ и $k_2 = n - 2m - 2$. Если $F_{\text{кр}} < F_{\text{набл}}$, то гипотеза H_0 отклоняется. В этом случае моделирование следует осуществлять с помощью кусочно-линейной модели. Если же $F_{\text{кр}} > F_{\text{набл}}$, то нет оснований отклонять нулевую гипотезу, а значит, ее моделирование следует осуществлять с помощью единого для всей совокупности уравнения.

4.9. Фиктивные переменные и сезонность

При экономическом моделировании фиктивные переменные используются для учета сезонной компоненты. При этом под *сезонной компонентой* понимается та составляющая, которая отражает повторяемость экономических процессов в течение не очень длительного периода (года, квартала, месяца).

Обычно сезонные колебания характерны для временных рядов. В таких моделях выделение и удаление сезонной компоненты позволяет сконцентрировать внимание на общем направлении развития экономического процесса.

Влияние сезонной компоненты может быть отражено как в аддитивном, так и мультипликативном виде. Например, аддитивная модель объема продаж y туристических путевок в зависимости от цены тура x с учетом времени года выглядит следующим образом: $y = a_0 + bx + a_1 z_1 + a_2 z_2 + a_3 z_3 + \varepsilon$, где

$$z_1 = \begin{cases} 1, \textit{ если наблюдение относится к зиме,} \\ 0, \textit{ во всех остальных случаях,} \end{cases}$$

$$z_2 = \begin{cases} 1, \textit{ если наблюдение относится к весне,} \\ 0, \textit{ во всех остальных случаях,} \end{cases}$$

$$z_3 = \begin{cases} 1, \textit{ если наблюдение относится к лету,} \\ 0, \textit{ во всех остальных случаях.} \end{cases}$$

Мультипликативная модель указанной задачи может быть описана следующим уравнением: $y = (a + bx)(a_1 z_1 + a_2 z_2 + a_3 z_3) + \varepsilon$.

Если первая модель сезонные различия отражает лишь в различии свободных членов сезонных моделей, то вторая модель затрагивает и изменения коэффициента пропорциональности b.

На практике строят несколько моделей, сравнивают их между собой и с точки зрения дисперсионного анализа выбирают лучшую.

4.10. Обзор некоторых вопросов и проблем множественной регрессии

Одной из ключевых проблем множественного регрессионного моделирования является проблема спецификации модели. Чтобы выбрать качественную модель, необходимо ответить на ряд вопросов, возникающих при ее анализе:

1. Какие ошибки спецификации встречаются, и каковы последствия данных ошибок?
2. Как обнаружить ошибку спецификации?
3. Каким образом можно исправить ошибку спецификации и перейти к более качественной модели?

Некоторые ответы на эти вопросы можно найти в [3] и [4].

Более обоснованным по сравнению с методами включения и исключения переменных является *метод пошагового отбора переменных*. Процедура его применения состоит в следующем:

1-й шаг. Из совокупности входных переменных выбирается переменная, имеющая наибольший парный коэффициент корреляции с переменной y. Для полученной модели парной регрессии вычисляется коэффициент детерминации R_1^2.

2-й шаг. К выбранной переменной добавляется следующая, выбираемая из условия, чтобы коэффициент двухфакторной модели был наибольшим. Коэффициент детерминации двухфакторной модели R_2^2 сравнивается с R_1^2. Если R_2^2 существенно больше R_1^2, то приступают к выбору третьей переменной. В противном случае удовлетворяются однофакторной моделью.

Последующие шаги осуществляются аналогично второму шагу. Процесс отбора заканчивается, когда очередная включаемая в модель переменная не дает существенного увеличения коэффициента детерминации.

В любом случае при выборе спецификации модели следует в первую очередь руководствоваться экономическим анализом. Иначе можно получить чрезвычайно хорошую, с точки зрения математики, модель, которая будет лишена какого-либо экономического смысла.

Как отмечено выше, ранжировать факторы, участвующие в линейной модели, можно с помощью стандартизованных коэффициентов регрессии, коэффициентов парной корреляции и средних коэффициентов эластичности. Такая же цель может быть достигнута с помощью частных коэффициентов корреляции. Методика использования таких показателей описана в [2,4].

Порядок расчета доверительного интервала прогноза классической линейной нормальной модели можно найти в [2-4]. Другие методы прогнозирования по множественной модели описаны в [10].

Проверка качества оцененной множественной регрессионной модели, кроме оценки тесноты связи рассматриваемого набора факторов с исследуемым признаком, проверки общего качества уравнения регрессии и проверки статистической значимости коэффициентов регрессии, включает проверку выполнимости предпосылок МНК. Этот вопрос рассматривается в главе 5.

Примеры решения типовых заданий

Пример 4.1. По статистическим данным таблицы 4.2:

1) на основании анализа матрицы парных коэффициентов корреляции из трех независимых переменных отобрать два наиболее существенных фактора;

2) для отобранных факторов построить двухфакторное уравнение линейной регрессии;

3) определить коэффициент множественной корреляции;

4) проверить значимость уравнения при уровнях значимости 0,05 и 0,01.

Таблица 4.2. Статистические данные примера 4.1

	y	x_1	x_2	x_3
1	113	10	1	77
2	124	5	2	64

3	124	10	2	77
4	122	13	2	66
5	128	9	1	71
6	140	14	6	81
7	117	12	1	58
8	113	15	3	66
9	122	13	2	73
10	139	27	14	81
11	126	8	6	73
12	120	8	3	65
13	125	24	6	66
14	118	8	1	74
15	122	8	4	64
16	133	15	5	79
17	136	12	4	71
18	146	16	9	68
19	148	23	5	78
20	136	16	8	74
21	138	10	3	64
22	124	12	7	74
23	123	8	3	71
24	149	29	8	87
25	130	9	4	56
26	117	91	3	65
27	126	12	1	61
28	110	7	1	35
29	98	6	0	26

Решение:

1) Построим матрицу парных коэффициентов корреляции, используя функцию «Сервис. Анализ данных. Корреляция» табличного процессора MS Excel (таблица 4.3).

Таблица 4.3. Матрица парных коэффициентов корреляции примера 4.1

	y	x_1	x_2	x_3
y	1			
x_1	0, 638	1		
x_2	0,680	0,710	1	
x_3	0,661	0,513	0,506	1

Из матрицы следует, что наблюдается явная коллинеарность между факторами x_1 и x_2, так как $r_{x_1x_2} = 0{,}710 > 0{,}7$. Для дальнейшего рассмотрения

оставляем фактор x_2, так как он меньше коррелирует с фактором x_3 ($r_{x_2x_3}$ = 0,506 < $r_{x_1x_3}$ = 0,513) и теснее связан с результативным фактором y.

Таким образом, далее будет строиться регрессия переменной *y* на факторы x_2 и x_3.

2) Для построения уравнения множественной линейной регрессии используем функцию «Сервис. Анализ данных. Регрессия». Задав соответствующие диапазоны данных, получим следующий набор таблиц А, Б, В.

Таблица А

Показатель	**Значение**	**Комментарии**
Множественный R	0,773	Множественный коэффициент корреляции
R-квадрат	0,597	
Нормированный R-квадрат	0,566	
Стандартная ошибка	7,768	Стандартная ошибка регрессии
Наблюдения	29	Число наблюдений

Таблица Б

	Число степеней свободы	Дисперсия	Дисперсия на 1 степень свободы	Статистика Фишера	
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	2326,1	1163,1	19,3	7,35E-06
Остаток	26	1569,1	60,3		
Итого	28	3895,2			

Таблица В

	Коэффициенты уравнения регрессии	Стандартная ошибка определения коэффициентов	*t*-статистика	Вероятность ошибки	Нижние 95%-пределы	Верхние 95%-пределы
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-значение*	*Нижние* 95%	*Верхние* 95%
Y-пересечение	92,585	8,351	11,087	0,0000	75,420	109,750
Переменная x_2	1,761	0,547	3,219	0,0030	0,637	2,886
Переменная x_3	0, 397	0,134	2,952	0,0070	0,120	0,673

Из таблицы В следует, что уравнение регрессии имеет вид

$$\tilde{y} = 92{,}585 + 1{,}761x_2 + 0{,}397x_3.$$

3) Коэффициент множественной корреляции определяется из таблицы А: $R = 0{,}773$.

4) Проверка значимости уравнения регрессии основана на использовании F-критерии Фишера. Фактическое значение критерия берется из таблицы Б: $F_{\text{набл}} = 19{,}3$.

Для определения табличных значений используем встроенную функцию MS Excel «FРАСПОБР», задавая параметры $k_1 = 2$, $k_2 = 29 - 2 - 1 = 26$, $\alpha = 0{,}05$ и $\alpha = 0{,}01$.

В результате получаем $F_{\text{кр.}0{,}05} = 3{,}369$, $F_{\text{кр.}0{,}01} = 5{,}526$.

Откуда следует, что уравнение регрессии значимо при $\alpha = 0{,}05$ и $\alpha = 0{,}01$.

Пример 4.2. По статистическим данным, приведенным в таблице 4.4, построить линейную регрессионную модель зависимости заработной платы y (доллары) рабочих некоторого предприятия от их возраста x (годы) и пола z (мужской или женский).

Таблица 4.4. Статистические данные примера 4.2

Заработная плата, y	Возраст, x	Пол, z
300	29	ж
400	40	м
300	36	ж
320	32	ж
200	23	м
350	45	м
350	38	ж
400	40	м
380	50	м
400	47	м
250	28	ж
350	30	м
200	25	м
400	48	м
220	30	ж
320	40	м
390	40	м
360	38	м
260	29	ж
250	25	м

Решение:

Переменная z является фиктивной:

$$z = \begin{cases} 1, \textit{ если рабочий – мужчина}, \\ 0, \textit{ если рабочий – женщина}. \end{cases}$$

Модель будем строить в виде $y = a + bx + cz + \varepsilon$. Параметризуя ее (например, с помощью табличного процессора MS Excel), найдем коэффициенты регрессии $a = 63{,}52$, $b = 7$, $c = 10{,}32$. Поэтому уравнение регрессии имеет вид $\tilde{y} = 63{,}52 + 7x + 10{,}32z$. При этом $R^2 = 0{,}732$ и коэффициент детерминации значим (наблюдаемое значение F-критерия больше критического). Правда, коэффициент регрессии при фиктивной переменной является незначимым (это может объясняться малым размером выборки).

Коэффициент c в уравнении регрессии интерпретируется следующим образом: при одном и том же возрасте заработная плата мужчин-рабочих на 10,32 доллара выше, чем у женщин-рабочих.

Пример 4.3. Построить производственную функцию Кобба-Дугласа для оценки национального дохода США по статистическим данным, представленным в таблице 4.5: y – национальный доход США, млрд. долл., K – капиталовложения, млрд. долл., L – общее число занятых в экономике, тыс. чел. (источник данных: www.economagic.com).

Таблица 4.5. Статистические данные примера 4.3

Год	y	K	L
1992	6337,75	5512,75	120596
1993	6657,4	5773,35	122038
1994	7072,23	6122,25	122762
1995	7397,65	6453,93	124862
1996	7816,83	6840,1	126501
1997	8304,33	7292,18	129353
1998	8746,98	7752,8	131282
1999	9268,43	8236,65	133317
2000	9816,98	8795,23	136788
2001	10100,83	9881,23	13714
2002	10480,83	9290,85	122874
2003	10985,45	9600,47	137586

Решение:

Логарифмируя обе части уравнения $y = A \cdot K^{\alpha} \cdot L^{\beta} \cdot \varepsilon$, приходим к линейной модели $\ln y = \ln A + \alpha \cdot \ln K + \beta \cdot \ln L + \ln \varepsilon$. По статистическим

данным таблицы 4.5 рассчитаем значения логарифмов. Результаты вычислений сведем в таблицу 4.6.

Таблица 4.6. Значения логарифмов

$\ln y$	$\ln K$	$\ln L$
8,754279	8,614819	11,7002
8,803484	8,661008	11,71209
8,863931	8,719685	11,718
8,908918	8,772445	11,73496
8,964034	8,830558	11,74801
9,024532	8,894558	11,7703
9,076464	8,955809	11,7851
9,134369	9,016349	11,80049
9,191869	9,081965	11,82619
9,220373	9,198392	9,526172
9,257303	9,136785	11,71891
9,304327	9,169567	11,832

Параметризуя модель $\ln y = \ln A + \alpha \cdot \ln K + \beta \cdot \ln L + \ln \varepsilon$ по данным значений логарифмов таблицы 4.6 (например, с помощью табличного процессора MS Excel), найдем $\ln A = 0{,}826$ (тогда $A = 2{,}367$), $\alpha = 0{,}956$, $\beta = 0{,}129$.

Таким образом, производственная функция имеет вид $\tilde{y} = 2{,}367 \cdot K^{0{,}956} \cdot L^{0{,}129}$. Это означает, что при увеличении капиталовложений на 1% национальный доход США увеличивается на 0,956%, а при увеличении численности занятых в экономике на 1% национальный доход увеличивается на 0,129%.

Пример 4.4. По 20 предприятиям региона (таблица 4.7) изучается зависимость выработки продукции на одного работника y (млн руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%).

Таблица 4.7. Статистические данные примера 4.4

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7,0	3,9	10,0	11	9,0	6,0	21,0
2	7,0	3,9	14,0	12	11,0	6,4	22,0
3	7,0	3,7	15,0	13	9,0	6,8	22,0
4	7,0	4,0	16,0	14	11,0	7,2	25,0
5	7,0	3,8	17,0	15	12,0	8,0	28,0

6	7,0	4,8	19,0	16	12,0	8,2	29,0
7	8,0	5,4	19,0	17	12,0	8,1	30,0
8	8,0	4,4	20,0	18	12,0	8,5	31,0
9	8,0	5,3	20,0	19	14,0	9,6	32,0
10	10,0	6,8	20,0	20	14,0	9,0	36,0

Требуется:

1) Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2) Найти коэффициенты парной и множественной корреляции. Проанализировать их. Обосновать включение обоих факторов в модель или исключение одного из факторов.

3) Вычислить коэффициент детерминации R^2. С помощью F-критерия Фишера оценить статистическую надежность уравнения множественной регрессии.

4) С помощью Стьюдента *t*-статистики оценить статистическую значимость коэффициентов линейной множественной регрессии.

Решение:

1) Вычислим параметры a_0, a_1 и a_2 линейного уравнения $\tilde{y} = a_0 + a_1 x_1 + a_2 x_2$ множественной регрессии. Решая систему уравнений (4.3), получим, что $a_0 = 1{,}835$, $a_1 = 0{,}946$, $a_2 = 0{,}0856$. Таким образом, получили уравнение множественной регрессии $\tilde{y} = 1{,}835 + 0{,}946 x_1 + 0{,}0856 x_2$.

Коэффициенты β_1 и β_2 стандартизованного уравнения регрессии $\tilde{t}_y = \beta_1 t_{x_1} + \beta_2 t_{x_2}$ находятся по формулам (предварительно вычислим $\sigma_{x_1} = 1{,}890$, $\sigma_{x_2} = 6{,}642$, $\sigma_y = 2{,}396$):

$$\beta_1 = a_1 \frac{\sigma_{x_1}}{\sigma_y} = 0{,}946 \frac{1{,}890}{2{,}396} = 0{,}746,$$

$$\beta_2 = a_2 \frac{\sigma_{x_2}}{\sigma_y} = 0{,}0856 \frac{6{,}642}{2{,}396} = 0{,}237.$$

Следовательно, уравнение множественной линейной регрессии в стандартизованном масштабе имеет вид $\tilde{t}_y = 0{,}746 t_{x_1} + 0{,}237 t_{x_2}$. Так как стандартизованные коэффициенты регрессии можно сравнивать между собой, то можно сказать, что ввод в действие новых основных фондов

оказывает большее влияние на выработку продукции, чем удельный вес рабочих высокой квалификации.

Ранжировать влияние факторов на переменную y можно также при помощи средних коэффициентов эластичности. Они рассчитываются по формуле $\overline{Э}_{yx_j} = a_j \dfrac{\overline{x}_j}{\overline{y}}$, $i = 1,2$. Так как $\overline{x}_1 = 6,19$, $\overline{x}_2 = 22,3$, $\overline{y} = 9,6$, то $\overline{Э}_1 = a_1 \dfrac{\overline{x}_1}{\overline{y}} = 0,61$, $\overline{Э}_2 = a_2 \dfrac{\overline{x}_2}{\overline{y}} = 0,20$. Следовательно, увеличение основных фондов (от своего среднего значения) на 1% увеличивает в среднем выработку продукции на 0,61%. Увеличение удельного веса рабочих высокой квалификации (от своего среднего значения) на 1% увеличивает в среднем выработку продукции 0,20%. Таким образом, и средние коэффициенты эластичности подтверждают большее влияние на результат y фактора x_1 по сравнению с фактором x_2.

2) Рассчитаем сначала парные коэффициенты корреляции:

$$r_{yx_1} = \frac{\operatorname{cov}(y; x_1)}{\sigma_y \sigma_{x_1}} = \frac{\overline{yx_1} - \overline{y}\,\overline{x_1}}{\sigma_y \sigma_{x_1}} = 0,970,$$

$$r_{yx_2} = \frac{\operatorname{cov}(y; x_2)}{\sigma_y \sigma_{x_2}} = 0,941,\ r_{x_1x_2} = \frac{\operatorname{cov}(x_1; x_2)}{\sigma_{x_1} \sigma_{x_2}} = 0,943.$$

Они указывают на весьма сильную связь каждого фактора с результативным признаком y, а также высокую межфакторную зависимость (факторы x_1 и x_2 явно коллинеарны, т.к. $r_{x_1x_2} = 0,943 > 0,7$). При такой сильной межфакторной зависимости рекомендуется один из факторов исключить из рассмотрения.

Этим фактором должен быть фактор x_2, так как: 1) стандартизованный коэффициент для фактора x_2 меньше, чем для x_1; 2) средний коэффициент эластичности для фактора x_2 меньше, чем для x_1; 3) коэффициент корреляции r_{yx_2} меньше коэффициента корреляции r_{yx_1}.

Коэффициент множественной корреляции $R_{yx_1x_2}$ вычисляется по формуле

$$R_{yx_1x_2} = \sqrt{1 - \frac{\sigma^2_{\text{ост}}}{\sigma^2_y}} = \sqrt{1 - \frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}} = 0,973.$$

Он характеризует тесноту совместного влияния факторов на результат: совокупная связь факторов с результативным признаком весьма высокая. Значение коэффициент множественной корреляции незначительно отличается от коэффициента парной корреляции $r_{yx_2} = 0{,}970$.

Общий вывод состоит в том, что множественная модель с факторами x_1 и x_2 содержит малоинформативный фактор x_2. Если исключить фактор x_2, то можно ограничиться уравнением парной регрессии, которое имеет вид $\tilde{y} = 1{,}99 + 1{,}23x_1$.

3) Для вычисления коэффициента детерминации воспользуемся формулой $R^2 = R^2_{yx_1x_2}$. Величина $R^2 \cdot 100\%$ показывает, что изменения выработки продукции на одного работника на 94,7% объясняются изменением факторных признаков, включенных в модель.

Отметим, что в случае парной линейной связи коэффициент детерминации $r^2_{yx_1}$ равен 0,941. Таким образом, изменения выработки продукции на одного работника на 94,1% объясняются изменением фактора x_1.

Наблюдаемое значение статистики Фишера $F_{\text{набл}}$ вычисляется по выборочным данным на основании формулы $F_{\text{набл}} = \dfrac{R^2}{1-R^2} \cdot \dfrac{n-m-1}{m}$, где $m = 2$ – число объясняющих переменных в модели, а $n = 20$ – число наблюдений. В нашем случае имеем $F_{\text{набл}} = 151{,}7$.

По таблицам критических точек F-распределения находится критическое значение статистики $F_{\text{кр}} = 3{,}59$ при заданном уровне значимости $\alpha = 0{,}05$. При этом число степеней свободы определяется значениями $k_1 = m = 2$ и $k_2 = n - m - 1 = 20 - 2 - 1 = 17$. Так как $F_{\text{кр}} < F_{\text{набл}}$, то уравнение множественной регрессии признается статистически надежным.

4) Гипотеза о статистической значимости коэффициентов линейной множественной регрессии проверяется с помощью *t*-статистики, имеющей распределение Стьюдента с числом степеней свободы, равным $n - m - 1 = 20 - 2 - 1 = 17$.

Наблюдаемые значения *t*-статистики $t_{\text{набл}}$ для коэффициентов регрессии a_0, a_1 и a_2 равны 3,9, 4,45 и 1,42 соответственно. Табличное значение t-статистики $t_{\text{кр}}$ равно 2,1. Так как для коэффициентов регрессии a_0 и a_1 выполняется неравенство $|t_{\text{набл}}| > t_{\text{кр}}$, то признается, что коэффициенты a_0 и a_1 линейного уравнения регрессии не случайно отличаются от нуля, а значит, они статистически значимы. Так как для коэффициента a_2 выполняется неравенство $|t_{\text{набл}}| < t_{\text{кр}}$, то коэффициент регрессии a_2 статистически не значим. Это дает нам дополнительные основания для того, чтобы исключить

фактор x_2 из уравнения множественной регрессии и перейти к уравнению парной линейной регрессии с регрессором x_1.

Пример 4.5. На основании статистических данных, представленных в таблице 4.8, построена линейная модель $y = -3{,}54 + 0{,}854x_1 + 0{,}367x_2 + \varepsilon$, устанавливающая зависимость сменной добычи угля в шахтах на одного рабочего y (т) от мощности пласта x_1 (м) и уровня механизации x_2 (%).

По построенной модели осуществить точный прогноз добычи угля на одного рабочего в шахте, где мощность пласта составляет 8 м, а уровень механизации равен 6%. Вычислить доверительный интервал прогноза с доверительной вероятностью 0,95.

Таблица 4.8. Статистические данные задания 4.5

x_1	8	11	12	9	8	8	9	9	8	12
x_2	5	8	8	5	7	8	6	4	5	7
y	5	10	10	7	5	6	6	5	6	8

Решение:

Обозначим

$$X = \begin{pmatrix} 1 & 8 & 5 \\ 1 & 11 & 8 \\ 1 & 12 & 8 \\ 1 & 9 & 5 \\ 1 & 8 & 7 \\ 1 & 8 & 8 \\ 1 & 9 & 6 \\ 1 & 9 & 4 \\ 1 & 8 & 5 \\ 1 & 12 & 7 \end{pmatrix}, \; X_0 = \begin{pmatrix} 1 \\ 8 \\ 6 \end{pmatrix}.$$

Точечный прогноз найдем по уравнению регрессию:

$$y_p = -3{,}54 + 0{,}854 \cdot 8 + 0{,}367 \cdot 6 = 5{,}49.$$

Стандартную ошибку регрессии определим по формуле (4.4) $m_p = s \cdot \sqrt{1 + X_0'(X'X)^{-1}X_0}$, где $s = \sqrt{\frac{1}{n-(m+1)}\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2} = 0{,}951$.

Вычислив обратную матрицу $(X'X)^{-1}$ и выполнив последовательно умножение соответствующих матриц, получим $X_0'(X'X)^{-1}X_0 = 0{,}187$. Значит, $m_p = 0{,}951\sqrt{1+0{,}187} = 1{,}036$. Так как $t_{\text{кр}} = 2{,}36$, то доверительный интервал прогноза имеет вид $(5{,}49 - 2{,}36 \cdot 1{,}036; 5{,}49 - 2{,}36 \cdot 1{,}036)$, т.е. вид $(3{,}05; 7{,}93)$.

Следовательно, с доверительной вероятностью 0,95 значение сменной добычи угля в шахте, где мощность пласта составляет 8 м, а уровень механизации равен 6%, находится в пределах от 3,05 до 7,93 тонны.

Контрольные задания

Задание 4.1. В таблице 4.9 приведены статистические данные о продаже квартир на вторичном рынке жилья в некотором городе. При этом:

y – цена квартиры, тыс. долл.;

x_1 – число комнат в квартире;

x_2 – район города (1 – центральный, 0 – периферийный);

x_3 – общая площадь квартиры ($м^2$);

x_4 – жилая площадь квартиры ($м^2$);

x_5 – площадь кухни ($м^2$);

x_6 – тип дома (1 – кирпичный, 0 – другой);

x_7 – расстояние от метро, минут пешком.

По представленным статистическим данным необходимо определить факторы, формирующие цену квартир на вторичном рынке жилья города. Для этого:

1) Составить матрицу парных коэффициентов корреляции.

2) Построить линейное уравнение регрессии, характеризующее зависимость цены от всех факторов.

3) Установить, какие факторы явно коллинеарны.

4) Оценить значимость полученного уравнения регрессии и определить факторы, которые значимо воздействуют на формирование цены квартиры в этой модели?

5) Определить, существует ли разница в ценах квартир, расположенных в центральных и в периферийных районах?

6) Определить, существует ли разница в ценах квартир кирпичных и других типов домов?

7) Построить модель формирования цены квартиры за счет значимых факторов.

Таблица 4.9. Статистические данные задания 4.1

№ п/п	y	x_1	x_2	x_3	x_4	x_5	x_6	x_7
1	13,0	1	1	37,0	21,5	6,5	0	20
2	16,5	1	1	60,0	27,0	22,4	0	10
3	17,0	1	1	60,0	30,0	15,0	0	10
4	15,0.	1	1	53,0	26,2	13,0	0	15
5	14,2	1	1	35,0	19,0	9,0	0	8
6	10,5	1	1	30,0	17,5	5,6	1	15
7	23,0	1	1	43,0	25,5	8,5	0	5
8	12,0	1	1	30,0	17,8	5,5	1	10
9	15,6	1	1	35,0	18,0	5,3	1	3
10	12,5	1	1	32,0	17,0	6,0	1	5
11	11,3	1	0	31,0	18,0	5,5	1	10
12	13,0	1	0	33,0	19,6	7,0	0	5
13	21,0	1	0	53,0	26,0	16,0	1	5
14	12,0	1	0	32,0	18,0	6,3	0	20
15	11,0	1	0	31,0	17,3	5,5	1	15
16	11,0	1	0	36,0	19,0	8,0	1	5
17	22,5	2	1	48,0	29,0	8,0	1	15
18	26,0	2	1	55,5	35,0	8,0	0	10
19	18,5	2	1	48,0	28,0	8,0	0	10
20	13,2	2	1	44,1	30,0	6,0	1	25
21	25,8	2	1	80,0	51,0	13,0	0	10
22	17,0	2	1	60,0	38,0	10,0	0	12
23	18,0	2	0	50,0	30,0	8,7	1	15
24	21,0	2	0	54,6	32,0	10,0	1	20
25	14,5	2	0	43,0	27,0	5,5	1	10
26	23,0	2	0	66,0	39,0	12,0	1	5
27	19,5	2	0	53,5	29,5	7,0	1	15
28	14,2	2	0	45,0	29,0	6,0	1	12
29	13,3	2	0	45,0	30,0	5,5	0	5
30	16,1	2	0	50,6	30,8	7,9	0	10
31	13,5	2	0	42,5	28,0	5,2	1	25
32	16,0	2	0	50,1	31,0	6,0	0	10
33	15,5	3	1	68,1	44,4	7,2	0	5
34	38,0	3	1	107,0	58,0	24,0	0	15
35	30,0	3	1	100,0	58,0	20,0	0	15
36	24,0	3	1	71,0	52,0	7,5	1	15
37	32,5	3	1	98,0	51,0	15,0	0	10
38	43,0	3	0	100,0	45,0	35,0	1	25
39	17,8	3	0	58,0	39,0	6,2	0	10
40	28,0	3	0	75,0	40,0	18,0	1	3
41	32,7	3	0	85,0	59,0	9,0	0	5
42	31,0	3	0	66,0	48,0	6,0	0	2
43	33,0	3	0	81,0	52,0	12,0	0	10
44	28,0	3	0	76,4	49,0	10,0	0	5

45	21,5	3	0	55,0	40,5	6,0	1	15
46	15,3	3	0	53,7	37,6	5,5	1	3
47	21,0	3	0	57,0	38,0	6,3	0	7
48	35,5	3	0	62,0	52,0	8,0	0	3

Задание 4.2. В таблице 4.10 приведены данные о выработке литья на одного работающего x_1, браке литья x_2 и себестоимости одной тонны литья y по 25 литейным цехам заводов.

Таблица 4.10. Статистические данные задания 4.2

x_1	x_2	y
14,6	4,2	239
13,5	6,7	254
21,5	5,5	262
17,4	7,7	251
44,8	1,2	158
111,9	2,2	101
20,1	8,4	259
28,1	1,4	186
22,3	4,2	204
25,3	0,9	198
56,0	1,3	170
40,2	1,8	173
40,6	3,3	197
75,8	3,4	172
27,6	1,1	201
88,4	0,1	130
16,6	4,1	251
33,4	2,3	195
17,0	9,3	282
33,1	3,3	196
30,1	3,5	186
65,2	1,0	176
22,6	5,2	238
33,4	2,3	204
19,7	2,7	205

Необходимо:

1) найти множественный коэффициент детерминации и пояснить его смысл;

2) найти уравнение множественной линейной регрессии y по x_1 и x_2, оценить значимость этого уравнения и его коэффициентов на уровне α = 0,05;

3) сравнить раздельное влияние на зависимую переменную каждой из объясняющих переменных, используя стандартизованные коэффициенты регрессии и коэффициенты эластичности;

4) найти 95%-ные доверительные интервалы для коэффициентов регрессии;

5) найти точечный и интервальный прогнозы себестоимости одной тонны литья в цехах, в которых выработка литья на одного работающего составляет 40 тонн, а брак литья – 5%.

Задание 4.3. В таблице 4.11 приведены статистические данные о годовых ставках месячных доходов по трем акциям за шестимесячный период.

Таблица 4.11. Статистические данные задания 4.3

Акция	Доходы по месяцам, %					
A	5,4	5,3	4,9	4,9	5,4	6,0
B	6,3	6,2	6,1	5,8	5,7	5,7
C	9,2	9,2	9,1	9,0	8,7	8,6

Есть основания предполагать, что доходы y по акции С зависят от доходов x_1 и x_2 по акциям А и В. Необходимо:

1) составить уравнение множественной линейной регрессии y по x_1 и x_2;

2) найти множественный коэффициент детерминации R^2 и пояснить его смысл;

3) проверить значимость полученного уравнения регрессии на уровне α=0,05;

4) оценить средний доход по акции С, если доходы по акциям А и В составили 5,5% и 6,0% соответственно.

Задание 4.4. В таблице 4.12 приведены статистические данные о расходах на питание, душевом доходе и размере семьи для девяти групп семей.

Таблица 4.12. Статистические данные задания 4.4

Расход на питание y, у.е.	Душевой доход x_1, у.е.	Размер семей x_2
426	611	1,4
611	1520	1,6
870	2580	1,8
1050	36015	2
1290	4585	2,2

1348	5820	2,4
1525	7110	2,5
1820	9140	2,7
2342	15980	3

Необходимо:

1) составить уравнение множественной линейной регрессии y по x_1 и x_2;

2) найти множественный коэффициент детерминации R^2 и пояснить его смысл;

3) проверить значимость полученного уравнения регрессии на уровне α=0,05;

4) найти точечный и интервальный прогнозы расходов на питание, если душевой доход и размер семей составили 4950 и 2,3 соответственно.

Задание 4.5. Применяя тест Чоу, установить, является ли существенным различие в оплате труда работников государственных и коммерческих предприятий. Данные об оплате труда в зависимости от стажа, возраста и образования работников государственных и коммерческих предприятий приведены в таблицах 4.13 и 4.14.

Таблица 4.13. Данные задания 4.5 по государственным предприятиям

Зарплата, у.е.	Стаж	Возраст	Образование (1 – высшее; 0 – среднее)
596	15	40	1
524	6	31	0
610	16	46	1
756	30	53	1
810	25	48	1
750	20	44	1
480	28	50	0
623	13	42	0
785	25	46	1
761	18	49	1
640	31	54	1
321	2	24	0

Таблица 4.14. Данные задания 4.5 по коммерческим предприятиям

Зарплата, у.е.	Стаж	Возраст	Образование (1 – высшее; 0 – среднее)
840	20	46	1
522	19	42	0
649	1	23	1
545	12	46	0

924	30	56	1
843	14	35	1
878	28	50	1
621	18	40	0
802	4	30	1
920	22	47	1
442	20	46	0
587	30	57	0
796	18	42	1
388	5	30	0

Задание 4.6. В таблице 4.15 по 15 предприятиям отрасли приведены данные о выпуске продукции Q в зависимости от капиталовложений K и трудозатрат L. Построить производственную модель Кобба-Дугласа $Q = A \cdot K^{\alpha} \cdot L^{\beta} \cdot \varepsilon$. Оценить общее качество модели. Рассчитать по модели выпуск продукции при $K = 1780$ и $L = 2450$.

Таблица 4.15. Статистические данные задания 4.6

Q	K	L
2410	1560	2342
2485	1840	2432
2112	1165	2220
2550	1940	2460
2640	2430	2565
2245	1320	2275
2420	1710	2370
2525	1850	2435
2560	1870	2445
2455	1780	2401
2288	1465	2305
2155	1230	2250
2390	1650	2365
2480	1850	2438
2602	1990	2460

Задание 4.7. По статистическим данным, представленным в таблице 4.16, о зависимости спроса на некоторый товар y у домашних хозяйств в зависимости от цены этого товара x_1 и дохода домохозяйств x_2:

1) составить уравнение линейной регрессии y по x_1 и x_2;

2) найти множественный коэффициент детерминации R^2;

3) проверить значимость уравнения регрессии на уровне α=0,05;

4) найти точечный прогноз спроса, если цена товара и доход домохозяйства составили 5,4 и 965 соответственно.

Таблица 4.16. Статистические данные задания 4.7

y	x_1	x_2
31,4	4,1	1050
30,4	4,2	1010
32,1	4,0	1070
30,0	4,6	1060
30,5	4,0	1000
29,8	5,0	1040
31,1	3,9	1030
31,7	4,4	1080
30,7	4,5	1050
29,7	4,8	1020

Задание 4.8. По статистическим данным по 8 магазинам, представленным в таблице 4.17, о зависимости товарооборота магазина y от численности работающих x_1 и площади подсобных помещений x_2:

1) составить уравнение множественной линейной регрессии y по x_1 и x_2;

2) найти множественный коэффициент детерминации R^2;

3) проверить значимость полученного уравнения регрессии на уровне α=0,05;

4) оценить относительный вклад в величину товарооборота факторов x_1 и x_2.

Таблица 4.17. Статистические данные задания 4.8

y	x_1	x_2
22,1	31	29,5
14,2	34	14,2
23,3	35	18,0
43,2	41	21,3
66,6	38	47,5
7,8	32	10,0
12,7	29	21,0
36,1	34	36,5

Задание 4.9. По 20 предприятиям региона (таблица 4.18) изучается зависимость выработки продукции на одного работника y (млн руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на

конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%).

Таблица 4.18. Статистические данные примера 4.9

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	6,0	3,6	9,0	11	9,0	6,3	21,0
2	6,0	3,6	12,0	12	11,0	6,4	22,0
3	6,0	3,9	14,0	13	11,0	7,0	24,0
4	7,0	4,1	17,0	14	12,0	7,5	25,0
5	7,0	3,9	18,0	15	12,0	7,9	28,0
6	7,0	4,5	19,0	16	13,0	8,2	30,0
7	8,0	5,3	19,0	17	13,0	8,0	30,0
8	8,0	5,3	19,0	18	13,0	8,6	31,0
9	9,0	5,6	20,0	19	14,0	9,5	33,0
10	10,0	6,8	21,0	20	14,0	9,0	36,0

Требуется:

1) Построить линейную модель множественной регрессии. Записать стандартизованное уравнение множественной регрессии. На основе стандартизованных коэффициентов регрессии и средних коэффициентов эластичности ранжировать факторы по степени их влияния на результат.

2) Найти коэффициенты парной и множественной корреляции. Проанализировать их. Обосновать включение обоих факторов в модель или исключение одного из факторов. В случае исключения из модели множественной регрессии одного из факторов построить парную линейную модель.

3) Вычислить коэффициент детерминации R^2. С помощью F-критерия Фишера оценить статистическую надежность уравнения множественной регрессии.

4) С помощью Стьюдента *t*-статистики оценить статистическую значимость коэффициентов линейной множественной регрессии.

Задание 4.10. По десяти кредитным учреждениям получены данные (таблица 4.19), характеризующие зависимость объема прибыли (y, млрд руб.) от среднегодовой ставки по кредитам (x_1, %) , ставки по депозитам (x_2, %) и размера внутрибанковских расходов (x_3, млрд руб.).

Таблица 4.19. Статистические данные примера 4.10

Объем прибыли, y	Среднегодовая ставка по кредитам, x_1	Ставка по депозитам, x_2	Внутрибанковские расходы, x_3

50	22	17	150
54	24	17	154
60	20	15	146
62	30	17	134
70	32	16	132
54	32	16	126
84	30	16	134
82	33	15	126
86	34	15	88
84	28	14	120

Требуется:

1. Осуществить выбор факторных признаков для построения двухфакторной линейной регрессионной модели.
2. Рассчитать параметры этой модели.
3. Для характеристики модели определить линейный коэффициент множественной корреляции, коэффициент детерминации, средние коэффициенты эластичности. Интерпретировать полученные характеристики модели.
4. Осуществить оценку надежности уравнения регрессии.
5. Оценить статистическую значимость коэффициентов уравнения множественной регрессии.

Задание 4.11. Некоторая фирма, занимающаяся торговлей компьютеров и ноутбуков, определила, что на количество продаж y основное влияние оказывают факторы: цена товара x_1 (млн руб.), затраты на рекламу x_2 (млн руб.) и число конкурирующих фирм в регионе x_3.

Таблица 4.20. Статистические данные примера 4.11

Количество продаж, y	Цена товара, x_1	Затраты на рекламу, x_2	Число конкурентов, x_3
112	5,1	51	16
132	5,0	44	15
129	4,9	48	17
134	5,0	50	15
132	4,8	39	14
137	4,8	45	14
139	4,6	37	14
139	4,7	45	13
138	4,7	40	11
143	4,6	29	12
141	4,5	36	11
146	4,5	44	11

148	4,6	31	12
152	4,4	38	13

По статистическим данным, представленным в таблице 4.21, необходимо:

1) Построить линейное уравнение множественной регрессии, которое включает все три фактора. Интерпретировать коэффициенты уравнения регрессии.

2) На основании анализа корреляционной матрицы из трех независимых переменных отобрать два наиболее существенных фактора. Для отобранных факторов построить двухфакторное уравнение линейной регрессии.

3) Для характеристики двухфакторной модели определить линейный коэффициент множественной корреляции, коэффициент детерминации, средние коэффициенты эластичности. Интерпретировать полученные характеристики модели.

4) Осуществить оценку надежности уравнения. Оценить статистическую значимость коэффициентов уравнения множественной регрессии.

5) Осуществить точечный и интервальный прогнозы, если цена товара $x_1 = 4{,}5$, затраты на рекламу $x_2 = 40$, а число конкурентов $x_3 = 15$.

Задание 4.12. Предприятие выпускает продукцию, ежемесячный объем которой y (тыс. шт.) зависит от затрат материальных ресурсов x_1 (тонны), трудозатрат x_2 (тыс. часов) и энергозатрат x_3 (млн кВт). В процессе развития производства наблюдалась эмпирическая зависимость между выпуском продукции y и затратами ресурсов x_1, x_2 и x_3, отраженная в таблице 4.21. Построить производственную функцию $y = A \cdot x_1^{\alpha} \cdot x_2^{\beta} x_3^{\gamma} \cdot \varepsilon$, оценить ее общее качество с помощью индекса детерминации и осуществить точечный прогноз, если затраты материальных ресурсов $x_1 = 4{,}5$, трудозатраты $x_2 = 40$, а энергозатраты $x_3 = 15$.

Таблица 4.21. Статистические данные примера 4.12

Объем продукции, y	Затраты материальных ресурсов, x_1	Трудозатраты, x_2	Энергозатраты, x_3
45,0	16,0	50,3	7,7
50,3	20,4	55,6	6,9
54,1	18,0	58,4	7,8
55,1	22,1	50,8	8,6
60,8	21,2	57,5	10,1
65,6	24,2	59,3	8,4
68,8	27,1	62,0	9,3
66,6	26,6	64,7	7,4

73,2	28,3	59,9	11,3
81,9	31,2	64,6	10,7
91,8	35,5	59,8	12,4
86,1	34,6	62,3	11,7
83,1	33,7	65,2	9,9
93,1	34,2	70,1	13,0

Задание 4.13. Торговый дом стройматериалов продает облицовочную плитку в двух городах: Гомеле и Могилеве. Маркетинговая служба хочет определить влияние отчислений на рекламу x (млн руб.) на количество проданной продукции y (млн штук). При этом предполагается, что зависимость между факторами x и y линейная и степень влияния факторов друг на друга в обоих городах примерно одинакова. На основании статистических данных, приведенных в таблице 4.22, построить регрессионную модель зависимости количества проданной продукции от отчислений на рекламу, включающую качественный фактор «город». Интерпретировать полученные результаты.

Таблица 4.22. Статистические данные примера 4.13

Гомель											
x	33,1	33,1	33,4	41,0	35,5	31,4	45,2	45,4	30,7	40,4	33,3
y	13,3	18,4	19,1	24,9	21,5	17,9	31,6	29,7	16,4	27,4	22,8
Могилев											
x	22,8	20,6	18,3	25,5	28,1	35,2	32,5	27,8	26,6	31,4	27,4
y	16,3	15,4	11,8	19,5	27,4	31,8	29,1	22,1	19,2	26,4	21,7

Задание 4.14. На основании статистических данных, приведенных в таблице 4.23, построить линейную модель зависимости стоимости квартиры y от ее общей площади x_1, жилой площади x_2 и расстояния от квартиры до метро x_3. Оценить статистическую значимость уравнения регрессии и ее коэффициентов. Сделать выводы. Найти точечный прогноз стоимости квартиры, если общая площадь, жилая площадь и расстояние от квартиры до метро составляют 90, 61 и 12 соответственно.

Таблица 4.23. Статистические данные задания 4.14

Общая площадь, кв. м, x_1	Жилая площадь, кв. м, x_2	Расстояние до метро, минут пешком, x_3	Стоимость, тыс. дол. y
80	64	3	66
62	37	8	52
70	52	18	53
79	55	28	50
96	71	8	64

90	62	8	62
102	78	7	85
87	66	10	56
112	78	10	82
116	82	8	83
90	62	8	70
118	82	12	93
107	78	10	80
93	66	18	59
96	68	8	68
92	72	10	56
74	49	18	53
106	76	10	46
88	61	3	36
74	48	11	24
74	54	10	28
118	76	8	54
92	62	18	30
110	80	8	45

Контрольные вопросы

1. Дайте определение множественной регрессионной модели.
2. Приведите примеры использования множественных регрессионных моделей в экономике.
3. Сформулируйте общую постановку задачи множественного эконометрического моделирования.
4. Каким требованиям должны удовлетворять факторы множественной регрессии?
5. Какие задачи решаются на этапе спецификации модели множественной линейной регрессии?
6. Опишите этапы отбора факторов множественной регрессии.
7. Как определяется матрица парных коэффициентов корреляции?
8. Опишите метод отбора факторов, основанный на анализе матрицы парных коэффициентов корреляции.
9. Какой вид имеет линейная модель множественной регрессии?
10. Приведите примеры нелинейных моделей множественной регрессии.
11. Что характеризуют коэффициенты уравнения множественной линейной регрессии?
12. В чем суть МНК для построения линейной модели множественной регрессии?
13. Сформулируйте теорему Гаусса-Маркова.
14. Какая модель множественной линейной регрессии называется классической?
15. Опишите процедуру параметризации модели Кобба-Дугласа.

16. Какие задачи решаются на этапе верификации модели множественной линейной регрессии?

17. По какой формуле вычисляется коэффициент множественной корреляции?

18. Запишите уравнение множественной линейной регрессии в стандартизованном масштабе.

19. Как можно ранжировать факторы по силе их воздействия на результат по величине стандартизованных коэффициентов линейной регрессии?

20. Как вычисляются средние коэффициенты эластичности для линейной множественной регрессии?

21. Как вычисляется коэффициент детерминации R^2 в случае множественной регрессии?

22. Что оценивает коэффициента детерминации R^2?

23. Как вычисляется и что оценивает показатель средней ошибки аппроксимации?

24. Как осуществляется анализ статистической значимости коэффициента детерминации?

25. Как определяется статистическая значимость коэффициентов множественной линейной регрессии?

26. В чем суть статистической значимости коэффициентов множественной линейной регрессии?

27. Перечислите предпосылки МНК.

28. Каковы последствия выполнимости или невыполнимости предпосылок МНК?

29. Как по множественной модели осуществляется точечный прогноз?

30. Приведите примеры зависимостей, содержащих качественные факторы.

31. Какие переменные называются значащими, а какие – фиктивными?

32. Какие переменные называются дихотомическими?

33. В чем проявляется необходимость введения фиктивных переменных?

34. Как дихотомическая переменная вводится в эконометрическую модель?

35. Почему при построении регрессионной модели, учитывающей фактор с k уровнями качества, вводится k-1 фиктивная переменная, а не k переменных?

36. В чем проявляется «ловушка фиктивной переменной»?

37. Для решения каких задач применяется тест Чоу?

38. В чем суть теста Чоу?

39. Изложите графическую интерпретацию возможных выводов, полученных на основании теста Чоу?

40. Как с помощью фиктивных переменных в регрессионной модели учитывается сезонность?

Тестовые задания

Выберите правильные ответы из предложенных вариантов:

1. Множественная регрессия используется, если:
а) между факторами существуют нелинейные соотношения;
б) в уравнение необходимо включить два и более фактора;
в) имеется доминирующий фактор.

2. Зависимость расходов на продукты питания y (тыс. руб.) от месячного дохода на одного члена семьи (x_1, тыс. руб.) и размера семьи (x_2, человек) характеризуется уравнением $\tilde{y} = 0{,}6 + 0{,}32 \cdot x_1 + 0{,}81 \cdot x_2$. Тогда:
а) с ростом дохода на одного члени семьи на 1 тыс. рублей расходы на питание возрастут в среднем на 0,32 тыс. руб. при том же среднем размере семьи;
б) с ростом дохода на одного члени семьи на 1 тыс. рублей расходы на питание возрастут в среднем на 0,81 тыс. руб. при том же среднем размере семьи;
в) с ростом дохода на одного члени семьи на 1 тыс. руб. расходы на питание возрастут в среднем на 0,6 тыс. рублей при том же среднем размере семьи.

3. Моделью множественной регрессии является:
а) модель $y = a + bx + \varepsilon$;
б) модель $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$;
в) модель $y = a_0 \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot ... \cdot x_m^{a_m} \cdot \varepsilon$;
г) модель $y = \dfrac{1}{a_0 + a_1 x_1} + \varepsilon$.

4. Моделью множественной линейной регрессии является:
а) модель $y = a + bx + \varepsilon$;
б) модель $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$;
в) модель $y = a_0 \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot ... \cdot x_m^{a_m} \cdot \varepsilon$;
г) модель $y = \dfrac{1}{a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m} + \varepsilon$.

5. Укажите требования к факторам, включаемым в модель множественной линейной регрессии:
а) между факторами не должна существовать высокая корреляция;
б) факторы должны быть количественно измеримы;
в) факторы должны иметь одинаковую размерность;
г) факторы должны представлять временные ряды.

6. Матрица парных коэффициентов корреляции используется:

а) для построения уравнения регрессии;
б) для отбора факторов регрессионной модели;
в) для параметризации модели;
г) для линеаризации модели.

7. Если в модели множественной регрессии опущена переменная, которая должна быть включена, то оценки регрессии:
а) часто оказываются смещенными;
б) могут быть неэффективными;
в) могут быть неэластичными;
г) часто оказываются нелинейными.

8. Уравнение множественной регрессии называется степенным, если оно имеет вид:
а) $\tilde{y} = a_0 \cdot x_1^{a_1} \cdot x_2^{a_2} \cdot \ldots \cdot x_m^{a_m}$;
б) $\tilde{y} = e^{a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m}$;
в) $\tilde{y} = a_0 + \frac{a_1}{x_1} + \frac{a_2}{x_2} + \ldots + \frac{a_m}{x_m}$;
г) $\tilde{y} = \frac{1}{a_0 + a_1 x_1 + a_2 x_2 + \ldots + a_m x_m}$.

9. Параметры уравнения множественной регрессии оцениваются:
а) двухшаговым методом наименьших квадратов;
б) косвенным методом наименьших квадратов;
в) методом наименьших квадратов.

10. Относительная сила влияния факторов на результативный признак оценивается:
а) индексом множественной корреляции;
б) частными коэффициентами эластичности;
в) параметрами при независимых переменных.

11. Верификация множественной регрессионной модели не включает:
а) проверку статистической значимости коэффициентов регрессии;
б) проверку общего качества уравнения регрессии;
в) проверку выполнимости предпосылок МНК;
г) проверку тесноты связи.

12. Оценка статистической значимости коэффициентов множественной линейной регрессии осуществляется с помощью:
а) критерия Стьюдента;
б) критерия Фишера;

в) критерия Пирсона;

г) критерия Дарбина-Уотсона.

13. Линейное уравнение множественной регрессии имеет вид $\tilde{y} = 20 + x_1 - 2x_2$. Определите какой из факторов x_1 или x_2 оказывает более сильное влияние на *y*:

а) x_1;

б) оказывают одинаковое влияние;

в) x_2;

г) по этому уравнению нельзя ответить на поставленный вопрос, так как коэффициенты регрессии несравнимы между собой.

14. Для уравнения линейной регрессии рассчитаны стандартизованные коэффициенты регрессии: $\beta_1 = 0{,}57$, $\beta_2 = 0{,}26$ и $\beta_3 = 0{,}045$. Какой фактор сильнее влияет на результативный признак:

а) первый; б) второй; в) третий.

15. Из пары коллинеарных факторов в эконометрическую модель включается тот фактор:

а) который при достаточно тесной связи с результатом имеет наибольшую связь с другими факторами;

б) который при отсутствии связи с результатом имеет максимальную связь с другими факторами;

в) который при отсутствии связи с результатом имеет наименьшую связь с другими факторами;

г) который при достаточно тесной связи с результатом имеет меньшую связь с другими факторами.

16. Фактор x_j линейно не связан с зависимой переменной и его рекомендуется исключить из уравнения множественной линейной регрессии, если для него:

а) $|t_{\text{набл}}| > t_{\text{кр}}$;

б) $|t_{\text{набл}}| < t_{\text{кр}}$;

в) $t_{\text{набл}} \in (-t_{\text{кр}}, t_{\text{кр}})$.

17. Оценка общего качества уравнения множественной регрессии осуществляется:

а) с помощью индекса множественной корреляции;

б) с помощью индекса детерминации;

в) по величине коэффициентов уравнения регрессии.

18. Проверка статистической значимости коэффициента детерминации осуществляется с помощью:

а) критерия Стьюдента;
б) критерия Фишера;
в) критерия Пирсона;
г) критерия Дарбина-Уотсона.

19. Величину изменения зависимой переменной с изменением на 1% фактора x_j при фиксированном значении других факторов в множественной линейной модели оценивает:
а) коэффициент уравнения регрессии при x_j;
б) средний коэффициент эластичности;
в) индекс множественной корреляции.

20. Два фактора x и y явно коллинеарны, если:
а) $r_{xy} \geq 0{,}3$;
б) $r_{xy} \geq 0{,}5$;
в) $r_{xy} \geq 0{,}7$;
г) $r_{xy} \leq 0{,}6$.

21. Построена линейная множественная регрессионная модель $y = 1 - 2x_1 + 2x_2 + \varepsilon$. Точечный прогноз по этой модели при $x_1 = 1$ и $x_2 = 1$ составляет:
а) 1; б) 0; в) -1; г) 4.

22. Добавление в уравнение множественной регрессии новой объясняющей переменной:
а) уменьшает значение коэффициента детерминации;
б) увеличивает значение коэффициента детерминации;
в) не оказывает никакого влияния на коэффициент детерминации.

23. Множественный коэффициент корреляции $R_{yx_1x_2}$ равен 0,9. Определите, какой процент дисперсии зависимой переменной y объясняется влиянием факторов x_1 и x_2:
а) 90%; б) 81%; в) 19%.

24. Для построения модели линейной множественной регрессии вида $y = a_0 + a_1x_1 + a_2x_2 + \varepsilon$ необходимое количество наблюдений должно быть не менее:
а) 2; б) 7; в) 14.

25. Стандартизованные коэффициенты регрессии β_i:
а) позволяют ранжировать факторы по силе их влияния на результат;

б) оценивают статистическую значимость факторов;
в) являются коэффициентами эластичности.

26. Фиктивные переменные – это:
а) атрибутивные признаки (например, как профессия, пол, образование), которым придали цифровые метки;
б) экономические переменные, принимающие количественные значения в некотором интервале;
в) значения зависимой переменной за предшествующий период времени.

27. Если качественный фактор имеет три градации, то необходимое число фиктивных переменных:

а) 4; б) 3; в) 2.

Ответы тестовых заданий

Номер задания	Ответы	Номер задания	Ответы	Номер задания	Ответы
1	б)	10	а)	19	б)
2	а)	11	г)	20	в)
3	б), в)	12	а)	21	а)
4	б)	13	г)	22	б)
5	а), б)	14	а)	23	а)
6	б)	15	г)	24	в)
7	а)	16	б)	25	а)
8	а)	17	б)	26	а)
9	в)	18	б)	27	в)

Глава 5
Эконометрический анализ классических модельных предположений

Основные понятия: *модельные предположения, несмещенная оценка, эффективная оценка, состоятельная оценка, гомоскедастичность, гетероскедастичность, тест Голдфелда-Квандта, тест ранговой корреляции Спирмена, взвешенный метод наименьших квадратов, автокорреляция, критерий Дарбина–Уотсона, обобщенный метод наименьших квадратов, преобразование Бокса-Дженкинса, полная и частичная мультиколлинеарность, определитель матрицы межфакторной корреляции.*

Литература: [2-4], [9], [13], [15-16].

5.1. О необходимости проверки модельных предположений

Метод наименьших квадратов является одним из самых популярных в эконометрике. Это связано не только с тем, что его практическая реализация встроена в большинство статистических и эконометрических компьютерных программ, но и со следующими соображениями.

Статистическая значимость коэффициентов регрессии и близкое к единице значение коэффициента детерминации R^2 еще не гарантируют высокое качество регрессионной модели. Даже при таких условиях оценки параметров a_0, a_1, …, a_m линейной модели $y = a_0 + a_1 x_1 + a_2 x_2 + ... + a_m x_m + \varepsilon$ не всегда являются надежными. Во многом это объясняется тем, что они зависят от случайной составляющей ε, которая в отличие от случайной величины y является ненаблюдаемой. Ненаблюдаемость ε и не позволяет в общем случае делать выводы о точности и достоверности оценок параметров регрессии.

Поэтому для получения надежных (несмещенных, эффективных и состоятельных) оценок параметров регрессии, мы вынуждены при моделировании зависимости y от факторов $x_1, x_2, ..., x_m$ потребовать для ε выполнения ряда дополнительных условий. Эти условия (таблица 5.1) называются *классическими предпосылками МНК* или *модельными предположениями* (они также известны как *условия Гаусса-Маркова*).

Как утверждает теорема Гаусса-Маркова, при совокупном выполнении условий 1-4 метод наименьших квадратов для линейной относительно параметров модели дает наилучшие из всех возможных результаты: оценки параметров регрессии являются несмещенными, состоятельными и эффективными, а модель адекватной и надежной. Эти свойства оценок имеют чрезвычайно важное практическое значение (в частности, для принятия решений и прогнозирования).

Несмещенность гарантирует правдоподобность результатов: математическое ожидание оценки каждого параметра регрессии равно его истинному значению, т.е. оценки центрируются вокруг значений истинных коэффициентов.

Эффективность обеспечивает точность: оценки коэффициентов регрессии наиболее компактно группируются вокруг истинных значений параметров. Никакой другой метод оценки коэффициентов не дает меньшую дисперсию для каждого из оцененных коэффициентов, чем МНК. В практических исследованиях свойство несмещенности оценки, дополненное свойством эффективности, создает возможность перехода от точечного оценивания к интервальному.

Состоятельность характеризует увеличение точности оценок с увеличением объема выборки. С ростом числа наблюдений, дисперсия становится меньше, и каждая оценка приближается к истинному значению параметра.

Таблица 5.1. Модельные предположения

1 условие	Математическое ожидание случайной переменной ε равно нулю	$M(\varepsilon) = 0$
2 условие	Дисперсия случайной переменной ε постоянна для всех наблюдений	$D(\varepsilon) = \sigma^2 = const$
3 условие	Отсутствует систематическая связь между значениями случайной переменной ε для любых двух наблюдений	$\mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0$
4 условие	Случайная переменная ε независима от объясняющих переменных $x_1, x_2, ..., x_m$	$\mathrm{cov}(x_i, \varepsilon_j) = 0$
5 условие	Случайная переменная ε имеет нормальный закон распределения вероятностей с нулевым математическим ожиданием и постоянной дисперсией	$\varepsilon \in N(0, \sigma^2)$

Если не выполняются условия 1 или 4, то появляется систематическое смещение; если не выполняются условия 2 или 3 – оценки становятся неэффективными. В обоих случаях модель некорректна.

Условие 4, кроме того, позволяет утверждать, что величина y состоит из двух составляющих: объясняемой и случайной. Невыполнение условия 4, в частности, не дает возможности при анализе общей дисперсии разграничивать вклады объясняющих переменных и случайных факторов. Условие 4 имеет значение, если факторные переменные являются случайными величинами. Оно автоматически выполняется, если переменные $x_1, x_2, ..., x_m$ являются неслучайными величинами.

Условие 5 необходимо для проведения проверки статистических гипотез и определения доверительных интервалов прогноза и коэффициентов регрессии. Невыполнение этого условия приводит к отказу от использования тестов.

Не следует забывать и о такой предпосылке МНК как правильность спецификации. Под этим понимается следующее:

1) В модели отсутствует недоопределённость (не упущены важные факторы) и переопределённость (не включены ненужные факторы).

2) Модель адекватна устройству данных. Например, если точки наблюдений явно расположены вдоль невидимой экспоненты, логарифма или любой нелинейной функции, то нет смысла строить линейное уравнение регрессии.

В случае множественной регрессии важной предпосылкой МНК является также условие отсутствия в модели мультиколлинеарности.

Если игнорировать проверку выполнимости модельных предположений, то регрессионная модель может оказаться статистически незначимой, а значит, прогнозы по ней будут подвергаться сомнению.

В связи с этим и возникает необходимость рассмотрения методов

обнаружения и устранения нарушений предпосылок МНК. Проверка их является важным и неотъемлемым этапом верификации регрессионной модели.

5.2. Первое модельное предположение

Если в регрессионное уравнение включен свободный член, то условие 1 (допущение о равенстве нулю математического ожидания случайной переменной) никогда не нарушается. В тех же случаях, когда возникает необходимость рассмотрения уравнения регрессии с нулевым свободным членом, первое модельное предположение может быть нарушено. Поэтому для устранения проблем с первой предпосылкой в модель регрессии следует включать свободный член.

Выполнение этой предпосылки может быть протестировано разными методами. Один из них заключается в визуальном анализе графика зависимости остатков $e_i = y_i - \tilde{y}_i$ результативного признака y. Если точки графика разбросаны хаотично по отношению к оси абсцисс, то первая предпосылка выполняется. Если же такой разброс имеет системный характер, то это дает основание для того, чтобы усомниться в выполнимости предпосылки, а следовательно, и в правильности спецификации модели.

Второй подход является статистическим. Как известно, несмещенной оценкой математического ожидания случайной величины является выборочное среднее. Поэтому для проверки предположения о равенстве нулю математического ожидания случайной переменной достаточно оценить величину среднего выборочного остатков регрессии $e_i = y_i - \tilde{y}_i$, $i = 1,2,...,n$, т.е. величину $\bar{e} = \frac{1}{n}\sum_{i=1}^{n} e_i$.

Можно также по остаткам e_i с помощью t-статистики проверить нулевую гипотезу о равенстве нулю математического ожидания случайной переменной ε (если она распределена по нормальному закону). Для этого сравниваются наблюдаемое значение $t_{\text{набл}} = \frac{|\bar{e}|\sqrt{n}}{s}$, где $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{e})^2}$ – стандартное отклонение, и критическое значение $t_{\text{кр}}$ статистики Стьюдента с числом степеней свободы, равным $n-1$. Если $|t_{\text{набл}}| < t_{\text{кр}}$, то нулевая гипотеза о равенстве нулю математического ожидания случайной переменной ε принимается, а значит, условие 1 выполняется.

5.3. Проблема гетероскедастичности

Предположение о постоянстве и конечности дисперсии остатков называется *свойством гомоскедастичности* остатков (рисунок 5.1). В практических исследованиях это свойство случайной ошибки модели регрессии не всегда выполняется и дисперсия остатков не является

постоянной величиной (рисунок 5.2). Такое явление называется *гетероскедастичностью*.

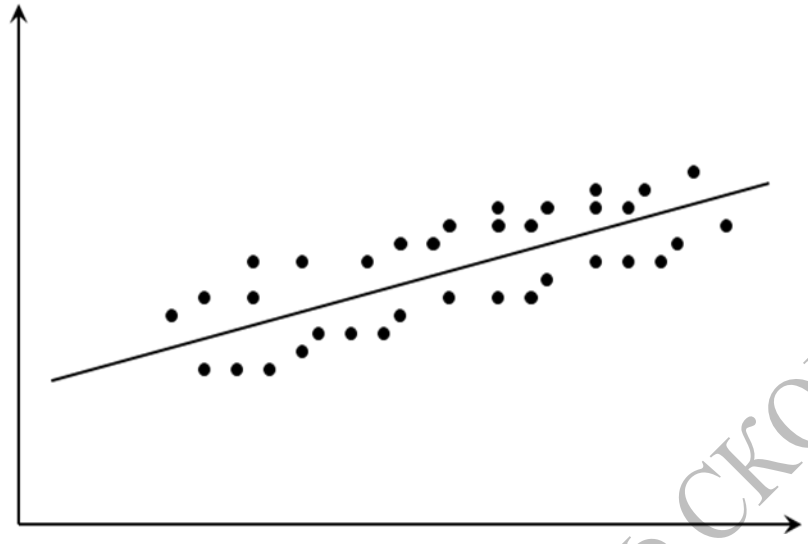


Рис. 5.1. Линейная модель с гомоскедастичностью

Гетероскедастичность часто вызывается ошибками спецификации, когда не учитывается в модели существенная переменная.

Гетероскедастичность приводит к тому, что оценки коэффициентов регрессии не являются эффективными, т.е. их дисперсии не будут наименьшими. Как следствие рассчитанные значения стандартных ошибок коэффициентов регрессии могут быть заниженными, а потому при проверке статистической значимости коэффициентов может быть ошибочно принято решение об их значимом отличии от нуля, тогда как на самом деле это не так.

Проблема гетероскедастичности характерна для пространственных данных, полученных от неоднородных объектов. Например, если исследуется зависимость расходов на питание в семье от ее общего дохода, то можно ожидать, что разброс данных будет выше для семей с более высоким доходом. Если исследуется зависимость оплаты труда сотрудников предприятий в зависимости от размера основных фондов предприятий и разряда работника, то понятно, что вариация оплаты труда на крупных предприятиях у сотрудников высокого разряда будет значительно превосходить его вариацию для сотрудников низких уровней на малых и средних предприятиях.

Гетероскедастичность иногда возникает и во временных рядах. Это происходит в тех случаях, когда зависимая переменная имеет большой интервал качественно неоднородных значений или высокий темп изменения (инфляция, технологические сдвиги, изменения в законодательстве, потребительские предпочтения и т.д.).

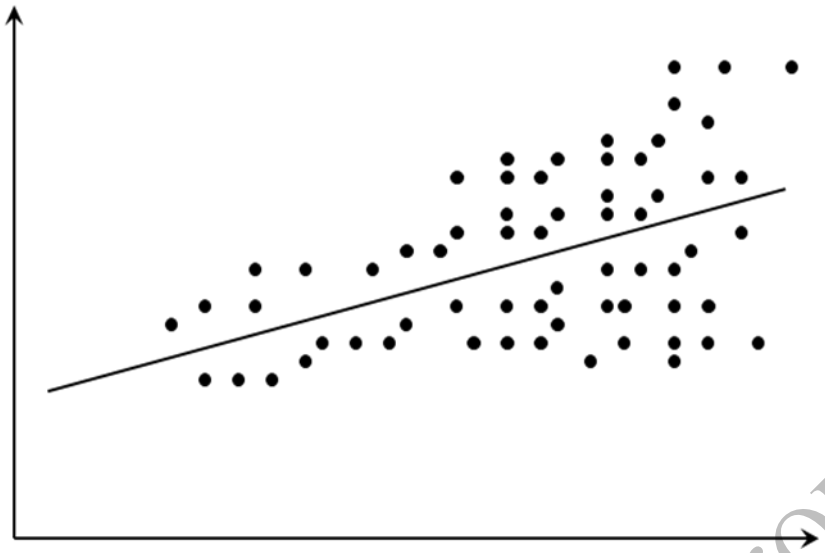


Рис. 5.2. Линейная модель с гетероскедастичностью

В настоящее время для оценки нарушения гомоскедастичности предложено большое число тестов. Чаще всего используются графический анализ отклонений, тест ранговой корреляции Спирмена и тест Голдфелда-Квандта.

Графический анализ отклонений заключается в визуальной оценке разброса точек корреляционного поля около линии регрессии: считается, что условие 2 выполняется, если точки наблюдений расположены внутри полосы постоянной ширины, окаймляющей линию регрессии (например, как на рисунке 5.1). Для множественной регрессии осуществляется графический анализ корреляционных полей объясняемой переменной y в зависимости от каждого из факторов $x_1, x_2, ..., x_m$.

Наиболее популярным тестом обнаружения гетероскедастичности является тест Голдфелда-Квандта. Тест применяется в том случае, если ошибки регрессии можно считать нормально распределенными случайными величинами. Кроме того, в основе его лежит предположение о пропорциональности дисперсий случайного члена значению выбранной объясняющей переменной. Тест проводится по следующей схеме.

1. На основе выборочных данных строится линейная модель множественной регрессии с m объясняющими переменными $x_1, x_2, ..., x_m$.

2. В модели множественной регрессии (например, на основе графического анализа) выбирается факторная переменная, от которой предположительно могут зависеть остатки. Значения этой переменной ранжируются, располагаются по возрастанию и делятся на три части объемами $k, (n-2k), k$ (обычно принимают $k \approx n/3$).

3. Для первой и третьей частей строятся две независимые модели регрессии.

4. По каждой из построенных моделей рассчитывают суммы квадратов остатков S_1 и S_3.

5. Осуществляется проверка основной гипотезы об отсутствии гетероскедастичности с помощью F-критерия Фишера. Наблюдаемое значение F-критерия рассчитывается следующим образом:

$$F_{\text{набл}} = \frac{S_3}{S_1}, \text{ если } S_3 > S_1,$$

или

$$F_{\text{набл}} = \frac{S_1}{S_3}, \text{ если } S_1 > S_3.$$

Если $F_{\text{набл}} > F_{\text{кр}}$, то в основной модели присутствует гетероскедастичность, зависящая от выбранной объясняющей переменной (число степеней свободы определяется значениями $k_1 = k - m - 1$ и $k_2 = k - m - 1$).

Если нет уверенности относительно выбора объясняющей переменной, вызывающей гетероскедастичность, то тест осуществляется для каждой из объясняющих переменных $x_1, x_2, ..., x_m$.

Наличие гетероскедастичности в остатках регрессии можно проверить и с помощью теста ранговой корреляции Спирмена. При выполнении теста предполагается, что абсолютные величины остатков и значения объясняющей переменной коррелированны. Эту корреляцию можно измерять с помощью коэффициента ранговой корреляции Спирмена:

$$r = 1 - \frac{6 \cdot \sum D_i^2}{n(n^2 - 1)},$$

где D_i – разность между рангом x_i и рангом модуля остатка $e_i = y_i - \tilde{y}_i$.

Тест проводится по следующей схеме.

1. Строится линейная модель регрессии.

2. Данные по x и модули остатков $|e_i|$ ранжируются по переменной x, определяются их ранги (ранг – это порядковый номер значений переменной в ранжированном ряду).

3. Осуществляется проверка основной гипотезы об отсутствии гетероскедастичности с помощью t-статистики с $n-2$ степенями свободы, где n – объем выборки. При этом наблюдаемое значение t-критерия определяется равенством $t_{\text{набл}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$. Если $|t_{\text{набл}}| > t_{\text{кр}}$, то нулевая

гипотеза об отсутствии гетероскедастичности отклоняется и имеет место гетероскедастичность в остатках регрессии, т.е. условие 2 не выполняется.

После установления в модели наличия гетероскедастичности возникает вопрос о том, в какой мере существенно она влияет на качество модели и следует ли вообще с гетероскедастичностью бороться. Ведь при гетероскедастичности оценки коэффициентов регрессии все равно остаются несмещенными и состоятельными, правда, не будут эффективными.

Если исследователь решил вступить в борьбу с гетероскедастичностью, то первый шаг на этом пути заключается в определении ее типа. Если гетероскедастичность вызвана ошибками спецификации, то для ее устранения необходимо включить в уравнение пропущенные существенные переменные и подобрать правильную функциональную форму. Если гетероскедастичность наблюдается в правильно специфицированных моделях (*чистая гетероскедастичность*), то можно воспользоваться *взвешенным методом наименьших квадратов* (ВМНК).

Данный метод применяется при известных для каждого наблюдения значениях дисперсиях σ_i^2. В этом случае можно устранить гетероскедастичность, разделив каждое наблюдаемое значение на соответствующее ему среднеквадратическое отклонение. Тем самым обеспечивается равномерный вклад остатков в общую сумму.

Таким образом, если при обычном МНК в случае парной линейной модели $y = a + bx + \varepsilon$ для нахождения ее параметров a и b минимизируется сумма $\sum_{i=1}^{n}(y_i - a - bx_i)^2$, то при ВМНК минимизируется сумма

$$\sum_{i=1}^{n}(\frac{y_i}{\sigma_i} - \frac{a}{\sigma_i} - b\frac{x_i}{\sigma_i})^2 = \sum_{i=1}^{n}\frac{1}{\sigma_i^2}(y_i - a - bx_i)^2 .$$

Применение ВМНК включает следующие этапы.

1. С помощью обычного МНК строится линейная регрессионная модель и доказывается наличие гетероскедастичности остатков.

2. Для каждого наблюдения устанавливаются фактические значения дисперсий σ_i^2 отклонений.

3. Значения каждой пары наблюдений (x_i, y_i) делятся на известную величину σ_i. Тем самым наблюдениям с наименьшими дисперсиями придаются наибольшие веса, а наблюдениям с наибольшими дисперсиями – наименьшие веса.

4. С помощью обычного МНК по преобразованным значениям $V_i = \frac{1}{\sigma_i}$, $X_i = \frac{x_i}{\sigma_i}$, $Y_i = \frac{y_i}{\sigma_i}$ (такое преобразование называется *масштабированием исходных данных*) оценивается двухфакторное уравнение регрессии $\tilde{Y} = aV + bX$ с нулевым свободным членом. Построенная модель гомоскедастична.

Описанный подход возможен и для уравнения множественной регрессии.

Главная проблема взвешенного метода наименьших квадратов состоит в необходимости знания среднеквадратических отклонений σ_i случайных ошибок регрессии. На практике дисперсии σ_i^2 неизвестны. В таком случае делаются реалистические предположения об их величине. В частности, принимаются предположения о том, что либо дисперсии σ_i^2 отклонений, либо сами среднеквадратичные отклонения σ_i пропорциональны значениям переменной x_i.

Следует иметь в виду, что коэффициенты a и b модели оценены по преобразованным данным, а потому изменяют свой первоначальный экономический смысл. В частности, если среднеквадратичные отклонения σ_i пропорциональны значениям x_i, то обычным МНК оценивается преобразованная модель $\frac{y}{x} = a\frac{1}{x} + b + \frac{\varepsilon}{x}$, в которой по сравнению с исходной моделью свободный член и угловой коэффициент как бы поменялись местами. Поэтому оценка свободного члена b преобразованной модели характеризует изменение показателя y при изменении фактора x на одну единицу.

5.4. Проблема автокорреляции

Одной из предпосылок регрессионного анализа является независимость случайного члена в любом наблюдении от его значений во всех других наблюдениях (условие 3). Неформально это означат, что «данные одного наблюдения не влияют на данные других наблюдений». Если данное условие не выполняется, то говорят, что случайный член подвержен *автокорреляции*.

Что касается моделей, построенных по пространственным данным, то для них автокорреляция, как правило, отсутствует. Пусть, например, обследуется выборка, полученная для различных предприятий, торговых организаций и т.д. Маловероятно, что при таком обследовании значение изучаемого показателя для какого-то объекта, окажется связанным со значением этого же показателя для другого объекта.

Автокорреляция обычно встречается в регрессионном анализе при использовании данных временных рядов. Поэтому под автокорреляцией

понимается в основном корреляционная зависимость между наблюдаемыми показателями во времени.

В экономических задачах чаще встречается положительная автокорреляция $[\mathrm{cov}(\varepsilon_{i-1}, \varepsilon_i) > 0]$, чем отрицательная $[\mathrm{cov}(\varepsilon_{i-1}, \varepsilon_i) < 0]$. Если неучтенная в уравнении переменная действует на зависимую переменную постоянно позитивно или негативно, то это приводит к положительной автокорреляции. Она более характерна для экономического анализа.

Графически положительная автокорреляция выражается в чередовании зон, в которых наблюдаемые значения оказываются выше объясненных регрессией, и зон, в которых наблюдаемые значения ниже (рисунок 5.3).

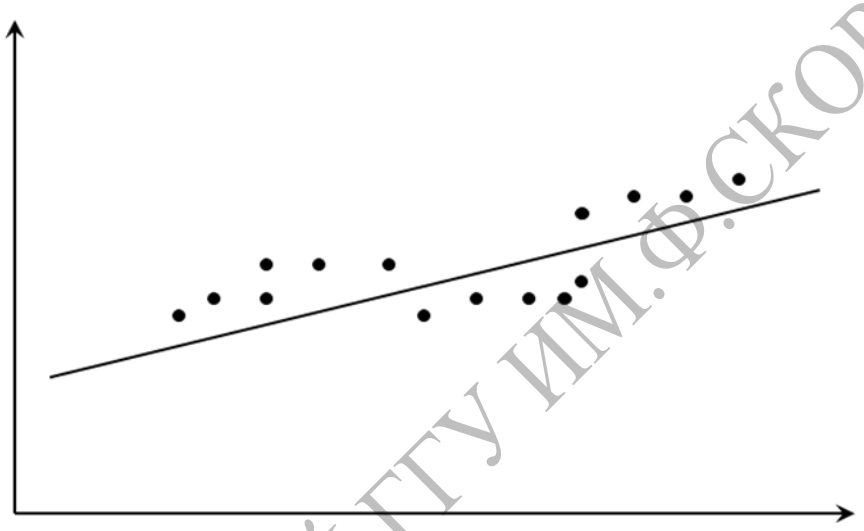


Рис. 5.3. Модель регрессии с положительной автокорреляцией

Отрицательная автокорреляция встречается в тех случаях, когда наблюдения действуют друг на друга по принципу маятника – завышенные значения в предыдущих наблюдениях приводят к занижению их в наблюдениях последующих. Графически это выражается в том, что результаты наблюдений слишком часто перескакивают через линию регрессии (рисунок 5.4).

Чаще всего причиной автокорреляции является либо неверная форма спецификации модели, либо наличие неучтенных факторов. Например, при

выборе линейной формы связи в ситуации, представленной на рисунке 5.3, когда имеет место экспоненциальная зависимость, возникает положительная автокорреляция.

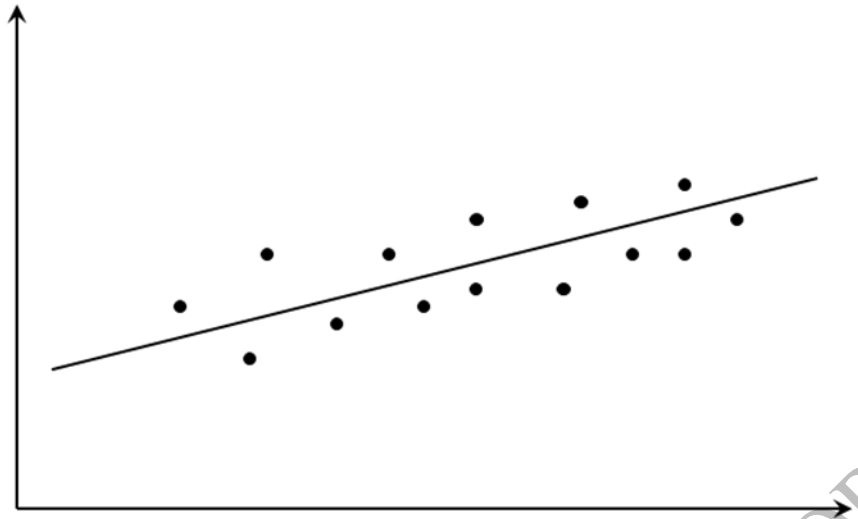


Рис. 5.4. Модель регрессии с отрицательной автокорреляцией

Автокорреляция может быть обусловлена также ошибками измерения результативного признака, цикличностью значений экономических показателей, запаздыванием изменения значений показателей по отношению к изменению экономических условий.

При наличии автокорреляции обычный метод наименьших квадратов дает несмещенные и состоятельные оценки параметров модели, которые однако неэффективны, т.е. их дисперсии не будут наименьшими. По сравнению с гетероскедастичностью автокорреляция приводит, наоборот, к завышению стандартных ошибок коэффициентов регрессии. На основе таких результатов может быть сделан ошибочный вывод о несущественном влиянии исследуемого фактора на зависимую переменную, в то время как на самом деле влияние фактора на нее значимо. Это может привести к ухудшению прогнозных качеств модели.

Игнорирование автокорреляции создает серьезные трудности для применения обыкновенного МНК. Поэтому важно владеть методами диагностики автокорреляции.

Существуют различные методы определения автокорреляции. Наиболее распространенными являются следующие два:

1. Построение графика остатков в зависимости от их порядковых номеров и визуальное определение наличия или отсутствия автокорреляции. Считается, что условие 3 выполняется, если точки наблюдений расположены возле линии регрессии хаотично без видимой закономерности (например, как на рисунке 5.5).

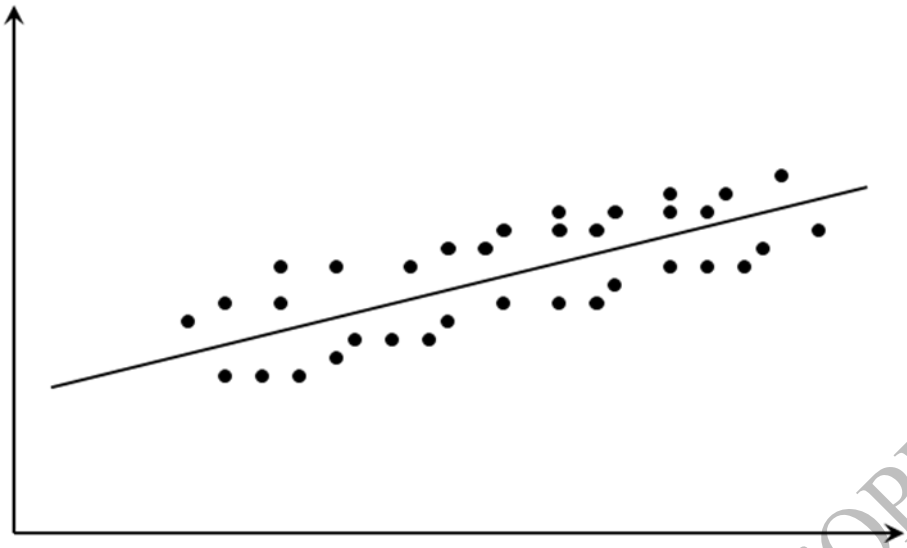


Рис. 5.5. Модель регрессии без автокорреляции

2. Использование *критерия Дарбина–Уотсона* и расчет величины

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

Статистическая независимость отклонений $e_i = y_i - \tilde{y}_i$ может быть количественно оценена по некоррелированности соседних величин отклонений с помощью коэффициента автокорреляции первого порядка r_1, вычисляемого по формуле $r_1 = \frac{\sum_i e_i e_{i-1}}{\sqrt{\sum e_i^2 \sum e_{i-1}^2}}$.

Считается, что $\bar{e}_i = \bar{e}_{i-1} = 0$, $\sum e_i^2 = \sum e_{i-1}^2$.

Поэтому критерий Дарбина–Уотсона основан на следующем соотношении, связывающим *DW*-статистику с коэффициентом автокорреляции первого порядка r_1:

$$DW = \frac{\sum (e_i - e_{i-1})^2}{\sum e_i^2} \approx 2 \cdot (1 - r_1).$$

Если $e_i = e_{i-1}$, то $DW = 0$, если $e_i = -e_{i-1}$, то $DW = 4$. Во всех случаях $0 \le DW \le 4$. Практически, если статистика Дарбина-Уотсона близка к двум, то автокорреляция отсутствует.

Общая схема критерия Дарбина–Уотсона следующая:

1. По построенному уравнению регрессии вычисляются значения отклонений $e_i = y_i - \tilde{y}_i$ для каждого наблюдения $i = 1,2,...,n$.

2. По формуле $DW = \dfrac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$ вычисляется величина *DW*.

3. По таблицам критических точек Дарбина–Уотсона определяются пороговые значения d_1 и d_2 в зависимости от уровня значимости, количества наблюдений и числа объясняющих переменных.

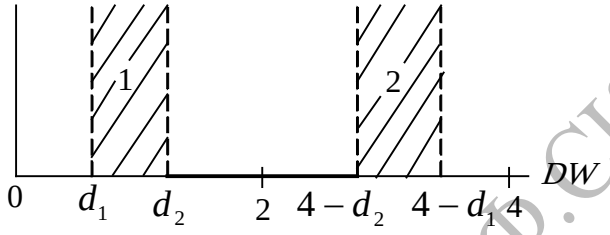


Рис. 5.6. Тест Дарбина–Уотсона

На рисунке 5.6 d_1 – граница для признания положительной автокорреляции остатков; d_2 – граница признания отсутствия автокорреляции остатков; 1 – зона неопределенности в случае предполагаемой положительной автокорреляции; 2 – зона неопределенности в случае предполагаемой отрицательной автокорреляции.

Далее наличие или отсутствие автокорреляции определяется тем, на какой участок отрезка $[0;4]$ попадает значение DW:

1) если $d_2 < DW < (4 - d_2)$, то признаются отсутствие автокорреляции;

2) если $0 < DW < d_1$, то имеется положительная автокорреляция;

3) если $(4 - d_1) < DW < 4$, то существует отрицательная автокорреляция;

4) если $d_1 < DW < d_2$ или $(4 - d_2) < DW < (4 - d_1)$, то имеет место зона неопределенности, когда нельзя ни отклонить, ни применить нулевую гипотезу об отсутствии автокорреляции (в этом случае требуется привлечение других критериев).

Недостатками критерия Дарбина–Уотсона является наличие области неопределенности критерия, а также то, что критические значения *DW*-статистики определены для объемов выборки не менее 15. Если фактическое значение *DW* критерия Дарбина-Уотсона попадает в зону неопределенности, то на практике предполагают существование автокорреляции остатков и отклоняют гипотезу об отсутствии автокорреляции.

Тест Дарбина-Уотсона построен в предположении, что объясняющие переменные некоррелированы со случайным членом. Поэтому тест

неприменим к моделям, включающим в качестве объясняющих переменных лаговые значения зависимой переменной.

Действия по устранению автокорреляции необходимо начинать с проверки спецификации модели, поскольку всегда существует вероятность того, что обнаруженная автокорреляция связана с пропущенной переменной или использованием неправильной функциональной формы уравнения.

Если ошибки спецификации устранены, то, возможно, это связано с внутренними свойствами ряда отклонений. Тогда для устранения автокорреляции можно воспользоваться *обобщенным методом наименьших квадратов* (ОМНК).

Для применения ОМНК необходимо специфицировать модель автокорреляции регрессионных остатков. Обычно в качестве такой модели используется авторегрессионный процесс первого порядка *AR*(1). Для простоты изложения ограничимся случаем парной регрессии.

Пусть исходная регрессионная модель $y = a + bx + \varepsilon$ содержит автокорреляцию случайных членов.

Допустим, что автокорреляция подчиняется *авторегрессионной схеме* первого порядка $\varepsilon_i = r_1\varepsilon_{i-1} + u_i$, где r_1 – коэффициент автокорреляции первого порядка, а u_i – случайный член, удовлетворяющий предпосылкам МНК.

Пусть r_1 известно.

Для значений i и $i-1$ справедливы равенства: $y_i = a + bx_i + \varepsilon_i$ и $y_{i-1} = a + bx_{i-1} + \varepsilon_{i-1}$.

Вычтем из первого равенства второе, умноженное на r_1:

$$y_i - r_1 y_{i-1} = (1 - r_1)a + b(x_i - r_1 x_{i-1}) + (\varepsilon_i - r_1\varepsilon_{i-1}).$$

Обозначим $y_i^* = y_i - r_1 y_{i-1}$, $x_i^* = x_i - r_1 x_{i-1}$, $a^* = (1 - r_1)a$.

Такое преобразование переменных и называется *авторегрессионым преобразованием первого порядка AR*(*1*) или *преобразованием Бокса-Дженкинса*.

Тогда преобразованное уравнение принимает вид

$$y_i^* = a^* + bx_i^* + u_i,$$

где $i \geq 2$. Это уравнение не содержит автокорреляцию и для оценки его параметров используется обычный МНК. Свободный член исходной модели имеет вид $a = \dfrac{a^*}{1 - r_1}$.

На практике величина r_1 неизвестна. Наиболее простой способ оценить r_1 – применить обычный МНК к регрессионному уравнению $\varepsilon_i = r_1\varepsilon_{i-1} + u_i$.

Коэффициент r_1 можно также приближенно оценить, используя статистику Дарбина-Уотсона: $DW \approx 2 \cdot (1 - r_1)$.

5.5. Проблема мультиколлинеарности

Мультиколлинеарность – это линейная зависимость между двумя или несколькими факторными переменными в уравнении множественной регрессии. Если такая зависимость является функциональной, то говорят о *полной мультиколлинеарности*. Если же она является корреляционной, то имеет место *частичная мультиколлинеарность*. Если полная мультиколлинеарность является скорее теоретической абстракцией (она проявляется, в частности, если фиктивную переменнную, имеющую *k* уровней качества, заменить на *k* дихотомических переменных), то частичная мультиколлинеарность весьма реальна и присутствует практически всегда. Речь может идти лишь о степени ее выраженности. Например, если в состав объясняющих переменных входят располагаемый доход и потребление, то обе эти переменные, конечно, будут сильно коррелированными.

Отсутствие мультиколлинеарности является одной из желательных предпосылок классической линейной множественной модели. Это связано со следующими соображениями:

1) В случае полной мультиколинеарности вообще невозможно построить оценки параметров линейной множественной регрессии с помощью МНК.

2) В случае частичной мультиколлинеарности оценки параметров регрессии могут быть ненадежными и, кроме того, затруднено определение изолированного вклада факторов в результативный показатель.

Главной причиной возникновения мультиколлинеарности является наличие в изучаемом объекте процессов, которые одновременно влияют на некоторые входные переменные, но не учтены в модели. Это может быть результатом некачественного исследования предметной области или сложности взаимосвязей параметров изучаемого объекта.

Подозрением наличия мультиколлинеарности служат:

– большое количество незначимых факторов в модели;

– большие стандартные ошибки параметров регрессии;

– неустойчивость оценок (небольшое изменение исходных данных приводит к их существенному изменению).

Один из подходов для определения наличия или отсутствия мультиколлинеарности заключается в анализе корреляционной матрицы

$$R = \begin{pmatrix} r_{yy} = 1 & r_{x_1 y} & \ldots & r_{x_m y} \\ r_{y x_1} & r_{x_1 x_1} = 1 & \ldots & r_{x_m x_1} \\ \ldots & \ldots & \ldots & \ldots \\ r_{y x_m} & r_{x_1 x_m} & \ldots & r_{x_m x_m} = 1 \end{pmatrix}$$

между объясняющими переменными $x_1, x_2, ..., x_m$ и выявлении пар факторов, имеющих высокие коэффициенты парной корреляции (обычно больше 0,7). Если такие факторы существуют, то говорят о явной коллинеарности между ними.

Однако парные коэффициенты корреляции, рассматриваемые индивидуально, не могут оценить совокупное взаимодействие нескольких факторов (а не только двух).

Поэтому для оценки наличия мультиколлинеарности в модели используется определитель матрицы парных коэффициентов корреляции между факторами (*определитель матрицы межфакторной корреляции*)

$$\Delta r = \begin{vmatrix} r_{x_1x_1}=1 & r_{x_2x_1} & \dots & r_{x_mx_1} \\ r_{x_1x_2} & r_{x_2x_2}=1 & \dots & r_{x_mx_2} \\ \dots & \dots & \dots & \dots \\ r_{x_1x_m} & r_{x_2x_m} & \dots & r_{x_mx_m}=1 \end{vmatrix}.$$

Чем ближе определитель матрицы межфакторной корреляции к 0, тем сильнее мультиколлинеарность, и наоборот, чем ближе определитель к 1, тем меньше мультиколлинеарность.

Статистическая значимость мультиколлинеарности факторов определяется проверкой нулевой гипотезы $H_0 : \Delta r = 1$ при альтернативной гипотезе $H_0 : \Delta r \neq 1$. Для проверки нулевой гипотезы используется распределение Пирсона χ^2 с $\frac{1}{2}m(m-1)$ степенями свободы. Наблюдаемое значение статистики находится по формуле $\chi^2_{\text{набл}} = n - 1 - \frac{1}{6}(2m+5)\lg \Delta r$, где n – число наблюдений, m – число факторов. Для заданного уровня значимости α по таблице критических точек распределения Пирсона определяется критическое значение $\chi^2_{\text{кр}}$. Если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$, то гипотеза H_0 отклоняется и считается, что в модели присутствует мультиколлинеарность факторов.

Выделить факторы, влияющие на мультиколлинеарность, позволяет также анализ коэффициентов множественной детерминации, вычисленных при условии, что каждый из факторов рассматривается в качестве зависимой переменной от других факторов: $R^2_{x_1x_2x_3\dots x_m}$, $R^2_{x_2x_1x_3\dots x_m}$, …, $R^2_{x_mx_1x_3\dots x_{m-1}}$. Чем ближе они к 1, тем сильнее мультиколлинеарность факторов. Значит, в уравнении следует оставлять факторы с минимальной величиной коэффициента множественной детерминации.

Что касается полной мультиколлинеарности, то с ней следует вести самую решительную борьбу: сразу же удалять из регрессионного уравнения

переменные, которые являются линейными комбинациями другихпеременных.

Частичная мультиколлинеарность не является таким уж серьезным злом, чтобы ее выявлять и устранять. Все зависит от целей исследования. Если основная задача моделирования – только прогнозирование значений зависимой переменной, то при достаточно большом коэффициенте детерминации ($R^2 \geq 0{,}9$) присутствие мультиколлинеарности не сказывается на прогнозных качествах модели. Если же целью моделирования является и определение вклада каждого фактора в изменение зависимой переменной, то наличие мультиколлинеарности является серьезной проблемой.

Простейшим методом устранения мультиколлинеарности является исключение из модели одной или ряда коррелированных переменных.

Поскольку мультиколлинеарность напрямую зависит от выборки, то, возможно, при другой выборке мультиколлинеарности не будет вообще либо она не будет настолько серьезной. Поэтому для уменьшения мультиколлинеарности в ряде случаев достаточно увеличить объем выборки.

Иногда проблема мультиколлинеарности может быть решена путем изменения спецификации модели: либо изменяется форма модели, либо добавляются факторы, не учтенные в первоначальной модели, но существенно влияющие на зависимую переменную.

В ряде случаев минимизировать либо совсем устранить мультиколлинеарность можно с помощью преобразования факторных переменных. При этом наиболее распространены следующие преобразования:

1. Линейная комбинация мультиколлинеарных переменных (например, $k_i \cdot x_i + k_j x_j$).

2. Замена мультиколлинеарной переменной x_i ее приращением $\Delta x_i = x_i - x_{i-1}$.

3. Деление одной коллинеарной переменной на другую.

5.6. Проверка предположения о нормальности распределения

Одно из требований классической регрессионной модели заключается в нормальности распределения случайной величины ε. Выполнимость или невыполнимость его не оказывает влияния на качество оценок параметров регрессии, но является достаточно важным. Оно позволяет использовать стандартные процедуры проверки статистических гипотез и построения доверительных интервалов.

Проверка на нормальность распределения без существенных вычислительных затрат может быть проведена с помощью стандартного теста, опирающегося на критерий Пирсона.

Второй способ тестирования имеет визуальный характер. Он ориентируется на характерные особенности графика плотности нормально распределенных случайных величин. Для этого строится гистограмма

остатков и соединяются середины верхних сторон прямоугольников гистограммы. Если ломаная линия приближенно напоминает кривую плотности нормального распределения, то можно сделать грубый вывод о том, что остатки распределены по нормальному закону.

Применение визуального метода рекомендуется в случаях, когда число наблюдений сравнительно невелико (от 10 до 25) и невозможно применить критерий Пирсона хи-квадрат, где предполагается, что $n > 70$.

Достаточно простым методом диагностики нормальности распределения остатков является тест Жарка-Бера. Идея метода состоит в том, что для совокупности остатков оценивается «скошенность» (асимметрия) и «вытянутость» (эксцесс) фактического распределения ряда остатков и сравнивается с нормальным. При этом за оценку «скошенности» распределения отвечает *коэффициент асимметрии* $K_{ас}$, а за оценку «вытянутость» распределения – *коэффициент эксцесса* $K_{экс}$.

Алгоритм теста Жарка-Бера следующий.

1) Выдвигается гипотеза H_0 о нормальном распределении остатков выборки.

2) Вычисляется наблюдаемое значение критерия по формуле

$$JB_{набл} = \frac{n}{6}((K_{ас})^2 + \frac{(K_{экс} - 3)^2}{4}).$$

3) По таблице критических точек распределения Пирсона (число степеней свободы 2) определяется критическое значение (на уровне значимости 0,05 оно равно 5,991).

4) Если $JB_{набл} > 5{,}991$, то гипотеза H_0 о нормальном распределении остатков отклоняется, т.е. распределение остатков не является нормальным. Если $JB_{набл} < 5{,}991$, то гипотеза H_0 о нормальном распределении остатков принимается.

Нахождение коэффициента асимметрии $K_{ас}$ и коэффициента эксцесса $K_{экс}$ можно осуществить с помощью вкладки «*Описательная статистика*» пакета «*Анализ данных*» Excel.

5.7. Обзор некоторых вопросов и проблем модельного анализа

С практической точки зрения рекомендуется следующая последовательность проверки модельных предположений.

1. Проверяется предположение о равенстве нулю математического ожидания случайной переменной, что гарантирует получение несмещенных оценок коэффициентов регрессии и прогноза.

2. Проверяется предположение о нормальности распределения остатков, что позволяет оценивать значимость модели, проверять статистические гипотезы и строить доверительные интервалы коэффициентов регрессии и

прогноза.

3. Предположение об отсутствии автокорреляции напрямую связано со всеми формулами, которые позволяют вычислять оценки параметров регрессии. Поэтому третьим шагом с помощью критерия Дарбина-Уотсона проверяется наличие в модели автокорреляции.

4. С помощью теста Голдфелда-Квандта или теста Спирмена проверяется однородность результатов наблюдений с целью обеспечения эффективных оценок регрессии.

Обычно в предпосылках регрессионного анализа считается, что факторы являются неслучайными величинами и что они не коррелируют со случайной переменной. При наличии такой корреляции МНК-оценки могут быть смещенными и несостоятельными. Для получения состоятельных оценок можно воспользоваться методом инструментальных переменных, описанным в [15].

Тесты Спирмена и Голдфелда-Квандта не дают количественной оценки зависимости дисперсии ошибок регрессии от соответствующих значений факторов, включенных в регрессию. Они позволяют лишь определить наличие или отсутствие гетероскедастичности. Поэтому если гетероскедастичность остатков установлена, можно количественно оценить зависимость дисперсии ошибок регрессии от значений факторов. С этой целью могут быть использованы тесты Уайта, Парка, Глейзера и другие (см., например, [2,3]).

Одним из возможных методов диагностики автокорреляции является метод рядов. Он связан с анализом знаков остатков регрессии $e_i = y_i - \tilde{y}_i$, $i = 1,2,...,n$. Описание метода можно найти в [3].

Примеры решения типовых заданий

Пример 5.1. По статистическим данным таблицы 5.1 для двенадцати транспортных компаний исследуется зависимость годового дохода (переменная y, млн. руб.) от среднегодового количества грузовых автомобилей (переменная x).

1. Построить парную линейную модель зависимости y от x.

2. Проверить наличие в модели гетероскедастичности с помощью графического анализа и методом Голдфелда–Квандта.

3. При обнаружении гетероскедастичности построить взвешенную модель регрессии.

Таблица 5.1. Статистические данные примера 5.1

	1	2	3	4	5	6	7	8	9	10	11	12
x	15	18	22	25	27	31	34	37	40	45	48	48
y	235	250	247	260	287	262	307	280	357	410	389	311

Решение:

1. По исходным данным строим линейную модель парной регрессии. Параметры ее оцениваем обычным методом наименьших квадратов. Коэффициенты уравнения регрессии определяем с помощью табличного процессора MS Excel: $a = 160{,}6$; $b = 4{,}277$. Таким образом, уравнение регрессии имеет вид $\tilde{y} = 160{,}6 + 4{,}277x$.

Уравнение регрессии статистически значимо на уровне α=0,05: наблюдаемое значение *F*-статистики равно 25,15; критическое значение *F*-критерия Фишера – 4,96; коэффициент детерминации равен 0,716.

Значение коэффициента *b* уравнения регрессии показывает, что увеличение количества автомобилей на одну единицу приводит к росту годового дохода в среднем на 4,277 млн. руб.

Визуальный анализ графика зависимости годового дохода от количества автомобилей дает основание предполагать наличие в модели гетероскедастичности. Из рисунка 5.5 видно, что отклонение от линии регрессии наблюдений, соответствующих крупным предприятиям, больше, чем для малых предприятий.

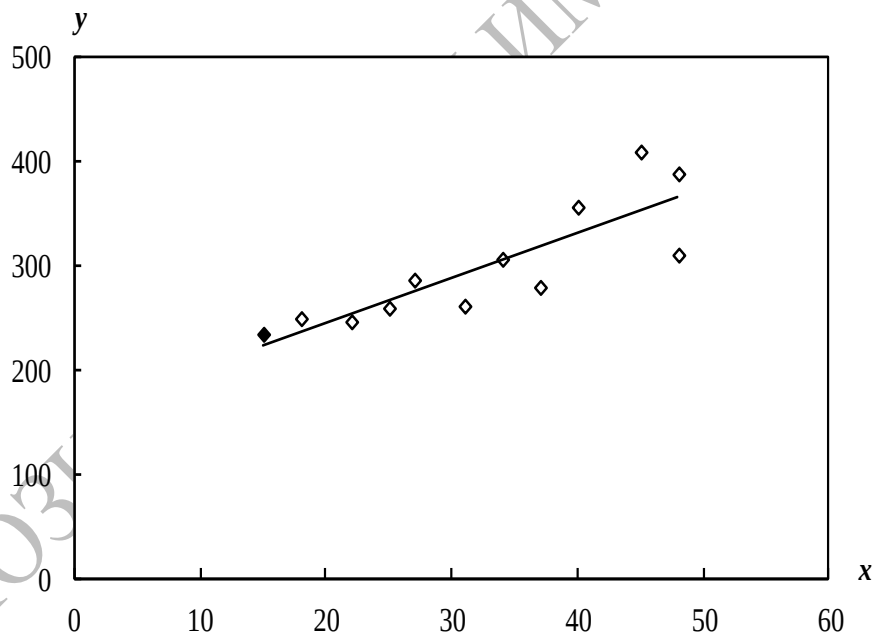


Рис. 5.5. График разброса точек наблюдений относительно линии регрессии

2. Строим график остатков и проводим его визуальный анализ. В таблице 5.2 приведены наблюдаемые y_i и теоретические $\tilde{y}_i$ значения переменной y, а также значения остатков $e_i = y_i - \tilde{y}_i$.

Таблица 5.2. Наблюдаемые и теоретические значения переменной y

	1	2	3	4	5	6	7	8	9	10	11	12
x_i	15	18	22	25	27	31	34	37	40	45	48	48
y_i	235	250	247	260	287	262	307	280	357	410	389	311
$\tilde{y}_i$	225	238	255	268	276	293	306	319	332	353	366	366
e_i	10	12	-8	-8	11	-31	1	-39	25	57	23	-55

График остатков представлен на рисунке 5.6.

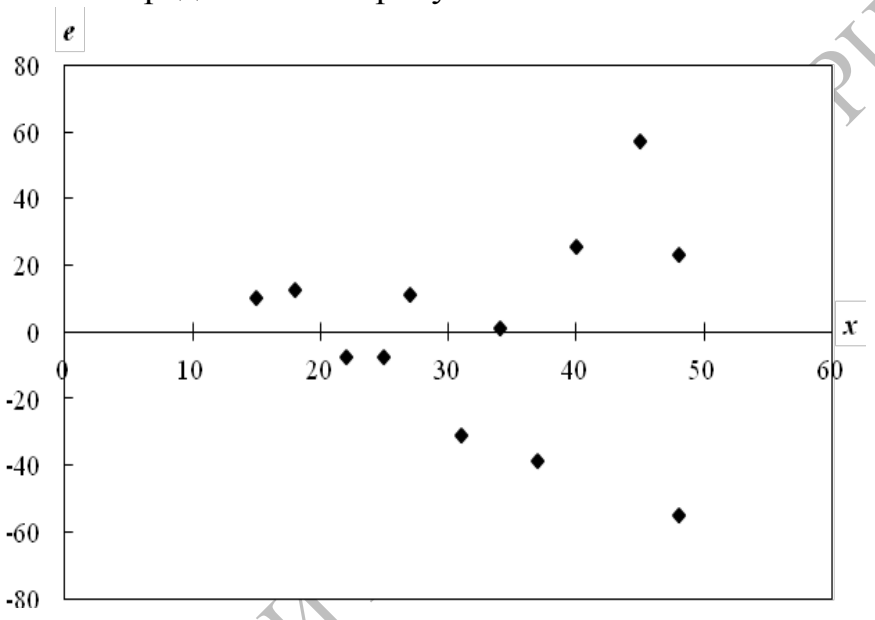


Рис. 5.6. График остатков

Графический анализ остатков показывает, что их разброс растет по мере увеличения фактора x, что может свидетельствовать о наличии гетероскедастичности. Проверим это предположение методом Голдфелда-Квандта. Будет считать, что остатки распределены по нормальному закону и их среднее квадратическое отклонение пропорционально значению фактора x. Все остатки уже упорядочены по x. Выбираем $k = n/3 = 12/3 = 4$ первых и последних остатков. Так как число наблюдений невелико, то оценим суммы квадратов остатков для крайних групп наблюдений по общей модели:

$$S_1 = e_1^2 + e_2^2 + e_3^2 + e_4^2 = 10^2 + 12^2 + (-8)^2 + (-8)^2 = 372;$$
$$S_3 = e_9^2 + e_{10}^2 + e_{11}^2 + e_{12}^2 = 25^2 + 57^2 + 23^2 + (-55)^2 = 7432.$$

Так как $S_3 > S_1$, то наблюдаемое значение F-статистики рассчитываем по формуле $F_{\text{набл}} = S_3 / S_1 = 7432/372 = 19{,}98$.

Критическое значение *F*-критерия Фишера для уровня значимости α=0,05 и степеней свободы $k_1 = k_2 = k - m - 1 = 4 - 1 - 1 = 2$ (где $m = 1$ – число факторов в модели) составляет 19.

Так как $F_{\text{набл}} = 19{,}98 > F_{\text{кр}} = 19$, то статистическая гипотеза об одинаковой дисперсии остатков отклоняется на уровне значимости α=0,05. Факт наличия гетероскедастичности в модели считается установленным.

3. Применим взвешенный МНК к исходной модели в предположении, что среднее квадратическое отклонение остатков пропорционально значению фактора x. Для этого масштабируем исходные данные по x. Результаты масштабирования сведем в таблицу 5.3.

Таблица 5.3. Результаты масштабирования

	1	2	3	4	5	6	7	8	9	10	11	12
$\frac{1}{x_i}$	0,0667	0,0556	0,0455	0,04	0,037	0,0323	0,0294	0,0270	0,0250	0,0222	0,0208	0,0208
$\frac{y_i}{x_i}$	15,67	13,89	11,23	10,4	10,63	8,45	9,03	7,57	8,93	9,11	8,10	6,48

Исходную модель преобразуем в модель $\frac{y_i}{x_i} = b + a\frac{1}{x_i} + \frac{\varepsilon_i}{x_i}$. Оцениваем параметры преобразованной модели b и a обычным методом наименьших квадратов. С помощь MS Excel определяем коэффициенты уравнения регрессии $\tilde{Y} = b + aX$, где $Y = \frac{y}{x}$, $X = \frac{1}{x}$: $b = 3{,}863$; $a = 173{,}2$. После этого уравнение принимает вид $\tilde{Y} = 3{,}863 + 173{,}2X$. Общее качество уравнения высокое: $R^2 = 0{,}914$. Уравнение регрессии статистически значимо на уровне α=0,05: наблюдаемое значение *F*-статистики равно 106 при критическом значении *F*-критерия Фишера 4,96.

Тест Голдфельда–Квандта, примененный к преобразованной модели $\tilde{Y} = 3{,}863 + 173{,}2X$, уже не выявляет гетероскедастичности ее остатков.

Используя преобразованное уравнение регрессии, делаем вывод, что увеличение количества автомобилей на одну единицу приводит к росту годового дохода в среднем на 3,863 млн. руб.

Пример 5.2. По статистическим данным за 12 месяцев, приведенным в таблице 5.4, исследуется зависимость цены акции предприятия (переменная y, дол.) от индекса фондового рынка (переменная x, пунктов).

1. Построить линейную парную регрессионную модель зависимости y от

фактора x.

2. Проверить наличие автокорреляции остатков модели с помощью графического анализа и методом Дарбина-Уотсона.

3. При обнаружении автокорреляции построить обобщенную модель регрессии.

Таблица 5.4. Статистические данные примера 5.2

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
x	244	222	201	186	215	248	256	255	217	224	263	292
y	152	154	149	136	139	148	152	156	152	156	169	176

Решение:

1. По статистическим данным строим модель парной регрессии, параметры которой оцениваем обычным методом наименьших квадратов. С помощью табличного процессора MS Excel определяем коэффициенты уравнения регрессии: $a = 81{,}8$; $b = 0{,}304$. Уравнение регрессии имеет вид $\tilde{y} = 81{,}8 + 0304x$.

Уравнение регрессии статистически значимо на уровне α=0,05: коэффициент детерминации имеет значение $R^2 = 0{,}671$; $F_{\text{набл}} = 20{,}41$; $F_{\text{кр}} = 4{,}96$.

Значение коэффициента $b = 0{,}304$ уравнения показывает, что при росте индекса рынка на 1 пункт цена акции возрастает в среднем на 0,304 дол.

График зависимости y от x представлен на рисунке 5.7.

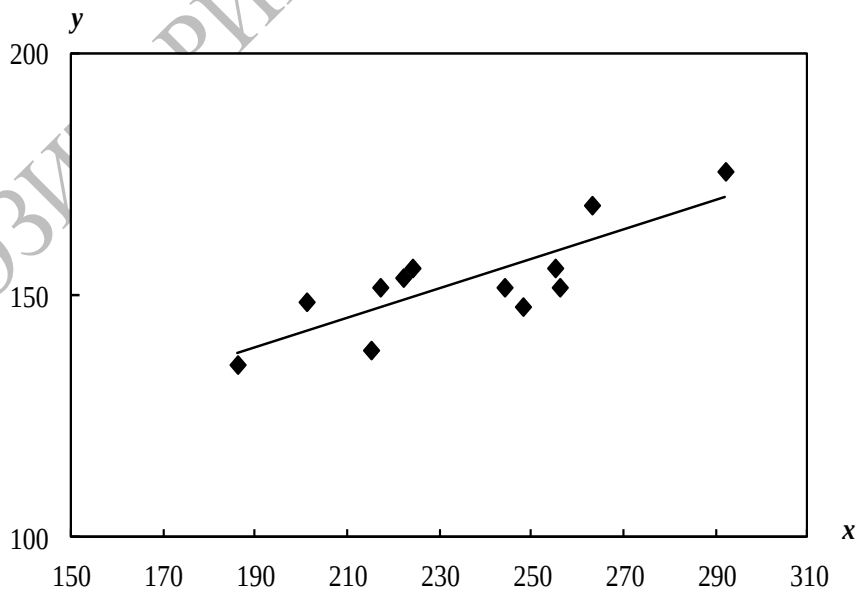


Рис. 5.7. График разброса точек наблюдений

относительно линии регрессии

2. Построим график временного ряда остатков регрессии (рисунок 5.8) и проведем его визуальный анализ. В таблице 5.5 приведены наблюдаемые y_i и теоретические $\tilde{y}_i$ значения переменной y, а также значения остатков $e_i = y_i - \tilde{y}_i$.

Таблица 5.5. Наблюдаемые и теоретические значения переменной y

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
x_i	244	222	201	186	215	248	256	255	217	224	263	292
y_i	152	154	149	136	139	148	152	156	152	156	169	176
$\tilde{y}_i$	156	149	143	138	147	157	160	159	148	150	162	170
e_i	-3,9	4,8	6,2	-2,3	-8,1	-9,1	-7,6	-3,2	4,3	6,2	7,3	5,5

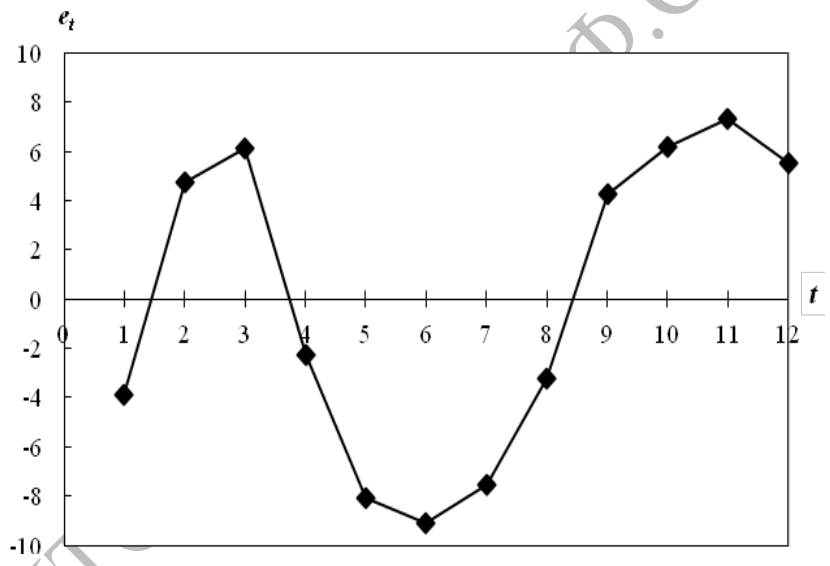


Рис. 5.8. График остатков

Визуальный анализ графика 5.8 указывает на положительную автокорреляцию остатков: видно, что на графике имеются чередующиеся зоны положительных и отрицательных остатков регрессии. Проверим это предположение методом Дарбина-Уотсона. Определяем DW-статистику по формуле

$$DW = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = \frac{(4{,}8-(-3{,}9))^2 + (6{,}2-4{,}8)^2 + \ldots + (5{,}5-7{,}3)^2}{(-3{,}9)^2 + 4{,}8^2 + \ldots + 5{,}5^2} = 0{,}61.$$

Критические значения DW-критерия для числа наблюдений $n=12$ и уровня значимости α=0,05 составляют $d_1=0{,}97$ и $d_2=1{,}33$. Так как $0<DW<d_1$, то это свидетельствует о наличии положительной автокорреляции остатков.

3. Применим обобщенный метод наименьших квадратов для оценки параметров исходной модели, для чего преобразуем исходные данные по формулам $y_i^*=y_i-r_1y_{i-1}$, $x_i^*=x_i-r_1x_{i-1}$, $a^*=(1-r_1)a$.

Преобразованные данные представим в виде таблицы 5.6.

Таблица 5.6. Преобразованные данные

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
x_i^*	-	65,7	58,8	57,3	95,9	110,3	97,2	91,0	53,7	85,0	119,5	123,6
y_i^*	-	56,7	50,4	40,6	51,9	59,0	57,2	58,7	52,1	58,7	69,1	67,8

Обычным методом наименьших квадратов определяем коэффициенты преобразованного уравнения регрессии $\tilde{y}_i^*=a^*+bx_i^*$: $a^*=34{,}2$; $b=0{,}257$. Свободный член исходного уравнения теперь находим из равенства

$$a=\frac{a^*}{1-r_1}=\frac{34{,}2}{1-0{,}640}=95{,}0.$$

Окончательно обобщенное уравнение линейной регрессии принимает вид $\tilde{y}=95{,}0+0{,}257x$.

Данное уравнение статистически значимо на уровне α=0,05: коэффициент детерминации имеет значение $R^2=0{,}666$; $F_{\text{набл}}=17{,}92$; $F_{\text{кр}}=5{,}12$.

Таким образом, при росте индекса рынка на 1 пункт цена акции возрастает в среднем на 0,257 дол.

Пример 5.3. В таблице 5.7 приведены данные о заработной плате y (доллары), возрасте x_1 (годы), стаже работы по специальности x_2 (годы), выработке x_3 (единицы за смену) для 20 рабочих предприятия.

1. Проверить наличие мультиколлинеарности между факторами.

2. Построить линейную регрессионную модель с полным набором факторов.

3. Определить факторы, ответственные за мультиколлинеарность.

4. Построить линейную регрессионную модель без факторов, ответственных за мультиколлинеарность.

5. Сравнить качество построенных моделей.

Таблица 5.7. Статистические данные примера 5.3

y	x_1	x_2	x_3
300	29	6	17
400	40	19	25
300	36	10	15
320	32	10	17
200	23	3	15
350	45	20	18
350	38	17	17
400	40	23	25
380	50	31	19
400	47	25	23
250	28	7	15
350	30	7	18
200	25	6	16
400	48	20	23
220	30	5	18
320	40	15	18
390	40	20	25
360	38	20	23
260	29	10	18
250	25	5	17

Решение:

1. С помощью табличного процессора MS Excel построим матрицу межфакторной корреляции:

$$r = \begin{pmatrix} 1 & 0{,}935263 & 0{,}615448 \\ 0{,}935263 & 1 & 0{,}69661 \\ 0{,}615448 & 0{,}69661 & 1 \end{pmatrix}.$$

Вычислим определитель этой матрицы: $\Delta r = 0{,}063187$. Близость к нулю определителя матрицы межфакторной корреляции указывает на наличие сильной мультиколлинеарности факторов.

Статистическую значимость мультиколлинеарности определим проверкой нулевой гипотезы $H_0 : \Delta r = 1$. Для проверки нулевой гипотезы используем распределение Пирсона χ^2 с $\frac{1}{2}m(m-1) = 3$ степенями свободы. Наблюдаемое значение статистики находится по формуле

$$\chi^2_{\text{набл}} = n - 1 - \frac{1}{6}(2m+5)\lg\Delta r = 20 - 1 - \frac{1}{6}(2\cdot 3+5)\lg 0{,}063187 = 21{,}2,$$

где $n = 20$ – число наблюдений, $m = 3$ – число факторов. Для заданного уровня значимости α по таблице критических точек распределения Пирсона определяется критическое значение $\chi^2_{\text{кр}} = 7{,}8$. Так как $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}$, то гипотеза H_0 отклоняется и считается, что присутствует мультиколлинеарность факторов.

Построим матрицу парных коэффициентов корреляции:

$$\begin{pmatrix} 1 & 0{,}853056 & 0{,}849877 & 0{,}778766 \\ 0{,}853056 & 1 & 0{,}935263 & 0{,}615448 \\ 0{,}849877 & 0{,}935263 & 1 & 0{,}69661 \\ 0{,}778766 & 0{,}615448 & 0{,}69661 & 1 \end{pmatrix}.$$

Из матрицы видно, что между факторами x_1 (возраст) и x_2 (стаж по специальности) имеет место сильная линейная зависимость $r_{x_1x_2} = 0{,}935$. Это также указывает на наличие мультиколлинеарности факторов.

2. С помощью табличного процессора MS Excel строим линейную регрессионную модель зависимости заработной платы от всех трех факторов. Она имеет вид $y = -10{,}9 + 4{,}92x_1 + 0{,}22x_2 + 7{,}98x_3 + \varepsilon$. Проверка значимости коэффициентов регрессии при факторах в уравнении показывает, что статистически значимыми будут коэффициенты регрессии только при x_1 и x_3. Поэтому полученное уравнение регрессии неприемлемо.

3. Из модели следует исключить фактор x_2, так как он теснее связан с третьим фактором x_3 нежели фактор x_1 (это следует из анализа матрицы парных коэффициентов корреляции). Кроме того, фактор x_1 теснее связан с переменной y. Отдавая предпочтение фактору x_1, мы учитываем также, что в построенной трехфакторной модели коэффициент при x_2 не является статистически значимым.

4. Построим теперь уравнение линейной регрессии, исключив фактор x_2. Это уравнение имеет вид $\tilde{y} = -16{,}04 + 5{,}1x_1 + 8{,}08x_3$.

5. В уравнении $\tilde{y} = -16{,}04 + 5{,}1x_1 + 8{,}08x_3$ коэффициенты регрессии при переменных x_1 и x_3 значимы. Так как $R^2 = 0{,}83$, то общее качество уравнения регрессии высокое. Новое уравнение значимо в целом, что подтверждается критерием Фишера.

Реализация с помощью ППП *Excel*

Расчет и анализ показателей множественной регрессии, а также проверка необходимых статистических гипотез могут быть осуществлены с помощью «Пакета анализа» табличного процессора *Excel*.

Методику применения ППП *Excel* проиллюстрируем на примере следующей задачи (при этом будем опираться на статистические данные, находящиеся в таблице 5.9, и исходить из того, что объем выборки n равен 20).

Задача. Для объяснения заработной платы y в зависимости от возраста x_1 и стажа по данной специальности x_2 построить и исследовать линейную регрессионную модель $y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon$. Применить построенную регрессионную модель для прогноза заработной платы при $x_1^* = 55$, $x_2^* = 25$.

Требуется:

1) ввести данные;

2) провести регрессионный анализ;

3) провести анализ общего качества уравнения регрессии;

4) указать стандартную ошибку регрессии;

5) указать стандартные ошибки коэффициентов регрессии;

6) проанализировать статистическую значимость коэффициентов при уровне значимости α = 0,05, при необходимости получить новое уравнение регрессии со значимыми коэффициентами;

7) найти точечные и интервальные оценки для коэффициентов регрессии;

8) дать точечный прогноз заработной платы по заданным значениям возраста и стажа;

9) рассчитать коэффициенты эластичности и сделать вывод о влиянии факторов на заработную плату.

Выяснить, выполняются ли условия теоремы Гаусса-Маркова для случайного члена:

1) проверить предположение о равенстве нулю математического ожидания случайного члена;

2) проверить предположение о наличии в модели гомоскедастичности остатков с помощью теста ранговой корреляции Спирмена, при необходимости построить взвешенную модель регрессии;

3) проверить предположение об отсутствии в модели автокорреляции остатков с помощью DW-статистики Дарбина-Уотсона, при необходимости построить обобщенную модель регрессии;

4) проверить гипотезу о нормальном распределении остатков;

5) сделать вывод о присутствии или отсутствии мультиколлинеарности факторов, при необходимости исключить один из факторов и построить парную линейную модель.

Результаты вычислений и анализа оформить в виде отчета (*форма отчета прилагается ниже*).

Порядок выполнения:

1) В ячейку A1 введите название Зарплата, в ячейку B1 – название Возраст, в ячейку C1 – название Стаж. В ячейки A2, A3, …, A21 введите данные первого столбца выбранного варианта задания, в ячейки B2, B3, …, B21 – данные второго столбца, в ячейки C2, C3, …, C21 – данные третьего столбца.

Введите новое название листа «Исходные данные», щелкнув правой кнопкой мыши на названии листа «Лист 1» и выбрав опцию *переименование* или двойным щелчком левой кнопкой мыши в поле «Лист 1».

2) Выберите в опциях меню *Сервис → Анализ данных → Регрессия → ОК*. Установите значения параметров в диалоговом окне следующим образом:

- *Входной интервал Y* — введите ссылки на ячейки A1:A21;
- *Входной интервал X* — введите ссылки на ячейки B1:C21;
- *Метки* — установите флажок;
- *Уровень надежности* — установите флажок;
- *Константа ноль* — оставьте пустым;
- *Параметры вывода* — установите флажок на *Новый рабочий лист* и в соответствующее поле введите его название «Регрессия»;
- *Остатки* — установите флажок;
- *Стандартизированные остатки* — оставьте пустым;
- *График остатков* — установите флажок;
- *График подбора* — установите флажок;
- *График нормальной вероятности* — оставьте пустым.

Нажмите *ОК*.

Расположите диаграммы рядом (на поле диаграммы нажмите левую кнопку мышки, затем поместите курсор на белое поле и при нажатой левой кнопке передвигайте диаграмму вниз) и растяните их (на поле диаграммы нажмите левую кнопку мышки, нижнюю линию границы диаграммы при нажатой левой клавише протяните вниз).

3) Значение коэффициента детерминации R^2 находится на листе «Регрессия» в ячейке B5. Наблюдаемое значение *F*-критерия Фишера $F_{\text{набл}}$ находится на листе «Регрессия» в ячейке E12.

Вычислите критическое значение $F_{кр}$ в свободной ячейке E15 следующим образом:

– нажмите на $\boldsymbol{f_x}$ (вставка функций);

– в поле *Категория* окна *Мастер функций* выберите *статистические*, из предложенных ниже функций выделите FРАСПОБР и нажмите *ОК*.

Откроется окно *Аргументы функций*. Заполните поля так:

- *Вероятность* — наберите значение 0,05;
- *Степени свободы 1* — установите курсор в поле и выделите ячейку B12 столбца *df* таблицы «Дисперсионный анализ»;

• *Степени свободы 2* — установите курсор в поле и выделите ячейку B13 столбца *df* таблицы «Дисперсионный анализ».

Нажмите *ОК*.

4) Значение стандартной ошибки s регрессии находится на листе «Регрессия» в ячейке B7.

5) Значения стандартных ошибок коэффициентов регрессии находятся на листе «Регрессия» в ячейках C17, C18 и C19 соответственно.

6) Наблюдаемые значения t-статистики $t_{\text{набл}}$ коэффициентов регрессии a_0, a_1 и a_2 находятся на листе «Регрессия» в ячейках D17, D18 и D19 соответственно.

Вычислите критическое значение $t_{кр}$ в свободной ячейке D20 следующим образом:

– нажмите на $\boldsymbol{f_x}$ (вставка функций);

– в поле «Категория» окна *Мастер функций* выберите *статистические*, из предложенных ниже функций выделите СТЬЮДРАСПОБР и нажмите *ОК*. Откроется окно *Аргументы функций*. Заполните поля:

• *Вероятность* – наберите значение 0,05;

• *Степени свободы* – введите 20-2-1, где 20 – число наблюдений, 2 – число факторов в уравнении регрессии, 1 – число свободных членов (a_0) в уравнении регрессии.

Нажмите *ОК*.

7) Точечные оценки коэффициентов регрессии a_0, a_1 и a_2 находятся на листе «Регрессия» в ячейках B17, B18 и B19 соответственно. Нижняя и верхняя границы доверительного интервала вычислены на листе «Регрессия» в ячейках F17 и G17 для коэффициента a_0, в ячейках F18 и G18 для коэффициента a_1 и в ячейках F19 и G19 для коэффициента a_2.

8) На листе «Регрессия» в ячейке D22 введите формулу

= B17+B18*55+B19*25.

9) На листе «Исходные данные» введите формулы:

в ячейку A23 =СРЗНАЧ(A2:A21);

в ячейку B23 =СРЗНАЧ(B2:B21);

в ячейку C23 =СРЗНАЧ(C2:C21).

С листа «Регрессия» скопируйте ячейку B18 в ячейку B24 листа «Исходные данные», ячейку B19 – в ячейку C24 (значения коэффициентов регрессии).

В ячейку B25 введите формулу =B24*B23/A23 (для вычисления коэффициента эластичности переменной *Возраст*).

В ячейку C25 введите формулу =C24*C23/A23 (для вычисления коэффициента эластичности переменной *Стаж*).

Проверка модельных предположений:

1) Для проверки условия о равенстве нулю математического ожидания случайной величины на листе «Регрессия» выберите в опциях меню *Сервис → Анализ данных → Описательная статистика → ОК*. Установите значения параметров в диалоговом окне следующим образом:

- *Входной интервал* – введите ссылки на ячейки C25:C45 (столбец *Остатки* с названием);
- *Группирование* – установите флажок *по столбцам*;
- *Метки* – установите флажок *в первой строке*;
- *Выходной диапазон* – установите флажок на *Новый рабочий лист* и в поле напротив введите «Статистика»;
- *Итоговая статистика* – установите флажок;
- *Уровень надежности* (95%) – установите флажок.

Нажмите *ОК*.

Для оценки значимости среднего на листе «Статистика» в ячейку D3 введите формулу =(B3-0)*КОРЕНЬ(20)/B7 (для подсчета наблюдаемого значения статистики $t_{\text{набл}}$). В ячейку D4 введите формулу =СТЬЮДРАСПОБР(0,05;20-1) (для подсчета критического значения распределения Стьюдента $t_{\text{кр}}$).

2) Для проверки условия гомоскедастичности остатков с листа «Исходные данные» ячейки B1:B21 скопируйте в ячейку A1 нового листа, который назовите «Спирмен». В ячейку B1 скопируйте с листа «Регрессия» столбец «Остатки» вместе с названием. В ячейку C1 введите название «Модули». Выделите C2:C21 и введите формулу массива (нажмите F2, введите формулу, нажмите Ctrl+Shift+Enter) {=ABS(B2:B21)}.

Выберите в опциях меню *Сервис → Анализ данных → Ранг и персентиль → ОК* и заполните диалоговое окно следующим образом:

- *Входной интервал* – введите ссылки на ячейки A1:A21;
- *Метки* – установите флажок;
- *Выходной интервал* – ячейка D1.

Нажмите *ОК*.

Выберите в опциях меню *Сервис → Анализ данных → Ранг и персентиль → ОК* и заполните диалоговое окно следующим образом:

- *Входной интервал* – введите ссылки на ячейки C1:C21;
- *Метки* – установите флажок;
- *Выходной интервал* – ячейка H1.

Нажмите *ОК*.

Выделите ячейки D2:G21 и нажмите кнопку *Сортировка по возрастанию* на панели инструментов. Выделите ячейки H2:K21 и нажмите кнопку *Сортировка по возрастанию*. Скопируйте ячейки F1:F21 в ячейку M1, ячейки J1:J21 в ячейку N1. Выделите ячейки O2:O21 и введите формулу массива {= (M2:M21-N2:N21)^2}.

Установите курсор на ячейке О22 и введите формулу

= 1-6*СУММ(О2:О21)/(20*(20^2-1)).

В ячейку О23 введите формулу = О22*КОРЕНЬ(19) (для вычисления $t_{набл}$)

В ячейку О24 введите формулу = СТЬЮДРАСПОБР(0,05;20-2) (для вычисления $t_{кр}$).

3) Для проверки условия об отсутствии автокорреляции с листа «Регрессия» скопируйте столбец «Остатки» в ячейку А1 нового листа, который назовите «Автокорреляция».

В ячейки В2:В20 введите формулу массива {=(А2:А20-А3:А21)^2}.

В ячейку В22 введите формулу = СУММ(В2:В20).

В ячейки С2:С21 введите формулу массива {=(А2:А21)^2}.

В ячейку С22 введите формулу = СУММ(С2:С21).

В ячейку С23 введите формулу =В22/С22 (для вычисления *DW*).

4) Для проверки гипотезы о нормальном распределении остатков вернитесь на лист «Регрессия». Выберите в опциях меню *Сервис → Анализ данных → Гистограмма → ОК.* Заполните значения параметров окна следующим образом:

- *Входной интервал* – введите ссылки на ячейки С25:С45 (столбец *Остатки* листа «Регрессия» с названием);
- *Интервал карманов* – не заполняйте;
- *Метки* — установите флажок;
- *Выходной диапазон* – введите ссылку на *Новый рабочий лист* «Гистограмма»;
- *Парето* – оставьте пустым;
- *Интегральный процент* – оставьте пустым;
- *Вывод графика* – установите флажок.

Нажмите *ОК*. Перенесите гистограмму вниз и растяните ее.

5) Для проверки условия об отсутствии мультиколлинеарности откройте лист «Исходные данные», выберите в опциях меню *Сервис → Анализ данных → Корреляция → ОК.* Заполните значения параметров окна следующим образом:

- *Входной интервал* – ячейки В1:С21;
- *Метки* – установите флажок;
- *Новый рабочий лист* – «Мульти».

Нажмите *ОК*.

Скопируйте ячейку В3 в ячейку С2.

В ячейку А5 введите математическую формулу = МОПРЕД(В2:С3).

В ячейку А6 введите формулу =20-1-(1/6)*9*LOG(А5;10) (для нахождения хи-квадрат наблюдаемого).

В ячейку А7 введите формулу =ХИ2ОБР(0,05;1) (для нахождения хи-квадрат критического).

Приложение: Отчет о результатах вычислений и анализа

1. *Постановочный этап*

Из экономической теории известно, что заработная плата зависит от многих факторов, например, от ..... (*перечислить основные факторы*).

Выделим два фактора: *Возраст* и *Стаж по специальности*, которые являются объясняющими факторами для результативного (объясняемого) признака – *Заработная плата*. Возникает задача количественного описания зависимости указанных показателей уравнением множественной регрессии на основании статистических данных.

2. *Спецификация модели*

Предположим, что зависимость заработной платы y от возраста x_1 и стажа по данной специальности x_2 описывается линейной регрессионной моделью $y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon$, где a_0, a_1, a_2 – неизвестные параметры модели, ε – случайный член, который включает в себя суммарное влияние всех неучтенных в модели факторов.

3. *Параметризация модели*

Для оценки параметров уравнения множественной регрессии применяется *метод наименьших квадратов* (МНК). В результате проведения регрессионного анализа получены точечные и интервальные оценки неизвестных параметров.

Точечная оценка параметра a_0 равна ..... Интервальная оценка имеет вид (....., ...).

Точечная оценка параметра a_1 равна ..... Интервальная оценка имеет вид (....., ...).

Точечная оценка параметра a_2 равна ..... Интервальная оценка имеет вид (..... , ...).

Таким образом, уравнение регрессии имеет вид (*записать уравнение линейной регрессии*).

4. *Верификация модели*

4.1. Значимость коэффициентов регрессии a_0 оценивается с помощью t-статистики.

Для коэффициента a_0 наблюдаемое значение статистики $t_{\text{набл}}$ равно ..... Критическое значение $t_{\text{кр}}$ равно ..... Так как $|t_{\text{набл}}|$ (*больше или меньше*) $t_{\text{кр}}$, то коэффициент a_0 (*значим или незначим*).

Для коэффициента регрессии a_1 наблюдаемое значение статистики $t_{\text{набл}}$ равно ..... Критическое значение $t_{\text{кр}}$ равно ..... Так как $|t_{\text{набл}}|$ (*больше или меньше*) $t_{\text{кр}}$, то коэффициент a_1 (*значим или незначим*).

Для коэффициента регрессии a_2 наблюдаемое значение статистики $t_{набл}$ равно ..... Критическое значение $t_{кр}$ равно ..... Так как $|t_{набл}|$ (*больше или меньше*) $t_{кр}$, то коэффициент a_2 (*значим или незначим*).

4.2. Качество построенной модели в целом оценивает коэффициент детерминации. В таблице «Регрессионная статистика» листа «Регрессия» коэффициент множественной детерминации *R*-квадрат равен ..... *Сделать вывод об общем качестве уравнения*.

Коэффициент множественной корреляции равен ..... Он оценивает тесноту совместного влияния факторов на результат. *Сделать вывод о силе совместного влияния факторов*.

Значимость коэффициента детерминации *R*-квадрат устанавливается с помощью критерия Фишера в таблице «Дисперсионный анализ» листа «Регрессия». Наблюдаемое значение $F_{набл}$ равно ..... Критическое значение $F_{кр}$ равно ..... Так как наблюдаемое значение $F_{набл}$ (*больше, меньше*) $F_{кр}$, то *R*-квадрат (*значим или незначим). Сделать вывод об общем качестве уравнения*.

4.3. Для того, чтобы оценки параметров линейного уравнения множественной регрессии были несмещенными, состоятельными и эффективными, необходимо выполнение условий Гаусса–Маркова для случайного члена.

4.3.1. Значение «Среднее» из таблицы на листе «Статистика» равно ..... Оно является несмещенной оценкой математического ожидания случайного члена. *Сделать вывод о выполнении предпосылки* 1.

Значимость среднего устанавливается с помощью критерия Стьюдента. Так как $|t_{набл}|$=..... (*больше или меньше*) $t_{кр}$ =....., то среднее (*значимо или незначимо*). *Сделать вывод о выполнении предпосылки* 1.

4.3.2. Для диагностики гетероскедастичности применяется тест ранговой корреляции Спирмена (лист «Спирмен»). Выдвигается нулевая гипотеза об отсутствии гетероскедастичности случайного члена. Так как $|t_{набл}|$=| ..... | (*больше или меньше*) $t_{кр}$=....., то гипотеза об отсутствии гетероскедастичности (*принимается или отвергается*). *Сделать вывод о выполнении предпосылки* 2.

4.3.3. Для проверки гипотезы об отсутствии автокорреляции используется *DW*-статистика Дарбина-Уотсона. Критические значения *DW*-статистики находятся из таблицы 5.8.

Таблица 5.8. Значения *DW*-статистики при уровне значимости 0,05
(k – число факторов)

n	$k=1$		$k=2$		$k=3$		$k=4$	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2

19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81

Так как *DW* = ... при d_1 = ... и d_2 = ... попадает в зону (*положительной автокорреляции, отрицательной автокорреляции, отсутствия автокорреляции, неопределенности*), то автокорреляция (*присутствует или отсутствует*).

Сделать вывод о выполнении предпосылки 3.

4.3.4. *Сделать вывод о нормальности распределения остатков* по виду гистограммы. *Дополнить графический анализ выводами на основе теста Жарка-Бера.*

5. Для оценки мультиколлинеарности факторов используется *определитель* Δr *матрицы парных коэффициентов корреляции.*Он равен ..... Анализ мультиколлинерности факторов осуществляется проверкой гипотезы $H_0 : \Delta r = 1$ на основании статистики Пирсона χ^2.

Наблюдаемое значение $\chi^2_{\text{набл}}$ равно ..... Критическое значение $\chi^2_{\text{кр}}$ равно ..... Так как наблюдаемое значение $\chi^2_{\text{набл}}$ (*больше, меньше*) $\chi^2_{\text{кр}}$, то гипотеза (*принимается или отвергается*).

Сделать вывод о мультиколлинеарности факторов.

6. *Частные коэффициенты эластичности* для линейной регрессии рассчитывается по формуле: $\overline{Э}_{yx_j} = b_j \dfrac{\overline{x}_j}{\overline{y}}$. Они равны ..... и ..... *Сделать вывод о влиянии факторов по величине коэффициентов эластичности*.

5. *Прогнозирование*

Если выполняются все условия верификации, то модель является качественной. В противном случае ее надо усовершенствовать: либо на этапе спецификации, либо варьировать выборку. По качественной модели можно прогнозировать объем заработной платы в зависимости от возраста и стажа по данной специальности. *Сделать вывод о возможности прогнозирования*.

Точечный прогноз экспорта равен ....., доверительный интервал прогноза имеет вид (....., ...), где центр интервала равен точечному прогнозу, концы интервала получены прибавлением и вычитанием произведения стандартной ошибки прогноза на критическое значение *t*-статистики. *Сделать вывод о качестве прогноза*.

Интегрированная задача

С помощью «Пакета анализа» табличного процессора Excel, опираясь на статистические данные, находящиеся в таблице 5.9, для объяснения заработной платы y в зависимости от возраста x_1 и стажа по специальности

x_2 построить и исследовать линейную регрессионную модель $y = a_0 + a_1 x_1 + a_2 x_2 + \varepsilon$. Применить построенную регрессионую модель для прогноза заработной платы при $x_1^* = 55,\ x_2^* = 25$. Расчеты и анализ провести в соответствии с требованиями и описанием задачи, изложенной в разделе «Реализация с помощью ППП *Excel*».

Результаты вычислений и анализа представить в виде отчета по форме, предложенной выше.

Таблица 5.9. Статистические данные интегрированной задачи

1 вариант			2 вариант		
заработная плата	возраст	стаж	заработная плата	возраст	стаж
699,7	28	10	564,2	26	7
679,7	26	8	554,8	25	6
739,7	30	14	582,1	28	9
749,6	35	15	635,2	35	15
759,5	41	16	570,2	40	9
779,5	45	18	654,8	45	18
629,7	27	3	523,1	27	3
729,6	35	13	615,1	35	13
699,7	30	10	590,2	30	10
619,7	23	2	517,1	23	2
669,7	29	7	560,1	30	7
709,6	33	11	596,8	33	11
799,6	40	20	600,1	40	12
809,5	41	21	678,9	41	20
659,5	41	6	539,1	41	6
609,7	23	1	507,6	23	1
719,6	32	12	608,2	32	12
729,6	37	13	613,4	37	13
669,6	31	7	558,5	31	7
689,7	30	9	588,1	22	9

3вариант			4 вариант		
заработная плата	возраст	стаж	заработная плата	возраст	стаж
725	20	5	830,5	20	5
752,5	35	6	858	35	6
791	28	10	896,5	28	10
775,3	35	8	880,8	35	8
793,5	42	9	899	42	9
801	38	10	906,5	38	10
703,5	20	3	809	20	3
809	35	11	914,5	35	11
793	30	10	898,5	30	10
695,5	23	2	801	23	2
759	30	7	864,5	30	7

806	32	11	911,5	32	11
825	40	12	930,5	40	12
813,5	28	12	919	28	12
758,7	41	6	864,2	41	6
684,3	23	1	789,8	23	1
815	32	12	920,5	32	12
833,5	37	13	939	37	13
753,5	24	7	859	24	7
775	22	9	880,5	22	9

5 вариант			6 вариант		
заработная плата	возраст	стаж	заработная плата	возраст	стаж
735,5	28	10	604,2	26	7
712,8	26	8	594,8	25	6
778,2	30	14	622,5	28	9
794,5	35	15	675,2	35	15
812,3	41	16	610,2	40	9
857,5	45	18	694,8	45	18
664,2	27	3	563,9	27	3
774,5	35	13	655,1	35	13
738,1	30	10	630	30	10
648,5	23	2	557,7	23	2
706,7	29	7	600,5	30	7
751,9	33	11	636,8	33	11
851,1	40	20	640,3	40	12
862,2	41	21	718,9	41	20
712,3	41	6	579,5	41	6
638,9	23	1	547,6	23	1
760,6	32	12	648,2	32	12
777,2	37	13	653,4	37	13
709,2	31	7	598,5	31	7
728,1	30	9	628,2	22	9

7 вариант			8 вариант		
заработная плата	возраст	стаж	заработная плата	возраст	стаж
725	20	5	824,8	26	7
752,5	35	6	815,4	25	6
791	28	10	843,1	28	9
775,3	35	8	895,8	35	15
793,5	42	9	830,8	40	9
801	38	10	915,4	45	18
703,5	20	3	784,5	27	3
809	35	11	875,7	35	13
793	30	10	850,6	30	10
695,5	23	2	778,3	23	2
759	30	7	821,1	30	7

806	32	11	857,4	33	11
825	40	12	860,9	40	12
813,5	28	12	939,5	41	20
758,7	41	6	800,1	41	6
684,3	23	1	768,2	23	1
815	32	12	868,8	32	12
833,5	37	13	874	37	13
753,5	24	7	819,1	31	7
775	22	9	848,8	22	9

Контрольные задания

Задание 5.1. На основании статистических данных, приведенных в таблице 5.10, исследуется зависимость урожайности зерновых культур y от следующих факторов производства:

x_1 – число тракторов на 100 га;

x_2 – число зерноуборочных комбайнов на 100 га;

x_3 – число орудий поверхностной обработки почвы на 100 га;

x_4 – количество удобрений, расходуемых на гектар (т/га);

x_5 – количество химических средств защиты растений (т/га).

Необходимо:

1) построить линейную регрессионную модель с полным набором факторов;

2) с помощью критерия χ^2 проверить наличие в построенной модели мультиколлинеарности;

3) в случае присутствия мультиколлинеарности в модели с полным набором факторов методом последовательного включения факторов построить множественную линейную модель, не имеющую мультиколлинеарности между факторами;

4) оценить и сравнить качество построенных моделей.

Таблица 5.10. Статистические данные задания 5.1

y	x_1	x_2	x_3	x_4	x_5
29,70	1,59	0,26	2,05	0,32	0,14
28,40	0,34	0,28	0,46	0,59	0,66
29,00	2,53	0,31	2,46	0,30	0,31
29,90	4,63	0,40	6,44	0,43	0,59
29,60	2,16	0,26	2,16	0,39	0,16
28,60	2,16	0,30	2,69	0,32	0,17
32,50	0,68	0,29	0,73	0,42	0,23
27,60	0,35	0,26	0,42	0,21	0,08
28,90	0,52	0,24	0,49	0,20	0,08
33,50	3,42	0,31	3,02	1,37	0,73
29,70	1,78	0,30	3,19	0,73	0,17

30,70	2,40	0,32	3,30	0,25	0,14
22,20	9,36	0,40	11,51	0,39	0,38
29,70	1,72	0,28	2,26	0,82	0,17
27,00	0,59	0,29	0,60	0,13	0,35
27,20	0,28	0,26	0,30	0,09	0,15
28,20	1,64	0,29	1,44	0,20	0,08
28,40	2,09	0,22	2,05	0,43	0,2
33,10	2,08	0,25	2,03	0,73	0,2
28,70	1,36	0,26	0,17	0,99	0,42

Задание 5.2. Торговое предприятие имеет сеть, состоящую из 12 магазинов. Статистические данные о работе магазинов представлены в таблице 5.11.

Необходимо:

1) построить линейную регрессионную модель зависимости товарооборота магазина от среднего числа посетителей в день;

2) графическим методом оценить выполнение условий Гаусса-Маркова;

3) проверить выполнение модельных предположений с помощью статистических методов;

4) сравнить результаты.

Таблица 5.11. Статистические данные задания 5.2

Номер магазина	Товарооборот, y (млн руб.)	Среднее число посетителей в день, x (тыс. чел.)
1	197,6	8,2
2	380,9	10,3
3	409,5	9,4
4	410,8	11,2
5	562,7	8,6
6	685,1	7,7
7	750,2	12,4
8	890,7	10,9
9	911,3	9,6
10	912,6	13,6
11	998,4	12,4
12	1083,3	13,9

Задание 5.3. В результате применения инструмента «Регрессия» получены график и гистограмма остатков (рисунки 5.12 и 5.13). Необходимо графическим способом оценить выполнение условий теоремы Гаусса-Маркова для соответствующей парной линейной модели регрессии.

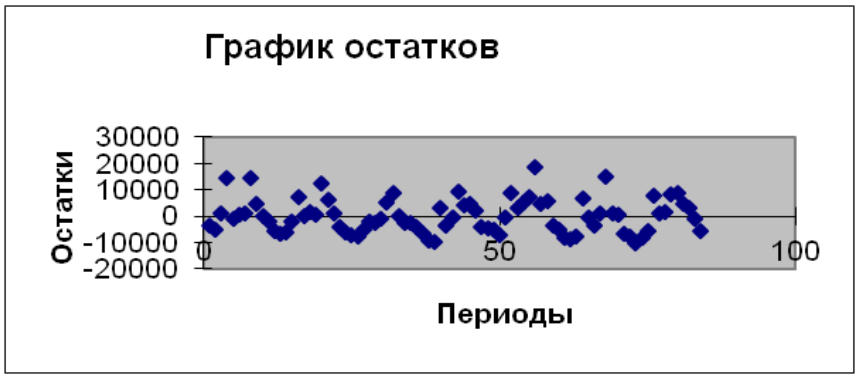


Рис. 5.12. График остатков задания 5.3

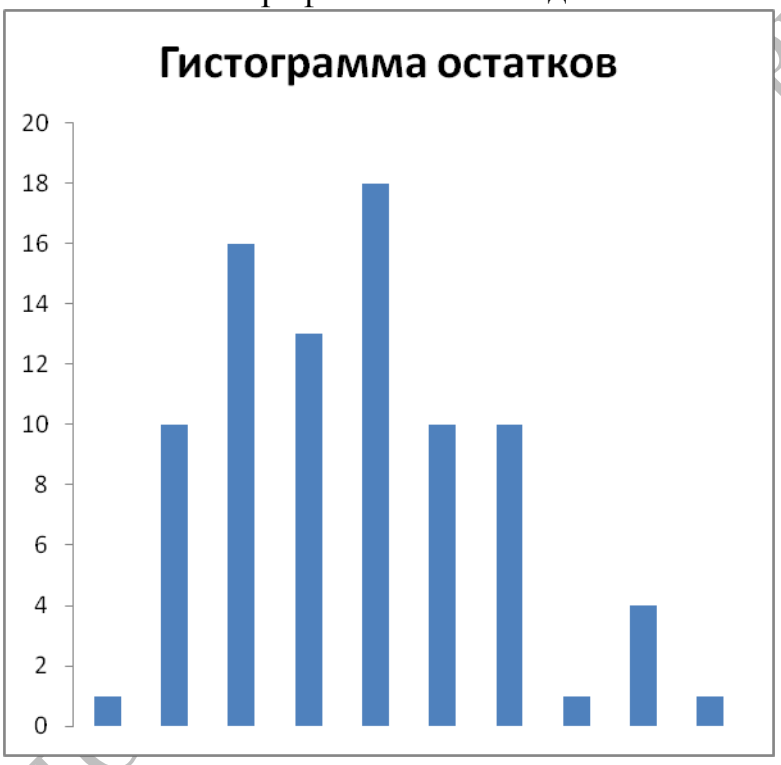


Рис. 5.13. Гистограмма остатков задания 5.3

Задание 5.4. По статистическим данным таблицы 5.14 для десяти сельскохозяйственных предприятий исследуется зависимость урожайности зерновых культур (переменная y, ц/га) от качества земли (переменная x, баллы).

1. Построить парную линейную модель зависимости y от x.

2. Проверить наличие в модели гетероскедастичности с помощью графического анализа и методом Голдфелда–Квандта.

3. При обнаружении гетероскедастичности построить взвешенную модель регрессии.

Таблица 5.14. Статистические данные задания 5.4

	1	2	3	4	5	6	7	8	9	10
x	25	27	28	29	30	29	31	32	32	34
y	23,0	23,3	24,0	24,5	24,2	25,0	27,1	28,8	29,3	30,2

Задание 5.5. По статистическим данным задания 5.4 построить парную линейную модель вида $y = bx + \varepsilon$. Для построенной модели проверить выполнение первого условия теоремы Гаусса-Маркова:

1) с помощью графического анализа остатков регрессии;

2) с помощью оценки величины среднего выборочного остатков регрессии;

3) с помощью проверки гипотезы о равенстве нулю математического ожидания случайного члена.

Задание 5.6. По статистическим данным таблицы 5.15 для двенадцати предприятий отрасли исследуется (в расчете на одного работающего) зависимость производительности труда (переменная y, млн руб.) от энерговооруженности труда (переменная x, кВт).

1. Построить парную линейную модель зависимости y от x.

2. Проверить наличие в модели гетероскедастичности с помощью графического анализа и методом Голдфелда–Квандта.

3. При обнаружении гетероскедастичности построить взвешенную модель регрессии.

Таблица 5.15. Статистические данные задания 5.6

x	2,7	2,3	2,9	3,4	3,3	3,6	4,1	4,7	5,8	5,7	6,6	9,6
y	6,9	7,0	7,4	7,6	8,5	8,9	9,2	9,9	10,9	10,8	12,5	15,8

Задание 5.7. По статистическим данным задания 5.6 построить парную линейную модель вида $y = bx + \varepsilon$. Для построенной модели проверить выполнение первого условия теоремы Гаусса-Маркова:

1) с помощью графического анализа остатков регрессии;

2) с помощью оценки величины среднего выборочного остатков регрессии;

3) с помощью проверки гипотезы о равенстве нулю математического ожидания случайного члена.

Задание 5.8. На основании статистических данных, приведенных в таблице 5.16, изучается зависимость стоимости квартиры y (тыс. долларов) от ее общей площади x ($м^2$).

Таблица 5.16. Статистические данные задания 5.8

x	27	29	32	34	41	44	45	50	53
y	15,1	19,2	25,3	31,4	33,6	35,6	34,8	37,4	38,8
x	58	63	65	72	75	81	84	89	93
y	41,8	45,1	44,8	48,9	51,7	53,6	51,9	64,7	68,1

Необходимо:

1) Построить парную линейную модель зависимости y от x.

2) Построить график и гистограмму остатков.

3) Сделать выводы о характере остатков модели, проиллюстрировать их графиками.

4) Вычислив коэффициент асимметрии и эксцесса, проверить условие о нормальности распределения остатков.

5) Методом Голдфелда–Квандта проверить наличие в модели гетероскедастичности.

Задание 5.9. На основании статистических данных, приведенных в таблице 5.17, изучается зависимость урожайности зерновых культур y (ц/га) от количества внесенных удобрений на один гектар x (тонны).

Таблица 5.17. Статистические данные задания 5.9

x	3,6	4,1	5,2	5,9	6,7	7,7	10,2	11,9	13,0	14,1
y	26,1	24,8	25,1	26,3	27,6	28,3	29,7	31,3	35,7	34,5

Необходимо:

1) Построить парную линейную модель зависимости y от x.

2) Построить график остатков.

3) На основе графического анализа сделать выводы о характере остатков модели.

4) Вычислив коэффициент асимметрии и эксцесса, проверить условие о нормальности распределения остатков.

5) Методом Голдфелда–Квандта проверить наличие в модели гетероскедастичности.

6) Проверить тест Дарбина-Уотсона на наличие в модели автокорреляции.

Задание 5.10. На основании статистических данных, приведенных в таблице 5.18, изучается зависимость объема спроса на товар y (кг) от его цены x (тыс. руб.).

Таблица 5.18. Статистические данные задания 5.10

x	9,8	8,3	7,8	6,9	6,5	5,9	5,7	5,3	5,5	4,9
y	1100	1220	1980	2120	2350	2460	2540	2640	2850	2980

Необходимо:

1) Построить парную линейную модель зависимости y от x.

2) Построить график остатков.

3) На основе графического анализа сделать выводы о характере остатков модели.

4) Вычислив коэффициент асимметрии и эксцесса, проверить условие о нормальности распределения остатков.

5) Методом Голдфелда–Квандта проверить наличие в модели гетероскедастичности.

6) Проверить тест Дарбина-Уотсона на наличие в модели автокорреляции.

Контрольные вопросы

1. Назовите основные допущения в нормальной классической линейной модели регрессии.
2. С какой целью формулируются и что обеспечивают модельные предположения?
3. В чем заключается несмещенность, эффективность и состоятельность оценок?
4. В связи с чем возникает необходимость рассмотрения методов обнаружения и устранения нарушений предпосылок МНК?
5. Как решается проблема выполнения первого условия теоремы Гаусса-Маркова?
6. Какими способами может быть протестировано условие равенства нулю математического ожидания случайной переменной?
7. В чем заключается проблема гетероскедастичности?
8. Назовите основные причины гетероскедастичности.
9. Каковы последствия гетероскедастичности?
10. Как проявляется гетероскедастичность при использовании пространственных выборок и при использовании данных временных рядов?
11. Как можно с помощью графического анализа проверить наличие гомо- или гетероскедастичности?
12. Опишите алгоритм теста Голдфелда-Квандта.
13. Опишите схему теста ранговой корреляции Спирмена.
14. Опишите методы устранения гетероскедастичности.
15. В чем суть взвешенного метода наименьших квадратов?
16. Что такое автокорреляция?
17. Как проявляется автокорреляция при использовании пространственных выборок и при использовании данных временных рядов?
18. Как определяется положительная и отрицательная автокорреляции?
19. Как отражается положительная и отрицательная автокорреляции на изображении точек корреляционного поля?

20. Назовите основные причины автокорреляции.
21. Каковы последствия автокорреляции?
22. Перечислите основные методы диагностики автокорреляции.
23. Опишите схему использования статистики Дарбина-Уотсона.
24. Опишите методы устранения автокорреляции.
25. В чем суть обобщенного метода наименьших квадратов?
26. Что такое мультиколлинеарность?
27. Назовите причины мультиколлинеарности.
28. К каким последствиям приводит мультиколлинеарность факторов, включенных в модель?
29. Как определяется матрица парных коэффициентов корреляции?
30. Опишите метод выявления явно коллинеарных факторов, основанный на анализе матрицы парных коэффициентов корреляции.
31. Опишите приемы диагностики мультиколлинеарности.
32. Перечислите основные методы устранения мультиколлинеарности.
33. Опишите рекомендуемую последовательность проверки модельных предположений.

Тестовые задания

Выберите правильные ответы из предложенных вариантов:

1. Если выполнены основные предпосылки МНК – условия Гаусса-Маркова, то коэффициенты уравнения регрессии как оценки параметров модели обладают свойствами:
а) несостоятельность;
б) несмещенность;
в) достоверность;
г) состоятельность;
д) эффективность;
е) смещенность.

2. Несмещенность оценки параметра регрессии, полученной по МНК, означает:
а) что она характеризуется наименьшей дисперсией;
б) что математическое ожидание случайной переменной равно нулю;
в) увеличение ее точности с увеличением объема выборки.

3. Эффективность оценки параметра регрессии, полученной по МНК, означает:
а) что она характеризуется наименьшей дисперсией;
б) что математическое ожидание остатков равно нулю;
в) увеличение ее точности с увеличением объема выборки.

4. Состоятельность оценки параметра регрессии, полученной по МНК, означает:

а) что она характеризуется наименьшей дисперсией;

б) что математическое ожидание остатков равно нулю;

в) увеличение ее точности с увеличением объема выборки.

5. К классическим предпосылкам МНК – условиям Гаусса-Маркова – относятся следующие условия:

а) гомоскедастичность;

б) равенство нулю дисперсий случайных отклонений;

в) наличие автокорреляции;

г) равенство нулю математического ожидания случайных отклонений;

д) гетероскедастичность;

е) независимость случайных отклонений между собой.

6. Согласно предпосылке теоремы Гаусса-Маркова дисперсии случайных возмущений во всех наблюдениях должны быть:

а) равными;

б) различными;

в) нулевыми.

7. Согласно предпосылке теоремы Гаусса-Маркова математическое ожидание случайного члена должны быть:

а) положительным;

б) отрицательным;

в) нулевым;

г) большим единицы.

8. Условие гетероскедастичности заключается:

а) в непостоянстве дисперсии остатков;

б) в постоянстве дисперсии остатков;

в) в равенстве нулю математического ожидания случайного члена;

г) в зависимости случайных отклонений разных наблюдений между собой.

9. Проблема гетероскедастичности характерна для:

а) пространственных данных, полученных от однородных объектов;

б) пространственных данных, полученных от неоднородных объектов;

в) временных рядов.

10. Гомоскедастичность остатков подразумевает:

а) рост дисперсии остатков с увеличением значения фактора;

б) максимальную дисперсию остатков при средних значениях фактора;

в) уменьшение дисперсии остаток с уменьшением значения фактора;

г) одинаковую дисперсию остатков при каждом значении фактора.

11. При нарушении предпосылки МНК о нормальном законе распределения остатков:

а) оценки параметров уравнения регрессии будут смещенными;

б) оценки параметров уравнения регрессии будут не эффективными;

в) возникнут проблемы при оценке точности уравнения регрессии и его коэффициентов;

г) исказится смысл коэффициентов регрессии.

12. Проблема автокорреляции характерна для:

а) пространственных данных, полученных от однородных объектов;

б) пространственных данных, полученных от неоднородных объектов;

в) временных рядов.

13. Для оценки наличия мультиколлинеарности в множественной линейной модели используется:

а) определитель матрицы парных коэффициентов корреляции;

б) коэффициент множественной корреляции;

в) коэффициент множественной детерминации.

14. Наличие мультиколлинеарности факторов в множественной линейной регрессионной модели оценивается с помощью:

а) критерия Стьюдента;

б) критерия Фишера;

в) критерия Пирсона χ-квадрат;

г) критерий Дарбина-Уотсона.

15. Два фактора x и y явно коллинеарны, если:

а) $r_{xy} \geq 0{,}3$;

б) $r_{xy} \geq 0{,}5$;

в) $r_{xy} \geq 0{,}7$;

г) $r_{xy} \leq 0{,}6$.

16. К простейшим методам устранения мультиколлинеарности относится:

а) изменение спецификации;

б) увеличение объема выборки;

в) преобразование факторных переменных;

г) устранение в уравнении свободного члена.

17. Нормальность распределения случайного члена может быть оценена с помощью:

а) теста Чоу;

б) теста Жарка-Бера;

в) теста Дарбина-Уотсона.

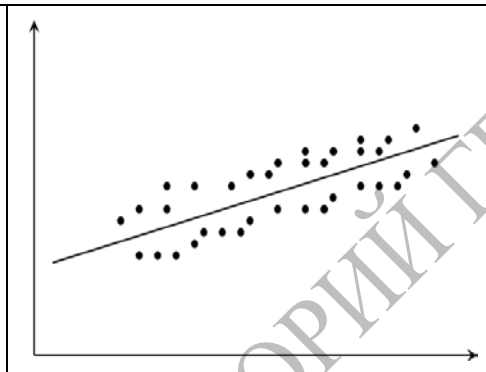
18. Нарушение первой предпосылки теоремы Гаусса-Маркова связано:
а) со смещенностью оценок;
б) с неэффективностью оценок;
в) с несостоятельностью оценок.

19. Проблема гетероскедастичности заключается:
а) в постоянстве и конечности дисперсии остатков;
б) в непостоянстве дисперсии остатков;
в) в равенстве нулю математического ожидания случайного члена;
г) в зависимости случайных отклонений разных наблюдений между собой.

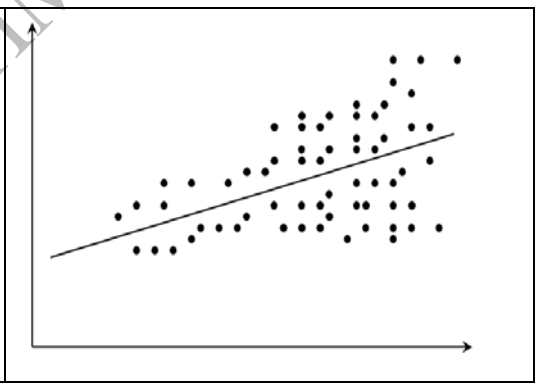
20. Для оценки нарушения гомоскедастичности используется:
а) критерий Дарбина-Уотсона;
б) тест ранговой корреляции Спирмена;
в) тест Голдфелда-Квандта;
г) теорема Гаусса-Маркова.

21. Какой график указывает на наличие в модели гетероскедастичности:

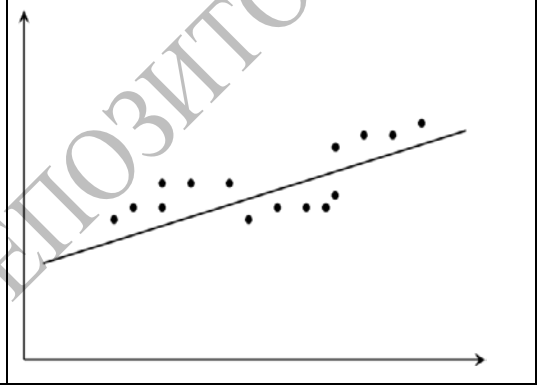
а)



б)



в)



22. Для определения автокорреляции используется:
а) критерий Дарбина-Уотсона;
б) тест ранговой корреляции Спирмена;
в) тест Голдфелда-Квандта;
г) теорема Гаусса-Маркова.

23. Мультиколлинеарность – это:

а) коррелированность двух или нескольких переменных в уравнении регрессии;

б) непостоянство дисперсии остатков;

в) равенство нулю математического ожидания случайного члена;

г) зависимость случайных отклонений разных наблюдений.

24. Равенство нулю математического ожидания случайной переменной гарантирует, что:

а) обеспечена возможность применения аппарата проверки статистических гипотез;

б) при возрастании числа *n* наблюдений дисперсия оценок параметров регрессии стремится к нулю;

в) оценки имеют наименьшую дисперсию по сравнению с любыми другими оценками;

г) в определении линии регрессии отсутствует систематическая ошибка.

Ответы тестовых заданий

Номер задания	Ответы	Номер задания	Ответы	Номер задания	Ответы
1	б), г), д)	9	б)	17	б)
2	б)	10	г)	18	а)
3	а)	11	в)	19	б)
4	в)	12	в)	20	б), в)
5	а), г), е)	13	а)	21	б)
6	а)	14	в)	22	а)
7	в)	15	в)	23	а)
8	а)	16	а), б), в)	24	г)

Глава 6
Моделирование временных рядов

Основные понятия: *временной ряд, модель временного ряда, тренд, сезонная, циклическая и случайная компоненты временного ряда, аддитивная и мультипликативная модели временного ряда, автокорреляция уровней временного ряда, автокорреляционная функция, коррелограмма, выравнивание временного ряда, кривая роста, «белый шум», структурный сдвиг, экстраполяция уровней временного ряда, динамическая модель, модель с распределенным лагом, авторегрессионная модель, коинтеграция.*

Литература: [2-4], [7,8], [12], [18,19].

6.1. Модель временного ряда

Во многих экономических исследованиях используются данные, которые характеризуют объект, процесс или явление за ряд последовательных промежутков времени. К таким данным относятся данные ежедневных, ежемесячных, ежеквартальных или ежегодных наблюдений за некоторыми микро- или макроэкономическими показателями. Например, речь может идти о ежемесячных показателях инфляции в одной из стран, о месячных данных по объему производства продукции на некотором предприятии, о ежедневных наблюдениях за курсом акций на биржевых торгах.

Совокупность данных наблюдений некоторого показателя y, упорядоченная по времени их получения, в экономике называется *временным* (*динамическим*) *рядом*. Отдельные наблюдения временного ряда называются *уровнями* ряда. Если y_t – значение переменной y в момент времени t, то последующие уровни показателя обозначаются y_{t+1}, y_{t+2}, …, а предыдущие y_{t-1}, y_{t-2}, ….

Характерным для временного ряда $y_{t_1}, y_{t_2}, ..., y_{t_n}$ является то, что порядок в последовательности $t_1, t_2, ..., t_n$ существенен для анализа, т.е. время выступает

как один из определяющих факторов. Это отличает временной ряд от пространственной выборки, где индексы служат лишь для удобства идентификации.

К уровням временного ряда предъявляется ряд требований: они должны быть *однородны* и *сопоставимы* (в первую очередь это касается стоимостных и ценовых характеристик), т.е. сформированы по одним методикам, иметь одинаковые единицы измерения и одинаковый шаг наблюдений. Кроме того, временной ряд должен иметь достаточную длину и в нем должны отсутствовать пропущенные наблюдения.

Однородность данных означает отсутствие *аномальных* (т.е. резко выделяющихся, нетипичных для данного ряда) наблюдений. Аномальные наблюдения проявляются в виде сильного изменения уровня (скачка или спада) с последующим приблизительным восстановлением предыдущего уровня. Наличие аномалии резко искажает результаты моделирования. Поэтому аномальные наблюдения необходимо исключить из временного ряда, заменив их расчетными значениями.

Появление аномальных значений может быть вызвано ошибками при сборе информации, записи или передаче информации (ошибки первого рода). Они не отражают никакой тенденции. Однако аномальные значения могут отражать и реальные процессы, например, скачок курса доллара или его падение. Такие аномальные уровни относят к ошибкам второго рода.

Для выявления аномальных уровней временного ряда можно применить *метод Ирвина*.

Этот метод предполагает использование для всех уровней временного ряда $\{y_t \mid t = 1,2,...,n\}$ (или только для уровней, подозреваемых в аномальности) показателя λ_t $(t = 2,3,...,n)$, рассчитываемого по следующей формуле:

$$\lambda_t = \frac{|y_t - y_{t-1}|}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}}. \qquad (6.1)$$

Расчетные значения $\lambda_2, \lambda_3, ..., \lambda_n$ сравниваются с табличными значениями критерия Ирвина λ_0; если некоторое значение λ_t оказывается больше табличного значения λ_0, то соответствующее значение y_t уровня t временного ряда считается аномальным.

Значения критерия Ирвина для уровня значимости $\alpha = 0{,}05$ приведены в таблице 6.1.

Таблица 6.1. Критические значения параметра Ирвина

n	10	20	30	50	100

λ_0	1,5	1,3	1,2	1,1	1,0

После выявления аномальных уровней временного ряда определяются причины их возникновения. Если аномалии вызваны ошибками первого рода, то соответствующие уровни корректируются (чаще всего аномальные значения заменяются средней арифметической величиной двух соседних уровней ряда). Ошибки, возникающие из-за воздействия факторов, имеющих объективный характер, не всегда подлежат устранению.

Эконометрическая модель, построенная на основе данных временного ряда, называется *моделью временного ряда*.

При моделировании по данным временных рядов необходимо учитывать следующие особенности:

- фактор времени естественным образом упорядочивает данные, т.е. важен порядок, в котором записаны данные временного ряда (в отличие от пространственной выборки);
- в отличие от пространственной выборки во временном ряду естественно допускать зависимость между уровнями ряда («эффект памяти»), т.е. наличие автокорреляции;
- часто приходится работать с временными рядами небольшой длины и нет возможности получить выборочные данные большого объема.

6.2. Компоненты временного ряда

Каждый уровень временного ряда формируется под воздействием большого числа факторов. Их условно можно разделить на три группы:

1) факторы, которые определяют изменения анализируемого показателя в длительной перспективе;

2) факторы, которые циклически изменяются во времени и формируют колебания уровней ряда;

3) факторы, которые не поддаются учету и регистрации и оказывают на анализируемый показатель несистематическое или случайное влияние.

Поэтому при исследовании экономического временного ряда y_t, t = 1, 2, …, n, выделяются несколько составляющих:

– *тренд*, описывающий общее направление развития показателя, устойчивую долговременную тенденцию изменения экономического показателя y;

– *сезонная компонента*, отражающая повторяемость экономических процессов в течение не очень значительного периода (года, месяца, недели);

– *циклическая компонента*, определяющая периодические колебания экономических процессов в течение длительных периодов (больше года);

– *случайная компонента*, отражающая влияние на уровни ряда случайных факторов.

Тренд – это систематическая составляющая долговременного действия. Обычно он описывается с помощью той или иной, как правило, монотонной

функции $f(t)$ (аргументом которой является время). Эту функцию называют *функцией тренда*.

Сезонная компонента связана с наличием факторов, действующих с короткой и, как правило, заранее известной периодичностью. В отличие от трендовой компоненты сезонная имеет фазу возрастания и убывания в каждом периоде. Ее циклический характер связан с флуктуацией экономической активности.

В настоящее время в экономике известны около 1500 длинных циклов: циклы перепроизводства (10-15 лет), цикл Кузнеца (15-25 лет, связан с демографическими процессами), цикл Жуглара (7-12 лет, связан с периодичностью инвестиционного процесса), цикл Кондратьева (50-60 лет, связан с закономерностями накопления и обновления научных знаний, их влиянием на экономическое развитие) и др.

Случайная компонента отражает воздействие на уровни ряда многочисленных факторов случайного характера. Она является обязательной составной частью любого временного ряда в экономике, так как случайные отклонения неизбежно сопутствуют любому экономическому явлению.

Первые три компоненты временного ряда являются закономерными (систематическими и неслучайными). Конечно, временной ряд может и не содержать одну или несколько закономерных компонент. Графические модели таких рядов приведены на рисунках 6.1-6.3. В то же время случайная компонента всегда присутствует во временном ряду.

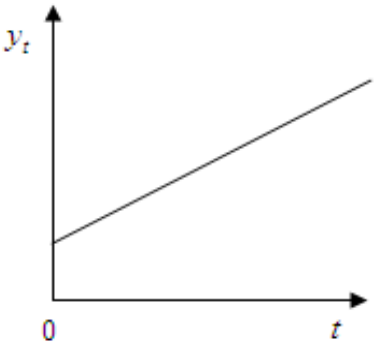


Рис. 6.1. Временной ряд, содержащий только тренд

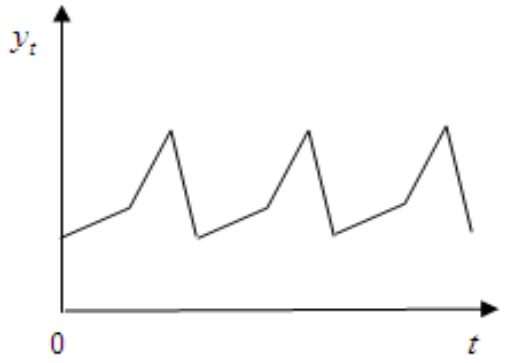


Рис. 6.2. Временной ряд, содержащий только сезонную компоненту

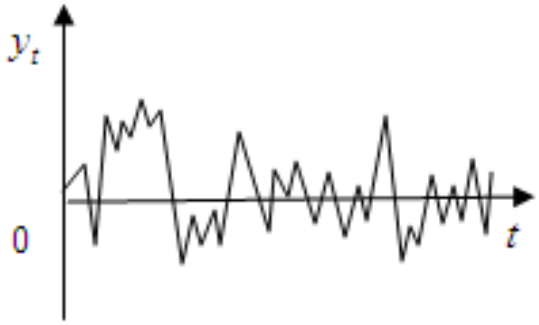


Рис. 6.3. Временной ряд, содержащий только случайную компоненту

В общем случае уровень y_t временного ряда можно представить как функцию вида $y_t = f(T, S, C, \varepsilon)$, где T – трендовая компонента (тенденция), S – сезонная компонента, C – циклическая компонента, ε – случайная компонента.

Модель $y_t = T(t) + S(t) + C(t) + \varepsilon(t)$, в которой временной ряд представлен в виде суммы всех компонент, называется *аддитивной моделью временного ряда.* Аддитивная модель применима в тех случаях, когда анализируемый временной ряд имеет приблизительно одинаковые изменения на протяжении всей длительности ряда (пример подобного изменения приведен на рис. 6.4).

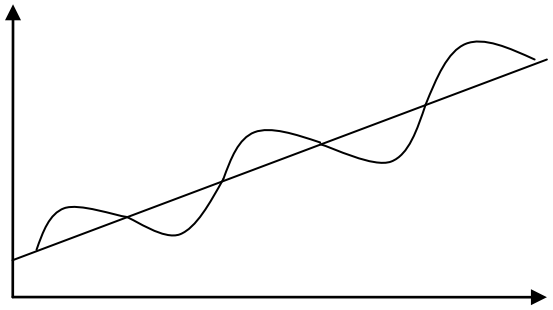


Рис. 6.4. Модель временного ряда с линейным трендом

и аддитивным сезонным эффектом

Модель $y_t = T(t) \cdot S(t) \cdot C(t) \cdot \varepsilon(t)$, в которой временной ряд представлен в виде произведения всех компонент, называется *мультипликативной моделью временного ряда.* Мультипликативная модель применяется, когда колебания уровней временного ряда со временем нарастают или затухают (рис. 6.5).

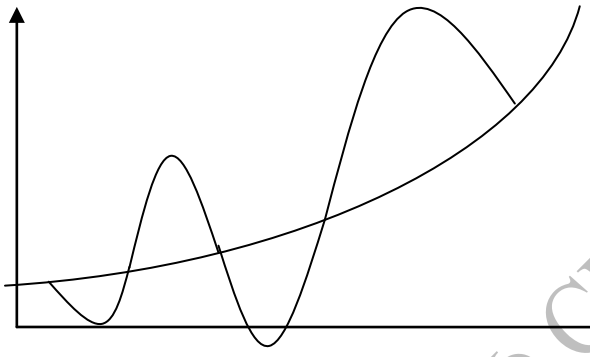


Рис. 6.5. Модель временного ряда с экспоненциальным трендом и мультипликативным сезонным эффектом

Используются и смешанные типы моделей, например, модель $y_t = T(t) \cdot S(t) \cdot C(t) + \varepsilon(t)$.

К основным целям анализа временных рядов относятся следующие: определение природы ряда и прогнозирование, т.е. предсказание будущих значений временного ряда по настоящим и прошлым значениям. Исходя из этого, центральными задачами эконометрического исследования отдельного временного ряда являются:

– выделение и количественное выражение закономерных компонент (тенденции, сезонности и цикличности);

– анализ случайной составляющей;

– нахождение прогнозных значений будущих уровней временного ряда.

6.3. Выявление структуры временного ряда

При наличии во временном ряде тренда и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционную зависимость между последовательными уровнями временного ряда, сдвинутыми на определенный промежуток времени k, называют *автокорреляцией уровней ряда*. При этом величина сдвига k называется *лагом*.

Количественно автокорреляцию можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого же ряда, сдвинутыми на лаг k (рекомендуемая максимальная

величина сдвига – не более $\frac{n}{4}$, где n – число наблюдений). Такой коэффициент называется *коэффициентом автокорреляции k-ого порядка* и обозначается через r_k (величина k лага определяет порядок коэффициента автокорреляции).

В частности, формула для расчета коэффициента автокорреляции r_1 первого порядка получается из формулы линейного коэффициента корреляции

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

если переменная x принимает значения $y_2, y_3, ..., y_n$, а переменная y принимает значения $y_1, y_2, ..., y_{n-1}$:

$$r_1 = \frac{\sum_{i=2}^{n}(y_i - \bar{y}_1)(y_{i-1} - \bar{y}_2)}{\sqrt{\sum_{i=2}^{n}(y_i - \bar{y}_1)^2 \cdot \sum_{i=2}^{n}(y_{i-1} - \bar{y}_2)^2}}, \; \bar{y}_1 = \frac{1}{n-1}\sum_{i=2}^{n} y_i \; , \; \bar{y}_2 = \frac{1}{n-1}\sum_{i=2}^{n} y_{i-1} \; .$$

Аналогично имеем:

$$r_2 = \frac{\sum_{i=3}^{n}(y_i - \bar{y}_3)(y_{i-2} - \bar{y}_4)}{\sqrt{\sum_{i=3}^{n}(y_i - \bar{y}_3)^2 \cdot \sum_{i=3}^{n}(y_{i-2} - \bar{y}_4)^2}}, \; \bar{y}_3 = \frac{1}{n-2}\sum_{i=3}^{n} y_i \; , \; \bar{y}_4 = \frac{1}{n-2}\sum_{i=3}^{n} y_{i-2} \; ;$$

.....

$$r_k = \frac{\sum_{i=k+1}^{n}(y_i - \bar{y}_{2k-1})(y_{i-k} - \bar{y}_{2k})}{\sqrt{\sum_{i=k+1}^{n}(y_i - \bar{y}_{2k-1})^2 \cdot \sum_{i=k+1}^{n}(y_{i-k} - \bar{y}_{2k})^2}}, \; \bar{y}_{2k-1} = \frac{1}{n-k}\sum_{i=k+1}^{n} y_i \; , \; \bar{y}_{2k} = \frac{1}{n-k}\sum_{i=k+1}^{n} y_{i-k} \; .$$

Для выявления структуры (состава компонент) временного ряда используют *автокорреляционную функцию*, под которой понимается последовательность

$$r_1, r_2, ..., r_k, ...$$

коэффициентов автокорреляции первого, второго и более высоких порядков.

График зависимости значений коэффициентов автокорреляции от величины лага (порядка коэффициента автокорреляции) называется *коррелограммой*.

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, а следовательно, и лаг, при котором связь между текущими и предыдущими уровнями наиболее тесная.

Если в последовательности $r_1, r_2, ..., r_k, ...$ самым высоким оказывается коэффициент r_1, то исследуемый ряд содержит тенденцию и, скорее всего, только ее. Если наиболее высоким оказывается коэффициент r_k, $k > 1$, то ряд содержит циклические колебания с периодом *k*.

Если ни один из коэффициентов r_k не является значимым, то можно сделать следующий альтернативный вывод:

- либо ряд не содержит тренда и периодических колебаний, а его уровень определяется только случайной компонентой;
- либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ.

Таким образом, автокорреляционную функцию целесообразно применять для выявления во временном ряде наличия или отсутствия трендовой компоненты $f(t)$ и сезонной компоненты $S(t)$.

Существуют и другие подходы идентификации временного ряда. Например, если тренд является монотонным (устойчиво возрастает или устойчиво убывает), то он может быть выявлен с помощью визуальной оценки уровней временного ряда или анализа его точечного графика.

Другой простой прием обнаружения тенденции развития явления – укрупнение интервала временного ряда. Смысл этого приема заключается в том, что первоначальный временной ряд преобразуется и заменяется другим, уровни которого относятся к большим по продолжительности периодам времени.

Для проверки гипотезы о существовании тренда во временном ряду могут быть использованы статистические подходы, в частности, – метод сравнения средних уровней временного ряда.

В соответствии с этим методом временной ряд, содержащий n наблюдений, разбивается на две приблизительно равные части объемами n_1 и n_2 ($n = n_1 + n_2$). Каждая из частей рассматривается как самостоятельный временной ряд, имеющий нормальное распределение. Если полный временной ряд имеет тренд, то средние, вычисленные для каждой

совокупности, должны существенно (значимо) различаться между собой. Если же расхождение несущественно (случайно), то временной ряд не имеет тренда. Гипотеза о наличии тренда проверяется при условии однородности выделенных частей временного ряда.

Условие однородности оценивается с помощью F-критерия Фишера. Пусть S_1^2, S_2^2 – выборочные дисперсии выделенных частей временного ряда. Если при $S_1^2 > S_2^2$ имеет место неравенство $F_{\text{набл}} = \frac{S_1^2}{S_2^2} < F_{\text{кр}}$, то нет оснований отвергать нулевую гипотезу о равенстве дисперсий частей временного ряда, дисперсии различаются незначимо и считается, что выделенные части временного являются однородными ($F_{\text{кр}}$ – критическое значение статистики Фишера при уровне значимости α и числе степеней свободы $k_1 = n_1 - 1$, $k_2 = n_2 - 1$).

Далее для частей временного ряда вычисляются средние арифметические значения $\bar{y}_1, \bar{y}_2$ и применяется t-статистика, имеющая распределение Стьюдента с числом степеней свободы, равным $n - 2$.

Тренд во временном ряду присутствует, если

$$t_{\text{набл}} = \frac{\left|\bar{y}_1 - \bar{y}_2\right|}{\sqrt{S_1^2(n_1 - 1) + S_2^2(n_2 - 1)}} \cdot \sqrt{\frac{n_1 n_2 (n-2)}{n}} \geq t_{\text{кр}}.$$

6.4. Выравнивание временного ряда

После выявления структуры временного ряда, первым этапом эконометрического моделирования по данным временного ряда является моделирование тренда – систематической составляющей, зависящей только от времени. Этот этап получил название *выравнивания временного ряда*.

Для выравнивания временного ряда используются методы механического и аналитического выравнивания.

Методы механического выравнивания (*метод скользящих средних*, *метод экспоненциального сглаживания* и др.) подробно рассматриваются в курсе статистики.

В частности, метод скользящей средней основан на переходе от заданных уровней временного ряда к их средним значениям на интервале времени, длина которого определена заранее (такое преобразование называется *фильтрованием*). При этом сам выбранный интервал времени “скользит” вдоль ряда. Получаемый таким образом ряд скользящих средних ведет себя более гладко, чем исходный, из-за усреднения отклонений ряда. Для выравнивания временного ряда методом скользящей средней в Excel используется опция «*Линейная фильтрация*» инструмента «*Подбор линии тренда*».

Метод скользящей средней (как и другие механические методы выравнивания) пригоден лишь для осреднения значений ряда и не может быть использован для количественного прогнозирования. В то же время получение достаточно гладкой траектории дает возможность визуально оценить наличие тенденции в условиях сильной зашумленности ряда случайной компонентой, а также получить ответы на некоторые качественные вопросы относительно тренда.

В эконометрике основное внимание уделяется методу аналитического выравнивания, так как он дает количественную модель изменений временного ряда. При этом под *аналитическим выравниванием временного ряда* понимается построение аналитической функции для моделирования тренда.

Для построения трендов чаще всего применяются следующие регрессионные уравнения:

- $\tilde{y}_t = a + bt$ (линейный тренд);
- $\tilde{y}_t = a + b_1 t + b_2 t^2 + ... + b_m t^m$ (полиномиальный тренд);
- $\tilde{y}_t = a + \frac{b}{t}$ (гиперболический тренд);
- $\tilde{y}_t = e^{a+bt}$ (экспоненциальный тренд).

Параметры модели определяются с помощью метода наименьших квадратов. При этом в качестве независимой переменной выступает переменная t, принимающая значения 1, 2, …, n, а в качестве зависимой переменной – уровни $y_1, y_2, \ldots, y_n$ временного ряда. Для нелинейных трендов предварительно проводится процедура линеаризации.

На стадии спецификации модели, т.е. при выборе аппроксимирующей функции $f(t)$, обычно отталкиваются либо от качественного анализа процесса, исходя из соображений экономической теории, либо от визуального анализа графика зависимости уровней ряда от времени. В большинстве случаев выбор наилучшего уравнения тенденции осуществляется путем перебора многих форм $f(t)$ с последующим сравнением коэффициента детерминации R^2 для каждой из них. Предпочтение отдается той форме, для которой значение R^2 больше. Реализация такого подхода относительно проста при компьютерной обработке данных. В частности, для нахождения наиболее адекватного уравнения тренда в Excel используется инструмент *«Подбор линии тренда»* из *Мастера диаграмм*.

Трендовую кривую, аппроксимирующую временной ряд, называют еще *кривой роста*.

Наиболее часто в практической работе используются кривые роста, которые позволяют описывать процессы трех основных типов: без предела роста; с пределом роста без точки перегиба; с пределом роста и точкой

перегиба.

Для описания *процессов без предела роста* чаще всего служат функции: прямая (полином первой степени) $\tilde{y}_t = a + bt$, парабола (полином второй степени) $\tilde{y}_t = a + bt + ct^2$, экспонента $\tilde{y}_t = e^{a+bt}$ и другие. Процессы развития такого типа характерны в основном для абсолютных объемных показателей.

Для описания *процессов с пределом роста* применяется, как правило, модифицированная экспонента $\tilde{y}_t = k + a \cdot b^t$ и гипербола $\tilde{y}_t = a + \frac{b}{t}$. При этом в случае экспоненты прямая $y = k$ является горизонтальной асимптотой. Коэффициент k подбирается исходя из свойств прогнозируемого процесса или на основании экспертных оценок. Процессы с пределом роста характерны для многих относительных показателей (душевое потребление продуктов питания, внесение удобрений на единицу площади, затраты на одну денежную единицу произведенной продукции и т.п.).

Для описания процессов третьего типа (*с пределом роста и точкой перегиба*) используются логистическая кривая (кривая Перла-Рида) $\tilde{y}_t = \frac{k}{1 + be^{-at}}$ и кривая Гомперца $\tilde{y}_t = k \cdot a^{b^t}$. Такой тип развития характерен для растущих рынков, в частности, для описания развития спроса на некоторые новые товары. На основании кривой Гомперца также описывается динамика показателей уровня жизни; модификации этой кривой используются в демографии для моделирования показателей смертности и т.д.

Параметры большинства кривых роста, как правило, оцениваются по методу наименьших квадратов, т.е. подбираются таким образом, чтобы график кривой роста располагался на минимальном удалении от точек исходных данных.

Предпочтение, как правило, отдается простым моделям, допускающим содержательную интерпретацию. К числу таких моделей, в первую очередь, относится линейная модель роста $\tilde{y}_t = a + bt$.

6.5. Моделирование сезонных и циклических колебаний

Сезонность часто наблюдается во временных рядах, полученных на основе месячных или квартальных (иногда недельных или дневных) данных. Например, явно выраженную сезонность имеют объемы продаж мороженого и количество постояльцев курортного отеля, что связано с погодными условиями. Сезонные колебания присутствуют во многих экономических рядах. Потребление электроэнергии, газа, продажа определенных видов товаров, деловая активность предприятий – все эти ряды в той или иной степени подвержены эффекту сезонности.

Анализ сезонных колебаний облегчается тем обстоятельством, что после анализа автокорреляционной функции становится известным число сезонов в

одном периоде колебаний. Кроме того, сезонные колебания, как правило, выражены на графике временного ряда так ярко, что нет необходимости доказывать их существование.

Моделирование циклической компоненты $C(t)$ в целом аналогично моделированию сезонных колебаний, поэтому рассмотрим лишь моделирование последних.

Наиболее простой подход моделирования сезонной компоненты связан с расчетом ее значений. Он начинается с выбора формы модели (аддитивной или мультипликативной) на основе анализа графика временного ряда. Если амплитуда сезонных колебаний приблизительно постоянна, то выбирается аддитивная форма. В противном случае строят мультипликативную модель.

В качестве сезонной компоненты для аддитивной модели применяют абсолютное отклонение S_i, а для мультипликативной модели – индекс сезонности I_i.

Пусть y_{ij} – значение исходного временного ряда для i-ого сезона в j-ом периоде колебаний (i изменяется от 1 до l, где l – число сезонов в одном периоде; j изменяется от 1 до m, где m – число периодов во временном ряду; если временной ряд содержит целое число периодов и число уровней исходного ряда равно n, то $n = lm$).

Перед расчетом сезонных компонент производится аналитическое выравнивание временного ряда (можно использовать также механическое выравнивание, например, методом скользящей средней).

Если $y_{ij}^{\text{тр}}$ – теоретические значения оцененного тренда временного ряда, то абсолютное отклонение в i-ом сезоне определяются как среднее арифметическое из отклонений фактического и выровненного уровней ряда:

$$S_i = \frac{1}{m}\sum_{j=1}^{m}(y_{ij} - y_{ij}^{\text{тр}}).$$

Индекс сезонности в i-ом сезоне определяются как среднее арифметическое из отношений фактического уровня ряда к выровненному:

$$I_i = \frac{1}{m}\sum_{j=1}^{m}\frac{y_{ij}}{y_{ij}^{\text{тр}}}.$$

При необходимости полученные значения сезонной компоненты корректируются ($S_i^{\text{кор}} = S_i - \frac{1}{l}\sum_{i=1}^{l} S_i$ в случае аддитивной модели и

$I_i^{\text{кор}} = I_i \cdot \frac{l}{\sum_{i=1}^{l} I_i}$ в случае мультипликативной модели), чтобы суммарное воздействие сезонности на динамику было нейтральным: для аддитивной модели сумма сезонных компонент равна нулю, для мультипликативной модели – числу сезонов в одном периоде, т.е. l.

Для моделирования сезонных колебаний могут быть использованы фиктивные переменные. Количество фиктивных переменных должно быть на единицу меньше числа сезонов внутри периода колебаний, т.е. $l-1$. Например, при моделировании поквартальных данных модель должна содержать наряду с фактором времени три фиктивные переменные.

Каждому сезону соответствует определенное сочетание значений фиктивных переменных. Тот сезон, для которого значения всех фиктивных переменных равны нулю, принимается за эталон сравнения. Для остальных сезонов одна из фиктивных переменных должна принимать значение, равное единице. В частности, при моделировании поквартальных данных значения фиктивных переменных z_1, z_2, z_3 задаются таблицей 6.2.

Таблица 6.2. Значения фиктивных переменных в случае описания поквартальной сезонности

Квартал	z_1	z_2	z_3
1	0	0	0
2	1	0	0
3	0	1	0
4	0	0	1

Общий вид модели в таком случае будет следующим:

$$y_t = f(t) + c_1 z_1 + c_2 z_2 + c_3 z_3 + \varepsilon(t).$$

Эта модель, по сути, и является аналогом аддитивной модели временного ряда, так как фактический уровень временного ряда y_t является суммой тренда $f(t)$, сезонной компоненты $S(t) = c_1 z_1 + c_2 z_2 + c_3 z_3$ и случайной компоненты $\varepsilon(t)$.

Уравнение тренда для каждого квартала будет иметь вид:

$y_t = f(t) + \varepsilon(t)$ (первый квартал);

$y_t = f(t) + c_1 + \varepsilon(t)$ (второй квартал);

$y_t = f(t) + c_2 + \varepsilon(t)$ (третий квартал);

$y_t = f(t) + c_3 + \varepsilon(t)$ (четвертый квартал).

Основной недостаток модели с фиктивными переменными для описания сезонных и циклических эффектов – наличие большого количества переменных, что требует, соответственно, большого объема выборочных данных (не менее 7 на каждую переменную).

6.6. Общая схема моделирования временного ряда

Как отмечалось выше, одна из основных задач эконометрического исследования временного ряда – выявление всех его компонент. Поэтому, с учетом изложенного, общая схема построения аддитивной (мультипликативной) модели отдельного временного ряда включает следующие шаги:

1) аналитическое выравнивание уровней и расчет значений $T(t)$ с использованием полученного уравнения тренда;

2) устранение трендовой компоненты $T(t)$ из исходных уровней ряда и получение скорректированных данных для аддитивной $(S(t)+\varepsilon(t))$ или мультипликативной $(S(t)\cdot\varepsilon(t))$ модели;

3) расчет значений $S(t)$ сезонной компоненты ряда;

4) расчет полученных по модели значений $(T(t)+S(t))$ или $(T(t)\cdot S(t))$;

5) анализ случайной компоненты $\varepsilon(t)=y_t-T(t)-S(t)$ (в случае аддитивной модели) и $\varepsilon(t)=\dfrac{y_t}{T(t)\cdot S(t)}$ (в случае мультипликативной модели);

6) оценка качества модели.

Отметим, что предложенная схема построения модели временного ряда достаточна условна. Возможны и другие подходы. Один из них, например, предполагает сначала вычисление сезонных значений для каждого уровня модели временного ряда, а затем оценку тренда и анализ случайной компоненты.

6.7. Анализ случайной компоненты временного ряда

После выделения тренда и периодических составляющих (сезонной и циклической) проводится исследование случайной компоненты. Это исследование осуществляется с целью решения двух основных задач:

1) для оценки правильности выбора трендовой и сезонной компонент модели;

2) для оценки стационарности случайного процесса.

По сути, речь идет о проверке для случайной компоненты выполнения модельных предпосылок (условий теоремы Гаусса-Маркова). Временной ряд остатков должен обладать свойствами:

1) случайности (изменение величины ε не связано с изменением t);

2) соответствия нормальному закону распределения с нулевым математическим ожиданием;

3) независимости значений между собой (отсутствие автокорреляции в остатках).

Для этого в случае аддитивной модели рассматривается последовательность остатков

$$\varepsilon_1 = y_1 - T(1) - S(1) - C(1),$$
$$\varepsilon_2 = y_2 - T(2) - S(2) - C(2),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$\varepsilon_n = y_n - T(n) - S(n) - C(n)$$

как временной ряд и строится график зависимости остатков от времени.

Если во временном ряду тренд и сезонная составляющая выделены качественно, то остаток такого ряда после удаления тренда и сезонных колебаний является «*белым шумом*» (т.е. остатки распределены по нормальному закону, в остатках ряда отсутствуют тренд, систематические изменения дисперсии, систематически изменяющиеся взаимосвязи между элементами ряда и, кроме того, математическое ожидание ряда остатков равно нулю). Если же остаток содержит тренд, скрытые колебания или имеет место автокорреляция, то это говорит о необходимости уточнения спецификации модели и нового (более точного) выделения в ней тренда и колебаний.

При проверке свойства случайности полезно провести графический анализ остаточной последовательности и проверку наличия в ней тенденции. При этом возможны следующие различные ситуации, которые отражены на рисунках 6.6-6.8.

Наиболее простой способ проверки случайности уровней сводится к расчету коэффициента корреляции между остатками и фактором времени t и проверке его значимости. Кроме того, могут быть использованы *критерий серий* и *критерий поворотных точек*.

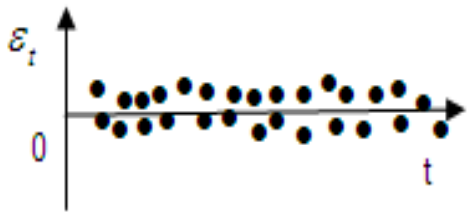


Рис. 6.6. Случайные остатки

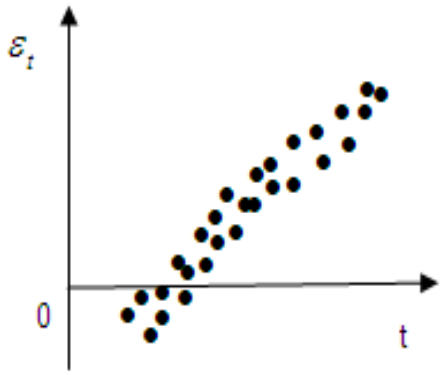


Рис. 6.7. Тенденция в остатках

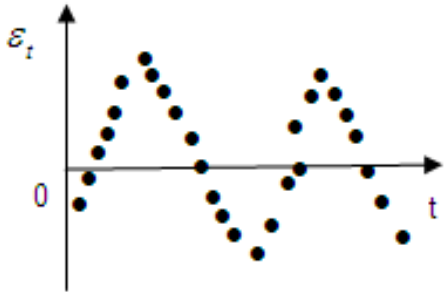


Рис. 6.8. Циклические колебания в остатках

Соответствие нормальному закону распределения можно проверить с помощью критерия Пирсона или более простыми средствами, например, с помощью теста Жарка-Бера оценки показателей асимметрии и эксцесса. Если они близки к нулю, то это дает основания считать ряд остатков нормально распределенным.

Проверка равенства нулю математического ожидания уровней ряда остатков осуществляется в ходе проверки соответствующей нулевой гипотезы с помощью статистики Стьюдента.

Проверка некоррелированности ряда остатков осуществляется с помощью критерия Дарбина-Уотсона.

6.8. Анализ структурной стабильности тенденции

От сезонных и циклических колебаний, оказывающих систематическое влияние на характер тенденции, отличаются *структурные сдвиги* – аномальные движения временного ряда, которые связаны с редко происходящими экономическими событиями, имеют скачкообразный характер и меняют тенденцию. К событиям, вызывающим структурные сдвиги, можно отнести, например, начало нового экономического курса, топливно-энергетические и финансовые кризисы, смену экономической ситуации в исследуемой области, а также другие факторы глобального характера.

В результате структурного сдвига, начиная с некоторого момента времени *t*, происходит изменение в развитии самого изучаемого показателя, а следовательно, изменяется поведение тренда, описывающего его развитие, например, на смену линейному характеру развития приходит кусочно-линейный.

Для моделирования ситуаций с изменением поведения тенденции, существуют специальные математические приемы. Один из них связан с применением теста Чоу, описанного в главе 4.

Суть теста Чоу для анализа структурной стабильности тенденции заключается в следующем:

1) полный временной ряд y_t длины *n* разбивается на две приблизительно равные части длины n_1 и n_2 $(n = n_1 + n_2)$;

2) для полного временного ряда и его частей оцениваются параметры линейных уравнений тренда: $\tilde{y}_t = a + bt$, $\tilde{y}_t = a_1 + b_1 t$, $\tilde{y}_t = a_2 + b_2 t$;

3) выдвигается и проверяется с помощью *F*-статистики гипотеза о структурной стабильности тенденции, а именно гипотеза $H_0 : a_1 = a_2, b_1 = b_2$.

Наблюдаемое значение статистики $F_{\text{набл}}$ вычисляется по выборочным данным на основании формулы $F_{\text{набл}} = \frac{S_0 - S_1 - S_2}{S_1 + S_2} \cdot \frac{n-4}{2}$, где S_0 – сумма квадратов отклонений уровней полного временного ряда y_t от соответствующих значений $\tilde{y}_t$, рассчитанных по уравнению регрессии $\tilde{y}_t = a + bt$, а S_1 и S_2 – суммы квадратов отклонений уровней частей временного ряда y_t от соответствующих теоретических значений, рассчитанных по уравнениям $\tilde{y}_t = a_1 + b_1 t$ и $\tilde{y}_t = a_2 + b_2 t$.

Рассматриваемая *F*-статистика имеет распределение Фишера с числами степеней свободы $k_1 = 2$ и $k_2 = n - 4$. Если $F_{\text{кр}} < F_{\text{набл}}$, то гипотеза H_0 отклоняется. В этом случае моделирование тренда следует осуществлять с помощью кусочно-линейной модели. Если же $F_{\text{кр}} > F_{\text{набл}}$, то нет оснований отклонять нулевую гипотезу, а значит, моделирование тренда следует осуществлять с помощью единого для всего временного ряда уравнения $\tilde{y}_t = a + bt$.

Графическая интерпретация выводов тестирования заключается в следующем. Если $F_{\text{кр}} > F_{\text{набл}}$, то по заданному корреляционному полю единая прямая хорошо моделирует ситуацию (рис. 6.9). Если же $F_{\text{кр}} < F_{\text{набл}}$, то по тому же корреляционному полю (рис. 6.10) следует отдать предпочтение моделированию с помощью некоторой ломаной линии.

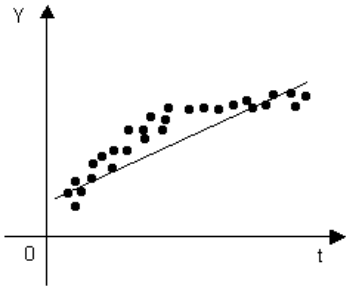


Рис. 6.9. Модель, представленная единой прямой

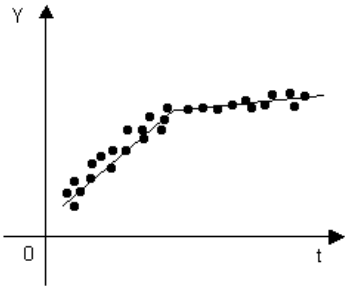


Рис. 6.10. Модель, представленная ломаной линией

6.9. Прогнозирование на основе модели временного ряда

Прогнозирование на основе модели временного ряда основано на идее *экстраполяции*, т.е. на предположении о том, что закономерность сложившейся связи между заданными уровнями временного ряда сохранится в будущем. По модели временного ряда строятся точечный и интервальный прогнозы. Обычно для точечного прогнозирования выбирается период, равный $n+1$ или $n+2$. Прогнозировать на период $t = n + p$, где $p > 2$, не рекомендуется из-за увеличивающейся расплывчатости прогноза.

Точечный прогноз y_p на основе трендовых моделей осуществляется подстановкой в уравнение тренда номера $t = n + p$ прогнозируемого периода. Точечный прогноз в моделях с сезонной компонентой получают следующим образом: находят прогнозное значение по трендовой модели и прибавляют к нему соответствующее значение сезонной компоненты в случае аддитивной модели или умножают его на значение сезонной компоненты в случае мультипликативной модели.

Конечно, совпадение фактических данных и прогнозных точечных оценок, полученных путем экстраполяции кривых, характеризующих тенденцию, имеет малую вероятность. Возникновение соответствующих отклонений объясняется следующими причинами:

1) Выбранная для прогнозирования кривая не является единственно возможной для описания тенденции. Можно подобрать такую кривую, которая дает более точные результаты.

2) Прогноз осуществляется на основании ограниченного числа исходных данных. Кроме того, каждый исходный уровень обладает еще и случайной компонентой. Поэтому и кривая, по которой осуществляется экстраполяция, также будет содержать случайную компоненту.

3) Тенденция характеризует изменение среднего уровня временного ряда, поэтому отдельные наблюдения могут от него отклоняться. Если такие отклонения наблюдались в прошлом, то они будут наблюдаться и в будущем.

Интервальный прогноз строится на основе точечного прогноза y_p. *Доверительным интервалом* называется такой интервал, относительно которого можно с заранее выбранной вероятностью утверждать, что он содержит значение прогнозируемого показателя. Ширина интервала зависит от качества модели, т.е. степени ее близости к фактическим данным, числа наблюдений, периода прогнозирования и выбранного пользователем уровня вероятности.

Для интервального прогноза на период $t = n + p$ по линейной трендовой модели $\tilde{y}_t = a + bt$ предварительно рассчитывается стандартная ошибка прогноза:

$$m_p = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(n + 2p - 1)^2}{\sum_{i=1}^{n} (2i - n - 1)^2}}, \qquad (6.2)$$

где $s = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}{n - 2}}$.

Затем строится доверительный интервал прогноза

$$(y_p - t_{\text{кр}} \cdot m_p; y_p + t_{\text{кр}} \cdot m_p),$$

т.е. определяются нижняя и верхняя границы интервала прогноза.

Если построенная модель адекватна, то с выбранной вероятностью можно утверждать, что при сохранении сложившихся закономерностей развития прогнозируемая величина попадает в интервал, образованный верхней и нижней границей.

6.10. Обзор некоторых вопросов и проблем моделирования временных рядов

Классификация кривых роста и методика их подбора, исходя из специфики экономических процессов, излагается в [5]. В [12,18] обсуждаются методы выбора вида кривой роста, основанные на анализе таких статистических показателей, как абсолютные приросты, темпы роста и темпы прироста.

В классе моделей временных рядов особое место занимают *динамические модели*, т.е. модели, которые в данный момент времени *t* учитывают значения входящих в них переменных как в текущий, так и в предыдущие моменты времени, а также само время *t*.

Выделяют два основных типа динамических эконометрических моделей:

1) модели с распределенным лагом;
2) модели авторегрессии.

Моделью с распределенным лагом называется модель, которая содержит в качестве объясняющих переменных текущие и лаговые (т.е. взятые с некоторым запаздыванием) значения лишь факторных переменных. Примером такой модели является модель

$$y_t = a + a_0 x_t + a_1 x_{t-1} + ... + a_m x_{t-m} + \varepsilon \qquad (6.3)$$

Авторегрессионной моделью называется модель, которая в качестве объясняющих переменных содержит как текущие и лаговые значения факторных переменных, так и лаговые значения зависимой переменной. Примером такой модели является модель

$$y_t = a + bx_t + cy_{t-1} + \varepsilon \qquad (6.4)$$

Если модель (6.4) применяется для описания зависимости потребления человека от его дохода, то она учитывает доход человека в данный момент и его потребление в предыдущий период. В отличие от (6.4), например, модель

$$y_t = a + bx_t + cx_{t-1} + \varepsilon,$$

описывая потребление человека, учитывает его доходы в период времени t и в предыдущий период $t-1$, но не учитывает расходы на потребление в предыдущий период.

Динамические модели обладают определенной спецификой, которая проявляется в следующем.

1. Параметризация динамических моделей требует в большинстве случаев специальных статистических методов, так как обычный МНК не применим ввиду нарушений условий Гаусса-Маркова.

2. Спецификация динамической модели является более сложной. Исследователю необходимо не только выбрать факторные переменные, но и определиться с «глубиной» их лагирования (т.е. необходимо решить, сколько предшествующих периодов следует учитывать в модели).

Указанная специфика динамических моделей требует особых методов оценки параметров регрессии. Некоторые из таких методов (метод Койка, метод Алмон, метод инструментальных переменных) изложены в [2-4].

Весьма сложной является эконометрическая задача, учитывающая фактор «ожидания». Соответствующие динамические модели получили название *моделей адаптивного ожидания*. В них зависимая переменная y_t связана не с текущим или предыдущим значением x_t и x_{t-1} объясняющей переменной *x,* а с ожидаемым значением ее в (t+1)-м периоде.

Модель имеет вид $y_t = a + bx^*_{t+1} + \varepsilon_t$, где x^*_{t+1} – ожидаемое значение ненаблюдаемой объясняющей переменной.

Модель, учитывающая фактор «ожидания», возникает, например, в случае, когда производитель принимает решение об объеме производимой в период t продукции y_t до того, как станет известной цена x_{t+1} на эту продукцию в следующем периоде. Поскольку цена x_{t+1} неизвестна в период $t+1$, то решение принимается на основе ожидаемого значения x^*_{t+1}.

Некоторые аспекты построения и анализа моделей адаптивного ожидания изложены в [2,4].

Если имеются данные о двух временных рядах x_t и y_t, то часто возникает вопрос: а есть ли зависимость между уровнями этих временных рядов (в частности, имеется ли линейная зависимость $y_t = a + bx_t + \varepsilon_t$). При ответе на этот вопрос обычные корреляционные подходы не работают, так как наличие тенденций в рядах x_t и y_t может привести к ложной корреляции и искажению показателей тесноты связи. Чтобы получить объективную картину нужно исключить тенденцию из каждого ряда. Это достигается различными методами. Наиболее распространены среди них метод отклонений от тренда, метод последовательных разностей и метод включения фактора времени. Описание этих методов можно найти в [2].

Один из подходов установления причинности зависимости между временными рядами связан с методом коинтеграции. Он основан на анализе графиков временных рядов x_t и y_t и их тенденций. Если на протяжении длинного промежутка времени имеет место устойчивая одинаковая (или противоположная) направленность тенденций, то естественно предположить, что коэффициент корреляции между уровнями рядов x_t и y_t характеризует не ложную, а истинную причинно-следственную связь.

Такое предположение было положено в основу теории коинтеграции временных рядов. *Коинтеграция* – это причинно-следственная зависимость в уровнях двух временных рядов, которая выражается в совпадении или противоположной направленности их тенденций и случайной колеблемости.

Один из тестов на проверку коинтеграции временных рядов предложили Р. Энгл и К. Гренджер. За этот тест в 2003 году им была присуждена Нобелевская премия по экономике.

Проверка нулевой гипотезы об отсутствии коинтеграции временных рядов x_t и y_t методом Энгла-Гренджера осуществляется по следующей схеме:

1) с помощью обычного МНК рассчитывается регрессия $\tilde{y}_t = a + bt$ и вычисляются остатки $\varepsilon_t = y_t - \tilde{y}_t$;

2) рассчитывается уравнение линейной регрессии $\Delta\varepsilon_t = A + B\varepsilon_{t-1}$, где $\Delta\varepsilon_t = \varepsilon_t - \varepsilon_{t-1}$ – первые разности остатков;

3) определяется наблюдаемое значение t-статистики Стьюдента для коэффициента регрессии B; если $t_{\text{набл}}$ для коэффициента B больше критического, то нулевая гипотеза об отсутствии коинтеграции отклоняется и принимается гипотеза о том, что между рядами x_t и y_t имеется коинтеграция.

Критические значения теста Ингла-Гренджера составляют 2,5899 (для уровня значимости 1%) и 1,9439 (для уровня значимости 5%).

Примеры решения типовых заданий

Пример 6.1. В таблице 6.3 приведены данные об уровнях временного ряда. Проверить с помощью метода Ирвина наличие аномальных уровней.

Таблица 6.3. Данные примера 6.1

t	1	2	3	4	5	6	7	8	9	10
y_t	1,6	1,9	2,1	2,4	4,5	2,8	3,1	3,3	3,6	3,8

Решение:

На рисунке 6.11 приведен точечный график временного ряда.

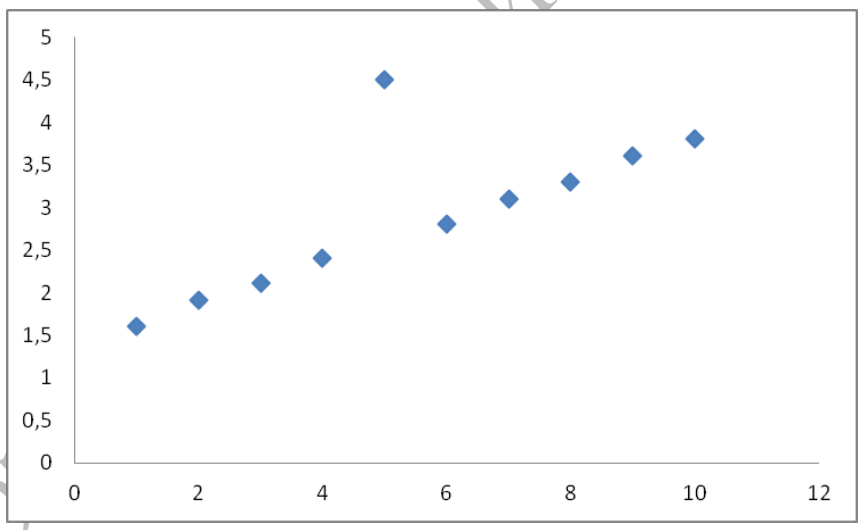


Рис. 6.11. График временного ряда

Из графика временного ряда можно сделать вывод, что пятому наблюдению временного ряда соответствует резкий выброс. Исследуем точку $t = 5$ на аномальное значение.

Для нахождения наблюдаемого значения критерия Ирвина воспользуемся формулой $\lambda_5 = \dfrac{|y_5 - y_4|}{\sqrt{\frac{1}{9}\sum_{i=1}^{10}(y_i - \bar{y})^2}}$.

Так как $\lambda_5 = 2{,}28$ и $\lambda_0 = 1{,}5$, то $\lambda_5 > \lambda_0$, а значит, уровень $t = 5$ является аномальным. Если аномалия уровня вызвана ошибками первого рода, то пятый уровень можно заменить на среднее арифметическое четвертого и шестого уровней, т.е. на значение $y_5 = (2{,}4 + 2{,}8)/2 = 2{,}6$.

Пример 6.2. В таблице 6.4 приведены поквартальные данные об объемах производства некоторого предприятия. С помощью анализа автокорреляционной функции и графика временного ряда определить структуру временного ряда.

Таблица 6.4. Данные примера 6.2

Год	Квартал	Период	Объем производства, млрд. руб., y_t
2007	1	1	410
	2	2	400
	3	3	715
	4	4	600
2008	1	5	585
	2	6	560
	3	7	975
	4	8	800
2009	1	9	765
	2	10	720
	3	11	1235
	4	12	1100

Решение:

Вычислим коэффициенты автокорреляции первого, второго, третьего, четвертого и пятого порядков.

Для вычисления коэффициента автокорреляции первого порядка по данным таблицы 6.5 найдем корреляцию между рядами y_t, где $t \in \{1,2,...,11\}$, и y_{t+1}, где $t \in \{1,2,...,11\}$.

Тогда $r_1 = 0{,}538$.

Таблица 6.5. Данные для расчета коэффициента автокорреляции первого порядка

y_t	y_{t+1}
410	400
400	715
715	600

600	585
585	560
560	975
975	800
800	765
765	720
720	1235
1235	1100

Для вычисления коэффициента автокорреляции второго порядка по данным таблицы 6.6 найдем корреляцию между рядами y_t, где $t \in \{1,2,...,10\}$, и y_{t+2}, где $t \in \{1,2,...,10\}$.

Таблица 6.6. Данные для расчета коэффициента автокорреляции второго порядка

y_t	y_{t+2}
410	715
400	600
715	585
600	560
585	975
560	800
975	765
800	720
765	1235
720	1100

Тогда $r_2 = 0{,}286$.

Аналогично рассчитываются коэффициенты автокорреляции третьего, четвертого и пятого порядков: $r_3 = 0{,}432$, $r_4 = 0{,}992$, $r_5 = 0{,}373$.

Так как из последовательности коэффициентов автокорреляции r_1, r_2, r_3, r_4, r_5 самым высоким оказался коэффициент $r_4 = 0{,}992$, то можно сделать вывод о том, что исследуемый временной ряд содержит периодические (сезонные) колебания с периодом, равным 4. Кроме того, из вида графика временного ряда (рисунок 6.12) можно сделать вывод о наличии тренда.

Рис. 6.12. График временного ряда

Пример 6.3. С помощью метода сравнения средних уровней определить наличие тренда во временном ряду y_t, заданном таблицей 6.7.

Таблица 6.7. Данные примера 6.3

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

Решение:

Из вида корреляционного поля (рисунок 6.13) трудно сделать однозначный вывод о наличии или отсутствии тренда в заданном временном ряду.

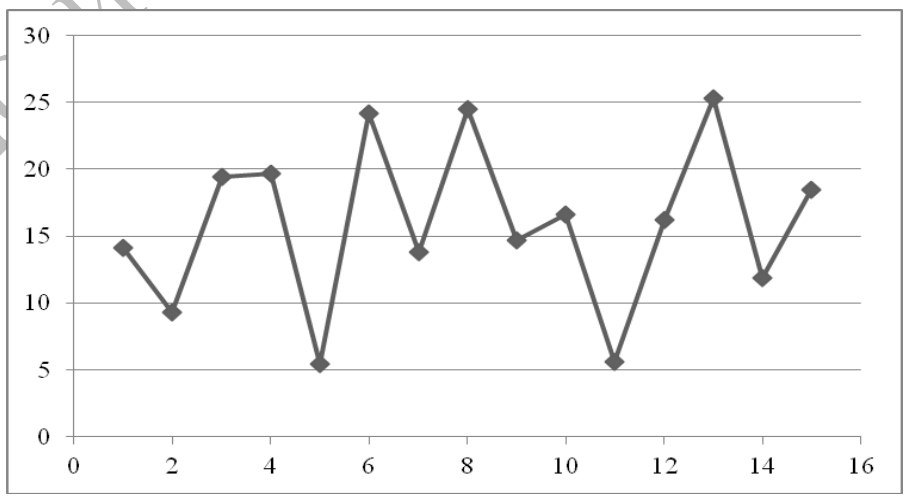


Рис. 6.13. График временного ряда

Для определения тренда применим метод сравнения средних уровней.

Разделим временной ряд на две части длины 7 и 8 соответственно. Оценим среднее и дисперсию первой и второй частей временного ряда: $\bar{y}_1 = 15{,}13$, $\bar{y}_2 = 16{,}66$, $S_1^2 = 42{,}15$, $S_2^2 = 41{,}22$.

Проверяем гипотезу о равенстве (однородности) дисперсий обеих частей ряда с помощью *F*-критерия Фишера. Для вычисления наблюдаемого значения *F*-критерия большую дисперсию разделим на меньшую: $F_{\text{набл}} = \frac{S_1^2}{S_2^2} = \frac{42{,}15}{41{,}22} = 1{,}022$. Табличное значение $F_{\text{кр}}$ статистики Фишера при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k_1 = 7-1 = 6$, $k_2 = 8-1 = 7$ составляет 3,87.

Так как $F_{\text{набл}} < F_{\text{кр}}$, то с вероятностью 95% нет оснований отвергать нулевую гипотезу. По данным наблюдений выборочные дисперсии различаются незначимо (расхождение между ними случайно). Значит, выделенные части временного ряда можно считать однородными.

Проверим основную гипотезу о равенстве средних значений с использованием t-критерия Стьюдента:

$$t_{\text{набл}} = \frac{|\bar{y}_1 - \bar{y}_2|}{\sqrt{S_1^2 \cdot (7-1) + S_2^2 \cdot (8-1)}} \cdot \sqrt{\frac{7 \cdot 8 \cdot (15-2)}{15}} = 0{,}459,\ t_{\text{кр}} = 2{,}16.$$

Так как $t_{\text{набл}} \le t_{\text{кр}}$, то нет оснований отвергать нулевую гипотезу о равенстве средних, а значит, расхождение между вычисленными средними незначимо. Отсюда следует, что тренд во временном ряду y_t отсутствует.

Пример 6.4. В таблице 6.8 отражены данные некоторой фирмы об объемах продаж товара. Подобрать линию тренда, которая лучше всего описывает фактические данные и на ее основе сделать прогноз на 3 недели вперед.

Таблица 6.8. Данные примера 6.4

Неделя	1	2	3	4	5	6	7	8	9	10	11
Количество проданных единиц	17	22	26	27	35	40	41	45	50	63	78

Решение:

1. *Ввод исходных данных задачи.* В ячейки A1 и B1 введем заголовки исходных данных, в ячейки A2:A12 – номера недель, а в ячейки B2:B12 – соответствующее количество продаж.

2. *Построение графика фактических значений показателя.* Выделим

ячейки B1:B12 (исходные данные вместе с заголовком) и нажмем кнопку *Мастер диаграмм* на панели инструментов. Построим с его помощью диаграмму типа *График* (рисунок 6.14).

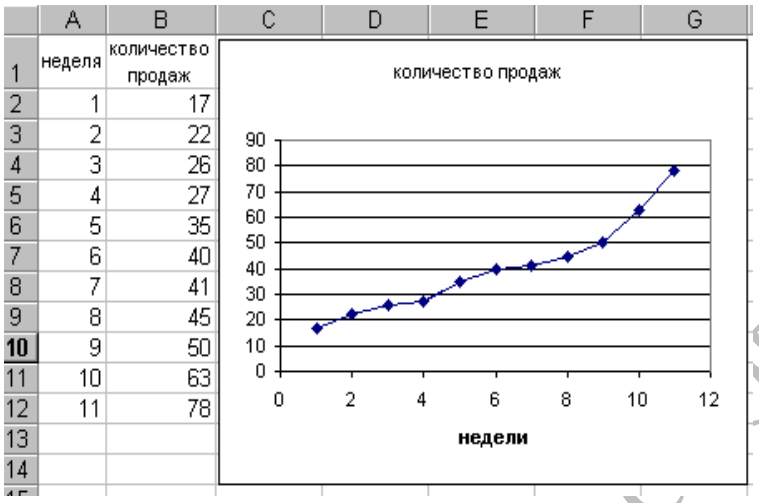


Рис. 6.14. График продаж

3. *Изображение на графике кривой роста линейной модели.* Выполним один щелчок по диаграмме для того, чтобы перейти в режим ее редактирования. Затем подведем курсор к какой-либо точке на графике и снова щелкнем левой кнопкой мыши. Нажмем правую кнопку мыши для вызова контекстного меню. В контекстном меню выберем команду *Добавить линию тренда* (рисунок 6.15).

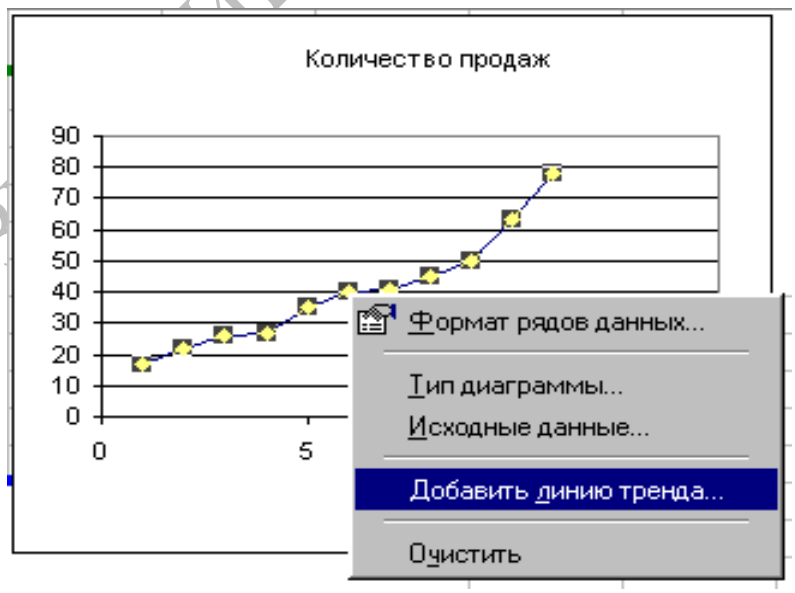


Рис. 6.15. Схема выбора меню *Добавить линию тренда*

На экране появляется окно *Линия тренда* (рисунок 6.16).

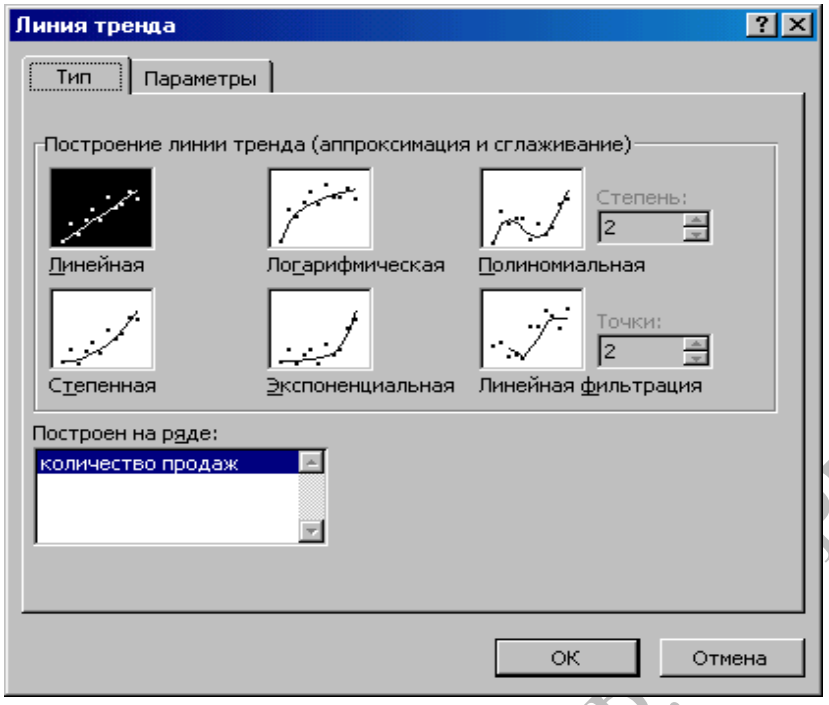


Рис. 6.16. Окно *Линия тренда*

В окне *Линия тренда* на вкладке *Тип* выберем *Линейная*, а на вкладке *Параметры* установим следующие флажки:

- показывать уравнение на диаграмме;
- поместить на диаграмму величину индекса детерминации (*R*^2).

После нажатия кнопки *OK* на графике наряду с фактическими значениями количества продаж будет показана линия тренда, уравнение и коэффициент детерминации (рисунок 6.17).

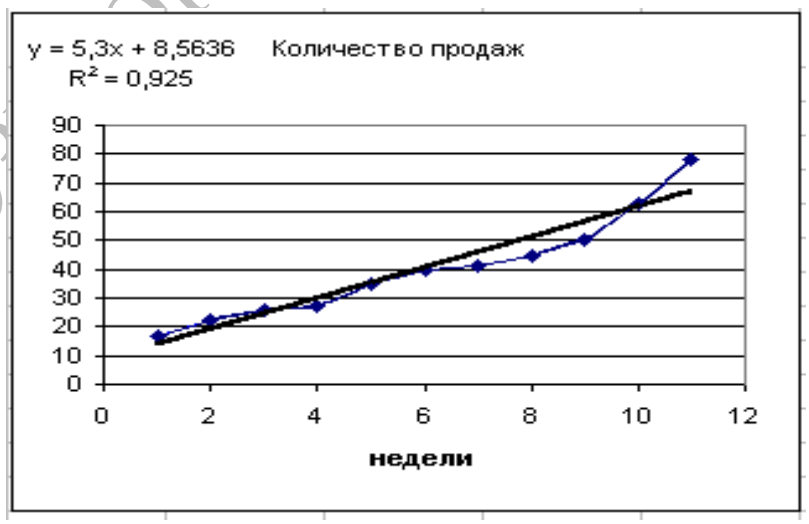


Рис. 6.17. График линейного тренда

4. *Подбор функции тренда, наиболее точно описывающей исходные*

данные. Аналогично следует испытать другие типы линий тренда. При добавлении каждой новой линии тренда на график нужно сравнить ее коэффициент детерминации с аналогичным показателем предыдущей модели. В результате перебора всех возможных (стандартных) линий тренда в данной задаче выбор останавливается на экспоненциальной модели, поскольку для нее коэффициент детерминации наибольший (рисунок 6.18).

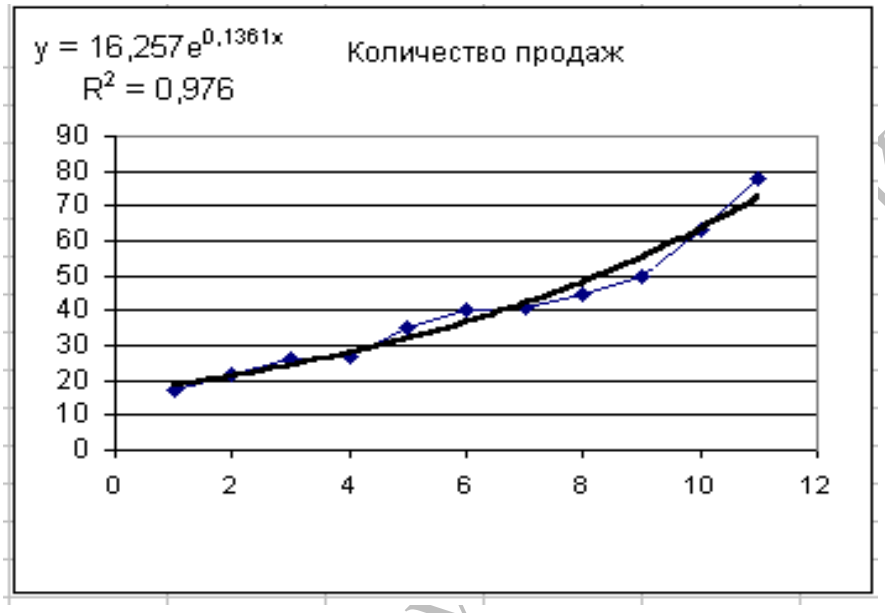


Рис. 6.18. График экспоненциального тренда

5. *Выполнение прогноза.* Поскольку нужно выполнить прогноз на 3 недели вперед, допишем номера этих недель (12, 13 и 14) в столбец А. В соответствующие ячейки в столбце В занесем формулы вычисления теоретического значения по функции тренда. Так как уравнение экспоненциальной кривой имеет вид $y = ae^{bt}$, то для расчетов необходимо использовать функцию EXP(), т.е. в ячейку В13 нужно записать формулу

=16,257*EXP(0,1361*A13).

Затем эту формулу можно скопировать в ячейки В14 и В15. В результате получим в ячейках В13:В15 следующие прогнозные значения:

- на 12-ю неделю – 83 продажи;
- на 13-ю неделю – 95 продаж;
- на 14-ю неделю – 109 продаж.

Пример 6.5. Имеются статистические данные (таблица 6.9) об объемах выпуска продукции y (млрд. руб.) на некотором предприятии за несколько лет.

Таблица 6.9. Данные примера 6.5

Год	2000	2001	2002	2003	2004	2005	2006	2007	2008
y	10	12	15	16	20	22	25	24	27

Проверить, имеется ли тенденция в изменении выпуска продукции. Выбрать тип тренда и рассчитать его параметры. Проверить качество построенной модели. Сделать точечный и интервальный прогнозы на 2009 год.

Решение:

Построим график выпуска продукции (рисунок 6.19). По виду диаграммы можно предположить, что во временном ряду присутствует линейный тренд.

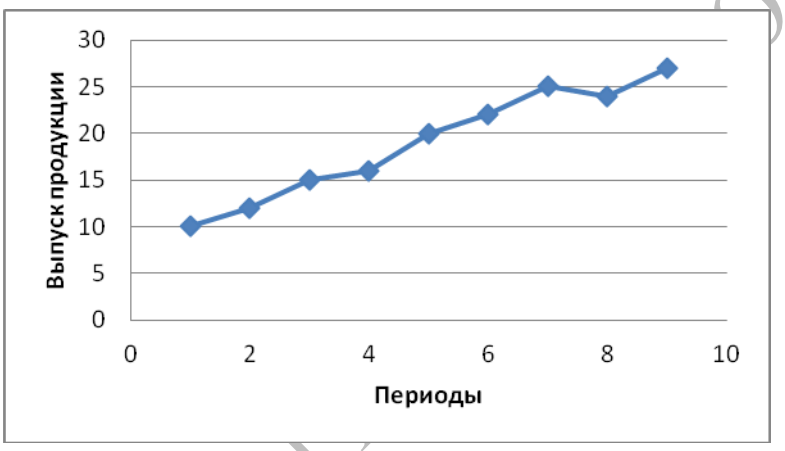


Рис. 6.19. График выпуска продукции

Конечно, выбор тренда можно осуществить и экспериментальным методом. При этом индекс детерминации у полиномиального тренда несколько выше, но выбор остановим на линейном тренде ввиду его простоты и ясной экономической интерпретации параметров регрессии.

Оценим параметры тренда методом наименьших квадратов. В качестве значений независимой переменной возьмем значения периодов.

Уравнение линейной регрессии имеет вид $\tilde{y} = 8{,}17 + 2{,}17 \cdot x$. Коэффициент при переменной x показывает, что за один год выпуск продукции на предприятии увеличивается в среднем на 2,17 млрд. руб.

Общее качество уравнения регрессии высокое ($R^2 = 0{,}971$).

Оценим статистическую значимость индекса детерминации. В нашем случае $F_{\text{набл}} = 236{,}6$, $F_{\text{кр}} = 5{,}59$. Так как $F_{\text{набл}} > F_{\text{кр}}$, то уравнение регрессии в целом статистически значимо.

Для коэффициентов регрессии модули наблюдаемых значений статистики Стьюдента (10,3 и 15,4) больше критического значения, равного 2,36. Значит, коэффициенты регрессии статистически значимы.

Точность модели оценим с помощью коэффициента аппроксимации $\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \tilde{y}_i}{y_i}\right| \cdot 100\,\%$. В нашем случае он составляет 4,34%, что говорит о высокой точности модели.

Так как среднее значение остатков (равное -9,86865E-16) является несмещенной оценкой математического ожидания, то можно считать (не прибегая к статистическим тестам), что математическое ожидание случайного члена равно нулю и первое условие теоремы Гаусса-Маркова.

Так как модель построена по данным временного ряда, она гомоскедастична.

Проверим модель на автокорреляцию. Для этого воспользуемся критерием Дарбина-Уотсона. Вычислим наблюдаемое значение статистики: $DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = 2{,}026$. Критические значения статистики имеют следующие значения: d_1 =0,824; d_2 =1,32. Так как наблюдаемое значение попадает в зону отсутствия автокорреляции, то соответствующее условие теоремы Гаусса-Маркова об отсутствии автокорреляции выполняется.

Исследование ряда остатков показывает адекватность и надежность построенной модели.

Выберем $t = 10$ (соответствует 2009-ому году) и подставим в уравнение тренда. Получим точечный прогноз, равный 29,8 (млрд. руб.).

Для интервального прогноза по модели предварительно рассчитаем стандартную ошибку прогноза:

$$m_p = s \cdot \sqrt{1 + \frac{1}{n} + \frac{(n + 2p - 1)^2}{\sum_{i=1}^{n}(2i - n - 1)^2}} = 1{,}091 \cdot \sqrt{1 + \frac{1}{9} + \frac{100}{\sum_{i=1}^{9}(2i - 10)^2}} = 1{,}348,$$

где $s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2}} = 1{,}091$ и $p = 1$, $n = 9$.

Затем по значениям $t_{\text{кр}} = 2{,}36$, $y_p = 29{,}8$ и $m_p = 1{,}348$ строится доверительный интервал прогноза $(y_p - t_{\text{кр}} \cdot m_p; y_p + t_{\text{кр}} \cdot m_p)$, т.е. определяются нижняя и верхняя границы интервала прогноза. В нашем случае эти границы имеют значения 26,62 и 32,98.

Пример 6.6. По статистическим данным РУП «Белоруснефть-Минскоблнефтепродукт» **(**таблица 6.10) о реализации дизельного топлива

организациям и предприятиям Минска и Минской области в 2004-2010 годах:

1) построить график реализации дизельного топлива;
2) провести анализ структуры временного ряда;
3) смоделировать тренд временного ряда;
4) рассчитать значения сезонной компоненты;
5) построить график модели временного ряда;
6) оценить точность модели с помощью ошибки аппроксимации.

Таблица 6.10. Данные примера 6.6

	2004 год	2005 год	2006 год	2007 год	2008 год	2009 год	2010 год
январь	5 656,5	5 322,6	7 555,6	11 283,5	15 369,9	15 232,7	16 122,2
февраль	4 575,1	5 963,0	7 700,9	8 630,4	13 775,8	15 256,1	18 212,8
март	10 524,8	10 501,8	10 368,0	8 677,2	20 710,9	16 086,6	21 044,2
апрель	24 638,2	19 879,1	13 724,0	21 442,1	29 942,0	31 164,5	34 991,5
май	8 948,6	12 874,2	13 110,3	15 089,4	24 888,2	23 869,4	28 488,6
июнь	11 098,6	14 624,8	15 369,7	18 561,8	27 098,9	21 392,0	29 021,1
июль	11 717,2	14 250,0	21 332,0	28 819,5	29 186,0	26 125,8	35 921,0
август	25 471,9	26 119,8	25 174,5	23 840,2	41 018,0	40 182,8	36 633,5
сентябрь	15 656,3	20 347,7	16 643,6	24 188,4	27 176,0	26 594,7	33 083,1
октябрь	11 659,1	15 497,3	14 487,9	21 821,1	28 725,7	26 129,8	31 881,8
ноябрь	9 440,5	10 225,2	14 739,6	16 209,7	19 275,5	18 995,2	27 850,2
декабрь	6 120,0	8 471,3	13 303,0	15 871,4	18 229,9	18 567,1	23 300,6

Решение:

1) По данным, представленным в таблице 6.10, построим график (рисунок 6.20).

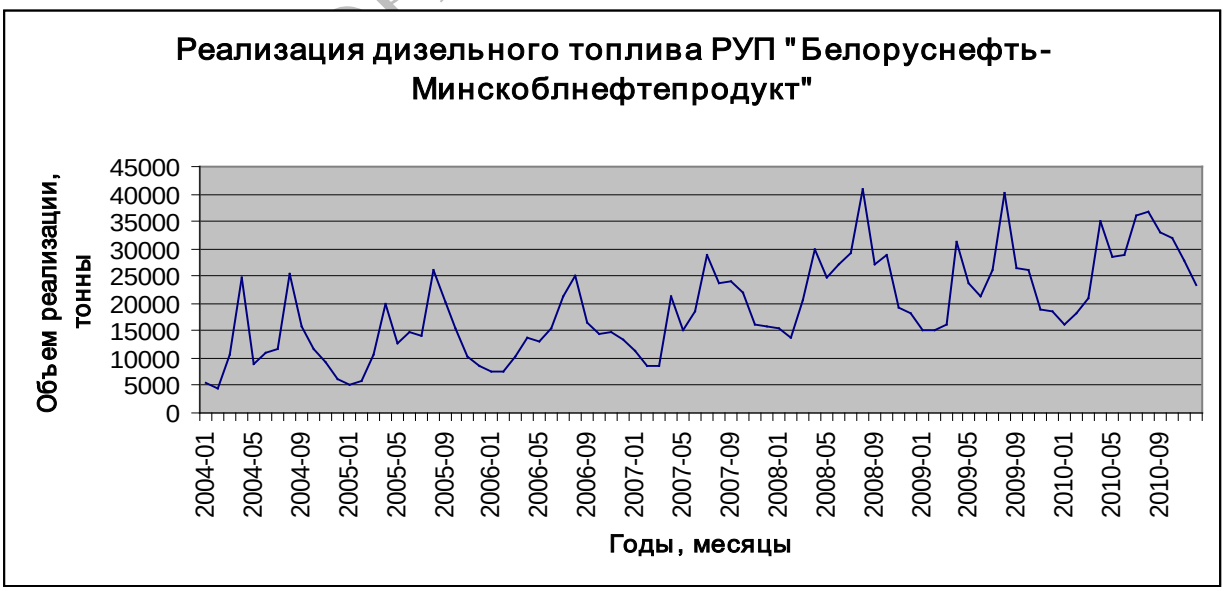


Рис. 6.20. График объемов реализации дизельного топлива

Визуальный анализ графика позволяет предположить, что в структуре временного ряда присутствует тренд и периодическая компонента.

2) Гипотеза о наличии во временном ряду тренда и цикличности подтверждается анализом автокорреляционной функции (таблица 6.11) и видом коррелограммы (рисунок 6.21).

Таблица 6.11. Значения коэффициентов автокорреляции

Величина лага k	Значение r_k	Величина лага k	Значение r_k
1	0,71702083	13	0,651380855
2	0,54776565	14	0,46330488
3	0,408553694	15	0,308968907
4	0,374505412	16	0,249877841
5	0,16680431	17	0,021476034
6	0,098540112	18	-0,041952295
7	0,083549887	19	-0,022349169
8	0,278848046	20	0,180584117
9	0,346140315	21	0,314749942
10	0,48576767	22	0,466136492
11	0,664128668	23	0,63515393
12	0,863952903	24	0,797762771

Так как самыми высокими являются двенадцатый и двадцать четвертый коэффициенты автокорреляции, то временной ряд содержит сезонную компоненту с периодом 12.

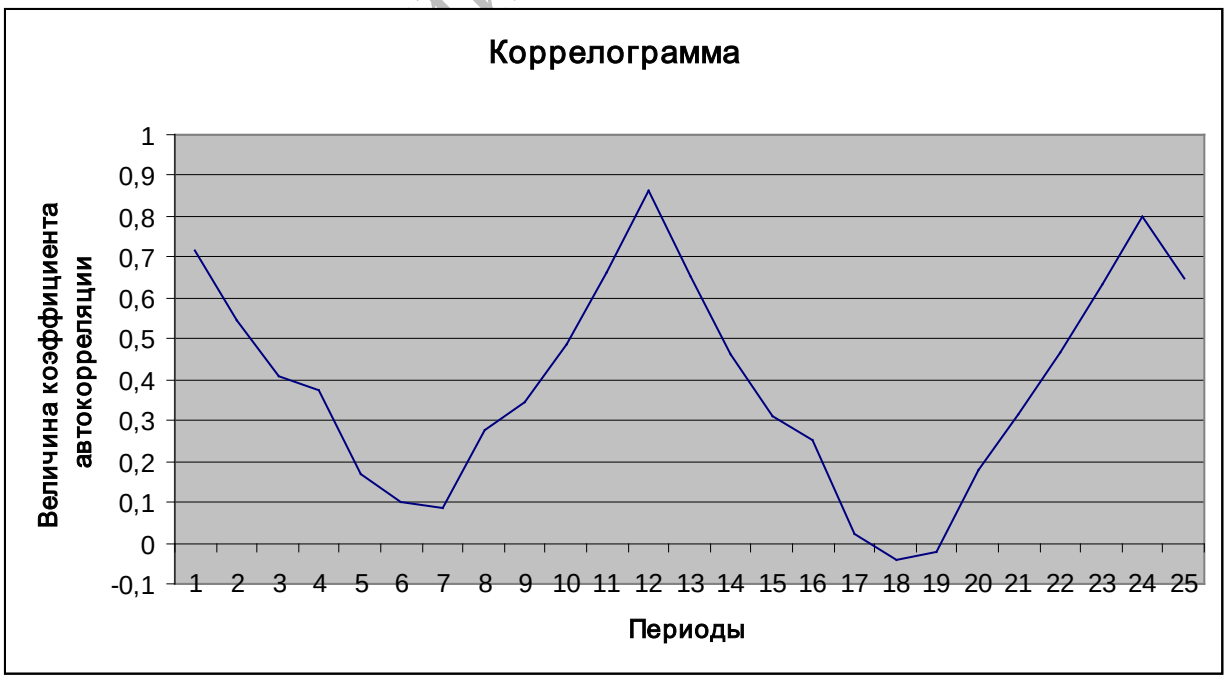


Рис. 6.21. Коррелограмма

3) Для моделирования тренда временного ряда воспользуемся экспериментальным методом. Для этого построим и сравним между собой несколько трендов, полученных на основе аналитического выравнивания (рисунки 6.22-6.26).

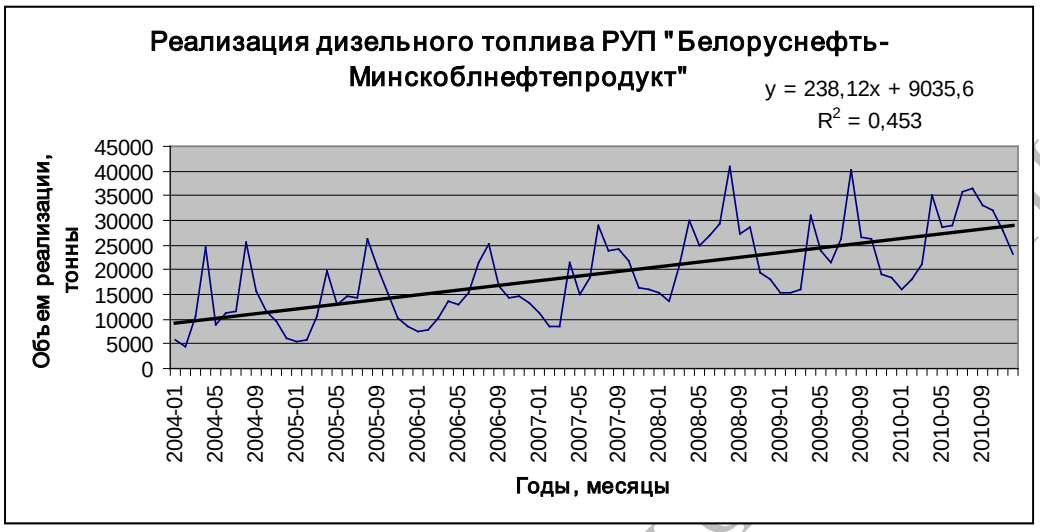


Рис. 6.22. График линейного тренда

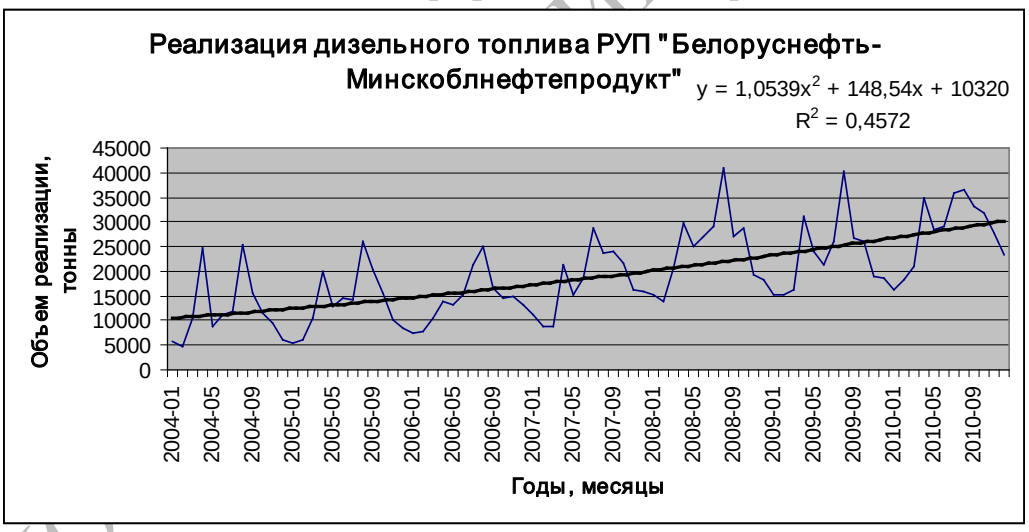


Рис. 6.23. График полиномиального тренда

Сравнивая индексы детерминации для линейной ($R^2 = 0{,}453$), полиномиальной ($R^2 = 0{,}4572$), степенной ($R^2 = 0{,}4211$), логарифмической ($R^2 = 0{,}3451$) и экспоненциальной ($R^2 = 0{,}4704$) функций, выбираем линейный тренд $\tilde{y} = 238{,}12t + 9035{,}6$ (ввиду его простоты и четкой экономической интерпретации коэффициентов регрессии). Из уравнения линейного тренда следует, что в среднем объем реализации дизельного топлива ежемесячно увеличивается на 238,12 тонны.

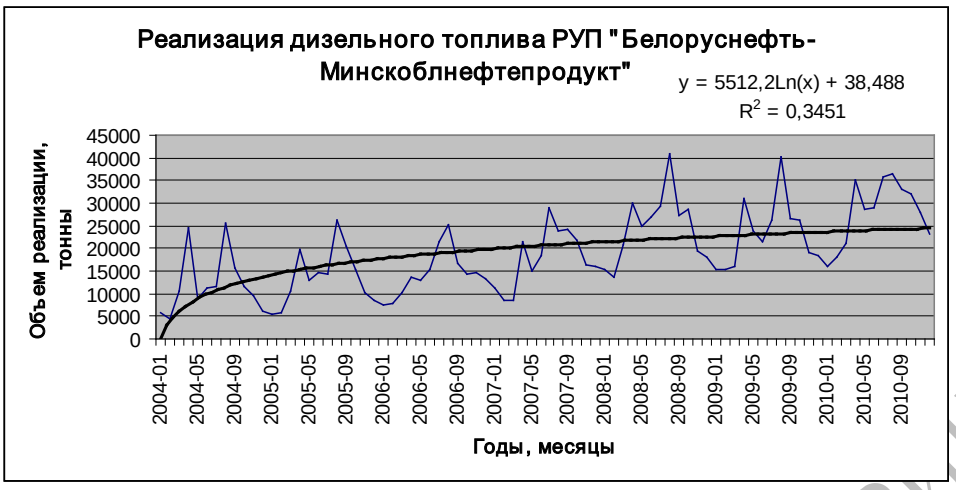


Рис. 6.24. График логарифмического тренда

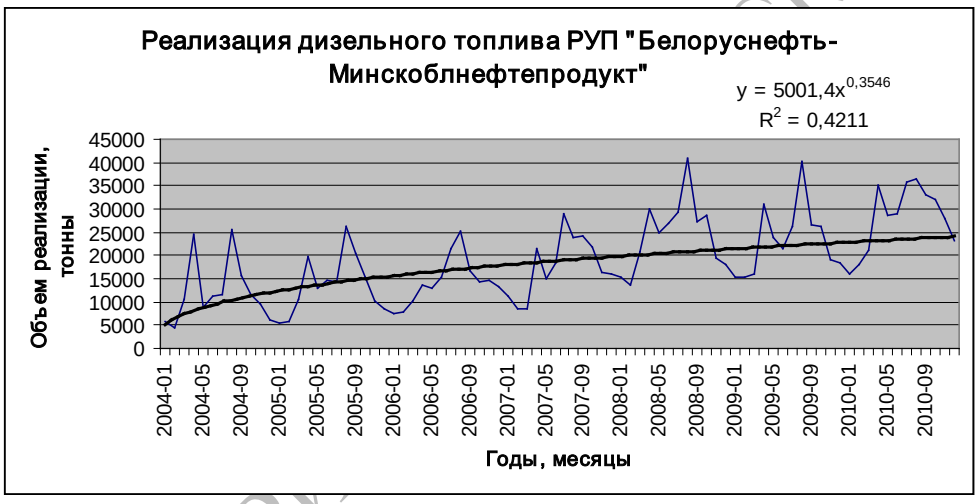


Рис. 6.25. График степенного тренда

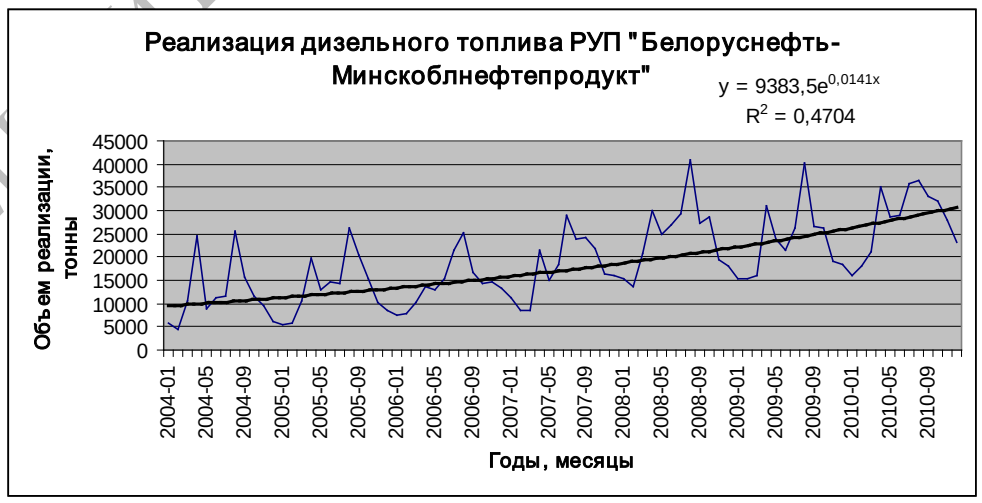


Рис. 6.26. График экспоненциального тренда

4) Из графика временного ряда следует также, что амплитуда колебаний со временем изменяется слабо. Поэтому в качестве итоговой формы модели временного ряда следует выбрать аддитивную форму.

Такой выбор при моделировании циклических колебаний предполагает использование абсолютных отклонений $S_i, i = 1, 2, \ldots, 12,$ фактического ряда от выровненного по каждому из двенадцати месяцев (в качестве выровненного ряда выбирается последовательность соответствующих значений линейного тренда $y = 238{,}12x + 9035{,}6$). Расчет абсолютных отклонений S_i, $i = 1, 2, \ldots, 12,$ осуществляется по формуле

$$S_i = \frac{1}{7}\sum_{j=1}^{7}(y_{ij} - y_{ij}^{\text{тр}}),$$

где y_{ij} – значение исходного временного ряда для i-ого месяца в j-ом году, $y_{ij}^{\text{тр}}$ – значение линейного тренда $y = 238{,}12x + 9035{,}6$ для i-ого месяца в j-ом году. Все результаты соответствующих вычислений сведены в таблице 6.12. Так как корректировочный коэффициент сезонности $\frac{1}{7}\sum_{i=1}^{7} S_i = \frac{-0{,}0878}{7} = -0{,}012$ близок к нулю, то корректировка значений сезонной компоненты не производится.

Таблица 6.12. Значения средних абсолютных отклонений

Месяц	Среднее абсолютное отклонение от тренда S_i
январь	-6911,33
февраль	-7496,43
март	-4334,64
апрель	6551,229
май	-617,277
июнь	558,6314
июль	4632,597
август	11692,93
сентябрь	3633,257
октябрь	1468,409
ноябрь	-3550,68
декабрь	-5627,75

5) Далее с учетом выбора аддитивной формы модели пересчитываются теоретические значения. Для этого складываются соответствующие значения тренда и средние абсолютные отклонения от него (таблица 6.13).

Таблица 6.13. Теоретические значения модельного временного ряда

2004 год	2005 год	2006 год	2007 год	2008 год	2009 год	2010 год
2362,394	5219,834	8077,274	10934,71	13792,15	16649,59	19507,03
2015,409	4872,849	7730,289	10587,73	13445,17	16302,61	19160,05
5415,323	8272,763	11130,2	13987,64	16845,08	19702,52	22559,96
16539,31	19396,75	22254,19	25111,63	27969,07	30826,51	33683,95
9608,923	12466,36	15323,8	18181,24	21038,68	23896,12	26753,56
11022,95	13880,39	16737,83	19595,27	22452,71	25310,15	28167,59
15335,04	18192,48	21049,92	23907,36	26764,8	29622,24	32479,68
22633,49	25490,93	28348,37	31205,81	34063,25	36920,69	39778,13
14811,94	17669,38	20526,82	23384,26	26241,7	29099,14	31956,58
12885,21	15742,65	18600,09	21457,53	24314,97	27172,41	30029,85
8104,237	10961,68	13819,12	16676,56	19534	22391,44	25248,88
6265,294	9122,734	11980,17	14837,61	17695,05	20552,49	23409,93

Построенная по этим данным аддитивная модель временного ряда имеет графическое изображение, представленное на рисунке 6.27.

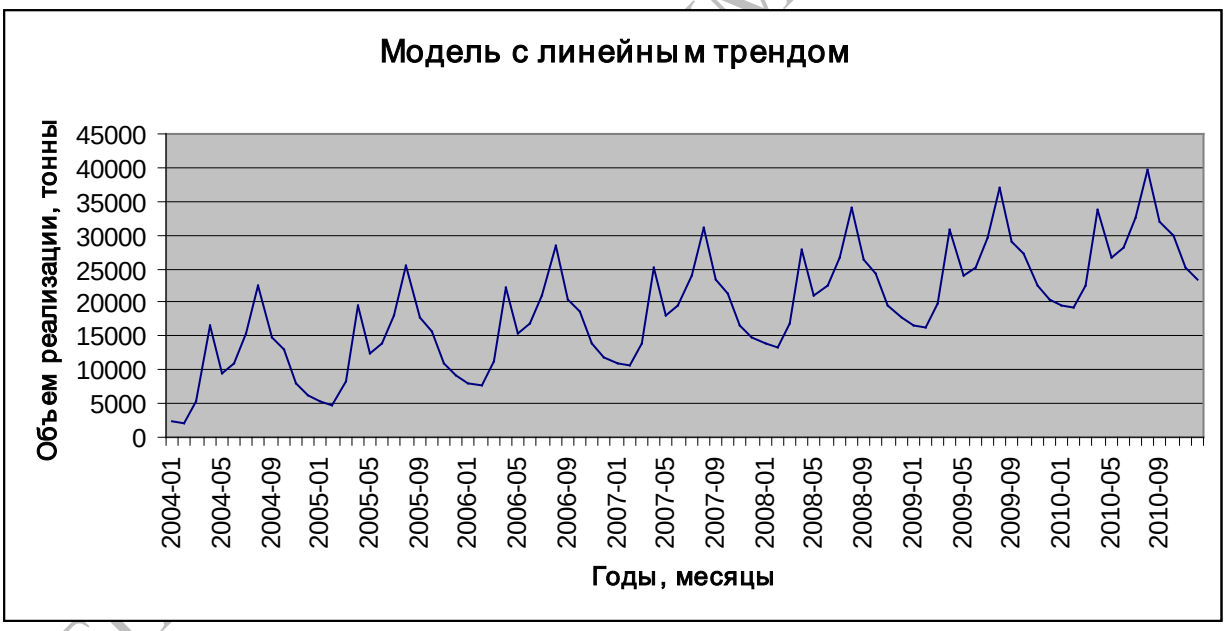


Рис. 6.27. Графическое изображение модельного временного ряда с линейным трендом

6) Средняя ошибка аппроксимации $\overline{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \tilde{y}_i}{y_i}\right| \cdot 100\%$ построенной модели составляет 12,9%. Таким образом, точность модели является удовлетворительной.

Реализация с помощью ППП *Excel*

Анализ структуры временного ряда, подбор линии тренда и прогнозирование удобно осуществлять с помощью «Пакета анализа» табличного процессора *Excel*. Отчасти это уже делалось в изложенных выше примерах.

Методику применения ППП *Excel* проиллюстрируем также на примере следующей задачи (при этом будем опираться на статистические данные, находящиеся в таблице 6.15, и исходить из того, что число наблюдений n равно 20).

Задача. Для осуществления прогноза исследовать совокупность квартальных данных о динамике выпуска продукции некоторого предприятия за пять лет.

Требуется:

1) ввести данные;

2) проверить гипотезу о существовании тренда с помощью метода сравнения средних уровней;

3) рассчитать значения коэффициентов автокорреляции r_1, r_2, r_3, r_4, r_5 для характеристики структуры временного ряда;

4) построить коррелограмму и с помощью автокорреляционной функции провести анализ структуры временного ряда;

5) провести аналитическое выравнивание временного ряда с использованием линейной, степенной, логарифмической, экспоненциальной и полиномиальной функций;

6) выбрать наилучшую форму тренда по максимальному значению коэффициента детерминации построенных моделей;

7) рассчитать по выбранной наилучшей форме тренда точечный прогноз выпуска продукции во втором квартале шестого года (с учетом сезонных колебаний, если они имеются во временном ряду).

Результаты вычислений и анализа оформить в виде отчета (*форма отчета прилагается ниже*)**.**

Порядок выполнения работы

1) В ячейку A1 занесите название – Квартал, в ячейку B1 – название Выпуск. В ячейки A2, A3, …, A21 введите данные первого столбца таблицы 6.15, в ячейки B2, B3, …, B21 – данные выбранного варианта задания.

Введите новое название листа «Исходные данные».

2) Проверьте гипотезу о существовании тренда с помощью метода сравнения средних уровней. Скопируйте ячейки A2:B11 в ячейку A1 нового листа, ячейки A12:B21 в ячейку D1 полученного листа. Новый лист назовите «Тест тренда».

Введите формулы:

в ячейку A13 =СРЗНАЧ(B1:B10);

в ячейку A14 =ДИСП(B1:B10);

в ячейку D13 =СРЗНАЧ(E1:E10);

в ячейку D14 =ДИСП(E1:E10).

В ячейку A15 введите формулу для вычисления наблюдаемого значения $F_{\text{набл}}$ критерия Фишера =A14/D14, если значение в ячейке A14 больше, чем в D14. В противном случае введите формулу =D14/A14. В ячейку A16 введите формулу =FРАСПОБР(0,05; 9; 9) для вычисления критического значения *F*-статистики $F_{\text{кр}}$ (при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $k_1 = k_2 = 9$).

Если $F_{\text{набл}} < F_{\text{кр}}$, то принимается гипотеза о равенстве дисперсий частей временного ряда, а значит, можно проверять гипотезу о наличии тренда.

В ячейку A17 введите формулу для вычисления наблюдаемого значения $t_{\text{набл}}$ статистики Сьюдента

=3*КОРЕНЬ(10)*ABS(A13-D13)/КОРЕНЬ(A14*(10-1)+B14*(10-1)).

В ячейку A18 введите формулу для вычисления критического значения $t_{\text{кр}}$ статистики Стьюдента =СТЬЮДРАСПОБР(0,05; 20-2).

Сравните значения $t_{\text{набл}}$ и $t_{\text{кр}}$. Если $t_{\text{набл}}$ больше $t_{\text{кр}}$, то сделайте вывод о присутствии тренда во временном ряду.

3) Вернитесь на лист «Исходные данные». В ячейку C1 введите название «Лаги». В ячейки C2:C6 введите значения 1, 2, 3, 4, 5.

В ячейку D1 введите название «Коэффициенты».

В ячейку D2 введите формулу коэффициента корреляции r_1

=КОРРЕЛ(B2:B20; B3:B21).

В ячейку D3 введите формулу коэффициента корреляции r_2

=КОРРЕЛ(B2:B19; B4:B21).

В ячейку D4 введите формулу коэффициента корреляции r_3

=КОРРЕЛ(B2:B18; B5:B21).

В ячейку D5 введите формулу коэффициента корреляции r_4

=КОРРЕЛ(B2:B17; B6:B21).

В ячейку D6 введите формулу коэффициента корреляции r_5

=КОРРЕЛ(B2:B16; B7:B21).

4) Постройте коррелограмму. Для этого выполните следующие действия:

– на панели инструментов активизируйте кнопку *Мастер диаграмм* (*шаг 1 из 4)*, в одноименном диалоговом окне среди стандартных типов выберите *График* и верхний левый вид диаграммы и нажмите кнопку *Далее>*;

– открывается диалоговое окно *Мастер диаграмм* (*шаг 2 из 4*), в котором во вкладке *Диапазон данных* в поле *Диапазон* введите ссылку на диапазон ячеек D2:D6; во вкладке *Ряд* в поле *Подписи оси X* введите ссылку на ячейки

C2:C6 значений лагов, в поле *Имя* введите название «Коэффициенты автокорреляции»; нажмите кнопку *Далее>*;

– открывается диалоговое окно *Мастер диаграмм* (*шаг 3 из 4*), в котором во вкладке *Заголовки* в поле *Ось X*(*категорий*) введите название «Лаги», в поле *Ось Y*(*значений*) – название «Коэффициенты автокорреляции»; во вкладке *Легенда* снимите флажок *Добавить легенду* и нажмите кнопку *Далее>*;

– открывается диалоговое окно *Мастер диаграмм* (*шаг 4 из 4*) в поле *имеющемся* установите флажок и нажмите кнопку *Готово*.

5) Проведите аналитическое выравнивание временного ряда с использованием линейной, степенной, логарифмической, экспоненциальной и полиномиальной функций. На листе «Исходные данные» выполняются следующие действия:

– на панели инструментов активизируйте кнопку *Мастер диаграмм* (*шаг 1 из 4*), в одноименном диалоговом окне среди стандартных типов выберите *График* и верхний левый вид диаграммы и нажмите кнопку *Далее>*;

– открывается диалоговое окно *Мастер диаграмм* (*шаг 2 из 4*), в котором во вкладке *Диапазон данных* в поле *Диапазон* введите ссылку на диапазон ячеек B2:B21; во вкладке *Ряд* в поле *Подписи оси X* введите ссылку на ячейки A2:A21, в поле *Имя* введите название «Выпуск продукции»; нажмите кнопку *Далее>*;

– открывается диалоговое окно *Мастер диаграмм* (*шаг 3 из 4*), в котором во вкладке *Заголовки* в поле *Ось X*(*категорий*) введите название «Кварталы», в поле *Ось Y*(*значений*) – название «Выпуск»; во вкладке *Легенда* снимите флажок *Добавить легенду* и нажмите кнопку *Далее>*;

– открывается диалоговое окно *Мастер диаграмм* (*шаг 4 из 4*) в поле *имеющемся* установите флажок и нажмите кнопку *Готово*.

Далее график с линией тренда пять раз копируется на новый лист «Тренды». На каждом из графиков курсор установите на кривой, нажмите левую клавишу мышки, затем нажмите правую кнопку мыши, выберите *Добавить линию тренда*. В диалоговом окне выделите один из типов трендов, во вкладке *Параметры* установите флажок в поле *показать уравнение на диаграмме, поместить на диаграмму R^2*. Повторите действия *Добавить линию тренда* для всех требуемых типов трендовых моделей.

6) Выберите наилучшую форму тренда по максимальному значению коэффициента детерминации построенных моделей.

7) Рассчитайте для выбранной трендовой модели точечный прогноз выпуска продукции. Для этого в ячейке B23 листа «Исходные данные» по уравнению лучшего тренда вычислите прогнозное значение при $x = 22$. При необходимости рассчитайте значения сезонной компоненты и учтите их при определении окончательного значения точечного прогноза.

Приложение: Отчет о результатах вычислений и анализа

1. *Спецификация, параметризация и верификация модели*

Анализ исходных данных позволяет сделать вывод о том, что они представляют собой временной ряд. Поэтому возникает задача анализа структуры этого временного ряда.

Как известно, при исследовании экономического временного ряда выделяются несколько составляющих: тренд, сезонная компонента, циклическая компонента, случайная компонента.

В данном случае может идти речь только о тренде, сезонной и случайной компонентах.

Для проверки гипотезы о существовании тренда используется метод сравнения средних уровней, который опирается на предположение об однородности исходного временного ряда. Это предположение и проверяется перед применением метода сравнения средних уровней.

Для этого исходный ряд разбивается на две равные части объема 10, для каждой из них вычисляются выборочные дисперсии, по которым находится значение $F_{\text{набл}}$ =..... и сравнивается со значением $F_{\text{кр}}$ =..... . Так как $F_{\text{набл}}$ (*больше или меньше*) $F_{\text{кр}}$, то гипотеза о равенстве дисперсий выделенных частей временного ряда (*принимается или отклоняется*).

Тренд по временному ряду присутствует, если

$$t_{\text{набл}} = \frac{\left|\bar{y}_1 - \bar{y}_2\right|}{\sqrt{S_1^2(10-1) + S_2^2(10-1)}} \cdot \sqrt{\frac{10 \cdot 10 \cdot (20-2)}{20}} > t_{\text{кр}}.$$

Так как $t_{\text{набл}}$ (*больше или меньше*) $t_{\text{кр}}$, то тренд во временном ряду (*присутствует или отсутствует*).

Анализ структуры временного ряда осуществляются на основании автокорреляционных коэффициентов, которые приведены в таблице 6.14.

Таблица 6.14. Коэффициенты автокорреляции

Лаги	Коэффициенты автокорреляции
1	
2	
3	
4	
5	

Наиболее высоким оказался коэффициент автокорреляции порядка (*1, 2, 3, 4, 5*). Анализ коррелограммы показывает, что исследуемый ряд содержит (*тренд, сезонную компоненту*). Кроме того, визуальный анализ точечного графика временного ряда подтверждает, что исследуемый ряд содержит (*тренд, сезонную компоненту*).

Из линейной, степенной, логарифмической, экспоненциальной и полиномиальной функций мы выбираем (*линейную, степенную, логарифмическую, экспоненциальную, полиномиальную*) функцию, так как у нее коэффициент детерминации ..... больше, чем у других.

Выбранная (*линейная, степенная, логарифмическая, экспоненциальная, полиномиальная*) модель тренда имеет вид (*записать уравнение тренда*).

2. Прогнозирование

Прогнозируемое по выбранной модели тренда значение выпуска продукции во втором квартале шестого года составляет ..... . Оно получается путем подстановки в выбранное уравнение тренда вместо x значения $x = 22$.

Интегрированная задача

С помощью «Пакета анализа» табличного процессора Excel, опираясь на находящиеся в таблице 6.15 квартальные данные о динамике выпуска продукции некоторого предприятия за пять лет, осуществить прогноз выпуска продукции во втором квартале шестого года. Расчеты и анализ провести в соответствии с требованиями и описанием задачи, изложенной в разделе «Реализация с помощью ППП *Excel*».

Результаты вычислений и анализа представить в виде отчета по форме, предложенной выше.

Таблица 6.15. Статистические данные интегрированной задачи

Квартал	Объем выпуска, вар. 1	Объем выпуска, вар. 2	Объем выпуска, вар. 3	Объем выпуска, вар. 4	Объем выпуска, вар. 5	Объем выпуска, вар. 6	Объем выпуска, вар. 7	Объем выпуска, вар. 8
1	7665	8650	3600	36010	3000	300	766	865
2	8570	9670	6570	65750	6070	607	857	967
3	11172	12135	9125	91220	9025	902	1117	1213
4	14150	15100	12104	121010	12004	1200	1415	1510
1	8465	9410	6400	64015	6000	600	846	941
2	9970	10870	7820	78205	7020	702	997	1087
3	12972	13900	10800	108025	10000	1000	1297	1390
4	15350	16310	13210	132130	13010	1301	1535	1631
1	9565	10510	7610	76112	7010	701	956	1051
2	10870	11820	8720	87245	8020	802	1087	1182
3	13972	14870	11650	116578	11050	1105	1397	1487
4	16350	17150	14050	140503	14150	1415	1635	1715
1	9965	10765	7760	77665	7060	706	996	1076
2	11470	12430	9410	94108	9010	901	1147	1243
3	14372	15300	12200	122045	12000	1200	1437	1530
4	16850	17800	14500	145056	14000	1400	1685	1780
1	10565	11255	8205	82058	8005	800	1056	1125
2	12070	12870	9810	98109	9010	901	1207	1287

3	14972	15771	12700	127034	12000	1200	1497	1577
4	17350	18100	15005	150052	15205	1520	1735	1810

Контрольные задания

Задание 6.1. Администрация банка изучает динамику депозитов физических лиц за ряд лет (млн. долл.). Исходные данные представлены в таблице 6.16.

1) Построить график временного ряда.

2) Построить уравнение линейного тренда и интерпретировать его параметры.

2) Определить коэффициент детерминации для линейного тренда.

Таблица 6.16. Данные задания 6.1

Год	2001	2002	2003	2004	2005	2006	2007
Депозиты физических лиц	4,2	6,4	7,1	8,3	10,5	12,6	13,2

Задание 6.2. Имеются данные (таблица 6.17) об урожайности зерновых культур (ц/га) в одном из районов области.

1) Построить график временного ряда и обосновать выбор тренда.

2) Рассчитать параметры уравнения выбранного тренда.

3) Дать точечный прогноз урожайности зерновых на 2010 год.

Таблица 6.17. Данные задания 6.2

Год	2002	2003	2004	2005	2006	2007	2008	2009
Урожайность	20,2	20,7	21,7	23,1	24,9	27,2	30,0	33,2

Задание 6.3. Имеются данные (таблица 6.18) об объемах продаж некоторой фирмы. Подобрать линию тренда, которая лучше всего описывает фактические данные, и на ее основе сделать прогноз на неделю вперед.

Таблица 6.18. Данные задания 6.3

Неделя	1	2	3	4	5	6	7
Стоимость проданной продукции	40	42	43	46	49	50	56

Задание 6.4. Имеются данные (таблица 6.19) об уровне продаж автомобилей некоторой фирмой за 9 недель ее работы. Подобрать линию тренда, которая лучше всего описывает фактические данные и на ее основе

сделать прогноз на две недели вперед.

Таблица 6.19. Данные задания 6.4

Неделя	1	2	3	4	5	6	7	8	9
Количество проданных единиц	3	10	11	18	21	30	32	37	42

Задание 6.5. Имеются данные (таблица 6.20) о продаже холодильников некоторым магазином за 10 недель его работы. Подобрать линию тренда, которая лучше всего описывает фактические данные, и на ее основе сделать прогноз на две недели вперед.

Таблица 6.20. Данные задания 6.5

Неделя	1	2	3	4	5	6	7	8	9	10
Количество проданных единиц	15	15	16	22	32	39	49	58	65	67

Задание 6.6. Имеются данные (таблица 6.21) о продаже телевизоров некоторым торговым центром за 11 месяцев его работы. Подобрать линию тренда, которая лучше всего описывает фактические данные, и на ее основе сделать прогноз на два месяца вперед.

Таблица 6.21. Данные задания 6.6

Месяц	1	2	3	4	5	6	7	8	9	10	11
Количество проданных телевизоров	18	25	27	26	30	37	35	40	48	53	57

Задание 6.7. Число общеобразовательных школ Республики Беларусь по годам приведено в таблице 6.22. Подобрать линию тренда, которая лучше всего описывает фактические данные, и на ее основе сделать прогноз на 2002 и 2003 годы.

Таблица 6.22. Данные задания 6.7

Годы	1995	1996	1997	1998	1999	2000	2001
Число школ	5007	4964	4918	4880	4830	4772	4718

Задание 6.8. В таблице 6.23 приведены данные о количестве студентов

вузов Республики Беларусь. Подобрать линию тренда, которая лучше всего описывает фактические данные, и на ее основе сделать прогноз на 2002 и 2003 годы.

Таблица 6.23. Данные задания 6.8

Годы	1995	1996	1997	1998	1999	2000	2001
Число студентов	197,4	208,9	224,5	244	262,1	281,7	301,8

Задание 6.9. В таблице 6.24 приведены поквартальные данные потребления электроэнергии на некотором предприятии за четыре года.

1) Построить график временного ряда.
2) Провести анализ структуры временного ряда.
3) Смоделировать тренд временного ряда.
4) Рассчитать сезонные (квартальные) компоненты.
5) Построить аддитивную модель временного ряда и ее график.
6) Оценить качество построенной модели.

Таблица 6.24. Данные задания 6.9

Номер квартала	1	2	3	4	5	6	7	8
y_t	6	4,4	5	9	7,2	4,8	6	10
Номер квартала	9	10	11	12	13	14	15	16
y_t	8	5,6	6,4	11	9	6,6	7	10,8

Задание 6.10. В таблице 6.25 приведены поквартальные данные предприятия об объемах выпуска некоторого товара за три года.

1) Построить график временного ряда.
2) Провести анализ структуры временного ряда.
3) Смоделировать тренд временного ряда.
4) Рассчитать сезонные (квартальные) компоненты.
5) Построить аддитивную модель временного ряда и ее график.
6) Оценить качество модели.
7) Осуществить прогноз объема выпуска товара во втором квартале четвертого года.

Таблица 6.25. Данные задания 6.10

Квартал	1	2	3	4
Выпуск	410	400	715	600
Квартал	5	6	7	8
Выпуск	585	560	975	800
Квартал	9	10	11	12

Выпуск	765	720	1235	1100

Задание 6.11. С помощью метода сравнения средних уровней определить наличие тренда во временном ряду y_t, заданном таблицей 6.26.

Таблица 6.26. Данные примера 6.11

Период	1	2	3	4	5	6	7	8
y_t	24,3	19,8	28,4	29,9	15,7	34,6	23,7	33,9
Период	9	10	11	12	13	14	15	16
y_t	25,1	26,8	15,8	26,7	35,2	21,7	28,9	17,3

Задание 6.12. Недельные объемы продаж некоторой фирмы (млн. руб.) представлены в таблице 6.27. Проверить с помощью метода Ирвина наличие аномальных уровней.

Таблица 6.27. Данные примера 6.12

t	y_t	t	y_t	t	y_t
1	71	5	81	9	87
2	68	6	78	10	75
3	65	7	27	11	78
4	72	8	83	12	81

Задание 6.13. В таблице 6.28 приведены данные, отражающие динамику курса акций некоторой компании.

Таблица 6.28. Данные примера 6.13

t	1	2	3	4	5	6	7	8	9	10	11
y_t	971	1166	1044	907	957	727	752	1019	972	815	823
t	12	13	14	15	17	17	18	19	20	21	22
y_t	1112	1386	1428	1364	1241	1145	1351	1325	1226	1189	1213

Дать точечный и интервальный прогнозы значения курса акций в момент $t = 23$.

Задание 6.14. Имеются поквартальные данные (таблица 6.29) об объемах потребления y_t некоторого продукта (тыс. тонн) в развивающемся городе за восемь лет.

1) Построить график временного ряда.

2) Провести визуальный анализ структуры временного ряда.

3) Методом сравнения средних уровней определить наличие тренда.

4) Смоделировать тренд временного ряда.

5) Рассчитать сезонные (квартальные) компоненты.

6) Построить аддитивную модель временного ряда и ее график.

7) Оценить качество модели.

8) Осуществить прогноз потребления продукта во втором квартале девятого года.

Таблица 6.29. Данные примера 6.14

Квартал	1	2	3	4	5	6	7	8
y_t	15	5	10	35	26	19	23	46
Квартал	9	10	11	12	13	14	15	16
y_t	38	31	34	58	51	41	46	70
Квартал	17	18	19	20	21	22	23	24
y_t	63	53	58	82	75	67	70	94
Квартал	25	26	27	28	29	30	31	32
y_t	86	77	84	105	98	89	94	117

Контрольные вопросы

1. Дайте определение временного ряда?
2. Приведите примеры временных рядов в экономике.
3. В чем заключается различие между данными пространственной выборки и данными временного ряда?
4. Какие требования предъявляются к уровням временного ряда?
5. В чем заключается требование однородности уровней временного ряда?
6. В чем заключается требование сопоставимости уровней временного ряда?
7. Как выявляются аномальные уровни временного ряда?
8. Изложите сущность метода Ирвина.
9. Как определяется модель временного ряда?
10. Какая модель называется динамической?
11. Приведите примеры экономических задач, требующих применения моделей с распределенным лагом и моделей авторегрессии.
12. В чем проявляется специфика динамических моделей по сравнению со статическими?
13. Перечислите основные составляющие временного ряда.
14. Как определяется аддитивная модель временного ряда?
15. Как строится мультипликативная модель временного ряда?
16. В чем заключаются основные задачи эконометрического моделирования временного ряда?

17. Перечислите основные методы выявления тренда во временных рядах.

18. Сформулируйте сущность метода сравнения средних уровней временного ряда.

19. Какой вид связи между соседними уровнями характеризует коэффициент автокорреляции?

20. Как вычисляются коэффициенты автокорреляции первого, второго и более высоких порядков?

21. Что такое автокорреляционная функция?

22. Как можно по коэффициентам автокорреляции определить структуру временного ряда?

23. Опишите процедуру сглаживания временного ряда с помощью метода скользящей средней.

24. В чем заключается аналитическое выравнивание временного ряда?

25. Перечислите основные виды трендов.

26. Охарактеризуйте основные типы кривых роста.

27. Перечислите основные методы выявления сезонной компоненты.

28. Как осуществляется моделирование сезонных колебаний временного ряда?

29. Как вычисляется абсолютное отклонение?

30. Как вычисляется индекс сезонности?

31. В каких случаях используется аддитивная, а в каких мультипликативная модель временного ряда?

32. Опишите схему использования фиктивных переменных для моделирования сезонных компонент временного ряда.

33. Сформулируйте основные этапы построения аддитивной (мультипликативной) модели временного ряда.

34. Перечислите основные методы оценки случайной компоненты.

35. Как с помощью остатков временного ряда определить качество моделирования тренда и периодических компонент временного ряда?

36. Как проверяется условие равенства нулю математического ожидания случайной компоненты?

37. Как проверяется условие нормальности распределения остатков временного ряда?

38. Как тестируется автокорреляция в остатках временного ряда?

39. Что такое структурный сдвиг?

40. Приведите примеры экономических событий, приводящих к структурным сдвигам.

41. Как с помощью теста Чоу проверить гипотезу о структурной стабильности временного ряда?

42. В чем заключается метод экстраполяции?

43. Раскройте сущность точечного и интервального прогнозов по модели временного ряда.

44. Как осуществляется точечный прогноз по трендовой модели?

45. Как осуществляется точечный прогноз по модели временного ряда, содержащей тренд и сезонную компоненту?

46. Что такое коинтеграция временных рядов?

47. Как тестируется коинтеграция временных рядов?

48. Перечислите основные методы исключения тенденции из временного ряда.

Тестовые задания

Выберите правильные ответы из предложенных вариантов:

1. Сезонная компонента временного ряда – это:

а) компонента, описывающая долговременную тенденцию изменения;

б) компонента, определяющая периодические колебания экономических процессов в течение длительных периодов;

в) компонента, отражающая повторяемость экономических процессов в течение не очень значительного периода;

г) компонента, отражающая влияние на уровни ряда случайных факторов.

2. Тренд временного ряда – это:

а) компонента, описывающая долговременную тенденцию изменения;

б) компонента, определяющая периодические колебания экономических процессов в течение длительных периодов;

в) компонента, отражающая повторяемость экономических процессов в течение не очень значительного периода;

г) компонента, отражающая влияние на уровни ряда случайных факторов.

3. Циклическая компонента временного ряда – это:

а) компонента, описывающая долговременную тенденцию изменения;

б) компонента, определяющая периодические колебания экономических процессов в течение длительных периодов;

в) компонента, отражающая повторяемость экономических процессов в течение не очень значительного периода;

г) компонента, отражающая влияние на уровни ряда случайных факторов.

4. Аддитивная модель временного ряда имеет вид:

а) $y_t = f(T, S, C, \varepsilon)$;

б) $y_t = T(t) + S(t) + C(t) + \varepsilon(t)$;

в) $y_t = T(t) \cdot S(t) \cdot C(t) \cdot \varepsilon(t)$;

г) $y_t = T(t) \cdot S(t) + C(t) \cdot \varepsilon(t)$.

5. Мультипликативная модель временного ряда имеет вид:

а) $y_t = f(T, S, C, \varepsilon)$;

б) $y_t = T(t) + S(t) + C(t) + \varepsilon(t)$;

в) $y_t = T(t) \cdot S(t) \cdot C(t) \cdot \varepsilon(t)$;

г) $y_t = T(t) \cdot S(t) + C(t) \cdot \varepsilon(t)$.

6. К основным задачам эконометрического исследования отдельного временного ряда относится:

а) задача выделения и количественного выражения закономерных компонент;

б) задача анализа случайной составляющей;

в) задача прогнозирования будущих уровней временного ряда;

г) задача параметризации модели;

д) задача спецификации временного ряда.

7. Для выявления структуры временного ряда используется:

а) тест Дарбина-Уотсона;

б) автокорреляционная функция;

в) корреляционная матрица.

8. Полиномиальный тренд имеет вид:

а) $\tilde{y}_t = a + bt$;

б) $\tilde{y}_t = a + b_1 t + b_2 t^2 + ... + b_m t^m$;

в) $\tilde{y}_t = a + \frac{b}{t}$;

г) $\tilde{y}_t = e^{a+bt}$.

9. Для выявления периодической компоненты во временном ряду используется:

а) тест Дарбина-Уотсона;

б) автокорреляционная функция;

в) тест Чоу;

г) корреляционная матрица.

10. На стадии спецификации тренда временного ряда чаще других используется:

а) графический метод;

б) экспериментальный метод;

в) аналитический метод;

г) метод наименьших квадратов.

11. Для моделирования сезонных колебаний могут быть использованы:

а) лаговые переменные;

б) факторные переменные;

в) фиктивные переменные;

г) эндогенные переменные.

12. В случае описания поквартальной сезонности количество используемых фиктивных переменных равно:
а) 2; б) 3; в) 4; г) 5.

13. Аддитивная модель временного ряда строится, если:
а) значения сезонной компоненты предполагаются постоянными для различных циклов;
б) амплитуда сезонных колебаний возрастает или уменьшается;
в) отсутствует линейная тенденция.

14. Мультипликативная модель временного ряда строится, если:
а) значения сезонной компоненты предполагаются постоянными для различных циклов;
б) амплитуда сезонных колебаний возрастает или уменьшается;
в) отсутствует линейная тенденция.

15. На основе поквартальных данных построена аддитивная модель временного ряда. Скорректированные значения сезонной компоненты за первые три квартала равны: 7 – I квартал, 9 – II квартал и –11 – III квартал. Значение сезонной компоненты за IV квартал равно:
а) 5; б) –4; в) –5.

16. На основе поквартальных данных построена мультипликативная модель временного ряда. Скорректированные значения сезонной компоненты за первые три квартала равны: 0,8 – I квартал, 1,2 – II квартал и 1,3 – III квартал. Значение сезонной компоненты за IV квартал равно:
а) 0,7; б) 1,7; в) 0,9.

17. Критерий Дарбина-Уотсона применяется для:
а) определения автокорреляции в остатках;
б) определения наличия сезонных колебаний;
в) для оценки существенности построенной модели.

18. В стационарном временном ряде трендовая компонента:
а) имеет линейную зависимость от времени;
б) отсутствует;
в) имеет нелинейную зависимость от времени;
г) присутствует.

19. Известны значения мультипликативной модели временного ряда: $y(t)=15$ – значение уровня ряда, $T(t)=5$ – значение тренда, $S(t)=3$ – значение сезонной компоненты. Определите значение случайной компоненты:

а) $\varepsilon(t) = -1$; б) $\varepsilon(t) = 3$; в) $\varepsilon(t) = 1$; г) $\varepsilon(t) = 0$.

20. Если наиболее высоким оказался коэффициент автокорреляции первого порядка, то исследуемый ряд содержит:
а) только тенденцию;
б) циклические колебания с периодичностью в один момент времени;
в) сильную нелинейную тенденцию;
г) случайную компоненту.

21. Коррелограммой называется:
а) график автокорреляционной функции;
б) аналитическое выражение автокорреляционной функции;
в) график временного ряда;
г) процесс нахождения значений автокорреляционной функции.

22. Известны значения аддитивной модели временного ряда: $y(t) = 30$ – значение уровня ряда, $T(t) = 15$ – значение тренда, $\varepsilon(t) = 2$ – значение случайной компоненты. Определите значение сезонной компоненты:
а) $S(t) = 0$; б) $S(t) = 13$; в) $S(t) = 1$; г) $S(t) = -1$.

23. Если наиболее высоким оказался коэффициент автокорреляции третьего порядка, то исследуемый ряд содержит:
а) сезонные колебания с периодичностью в три момента времени;
б) линейный тренд, проявляющийся в каждом третьем уровне ряда;
в) случайную величину, влияющую на каждый третий уровень ряда;
г) нелинейную тенденцию в виде полинома третьего порядка.

24. Уровнем временного ряда является:
а) значение временного ряда в конкретный момент (период) времени;
б) среднее значение временного ряда;
в) совокупность значений временного ряда;
г) значение конкретного момента (периода) времени.

Ответы тестовых заданий

Номер задания	Ответы	Номер задания	Ответы	Номер задания	Ответы
1	в)	9	б)	17	а)
2	а)	10	а), б)	18	б)
3	б)	11	в)	19	в)
4	б)	12	б)	20	а)
5	в)	13	а)	21	а)
6	а), б), в)	14	б)	22	б)
7	б)	15	в)	23	а)

8	б)	16	а)	24	а)

Глава 7
Системы эконометрических уравнений

Основные понятия: *система одновременных уравнений, система независимых уравнений, система рекурсивных уравнений, система взаимосвязанных уравнений, эндогенные, экзогенные и лаговые переменные, расширенная форма модели, структурная форма модели, приведенная форма модели, проблема идентифицируемости, идентифицируемая, неидентифицируемая и сверхидентифицируемая модель, косвенный метод наименьших квадратов, двухшаговый метод наименьших квадратов.*

Литература: [2-4], [7], [14].

7.1. Системы уравнений, используемые в эконометрике

Отдельно взятое уравнение множественной регрессии не всегда адекватно характеризует сложное экономическое явление. Это связано, прежде всего, с наличием внешних факторов, которые *одновременно* воздействуют на регрессоры и объясняемые переменные. В результате выбранные переменные оказывают взаимное влияние друг на друга: изменение одних влечет изменения других. При этом порой трудно даже определить, какая из переменных является регрессором, а какая независимой переменной.

Обратимся к классическому примеру. Как известно, формирование предложения S_t и спроса D_t товара осуществляется в зависимости от его цены P_t и дохода населения I_t: $S_t = a_0 + a_1 P_t + \varepsilon_1$; $D_t = b_0 + b_1 P_t + b_2 I_t + \varepsilon_2$. Поэтому, анализируя спрос и предложение отдельно друг от друга, мы можем считать S_t и D_t объясняемыми переменными, а цену P_t и величину дохода I_t – факторами. В таком случае вполне уместно использовать регрессионные модели в виде одного уравнения.

Ситуация принципиально меняется в случае рассмотрения взаимодействия (равновесия) спроса и предложения. В этом случае к уравнениям предложения и спроса добавляется балансовое равенство $S_t = D_t = Q_t$ и, кроме того, наблюдаемое значение P_t – это равновесная цена, которая формируется одновременно со спросом и предложением. Поэтому здесь объем спроса и предложения Q_t и равновесная цена P_t рассматриваются уже в качестве объясняемых переменных, а величина дохода I_t – в качестве объясняющей переменной.

Таким образом, в результате одновременного действия внешних факторов модель множественной линейной регрессии может оказаться неполной. Поэтому при рассмотрении достаточно сложных экономических явлений

приходится строить модель с помощью системы одновременных уравнений. В одних уравнениях такой системы одна и та же переменная может выступать в качестве объясняемой переменной, а в других быть регрессором.

Системой одновременных уравнений называется такая эконометрическая модель, которая представлена системой линейных уравнений, хотя бы одно из которых содержит случайную составляющую (уравнения, которые не содержат случайную составляющую, называются *балансовыми*). Регрессионные уравнения системы одновременных уравнений называются *поведенческими*.

В системе

$$\begin{cases} S_t = a_0 + a_1 P_t + \varepsilon_1 - \textit{уравнение предложения}, \\ D_t = b_0 + b_1 P_t + b_2 I_t + \varepsilon_2 - \textit{уравнение спроса}, \\ S_t = D_t - \textit{уравнение равновесия}, \end{cases} \qquad (7.1)$$

описывающей равновесие между спросом и предложением, первое и второе уравнения являются поведенческими, а третье – балансовым.

Различают три типа эконометрических систем:

– система независимых уравнений;

– система рекурсивных уравнений;

– система взаимозависимых уравнений.

В *системе независимых уравнений* каждая из зависимых переменных y_1, y_2, …, y_n рассматривается как линейная функция от одних и тех же факторов x_1, x_2, …, x_m. Система независимых уравнений имеет вид:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ ..... \\ y_n = a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases} \qquad (7.2)$$

Каждое из уравнений такой системы может рассматриваться самостоятельно, а потому параметризация системы сводится к оценке параметров каждого уравнения с помощью МНК.

Система рекурсивных уравнений определяется таким образом, что каждое ее уравнение в качестве факторов содержит все зависимые переменные предшествующих уравнений, а также переменные $x_1, x_2, ..., x_m$. Система имеет вид:

$$\begin{cases} y_1 = & + a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + & + a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + & + a_{31}x_1 + a_{32}x_2 + ... + a_{3m}x_m + \varepsilon_3, \\ ..... \\ y_n = b_{n1}y_1 + b_{n2}y_2 + ... + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases} \quad (7.3)$$

В таких моделях параметризация осуществляется последовательно (от первого уравнения к последнему) с помощью МНК.

Наиболее распространена в экономических исследованиях *система взаимосвязанных* (*одновременных*) *уравнений.* В ней одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других уравнениях – в правую часть системы:

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + ... + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + ... + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ ..... \\ y_n = b_{n1}y_1 + b_{n2}y_2 + ... + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases} \quad (7.4)$$

Для нахождения параметров уравнений системы (7.4) традиционный МНК неприменим, так как нарушаются модельные предположения:

1) факторы в системе (7.4) мультиколлинеарны;

2) случайные составляющие оказываются коррелированными с объясняющими переменными.

Если применить к уравнениям этой системы обычный метод наименьших квадратов, то получатся смещенные и несостоятельные оценки параметров. Поэтому для оценивания систем одновременных уравнений используются специальные методы.

7.2. Структурная и приведенная формы моделей

Различают расширенную, структурную и приведенную формы записи системы эконометрических уравнений.

Расширенная форма модели – это такая система уравнений, которая содержит балансовые уравнения.

Пример расширенной формы дает модель (7.1) взаимодействия спроса, предложения и цены, содержащая одно балансовое уравнение.

За счет балансовых уравнений некоторые из переменных могут быть исключены, что позволяет в конечном итоге перейти к рассмотрению системы уравнений меньшей размерности. Например, за счет балансового равенства $S_t = D_t = Q_t$ в системе (7.1) можно перейти к системе

$$\begin{cases} Q_t = A_0 + A_1 I_t + \varepsilon_3, \\ P_t = B_0 + B_1 I_t + \varepsilon_4, \end{cases} \quad (7.5)$$

содержащей только два уравнения.

Структурной формой модели называется система уравнений (7.4). Уравнения, входящие в структурную форму, называются *структурными уравнениями*. Коэффициенты уравнений структурной формы называются *структурными коэффициентами*.

Для простоты все переменные в модели (7.4) выражены в отклонениях от среднего уровня, т.е. под x_i подразумевается $x_i - \bar{x}$, а под y_i – соответственно $y_i - \bar{y}$. Поэтому свободный член в каждом уравнении системы (7.4) отсутствует.

В процессе оценивания параметров системы (7.4) следует различать два класса переменных: эндогенные и экзогенные. *Эндогенными* называются переменные, значения которых определяется внутри модели. В системе (7.3) эндогенными являются переменные $y_1, y_2, \ldots, y_n$. Их число совпадает с числом уравнений системы. *Экзогенными* называются те переменные, значения которых определяется вне модели. Это предопределенные переменные, которые влияют на эндогенные переменные, но не зависят от них. В динамических моделях в качестве экзогенных могут рассматриваться значения эндогенных переменных за предыдущие периоды времени – *лаговые* переменные.

С математической точки зрения главное отличие между экзогенными и эндогенными переменными заключается в том, что экзогенные переменные не коррелируют с ошибками регрессии, а эндогенные, как правило, коррелируют.

Разделение переменных на эндогенные и экзогенные во многом зависит от теоретической концепции принятой модели. Экономические переменные могут выступать в одних моделях как эндогенные, а в других как экзогенные переменные. Внеэкономические переменные (социальное положение, пол, возрастная категория) входят в систему только как экзогенные переменные.

Структурная форма модели позволяет увидеть влияние изменений любой экзогенной переменной на значения эндогенной переменной. Поэтому целесообразно в качестве экзогенных переменных выбирать такие переменные, которые могут быть объектом регулирования. Меняя их или управляя ими, можно получать целевые значения эндогенных переменных.

Приведенной формой модели называется система уравнений, в каждом из которых эндогенные переменные выражены только через экзогенные переменные и случайные составляющие:

$$\begin{cases} y_1 = c_{11}x_1 + c_{12}x_2 + \ldots + c_{1m}x_m + \varepsilon_1, \\ y_2 = c_{21}x_1 + c_{22}x_2 + \ldots + c_{2m}x_m + \varepsilon_2, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ y_n = c_{n1}x_1 + c_{n2}x_2 + \ldots + c_{nm}x_m + \varepsilon_n. \end{cases} \quad (7.6)$$

Коэффициенты приведенной формы (7.6) называются *приведенными коэффициентами*. Они оцениваются с помощью обычного МНК, так как экзогенные переменные не коррелированны со случайными составляющими.

Приведенная форма строится для того, чтобы через МНК-оценки ее параметров выразить оценки структурных коэффициентов. Возможность такого выражения проиллюстрируем на следующем примере.

Пусть задана структурная модель

$$\begin{cases} y_1 = b_{12} y_2 + a_{11} x_1 + \varepsilon_1, \\ y_2 = b_{21} y_1 + a_{22} x_2 + \varepsilon_2. \end{cases} \tag{7.7}$$

Тогда приведенная форма модели имеет вид

$$\begin{cases} y_1 = c_{11} x_1 + c_{12} x_2 + \varepsilon_1, \\ y_2 = c_{21} x_1 + c_{22} x_2 + \varepsilon_2. \end{cases} \tag{7.8}$$

Из второго уравнения системы (7.8) выразим x_2 следующим образом:

$$x_2 = \frac{y_2 - c_{21} x_1}{c_{22}}.$$

Подставляя результат в первое уравнение системы (7.8), имеем

$$y_1 = c_{11} x_1 + c_{12} x_2 = c_{11} x_1 + c_{12} \frac{y_2 - c_{21} x_1}{c_{22}},$$

Откуда

$$y_1 = \frac{c_{12}}{c_{22}} y_2 + \frac{c_{11} c_{22} - c_{12} c_{21}}{c_{22}} x_1.$$

Выражая далее x_1 из первого уравнения системы (7.8) и подставляя результат во второе уравнение системы (7.8), аналогично получим

$$y_2 = \frac{c_{21}}{c_{11}} y_1 + \frac{c_{11} c_{22} - c_{12} c_{21}}{c_{11}} x_2.$$

Таким образом, коэффициенты приведенной формы (7.8) модели связаны с коэффициентами структурной формы (7.7) следующим образом:

$$b_{12} = \frac{c_{12}}{c_{22}},\ b_{21} = \frac{c_{21}}{c_{11}},\ a_{11} = \frac{c_{11}c_{22} - c_{12}c_{21}}{c_{22}},\ a_{22} = \frac{c_{11}c_{22} - c_{12}c_{21}}{c_{11}}.$$

Теперь зная оценки параметров модели (7.8), по полученным формулам можно получить оценки параметров структурной модели (7.7). Такой метод оценивания структурных коэффициентов называется *косвенным методом наименьших квадратов*.

7.3. Проблема идентифицируемости модели

В ходе анализа приведенного примера возникают два естественных вопроса:

1) Всегда ли можно структурные коэффициенты алгебраически выразить через приведенные?

2) Является ли такое выражение единственным?

Эти вопросы и составляют суть *проблемы идентифицируемости* модели, т.е. проблемы возможности и единственности численной оценки структурных коэффициентов по оценкам приведенных коэффициентов.

Ответы на вопросы достаточно очевидны. Структурная модель (7.4) в полном виде содержит $(n-1)n + mn$ параметров, а приведенная модель (7.6) – только mn параметров, т.е. в полном виде структурная модель содержит больше параметров, чем приведенная модель. Поэтому параметры структурной модели не всегда идентифицируемы, т.е. структурные коэффициенты не всегда могут быть определены по оценкам приведенных коэффициентов (а тем более однозначно).

С позиции идентифицируемости структурные модели подразделяются на три вида:

- идентифицируемые;
- неидентифицируемые;
- сверхидентифицируемые.

Модель является *идентифицируемой* (или *точно идентифицируемой*), если все ее структурные коэффициенты определяются однозначно по коэффициентам приведенной формы модели, т.е. если число параметров структурной модели равно числу параметров приведенной формы модели.

Модель является *неидентифицируемой*, если число параметров структурной модели больше числа параметров приведенной формы модели. В таком случае структурные коэффициенты не могут быть оценены через коэффициенты приведенной формы модели. Структурная модель (7.4) в полном виде, содержащая n эндогенных и m экзогенных переменных в каждом уравнении системы, всегда неидентифицируема.

Модель является *сверхидентифицируемой*, если число параметров структурной модели меньше числа параметров приведенной формы модели.

В этом случае на основе коэффициентов приведенной формы можно получить два и более значения одного структурного коэффициента. Сверхидентифицируемая модель в отличие от неидентифицируемой модели практически решаема, но требует для этого специальных методов оценки параметров.

Для определения вида модели необходимо рассматривать каждое уравнение. Если каждое уравнение системы идентифицируемо, то и модель считается идентифицируемой. Если хотя бы одно уравнение в системе не идентифицируемо, то вся модель считается неидентифицируемой. Модель считается сверхидентифицируемой, если хотя бы одно ее уравнение сверхидентифицируемо.

Для быстрого определения идентифицируемости структурной модели применяются следующие необходимые и достаточные условия.

Необходимое условие идентифицируемости. *Пусть* H *– число эндогенных переменных в уравнении, а* D *– число экзогенных переменных, отсутствующих в уравнении, но присутствующих в системе. Тогда справедливо следующее счетное правило:*

1) *если уравнение идентифицируемо, то* $D = H - 1$;

2) *если уравнение сверхидентифицируемо, то* $D > H - 1$;

3) *если уравнение неидентифицируемо, то* $D < H - 1$.

Предложенное правило, являясь необходимым, не достаточно для идентификации модели. Возможна ситуация, когда для каждого уравнения системы выполнено условие $D = H - 1$, но система уравнений не является идентифицируемой.

Достаточное условие идентифицируемости. *Уравнение системы идентифицируемо, если определитель матрицы* A*, составленной из коэффициентов при переменных* (*эндогенных и экзогенных*)*, отсутствующих в исследуемом уравнении, отличен от нуля, а ее ранг не меньше* $n - 1$*, где* n *– число эндогенных переменных системы.*

Приведем теперь (таблица 7.1) ***необходимые и достаточные условия,*** которые позволяют определить характер идентифицируемости структурной модели, содержащей n эндогенных переменных.

Таблица 7.1. Необходимое и достаточное условие идентифицируемости уравнения модели

Счетное правило	Ранг матрицы A	Вывод о идентифицируемости
$D > H - 1$	равен $n - 1$	уравнение сверхидентифицируемо
$D = H - 1$	равен $n - 1$	уравнение идентифицируемо
$D \geq H - 1$	меньше $n - 1$	уравнение неидентифицируемо
$D < H - 1$	меньше $n - 1$	уравнение неидентифицируемо

Проблема неидентифицируемости – это проблема структуры модели. В случае неидентифицируемой модели для получения статистического решения, описывающего изучаемое экономическое явление, сама модель должна быть соответствующим образом скорректирована и доведена до состояния идентифицируемости. Например, можно предположить, что некоторые из структурных коэффициентов модели ввиду слабой взаимосвязи соответствующих переменных с эндогенной переменной из левой части системы равны нулю. Тем самым уменьшится число структурных коэффициентов модели. Уменьшение числа структурных коэффициентов модели возможно и другим путем: например, уравниванием некоторых коэффициентов на основе предположения, что воздействие соответствующих переменных на формируемую эндогенную переменную одинаково.

7.4. Методы оценивания параметров структурной модели

После того, как решена проблема идентифицируемости рассматриваемой структурной модели, осуществляется оценка параметров этой модели. Наибольшее распространение получили следующие методы установления оценок коэффициентов структурной модели:

– *косвенный метод наименьших квадратов* (*КМНК*);

– *двухшаговый метод наименьших квадратов* (*ДМНК*).

КМНК используется в случае идентифицируемой структурной модели, а для параметризации сверхидентифицируемых структурных моделей используется ДМНК.

Косвенный МНК состоит в следующем:

1. Составляется приведенная форма модели.

2. Для каждого приведенного уравнения обычным МНК оцениваются приведенные коэффициенты.

3. С помощью алгебраических преобразований осуществляется переход от приведенной формы к уравнениям структурной формы модели. При этом приведенные коэффициенты трансформируются в численные оценки структурных параметров.

Смысл метода заложен в названии «косвенный МНК», которое подчеркивает, что структурные коэффициенты исходной модели находятся косвенно, т.е. через оценки приведенных коэффициентов.

Двухшаговый МНК заключается в следующем:

1. Составляется приведенная форма модели и с помощью обычного МНК определяются численные значения приведенных коэффициентов.

2. Выявляются эндогенные переменные из правой части структурного уравнения; находятся теоретические значения этих эндогенных переменных по соответствующим уравнениям приведенной формы модели.

3. Теоретические значения эндогенных переменных, стоящих в правых частях уравнений, подставляются вместо их фактических значений и с помощью обычного МНК определяются структурные коэффициенты исходной модели.

Метод получил название «двухшаговый МНК» потому, что метод наименьших квадратов используется в нем дважды: на первом шаге при определении приведенных коэффициентов и на втором шаге при определении структурных коэффициентов модели по данным теоретических значений эндогенных переменных.

Двухшаговый МНК является наиболее распространенным методом решения системы одновременных уравнений. Для идентифицируемых моделей двухшаговый МНК дает тот же результат, что и косвенный МНК.

7.5. Практика применения систем одновременных уравнений в макроэкономическом анализе

Системы одновременных уравнений широко применяются для построения макроэкономических моделей функционирования экономики той или иной страны. В этом направлении следует выделить модели Кейнса. Наиболее простая статическая модель Кейнса для описания экономики страны имеет вид:

$$\begin{cases} C_t = a + by_t + \varepsilon_t \ (\textit{функция потребления}), \\ y_t = C_t + I_t \ (\textit{балансовое уравнение потребления}), \end{cases}$$

где C_t – совокупное потребление, y_t – национальный доход, I_t – инвестиции, ε_t – случайная составляющая.

Указанная модель описывает закрытую экономику без вмешательства государства. Она содержит одно поведенческое уравнение и одно уравнение баланса.

Данная модель Кейнса идентифицируема. Ее приведенная форма имеет вид $y_t = \dfrac{a}{1-b} + \dfrac{1}{1-b} I_t + \dfrac{\varepsilon_t}{1-b}$.

В более поздних исследованиях статическая модель Кейнса включала дополнительно и функцию сбережений:

$$\begin{cases} C_t = a + by_t + \varepsilon_1 \ (\textit{функция потребления}), \\ r_t = a_1 + b_1(C_t + I_t) + \varepsilon_2 \ (\textit{функция сбережений}), \\ y_t = C_t + I_t - r_t \ (\textit{балансовое уравнение дохода}), \end{cases}$$

где r_t – сбережения.

Данная модель содержит три эндогенных переменных C_t, r_t, y_t, а также одну экзогенную переменную I_t. Система является идентифицируемой и решается с помощью КМНК.

Наряду со статическими моделями широкое распространение получили и динамические модели экономики. Такие модели содержат лаговые

переменные и учитывают тенденцию (фактор времени). Одна из таких моделей Кейнса имеет вид:

$$\begin{cases} C_t = a_1 + a_2 y_t + a_3 C_{t-1} + \varepsilon_1 \ (\textit{функция потребления}), \\ I_t = b_1 + b_2 y_t + \varepsilon_2 \ \ (\textit{функция сбережений}), \\ y_t = C_t + I_t + G_t \ (\textit{балансовое уравнение дохода}), \end{cases}$$

где C_t – потребление, y_t – ВВП, I_t – валовые инвестиции, G_t – государственные расходы.

В данной модели используется потребление предыдущего периода, т.е. считается, что потребление текущего года зависит не только от дохода, но и от достигнутого в предыдущий период уровня потребления.

Другим примером динамической модели является модель Клейна для экономики США в 1950-1960 годах:

$$\begin{cases} C_t = b_1 S_t + b_2 P_t + b_3 + \varepsilon_1, \\ I_t = b_4 P_t + b_5 P_{t-1} + b_6 + \varepsilon_2, \\ S_t = b_7 R_t + b_8 R_{t-1} + b_9 t + b_{10} + \varepsilon_3, \\ R_t = S_t + P_t + T_t, \\ R_t = C_t + I_t + G_t, \end{cases}$$

где C_t – потребление в период t,

S_t – заработная плата в период t,

P_t – прибыль в период t,

P_{t-1} – прибыль в предыдущий период $t-1$,

R_t – общий доход в период t,

R_{t-1} – общий доход в предыдущий период $t-1$,

t – время,

T_t – трансферты в период t,

I_t – капиталовложения в период t,

G_t – правительственные расходы в период времени t.

Модель содержит пять эндогенных переменных C_t, I_t, S_t, R_t и P_t, три экзогенные переменные T_t, G_t, t и две лаговые переменные P_{t-1} и R_{t-1}. Модель является сверхидентифицируемой и решается с помощью двухшагового метода наименьших квадратов.

Примеры решения типовых заданий

Пример 7.1. Дана расширенная модель формирования спроса и предложения:

$$\begin{cases} S_t = a_0 + a_1 P_t + a_2 P_{t-1} + \varepsilon_1 - \textit{уравнение предложения}, \\ D_t = b_0 + b_1 P_t + b_2 I_t + \varepsilon_2 - \textit{уравнение спроса}, \\ S_t = D_t - \textit{уравнение равновесия}, \end{cases}$$

где P_t – цена, P_{t-1} – цена в предыдущий момент времени, S_t – предложение товара, D_t – спрос на товар, I_t – доход.

Составить структурную и приведенную формы модели.

Решение:

Учитывая уравнение равновесия, перейдем от расширенной формы модели к модели:

$$\begin{cases} Q_t = a_0 + a_1 P_t + a_2 P_{t-1} + \varepsilon_1, \\ Q_t = b_0 + b_1 P_t + b_2 I_t + \varepsilon_2, \end{cases}$$

где Q_t – количество товара (производимого и потребляемого).

В данной модели эндогенными переменными (т.е. определяемыми внутри модели) являются переменные Q_t и P_t.

Предопределенными переменными являются экзогенная переменная I_t и лаговая эндогенная переменная P_{t-1}.

Поэтому структурная форма модели имеет вид:

$$\begin{cases} Q_t = a_0 + a_1 P_t + a_2 P_{t-1} + \varepsilon_1, \\ P_t = -\dfrac{b_0}{b_1} + \dfrac{1}{b_1} Q_t - \dfrac{b_2}{b_1} I_t + \dfrac{\varepsilon_2}{b_1}. \end{cases}$$

Приведенная форма содержит два уравнения (по числу эндогенных переменных модели). Каждое уравнение приведенной формы представляет собой зависимость эндогенной переменной от предопределенных переменных модели (дохода и цены в предыдущий период). В результате имеем приведенную форму:

$$\begin{cases} Q_t = c_0 + c_1 I_t + a_2 P_{t-1} + \varepsilon_3, \\ P_t = d_0 + d_1 I_t + d_2 P_{t-1} + \varepsilon_4. \end{cases}$$

Пример 7.2. Идентифицировать каждое уравнение системы и саму систему в целом

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + \varepsilon_2, \\ y_3 = b_{32}y_2 + a_{31}x_1 + a_{33}x_3 + \varepsilon_3. \end{cases}$$

Решение:

Для первого уравнения $H = 3$, $D = 1$. Так как $D < H - 1$, то ввиду необходимого и достаточного условия (таблица 7.1) уравнение является неидентифицируемым.

Для второго уравнения $H = 2$, $D = 2$, т.е. выполняется неравенство $D > H - 1$.

Кроме того, ранг матрицы, составленной из коэффициентов первого и третьего уравнений при переменных (эндогенных и экзогенных), отсутствующих во втором уравнении, равен двум. Следовательно, на основании необходимого и достаточного условия (таблица 7.1) второе уравнение сверхидентифицируемо.

Для третьего уравнения $H = 2$, $D = 1$, т.е. выполняется равенство $D = H - 1$.

Кроме того, ранг матрицы, составленной из коэффициентов первого и второго уравнений при переменных (эндогенных и экзогенных), отсутствующих в третьем уравнении, равен двум. Следовательно, на основании необходимого и достаточного условия (таблица 7.1) третье уравнение идентифицируемо.

Так как первое уравнение в системе не идентифицируемо, то вся модель является неидентифицируемой.

Пример 7.3. Идентифицировать следующую структурную модель:

$$\begin{cases} y_1 = b_{13}y_3 + a_{11}x_1 + a_{13}x_3 + \varepsilon_1, \\ y_2 = b_{21}y_1 + b_{23}y_3 + a_{22}x_2 + \varepsilon_2, \\ y_3 = b_{32}y_2 + a_{31}x_1 + a_{33}x_3 + \varepsilon_3. \end{cases}$$

Исходя из приведенной формы модели

$$\begin{cases} y_1 = 2x_1 + 4x_2 + 10x_3 + \varepsilon_1, \\ y_2 = 3x_1 - 6x_2 + 2x_3 + \varepsilon_2, \\ y_3 = -5x_1 + 8x_2 + 5x_3 + \varepsilon_3, \end{cases}$$

найти структурные коэффициенты.

Решение:

Модель имеет три эндогенные (y_1, y_2, y_3) и три экзогенные (x_1, x_2, x_3) переменные.

Проверим для каждого уравнения структурной модели выполнимость необходимого и достаточного условия идентификации, приведенного в таблице (7.1).

Первое уравнение содержит две эндогенные переменные y_1 и y_3; в нем отсутствует одна экзогенная переменная x_2. Значит, $H = 2$, $D = 1$ и выполняется равенство $D = H - 1$.

Построим матрицу A из коэффициентов при переменных y_2 и x_2 во втором и третьем уравнениях системы: $A = \begin{pmatrix} -1 & a_{22} \\ b_{32} & 0 \end{pmatrix}$. Так как $DetA = -1 \cdot 0 - b_{32} \cdot a_{22} \neq 0$, то ранг матрицы равен 2.

Следовательно, на основании необходимого и достаточного условия (таблица 7.1) первое уравнение идентифицируемо.

Второе уравнение содержит три эндогенные переменные y_1, y_2 и y_3; в нем отсутствуют две экзогенные переменные x_1 и x_3. Значит, $H = 3$, $D = 2$ и выполняется равенство $D = H - 1$.

Построим матрицу A из коэффициентов при переменных x_1 и x_3 в первом и третьем уравнениях системы: $A = \begin{pmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{pmatrix}$. Так как $DetA = a_{11} \cdot a_{33} - a_{31} \cdot a_{13} \neq 0$, то ранг матрицы равен 2.

Следовательно, на основании необходимого и достаточного условия (таблица 7.1) второе уравнение идентифицируемо.

Третье уравнение содержит две эндогенные переменные y_2 и y_3; в нем отсутствует одна экзогенная переменная x_2. Значит, $H = 2$, $D = 1$ и выполняется равенство $D = H - 1$.

Построим матрицу A из коэффициентов при переменные y_1 и x_2 в первом и втором уравнениях системы: $A = \begin{pmatrix} -1 & 0 \\ b_{21} & a_{22} \end{pmatrix}$. Так как $DetA = -1 \cdot a_{22} - b_{21} \cdot 0 \neq 0$, то ранг матрицы равен 2.

Следовательно, на основании необходимого и достаточного условия (таблица 7.1) третье уравнение идентифицируемо.

Таким образом, исследуемая система идентифицируема. Поэтому для ее решения применим косвенный метод наименьших квадратов: структурные коэффициенты модели с помощью алгебраических преобразований выразим через приведенные коэффициенты.

1. Из третьего уравнения приведенной формы выразим x_2 (так как его нет в первом уравнении структурной формы):

$$x_2 = \frac{y_3 + 5x_1 - 5x_3}{8}.$$

Данное выражение содержит переменные y_3, x_1 и x_3, которые нужны для первого уравнения структурной формы модели. Подставим полученное выражение x_2 в первое уравнение приведенной формы модели:

$$y_1 = 2 \cdot x_1 + 4 \cdot \frac{y_3 + 5x_1 - 5x_3}{8} + 10 \cdot x_3,$$
$$y_1 = 0{,}5y_3 + 4{,}5x_1 + 7{,}5x_3.$$

Получили первое уравнение структурной формы модели.

2. Во втором уравнении структурной формы модели нет переменных x_1 и x_3. Параметры второго уравнения структурной формы определим в два этапа.

На первом этапе выразим x_1 из первого уравнения приведенной формы модели:

$$x_1 = \frac{y_1 - 4x_2 - 10x_3}{2} = 0{,}5y_1 - 2x_2 - 5x_3.$$

Кроме того, выразим x_3 из третьего уравнения приведенной формы модели:

$$x_3 = \frac{y_3 + 5x_1 - 8x_2}{5}.$$

Подставим значение x_3 в выражение для x_1:

$$x_1 = 0{,}5y_1 - 2x_2 - 5\left(\frac{y_3 + 5x_1 - 8x_2}{5}\right) = 0{,}5y_1 - y_3 + 6x_2 - 5x_1,$$
$$x_1 = \frac{0{,}5y_1 - y_3 + 6x_2}{6}.$$

На втором этапе аналогично в выражение для x_3 подставим значение x_1, полученное из первого уравнения приведенной формы модели:

$$x_3 = \frac{y_3 + 5(0{,}5y_1 - 2x_2 - 5x_3) - 8x_2}{5} = 0{,}2y_3 + 0{,}5y_1 - 3{,}6x_2 - 5x_3.$$

Следовательно, $x_3 = 0{,}033y_3 + 0{,}083y_1 - 0{,}6x_2$.

Подставим теперь полученные значения x_1 и x_3 во второе уравнение приведенной формы модели:

$$y_2 = 3 \cdot \frac{0{,}5y_1 - y_3 + 6x_2}{6} - 6x_2 + 2 \cdot (0{,}033y_3 + 0{,}083y_1 - 0{,}6x_2),$$

$$y_2 = 0{,}416y_1 - 0{,}434y_3 - 4{,}2x_2.$$

Получили второе уравнение структурной формы модели.

3. Из второго уравнения приведенной формы модели выразим x_2:

$$x_2 = \frac{-y_2 + 3x_1 + 2x_3}{6} = -0{,}167y_2 + 0{,}5x_1 + 0{,}333x_3.$$

Подставим полученное выражение в третье уравнение приведенной формы модели:

$$y_3 = -5x_1 + 8(-0{,}167y_2 + 0{,}5x_1 + 0{,}333x_3) + 5x_3, \quad y_3 = -1{,}336y_2 - x_1 + 7{,}664x_3.$$

Получили третье уравнение структурной формы модели.

Таким образом, структурная форма модели имеет вид:

$$\begin{cases} y_1 = 0{,}5y_3 + 4{,}5x_1 + 7{,}5x_3 + \varepsilon_1, \\ y_2 = 0{,}416y_1 - 0{,}434y_3 - 4{,}2x_2 + \varepsilon_2, \\ y_3 = -1{,}336y_2 - x_1 + 7{,}664x_3 + \varepsilon_3. \end{cases}$$

Пример 7.4. Идентифицировать следующую структурную модель:

$$\begin{cases} y_1 = a_0 + a_1y_2 + a_2x_1 + \varepsilon_1, \\ y_2 = b_0 + b_1y_1 + b_2x_2 + \varepsilon_2. \end{cases}$$

На основании статистических данных, представленных в таблице 7.2, с помощью двухшагового метода наименьших квадратов найти структурные коэффициенты модели.

Таблица 7.2. Статистические данные примера 7.4

y_1	y_2	x_1	x_2
3,1	7,4	6,8	46,7

22,8	30,4	22,4	3,1
7,8	1,3	17,3	22,8
21,4	8,7	12,0	7,8
17,8	25,8	5,9	21,4
37,2	8,6	44,7	17,8
35,7	30,0	23,1	37,2
46,6	31,4	51,2	35,7
56,0	39,1	32,3	46,6

Решение:

Проверим уравнения структурной модели на идентифицируемость. Модель имеет две эндогенные y_1 и y_2 и две экзогенные x_1 и x_2 переменные.

Первое уравнение содержит две эндогенные переменные y_1 и y_2; в нем отсутствует одна экзогенная переменная x_2. Значит, $H=2$, $D=1$ и выполняется равенство $D=H-1$.

Построим матрицу A из коэффициентов при переменной x_2 во втором уравнении системы: $A=(b_2)$. Так как $DetA=b_2\neq 0$, то ранг матрицы A равен 1.

Следовательно, на основании необходимого и достаточного условия (таблица 7.1) первое уравнение системы идентифицируемо.

Второе уравнение содержит две эндогенные переменные y_1 и y_2; в нем отсутствует одна экзогенная переменная x_1. Значит, $H=2$, $D=1$ и выполняется равенство $D=H-1$.

Построим матрицу A из коэффициентов при переменной x_1 в первом уравнении системы: $A=(a_2)$. Так как $DetA=a_2\neq 0$, то ранг матрицы A равен 1.

Следовательно, на основании необходимого и достаточного условия (таблица 7.1) второе уравнение системы идентифицируемо.

Таким образом, исследуемая система является идентифицируемой.

Приведенная форма модели имеет вид:

$$\begin{cases} y_1 = c_0 + c_1x_1 + c_2x_2 + \varepsilon_3, \\ y_2 = d_0 + d_1x_1 + d_2x_2 + \varepsilon_4. \end{cases}$$

Приведенные коэффициенты вычислим обычным МНК с помощью инструмента «Регрессия» табличного процессора Excel (используя статистические данные таблицы 7.2):

$$\begin{cases} \tilde{y}_1 = 2{,}54 + 0{,}83x_1 + 0{,}19x_2, \\ \tilde{y}_2 = 9{,}1 + 0{,}26x_1 + 0{,}19x_2. \end{cases}$$

На основе уравнения $\tilde{y}_1 = 2{,}54 + 0{,}83x_1 + 0{,}19x_2$ найдем теоретические значения для эндогенной переменной y_1. Для этого подставим в уравнение значения переменных x_1 и x_2. Аналогично на основе уравнения регрессии $\tilde{y}_2 = 9{,}1 + 0{,}26x_1 + 0{,}19x_2$ найдем теоретические оценки для эндогенной переменной y_2. Результаты вычислений сведем в таблицу 7.3.

Таблица 7.3. Теоретические значения переменных y_1 и y_2

x_1	x_2	$\tilde{y}_1$	$\tilde{y}_2$
6,8	46,7	17,2	19,6
22,4	3,1	21,8	15,5
17,3	22,8	21,3	17,9
12,0	7,8	14,0	13,7
5,9	21,4	11,6	14,7
44,7	17,8	43,2	24,0
23,1	37,2	28,9	22,1
51,2	35,7	52,0	29,1
32,3	46,6	38,4	26,2

Наконец, используя статистические данные таблицы 7.3, с помощью инструмента «Регрессия» обычным МНК вычислим структурные коэффициенты.

В результате структурная форма модели имеет вид:

$$\begin{cases} y_1 = -6{,}93 + 1{,}03y_2 + 0{,}56x_1 + \varepsilon_1, \\ y_2 = 8{,}30 + 0{,}31y_1 + 0{,}13x_2 + \varepsilon_2. \end{cases}$$

Контрольные задания

Задание 7.1. Применив необходимое и достаточное условие идентификации для модели

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{33}x_3 + a_{24}x_4 + \varepsilon_3, \end{cases}$$

определить, идентифицируемо ли каждое из уравнений модели. Определить метод оценки параметров модели. Записать приведенную форму модели.

Задание 7.2. Применив необходимое и достаточное условие идентификации для следующей макроэкономической модели (упрощенной версии модели Клейна):

$$\begin{cases} C_t = a_1 + b_{11}Y_t + b_{12}T_t + \varepsilon_1, \\ I_t = a_2 + b_{21}Y_t + b_{22}K_{t-1} + \varepsilon_2, \\ Y_t = C_t + I_t, \end{cases}$$

где C – потребление; I – инвестиции; Y – доход; T – налоги; K – запас капитала; t – текущий период; t-1 – предыдущий период, определить, идентифицируемо ли каждое из уравнений модели.

Определить метод оценки параметров модели. Записать приведенную форму модели.

Задание 7.3. Предложение и спрос на рынке характеризуется следующей моделью:

$$\begin{cases} q_1 = a_1 + b_1 p + \varepsilon_1, \\ q_2 = a_2 + b_2 p + \varepsilon_2, \end{cases}$$

где q_1 – спрос на товар; q_2 – предложение количества товара; p – цена, по которой заключаются сделки.

Применив необходимое и достаточное условие идентифицируемости для этой модели, определить, идентифицируемо ли каждое из уравнений модели. Определить метод оценки параметров модели. Записать приведенную форму модели.

Задание 7.4. Проверить, является ли идентифицируемой эконометрическая модель

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{33}x_3 + a_{24}x_4 + \varepsilon_3. \end{cases}$$

Задание 7.5. Используя статистические данные таблицы 7.4, построить эконометрическую модель вида

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

Таблица 7.4. Статистические данные задания 7.5

y_1	y_2	x_1	x_2
2	5	1	3
3	6	2	1
4	7	3	2
5	8	2	5
6	5	4	6

Задание 7.6. Исследуется зависимость спроса D_t и предложения S_t некоторого товара от его цены P_t, дохода I_t в текущий период, дохода I_{t-1} в предыдущий период и процентной ставки R_t:

$$\begin{cases} D_t = b_0 + b_1 P_t + b_2 I_t + b_3 I_{t-1} + \varepsilon_1, \\ S_t = a_0 + a_1 P_t + a_2 R_t + \varepsilon_2, \\ S_t = D_t = Q_t. \end{cases}$$

На основаниии статистических данных таблицы 7.5 за восемь периодов:

1) применив необходимое и достаточное условие идентифицируемости, определить, идентифицируемо ли каждое из уравнений модели;

2) определить метод оценки параметров модели;

3) записать приведенную форму модели;

4) рассчитать параметры структурной модели.

Таблица 7.5. Статистические данные задания 7.6

Год	Q_t	P_t	R_t	I_t	I_{t-1}
1	40	6	3,0	13	6
2	45	6	3,0	15	6
3	40	5	2,0	15	5
4	50	8	3,5	18	8
5	35	5	2,5	20	5
6	45	9	4,0	18	9
7	50	10	3,5	22	10
8	45	9	3,5	21	9

Задание 7.7. Определить (если это возможно) структурные коэффициенты модели

$$\begin{cases} y_1 = a_1 + b_{11}x_1 + b_{12}x_2 + c_{12}y_2 + \varepsilon_1 \\ y_2 = a_2 + b_{22}x_2 + b_{23}x_3 + c_{21}y_1 + \varepsilon_2, \\ y_3 = a_3 + b_{31}x_1 + b_{33}x_3 + c_{31}y_1 + \varepsilon_3 \end{cases}$$

по ее приведенной форме

$$\begin{cases} y_1 = 6 + 8x_1 + 10x_2 + 4x_3 + \delta_1, \\ y_2 = 16 - 12x_1 - 70x_2 + 8x_3 + \delta_2, \\ y_3 = 10 - 5x_1 - 22x_2 + 5x_3 + \delta_3. \end{cases}$$

Задание 7.8. Дана модель денежного и товарного рынков:

$$\begin{cases} R_t = a_1 + b_{12}Y_t + b_{14}M_t + \varepsilon_1, \\ Y_t = a_2 + b_{21}R_t + b_{23}I_t + b_{25}G_t + \varepsilon_2, \\ I_t = a_3 + b_{31}R_t + \varepsilon_3, \end{cases}$$

где R – процентные ставки; Y – реальный ВВП; M – денежная масса; I – внутренние инвестиции; G – реальные государственные расходы.

Необходимо:

1) применив необходимое и достаточное условие идентификации, определить, идентифицируемо ли каждое из уравнений модели;

2) определить метод оценки параметров модели;

3) записать в общем виде приведенную форму модели.

Задание 7.9. Дана модель денежного рынка:

$$\begin{cases} R_t = a_1 + b_{11}M_t + b_{12}Y_t + +\varepsilon_1, \\ Y_t = a_2 + b_{21}R_t + b_{22}I_t + \varepsilon_2, \\ I_t = a_3 + b_{33}R_t + \varepsilon_3, \end{cases}$$

где R – процентные ставки; Y – ВВП; M – денежная масса; I – внутренние инвестиции.

Необходимо:

1) применив необходимое и достаточное условие идентификации, определить, идентифицируемо ли каждое из уравнений модели;

2) определить метод оценки параметров модели;

3) записать в общем виде приведенную форму модели.

Задание 7.10. Проверить, на идентифицируемость следующую модель

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + a_{11}x_1 + a_{12}x_2 + \varepsilon_1, \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{33}x_3 + a_{34}x_4 + \varepsilon_3. \end{cases}$$

Контрольные вопросы

1. Дайте определение системы одновременных уравнений.
2. Каковы основные причины использования одновременных уравнений в экономике?
3. Как классифицируются системы одновременных уравнений?
4. Как определяется система независимых уравнений?
5. Каким способом осуществляется оценка параметров системы независимых уравнений?
6. Как определяется система рекурсивных уравнений?
7. Каким способом осуществляется оценка параметров системы рекурсивных уравнений?
8. Как определяется система взаимосвязанных уравнений?
9. Почему обычный МНК практически не используется для оценки систем одновременных уравнений?
10. Как определяются структурная и приведенная формы модели? В чем состоит их основное различие?
11. Как связаны между собой структурная и приведенная формы модели?
12. В чем различие между эндогенными и экзогенными переменными системы уравнений?
13. В чем состоит проблема идентифицируемости модели?
14. Приведите причины неидентифицируемости и сверхидентифицируемости модели.
15. Приведите необходимое условие идентифицируемости модели.
16. Приведите достаточное условие идентифицируемости модели.
17. Приведите необходимые и достаточные условия идентифицируемости модели.
18. Какими способами можно устранить проблему неидентифицируемости модели?
19. В чем состоит сущность косвенного МНК?
20. Для решения каких систем используется КМНК?
21. В чем состоит сущность двухшагового МНК?
22. Для решения каких систем используется ДМНК?
23. Приведите пример статической модели Кейнса.
24. Приведите примеры динамических моделей Кейнса и Клейна.

Тестовые задания

Выберите правильные ответы из предложенных вариантов:

1. Система

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ ..... \\ y_n = a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases}$$

является:

а) системой независимых уравнений;
б) системой рекурсивных уравнений;
в) системой взаимозависимых уравнений.
г) системой нелинейных уравнений.

2. Система

$$\begin{cases} y_1 = & + a_{11}x_1 + a_{12}x_2 + ... + a_{1m}x_m + \varepsilon_1, \\ y_2 = b_{21}y_1 + & + a_{21}x_1 + a_{22}x_2 + ... + a_{2m}x_m + \varepsilon_2, \\ y_3 = b_{31}y_1 + b_{32}y_2 + & + a_{31}x_1 + a_{32}x_2 + ... + a_{3m}x_m + \varepsilon_3, \\ ..... \\ y_n = b_{n1}y_1 + b_{n2}y_2 + ... + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + ... + a_{nm}x_m + \varepsilon_n. \end{cases}$$

является:

а) системой независимых уравнений;
б) системой рекурсивных уравнений;
в) системой взаимозависимых уравнений;
г) системой нелинейных уравнений.

3. Уравнение системы идентифицируемо, если:

а) $D = H$;
б) $D = H - 1$;
в) $D > H - 1$;
г) $D < H - 1$.

4. В системе одновременных уравнений переменные, взятые в предыдущий момент времени, называются:

а) лаговыми;
б) экзогенными;
в) фиктивными;
г) предопределенными.

5. Для параметризации сверхидентифицируемых структурных моделей используется:

а) обычный МНК;

б) косвенный МНК;

в) двухшаговый МНК;

г) взвешенный МНК.

6. Эндогенные переменные – это:

а) предопределенные переменные, влияющие на зависимые переменные, но не зависящие от них;

б) переменные, значения которых определяется внутри модели;

в) значения зависимых переменных за предшествующий период времени;

г) случайные переменные.

7. Экзогенные переменные – это:

а) переменные, значения которых определяется вне модели;

б) зависимые переменные, число которых равно числу уравнений в системе;

в) случайные переменные.

8. Определите, какие из переменных данной системы

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

являются эндогенными:

а) x_1 и x_2; б) y_1 и y_2; в) ε_1 и ε_2.

9. Определите, какие из переменных данной системы

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

являются экзогенными:

а) x_1 и x_2; б) y_1 и y_2; в) ε_1 и ε_2.

10. Лаговые переменные – это:

а) предопределенные переменные, влияющие на зависимые переменные, но не зависящие от них;

б) зависимые переменные, число которых равно числу уравнений в системе;

в) переменные, влияние которых в модели характеризуется некоторым запаздыванием.

11. Для определения параметров структурную форму модели необходимо преобразовать в:

а) приведенную форму модели;

б) рекурсивную форму модели;

в) независимую форму модели;

г) расширенную форму модели.

12. Модель идентифицируема, если:

а) число приведенных коэффициентов меньше числа структурных коэффициентов;

б) если число приведенных коэффициентов больше числа структурных коэффициентов;

в) если число параметров структурной модели равно числу параметров приведенной формы модели.

13. Модель неидентифицируема, если:

а) число приведенных коэффициентов меньше числа структурных коэффициентов;

б) если число приведенных коэффициентов больше числа структурных коэффициентов;

в) если число параметров структурной модели равно числу параметров приведенной формы модели.

14. Модель сверхидентифицируема, если:

а) число приведенных коэффициентов меньше числа структурных коэффициентов;

б) если число приведенных коэффициентов больше числа структурных коэффициентов;

в) если число параметров структурной модели равно числу параметров приведенной формы модели.

15. Уравнение неидентифицируемо, если:

а) $D = H - 1$;

б) $D > H - 1$;

в) $D < H - 1$.

16. Для определения параметров идентифицируемой модели:

а) применяется двушаговый МНК;

б) применяется косвенный МНК;

в) ни один из существующих методов применить нельзя.

17. Система вида

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

является:

а) структурной формой;

б) расширенной формой;

в) приведенной формой.

18. Для определения параметров неидентифицируемой модели:

а) применяется двушаговый МНК;

б) применяется косвенный МНК;

в) ни один из существующих методов применить нельзя.

19. Выделяют три класса систем эконометрических уравнений:

а) система независимых уравнений, системы изолированных уравнений и системы рекурсивных уравнений;

б) системы взаимозависимых уравнений, системы возвратных уравнений и системы рекурсивных уравнений;

в) системы взаимозависимых уравнений, системы нелинейных уравнений и системы рекурсивных уравнений;

г) система независимых уравнений, системы взаимозависимых уравнений и системы рекурсивных уравнений.

20. При оценке параметров приведенной формы модели косвенный метод наименьших квадратов использует алгоритм:

а) расчета средней взвешенной величины;

б) метода главных компонент;

в) метода максимального правдоподобия;

г) обычного метода наименьших квадратов.

21. Система вида

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

имеет:

а) структурную форму;

б) расширенную форму;

б) приведенную форму.

Ответы тестовых заданий

Номер задания	Ответы	Номер задания	Ответы	Номер задания	Ответы

1	а)	8	б)	15	в)
2	в)	9	а)	16	а), б)
3	б)	10	в)	17	в)
4	а)	11	а)	18	в)
5	в)	12	в)	19	г)
6	б)	13	а)	20	г)
7	а)	14	б)	21	а)

Литература

1. Экономико-математические методы и модели: учебное пособие / Под общей редакцией А.В. Кузнецова. Мн.: БГЭУ, 2000.

2. Елисеева И.И. Эконометрика: учебник. М.: Финансы и статистика, 2008.

3. Бородич С.А. Эконометрика: учебное пособие. Мн.: Новое знание, 2001.

4. Кремер Н.Ш. Эконометрика: учебник для студентов вузов. М.: ЮНИТИ-ДАНА, 2008.

5. Красс М.С., Чупрынов Б.П. Математика для экономистов. СПб.: Питер, 2008.

6. Замков О.О., Толстопятенко А.В., Черемных Ю.Н. Математические методы в экономике: учебник. М.: Дело и Сервис, 2009.

7. Практикум по эконометрике: учебное пособие / Под редакцией И.И. Елисеевой. М.: Финансы и статистика, 2003.

8. Юферева О.Д. Экономико-математические методы и модели: сборник задач. Минск.: БГЭУ, 2002.

9. Доугерти К. Введение в эконометрику. М.: Финансы и статистика, 1999.

10. Айвазян С.А., Иванова С.С. Эконометрика: учебное пособие. М.: Маркет ДС, 2007.

11. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика / Начальный курс: учебник. М.: Дело, 2004.

12. Четыркин Е.М. Статистические методы прогнозирования. М.: Статистика, 1978.

13. Тихомиров Н.П., Дорохина Е.Ю. Эконометрика: учебник. М.: Экзамен, 2003.

14. Хацкевич Г.А., Гедранович А.Б. Эконометрика: учебно-методический комплекс. Мн.: МИУ, 2007.

15. Новиков А.И. Эконометрика: учебное пособие. М.: ИНФА-М, 2008.

16. Калмыкова Т.Ф., Моисеева Т.М. Эконометрика: пособие. Гомель, БТЭУПК, 2008.
17. Чудаков А.Д. Логистика: учебно-практическое пособие. М.: Альфа-Пресс, 2008.
18. Афанасьев В.Н., Юзбашев М.М. Анализ временных рядов и прогнозирование: учебник. М.: Финансы и статистика, 2001.
19. Кремер Н.Ш. Теория вероятностей и математическая статистика: учебник. М.: ЮНИТИ-ДАНА, 2000.