Министерство образования и науки Российской Федерации

Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Оренбургский государственный университет»

ЭКОНОМЕТРИКА

под редакцией
профессора В.Н. Афанасьева

Рекомендовано Ученым советом федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Оренбургский государственный университет» в качестве учебника для студентов, обучающихся по программам высшего профессионального образования по направлению подготовки 080100.62 Экономика

Оренбург
2012

УДК 330.4 (075.8)
ББК 65в631я73
Э-94

Рецензенты
профессор, доктор экономических наук, заведующий кафедрой математической статистики и эконометрики МЭСИ В.С. Мхитарян
профессор, доктор экономических наук, заведующий кафедрой статистики Новосибирского государственного университета экономики и управления В.В. Глинский

Авторы: В.Н. Афанасьев, Т.В. Леушина, Т.В. Лебедева, А.П. Цыпин

Эконометрика: учебник / В.Н. Афанасьев, Т.В. Леушина, Т.В. Лебедева,
Э-94 А.П. Цыпин; под ред. проф. В.Н. Афанасьева; Оренбургский гос. ун-т. – Оренбург: ОГУ, 2012. – 402 с.: ил.
ISBN

Учебник содержит 11 глав, включающих основы эконометрики: парную и множественную регрессии, нелинейные модели, модели с фиктивными переменными, моделирование одномерного временного ряда, динамические эконометрические модели, методы измерения корреляции и регрессии во временных рядах. В каждой главе даются вопросы для самоконтроля и тесты.

Для студентов очной и заочной форм обучения, по направлению 080100.62 Экономика, соответственно всех профилей бакалавров.

УДК 330.4 (075.8)
ББК 65в631я73

ISBN

Содержание

Введение

Что такое эконометрика?

Термин «эконометрика» имеет в своей основе два слова: «экономика» и «метрика» (от гр. *metron* — «метод расчета определения расстояния между двумя точками в пространстве»).

Эконометрика **—** это наука, которая на основе выявленных статистических закономерностей количественно характеризует взаимозависимые экономические явления и процессы. В общем случае эконометрику можно определить как науку об экономических измерениях для целей управления, регулирования этих явлений и процессов. Предмет исследования эконометрики — это массовые экономические явления и процессы.

Цель эконометрики — это количественная характеристика экономических закономерностей, выявляемых экономической теорией и статистикой. Знание эконометрики необходимо, прежде всего, тем специалистам, которые занимают соответствующие должности системы управления в масштабах, как отдельного предприятия, фирмы, так и региона, отрасли и экономики страны в целом.

Экономические закономерности выражаются в массовых, а не в единичных фактах - актах продажи и приобретения товаров и услуг, их использования в процессах производства и потребления, взаимодействия и взаимосвязи между предприятиями, работниками, населением и с окружающей природной средой. Массовость этих явлений требует применения статистических методов исследования и статистических показателей, характеризующих массовые варьирующие явления и связи между ними, - средних величин, характеристик распределения, корреляции, динамики и структуры массовых явлений. Поэтому основой методов эконометрики служит статистика, изучающая и разрабатывающая общие методы исследования массовых явлений и процессов, независимо от их материального содержания. Но поскольку экономические явления имеют свои качественные особенности, эко-

нометрика должна их учитывать и приспосабливать общие методы статистики к своему специфическому предмету. Например, экономические явления не могут быть, как это делается в других науках, экспериментально воспроизведены по желанию исследователя заново или в специальных условиях, исключающих случайные помехи.

Обрабатываемые эконометрикой данные - это, как правило, невоспроизводимые заново данные бухгалтерского учета и статистической отчетности предприятий, фирм, банков, их сводки по отраслям и регионам страны. Наряду с закономерностями экономических процессов эти данные неизбежно заключают в себе элементы случайных отклонений. Например, известно, что применение новых технологий закономерно увеличивает производительность труда, что, в свою очередь, приводит к улучшению результатов хозяйственной деятельности предприятий. Однако реальная жизнь неизбежно сложнее, и в условиях роста себестоимости продукции и услуг по новым технологиям в сравнении с себестоимостью одноименных товаров, отдельные предприятия, отказавшись от новых технологий (например, при монополии производства или низкой платежеспособностью покупателей) могут и при более низкой производительности труда получить лучший финансовый результат (большую прибыль или меньший убыток), чем предприятие, использующее новые технологии.

Сложность, многофакторный характер связей и зависимостей явлений в экономике приводят к тому, что измеряемые эконометрикой количественные характеристики (показатели) имеют вероятный характер, т. е. являются не абсолютно достоверными, а лишь с некоторой вероятностью, желательно, достаточно близкой к единице. Отсюда вытекает вывод о том, что методы эконометрики должны включать этап проверки полученных выводов на степень их надежности с помощью специальных статистических критериев (*t*-критерий Стьюдента, *F*-критерий Фишера, критерий Дарбина - Уотсона и др.).

Как было упомянуто ранее, одной из целей современной эконометрики является моделирование изучаемых процессов, взаимосвязей и тенденций развития. Эконометрические модели - это, как правило, математико-статистические выраже-

ния (формулы, уравнения, графические образы), характеризующие количественно те или иные закономерности экономических явлений и процессов, взаимосвязей, тенденций развития в конкретных условиях места и времени, обладающие достаточно высокой надежностью и пригодные для анализа и прогнозирования отображаемых явлений и процессов. В этом определении заключены основные требования (принципы), которые должны быть свойственны эконометрической модели:

1) соответствие эконометрической модели общим законам экономики;

2) конкретность, учет реальных условий (типа хозяйств, природной зоны, этапа развития);

3) достаточно высокая вероятность, скажем 0,9 или 0,95 того, что показатели моделируемого признака, прогнозируемые на основе модели, не окажутся вне указанных доверительных границ (или что ошибка предсказания по модели не превысит заданной величины).

Связь эконометрики с экономической теорией, математикой и другими дисциплинами

Все составляющие эконометрики — экономическая, математико-статистическая, информационная - тесно связаны. Но первенство все же следует отдать экономической сущности решаемой задачи. Без ясного понимания экономического содержания моделируемого показателя невозможно построить хорошую модель и правильно интерпретировать динамику. Например, в состав продукции предприятий топливно-энергетического комплекса включается сумма стоимости произведенной электроэнергии (в том числе и стоимость потребленной этим комплексом), стоимость произведенных нефти и газа, потребляемых для выработки электроэнергии, т. е. существует двойной счет стоимости электроэнергии и сырья для ее производства. В то же время в составе товарной и реализованной продукции стоимость произведенной и потребленной электроэнергии, сырья будет учтена не дважды, а лишь один раз. Поэтому при раздельном исследовании динамических, трендовых

моделей объема реализованной продукции предприятий производящих сырье для ТЭЦ и электроэнергию сумма стоимостей реализованной продукции возрастет, а в целом по ТЭК страны будет неверной.

Из первой роли экономического содержания, модели вытекает, что в случае противоречивости экономической и математико-статистической оценки роли того или иного фактора в модели следует предпочесть первую и допустить не очень значительную погрешность в математической оценке. Пусть в результате оценки надежности установления влияния фактора «энергообеспеченность», на выход продукции предприятий топливно-энергетического комплекса оказалось, что критерий Стьюдента ниже требуемого для надежности влияния фактора на уровне 0,95 (т. е. вероятность нулевой гипотезы - об отсутствии влияния больше 0,05, скажем 0,08 или даже 0,11). Технологически и экономически фактор «энергообеспеченность» в ТЭК весьма важен, и будет правильнее поступиться математической нормой и включить данный фактор в модель. Ведь все же 92 шанса из 100 (или 89 из 100) говорят за то что, даже по имеющимся данным фактор влияет на результат, и было бы весьма формальным подходом ориентироваться исключительно на величину вероятности «нулевой гипотезы».

Положения эконометрики - это не строго подлежащая исполнению «инструкция», наподобие инструкций налоговой инспекции или инструкции Минфина по ведению бухгалтерского учета, а лишь общие указания о путях и методах возможных решений практических задач. Изучившему их специалисту, менеджеру предприятия самому придется решать, насколько в реальной обстановке применима та или иная методика, насколько надежна и полна имеющаяся информация, какими требованиями «чистой» науки можно поступиться, а какими нельзя пренебречь ни в коем случае. Иногда, в неблагоприятных условиях, лучше вовсе отказаться от того или иного метода, чем получить сомнительные результаты, дискредитирующие науку в глазах практиков. Различие между «чистой» и «прикладной» наукой в том, что первая решает так, как нужно, то, что можно решить, строго соблюдая требования теории, а вторая решает то, что нужно, так, как можно, т. е. допуская отступления от чистой теории. Учитывая далеко не блестящее состояние информационной базы, немного

нашлось бы реальных задач, моделей, которые можно было построить и применять, если строго соблюдать все принципы математической статистики.

Излагаемые в учебниках эконометрики методы не исчерпывают всех путей анализа количественных связей и зависимостей в производстве товаров и услуг. Существуют и другие методы количественного анализа и моделирования, как, например, метод индексов, метод математико-статистической оптимизации плановых решений, основанные на линейной алгебре. В каком соотношении они находятся с эконометрикой?

Методы индексов применяются к системам признаков, связанных строго функциональной, жесткой зависимостью. Такие системы образуются «по определению»: если цена как признак определяется отношением выручки от реализации к объему реализованной продукции, то для любого предприятия выручка строго равна произведению реализованной продукции на цену. Аналогично рентабельность реализации определенного вида продукции равна частному от деления разности между средней ценой реализации единицы продукции, и ее себестоимостью, на эту себестоимость.

Как правило, разложение результативного признака на жестко связанные элементы методом индексов является первой стадией моделирования, которая, однако, не включается в предмет эконометрики. Второй стадией анализа и моделирования будет исследование связи каждого из жестко связанных элементов с реальными не по определению, а в силу социально-экономических свойств связанными факторами: с условиями производства и т.д., а цены - спросом и предложением, качеством продукции и услуг, себестоимостью и т.д.

Именно эти зависимости «второго порядка» и составляют предмет эконометрики, потому что они имеют статистический характер, проявляются в большой совокупности случаев, в разной, варьирующей, степени, измеряются с определенной вероятностью.

Взаимодействие эконометрических методов с методами оптимизации, основанными на линейной алгебре, состоит в том, что оптимальное с точки зрения заданного критерия решение достигается при наличии заданных или прогнозируемых

значений технологических, экономических и природных факторов. Для получения этих прогнозов используются эконометрические модели, играющие роль поставщика необходимых исходных данных для решения оптимизационной задачи. Иногда сами эконометрические модели можно применить для оптимизации значений фактора, не прибегая к методам линейной алгебры. Например, если зависимость производительности физического труда рабочего от его возраста имеет параболический характер, то оптимальную производительность труда можно вычислить, найдя максимум этой параболы. Оптимальные значения факторов производства, найденные (вычисленные) методами линейного программирования, могут, в свою очередь, использоваться в эконометрической модели результативного показателя экономики. Потому-то студент изучает различные методы исследования и управления экономикой, чтобы комплексно их использовать в деятельности руководителя предприятия, фирмы, экономиста-аналитика, специалиста статистических, административно-управленческих органов.

Дальнейшее развитие эконометрики, по мнению видного российского ученого в данной области науки С.А. Айвазяна идет, прежде всего, по пути углубления экономико-теоретического анализа содержательной сущности решаемых задач. Именно такой анализ должен предшествовать и обосновывать выбор математико-статистических методик решения, типа уравнений регрессии или трендовых моделей.

Следует также отметить, что происходящий в XXI веке все ускоряющийся процесс глобализации экономики, вступление РФ в ВТО, т. е. растущая тесная связь между развитием экономической конъюнктуры разных государств, приводит к тому, что в модели макроэкономических показателей страны следует включать факторы не только «свои», но и мирового рынка. Например, от мировых цен на нефть и газ зависят такие экономические показатели России, как ожидаемые темпы прироста валового внутреннего продукта на предстоящий год, а, следовательно, и темпы роста душевого дохода граждан, и другие важные параметры развития.

Авторы имеют основание полагать, что подготовленный ими учебник может быть полезен не только для бакалавров различных профилей направления «Эконо-

мика», но и ряда других направлений высших и средних учебных заведений. Нужны разные учебники, рассчитанные на годичный или полуторагодичный курс эконометрики для управленцев различных сфер деятельности, работников страховых компаний, Центробанка и других учреждений, имеющих дело с макроэкономическим моделированием и прогнозированием, и мы надеемся, что и это издание встанет в ряд востребованных. Более сложные методы и модели в эконометрических исследованиях будут нами освещены в учебнике для магистерских программ направления «Экономика».

Представленный учебник «Эконометрика» является переработанной и дополненной версией одноименного учебника «Эконометрика» 2006 года издания, издательства «Финансы и статистика», под редакцией профессора В.Н. Афанасьева. Учебник соответствует новому образовательному стандарту для подготовки бакалавров различных профилей направления «Экономика» и рассчитан на существующий уровень математической подготовки студентов.

Главы 1, 2, 3, 4, 5, 6, 7 подготовлены доцентом Т.В. Леушиной, главы 8, 9, 10, 11 – доцентом Т.В. Лебедевой, в работе над главами 4, 6, принимал участие доцент А.П. Цыпин. Общая редакция всех глав - профессора В.Н. Афанасьева.

Выражаем глубокую признательность доктору экономических наук, профессору В.С. Мхитаряну и доктору экономических наук, профессору В.В. Глинскому за ценные замечания, сделанные ими при рецензировании рукописи настоящего учебника.

Авторы будут благодарны всем, кто пожелает высказать свои предложения по улучшению учебника для бакалавров. Предложения просим присылать по адресу: 460018, г. Оренбург, проспект Победы, 13, кафедра статистики и эконометрики Оренбургского государственного университета.

Доктор экономических наук, профессор В.Н. Афанасьев

1 Анализ рядов распределения

Что необходимо знать из 1 главы:

1. Понятие, цель и правила составления рядов распределения.
2. Классификация рядов распределения, их отличительные особенности.
3. Методы анализа неравномерности распределения единиц совокупности в различных рядах распределения.

1.1 Понятие и виды рядов распределения

С целью, выявления характера распределения единиц совокупности по группирующему признаку и определения закономерности в полученном распределении, строят специальные таблицы, где единицы изучаемой совокупности упорядочены по величине изучаемого (количественного) признака. Такие таблицы носят название *рядов распределения*.

Ряд распределения – это первичная характеристика массовой статистической совокупности, в которой находят количественное выражение *закономерности вариации* массовых явлений и процессов общественной жизни. Ряды распределения дают возможность судить о закономерности распределения и о границах варьирования совокупности. Различные обобщающие показатели – средние, мода, медиана, дисперсия и т.д. исчисляются на основе ряда распределения.

Ряд распределения, образующийся в результате группировки единиц наблюдения по значению варьирующего признака, является наиболее фундаментальной характеристикой совокупности. Он дает наиболее полное представление о результатах действия и взаимодействия всех факторов явления (основных и случайных), о сложившейся под их влиянием закономерности ряда распределения, о свойственных явлению индивидуальных чертах и особенностях. Изучение ряда распределения позволяет установить связь единичного и массового, частного и общего, случайного и закономерного [1, с. 247].

По дискретному признаку, имеющему ограниченное число вариантов, ряд распределения составляется так: из совокупности имеющихся данных выбираются все варианты и записываются в ряд в порядке их возрастания или убывания. Одновременно подсчитывается повторяемость (частота) каждого варианта в данной совокупности, которая записывается напротив или рядом с вариантом. В таком случае будет построен *ранжированный ряд.* Наряду с частотами могут быть вычислены частости путем деления частоты каждого варианта на сумму всех частот ряда. Соответственно, сумма частостей равна 1 или 100 %.

Примером ранжированного дискретного ряда может служить распределение женщин по числу рожденных детей, полученное в ходе проведенного Росстатом в 2009 г. выборочного обследования «Семья и рождаемость» (таблица 1.1).

Таблица 1.1 – Распределение женщин по числу рожденных детей

Число рожденных детей	Доля женщин, %
0	9,4
1	58,3
2	27,8
3	3,8
4	0,6
5	0,1
Итого	100,0
Источник: http://www.gks.ru	

Если ряд распределения составляется по дискретному признаку с большим числом вариантов или по непрерывно изменяющемуся признаку, то варианты объединяются в интервалы методом группировки единиц совокупности. Если вариация признака слабая или умеренная, то применяется *равноинтервальный ряд распределения*. Равноинтервальный ряд используется при достаточно однородной совокупности и близости распределения к нормальному. В курсах теории статистики излагается группировка на *n* групп с равными интервалами с применением формулы *Стерджесса* для определения числа групп и длины интервала.

Как правило, слабая и умеренная вариация наблюдается у вторичных, качественных признаков: доли экономически активного населения, себестоимости единицы продукции, распределении домашних хозяйств по числу детей и т.п. Слабая вариация может также наблюдаться и у некоторых первичных признаков, например, в росте и весе лиц определенного пола и возраста.

Пример равноинтервального ряда распределения дается в таблице 1.2.

Таблица 1.2 – Распределение населения Российской Федерации по возрастным группам на начало 2010 г.

Возрастные группы, лет	Численность населения, тыс. человек	Возрастные группы, лет	Численность населения, тыс. человек
0-4	7956	40-44	9193
5-9	6881	45-49	11247
10-14	6564	50-54	11261
15-19	8496	55-59	9748
20-24	12256	60-64	6897
25-29	12257	65-69	4479
30-34	10799	70 и более	13811
35-39	10069	Все население	141914
Источник: http://www.gks.ru			

По равноинтервальному ряду легче определить модальную величину признака, но требуются достаточно сложные формулы для вычисления квантилей признака. Графическое изображение равноинтервального ряда в форме гистограммы позволяет определить модальное (наиболее часто встречающееся значение признака) без вычислений (рисунок 1.1).

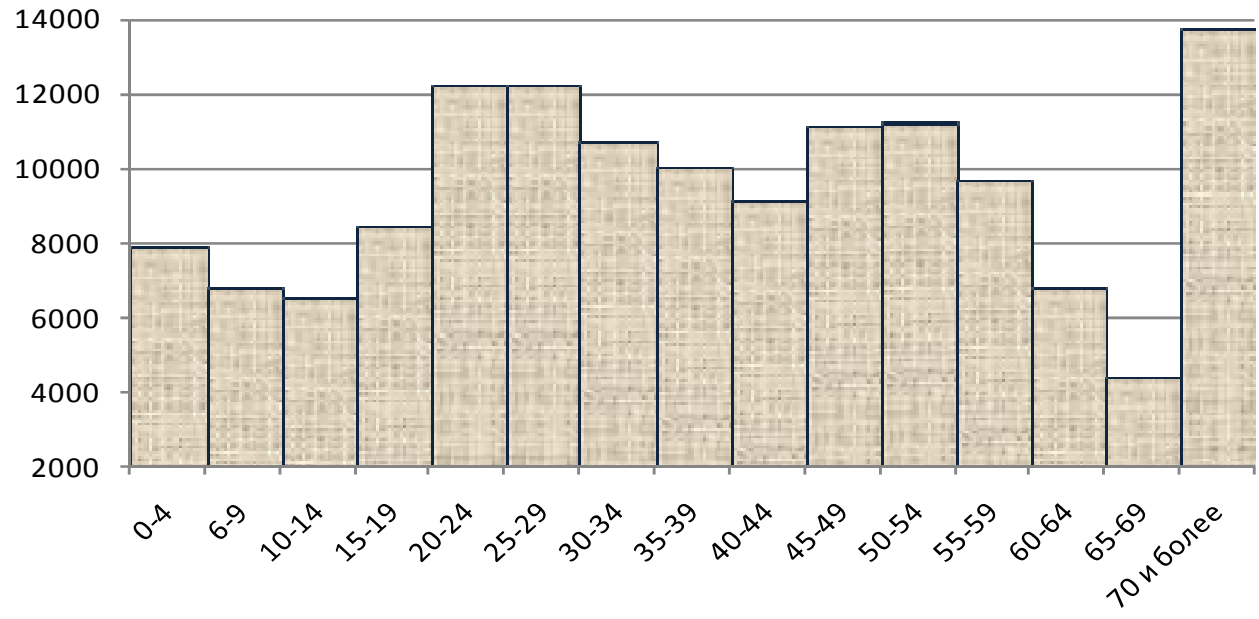


Рисунок 1.1 – Распределение населения РФ по возрасту
на начало 2010 г., тыс. человек

Гистограмма отражает, что на начало 2010 г. в России по 5-летним возрастным группам наибольшую долю занимало население 70-ти лет и старше, т.е. налицо демографическое старение населения страны.

В случаях, когда вариация признака достаточно сильная, разброс значений значителен, применение формулы Стерджесса не дает хороших результатов, так как значительное число единиц совокупности может быть сосредоточено в одном-двух интервалах, а другие группы могут быть малонаполненными или вообще не содержать ни одной единицы совокупности.

Так произойдет, например, при распределении банков по величине активов, распределении домашних хозяйств по величине среднедушевого дохода, распределении организаций по объему выпускаемой продукции и т.д. Распределение признаков, обладающих сильной вариацией, *не может* изучаться с помощью равноинтервальных рядов.

В таких случаях приемлемое распределение с сохранением информации об «аномальных» значениях можно получить, построив *равночастотный ряд*. Схема построения такого ряда следующая: совокупность, ранжированная по значениям

изучаемого признака, делится на *n* групп равной численности или равной частости, например, на 10 групп по 10 % единиц совокупности в каждой (децильные группы), либо на 20 групп по 5 %, или на 5 групп по 20 %. В последнем случае распределение признака будет отражено хуже.

В качестве примера равночастотного ряда представлено распределение 100 банков России по величине активов-нетто (таблица 1.3).

Таблица 1.3 – Распределение 100 банков РФ по величине активов-нетто (по состоянию на 1.08.2012 г.)

Группы банков по величине активов-нетто, млн. р.	Число банков	Величина активов-нетто, млн. р.	Процент от величины активов-нетто	Плотность распределения
40022,8-47986,7	10	44004,771	0,56	0,12557
48625,1-53761,1	10	51193,059	0,66	0,19470
55832,7-71201,9	10	63517,284	0,81	0,06506
73258,2-83676,3	10	78467,288	1,01	0,09599
84615,3-100479,1	10	93547,216	1,20	0,06304
102537,5-121894,2	10	112215,846	1,44	0,05166
123475,5-200105,7	10	161790,618	2,07	0,01305
201135,6-267632,1	10	234383,806	3,00	0,01504
271371,5-532897,8	10	452134,619	5,80	0,00277
682668,0-12336754,7*	10	6509711,358	83,45	0,00009
И т о г о	100	7800965,865	100	–
*Сбербанк России				
Источник: http://www.banki.ru . Расчеты автора.				

Границы интервалов в данном распределении равны фактическим величинам активов-нетто первого, десятого, одиннадцатого, двадцатого и так далее банков. Ввиду того, что все группы равночисленны, все расчеты характеристик распределения проводятся без взвешивания, по формулам простых средних. Квантили распределения определяются следующим образом: при четном числе групп медиана есть

простая средняя из середины интервала с номерами $\frac{n}{2}$ и $\frac{n}{2}+1$, а при нечетном числе интервалов медиана равна середине интервала с номером $\frac{n+1}{2}$. Соответственно определяются и другие квантили. Модальную величину определяем по группе с наибольшей плотностью распределения, то есть с наименьшим по ширине интервалом. Для нашего примера это вторая группа банков с величиной активов-нетто 48625,1-53761,1 млн. р.

Из менее употребляемых видов рядов распределения можно отметить ряды с прогрессивно возрастающими или прогрессивно убывающими по величине интервалами.

В качестве примера ряда с прогрессивно возрастающими уровнями можно привести распределение населения РФ по уровню среднедушевых денежных доходов, полученное на основе материалов выборочного обследования бюджетов домашних хозяйств и макроэкономического показателя среднедушевых денежных доходов населения (таблица 1.4).

Таблица 1.4 – Распределение населения по величине среднедушевых денежных доходов в 2010 г.

В процентах

Все население	100
в том числе со среднедушевыми денежными доходами в месяц, р.:	
до 3500,0	3,9
3500,1-5000,0	5,6
5000,1-7000,0	9,4
7000,1-10000,0	14,7
10000,1-15000,0	20,2
15000,1-25000,0	23,5
25000,1-35000,0	10,8
свыше 35000,0	11,9
Источник: http://www.gks.ru	

1.2 Анализ ранжированного ряда

Проанализируем однородность распределения в ранжированном ряду на примере распределения 30-ти банков Санкт-Петербурга по величине рентабельности активов-нетто. Исходные данные представлены в таблице 1.5.

Таблица 1.5 – Распределение 30-ти банков Санкт-Петербурга по величине рентабельности активов-нетто (ROA)

Банк	ROA	Банк	ROA
1 Горбанк	12,83	16 Севзапинвестпромбанк	1,64
2 Клиринговый дом	6,57	17 Данске Банк	1,41
3 Советский	4,25	18 Промсервисбанк	1,36
4 Санкт-Петербургский Банк Инвестиций	4,14	19 Прайм Финанс	1,17
5 Невский Банк	3,65	20 Турбобанк	1,15
6 Тетраполис	3,51	21 Банк БФА	1,13
7 Невастройинвест	3,05	22 Экси-Банк	0,96
8 Ганзакомбанк	2,90	23 ВТБ	0,94
9 Таврический	2,59	24 Россия	0,92
10 Викин	2,33	25 Балтика	0,92
11 Петербургский Социальный Коммерческий Банк	2,14	26 Объединенный капитал	0,91
12 Энергомашбанк	2,13	27 Банкирский Дом	0,80
13 Финансовый капитал	1,81	28 Констанс Банк	0,79
14 СИАБ	1,79	29 СЭБ Банк	0,71
15 КИТ Финанс Инвестиционный Банк	1,71	30 Балтинвестбанк	0,66
Источник: http://www.banki.ru . Данные представлены на 1.07.2012 г.			

Расчет средней величины, медианы и среднего квадратического отклонения проведем с использованием электронных таблиц Excel (меню «Анализ данных») (таблица 1.6).

Таблица 1.6 – Описательная статистика ранжированного ряда по 30-ти банкам

Показатель	значение	Показатель	значение
Среднее	2,36	Асимметричность	3,214
Стандартная ошибка	0,4376	Интервал	12,17
Медиана	1,68	Минимум	0,66
Мода	0,92	Максимум	12,83
Стандартное отклонение	2,397	Сумма	70,87
Дисперсия выборки	5,745	Счет	30
Эксцесс	12,608	Уровень надежности (95,0 %)	0,8950

Средняя величина рентабельности активов-нетто по исследуемой группе банков составила 2,36. Она находится между девятой и десятой величиной из 30-ти, т.е. в первой трети ряда.

Медиана ряда составила 1,68. Это значение в 1,4 раза меньше рассчитанной средней. Достаточно сильное различие средней величины и медианы может свидетельствовать о неоднородности распределения исследуемой группы банков по величине рентабельности активов-нетто.

Среднее квадратическое отклонение составило 2,397, следовательно, коэффициент вариации равен:

$$V = \frac{2,397}{2,36} \times 100 = 101,6\ \%.$$

В общей теории статистики исходят из того, что величина коэффициента вариации более 33 % уже свидетельствует о неоднородности совокупности, в то время как высокие значения данного коэффициента достаточно часто встречаются в распределении экономических признаков.

Полученные значения асимметрии и эксцесса свидетельствуют о правосторонней островершинной асимметрии анализируемого ряда распределения. При правосторонней асимметрии между показателями центра распределения существует соотношение: $M_o < M_e < \bar{x}$ [2, с. 139]. Для нашего примера имеем: 0,92<1,68<2,36.

Гипотезу о неоднородности распределения единиц анализируемой совокупности подтверждает также расчет *коэффициента Лоренца*:

$$L = \frac{\sum_{1}^{k} \left| d_j - dx_j \right|}{2}, \qquad (1.1)$$

где d_j – доля единиц совокупности в *j*–м интервале;

dx_j – доля анализируемого показателя в *j*-м интервале.

Для равномерного распределения коэффициент концентрации Лоренца равен нулю, в условиях абсолютно неравномерного распределения он равен единице. Значение коэффициента для исследуемого ряда распределения составило 0,314, что свидетельствует о достаточно сильной неоднородности распределения.

Для вычисления отношения 10 % единиц совокупности с наивысшими показателями следует сложить значения по Горбанку (3,33 % от 30-ти банков), Клиринговому Дому и Советскому (12,63+6,57+4,25=23,65).

Аналогично получим необходимое значение по 10 % банков с наименьшими анализируемыми показателями – Констанс Банку, СЭБ Банку и Балтинвестбанку (0,79+0,71+0,66=2,16). Отношение 10 % банков с самой высокой рентабельностью активов-нетто к 10 % банков с самыми низкими показателями рентабельности составило 23,65:2,16=10,95 раза, что также говорит о сильной неоднородности анализируемого ряда распределения.

В связи с тем, что распределение далеко от нормального закона, оно не может быть использовано для измерения связей методами корреляции.

Чтобы распределение группы анализируемых банков стало близким к нормальному, необходимо исключить наблюдения, наиболее сильно отклоняющиеся от основной массы единиц совокупности. В нашем примере это значение рентабельности активов в Горбанке. Исключим Горбанк из состава анализируемой совокупности. Получим следующие показатели распределения (таблица 1.7).

Таблица 1.7 – Описательная статистика ранжированного ряда по 29-ти банкам

Показатель	значение	Показатель	значение
Среднее	2,0	Асимметричность	1,61
Стандартная ошибка	0,2561	Интервал	5,91
Медиана	1,64	Минимум	0,66
Мода	0,92	Максимум	6,57
Стандартное отклонение	1,3790	Сумма	58,04
Дисперсия выборки	1,9016	Счет	29
Эксцесс	2,96	Уровень надежности (95,0 %)	0,5245

Средняя величина рентабельности составила 2,0, медиана равна 16-ой варианте из 30-ти, отличие от средней - 18 %. Среднее квадратическое отклонение составило 1,379, а коэффициент вариации $V = \frac{1,379}{2,0} \times 100 = 69,0\ \%$. Эксцесс равен 2,96 (распределение островершинное) а коэффициент асимметрии составил 1,61, что свидетельствует о сильной скошенности распределения [3, с. 162].

Распределение все еще далеко от нормального. Можно разбить анализируемую совокупность на две совокупности, исключив Горбанк и Клиринговый дом как единицы с аномальными значениями. В первую совокупность включим наблюдения 3-15 (13 наблюдений), а во вторую – 16-30 (15 наблюдений). В результате получим следующие показатели распределения (таблица 1.8).

Таблица 1.8 – Характеристики распределения по двум совокупностям

Показатель	Единицы наблюдения	
	№№ 3-15	№№ 16-30
Средняя величина	2,769	1,031
Медиана	2,59	0,94
Стандартное отклонение	0,89082	0,276567
Эксцесс	-1,1609	0,1136
Асимметричность	0,46334	0,79528
Коэффициент вариации	32,2	26,8

Значения медианы в каждой совокупности близко к значению средней величины, рассчитанные коэффициенты вариации менее 33 %. Невысоки показатели асимметрии и эксцесса. В результате мы получили близкое к нормальному распределение двух совокупностей с умеренной вариацией, которые могут быть использованы для измерения связей методами корреляции.

1.3 Проверка на соответствие нормальному закону распределения равноинтервального ряда

Ряды распределения позволяют решить такую важную задачу статистического анализа как характеристика *закономерностей распределения*.

Если увеличивать число наблюдений и одновременно с этим уменьшать величину интервала, то полигон и гистограмма распределения в пределе будут приближаться к кривой распределения. Она дает четкое представление о форме теоретического распределения единиц совокупности по величине варьирующего признака. Так как каждому ряду распределения достаточно большой совокупности объективно свойственна определенная закономерность, то кривая распределения является ее выражением. Она выражает зависимость между вариантами и частотами. Процесс нахождения функции кривой распределения (аппроксимация) заключается в следующем:

а) подбирается и теоретически обосновывается предельная теоретическая кривая плотности распределения, достаточно точно выражающая свойственную явлению закономерность;

б) определяются параметры функции кривой распределения;

в) оценивается близость эмпирического и теоретического распределения.

Проверка исходных данных на соответствие нормальному закону распределения является необходимым требованием большинства методов статистики и эконометрики. В XIX в. нормальное распределение называли «нормальной кривой ошибок». Нормальное распределение было открыто в 1711 г. в Англии Абрахом де Му-

авром[1]. Иногда его называют распределением Гаусса в честь немецкого математика XIX в. Карла Фридриха Гаусса.

Различные статистические критерии позволяют оценить близость распределения к нормальному: Пирсона, Романовского, Колмогорова-Смирнова, Лиллиефорса, Шапиро-Уилкса. В отечественной практике статистико-эконометрических исследований наиболее часто используются первые три критерия, в зарубежной – критерии Колмогорова-Смирнова, Лиллиефорса, Шапиро-Уилкса. Основой вышеперечисленных критериев является осуществление проверки на близость теоретических частот эмпирическим.

Рассмотрим методику анализа распределения и его близости к нормальному закону (Гаусса-Лапласа) с использованием критерия χ^2 - критерия английского статистика Карла (Чарльза) Пирсона:

$$\chi^2 = \sum_{j}^{H} \frac{(f_j - f_{Hj})^2}{f_{Hj}}, \qquad (1.2)$$

где f_j – эмпирические частоты;

f_{Hj} – теоретические частоты.

Чем меньше отклонение между эмпирическими и теоретическими частотами, тем меньше значение χ^2, а значит, теоретическое распределение лучше воспроизводит эмпирическое, и наоборот. Если эмпирические частоты совпадают с теоретическими, то значение критерия равно нулю. Предварительно следует отметить, что применение данного критерия должно удовлетворять следующим условиям:

---

[1] А. Муавр внес большой вклад в теорию вероятностей. Он доказал частный случай теоремы Лапласа, провел вероятностное исследование азартных игр и ряда статистических данных по народонаселению.

Есть легенда, согласно которой Муавр точно предсказал день собственной смерти: он обнаружил, что продолжительность его сна стала увеличиваться в арифметической прогрессии, и легко вычислил, когда она достигнет 24 часов, и, как всегда, не ошибся…

- результаты наблюдений должны быть независимыми;

- чтобы при малой величине теоретической частоты небольшое абсолютное отклонение не дало очень большой относительной величины, группы объединяются таким образом, чтобы ожидаемая частота была не менее 6 (поправка Йейтса);

- объем исследуемой совокупности должен составлять не менее 50-ти наблюдений.

Этапы проверки по данному критерию следующие:

1) рассчитывается расчетная величина критерия по формуле (1.2);

2) по табулированным значениям χ^2 - критерия находим его критическое значение с соответствующим уровнем значимости α (0,1; 0,05, 0,01) и числом степеней свободы, равных числу слагаемых критерия (число интервальных групп ряда) минус 3 (т.к. при расчете нормального распределения три параметра были фиксированы: $\sum f_j$, $\bar{x}$, σ_x). Чем меньше значение уровня значимости α, тем выше вероятность принятия верного решения;

3) сравниваем расчетное и критическое значения критерия χ^2. Случай, когда $\chi^2_{расч} > \chi^2_{крит}$ свидетельствует о том, что расхождение между эмпирическими и теоретическими частотами существенно и гипотеза о близости эмпирического распределения к нормальному отвергается. Если $\chi^2_{расч} < \chi^2_{крит}$, то расхождение между эмпирическими и теоретическими частотами объясняется случайными колебаниями результатов наблюдений и гипотеза о нормальном законе распределения принимается с вероятностью 1-α.

Проанализируем имеющийся равноинтервальный ряд распределения субъектов РФ по величине расходов домашних хозяйств на оплату услуг в 1 кв. 2010 г. (таблица 1.9). Средний размер расходов на услуги в представленной группировке найдем по формуле средней арифметической взвешенной:

$$\overline{X} = \frac{\sum_{j=1}^{k} x_j^{'} \cdot f_j}{\sum_{1}^{k} f_j} = \frac{172908{,}1}{80} = 2161{,}4 \ p.$$

Таблица 1.9 – Распределение субъектов РФ по величине расходов домашних хозяйств на оплату услуг в 1 кв. 2010 г. (в среднем на члена домохозяйства в месяц)

Группа субъектов по величине расходов на оплату услуг, р.	Число субъектов f_j	Накопленная частота, f'_j	Середина интервала, x'_j	$x'_j \times f_j$	$x'_j - \bar{x}$
330,9-943,6	3	3	637,25	1911,75	-1524,15
943,6-1556,3	13	16	1249,95	16249,35	-911,45
1556,3-2169,0	37	53	1862,65	68918,05	-298,75
2169,0-2781,7	11	64	2475,35	27228,85	313,95
2781,7-3394,4	6	70	3088,05	18528,3	926,65
3394,4-4007,1	5	75	3700,75	18503,75	1539,35
4007,1-4620,1	5	80	4313,60	21568,00	2152,20
Итого	80	x	x	172908,10	x
Источник: http://www.gks.ru . Расчеты автора.					

Медиана ряда составит:

$$Me = x_{Me} + h_{Me} \frac{0{,}5f - S_{Me-1}}{f_{Me}} = 1556{,}3 + 612{,}7 \times \frac{80 \div 2 - 16}{37} = 1953{,}7 \ p.,$$

где x_{Me} - начальное значение интервала, содержащего медиану;

h_{Me} - величина медианного интервала;

f - сумма частот ряда;

S_{Me-1} - сумма накопленных частот, предшествующих медианному интервалу;

f_{Me} - частота медианного интервала.

Мода расходов на оплату услуг составит:

$$Mo = x_{Mo} + h_{Mo} \times \frac{f_{Mo} - f_{Mo-1}}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})} = 15563 + 6127 \times \frac{37-13}{37-13+37-11} = 18504\ p.$$

где x_{Mo} - начальное значение интервала, содержащего моду;

h_{Mo} - величина модального интервала;

f_{Mo} - частота модального интервала;

f_{Mo-1} - частота интервала, предшествующего модальному;

f_{Mo+1} - частота интервала, следующего за модальным.

Рассчитаем критерий согласия Пирсона, который измеряет степень отличия частоты фактического распределения от частоты нормального распределения при той же численности единиц совокупности, той же средней величине признака и том же среднем квадратическом отклонении.

Для построения ряда с нормальным распределением, для каждой из границ интервалов признака в таблице 1.9 необходимо вычислить критерий *t* как отношение разности между этой границей интервала и средней величиной признака к среднему квадратическому отклонению.

Среднее квадратическое отклонение по рассматриваемому ряду составило:

$$\sigma = \sqrt{\frac{\sum (x_j - \bar{x})^2 f_j}{\sum f_j}} = 882{,}6\ p.$$

Для начала первого интервала имеем: (330,9-2161,4):882,6=-2,074; для верхней границы первого и нижней границы второго интервала: (943,6-2161,4):882,6=-1,38 и т.д. По значениям критериев *t* для конца и начала каждого интервала групп рассчитывается вероятность попадания единицы совокупности в данный интервал (при условии нормального закона распределения). Эта вероятность (P_{Hj}) равна половине

разности между функцией *F(t)* для большего по абсолютной величине значения *t* и *F(t)* для меньшей по абсолютной величине границы интервала. Если знаки *t* для границ одного из интервалов (среднего из них) разные, то вместо разности берется сумма.

Для первого интервала (таблица 1.10) вероятность попадания в этот интервал при нормальном законе равна:

$$[F(2{,}074) - F(1{,}38)]:2 = (0{,}9616 - 0{,}8324):2 = 0{,}0646.$$

Таблица 1.10 – Расчет критерия χ^2

Группа субъектов по величине расходов на оплату услуг, р.	f_j	t_j	P_{Hj}	f_{Hj}	$f_j - f_{Hj}$	$\frac{(f_j - f_{Hj})^2}{f_{Hj}}$
330,9-943,6	3	от –2,07 до –1,38	0,0646	5,2	-2,2	0,909
943,6-1556,3	13	от -1,38 до –0,69	0,1613	12,9	0,1	0,001
1556,3-2169,0	37	от –0,69 до +0,01	0,2589	20,7	16,3	12,809
2169,0-2781,7	11	от +0,01 до +0,70	0,2541	20,3	-9,3	4,280
2781,7-3394,4	6	от +0,70 до +1,40	0,1612	12,9	-6,9	3,688
3394,4-4007,1	5	от +1,40 до +2,09	0,6245	5,0	0,0	2,173
4007,1-4620,1	5	от +2,09 до +2,79	0,01565	1,3	3,7	
Σ	80	x	0,9782	78,3	x	23,860

Сумма полученных вероятностей для всех интервалов меньше единицы в связи с тем, что при нормальном законе часть единиц совокупности имела бы значения

признака, выходящие за границы фактического размаха вариации. Полученные вероятности для нормального распределения умножаются на общую численность единиц совокупности ($\sum_{j=1}^{H} f_j$), и в результате получаем частоты нормального распределения f_{Hj}.

Последние две группы согласно поправке Йейтса объединяются в одну при расчете χ^2. Получим расчетное значение критерия, равное 23,86. Число степеней свободы составляет 6-3=3. Табличное значение критерия при уровне значимости 0,05 и числе степеней свободы 3 составляет 7,8 (приложение А). Исходя из полученных расчетов, можно сделать вывод, что гипотеза о нормальном характере эмпирического распределения отклоняется.

1.4 Показатели степени неравномерности распределения равночастотного ряда

Чем больше число равных групп, на которые разделены единицы совокупности, тем больше информации можно получить о характере распределения. Так, Росстат публикует информацию о распределении населения России по душевому доходу всего по пяти 20 %-м группам (таблица 1.11). Из пятой 10-процентной группы населения с наибольшими доходами выделяется 10 % населения с наивысшими доходами. Эти данные публикуются в ежегодниках «Социальное положение и уровень жизни населения России», «Российский статистический ежегодник», статистическом бюллетене «Социально-экономические индикаторы бедности».

Такого рода распределение не позволяет разделить среднедоходную группу и гораздо менее многочисленную группу действительно богатых граждан. Соответственно не выделяется и группа наиболее бедных, низкодоходных граждан. Гораздо информативнее был бы ряд распределения на 20 групп по 5 % населения в каждой, а для наиболее высокодоходных - дополнительные группировки на 5 подгрупп по 1 % в каждой, т.к. этот последний 1 % самых высокодоходных граждан может занимать в общей сумме дохода всего населения несколько десятков процентов.

Таблица 1.11 – Распределение общего объема денежных доходов по 20-процентным группам населения

Показатели	Год					
	2005	2006	2007	2008	2009	2010
Денежные доходы – всего, %	100,0	100,0	100,0	100,0	100,0	100,0
в том числе по 20-процентным группам населения:						
первая (с наименьшими доходами)	5,4	5,3	5,1	5,1	5,1	5,2
вторая	10,1	9,9	9,7	9,8	9,8	9,8
третья	15,1	14,9	14,8	14,8	14,8	14,8
четвертая	22,7	22,6	22,5	22,5	22,5	22,5
пятая (с наибольшими доходами)	46,7	47,3	47,9	47,8	47,8	47,7
из нее 10 % населения с наивысшими доходами	30,1	30,6	31,1	31,1	31,0	30,9

Обратимся к представленной в первом параграфе таблице 1.3, отражающей распределение 100 банков РФ по величине активов-нетто, дополнив ее необходимыми для расчета показателей неравномерности распределения графами.

Данные пятой графы таблицы 1.12 показывают, что 10 % банков из представленной группы с наименьшими показателями активов-нетто располагают менее 1 % (0,56 %) всех активов, представленных в распределении. В свою очередь в 10 % банков с наибольшими показателями величины активов-нетто сосредоточено 83,45 % активов всех представленных банков.

Оценить неравномерность распределения ряда можно также с помощью коэффициента фондов, отражающего отношение доли анализируемых показателей 10 % высшей группы к доле 10 % низшей группы. В нашем случае это отношение составляет 83,45:0,56=149 раз – сильная неравномерность распределения банков по величине активов-нетто очевидна.

Таблица 1.12 – Данные для анализа ряда распределения 100 банков РФ по величине активов-нетто (по состоянию на 1.08.2012 г.)

Интервал активов-нетто, млн. р.	Доля банков, d_j, %	Середина интервала, x_j'	$x_j' \times d_j$	Доля в активах, dx_j, %	$\lvert d_j - dx_j \rvert$	Нарастающие доли, $d'x$, %	Нарастающая доля банков, d_j', %
40022,8-47986,7	10	44004,771	440047,71	0,56	9,44	0,56	10
48625,1-53761,1	10	51193,059	511930,59	0,66	9,34	1,22	20
55832,7-71201,9	10	63517,284	635172,84	0,81	9,19	2,03	30
73258,2-83676,3	10	78467,288	784672,88	1,01	8,99	3,04	40
84615,3-100479,1	10	93547,216	935472,16	1,20	8,80	4,24	50
102537,5-121894,2	10	112215,846	1122158,50	1,44	8,56	5,68	60
123475,5-200105,7	10	161790,618	1617906,20	2,07	7,93	7,75	70
201135,6-267632,1	10	234383,806	2343838,10	3,00	7,00	10,75	80
271371,5-532897,8	10	452134,619	4521346,20	5,80	4,20	16,55	90
682668,0-12336754,7	10	6509711,358	65097114,00	83,45	73,45	100,00	100
И т о г о	100	7800965,865	78009659,00	100,00	146,90	151,82	х

Рассчитаем далее коэффициент Лоренца. По данным таблицы 1.12 его значение составит 146,9:2= 73,45. Коэффициент Лоренца изменяется в пределах от 0 до 1, поэтому 0,7345 (73,45 %) – это сильная степень неравномерности. Если бы все акти-

вы-нетто были сосредоточен у 10 % банков наивысшей группы, то коэффициент Лоренца составил бы 0,9 или 90 %.

Графически анализируемая ситуация представлена на *диаграмме Лоренца* (рисунок 1.2).

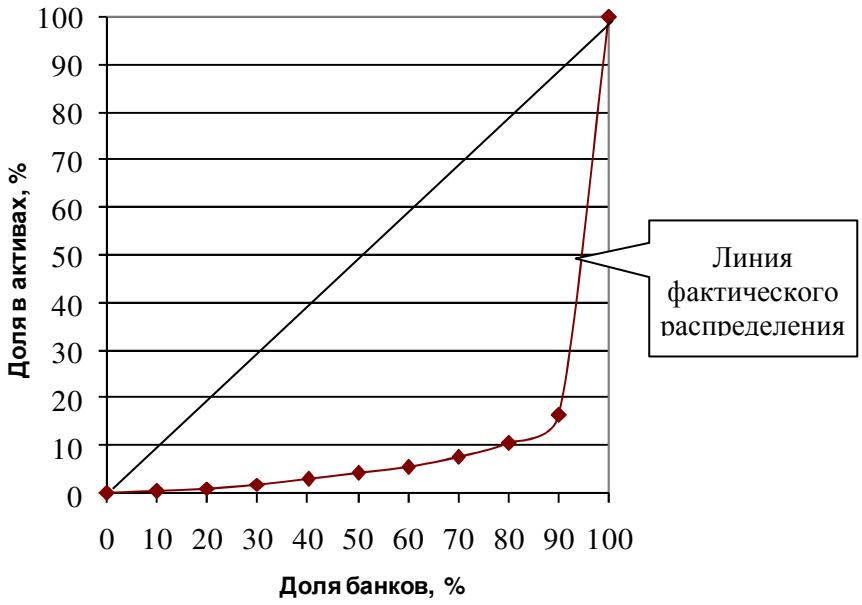


Рисунок 1.2 – Диаграмма Лоренца

Для построения диаграммы Лоренца используем кумулятивные итоги доли банков и доли активов. Если банки по доле активов были распределены равномерно, т.е. процесс концентрации отсутствовал полностью, график представлял бы собой прямую линию по диагонали квадрата. При неравномерном распределении линия концентрации отходит от прямой линии и представляет собой вогнутую кривую, причем, чем выше уровень концентрации, тем дальше отходит линия Лоренца от линии равномерного распределения, тем больше ее кривизна.

Среди показателей неравномерности распределения можно также рассчитать коэффициент, предложенный итальянским статистиком Коррадо Джини, рассчитываемый по формуле:

$$Д = 1 - 2\sum_{1}^{k} d_j \cdot dx_j^{'} + \sum_{1}^{k} d_j \cdot dx_j \qquad (1.3)$$

Для равночастотных рядов расчет коэффициента Джини упрощается: выносим постоянные доли за знак сумм, учитывая, что $\sum_{1}^{k} d_j = \sum_{1}^{k} dx_j = 1$, получим:

$$Д = 1 - 2d_j \cdot \sum_{1}^{k} d' x_j + d_j = 1 - 2 \cdot 0{,}1 \cdot 1{,}5182 + 0{,}1 = 0{,}7964 \text{ или } 79{,}64\ \%.$$

Коэффициент Джини всегда больше, чем коэффициент Лоренца и изменяется также в пределах от 0 до 1.

Рассчитаем структурные средние ряда – моду и медиану. В нашем примере число групп четное, поэтому медиана находится в середине ряда между пятым и шестым интервалом, ее значение равно:

$$Me = (93547{,}216 + 112215{,}846) \div 2 = 102881{,}531\ млн.\ р..$$

В случае если число интервалов в ряду нечетное, то медиана равна середине интервала с номером $\frac{n+1}{2}$, то есть в обоих случаях определить медиану достаточно просто.

При расчете моды в равночастотном ряду исходят из того, что мода распределения – это варианта с наибольшей величиной плотности распределения. Плотность распределения есть отношение частоты к ширине интервала. Так как частота во всех интервалах одинаковая, то плотность больше в том интервале, в котором меньше ширина.

В нашем примере (таблица 1.3) это интервал второй группы, с шириной, равной 53761,1-48625,1=5136 млн. р. Дальнейший расчет моды приводится по обычной ее формуле, где вместо частоты в нее входят плотности интервалов p_j:

$$Mo = x_{Mo} + \frac{(P_{мо} - P_{мо-1}) \times h_{мо}}{(P_{мо} - P_{мо-1}) + (P_{мо} - P_{мо+1})} = 48625{,}1 + \frac{(0{,}1947 - 0{,}12557) \times 5136}{(0{,}1947 - 0{,}12557) + (0{,}1947 - 0{,}06506)} =$$

$$= 50411{,}344\ млн.\ р.$$

В имеющемся равночастотном распределении нет открытых интервалов, поэтому среднюю величину активов-нетто вычислим по формуле средней арифметической простой из середин всех интервалов x'_j:

$$\bar{x} = \frac{\sum_{j=1}^{k} x'_j}{k} = \frac{7800965{,}865}{10} = 780096{,}587 \text{ } млн.\ р.$$

Большие расхождения между величинами средней, моды и медианы свидетельствуют о крайней неоднородности распределения представленной группы банков по величине активов-нетто и значительной асимметрии. Проведение корреляционно-регрессионного анализа связи величины активов-нетто с другими признаками требует оговорки о невозможности дать вероятностную оценку результатов *корреляционного* анализа из-за нарушения условий метода наименьших квадратов, расхождения распределения с нормальным законом распределения вероятностей.

1.5 Вопросы для самоконтроля

1. Дайте определение ряда распределения. Каковы правила составления рядов распределения?

2. Приведите классификацию рядов распределения и их отличительные особенности.

3. Каким образом проводится проверка близости распределения к нормальному в ранжированном ряду?

4. Какова схема расчета критерия Пирсона?

5. Как оценить степень неравномерности распределения в равночастотном ряду?

1.6 Тесты

1. Какой ряд даст наиболее верную информацию о распределении при сильной вариации признака

а) ранжированный;

б) равночастотный;

в) равноинтервальный;

г) с прогрессивно возрастающими уровнями;

д) с прогрессивно убывающими уровнями.

2. В ряду с нормальным распределением примерно равны величины

а) моды, медианы и средней;

б) только моды и медианы;

в) только моды и средней;

г) только медианы и средней.

3. Отрицательное значение эксцесса свидетельствует о том, что распределение

а) правостороннее;

б) левостороннее;

в) островершинное;

г) плосковершинное

д) нормальное.

4. Характеристиками неравномерности распределения могут служить

а) коэффициент Джини;

б) коэффициент фондов;

в) коэффициент корреляции;

г) коэффициент Лоренца.

5. Для расчета моды в равночастотном ряду, необходимо значение

а) частости;

б) частоты;

в) величины интервалов в группах;

г) вероятности попадания единицы совокупности в интервал.

2 Введение в регрессионный анализ. Классическая модель линейной регрессии

Что необходимо знать из 2 главы:

1. Понятие, цель, задачи проведения регрессионного анализа и проблемы спецификации регрессионной модели.

2. Предпосылки применения метода наименьших квадратов, свойства МНК-оценок.

3. Схема проведения дисперсионного анализа.

4. Порядок проверки значимости параметров уравнения и коэффициента корреляции. Построение доверительных интервалов.

2.1 Основные задачи, понятия и этапы проведения регрессионного анализа

Теоретические основы корреляционно-регрессионного анализа были предложены в XIX в. английским психологом и антропологом Фрэнсисом Гальтоном[1], а методы и модели регрессионного анализа занимают ведущее место в математико-статистическом аппарате эконометрики. Каждый изучающий экономику сталкивает-

[1] Ф. Гальтон разработал методы статистической обработки результатов исследований (в частности, метод исчисления корреляций между переменными); ввел коэффициент корреляции; создал т.н. биометрическую школу.

ся с принципиальной идеей взаимосвязи между явлениями и, как следствие, возникает задача количественного описания таких взаимосвязей. Не изучив характер, особенности, меру связи между явлениями и процессами невозможно адекватное управление ими и прогнозирование их дальнейшего развития. Интерес представляет не только определение характеристик распределения каждого признака, но и то, как они связаны между собой, и можно ли оценить зависимость значений одного признака от другого. Например, на микроэкономическом уровне нас может интересовать, как среднедушевой доход домохозяйства и предыдущий уровень потребления влияют на текущие потребительские расходы; зависимость стоимости квартиры от ее местоположения, этажа, благоустройства; каким образом объем реализации продукции связан с размерами товарных запасов, торговой площадью, уровнем квалификации работников и т.п. На макроэкономическом уровне - какие факторы, и в какой степени оказывают влияние на ожидаемую продолжительность жизни; цены потребительского рынка можно рассматривать как функцию от цен на энергоносители и т.д. Для статистического исследования взаимосвязей между изучаемыми явлениями, показателями, системами и предназначен математический инструмент *регрессионного анализа*.

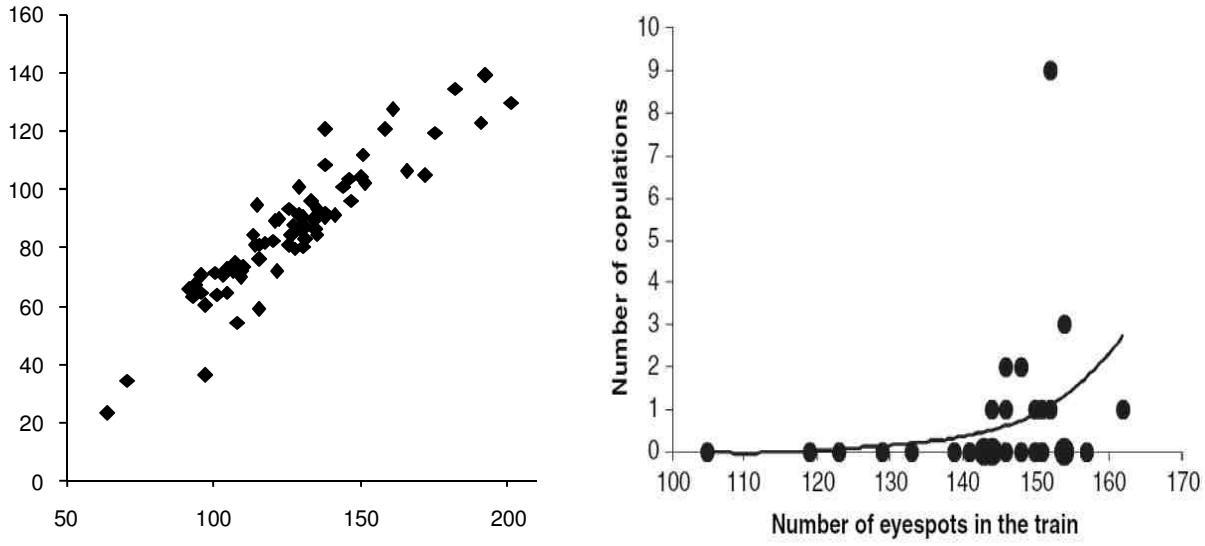
Предварительно необходимо отметить, что в исследовании явлений естественных наук (математике, физике, химии) чаще всего речь ведут о *функциональной* (детерминированной) зависимости (связи).

Большая часть традиционных экономических теорий, в которых связи между экономическими категориями отражаются с помощью диаграмм и алгебраических формул, имеет дело с точными функциональными соотношениями [4, с. 18].

В случае детерминированной зависимости между неслучайными переменными каждому значению одной переменной соответствует вполне определенное значение другой, т.е. одна переменная является функцией от другой или нескольких переменных. Анализируя экономические явления и процессы, исследователи в большинстве случаев сталкиваются не с функциональной, а со *стохастической* (вероятностной) связью. В чем ее отличие от функциональной? В случае стохастической зависимости каждому значению одной переменной соответствует не какое-то

определенное, а множество возможных значений другой переменной, т.е. определенное (условное) распределение другой переменной. Зависимая переменная может подвергаться влиянию неучтенных или неконтролируемых факторов. Кроме того, измеренные в ходе статистического наблюдения значения переменных могут быть несвободны от некоторых случайных ошибок наблюдения и измерения. Вероятностная связь потому так и называется, что мы делаем предположение о существующей зависимости между явлениями с некоторой долей вероятности. Регрессионный анализ в свою очередь применяется только в случаях, когда анализируемые зависимости имеют стохастическую природу и выявляются на основании статистического наблюдения.

Примеры стохастических зависимостей приведены на рисунке 2.1.



а б

а – линейная; б – нелинейная.

Рисунок 2.1 – Примеры стохастических зависимостей

В связи с вероятностной природой статистической зависимости между X и Y для исследователя представляет интерес усредненная по X схема зависимости – закономерность в измерении условного математического ожидания $M_x(Y)$ или

$M(Y/X) = x$ (математического ожидания случайной переменной *Y*, вычисленного в предположении, что переменная *X* приняла значение *x*) в зависимости от *x*.

Корреляционная связь предполагает, что каждому значению одной переменной соответствует определенное условное математическое ожидание (среднее значение) другой:

$$M_x(Y) = f(x), \tag{2.1}$$

где $f(x) \neq const$.

Несмотря на то, что понятия корреляции и регрессии достаточно близки – оба метода направлены на анализ статистической связи между признаками, между ними есть принципиальные различия. Невысокая величина коэффициента корреляции не исключает тесную линейную связь между анализируемыми признаками. В то же время высокое значение коэффициента может не содержать никакой полезной информации или противоречить экономическому смыслу. Кроме того, корреляция лишь обозначает, что изменение одного признака в среднем зависит от изменения другого, ничего не сообщая о причинной зависимости (обусловленности) между признаками. В то время как регрессионный анализ отражает именно *обусловленность* изменения одного признака от изменения другого.

Прикладной регрессионный анализ занимается решением следующих задач:

1. Установление факта наличия/отсутствия связи между переменными.

2. Установление формы зависимости между переменными.

3. Оценка функции регрессии.

4. Прогнозирование неизвестных значений зависимой переменной по заданным значениям независимых переменных.

Зависимую переменную обозначают *Y* и называют также эндогенной (объясняемой, выходной, функцией отклика, результативным признаком). Независимую переменную обозначают *X*, она носит название экзогенной (факторной, результирующей, предиктора, регрессора, входной, предсказывающей).

В регрессионном анализе эндогенная переменная выступает в роли функции, значения которой всегда стохастичны по своей природе. Экзогенная переменная играет роль аргумента той функции, в качестве которой рассматривается эндогенная (зависимая) переменная. По своей природе регрессоры могут быть как случайными, так и неслучайными.

Односторонняя зависимость случайной переменной *Y* от одной (или нескольких) неслучайной переменной *X* называют регрессионной и она может быть выражена в виде *модельного уравнения регрессии*.

При регрессионной модели предполагается, что:

1) каждому отдельно взятому значению *x* соответствует *нормальное распределение y*, из которого случайно отбираются выборочные y_i;

2) средние $\tilde{y}_i$ всех таких подсовокупностей лежат на линии регрессии;

3) все совокупности, образованные из элементов, имеющих одинаковые значения *x*, из которых берутся выборки, имеют нормальное распределение *y* с общей для них дисперсией [5. с. 133-140, 161-167].

В общем виде регрессионная модель имеет вид

$$Y = f(X_1, X_2, ..., X_i, \varepsilon). \qquad (2.2)$$

Подстрочный индекс *i* соответствует конкретному наблюдению. Функция *f* называется функцией регрессии *Y* по *X*. Она описывает аналитическую форму функции объясняющих переменных, определяемую в процессе построения модели.

Дж. Джонстон в работе «Эконометрические методы» замечает, что «…даже самое элементарное знакомство с экономическими данными показывает, что их отдельные значения не укладываются точно на прямую или другую гладкую линию. Поэтому формализация типа $Y = f(X)$, как и любые ее конкретизации, оказывается неадекватна целям, связанным с измерениями в экономике и с испытанием тех или иных форм зависимостей между переменными. Решение подобных задач становится возможным в результате введения в экономические соотношения стохастического члена» [4, с. 18].

Стохастический член (случайные отклонения) ε отражает вероятностный характер регрессионной модели. В связи со сложностью социально-экономических явлений практически невозможно учесть все факторы, влияющие на формирование и изменение эндогенной переменной. Поэтому эмпирические значения этой переменной никогда не бывают строго равны модельным (теоретическим) значениям, полученным по уравнению регрессии. Результат воздействия случайных, неучтенных факторов – это и есть разница между фактическими и расчетными значениями объясняемой переменной. Чем меньше эта разница, тем лучше полученная модель отражает исследуемую действительность. Величина остаточной компоненты зависит также от правильности спецификации регрессионной модели – в модель могут быть не включены важные объясняющие переменные, переменные агрегированы, неправильно описана структура модели или ее функциональная спецификация. Возможны ситуации, когда имеются переменные, которые мы хотели бы включить в регрессионное уравнение, но не можем их измерить (например, психологические факторы). Могут существовать объясняющие переменные, которые являются существенными, но из-за отсутствия опыта мы их таковыми не считаем.

Говоря об агрегировании переменных, заметим, что во многих случаях рассматриваемая зависимость – это попытка объединить вместе некоторое число микроэкономических соотношений. Например, функция суммарного потребления – это попытка общего выражения совокупности решений отдельных индивидов о расходах. Так как отдельные соотношения, вероятно, имеют разные параметры, любая попытка определить соотношение между совокупными расходами и доходом является лишь аппроксимацией [6, с. 55]. Ошибки измерения при проведении статистического наблюдения также могут существенно влиять на величину случайных отклонений.

Итак, чтобы точно описать уравнение регрессии, исследователь должен знать условный закон распределения эндогенной переменной Y при том, что экзогенная переменная X примет значение x ($X=x$). На практике наблюденные значения зависимой переменной представляют собой некую выборку объема n. Даже если данные обследования охватывают все изучаемые экономические объекты на момент време-

ни, к этим данным нужно относиться как к выборочным. Это связано с тем, что наблюденные значения зависимой переменной соответствуют только некоторым значениям ненаблюдаемых факторов, влияние которых описывается случайными отклонениями ε. Поэтому речь уже идет об оценке (приближенном значении или аппроксимации) по выборке функции регрессии, а уравнение в это случае называется выборочным уравнением регрессии:

$$\tilde{y} = \tilde{f}(x, b_0, b_1, ..., b_k), \qquad (2.3)$$

где $\tilde{y}$ - условная (групповая) средняя переменной Y при фиксированном значении $X=x$;

$b_0, b_1, ..., b_k$ - параметры аппроксимирующей функции.

Если аппроксимирующая функция адекватна исходным данным, а $n \to \infty$ (увеличение объема выборки), то $\tilde{f}$ будет сходиться по вероятности к функции регрессии *f*. Уравнение (2.3) является строгой функцией, поэтому не содержит стохастического (остаточного) члена.

В заключение остановимся на основных этапах проведения регрессионного анализа. Достаточно развернуто они представлены в работе Р. Винна и К. Холдена «Введение в прикладной эконометрический анализ».

Э. Новак во «Введении в методы эконометрики» выделяет следующие основные шаги проведения анализа зависимостей. На предварительном этапе определяется исследуемое явление, что равнозначно выбору переменной, объясняемой моделью. На первом этапе из множества факторов, влияющих на объясняемую переменную, выбираются объясняющие переменные. Второй этап – выбор аналитической формы модели, т.е. выбор конкретной математической функции, описывающей зависимость объясняемой переменной от объясняющих переменных. На третьем этапе оцениваются параметры модели, т.е. рассчитываются оценки значений каждого параметра. На четвертом этапе выполняется верификация модели, цель которой заключается в проверке, насколько хорошо построенная модель описывает экономи-

ческие реалии. Последний этап – принятие решений с помощью модели, т.е. ее практическое использование. Принимаемые решения могут относиться к одному из двух видов деятельности: к экономическому анализу или к прогнозированию [7, с. 12-13].

Предварительный, первый и второй этапы из вышеперечисленных относятся к задачам спецификации модели.

2.2 Проблемы спецификации модели

Спецификацией модели называют ее концептуальную функциональную форму. В практике регрессионного анализа не существует универсальной схемы подбора наилучшей с точки зрения аппроксимации функции регрессии. На предварительном этапе исследования опираются на имеющуюся априорную информацию об изучаемом явлении, проводится качественный анализ сущности явления, согласующийся с основными положениями экономической теории, социологии, специфики вида экономической деятельности. Оцениваются существующие точки зрения на изучаемое явление.

Далее из всего круга факторов, оказывающих влияние на результат, необходимо выделить наиболее существенные. Парная (простая) регрессия достаточна, если имеется доминирующий экзогенный признак, который используется в качестве объясняющей переменной. Уравнение простой регрессии характеризует связь между двумя переменными, которая проявляется как закономерность в среднем и целом по совокупности наблюдений [8, с. 47].

Уравнение взаимосвязи двух переменных (парная регрессионная модель) может быть представлено как:

$$Y = f(X) + \varepsilon . \qquad (2.4)$$

В случае парной регрессии уравнение может быть выражено различными классами математических функций. Различают линейную и нелинейные регрессии, которые в свою очередь делятся на 2 класса:

1. Регрессии, нелинейные относительно включенных в анализ факторных переменных, но линейные по оцениваемым параметрам.

2. Регрессии, нелинейные по оцениваемым параметрам.

В таблице 2.1. приведены основные классы математических функций, используемых при количественной оценке связи между двумя переменными.

Таблица 2.1 – Классы основных математических функций, используемых в парной регрессии

Линейная	$y = b_0 + b_1 x$
I класс нелинейных регрессий	
Полиномы различных степеней	$y = b_0 + b_1 x + b_2 x^2 + ... + b_n x^n$
Равносторонняя гипербола	$y = b_0 + \frac{b_1}{x}$
II класс нелинейных регрессий	
Степенная	$y = b_0 \times x^{b_1}$
Показательная	$y = b_0 \times b_1^x$
Экспоненциальная	$y = e^{b_0 + b_1 x}$

Когда связь между переменными хорошо выражена, удобно, чтобы она оказалась линейной. Интерполяция и интерпретация становятся легкими, анализ остатков от такой регрессии много проще. Стандартные пакеты прикладных программ предусматривают процедуры линеаризации нелинейных моделей, позволяющие работать с линейными моделями, построенными по преобразованным данным.

Например, функция $y = b_0 \times x^{b_1}$ при x>0 путем логарифмирования и замены переменных преобразуется как $\ln y = \ln b_0 + b_1 \ln x$. После замены переменных $y' = \ln y; b_0' = \ln b_0; x' = \ln x$, получают линейную по параметрам функцию $y' = b_0' + b_1 x'$.

Кроме приведенной в примере степенной зависимости линеаризации поддаются экспоненциальные, логарифмические, гиперболические зависимости.

Выбор вида уравнения может осуществляться путем сравнения рассчитанной при разных моделях остаточной дисперсии (дисперсия возмущений, ошибок) (см., например, [9]). Чем меньше величина данного показателя, тем меньше степень влияния не учтенных в модели факторов и тем адекватнее полученная модель фактическим данным.

Парные связи встречаются в экономике достаточно редко, чаще эндогенная переменная обусловлена несколькими экзогенными. Регрессия результативного признака с двумя или более факторными называется множественной. При отборе факторов для уравнения множественной регрессии необходимо учитывать следующие основные условия.

1. Данные должны быть количественно измеримыми, достоверными, а изучаемая совокупность достаточно большой, так как для статистической методологии важное значение имеет закон больших чисел. Согласно закону больших чисел в массе индивидуальных явлений общая закономерность проявляется тем полнее и точнее, чем больше их охвачено наблюдением, только в этом случае происходит взаимопогашение индивидуальных значений признака от средней величины. Если есть необходимость включения в модель качественного фактора, следует придать ему количественную определенность, например, проранжировав наблюдения, либо присвоив им определенный вес (балл).

Достаточным объемом совокупности (выборки) для установления надежной связи между признаками x и y следует считать такую численность единиц совокупности n, при которой величина коэффициента корреляции r превосходит его среднюю ошибку репрезентативности s_r не менее чем в t раз, где t – критерий Стьюдента при значимости (вероятности нулевой гипотезы об отсутствии связи) 0,05. При численности совокупности более 30 единиц t- критерий можно считать равным 2, при меньшей численности величину критерия следует определить по таблице t-распределения при числе степеней свободы (*d.f.*), равном n-2.

Имеем условие:

$$t = \frac{r}{s_r} = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}. \quad (2.5)$$

Избавляясь от радикалов, возведем обе стороны равенства в квадрат:

$$t^2 = \frac{r^2(n-2)}{1-r^2}. \quad (2.6)$$

Решая уравнение 2.6 относительно *r*, получаем:

$$r = \sqrt{\frac{1}{(n-2) \div t^2 + 1}}. \quad (2.7)$$

Значения *r*, вычисленные по формуле (2.7) приведены в таблице Б.1 приложения Б.

Приведенные в таблице значения коэффициентов корреляции означают минимальную величину, которая может быть надежной при численности совокупности $n = d.f. + 2$. Так, при объеме выборки 10 единиц, могут быть надежно отличными от нуля только те коэффициенты, которые равны или больше 0,6319. Если же требуется уровень надежности 0,99, то лишь коэффициенты, большие или равные 0,7646 будут надежно говорить о наличии связи, а, например, при $r = 0{,}2$ нужна большая совокупность, не менее 95 единиц.

Если необходимо не только установить наличие связи, но еще и достаточно точно измерить ее тесноту, то средняя ошибка коэффициента должна быть меньше самого коэффициента в несколько раз, например, в 4 раза. А для этого объем выборки должен возрасти в 4^2 раза, т.е. в 16 раз, так как ошибка снижается как корень квадратный из *n*. Чтобы с надежностью 0,95 (значимость 0,05) сказать, что коэффициент корреляции заключен в границах 0,6 ± 0,15 объем выборки должен быть не менее чем $9 \times 4^2 = 144$ единицы [10, с. 40-41].

2. Качественная однородность единиц совокупности – каждая единица совокупности должна в равной степени обладать характерными признаками определенного типа.

3. Факторные переменные должны иметь высокую вариабельность, сильную степень корреляции с результативным признаком и не должны сильно коррелировать между собой, а тем более находиться в точной функциональной связи.

4. Отдельные наблюдения должны быть независимыми, т.е. результаты, полученные в отдельном наблюдении не должны содержать информацию о предыдущих наблюдениях и не должны быть связаны с будущими.

5. Распределения факторных и результативного признаков должно подчиняться нормальному закону распределения вероятностей. Это обусловлено применением метода наименьших квадратов для расчета параметров уравнения. При большом объеме выборки проверить соответствие распределения нормальному можно по критерию К. Пирсона (см. главу «Анализ рядов распределения»). При малой выборке, используя статистические пакеты прикладных программ, следует получить показатели «стандартизованная асимметрия» и «стандартизованный эксцесс», являющиеся отношениями показателей асимметрии и эксцесса к их средним ошибкам, то есть *t*-критерии Стьюдента. Они должны быть не больше, чем критические табличные при значимости 0,05 и *n*-2 степенях свободы. В случае, если расчетные значения значительно превышают табличные, из состава совокупности следует исключить резко выделяющиеся (аномальные) единицы совокупности.

Для оценки тесноты связи зависимой переменной с каждой из независимых переменных можно визуализировать их с помощью диаграмм рассеяния - поля корреляции в виде точек в декартовой системе координат (см. рисунок. 2.1). О применении графического метода для подтверждения гипотезы о возможных видах связи указывалось американским экономистом Ф. Миллсом в работе «Статистические методы» [11].

При построении диаграмм рассеяния рекомендуется масштабы по осям абсцисс и ординат выбирать таким образом, чтобы значения обоих анализируемых признаков укладывались на отрезках приблизительно равной длины. Диаграмма отразит

существование/отсутствие зависимости, а при наличии зависимости – вид и тесноту связи между парами анализируемых признаков. После отбора признаков осуществляют сбор и контроль анализируемого материала методами статистического наблюдения в соответствии с задачами исследования. На практике работа, связанная со сбором статистической информации, зачастую проводится в рамках самостоятельного исследования.

Определившись с набором переменных и структурной формой анализируемой зависимости, переходят к определению ее функционального вида или аналитической формы связи. Для множественной регрессии, также как и для парной, существуют различные классы аппроксимирующих функций, как линейные, так и нелинейные:

- $y = b_0 + b_1 x + b_2 x_2 + ... + b_n x_n$ (линейная);
- $y = b_0 \times x_1^{b_1} \times x_2^{b_2} \times x_3^{b_3} \times ... \times x_n^{b_n}$ (степенная);
- $y = e^{b_0 + b_1 x_1 + b_2 x_2 + ... + b_n x_n}$ (экспоненциальная);
- $y = \frac{1}{b_0 + b_1 x + b_2 x_2 + ... + b_n x_n}$ (гиперболическая) и др.

В связи с наиболее четкой интерпретацией параметров чаще всего используются линейная и степенная функции. Последняя получила наибольшее распространение в исследованиях спроса и потребления, а также в производственных функциях.

Следует отметить, что чем сложнее функция, тем ее параметры экономически менее интерпретируемы, а сама модель может выражать второстепенные взаимосвязи между переменными в ущерб остальным, поэтому главным правилом построения регрессионной модели является движение *от простого к сложному*. Кроме того, слишком большой набор объясняющих переменных в модели требует соответственно большого числа наблюдений, поэтому полиномы выше третьей степени в экономическом анализе используются редко.

С.А. Айвазян и В.С. Мхитарян по этому поводу замечают, что «следует добиваться компромисса между сложностью регрессионной модели и точностью ее оценивания. Из общих результатов математической статистики, относящихся к анализу

точности оценивания исследуемой модели при ограниченных объемах выборки, следует, что с увеличением сложности модели точность оценивания падает» [12].

Регрессионный анализ принято начинать с простейшего случая аппроксимации неизвестной функции - построения линейной модели регрессии.

2.3 Линейная парная регрессия. Метод наименьших квадратов

Рассмотрим линейную регрессию, для которой функция $f(X)$ линейна относительно оцениваемых параметров. Данная функция имеет вид:

$$\tilde{y} = \beta_0 + \beta_1 x . \tag{2.8}$$

Поскольку $\tilde{y}$ неизвестно, то перейдем к модели:

$$y_i = \beta_0 + \beta_1 x + \varepsilon_i , \qquad i = \overline{1,n} \tag{2.9}$$

где ε_i - регрессионные остатки (случайная ошибка модели регрессии), характеризующие расхождение между наблюденным значением y_i и "осредненным" значением $\tilde{y}_i$.

Для того чтобы линейная регрессионная модель называлась классической, необходимо, чтобы она удовлетворяла ряду условий (допущений), которые относятся к свойствам регрессоров и остатков. Рассмотрим основные предпосылки регрессионного анализа, известные как *условия Гаусса-Маркова*.

1. Математическое ожидание случайных отклонений ε_i в любом наблюдении равно нулю:

$$M\varepsilon_i = 0 , \; i = \overline{1,n} . \tag{2.10}$$

Или математическое ожидание зависимой переменной y_i равно линейной функции регрессии $M(y_i) = \beta_0 + \beta_1 x_i$ (нет систематических ошибок в измерении *y*).

Случайные отклонения могут быть положительными или отрицательными, но они не должны иметь систематического смещения ни в одном из двух возможных направлений.

2. Дисперсия остатков (или зависимой переменной y_i) постоянна для любого *i* (условие гомоскедастичности (равноизменчивости)):

$$D\varepsilon_i = M\varepsilon_i^2 = \sigma^2, \; i = \overline{1,n}. \qquad (2.11)$$

Иногда случайная величина будет больше, иногда меньше, но не должно быть априорной причины для того, чтобы она порождала большую ошибку в одних наблюдениях, чем в других. Величина этой дисперсии, конечно, неизвестна. Одной из задач регрессионного анализа является оценка стандартного отклонения случайной величины.

3. Случайные отклонения ε_i и ε_j (или переменные y_i, y_j) не должны быть коррелированны:

$$\operatorname{cov}(\varepsilon_i, \varepsilon_j) = M(\varepsilon_i \cdot \varepsilon_j) = 0, \; i \neq j, \; i = \overline{1,n} \; j = \overline{1,n}. \qquad (2.12)$$

Данное условие предполагает отсутствие систематической связи между значениями случайной величины в любых двух наблюдениях. Так, если величина случайного члена большая положительная величина, это не должно обусловливать систематическую тенденцию к тому, что она будет велика и положительна и в следующем наблюдении (также как и велика и отрицательна или мала и положительна, или мала и отрицательна). Случайные отклонения должны быть абсолютно независимы друг от друга.

4. В модели (2.9) регрессионные остатки ε_i (или зависимая переменная y_i) есть величина случайная, а величина x_i - величина детерминированная, не имеющая

случайной составляющей. Значение любой независимой переменной в каждом наблюдении должно считаться экзогенным, полностью определяемым внешними причинами, не учитываемыми в уравнении регрессии. Это условие можно записать в виде:

$$M(x_i \varepsilon_i) = 0 . \tag{2.13}$$

5. Наряду с условиями Гаусса-Маркова обычно также предполагается, что ε_i есть нормально распределенная случайная величина

$$\varepsilon_i \in N(0, \sigma^2) . \tag{2.14}$$

Если остатки нормально распределены, то так же будут распределены и параметры регрессии. Предположение о нормальности основывается на центральной предельной теореме, суть которой в следующем утверждении: если случайная величина является результатом взаимодействия большого числа других случайных величин, ни одна из которых не является доминирующей, то она будет иметь приблизительно нормальное распределение, даже если отдельные составляющие не имеют нормального распределения. Случайная величина ε_i и определяется несколькими факторами, которые не входят в уравнение регрессии. Поэтому даже если исследователь не располагает данными о распределении этих факторов или даже об их сущности, он имеет право предположить, что они распределены нормально.

При соблюдении перечисленных условий модель (2.9) называется классической нормальной линейной регрессионной моделью (*Classical Normal Linear Regression model*).

Оценкой модели (2.9) по выборке является уравнение регрессии

$$\tilde{y} = b_0 + b_1 x , \tag{2.15}$$

где b_0 - свободный член уравнения (постоянная);

b_1 - коэффициент (параметр) регрессии, измеряющий среднее отклонение результативного признака от его средней величины при отклонении факторного признака от своей средней на одну единицу его измерения (*вариация y*, приходящаяся на единицу *вариации x*).

Почти во всех случаях исследования связей в экономике свободный член уравнения регрессии не имеет элементарной интерпретации. Например, если он отрицателен, то его нельзя считать средним значением результативного признака при условии, что факторный (или факторные) равен нулю, ибо большинство результатов хозяйственной деятельности по своей природе могут быть только положительными величинами. Утверждение, что свободный член уравнения регрессии характеризует среднее значение результативного признака при нулевом значении факторного, обычно звучит неубедительно и тогда, когда свободный член положителен. Свободный член уравнения регрессии имеет элементарную экономическую интерпретацию только в том случае, если нулевое значение единственного фактора в парном уравнении или нулевые значения всех факторов множественного уравнении регрессии входят в область существования данной модели. Для множественного уравнения это практически немыслимо [13, с. 186].

Свободный член уравнения графически представляет отрезок ординаты (*y*) в системе прямоугольных координат. Параметр b_1 с точки зрения аналитической геометрии - угловой коэффициент, определяющий наклон линии регрессии по отношению к осям координат (рисунок 2.2).

Линии регрессии, как показано на рисунке 2.2, пересекаются в точке 0 $(\bar{x}, \bar{y})$, соответствующей средним арифметическим значениям корреляционно связанных друг с другом признаков *Y* и *X*. Линия *AB*, проходящая через эту точку, изображает полную (функциональную) зависимость между переменными величинами *Y* и *X*. Чем сильнее связь между *Y* и *X*, тем ближе линии регрессии к *AB*, и, наоборот, чем слабее связь между варьирующими признаками, тем более удалены линии рег-

рессии от *AB*. При отсутствии связи между признаками линии регрессии оказываются под прямым углом по отношению друг к другу.

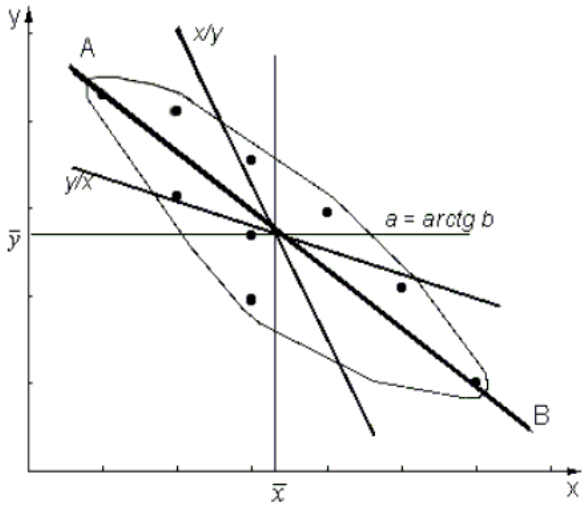


Рисунок 2.2 - Линии регрессии *Y* по *X* и *X* по *Y* в системе прямоугольных координат

Уравнение регрессии тем лучше описывает зависимость, чем меньше рассеяние диаграммы и больше теснота взаимосвязи.

С самого начала необходимо признать, что мы никогда не сможем рассчитать истинные значения β_0 и β_1 при попытке построить прямую и определить положение линии регрессии. Мы можем получить только *оценки*, которые могут быть хорошими или плохими. В результате случайного совпадения оценки могут быть абсолютно точными, но даже в этом случае у нас не будет способа узнать, что они абсолютно точны. Это справедливо и при использовании более совершенных методов.

Надежность получаемых по уравнению регрессии расчетных значений во многом определяется рассеянием наблюдений вокруг линии регрессии. Минимизировать сумму остатков при выполнении определенных условий позволяет обычный *метод наименьших квадратов* (МНК или OLS – *ordinary least squares*).

Воздействие неучтенных факторов (характеристика меры рассеяния) определяется с помощью дисперсии возмущений (ошибок) или остаточной дисперсии σ^2,

несмещенной оценкой которой является выборочная остаточная дисперсия (дисперсия относительно регрессии):

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}, \qquad (2.16)$$

где $\tilde{y}_i$ - групповая средняя, полученная по уровню уравнению регрессии;

$e_i = \tilde{y}_i - y_i$ - выборочная оценка возмущения (случайного члена) ε_i или остаток регрессии.

В математической статистике для получения несмещенной оценки дисперсии случайной величины соответствующую сумму квадратов отклонений делят не на число наблюдений *n*, а на число степеней свободы (*degress of freedom – d.f.*), равное разности между числом независимых наблюдений случайной величины и числом уравнений, связывающих эти наблюдения. При определении двух параметров прямой из системы нормальных уравнений (они будут рассмотрены ниже) две степени свободы теряются, поэтому в знаменателе формулы (2.16) стоит число степеней свободы *n*-2.

Теорема Гаусса-Маркова дает ответ на вопрос, являются ли оценки параметров β_0, β_1 и σ^2 - b_0, b_1 и s^2 наилучшими. Если регрессионная модель удовлетворяет предпосылкам МНК, то оценки b_0 и b_1 имеют наименьшую дисперсию в классе всех линейных несмещенных оценок (*Best Linear Unbiased Estimator*, или *BLUE* - наилучшая линейная несмещенная оценка). Свойства оценок параметров классической регрессионной модели – требования их состоятельности, несмещенности и эффективности более подробно будут рассмотрены в параграфе, посвященном классической линейной модели множественной регрессии.

Пример линейной зависимости представлен на рисунке 2.1. На нем видно, что через точки фактических значений можно провести бесчисленное множество прямых, но для качественной аппроксимации нужно выбрать одну, дающую наилучшее

приближение эмпирическим данным. Для решения этой задачи и применяется подход, получивший название метода наименьших квадратов.

Согласно МНК неизвестные параметры уравнения b_0 и b_1 выбираются таким образом, чтобы сумма квадратов отклонений фактических значений результативного признака от значений, найденных по уравнению регрессии (расчетных, теоретических, модельных) была минимальной:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (b_0 + b_1 x_i - y_i)^2 \to \min. \qquad (2.17)$$

Из множества линий регрессии мы выбираем такую, чтобы сумма квадратов расстояний по вертикали между точкой и этой линией была минимальной.

Обозначим сумму квадратов остатков модели через *S*: $S = \sum_{i=1}^{n} \varepsilon_i^2$ и определим минимум функции. На основании необходимого условия экстремума функции двух переменных $S = S(b_0, b_1)$ частные производные каждого параметра приравняем к нулю:

$$\begin{cases} \dfrac{dS}{da_0} = -2\sum_{i=1}^{n} (b_0 + b_1 x_i - y_i) = 0; \\ \dfrac{dS}{da_1} = -2\sum_{i=1}^{n} (b_1 + b_1 x_i - y_i) x_i = 0. \end{cases} \qquad (2.18)$$

После преобразований получим систему нормальных уравнений для расчета параметров линейной регрессии:

$$\begin{cases} b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i; \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i. \end{cases} \qquad (2.19)$$

Разделим обе части уравнений (2.19) на *n*:

$$\begin{cases} b_0 + b_1\bar{x} = \bar{y}; \\ b_0\bar{x} + b_1\bar{x}^2 = \overline{xy}. \end{cases} \tag{2.20}$$

Подставив значение $b_0 = \bar{y} - b_1\bar{x}$ из первого уравнения системы в уравнение (2.15), получим

$$\tilde{y} = \bar{y} - b_1\bar{x} + b_1 x \ , \tag{2.21}$$

$$\tilde{y} - \bar{y} = b_1(x - \bar{x}) \tag{2.22}$$

или

$$\bar{y} = b_0 + b_1\bar{x} \, . \tag{2.23}$$

Согласно (2.23) получаем, что линия регрессии проходит через точку $(\bar{x}, \bar{y})$.

Решая систему (2.20), найдем значение коэффициента регрессии b_1

$$b_1 = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{x^2} - \bar{x}^2}, \tag{2.24}$$

Числитель формулы (2.24) представляет собой выборочную ковариацию - $Cov(X,Y)$, а знаменатель – выборочную дисперсию переменной X - s_x^2.

Оценивая тесноту корреляционной зависимости *Y* от *X*, может показаться, что «хорошим» измерителем связи в линейном уравнении является параметр b_1, т.к. он характеризует среднее изменение результативного признака при изменении факторного на единицу его измерения. Однако, если факторный признак увеличить/уменьшить в *n* раз, то и параметр b_1 также увеличится/уменьшится. Чтобы данные по различным характеристикам были сравнимы между собой, в качестве единицы измерения переменной используют ее среднее квадратическое отклонение *s*. Представим уравнение (2.22) в виде:

$$\frac{\tilde{y} - \bar{y}}{s_y} = b_1 \frac{s_x}{s_y} \frac{x - \bar{x}}{s_x} \, . \tag{2.25}$$

Величина $r = b_1 \frac{s_x}{s_y}$ показывает, на сколько величин s_y изменится в среднем *Y*, когда *X* увеличится на одно s_x и носит название выборочного коэффициента корреляции[1]. Есть и другие модификации коэффициента корреляции:

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n s_x s_y}, \tag{2.26}$$

$$r = \frac{\overline{xy} - \overline{y}\overline{x}}{\sqrt{\overline{y^2} - (\overline{y})^2}\sqrt{\overline{x^2} - (\overline{x})^2}}. \tag{2.27}$$

Коэффициент корреляции применяется для обобщенного описания отношений между двумя переменными и при наличии двумерного нормального распределения является мерой линейной согласованности между переменными, их взаимного *варьирования*.

В практике статистического анализа не являются исключением случаи, когда с помощью корреляционного анализа обнаруживают существование достаточно сильной «зависимости» признаков, в действительности не имеющих причинной связи между собой. Такие корреляции принято называть *ложными* или бессмысленными. Как правило, бессмысленные корреляции получают при коррелировании временных рядов двух признаков, не связанных причинной зависимостью. В дальнейшем будем полагать, что между рассматриваемыми переменными существует причинная зависимость и, следовательно, применение теории корреляции имеет логическое основание [14, с. 228].

[1] Термин «корреляция» впервые применил французский палеонтолог Ж. Кювье, который вывел «закон корреляции частей и органов животных» (этот закон позволяет восстанавливать по найденным частям тела облик всего животного). Общая идея корреляции в значительной степени была обоснована Френсисом Гальтоном в 80-х годах XIX в. (не просто «связь» – *relation*, а «как бы связь» – *corelation*). Но тот показатель корреляции, который используется в настоящее время (коэффициент корреляции «произведения моментов») был введен Карлом Пирсоном в 1898 г.

Чем ближе точки корреляционного поля к прямой регрессии Y по X, тем выше значение коэффициента корреляции и теснее связь между переменными. Возможные значения коэффициента варьируются в пределах от -1 до +1, то есть от полной отрицательной до полной положительной корреляции. Принято считать, что если коэффициент корреляции по модулю находится в пределах: $r<0,3$ – связь слабая; $0,3>r<0,7$ – средняя; $r\geq 0,7$ – сильная связь. Если значение коэффициента корреляции (и параметра b_1) в уравнении парной линейной регрессии положительно, то связь называют прямой.

Это значит, что результативный и факторный признаки изменяются в одном направлении – увеличение/уменьшение факторной переменной ведет к увеличению/уменьшению условной (групповой) средней результативной переменной. Если же коэффициент корреляции (и параметра b_1) отрицателен, то направления изменений признаков обратные и связь называется обратной.

В случае равенства значения коэффициента по модулю единице, корреляционная связь представляет линейную функциональную зависимость. Полная корреляция соответствует случаю, когда все наблюдения находятся точно на прямой линии, имеющей положительный или отрицательный наклон. Если значение коэффициента корреляции близко к нулю, это свидетельствует об отсутствии сколько-нибудь существенной тенденции к совместному изменению значений x и y, а в случае равенства коэффициента нулю линейная корреляционная связь отсутствует, линия регрессии параллельна оси $0x$.

Графически примеры различных значений коэффициента корреляции отражены на рисунке 2.3.

Следует отметить в связи с вышесказанным одно важное замечание.

Из того, что значение коэффициента корреляции высоки, *нельзя* вывести ни одно из следующих утверждений:

1) Y зависит от X;

2) X зависит от Y;

3) X и Y совместно зависят от какой-то третьей переменной.

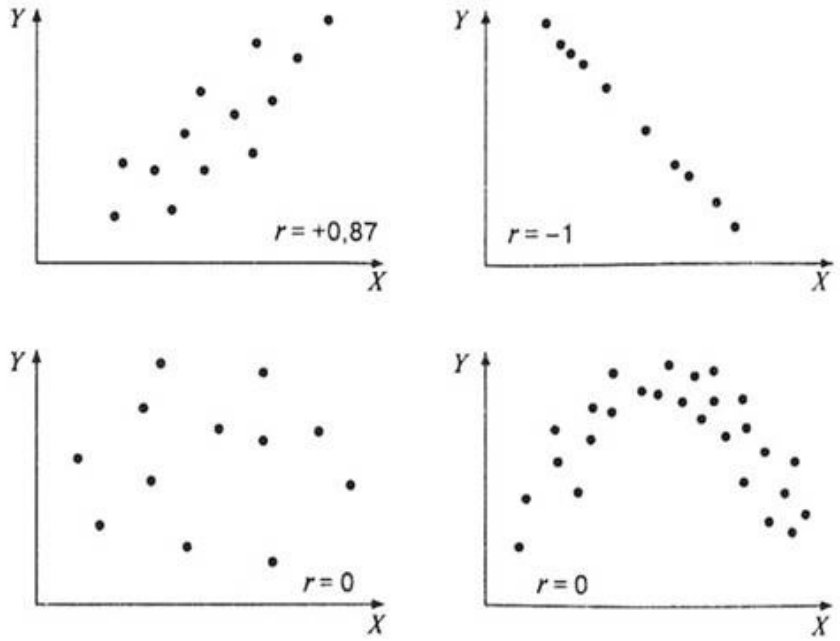


Рисунок 2.3 – Примеры поля корреляции при различных значениях коэффициента корреляции

Величина статистического показателя r абсолютно ничего не говорит о направлении причинно-следственных связей (о том, какая из рассматриваемых переменных является независимой и определяет поведение другой переменной). Эти вопросы должны быть решены в ходе теоретического анализа, т.е. априори. Высокая корреляция может свидетельствовать лишь о линейной форме связи [15, с. 144].

Выборочный коэффициент корреляции является непосредственной оценкой генерального коэффициента корреляции ρ между X и Y лишь в случае двумерного нормального закона распределения случайных величин X и Y.

В других случаях (когда распределения X и Y отличаются от нормального, одна из исследуемых величин, например, X не является случайной и т.п.) выборочный коэффициент корреляции не следует рассматривать как строгую меру взаимосвязи переменных [16, с. 159].

Рассмотрим в качестве примера зависимость ввода в действие жилых домов на 1000 человек населения (м. кв.) (Y) от объема инвестиций в жилищное строительство на душу населения (р.) (X). Данные представлены по городам Оренбургской области за 2010 г. (таблица 2.2).

Таблица 2.2 – Исходные данные для построения парного уравнения регрессии

Города Оренбургской области	Инвестиции в жилищное строительство на душу населения, р.	Ввод в действие жилых домов на 1000 чел. населения, м. кв.
г.Абдулино	6361,7	378,1
г.Бугуруслан	4200,8	374,4
г.Бузулук	5910,4	398,4
г.Гай	2686,3	237,0
г.Кувандык	3614,9	396,6
г.Медногорск	2931,5	88,6
г.Новотроицк	2176,9	119,0
г.Оренбург	2613,4	288,8
г.Орск	2596,5	150,0
г.Соль-Илецк	9341,3	764,0
г.Сорочинск	4387,8	313,0
г.Ясный	1109,0	65,9
Источник: Города и районы Оренбургской области: Стат.сб. / Территориальный орган Федеральной службы государственной статистики по Оренбургской области. – Оренбург, 2011.		

Построим диаграмму рассеяния, чтобы сделать предположения о наличии зависимости между переменными *X* и *Y* (рисунок 2.4).

Расположение точек на поле корреляции позволяет предположить наличие линейной регрессионной зависимости между переменными *X* и *Y*. Составим расчетную таблицу для вычисления параметров уравнения (таблица 2.3).

Получим:

$$b_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2} = \frac{1554115{,}73 - 3994{,}21 \times 297{,}82}{20665108{,}22 - 3994{,}21^2} = 0{,}077378;$$

$$b_0 = \overline{y} - b_1 \overline{x} = 297{,}82 - 0{,}077378 \times 3994{,}21 = -11{,}24398.$$

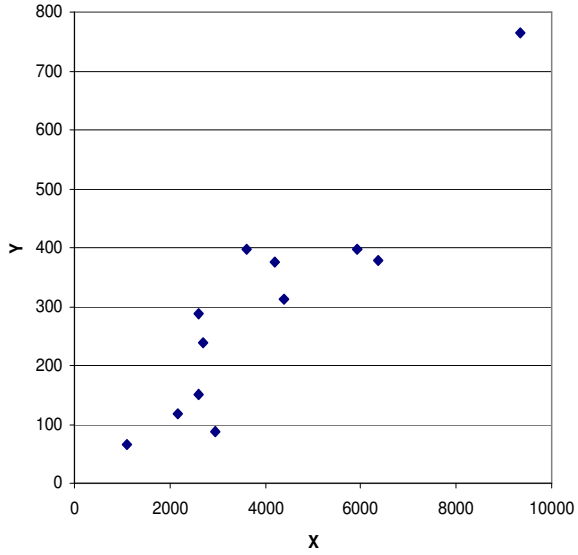


Рисунок 2.4 – Диаграмма рассеяния переменных *X* и *Y*

Таблица 2.3 - Расчетная таблица для вычисления параметров уравнения парной линейной регрессии

№ наблюдения	x	y	x^2	y^2	xy	$\tilde{y}$	$(y_i - \tilde{y})^2$	$(x_i - \bar{x})^2$
1	6361,7	378,1	40471226,9	142959,6	2405358,8	481,0	10590,5	5605016,8
2	4200,8	374,4	17646720,6	140175,4	1572779,5	313,8	3671,1	42680,1
3	5910,4	398,4	34932828,2	158722,6	2354703,4	446,1	2274,3	3671790,5
4	2686,3	237	7216207,7	56169,0	636653,1	196,6	1630,5	1710624,2
5	3614,9	396,6	13067502,0	157291,6	1433669,3	268,5	16417,3	143874,8
6	2931,5	88,6	8593692,3	7850,0	259730,9	215,6	16126,5	1129349,0
7	2176,9	119	4738893,6	14161,0	259051,1	157,2	1459,2	3302609,6
8	2613,4	288,8	6829859,6	83405,4	754749,9	191,0	9568,8	1906631,7
9	2596,5	150	6741812,3	22500,0	389475,0	189,7	1573,7	1953588,6
10	9341,3	764	87259885,7	583696,0	7136753,2	711,6	2748,9	28591389,3
11	4387,8	313	19252788,8	97969,0	1373381,4	328,3	233,5	154914,4
12	1109	65,9	1229881,0	4342,8	73083,1	74,6	75,1	8324427,1
Итого	47930,5	3573,8	247981298,6	1469242,3	18649388,7	3573,8	66369,5	56536896,1
В среднем	3994,21	297,82	20665108,22	122436,86	1554115,73	297,82	-	-

Уравнение регрессии *Y* по *X* имеет вид:

$$\tilde{y} = -11{,}24398 + 0{,}077378x\,.$$

В 7-й графе таблицы 2.3. рассчитаны теоретические значения результативной переменной по полученному уравнению регрессии. Данные, приведенные в графе 8, понадобятся для расчета дисперсии ошибки, а данные гр. 9 – для оценки значимости параметра.

На рисунке 2.5 отражены фактические значения и теоретическая линия регрессии.

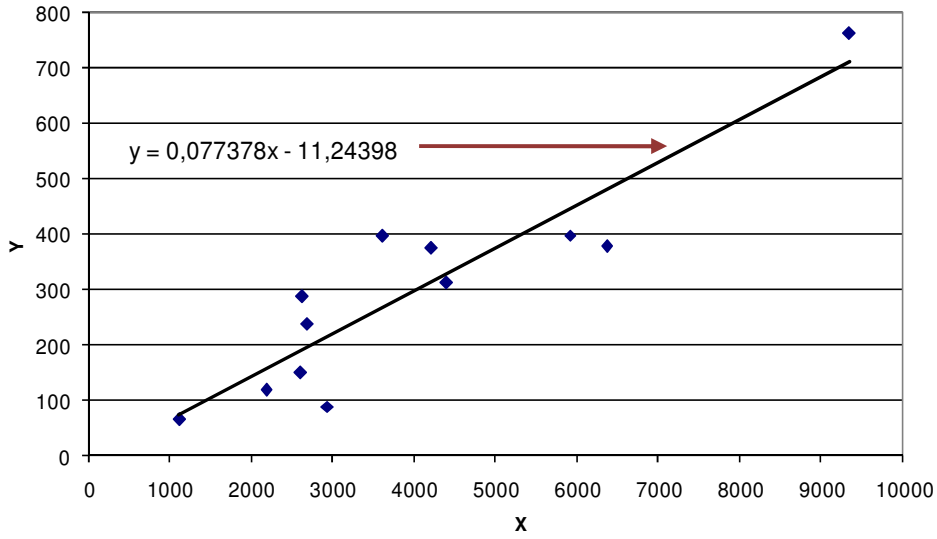


Рисунок 2.5 – Фактические уровни и теоретическая линия парной линейной регрессии

Выше мы говорили об оценке дисперсии ошибки (2.16), которая служит мерой среднего рассеяния наблюденных значений вокруг подобранной линии регрессии. Эта оценка может дать нам представление о возможных достоинствах выбранной регрессии. Остаточная дисперсия составила:

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{n-2} = \frac{66369{,}47}{10} = 6636{,}947 \approx 6636{,}9 \text{ м}^2.$$

Среднее квадратическое отклонение:

$$s = \sqrt{s^2} = \sqrt{6636{,}947} = 81{,}47 \text{ м}^2.$$

Это значит, что фактический ввод жилья на 1000 человек населения отличается от теоретического (рассчитанного по модели) на 81,47 м^2.

Рассчитаем выборочный коэффициент корреляции:

$$r = \frac{\overline{xy} - \overline{y}\,\overline{x}}{\sqrt{\overline{y^2} - (\overline{y})^2}\sqrt{\overline{x^2} - (\overline{x})^2}} = \frac{1554115{,}73 - 3994{,}21 \times 297{,}82}{\sqrt{122436{,}86 - 297{,}82^2}\sqrt{20665108{,}22 - 3994{,}21^2}} = 0{,}9144.$$

Определим средний коэффициент эластичности, показывающий, на сколько процентов в среднем по совокупности изменится результат *y* от своей средней величины при изменении фактора *x* на 1 % от своего среднего значения. Для парной прямолинейной зависимости коэффициент эластичности рассчитывается по формуле:

$$\overline{E} = b_1 \frac{\overline{x}}{\overline{y}}. \qquad (2.28)$$

$$\overline{E} = 0{,}077378 \times \frac{3994{,}21}{297{,}82} = 1{,}04 \text{ %}.$$

Полученное значение говорит о том, что при увеличении среднедушевых инвестиций в жилищное строительство в среднем на 1 %, ввод в действие жилья на 1000 человек увеличивается на 1,04 %.

Существуют два этапа интерпретации уравнения регрессии. Первый этап состоит в словесном истолковании уравнения так, чтобы это было понятно человеку, не являющемуся специалистом в области статистики. На втором этапе необходимо решить, следует ли ограничиться этим или провести более детальное исследование [6, с. 65].

Экономическая интерпретация полученного уравнения следующая: при увеличении среднедушевых инвестиций в жилищное строительство на 1000 р., ввод в

действие жилых домов на 1000 жителей увеличится *в среднем* на 77,4 м^2. Для более простого выражения результатов при интерпретации уравнения в качестве единиц измерения для *y* использованы не рубли, а тысячи рублей.

2.4 Оценка значимости и доверительные интервалы уравнения регрессии и его параметров

Прежде чем утвердиться в возможности применения полученного уравнения регрессии в экономическом анализе и прогнозировании, необходимо оценить качество модели в целом и ее параметров.

Существует четыре основных способа, помогающих решить этот вопрос [15, с. 28]:

1) анализ дисперсии;

2) построение доверительных интервалов для неизвестного углового коэффициента прямой - β;

3) определение области прогноза (двумерной) на плоскости *XY*;

4) проверка существенности выборочного коэффициента корреляции.

При гипотезе парной корреляционной зависимости первый, второй и четвертый методы оказываются полностью эквивалентны. Третий метод предназначен для получения прогнозов на будущее для характеристики последующего (за пределами выборки) поведения зависимой переменной *Y*.

Дисперсионный анализ является основой проверки значимости уравнения регрессии. Основная идея дисперсионного анализа заключается в том, что общая сумма квадратов отклонений зависимой переменной от средней (*SST* - sum of squares total) равна сумме двух дисперсий – сумме квадратов, обусловленных регрессией (*SSE* - sum of squares explained) и остаточной сумме квадратов, которая характеризует влияние неучтенных в модели факторов (*SSR* - sum of squares residual)[1]:

[1] Теория, лежащая в основе этого подхода, описывается в работе А. Муда и Ф. Грейбилла [Mood A.M., Graybill F.P. Intoduction to the Theory of Statistics, 1963].

$$SST = SSE + SSR \qquad (2.29)$$

Такое представление суммы квадратов отклонений *SST* позволяет непосредственно перейти к статистической проверке рассматриваемой регрессии. Общая схема дисперсионного анализа представлена в таблице 2.4.

Таблица 2.4 – Схема дисперсионного анализа

Компоненты дисперсии	Число степеней свободы (d.f.)	Сумма квадратов (SS)	Среднее значение суммы квадратов (MS)**
1	2	3	4
Регрессия	m-1	$SSE = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2$	$s_R^2 = \frac{SSE}{m-1}$
Остаточная	n-m	$SSR = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$	$s^2 = \frac{SSR}{n-m}$
Общая	n-1	$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$	-

В первом столбце содержатся заголовки приводимых в таблице строк; во втором столбце записаны числа степеней свободы, характеризующие суммы квадратов на пересечении следующего столбца и соответствующей строки; в третьем столбце представлены суммы квадратов, соответствующие источнику их «происхождения»; в четвертом столбце отражено соответствующее строкам среднее значение суммы квадратов (частное от деления гр. 3 на гр. 2). Число оцениваемых параметров обозначено через m, число наблюдений – n.

Средние квадраты s_R^2, s^2 являются несмещенными оценками дисперсии *Y*, обусловленной регрессией (факторной переменной *X*) и воздействием неучтенных в модели факторов. Если линейная зависимость между результативной и факторной переменными отсутствует, то случайные величины s_R^2, s^2 имеют χ^2 - распределение соответственно с (m-1) и (n-m) степенями свободы, а их отношение – *F*-

распределение с теми же степенями свободы (значения *F*-критерия приведены в приложении В).

F-статистика для проверки качества оценивания регрессии записывается как отношение объясненной суммы квадратов отклонений (в расчете на одну независимую переменную к остаточной сумме квадратов) в расчете на одну степень свободы:

$$F = \frac{SSE(n-m)}{SSR(m-1)} = \frac{s_R^2}{s^2} > F_{\alpha,k_1,k_2}, \tag{2.30}$$

где F_{α,k_1,k_2} - табличное значение *F*-критерия Фишера-Снедекора на уровне значимости α при $k_1=m-1$, $k_2=n-m$ степенях свободы.

Если данное неравенство выполняется, то нулевая гипотеза о незначимости уравнения отклоняется, а имеющееся «объяснение» поведения результативной переменной лучше, чем можно было бы получить чисто случайно. Другими словами, значение критерия показывает, насколько лучше регрессия оценивает значение результативной переменной по сравнению с ее средней.

В парной линейной регрессии число оцениваемых параметров $m=2$, поэтому значимость уравнения принимается при условии

$$F = \frac{SSE(n-2)}{SSR} > F_{\alpha,k_1,k_2}. \tag{2.31}$$

Эффективной оценкой регрессионной модели, мерой ее качества и характеристикой прогностической силы выступает коэффициент детерминации R^2, показывающий, какая часть вариации результативной переменной обусловлена вариацией факторной переменной:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}. \tag{2.32}$$

Вследствие того, что $0 \le SSE \le SST$, получаем: $0 \le R^2 \le 1$.

Максимальное значение коэффициента детерминации равно единице. В этом случае линия регрессии точно соответствует всем наблюдениям, так что $\tilde{y}_i = y_i$ для всех *i* и все остатки равны нулю. Если в выборке отсутствует видимая связь между *Y* и *X*, то величина коэффициента детерминации будет близка к нулю. Коэффициент R^2 следует рассматривать только при наличии в переменной свободного члена, так как только в этом случае верны равенства (2.29) и (2.32).

Допустимые значения для коэффициента детерминации следующие: 0,01-0,09 – связь слабая; 0,09-0,49 – связь средняя; 0,49-1,00 – связь достаточно сильная (использование полученной регрессионной модели в анализе теоретически обосновано).

В случае парной линейной регрессии коэффициент детерминации равен квадрату коэффициента корреляции ($R^2 = r^2$). Нетрудно заметить связь между коэффициентом детерминации и *F*-критерием Фишера:

$$F = \frac{R^2 (n-m)}{(1-R^2)(m-1)}. \qquad (2.33)$$

Хотя коэффициент детерминации и критерий *F* связаны между собой, они используются в разных целях: первый применяется для измерения степени согласованности оцененной модели с имеющимися данными выборочного наблюдения, а второй – для проверки гипотезы о том, что ни одна из объясняющих переменных не связана линейно с истинным значением Y_i [17, с. 16].

Поскольку *F*-критерий представляет собой соотношение дисперсий, он обладает некоторыми преимуществами по сравнению с коэффициентом детерминации, т.к. позволяет принять во внимание степени свободы в числителе и знаменателе (2.33).

Недостатком коэффициента детерминации является то, что при добавлении в уравнение новой независимой переменной, не имеющей отношения к анализируемой связи, - переменной, для которой истинное значение параметра равно нулю, R^2 в

лучшем случае сохранит свою величину, либо будет наблюдаться его увеличение, обусловленное использованием выборочных наблюдений. Данный недостаток можно устранить, внеся при исчислении R^2 поправку на число степеней свободы:

$$\overline{R}^2 = 1 - \frac{n-1}{n-m}\left(1 - R^2\right) \tag{2.34}$$

Этот коэффициент носит название скорректированного коэффициента детерминации. При добавлении переменных $\overline{R}^2$ будет увеличиваться только в том случае, если рост R^2 будет «перевешивать» увеличение количества переменных, поэтому скорректированный коэффициент детерминации можно использовать в качестве критерия для принятия решения о включении или невключении в модель дополнительных переменных.

В нашем примере коэффициент детерминации равен:

$$R^2 = r^2 = 0{,}9144^2 = 0{,}836 .$$

Следовательно, 83,6 % вариации результативного признака «Ввод в действие жилых домов на 1000 человек» обусловлено вариацией фактора «Среднедушевые инвестиции в жилищное строительство». Скорректированный коэффициент детерминации составит:

$$\overline{R}^2 = 1 - \frac{12-1}{12-2}\left(1 - 0{,}836\right) = 0{,}8196 .$$

Расчетная величина F-критерия:

$$F = \frac{0{,}836(12-2)}{(1-0{,}836)(2-1)} = 50{,}98 .$$

Табличное значение *F*-критерия на уровне значимости $\alpha = 0{,}05$ с числом степеней свободы k_1=1, k_2=10 составляет 4,96. Полученные расчеты позволяют сделать вывод о значимости полученного уравнения и возможности его применения в экономическом анализе и прогнозировании.

Наряду с проверкой уравнения парной линейной регрессии в целом, необходимо оценить значимость его параметров и коэффициента корреляции. При проверке значимости коэффициентов регрессии и корреляции в качестве нулевой гипотезы H_0 берется предположение о равенстве соответствующего коэффициента нулю для всей рассматриваемой совокупности. При нормально распределенных независимых остатках формула стандартной ошибки для углового коэффициента имеет вид[1]:

$$s_{b_1} = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}. \qquad (2.35)$$

Эту формулу можно применять непосредственно, не обращаясь к таблицам анализа дисперсии для построения доверительных интервалов углового коэффициента β.

Далее находим *t*-статистику Стьюдента, разделив коэффициент регрессии на его стандартную ошибку. Коэффициент признается значимым при условии

$$t_{b_1} = \frac{b_1}{s_{b_1}} > t_{\alpha;n-m}. \qquad (2.36)$$

В рассматриваемом примере:

$$s_{b_1} = \sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}} = \frac{s^2}{\sqrt{\sum_{n=1}^{n}(x_i - \bar{x})^2}} = \sqrt{\frac{6636{,}947}{56536905{,}1}} = 0{,}010835.$$

[1] Доказательства эквивалентных выражений можно найти в работе Дж. Томаса [Thomas J. James. An Introduction to Statistical Analysis for Economists, 1983].

$$t_{b_1} = \frac{0{,}077378}{0{,}010835} = 7{,}14 > t_{0{,}05;10} = 2{,}2281 .$$

Следовательно, параметр уравнения значим (табулированные значения t – критерия приведены в таблице Г.1 приложения Г).

Стандартная ошибка коэффициента корреляции произведения моментов r для выборки объемом в n наблюдений из совокупности, распределение которой близко к двумерному нормальному распределению и обладает нулевой корреляцией для всех ее наблюдений, рассчитывается с помощью отношения $1-r^2/\sqrt{n-1}$. Выборочное распределение r будет близко к нормальному для выборок большого объема ($n \geq 100$).

Если объем выборки менее 100 наблюдений, можно воспользоваться тем, что величина $r\sqrt{\frac{n-2}{1-r^2}}$ подчиняется t-распределению Стьюдента с (n-2) степенями свободы, если значение ρ (коэффициент корреляции генеральной совокупности), которое рассчитано с участием всех наблюдений исследуемой совокупности, равно нулю.

Как показал в 1915 г. Р. Фишер, проверка гипотезы о не равном нулю значении коэффициента ρ в двумерной нормально распределенной совокупности может быть произведена на основе того факта, что для различных выборочных значений r распределение величины $\frac{1}{2}\ln\frac{1+r}{1-r}$ с высокой степенью точности аппроксимируется нормальным распределением со средней $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ и дисперсией $\frac{1}{\sqrt{n-3}}$. Таким образом, для проверки нулевой гипотезы о не равном нулю коэффициенте корреляции может быть использовано это выборочное распределение [18, с. 354-355].

Значения $\frac{1}{2}\ln\frac{1+r}{1-r}$ (z -преобразование) приведены в приложении Д.

Для оценки значимости коэффициента корреляции r исходим из того, что при отсутствии корреляционной связи статистика

$$t = \frac{|r|\sqrt{n-2}}{1-r^2} \tag{2.37}$$

имеет *t*-распределение Стьюдента с *n*-2 степенями свободы. Гипотеза H_0 о равенстве генерального коэффициента корреляции нулю ($H_0 : \rho = 0$) отвергается на уровне значимости α, если выполняется следующее условие

$$|t| = \frac{|r|\sqrt{n-2}}{1-r^2} > t_{1-\alpha;n-2}. \tag{2.38}$$

Получим:

$$|t| = \frac{0{,}9144\sqrt{10}}{1-0{,}8361} = 17{,}64 > t_{0{,}05;10} = 2{,}2281.$$

Значение *t*-критерия Стьюдента превышает его критическое значение для значимости 0,05, следовательно, связь установлена надежно.

Утвердившись в значимости уравнения, его параметров и коэффициента корреляции, необходимо построить доверительные интервалы[1] прогноза для функции регрессии, индивидуальных значений зависимой переменной и параметров. Как мы уже отмечали, оценка, полученная по выборочным данным, не будет точно равна соответствующему значению исходной совокупности. Если выборочное распределение близко к нормальному, то с использованием стандартной ошибки полученных оценок можно определить доверительные интервалы.

Доверительный интервал для функции регрессии – это интервал для условного математического ожидания $M_x(Y)$, который с заданной доверительной вероятностью $\gamma = 1-\alpha$ накрывает неизвестное значение $M_x(Y)$.

Дисперсия групповой средней $\tilde{y}$ представляет собой выборочную оценку $M_x(Y)$. Представим уравнение регрессии (2.22) в виде:

[1] Подробно этот вопрос рассмотрен в работе Т. Уоннакота и Р. Уоннакота [Wonnacott, Wonnacott, 1985, глава 8].

$$\tilde{y} = \bar{y} + b_1 (x - \bar{x}) . \qquad (2.39)$$

Тогда дисперсия групповой средней:

$$\sigma_{\tilde{y}}^2 = \sigma_{\bar{y}}^2 + \sigma_{b_1}^2 (x - \bar{x})^2 . \qquad (2.40)$$

Дисперсия выборочной средней $\bar{y}$ равна

$$\sigma_{\bar{y}}^2 = \sigma^2 \left(\frac{\sum_{i=1}^{n} y_i}{n} \right) = \frac{\sum_{i=1}^{n} \sigma_{y_i}^2}{n^2} = \frac{\sum_{i=1}^{n} \sigma^2}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} . \qquad (2.41)$$

Для нахождения $\sigma_{b_1}^2$ коэффициент регрессии b_1 представим как

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} . \qquad (2.42)$$

Получим далее:

$$\sigma_{b_1}^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sigma^2}{\left(\sum_{i=1}^{n} (x_i - \bar{x})^2 \right)^2} = \frac{\sigma^2}{\sum_{i=1}^{1} (x_i - \bar{x})^2} . \qquad (2.43)$$

Тогда оценка дисперсии (2.40) с учетом (2.41) и (2.43) имеет вид:

$$s_{\tilde{y}}^2 = s^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right) . \qquad (2.44)$$

Согласно предпосылкам (1-5) Гаусса-Маркова статистика $t = \dfrac{\tilde{y} - M_x(Y)}{s_{\tilde{y}}}$ имеет *t*-распределение Стьюдента с $k=n-2$ степенями свободы. Доверительный интервал для условного математического ожидания $M_x(Y)$:

$$\tilde{y} - t_{1-\alpha;k} \times s_{\tilde{y}} \le M_x(Y) \le \tilde{y} + t_{1-\alpha;k} \times s_{\tilde{y}} . \qquad (2.45)$$

Построим доверительные интервалы функции регрессии для нашего примера (используем ППП Statistica) (рисунок 2.6).

Рисунок наглядно иллюстрирует, что по мере удаления факторной переменной *x* от ее средней, величина доверительного интервала увеличивается. Отсюда следует, что прогнозировать (экстраполировать) зависимую переменную *Y* с заданной вероятностью можно лишь в случае, когда значение *x* факторной переменной *X* не выйдет за диапазон ее значений по выборке, а наилучшие результаты прогноза будут в случае, когда значение *x* будет находиться в центре области наблюдений *X*.

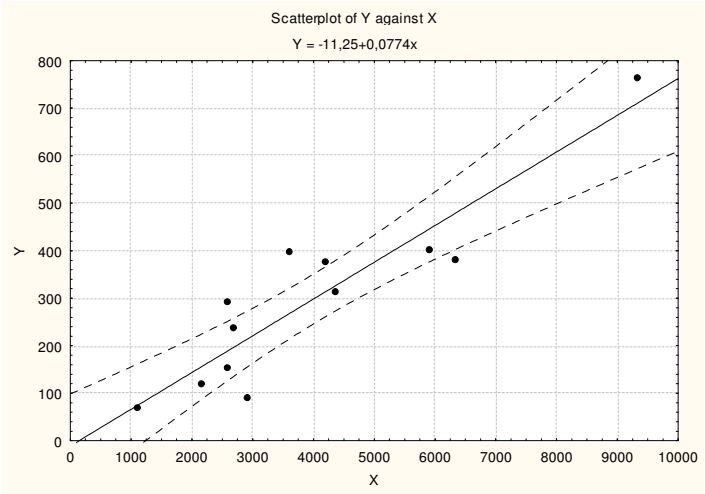


Рисунок 2.6 - График доверительных границ функции парной регрессии

При определении доверительного интервала для индивидуального значения признака y^* нужно учесть еще один источник вариации – оценку суммарной дисперсии. Тогда оценка индивидуальных значений y_0 при $x=x_0$ равна:

$$s^2_{\tilde{y}_0} = s^2\left(1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right). \qquad (2.46)$$

Интервальную оценку прогнозного значения для y_0^* определим по формуле

$$\tilde{y}_0 - t_{1-\alpha;n-2}\times s_{\tilde{y}_0} \le y_0^* \le \tilde{y}_0 + t_{1-\alpha;n-2}\times s_{\tilde{y}_0}. \qquad (2.47)$$

Выше мы привели формулу расчета стандартной ошибки параметра регрессии (2.35). Интервальная оценка параметра β_1 на уровне значимости α:

$$b_1 - t_{1-\alpha;n-2}\times\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}} \le \beta_1 \le b_1 + t_{1-\alpha;n-2}\times\sqrt{\frac{s^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}. \qquad (2.48)$$

Для нашего примера:

$$0{,}077378 - 2{,}2281\times 0{,}010835 \le \beta_1 \le 0{,}077378 + 2{,}2281\times 0{,}010835;$$

$$0{,}05324 \le \beta_1 \le 0{,}10152.$$

Из полученного выражения можно сделать вывод, что с надежностью 95 % при изменении среднедушевых инвестиций в жилищное строительство на 1000 р., ввод в действие жилья на 1000 человек будет изменяться на величину, заключенную в интервале от 53 до 101,5 м^2.

2.5 Вопросы для самоконтроля

1. Сформулируйте задачи и условия проведения регрессионного анализа.

2. С какими проблемами приходится сталкиваться при спецификации регрессионной модели?

3. Каковы основные предпосылки метода наименьших квадратов?

4. Каким образом оценивается значимость регрессионного уравнения в целом?

5. Приведите схему анализа значимости параметров регрессионного уравнения и построения доверительных интервалов функции регрессии и параметров.

2.6 Тесты

1. По отношению к выбранной спецификации модели все экономические переменные объекта подразделяются на два типа:

а) эндогенные и экзогенные;

б) дискретные и непрерывные;

в) случайные и детерминированные.

2. Чтобы получить качественные оценки множественного уравнения регрессии, необходимо выполнение следующих предпосылок МНК:

а) отклонения ε_i не должные коррелировать друг с другом;

б) отклонения ε_i должны иметь биномиальный закон распределения;

в) отклонения ε_i должны иметь показательный закон распределения;

г) отклонения ε_i должны быть нормально распределенными случайными величины с нулевым математическим ожиданием и постоянной дисперсией.

3. Если в уравнении регрессии имеется незначимая переменная, то ее можно определить по низкому значению

а) *t*-статистики;

б) коэффициента детерминации;

в) *F*-статистики.

4. Для определения доли вариации, обусловленной изменением величины изучаемого фактора, используется

а) коэффициент вариации;

б) коэффициент корреляции;

в) коэффициент детерминации;

г) коэффициент эластичности.

5. Параметр регрессии показывает

а) на сколько % увеличится или уменьшится в среднем y при увеличении x на 1 %;

б) часть дисперсии одной случайной величины, обусловленную вариацией другой случайной величины;

в) на сколько единиц своего измерения увеличится или уменьшится в среднем y при увеличении x на единицу своего измерения.

6. Коэффициент корреляции считается значимым с вероятностью $(1-\alpha)$, если

а) $|t_{набл}| < t_{кр}$;

б) $|t_{набл}| > t_{кр}$;

в) не имеет значения.

7. Суть метода наименьших квадратов заключается в том, что

а) оценка определяется из условия минимизации суммы квадратов отклонений выборочных данных от определяемой оценки;

б) оценка определяется из условия минимизации суммы отклонений выборочных данных от определяемой оценки;

в) оценка определяется из условия минимизации суммы квадратов отклонений выборочной средней от выборочной дисперсии.

8. Коэффициент корреляции, равный единице, означает, что между переменными наблюдается

а) линейная связь;

б) функциональная связь;

в) параболическая связь;

г) отсутствие связи.

9. По формуле $r = \frac{\overline{xy} - \bar{x} \times \bar{y}}{s_x s_y}$ рассчитывается

а) частный коэффициент корреляции;

б) парный коэффициент корреляции;

в) коэффициент детерминации;

г) множественный коэффициент корреляции.

10. Какой критерий используется для проверки статистической значимости уравнения регрессии:

а) F – критерий Фишера;

б) t – критерий Стьюдента;

в) критерий Дарбина-Уотсона;

г) χ^2.

11. Коэффициент эластичности показывает

а) на сколько процентов изменяется функция с изменением аргумента на одну единицу своего измерения;

б) на сколько процентов изменяется функция с изменением аргумента на 1 %;

в) на сколько единиц своего измерения изменяется функция с изменением аргумента на 1 %.

12. С увеличением объема выборки

а) расширяются интервальные оценки;

б) уменьшается ошибка регрессии;

в) увеличивается точность оценок.

3 Множественный регрессионный анализ

Что необходимо знать из 3 главы:

1. Цель и задачи проведения многофакторного регрессионного анализа.
2. Свойства МНК-оценок.
3. Порядок проведения проверки значимости классической модели множественной регрессии.
4. Построение частных уравнений регрессии и расчет частных коэффициентов корреляции.

3.1 Классическая модель множественной линейной регрессии

До сих пор основное внимание уделялось обсуждению линейной модели первого порядка с одной предикторной переменной. Развитием парного регрессионного анализа применительно к случаям, когда зависимая переменная гипотетически связана с более чем одной объясняющей переменной, является множественный (многофакторный) регрессионный анализ, предполагающий изучение зависимости одного результативного признака от нескольких факторных.

В прошлом статистический анализ более чем одной переменной сводили к рассмотрению каждой переменной в отдельности. Такой подход обладает ограниченными возможностями, поскольку выводы относительно совокупности переменных, как правило, не могут быть получены из выводов относительно каждой переменной в отдельности. Возможность получать такие общие выводы дает многомерный анализ [19, с. 313].

Множественная взаимосвязь на практике часто выражается при помощи линейного уравнения регрессии, которое в экономических исследованиях достаточно точно отражает большинство исследуемых связей.

Вероятностный характер природы наблюдаемых и описываемых с помощью регрессионного анализа объектов требует поиска по возможности наиболее простой теоретической формы представления признаковых связей и статистической оценки надежности как самих моделей, так и модельных параметров. С этой точки зрения особое значение приобретают линейные регрессионные модели, а также исходное предположение о нормальности распределения параметрических оценок. Линейные модели отличаются простой интерпретируемостью и хорошо разработанными приемами оценивания коэффициентов регрессии [20, с. 315-316].

Предположив существование связи между некоторой переменной и рядом других переменных, и применив к соответствующим данным метод наименьших квадратов, можно получить уравнение *множественной регрессии*. От уравнения простой (парной) регрессии линейное уравнение множественной регрессии отличается дополнительными членами, число которых определяется числом объясняющих (факторных) переменных. Особое достоинство метода множественной регрессии – это возможность выделить влияние каждой из объясняющих переменных. Степень этого влияния характеризуется оценками угловых коэффициентов b_j, называемых частными коэффициентами регрессии.

Число факторных признаков теоретически не ограничивается. Однако для практической работы целесообразно ограничиться тремя-восьмью (реже десятью) факторами. Для дальнейшего увеличения числа факторов необходимо значительное увеличение совокупности, по которой обрабатываются данные. В первом приближении можно считать, что число единиц совокупности должно быть по крайней мере в десять раз больше, чем число факторов. Если факторы имеют тесную корреляционную связь между собой, то и десятикратное превышение числа единиц совокупности над числом факторов может оказаться недостаточным [13, с. 56-57].

Рассмотрим классическую линейную модель множественной регрессии.

Пусть выбраны результативный признак Y и независимые переменные $X_1, X_2, ..., X_k$. Требуется оценить, какие факторы значимо влияют на результативный признак. Предположим, что для оценки уравнения регрессии взята выборка объемом

n. Результаты наблюдений над результативным признаком представлены вектором $Y = (y_1,..., y_n)^T$.

Модель множественной регрессии для своего описания и анализа требует использования матричной алгебры. Объясняющие переменные представим в виде:

$$X_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ ... \\ x_{n1} \end{pmatrix} \qquad X_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ ... \\ x_{n2} \end{pmatrix} \quad ..... \quad X_k = \begin{pmatrix} x_{1k} \\ x_{2k} \\ ... \\ x_{nk} \end{pmatrix}. \tag{3.1}$$

То есть наблюденные значения признаков $X_1, X_2,..., X_K$ представляются матрицей X типа «объект-свойство»:

$$X_{n\times k} = \begin{pmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ ... & ... & ... & ... \\ x_{n1} & ... & ... & x_{nk} \end{pmatrix}, \tag{3.2}$$

где x_{ij} – значение j-го признака на i-м объекте наблюдения.

Функция регрессии имеет вид:

$$\tilde{Y} = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k . \tag{3.3}$$

Поскольку $\tilde{Y}$ неизвестно, то переходим к модели вида:

$$y_i = \beta_0 + \beta_1 x_{i_1} + \beta_2 x_{i_2} + ... + \beta_k x_{i_k} + \varepsilon_i , \qquad i = \overline{1,n} . \tag{3.4}$$

Если обозначить

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix},\ \bar{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix},\ X = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix},\ \bar{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_k \end{pmatrix}, \quad (3.5)$$

то модель множественной регрессии в матричном виде:

$$Y = X\bar{\beta} + \varepsilon, \quad (3.6)$$

где Y – матрица размерности $n \times 1$;

X - матрица размерности $n \times k$;

$\bar{\beta}$ - матрица размерности $k \times 1$;

$\bar{\varepsilon}$ - матрица размерности $n \times 1$.

Матричные исчисления во множественном регрессионном анализе достаточно трудоемки. Для построения многофакторных уравнений, оценки значимости уравнения и его параметров, определения доверительных интервалов используют электронные таблицы Excel и специальные статистические пакеты прикладных программ (Statistica, SPSS, Stadia, StatVew, Stata и др.).

Система линейных уравнений (3.4) и (3.6) называется линейной моделью множественной регрессии (ЛММР). В случае, когда $к$=1, речь идет о парной (двумерной) модели регрессии. Линейная модель множественной регрессии, удовлетворяющая условиям Гаусса-Маркова, называется классической ЛММР (КЛММР) (условия 1-2 можно заменить одним в векторной форме $\Sigma_\varepsilon = M\bar{\varepsilon}\bar{\varepsilon}^T = \sigma^2 E_n$).

Как отмечает С.С. Валландер в «Заметках по эконометрике» (С.С Валландер, 2001, с. 11) спецификация (3.4) подразумевает некоторую теоретическую концепцию – мы считаем, что существуют истинные значения коэффициентов $\beta_{0,true}, \beta_{1,true}, ..., \beta_{k,true}$, но они неизвестны и могут обсуждаться лишь умозрительно.

Для решения уравнения (3.6) относительно вектора оценок параметров b нужно ввести еще одну предпосылку для множественного регрессионного анализа:

6. Векторы значений объясняющих переменных, или столбцы матрицы X, должны быть *линейно независимыми*, т.е. ранг матрицы X – максимальный ($r(X)=k+1$). Так, например, при наличии двух объясняющих линейно независимых переменных наблюдения расположатся в трехмерном пространстве, для них можно будет подобрать плоскость $\tilde{y} = \beta_0 + \beta_1 x_{1_1} + \beta_2 x_2$, причем сумма квадратов отклонений наблюдений от этой плоскости будет минимальной и для данной их совокупности плоскость будет единственной.

Если спроецировать точки, соответствующие наблюдениям, на плоскость x_1x_2, то они займут на ней некоторую область (рисунок 3.1).

Полагают также, что число имеющихся наблюдений (значений) каждой из объясняющей и зависимой переменных превосходит ранг матрицы X, т.е. $n>r$ или $n> k+1$, ибо в противном случае в принципе невозможно получение сколько-нибудь надежных статистических выводов [16, с. 86].

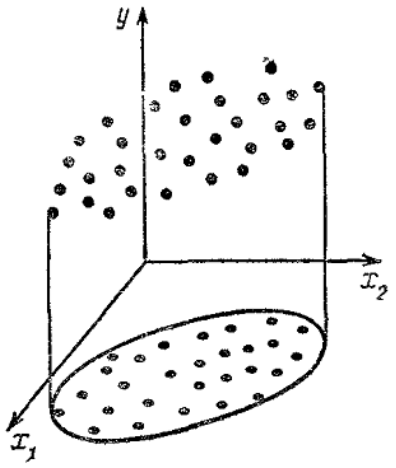


Рисунок 3.1 – Регрессионная зависимость y от x_1 и x_2

В теореме Гаусса-Маркова для множественного регрессионного анализа доказывается, что, как и для парной регрессии, метод наименьших квадратов дает наиболее эффективные линейные оценки. Это значит, что при выполнении условий Гаусса-Маркова на основе той же самой выборочной совокупности нельзя найти другие несмещенные оценки, дисперсии которых будут меньшими. Коэффициенты рег-

рессии являются более точными, чем больше число наблюдений в исследуемой выборке, чем больше дисперсия выборки независимых переменных, чем меньше теоретическая дисперсия стохастического члена и чем меньше связаны между собой независимые переменные (последнее условие, о котором мы говорили выше – для случая многофакторной регрессии).

Оценку коэффициентов $\bar{\beta}$ уравнения регрессии можно искать, исходя из требований минимума модуля отклонения наблюденных значений y_i от "значений" функции регрессии, либо (обычно) из критерия минимума суммы квадратов отклонений наблюденных значений y_i от "значений" функции регрессии (МНК), более удобного с позиций технической реализации.

Перейдем к оценке коэффициентов методом наименьших квадратов. Выпишем квадратичный функционал, обозначив через $\bar{b} = (b_0, b_1, ..., b_k)^T$ оценку вектора $\bar{\beta}$

$$F = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2} - ... - b_k x_{in})^2 = (Y - X\bar{b})^T (Y - X\bar{b}) =$$
$$= Y^T Y - \bar{b}^T X^T Y - Y^T X\bar{b} + \bar{b}^T X^T X\bar{b} = Y^T Y - 2\bar{b}^T X^T Y + \bar{b}^T X^T X\bar{b} \to \min \quad . \qquad (3.7)$$

Воспользовавшись необходимым условием существования экстремума, найдем

$$2X^T X\bar{b} - 2X^T Y = 0 . \qquad (3.8)$$

Тогда система нормальных уравнений в матричной форме для определения вектора $\bar{b}$ имеет вид

$$X^T X\bar{b} = X^T Y . \qquad (3.9)$$

В силу предположения о справедливости условий Гаусса-Маркова, в частности ($X=k+1$), матрица $X^T X$ – не вырождена и из (3.9) получим МНК - оценки для вектора $\bar{\beta}$:

$$\overline{b}_{\text{МНК}} \equiv \overline{b} = (X^T X)^{-1} X^T Y . \qquad (3.10)$$

Тогда оценка $\tilde{y}$ уравнения регрессии имеет вид:

$$\tilde{y} = b_0 + b_1 x_1 + b_2 x_2 + ... + b_k x_k . \qquad (3.11)$$

Коэффициенты уравнения показывают, на сколько натуральных единиц изменится результативный признак при изменении соответствующего фактора на одну единицу своего измерения.

На практике исследователю необходимо выявить влияние объясняющих переменных, которые, как правило, выражены в различных единицах измерения. Если изменить единицу измерения одной или нескольких переменных, то соответствующим образом изменятся коэффициенты регрессии, т.е. в общем случае коэффициенты регрессии между собой несопоставимы.

Такое сопоставление возможно лишь при одних и тех же единицах измерения одноименных коэффициентов в двух уравнениях регрессии. Для сравнительной оценки влияния факторов на результативный признак коэффициенты регрессии следует представить в стандартизированных единицах. Для выражения коэффициентов регрессии в стандартизированном масштабе необходима стандартизация всех переменных, т.е. как результативных, так и факторных признаков. С этой целью все переменные выражают в стандартных отклонениях от соответствующих средних арифметических [13, с. 65-66]:

$$b_j' = b_j \frac{s_{x_j}}{s_y} . \qquad (3.12)$$

Стандартизированный коэффициент регрессии показывает, на сколько величин s_y в среднем изменится результативная переменная *Y* при увеличении только *j*-й факторной переменной на одну s_{x_j}. Сравнивая стандартизированные коэффициенты

регрессии, можно ранжировать объясняющие переменные по силе их воздействия на результат. Это основное достоинство данных коэффициентов в отличие от несравнимых между собой коэффициентов «чистой регрессии».

Кроме того, для сравнительной оценки влияния факторов на результативный признак используют стандартизированные коэффициенты эластичности $E_j \left(j = \overline{1,k}\right)$:

$$E_j = b_j \frac{\overline{x}_j}{\overline{y}}. \tag{3.13}$$

Коэффициент эластичности отражает, на сколько процентов (от средней) изменится в среднем *Y* при увеличении только X_j на 1 %.

3.2 Оценка значимости КЛММР

В качестве характеристики степени рассеяния случайной величины *Y* относительно функции регрессии в случае нелинейной связи используется корреляционное отношение

$$\rho^2_{y/x_1,\ldots,x_n} = 1 - \frac{M(y - f_y(\overline{X}))^2}{\sigma_y^2} = \frac{M(f_y(\overline{X})M)^2}{\sigma_y^2}, \tag{3.14}$$

которое характеризует качество подгонки функции регрессии под выборочные данные. В случае линейной регрессии $\rho_y(\overline{X})$, называется коэффициентом детерминации $R^2_{y/x_1,\ldots,x_n} \equiv R^2$.

Как мы уже упоминали в предыдущих параграфах, коэффициент детерминации получается как отношение факторной дисперсии (обусловленной варьированием значений объясняющих переменных $\overline{X} = (x_1,..,x_k)^T$) к величине остаточной дисперсии (обусловленной вариацией случайной величины относительно функции регрессии):

$$\tilde{R}^2_{y/x_1,\dots,x_n} = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} = 1 - (\sum_{i=1}^{n}(y_i - b_0 - b_1 x_{i1} - \dots - b_k x_{ik})^2) / (\sum_{i=1}^{n}(y_i - \bar{y})^2), \qquad (3.15)$$

где $SSR = \sum_{i=1}^{n}(y_i - \tilde{y}_i)^2 \equiv \sum_{i=1}^{n} e_i^2$, $e_i = y_i - \tilde{y}_i$;

$SSE = \sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2$;

$SST = SSR + SSE$.

Выборочный коэффициент детерминации характеризует долю общей вариации результативного признака y, объясняемую вариацией выборочной функции регрессии $\tilde{f}(\bar{X})$.

Формула скорректированной на несмещенность оценки $\tilde{R}^{*2}_{y/x_1,\dots,x_n}$ коэффициента детерминации $R^2_{y/x_1,\dots,x_n}$ имеет вид

$$\tilde{R}^{*2}_{y/x_1,\dots,x_n} \approx 1 - (1 - \tilde{R}^2_{y/x_1,\dots,x_n}) \frac{n-1}{n-k-1}. \qquad (3.16)$$

Коэффициент множественной детерминации измеряет долю всей вариации результативного признака, которая объясняется за счет вариации всего комплекса факторов, входящих в уравнение регрессии. Но понятно, что желательно знать меру влияния каждого из факторов, на вариацию результативного признака. При этом нужно учесть, что влияние всего комплекса факторов нельзя считать равным простой сумме влияний каждого фактора в отдельности. Факторы, как правило, это *система* взаимосвязанных переменных, связь их в том состоит, что один фактор может либо усиливать влияние другого, либо, наоборот, препятствовать влиянию другого (других) факторов. Например, если в лучших предприятиях выше энерговооруженность работников и одновременно, выше их квалификация, то один фактор увеличивает и влияние другого. Если же в районе, где выше плодородие почв, но в

то лето выпало меньше осадков, а где ниже плодородие, там выпало больше осадков, то один фактор мешает влиянию другого, налицо «антисистема».

Следовательно, используя «системный подход», теорию систем, как требует современная наука, следует так разложить R^2 на доли факторов, чтобы выявить и измерить отдельно «системный эффект» факторов, как системы, а не простой суммы. Прежде всего, для этой цели необходимо знать, как измерить изолированное влияние отдельного фактора на вариацию результативного признака. Ни парный коэффициент детерминации, ни, тем более, частный коэффициент детерминации этой задачи не решают. В парном коэффициенте детерминации (и корреляции) включено влияние других факторов, если они варьируют параллельно с данным, если между ними есть связь, а это всегда так. Частные коэффициенты детерминации – доли не от всей вариации результативного признака, а той ее части, которая оставалась не объясненной вариацией прочих факторов, то есть это доли от разных величин.

Иногда предлагают считать мерой влияния изолированного фактора произведение его парного коэффициента корреляции на его стандартизованный коэффициент регрессии $r_{x_j} y \cdot \beta_j$. Сумма этих произведений по всем факторам равна R^2, но где же «системный эффект»? Он по частям разбросан по отдельным факторам, преувеличивая роль каждого из них, т.к. входящий в меру парный коэффициент корреляции не свободен от влияния других факторов.

Докажем теорему о том, что чистой мерой влияния вариации изолированного фактора на вариацию результативного признака является квадрат стандартизованного коэффициента регрессии, то есть β_j^2.

Для этого предложим, что из всех факторов, входящих в уравнение регрессии $\tilde{y} = b_0 + \sum_{j=1}^{K} b_j x_j$ варьирует только один, например, x_1:

$$\tilde{y}_{(x_1)i} = b_0 + b_1 x_{1i} + \sum_{2}^{K} b_j \bar{x}_j , \qquad (3.17)$$

а все факторы, кроме x_1, закреплены на среднем уровне, то есть не варьируют по i-тым единицам совокупности.

При этом индивидуальное значение x_{1i} можно выразить через $\bar{x}_1$ и индивидуальное отклонение от среднего для каждой i-той единицы совокупности: $x_{1i} = \bar{x}_1 + \Delta x_{1i}$.

Подставив это выражение в (3.16), имеем:

$$\tilde{y}_{(x_1)_i} = b_0 + b_1(\bar{x}_1 + \Delta x_{1i}) + \sum_2^K b_j \bar{x}_j = b_0 + \sum_{j=1}^K b_j \bar{x}_j + b_1 \Delta x_{1i} = \bar{y} + b_1 \Delta x_{1i} \qquad (3.17)$$

так как $b_0 + \sum_{j=1}^K b_j \bar{x}_j$ равно $\bar{y}$. Сумма квадратов отклонений значений $\tilde{y}_{(x_1)_i}$, то есть варьирующих только за счет вариации x_{1i}, имеет вид: $\sum_{i=1}^n (\tilde{y}_{(x_1)_i} - \bar{y})^2$.

Подставив в это выражение $\tilde{y}_{(x_1)_i}$ из (3.17) получим:

$$\sum_{i=1}^n (\bar{y} + b_1 \Delta x_{1i} - \bar{y})^2 = \sum_{i=1}^n (b_1 \Delta x_{1i})^2 = b_1^2 \sum_{i=1}^n \Delta^2 x_{1i} = \underline{b_1^2 n \sigma_{x_1}^2}, \qquad (3.18)$$

т.е. сумма квадратов отклонений $\tilde{y}_{(x_1)_i}$ от $\bar{x}_1$ (сумма $\Delta x_{1\,i}^2$) есть $n\sigma_{x_1}^2$. Какова же доля всей вариации результативного признака, объясняемая только за счет вариации одного x_1? Разделим выражение (3.18) на всю сумму квадратов отношений y_i от $\bar{y}$:

$$\frac{b_1^2 n \sigma_{x_1}^2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{b_1^2 n \sigma_{x_1}^2}{n \sigma_y^2} = (b \frac{\sigma_{x_1}}{\sigma_y})^2 = \beta_1^2, \qquad (3.19)$$

что и требовалось доказать. Итак, мерой доли вариации результативного признака, объясняемой только за счет вариации изолированного фактора, является квадрат стандартизованного коэффициента этого фактора, β_j^2.

Если факторы были бы независимы друг от друга, не представляли бы системы, то коэффициент детерминации R^2 был бы равен сумме β_j^2. Когда же факторы

образуют систему, то R^2 может быть больше $\sum_{j=1}^{K} \beta_j^2$, если системный эффект положителен, либо меньше, если он отрицателен.

Мерой системного эффекта факторов, входящих в уравнение регрессии, является величина (обозначим ее η_s^2 - «ню квадрат системный»):

$$\eta_s^2 = R^2_{yx_1 \ldots x_n} - \sum_{j=1}^{K} \beta_j^2 \,. \qquad (3.20)$$

Системный эффект может оказаться и отрицательной величиной, что свидетельствует о противоречивости влияния факторов, например, если два фактора имеют прямую связь с результативным признаком, но обратную связь друг с другом, то есть «мешают» друг другу положительно влиять на результат

Если же отрицательный системный эффект возникает в системе управляемых факторов, то это говорит об ошибке менеджмента, например, если неправильно построенная система оплаты труда поощряет расточительное использование горючего или других материалов, - это мешает снижению себестоимости. Отрицательный системный эффект – сигнал о неблагополучии в производстве, этим он и важен для управления [10].

Обратимся далее к статистическим свойствам МНК - оценок КЛММР. Это несмещенность, состоятельность и эффективность.

1. *Несмещенность*. Поскольку оценки являются случайными переменными, их значения не могут в точности равняться характеристикам генеральной совокупности – будет присутствовать определенная ошибка, которая может быть велика или мала, положительна или отрицательна. Разница между математическим ожиданием оценки и соответствующей теоретической характеристикой генеральной совокупности называется смещением. Исследователю требуется, чтобы математическое ожидание оценки равнялось бы соответствующей характеристике генеральной совокупности, т.е. чтобы оценка была несмещенной.

МНК – оценка $\overline{b}$ является несмещенной оценкой вектора $\overline{\beta}$:

$$\bar{b} = (X^T X)^{-1} X^T (X\bar{\beta} + \bar{\varepsilon}) = (X^T X)^{-1}(X^T X)\bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon} = \bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}\,, \quad (3.21)$$

$$M\bar{b} = M(\bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T M\varepsilon = \bar{\beta}\,. \quad (3.22)$$

2. *Состоятельность*. Если предел оценки по вероятности равен истинному значению генеральной совокупности, то эта оценка называется состоятельной (доказывается исходя из того, что наименьшее собственное число матрицы $X^T X$ при $n \to \infty$ стремится к ∞).

Свойство состоятельности означает, что при увеличении объема наблюдения оценки параметров становятся более надежными в вероятностном смысле, т.е. с ростом *n* оценки концентрируются вокруг истинных неизвестных значений параметров [14, с. 228]. Другими словами, состоятельной называется такая оценка, которая дает точное значение для большой выборки независимо от входящих в нее конкретных наблюдений.

К. Доугерти по этому поводу отмечал следующее. Иногда бывает, что оценка, смещенная на малых выборках, является состоятельной (иногда состоятельной может быть даже оценка, не имеющая на малых выборках конечного математического ожидания). Иногда невозможно найти оценку, несмещенную на малых выборках. Если при этом вы можете найти хотя бы состоятельную оценку, это может быть лучше, чем не иметь никакой оценки, особенно если вы можете предположить направление смещения на малых выборках. Нужно, однако, иметь в виду, что состоятельная оценка в принципе может на малых выборках работать хуже, чем несостоятельная (например, иметь большую среднеквадратическую ошибку), и поэтому требуется осторожность. Подобно тому, как вы можете предпочесть смещенную оценку несмещенной, если ее дисперсия меньше, вы можете предпочесть состоятельную, но смещенную оценку несмещенной или несостоятельную оценку им обоим (также в случае меньшей дисперсии) [6, с. 27].

3. *Эффективность*. Оценка с максимально возможной вероятностью должна давать близкое значение к теоретической характеристике (получить функцию плотности вероятности как можно более «сжатую» вокруг истинного значения), т.е. эф-

фективная оценка – это несмещенная оценка, обладающая наименьшей дисперсией по сравнению с любыми другими линейными и несмещенными оценками параметра. Выборочное среднее имеет наименьшую дисперсию – это наиболее эффективная оценка среди всех несмещенных оценок.

Следует отметить, что эффективность оценок можно сравнивать только в том случае, когда они используют один и тот же набор переменных. Если одна из оценок использует объем информации в несколько раз превышающий информацию, используемую другой оценкой, то она вполне может иметь дисперсию, меньшую по величине, но считать такую оценку более эффективной неправильно.

Для проверки значимости построенного уравнения регрессии выдвигается гипотеза H_0: линейная модель множественной регрессии не адекватна выборочным данным, что формально можно сформулировать как равенство параметров модели нулю H_0: $\beta_1=\beta_2=\ldots=\beta_к=0$. Альтернативная гипотеза H_1: ЛММР адекватна выборочным данным или H_1: $\exists j \in [\overline{1,n}]: \quad \beta_j \neq 0$.

Для проверки гипотезы H_0 используем F- статистику:

$$F = \frac{SSE/k}{SSR/(n-k-1)} = \frac{\tilde{R}^2_{y/x_1,\ldots,x_n}/k}{(1-\tilde{R}^2_{y/x_1,\ldots,x_n})/(n-k-1)}, \qquad (3.23)$$

которая в случае справедливости H_0 имеет распределение Фишера – Снедекора с числом степеней свободы $\nu_1 = k \; u \; \nu_2 = n-k-1$. Затем проверяем гипотезу по стандартной схеме – либо сравнивая $F_{набл}$ и $F_{кр}$, либо сравнивая значимость нулевой гипотезы с заданным уровнем 0,05.

Надежность уравнения регрессии можно также оценить с помощью коэффициента аппроксимации (средней относительной величины модельной ошибки):

$$A = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \tilde{y}_i|}{\tilde{y}} \times 100. \qquad (3.24)$$

Пороговые значения коэффициента аппроксимации отражены на рисунке 3.2. Модель регрессии с ошибкой аппроксимации менее 10 % считается надежной.

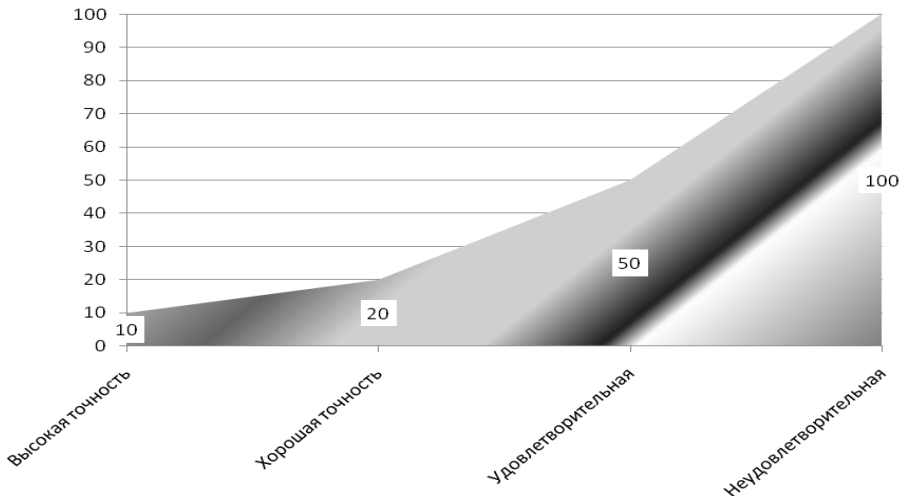


Рисунок 3.2 – Пороговые значения коэффициента аппроксимации, процентов

Если нулевая гипотеза о незначимости уравнения регрессии отвергнута, проверяем гипотезы о значимости параметров уравнения регрессии.

Как и в случае парной регрессии, основные предположения заключаются в том, что отдельные значения ошибок:

1) являются независимыми от величины всех X_i и от других значений ошибок по всей выборке;

2) характеризуются нормальным распределением с нулевым математическим ожиданием и постоянной конечной дисперсией [15, с. 69].

Для коэффициентов множественной регрессии *t*-тесты выполняются так же, как и в парном регрессионном анализе. Критический уровень *t* при любом уровне значимости зависит от числа степеней свободы, которое равно *(n-k-1)*, т.е. число наблюдений минус число оцениваемых параметров. Доверительные интервалы также рассчитываются аналогично парному регрессионному анализу с учетом указанных степеней свободы.

Схема проверки значимости параметров уравнения регрессии следующая:

Выдвигаем гипотезы вида:

- H_0: коэффициент β_j незначимо отличен от нуля (или формально $\beta_j=0$);

- альтернативная гипотеза H_1: параметр β_j – значимо отличен от нуля ($\beta_j \neq 0$).

Для проверки гипотез строятся статистики

$$t = \frac{|b_j|}{s_{b_j}}, \quad j = 1,2,...,k, \tag{3.25}$$

$$s_{b_j} = s\sqrt{\left[\left(X^T X\right)^{-1}\right]_{jj}}, \tag{3.26}$$

которые в случае справедливости H_0, имеют распределение Стьюдента с $\nu = n - k - 1$ степенями свободы. Затем либо сравниваем $t_{набл}$ с $t_{кр}(\alpha)$, либо значимость нулевой гипотезы с заданным уровнем.

Для коэффициентов уравнения регрессии, значимо отличных от нуля, находят доверительные интервалы, используя t-статистику

$$|t| = \frac{|b_j - \beta_j|}{s_{b_j}} > t_{1-\alpha;n-k-1}. \tag{3.27}$$

Доверительный интервал для параметра b_j:

$$b_j - t_{1-\alpha;n-k-1} s_{b_j} \leq \beta_j \leq b_j + t_{1-\alpha;n-k-1} s_{b_j}. \tag{3.28}$$

Доверительный интервал для функции регрессии (для условного математического ожидания зависимой переменной $M_x(Y)$) есть

$$\tilde{y} - t_{1-\alpha;n-k-1} s_{\tilde{y}} \leq M(Y) \leq \tilde{y} + t_{1-\alpha;n-k-1} s_{\tilde{y}}, \tag{3.29}$$

где $\tilde{y}$ - групповая средняя, определяемая по уравнению регрессии;

$s_{\tilde{y}} = s\sqrt{X_0^T \left(X^T X\right)^{-1} X_0}$ - стандартная ошибка групповой средней.

При построении интервала предсказания для индивидуальных значений зависимой переменной используем *t*-статистику и стандартную ошибку индивидуального значения, рассчитываемую по формуле

$$s_{\tilde{y}_0} = s\sqrt{1 + X_0^T \left(X^T X\right)^{-1} X_0}\ , \tag{3.30}$$

где X_0 - точка, в которой мы хотим построить доверительный интервал.

Тогда:

$$\tilde{y}_0 - t_{1-\alpha;n-k-1} s_{\tilde{y}_0} \le y_0^* \le \tilde{y}_0 + t_{1-\alpha;n-k-1} s_{\tilde{y}_0}\ . \tag{3.31}$$

Прогнозирование на основе регрессионной модели исходит из предположения (гипотезы), что факторы управляемы и могут принять то или иное плановое, ожидаемое значение, а прочие неизвестные условия сохранятся на среднем по совокупности уровне. Управляемость факторов не означает, что при прогнозе в модель можно подставлять любые их значения. Уравнение регрессии отражает те условия, которые существовали в совокупности, по данным которой уравнение получено. Если бы значения факторных признаков были в 2–3 раза и более высокими, то нельзя ручаться, что коэффициенты условно-чистой регрессии остались бы теми же. Более вероятно, что есть статистическая связь между величиной факторов и значениями коэффициентов; связь близкая и линейная на ограниченном пространстве вариации факторов вполне может оказаться нелинейной на значительно большем пространстве вариации факторов. Поэтому рекомендуется при прогнозировании по уравнению регрессии *не выходить за пределы реально наблюдаемых значений факторов* в совокупности или выходить за эти границы не более чем на 10–15 % средних величин [10, с. 92].

3.3 Частная регрессия и корреляция

В отличие от парной регрессии в ходе изучения множественной регрессии вместо одного фактора выступает группа объясняющих переменных, влияющих на результативный признак. В ситуации, когда исследователь заинтересован в получении ответа на вопрос, какая связь существует между интересующим его фактором и результатом при условии, что остальные факторы остаются неизменными, нет возможности решить поставленную таким образом задачу полностью, поскольку факторов, влияющих на результативный признак, очень много. Однако частное решение, при котором на неизменном уровне закрепляется несколько важных объясняющих переменных, возможно. Такого рода задачи решаются методом частной регрессии и корреляции.

Если переменные коррелируют друг с другом, то на величине парного коэффициента корреляции частично сказывается влияние других переменных. Если, например, между x_1 и x_2 существует тесная связь, и, кроме того, y зависит от x_1, то y будет также коррелировать с x_2. Вполне возможно, что корреляция между y и x_2 не прямая, а косвенная, возникающая вследствие воздействия x_1. Поэтому необходимо исследовать частную корреляцию между y и x_2 при исключении влияния x_1 на y. Исключаемые переменные могут закрепляться как на средних, так и на других уровнях, выбранных в соответствии с интересующими нас участками изменения переменных, между которыми определяется связь в «чистой» форме. Здесь следует учитывать профессионально-теоретические соображения об изучаемом явлении [21, с. 132]. Коэффициент частной регрессии совпадает с соответствующим коэффициентом множественной регрессии и имеет такое же экономическое содержание.

Частные уравнения регрессии имеют следующий вид:

$$\begin{cases} y_{x_1/x_2x_3,\ldots,x_k} = b_0 + b_1x_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_k\bar{x}_k + \varepsilon; \\ y_{x_2/x_1x_3,\ldots,xk} = b_0 + b_1\bar{x}_1 + b_2x_2 + b_3\bar{x}_3 + \ldots + b_k\bar{x}_k + \varepsilon; \\ \ldots\ldots\ldots\ldots\ldots\ldots, \\ y_{x_k/x_1x_2,\ldots,x_{k-1}} = b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_kx_k + \varepsilon. \end{cases} \quad (3.32)$$

При подстановке в уравнения (3.32) средних значений соответствующих факторов они принимают вид парных уравнений линейной регрессии:

$$\begin{cases} y_{x_1/x_2x_3,\ldots,x_k} = A_1 + b_1 x_1 + \varepsilon; \\ y_{x_2/x_1x_3,\ldots,x_k} = A_2 + b_2 x_2 + \varepsilon; \\ \ldots\ldots\ldots\ldots\ldots\ldots, \\ y_{x_k/x_1x_2,\ldots,x_{k-1}} = A_k + b_k x_k + \varepsilon. \end{cases}, \qquad (3.33)$$

где $\begin{cases} A_1 = b_0 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_k\bar{x}_k, \\ A_2 = b_0 + b_1\bar{x}_1 + b_3\bar{x}_3 + \ldots + b_k\bar{x}_k, \\ \ldots\ldots\ldots\ldots\ldots\ldots, \\ A_k = b_0 + b_1\bar{x}_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_{k-1}\bar{x}_{k-1} \end{cases}$.

На основе частных уравнений регрессии можно рассчитать частные коэффициенты эластичности:

$$E_{y_{x_i}} = b_i \times \frac{x_i}{\tilde{y}_{x_i/x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k}}, \qquad (3.34)$$

где b_i - коэффициенты регрессии для фактора *x*, в уравнении множественной регрессии;

$\tilde{y}_{x_i \cdot x_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k}$ - частное уравнение регрессии.

Ранжирование факторов, участвующих во множественной линейной регрессии, может быть проведено с помощью частных коэффициентов корреляции для линейных связей. При нелинейной взаимосвязи исследуемых признаков эту функцию выполняют частные индексы детерминации. Кроме того, частные показатели корреляции широко используются при отборе факторов: целесообразность включения того или иного фактора в модель доказывается величиной показателя частной корреляции. Частные коэффициенты (или индексы) корреляции характеризуют тесноту связи между результатом и соответствующим фактором при неизменном уровне других факторов, использованных в уравнении регрессии.

Базой для вывода формул коэффициентов частной корреляции служит коэффициент простой корреляции. Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель. Если рассматривается регрессия с числом факторов k, то возможны частные коэффициенты корреляции не только первого, но и второго, третьего,..., (k - 1) порядка, т. е. влияние фактора x_1 можно оценить при разных условиях независимости действия других факторов:

- $r_{yx_1 \cdot x_2}$ - при постоянном действии фактора x_2;

- $r_{yx_1 \cdot x_2 x_3}$ - при постоянном действии факторов x_2 и x_3;

- $r_{yx_1 \cdot x_2 \ldots x_p}$ - при неизменном действии факторов, включенных в уравнении регрессии.

Сопоставление коэффициентов частной корреляции разных порядков по мере увеличения числа включаемых факторов показывает процесс «очищения» связи результативного признака с исследуемым фактором.

В общем виде при наличии k факторов для уравнения

$$\tilde{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + \ldots + b_k \cdot x_k \qquad (3.35)$$

коэффициент частной корреляции, измеряющий влияние на y фактора x при неизменном уровне других факторов, можно определить по формуле:

$$r_{yx_i / x_1 x_2 \ldots x_{i-1} x_{i+1} \ldots x_k} = \sqrt{1 - \frac{1 - R^2_{yx_1 / x_2 \ldots x_i \ldots x_k}}{1 - R^2_{yx_1 / x_2 \ldots x_{i-1} x_{i+1} \ldots x_k}}}, \qquad (3.36)$$

где $R^2_{yx_1 / x_2 \ldots x_i \ldots x_k}$ - множественный коэффициент детерминации всего комплекса k факторов с результатом;

$R^2_{yx_1 / x_2 \ldots x_{i-1} x_{i+1} \ldots x_k}$ - тот же показатель детерминации, но без введения в модель фактора x_i.

В случае, когда $i = 1$, формула коэффициента частной корреляции примет вид:

$$r_{yx_i / x_2 \ldots x_k} = \sqrt{1 - \frac{1 - R^2_{yx_1 / x_2 \ldots x_k}}{1 - R^2_{yx_2 \ldots x_k}}}\,. \qquad (3.37)$$

Коэффициент частной корреляции (3.37) позволяет измерить тесноту связи между y и x_1 при неизменном уровне всех других факторов, включенных в уравнение регрессии.

Порядок частного коэффициента корреляции определяется количеством факторов, влияние которых исключается. Например, $r_{yx_1 \cdot x_2}$ — коэффициент частной корреляции первого порядка. Соответственно коэффициенты парной корреляции называются коэффициентами нулевого порядка. Каждый коэффициент частной корреляции может быть вычислен на основе коэффициентов ближайшего низшего порядка:

$$r_{yx_i / x_1 x_2 \ldots x_k} = \frac{r_{yx_i / x_1 x_2 \ldots x_{p-1}} - r_{yx_p / x_1 x_2 \ldots x_{k-1}} \times r_{x_i x_p / x_1 x_2 \ldots x_{k-1}}}{\sqrt{\left(1 - r^2_{yx_p / x_1 x_2 \ldots x_{k-1}}\right) \times \left(1 - r^2_{x_i x_p / x_1 x_2 \ldots x_{k-1}}\right)}}\,. \qquad (3.38)$$

При двух факторах и $i = 1$ (коэффициент первого порядка) формула (3.38) примет вид:

$$r_{yx_1 / x_2} = \frac{r_{yx_1} - r_{yx_2} \times r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_2}\right) \times \left(1 - r^2_{x_1 x_2}\right)}}. \qquad (3.39)$$

Соответственно при $i = 2$ и двух факторах частный коэффициент корреляции у с фактором x_2 можно определить по формуле

$$r_{yx_2 / x_1} = \frac{r_{yx_2} - r_{yx_1} \times r_{x_1 x_2}}{\sqrt{\left(1 - r^2_{yx_1}\right) \times \left(1 - r^2_{x_1 x_2}\right)}}. \qquad (3.40)$$

Для уравнения регрессии с тремя факторами частные коэффициенты корреляции второго порядка определяются на основе частных коэффициентов корреляции первого порядка. По уравнению можно исчислить три частных коэффициента корреляции второго порядка: $r_{yx_1/x_2x_3}; r_{yx_2/x_1x_3}; r_{yx_3/x_1x_2}$, каждый из которых определяется по рекуррентной формуле. Так, при *i= 1* получим формулу для расчета $r_{yx_1 \cdot x_2x_3}$:

$$r_{yx_1/x_2x_3} = \frac{r_{yx_1/x_2} - r_{yx_3/x_2} \times r_{x_1x_3/x_2}}{\sqrt{\left(1 - r^2_{yx_3/x_2}\right) \times \left(1 - r^2_{x_1x_3/x_2}\right)}}. \qquad (3.41)$$

Рассчитанные по рекуррентной формуле частные коэффициенты корреляции изменяются в пределах от минус 1 до плюс 1, а по формулам через множественные коэффициенты детерминации от 0 до 1.

Приведенные формулы частных коэффициентов корреляции удобны для расчетов, если уравнение множественной регрессии включает 2-3 фактора. В случае, когда объясняющих переменных больше, коэффициенты частной корреляции удобнее вычислять из элементов обратных матриц коэффициентов системы нормальных уравнений.

Из формул частных коэффициентов корреляции видна связь этих показателей с множественным коэффициентом корреляции. Зная частные коэффициенты корреляции (последовательно первого, второго и более высокого порядка), можно определить множественный коэффициент корреляции по формуле

$$R_{yx_1/x_2 \ldots x_k} = \left(1 - \left(1 - r^2_{yx_1}\right) \times \left(1 - r^2_{yx_2/x_1}\right) \times \left(1 - r^2_{yx_3/x_1x_2}\right) \times \ldots \times \left(1 - r^2_{yx_p/x_1x_2 \ldots x_{k-1}}\right)\right)^{1/2}. \qquad (3.42)$$

Сумма частных коэффициентов детерминации равна квадрату множественного коэффициента детерминации (доказательство см., например, [11, с. 649-659]).

3.4 Вопросы для самоконтроля

1. Сформулируйте цели и порядок проведения многофакторного регрессионного анализа.

2. Каким образом проводится оценивание КЛММР?

3. Как проводится построение доверительных интервалов регрессии и параметров уравнения?

4. Опишите схему разложения коэффициента множественной детерминации по отдельным факторам и измерения их системного эффекта.

5. Каковы статистические свойства МНК - оценок?

6. С какой целью строятся частные уравнения регрессии, что они характеризуют?

7. Как рассчитываются частные коэффициенты корреляции?

3.5 Тесты

1. Для проверки значимости параметров уравнения множественной регрессии используется распределение

а) нормальное;

б) Стьюдента;

в) биномиальное;

г) Фишера-Снедекора.

2. Какие требования при построении модели регрессии предъявляются к математическому ожиданию и дисперсии случайных отклонений:

а) $M\varepsilon_i = 1; D\varepsilon_i = 0$;

б) $M\varepsilon_i = 0; D\varepsilon_i = 1$;

в) $M\varepsilon_i = 1; D\varepsilon_i = \sigma^2$;

г) $M\varepsilon_i = 0; D\varepsilon_i = \sigma^2$.

3. При добавлении в уравнение регрессии еще одного объясняющего фактора множественный коэффициент корреляции:

а) уменьшится;

б) возрастет;

в) не изменится.

4. Значимость парных и частных коэффициентов корреляции проверяется с помощью

а) F – критерия Фишера;

б) t – критерия Стьюдента;

в) нормального закона распределения.

5. Известно, что при фиксированном значении X_2 между величинами X_1 и X_3 существует положительная связь. Какое значение может принять частный коэффициент корреляции $\rho_{13/2}$:

а) 1,2;

б) -0,33;

в) 0;

г) 0,5.

6. Коэффициент детерминации – это:

а) квадрат частного коэффициента корреляции

б); квадрат парного коэффициента корреляции;

в) квадрат множественного коэффициента корреляции.

7. В хорошо подобранной модели остатки должны

а) не коррелировать друг с другом;

б) иметь логнормальное распределение;

в) иметь нормальный закон распределения с нулевым математическим ожиданием и постоянной дисперсией;

г) иметь экспоненциальный закон распределения;

д) форма и вид распределения не важны.

8. В каких пределах меняется коэффициент детерминации?

а) от 0 до минус 1;

б) от минус ∞ до $+\infty$;

в) от 0 до + 1;

г) от минус 1 до +1.

9. Каковы последствия нарушения допущения МНК «математическое ожидание регрессионных остатков равно нулю»?

а) смещенные оценки коэффициентов регрессии;

б) эффективные, но несостоятельные оценки коэффициентов регрессии;

в) неэффективные оценки коэффициентов регрессии;

г) несостоятельные оценки коэффициентов регрессии.

10. Во множественном линейном уравнении регрессии строятся доверительные интервалы для коэффициентов регрессии с помощью распределения

а) нормального;

б) Стьюдента;

в) Пуассона;

г) Фишера-Снедекора.

11. Матрица R парных коэффициентов корреляции является

а) симметричной;

б) положительно определенной;

в) обратной;

б) транспонированной.

4 Нарушение допущений классической линейной модели

Что необходимо знать из 4 главы:

1. Последствия нарушения предпосылок МНК.
2. Понятие, способы обнаружения и смягчения мультиколлинеарности.
3. Гетероскедастичность пространственной выборки.
4. Автокорреляция регрессионных остатков.
5. Отбор переменных в спецификации модели.

4.1 Мультиколлинеарность

Сам термин *коллинеарность*, означающий линейное соответствие, линейную зависимость, определяет суть проблемы *мультиколлинеарности*. Данное явление в генеральной или выборочной совокупностях возникает тогда, когда различные объясняющие переменные связаны линейной зависимостью. Сразу стоит отметить, что мультиколлениарность может быть проблемой лишь в случае множественной регрессии.

Если выражать точно, то первая часть термина *мульти* предполагает, что коллинеарны более чем две переменные. Именно такая возможность порождает серьезные трудности при поисках удовлетворительного статистического критерия для проверки данных на мультиколлинеарность. Тем не менее общепринятое употребление этого термина предполагает, что он охватывает и коллинеарность пары «объясняющих» переменных как частный случай [15, с. 85].

Мультиколлинеарность и ее последствия в различных аспектах рассматриваются как в специальной литературе (см., например [17, с. 24-27]), так и в прикладных эконометрических исследованиях.

Понятие мультиколлениарности используется для описания проблемы, когда нестрогая линейная зависимость между объясняющими переменными приводит к

получению ненадежных оценок регрессии. Следует отметить, что такая зависимость необязательно дает неудовлетворительные оценки. Так, если все другие предпосылки выполняются (число наблюдений выборочной совокупности и выборочные дисперсии факторных переменных велики, а дисперсия случайных отклонений мала), то в итоге можно получить вполне хорошие оценки. Мультиколлинеарность должна вызываться сочетанием нестрогой зависимости и одного (или более) неблагоприятного условия.

В случае если корреляционная зависимость между объясняющими переменными не очень тесная, она влечет за собой увеличение дисперсии оценок параметров регрессионной модели, а оценки значимости параметров будут смещены (выводы о незначительном отличии этих оценок от нуля будут ложными). Чем сильнее мультиколлинеарность, тем более произвольно и ненадежно удается распределить сумму объясненных вариаций по отдельным факторным переменным с помощью МНК. В предельном случае, когда между объясняющими переменными существует функциональная зависимость и изменению одной факторной переменной однозначно соответствует изменение другой факторной переменной или линейной комбинации других объясняющих переменных, невозможно разделить степень влияния каждой из них на результативный признак, а метод МНК-оценок становится непригодным. Приведенные положения наглядно демонстрируются с помощью *диаграммы Венна* (рисунок 4.1).

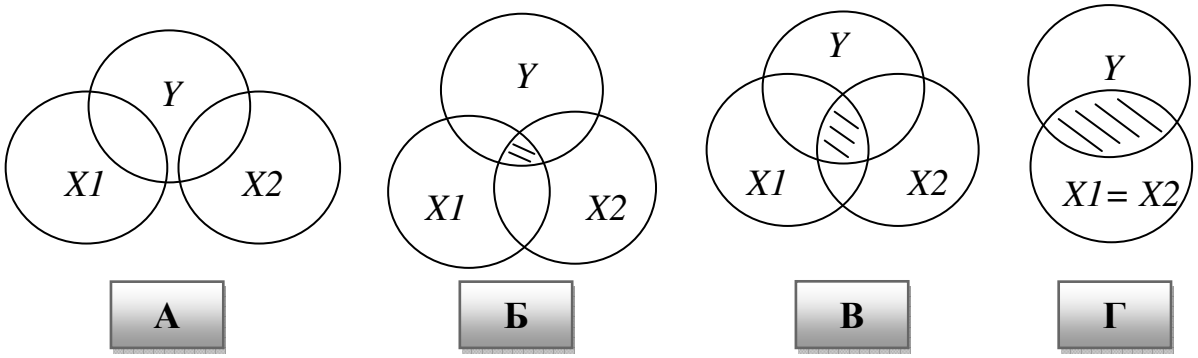


Рисунок 4.1 – Диаграмма Венна

Влияние каждой из объясняющих переменных на *Y* находит отражение в наложении кругов *X1* и *X2* на круг *Y*. Вариант А характеризует отсутствие коррелированности между объясняющими переменными *X1* и *X2*. По мере усиления линейной зависимости между *X1* и *X2* соответствующие круги все больше накладываются друг на друга. Заштрихованная область отражает совпадающие части влияния *X1* и *X2* на *Y*. Вариант Г наглядно показывает, что при совершенной мультиколлинеарности невозможно разграничить степени индивидуального влияния объясняющих переменных *X1* и *X2* на зависимую переменную *Y*.

Матрица $X^T X$ должна быть обратима, что означает неравенство нулю ее определителя ($\left|X^T X\right| \neq 0$). Это условие имеет место только тогда, когда векторы, составляющие матрицу, линейно независимы. В случае если хотя бы два вектора матрицы X строго линейно зависимы, наблюдается строгая (полная) мультиколлинеарность, $\left|X^T X\right| = 0$ и оценки по МНК найти нельзя.

В случае нестрогой (стохастической) мультиколлинеарности, когда $\left|X^T X\right| \approx 0$, формально можно определить вектор оценок параметров $\bar{b}$ (если вычисления $\left|X^T X\right|$ вести с большим числом знаков), но полученные оценки будут ненадежны. Как результат – значительные стандартные ошибки параметров регрессии и оценка их значимости по *t*-критерию Стьюдента бессмысленна (хотя в целом регрессионная модель по *F*-критерию может быть значима).

Основными последствиями явления мультиколлинеарности можно назвать следующие:

1) оценки параметров обнаруживают необычно большие стандартные ошибки, что затрудняет нахождение истинных значений определяемых величин и расширяет интервальные оценки, ухудшая их точность;

2) уменьшаются *t*-статистики коэффициентов, что может привести к неоправданному выводу о степени влияния соответствующей объясняющей на зависимую переменную;

3) МНК-оценки параметров становятся неустойчивыми - крайне чувствительными к малейшим изменениям данных.

4) затрудняется определение вклада каждой из независимых переменных в объясняемую уравнением регрессии дисперсию зависимой переменной;

5) возможно получение неверного знака у параметра регрессии.

Если основная цель построения регрессионной модели - прогноз будущих значений зависимой переменной, то при достаточно большом коэффициенте детерминации R^2 (>0,9) наличие мультиколлинеарности обычно не сказывается на прогнозных качествах модели. Если же целью исследования является определение степени влияния каждой из объясняющих переменных на зависимую переменную, то наличие мультиколлинеарности, приводящее к увеличению стандартных ошибок, скорее всего, исказит истинные зависимости между переменными. В этой ситуации мультиколлинеарность становится серьезной проблемой.

Мультиколлинеарность может возникать в силу разных причин. Например, несколько независимых переменных могут иметь общий временной тренд, относительно которого они совершают малые колебания. В частности, так может случиться, когда значения одной независимой переменной являются лагированными значениями другой [22, с. 111].

Универсальных критериев обнаружения наличия/отсутствия мультиколлинеарности не существует, однако, имеются некоторые эвристические подходы по ее выявлению. Обнаружить наличие мультиколлениарности можно, опираясь на следующие правила:

1. Анализ *матрицы парных коэффициентов корреляции*. Этот общепринятый метод заключается в вычислении матрицы парных коэффициентов корреляции, охватывающей все сочетания переменных. Если в корреляционной матрице между объясняющими переменными наблюдаются значения больше 0,8, то предполагают присутствие мультиколлинеарности.

Рассмотрим корреляционную матрицу (таблица 4.1), составленную для трех переменных: результативного признака y и двух факторных признаков – x_1 и x_2:

Таблица 4.1

1 случай			
	y	x_1	x_2
y	1	0,6	0,7
x_1	0,6	1	0,1
x_2	0,7	0,1	1

Связь обоих факторов с результативным признаком значима, но корреляционная связь между самими факторами невелика, что свидетельствует об отсутствии мультиколлениарности.

Это наилучший вариант для построения множественного уравнения регрессии, при этом факторы x_1 и x_2 в уравнении будут статистически значимы.

Рассмотрим следующий случай (таблица 4.2).

Таблица 4.2

2 случай			
	y	x_1	x_2
y	1	0,6	0,7
x_1	**0,6**	1	0,5
x_2	**0,7**	**0,5**	1

Связь обоих факторов с результативным признаком значима, но также значима и корреляционная связь между самими факторами. Возможно наличие мультиколлениарности.

Третий случай, когда связь между факторами более тесная, чем между результативным признаком и обоими (или только одним) факторами (таблица 4.3).

Данный вариант указывает на наличие мультиколлениарности.

Таблица 4.3

3 случай			
	y	x_1	x_2
y	1	0,6	0,7
x_1	**0,6**	1	0,95
x_2	**0,7**	**0,95**	1

2. Рассчитывают оценки $\overline{R}^2$ для регрессионных зависимостей между каждой из факторных переменных и остальными объясняющими переменными. При этом в случае получения высоких значений (более 0,6) оценки коэффициента детерминации, делают вывод о наличии мультиколлениарности.

Пример 4.1 - В ходе исследования наличия мультиколлениарности были полученные следующие результаты, свидетельствующие о наличии мультиколлинеарности (таблица 4.4):

Таблица 4.4 – Результаты регрессионного анализа, свидетельствующие о наличии мультиколлинеарности

Зависимая переменная	Независимые переменные	Регрессионное уравнение	$\overline{R}^2$
x_1	x_2, x_3	$\tilde{x}_{1_i} = b_0 + b_1 x_{2_i} + b_2 x_{3_i}$	0,47
x_2	x_1, x_3	$\tilde{x}_{2_i} = b_0 + b_1 x_{1_i} + b_2 x_{3_i}$	0,25
x_3	x_1, x_2	$\tilde{x}_{3_i} = b_0 + b_1 x_{1_i} + b_2 x_{2_i}$	**0,71**

3. На существование мультиколлинеарности в модели множественной регрессии указывает низкое значение *t*-критерия для параметров регрессии при высоком значении оценок коэффициента детерминации в тех случаях, когда значение *F*-критерия применительно ко всей совокупности независимых переменных существенно отлично от нуля. В этой ситуации результативная переменная в действительности может быть тесно связана с одной или несколькими объясняющими перемен-

ными, но тесная взаимозависимость между факторными переменными «маскирует» связи некоторых из них с результативной.

4. Если оценки параметров при объясняющих переменных сильно изменяются в зависимости от того, включает или не включает спецификация модели ту или иную факторную переменную, это также наводит на мысль о существовании мультиколлинеарности.

5. Когда оценки параметров регрессии имеют неправильные с точки зрения теории знаки или неоправданно большие значения.

6. Высокие частные коэффициенты корреляции, в случае большего количества малозначимых объясняющих переменных.

7. Расчет показателя А.Е. Хорла [23]. Он основан на использовании для измерения мультиколлинеарности числителя формулы коэффициента множественной детерминации. В предположении множественной регрессии числитель коэффициента множественной детерминации можно представить следующим образом:

$$\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2 = \sum_k b_k^2 \sum_i (x_{ik} - \bar{x}_k)^2 + \sum_{j,k} b_i b_k \sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \qquad (4.1)$$

для $j, k = 1,2,...,m; i = 1,2,...,n \; u \; j \neq k$.

Выражение

$$\sum_i (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k). \qquad (4.2)$$

является числителем коэффициента парной корреляции между переменными x_j и x_k. При отсутствии коллинеарности между этими переменными он равен нулю. Поэтому в качестве общего показателя мультиколлинеарности можно использовать разность M_2:

$$M_2 = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2 - \sum_k b_k^2 \sum_i (x_{ik} - \bar{x}_k)^2 . \qquad (4.3)$$

Если значение M_2 мало, то считаем, что мультиколлинеарность тоже незначительна [21, с. 216].

Как отмечают Р. Винн и К. Холден «все способы проверки обладают одним общим недостатком: ни один из них не проводит четкого недвусмысленного различия между тем, что считать «серьезной» мультиколлинеарностью, и тем, что можно считать обычной и «приемлемой» степенью связи между независимыми переменными при работе с выборочными данными. И все же эти критерии в совокупности дают исследователю, занимающемуся прикладным эконометрическим анализом, а также тем, кто знакомится с результатами его работы, достаточно ясное представление о том, насколько серьезно мультиколлинеарность переменных может повлиять на эти результаты» [17, с. 26].

Рассмотрим далее основные методы, разработанные с целью элиминирования отрицательных последствий мультиколлинеарности (рисунок 4.2).

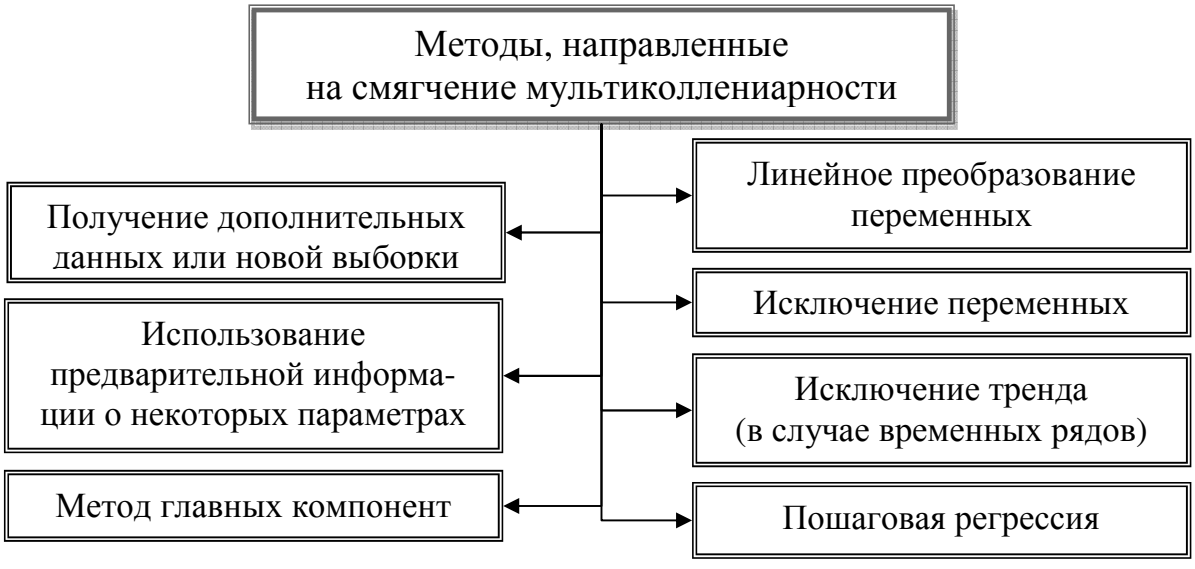


Рисунок 4.2 – Методы, направленные на смягчение мультиколлинеарности

Получение дополнительных данных или новой выборки.

Поскольку мультиколлинеарность напрямую зависит от характера выборки, то мы можем попытаться получить новые данные. Возможно, при наличии другой выборки мультиколлинеарности не будет, либо она не будет столь серьезной. Иногда для уменьшения мультиколлинеарности достаточно увеличить объем выборки. Увеличение количества данных сокращает дисперсии коэффициентов регрессии и тем самым увеличивает их статистическую значимость. Однако получение новой выборки или получение дополнительной статистической информации не всегда возможно или связано с серьезными издержками. Кроме того, увеличивая расходы на получение дополнительной информации, следует помнить, что такой подход имеет уменьшающуюся предельную отдачу – стандартные отклонения коэффициентов регрессии обратно пропорциональны величине $\sqrt{n}$, а расходы будут прямо пропорциональны n.

Использование предварительной информации о некоторых параметрах.

Обычно на основе ранее проведенного регрессионного анализа или в результате экономических исследований уже имеется более или менее точное представление о величине или соотношении двух или нескольких коэффициентов регрессии. Эта предварительная или вневыборочная информация может быть использована исследователем при построении регрессии. В связи с тем, что часть оценок, полученных на основе вневыборочных данных, уже имеет достаточно четкую интерпретацию, это облегчает путь обнаружения взаимных влияний изменений различных переменных. Ограниченность использования данного метода обусловлена тем, что, во-первых, получение предварительной информации зачастую затруднительно, а во-вторых, вероятность того, что выделенный коэффициент регрессии будет одним и тем же для различных моделей, невысока.

Исключение переменных.

Этот метод заключается в том, что высоко коррелированные объясняющие переменные устраняются из регрессии, и она заново оценивается. Отбор переменных, подлежащих исключению, производится с помощью коэффициентов корреляции. Для этого производится оценка значимости коэффициентов парной корреляции $r_{x_1 x_2}$

между объясняющими переменными x_1 и x_2 (в случае двух независимых переменных). Опыт показывает, что если $|r_{x_1x_2}| > 0{,}8$, то одну из переменных можно исключить, но какую именно переменную нужно удалить, решают, исходя экономических соображений. Из-за отсутствия теоретического обоснования этот подход весьма приближенный. Кроме того, в этой ситуации возможны ошибки спецификации. Поэтому в прикладных эконометрических моделях желательно не исключать объясняющие переменные до тех пор, пока коллинеарность не станет серьезной проблемой. В случае, если с экономической точки зрения нельзя отдать предпочтение ни одной из переменных, из рассмотрения удаляют ту, которая менее коррелированна с результативным признаком.

Другой способ исключения переменных был предложен Фарраром и Глубером [24]. Согласно данному подходу, процедура отбора переменных, подлежащих исключению, состоит из трех этапов (при этом предполагается близость распределения остатков к нормальному).

На первом этапе мультиколлинеарность выявляется лишь в общем виде. Для этого строится матрица *R* коэффициентов парной корреляции между объясняющими переменными и вычисляется ее определитель:

$$\Delta = \begin{vmatrix} 1 & r_{x_1x_2} & \dots & r_{x_1x_m} \\ r_{x_2x_1} & 1 & \dots & r_{x_2x_m} \\ \dots & \dots & \dots & \dots \\ r_{x_mx_1} & r_{x_mx_2} & \dots & 1 \end{vmatrix} \qquad (4.4)$$

Далее для проверки наличия мультиколлинеарности вообще среди объясняющих переменных применяется критерий χ^2.

Выдвигается нулевая гипотеза H_0: между объясняющими переменными мультиколлинеарность отсутствует. Альтернативная гипотеза H_1: между объясняющими переменными имеется мультиколлинеарность.

В качестве критерия используется величина:

$$\chi^2 = -\left(n - 1n\,\tfrac{1}{6}(2m+5)\right)\ln\Delta, \qquad (4.5)$$

имеющая χ^2-распределение с $v = 1/2m(m - 1)$ степенями свободы.

Если фактическое значение $\chi^2 \leq \chi^2_{\text{табл}}$, то нулевая гипотеза принимается. Считаем, что мультиколлинеарность между объясняющими переменными отсутствует. Если $\chi^2 > \chi^2_{\text{табл}}$, то гипотеза о наличии мультиколлинеарности не противоречит исходным данным. Между какими переменными она возникает, решается на втором и третьем этапах процедуры.

На втором этапе используются коэффициенты детерминации между объясняющими переменными R^2 (т.е. вначале необходимо оценить уравнение, где конкретная x_j будет выступать в роли зависимой переменной). Оценка мультиколлинеарности основана на том, что величина: $F = \dfrac{R^2(n-m)}{(1-R^2)(m-1)}$ имеет F-распределение с $v_1 = m - 1$ и $v_2 = n-m$ степенями свободы.

Если $F > F_{\text{табл}}$ (α; $v_1 = m - 1$ и $v_2 = n-m$), то переменной x_j в наибольшей степени присуща мультиколлинеарность. По Феррару и Глауберу изучение m-штук значений F-статистик должно показать, какие из объясняющих переменных в большей мере подвержены мультиколлинеарности.

На третьем этапе исследуется, какая объясняющая переменная порождает мультиколлинеарность, и решается вопрос об ее исключении из анализа. Для этой цели привлекаются коэффициенты частной корреляции $r_{x_j x_j} \ (j = 1,2,\ldots,m)$ между объясняющими переменными. В качестве критерия используется величина:

$$t_j = \frac{r_{x_j x_j}\sqrt{n-m}}{\sqrt{1-r^2_{x_j x_j}}}, \qquad (4.6)$$

имеющая t-распределение с $v = n-m$ степенями свободы.

Если $t_{\text{факт}} > t_{\text{табл}}$, то между переменными существует коллинеарность и одна из переменных должна быть исключена. При исключении переменной исследователь должен опираться как на собственную интуицию, так и на содержательную теорию

явления. Если $t_{факт} \le t_{табл}$, то данные не подтверждают наличие коллинеарности между переменными x_j и x_k.

При столкновении с проблемой мультиколлинеарности может возникнуть естественное желание отбросить «лишние» независимые переменные, которые, возможно, служат ее причиной. Однако следует помнить, что при этом могут возникнуть новые трудности. Во-первых, далеко не всегда ясно, какие переменные являются лишними в указанном смысле. Мультиколлинеарность означает лишь приблизительную линейную зависимость между столбцами матрицы X, но это не всегда выделяет «лишние» переменные. Во-вторых, во многих ситуациях удаление каких-либо независимых переменных может отразиться на содержательном смысле модели. Наконец, отбрасывание так называемых существенных переменных, т.е. независимых переменных, которые реально влияют на изучаемую зависимую переменную, приводит к смещенности МНК-оценок [22, с. 95].

Линейное преобразование переменных.

Другой способ уменьшения или устранения мультиколлинеарности заключается в переходе к регрессии приведенной формы путем замены переменных, которым присуща коллинеарность их линейной комбинацией. Например, следует построить уравнение регрессии в виде: $\tilde{y}_i = b_0 + b_1 x_{1i} + b_2 x_{2i}$ и установлено, что переменные x_1 и x_2 высоко коррелированы. Анализ явления и результаты наблюдений позволяют постулировать дополнительное уравнение связи между объясняющими переменными x_1 и x_2, фигурирующими в исходной гипотезе, а именно $x_2^* = x_1 - x_2$. Переменную x^* подставляем в уравнение регрессии и получаем:

$$\tilde{y}_i = b_0^* + b_1^* x_{1i} + b_2^* x_{2i}^* . \qquad (4.7)$$

В общем случае переменные x_1 и x_2^* не сильно коррелируют. Таким образом достигается снижение или даже полное устранение мультиколлинеарности.

Исключение тренда. При построении регрессии по данным, полученным из временных рядов, рекомендуется исключить тренд или компенсировать изменение последовательных значений переменных (прирост). Этим достигается соблюдение

предпосылок регрессионного анализа - независимость наблюдений и уменьшение мультиколлинеарности.

Пошаговая регрессия. Процедура применения пошаговой регрессии начинается с построения простой регрессии. В анализ последовательно включают по одной объясняющей переменной. На каждом шаге проверяется значимость коэффициентов регрессии и оценивается мультиколлинеарность переменных. Если оценка коэффициента получается незначимой, то переменная исключается, и рассматривают другую объясняющую переменную. Если оценка коэффициента регрессии значима, а мультиколлинеарность отсутствует, то в анализ включают следующую переменную. Таким образом, постепенно определяются все составляющие регрессии без нарушения предположения об отсутствии мультиколлинеарности.

Метод главных компонент[1]. Метод главных компонент достаточно давно применяется для исключения или уменьшения мультиколлинеарности объясняющих переменных регрессии. В общих чертах суть метода сводится к следующему.

Поскольку мультиколлинеарность связана с высокой степенью корреляции между объясняющими переменными, можно попытаться обойти эту трудность, используя в качестве новых объясняющих переменных некоторые линейные комбинации исходных, выбранные так, чтобы корреляции между вновь введенными переменными были малы или вообще отсутствовали.

Основная идея заключается в сокращении числа объясняющих переменных до наиболее существенно влияющих факторов. Это достигается путем линейного преобразования всех объясняющих переменных x_j (j = 1,2 ..., m) в новые переменные, так называемые главные компоненты. При этом требуется, чтобы выделению первой главной компоненты соответствовал максимум общей дисперсии всех объясняющих переменных x_j (j = 1,2 ..., m), второй компоненте - максимум оставшейся дисперсии, после того как влияние первой главной компоненты исключается, и т.д. Таким образом, выполненное преобразование содействует уменьшению мультиколлинеарности

[1] Метод главных компонент был предложен впервые в 1901 г. К. Пирсоном, а затем развит, доработан, описан и обоснован в работах Г. Хоттелинга.

новых выделенных переменных по сравнению с мультиколлинеарностью набора исходных переменных x_j (j = 1,2 ..., m).

Решение задачи методом главных компонент сводится к поэтапному преобразованию матрицы исходных данных X (рисунок 4.3).

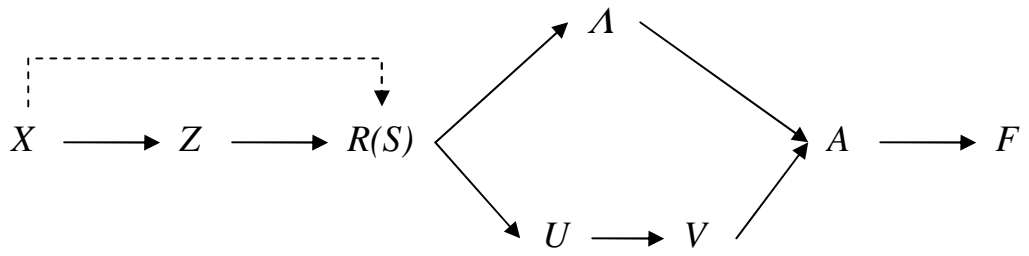


Рисунок 4.3 – Схема математических преобразований при использовании метода главных компонент

Здесь X – матрица исходных данных размерностью $n \times m$ (n – число объектов наблюдения, m – число элементарных аналитических признаков);

Z – матрица центрированных и нормированных значений признаков, элементы матрицы вычисляют по одной из формул:

$$z_{ij} = \frac{x_{ij}}{x_{\max_{ij}}}, \tag{4.8}$$

$$z_{ij} = \frac{x_{ij}}{x_{\min_{ij}}}, \tag{4.9}$$

$$z_{ij} = \frac{x_{ij}}{\bar{x}_j}, \tag{4.10}$$

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{\sigma_j}. \tag{4.11}$$

R – матрица парных корреляций:

$$R = (1/n) \times Z^T \times Z . \qquad (4.12)$$

Если предварительная стандартизация данных не проводилась, то на данном шаге получают матрицу $S = (1/n) \times X^T \times X$, элементы матрицы X для расчета будут центрированными величинами;

Λ – диагональная матрица собственных (характеристических) чисел.

Множество решений λ_j находят решением характеристического уравнения $|R - \lambda E| = 0$, λ_j – это характеристики вариации, точнее, показатели дисперсии каждой главной компоненты. Суммарное значение $\Sigma\lambda_j$ равно сумме дисперсий элементарных признаков X_j. При условии стандартизации исходных данных, эта сумма равна числу элементарных признаков m.

Решив характеристическое уравнение, находят его корни λ_j. После этого вычисляют собственные векторы матрицы R, т.е. решают m систем линейных уравнений для каждого λ_j при $j = 1,2,\ldots, m$.

Система имеет вид:

$$\begin{aligned}
&(1 - \lambda_j) \times \varepsilon_{1j} + r_{12} \times \varepsilon_{2j} + \ldots + r_{1m} \times \varepsilon_{mj} = 0 ; \\
&r_{21} \times \varepsilon_{1j} + (1 - \lambda_j) \times \varepsilon_{2j} + \ldots + r_{2m} \times \varepsilon_{mj} = 0 ; \\
&\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\
&r_{m1} \times \varepsilon_{1j} + r_{m2} \times \varepsilon_{2j} + \ldots + (1 - \lambda_j) \times \varepsilon_{mj} = 0 .
\end{aligned} \qquad (4.13)$$

Приведенная система объединяет однородные линейные уравнения, и так как число ее уравнений равно числу неизвестных, она имеет бесконечное множество решений. Конкретные значения собственных векторов при этом можно найти, задавая произвольно, по крайней мере, величину одной компоненты каждого вектора.

A – матрица факторного отображения, ее элементы a_{rj} – весовые коэффициенты. Вначале A имеет размерность $m \cdot m$ – по числу элементарных признаков X_j, затем в анализе остается r наиболее значимых компонент, $r \leq m$, Вычисляют матрицу A по известным данным матрицы собственных чисел Λ и нормированных собственных векторов V по формуле $A = V\Lambda^{1/2}$.

F – матрица значений главных компонент размерностью $r \times n$, $F = A^{-1}Z'$. Эта матрица в общем виде записывается:

$$F = \begin{vmatrix} & n_1 & n_2 & \cdots & n_n \\ F_1 & f_{11} & f_{12} & \cdots & f_{1n} \\ F_1 & f_{12} & f_{22} & \cdots & f_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ F_1 & f_{r1} & f_{r2} & \cdots & f_{rn} \end{vmatrix} \qquad (4.14)$$

При всех своих преимуществах (уменьшение мультиколлинеарности объясняющих переменных) метод главных компонент обладает и некоторыми недостатками:

1) главным компонентам, как правило, трудно подобрать экономические аналоги, поэтому вызывает затруднение экономическая интерпретация оценок параметров регрессии;

2) оценки параметров регрессии получают не по исходным объясняющим переменным, а по главным компонентам.

Метод главных компонент применяется, в основном, для оценки значений регрессии и для определения прогнозных значений зависимой переменной, что также является целью регрессионного анализа.

Коллинеарность и мультиколлинеарность факторов в экономических системах возникает не случайно. В совокупности однородных предприятий или регионов, как правило, в силу законов экономики, возникает параллельная вариация факторных признаков: те предприятия, которые имеют лучшие значения одних факторов, например – лучшие природные условия – одновременно имеют и более высокую фондо- и энерговооруженность (обеспеченность), более высокую квалификацию персо-

нала, лучшую технологию и т.п. Отсюда и неизбежная большая или меньшая коллинеарность всех факторов производства, либо социально-экономических условий жизни. Поэтому “бороться с коллинеарностью” следует осторожно, чтобы, как говорят, вместе с водой не выплеснуть из ванночки и ребенка! Первенство должен иметь экономический смысл модели, а не погоня за абсолютной чистотой математического аппарата исследования.

Спорной стороной данной проблемы остается и то, что между факторами может существовать, хотя и тесная, но нелинейная зависимость. Отсутствие коллинеарности не гарантирует от искажения математических условий МНК, если связь между факторами нелинейная [10, с. 69-70].

Пример 4.2 - Исследуется влияние факторов на заболеваемость населения субъектов РФ в 2010 г. (исходные данные представлены в таблице Е.1 приложения Е). В качестве объясняющих переменных отобраны следующие: *X1* – численность населения на 1 врача; *X2* – численность безработных, чел.; *X3* – доля населения с доходами ниже прожиточного минимума, %; *X4* – выбросы загрязняющих веществ в атмосферный воздух, отходящих от стационарных источников, тыс.т.; *X5* – использование свежей воды, млн. куб. м.; *X6* – сброс загрязненных сточных вод в поверхностные водоемы, млн. куб. м.; *X7* – удельный вес жилой площади, оборудованной водоотведением (канализацией), %; *X8* – число детей на 100 мест в дошкольных учреждениях.

В качестве результативного выступил показатель заболеваемости населения на 100000 человек населения. Обработка данных проводилась с использованием ППП Statistica.

Предварительно совокупность была исследована на нормальность распределения (рисунок 4.4). Близость распределения рассматриваемой совокупности к нормальному позволила проводить дальнейший регрессионный анализ.

Далее была построена матрица парных коэффициентов корреляции (рисунок 4.5). Корреляционная матрица отразила, что результативный показатель имеет тесную статистическую связь с факторами X5 и X8.

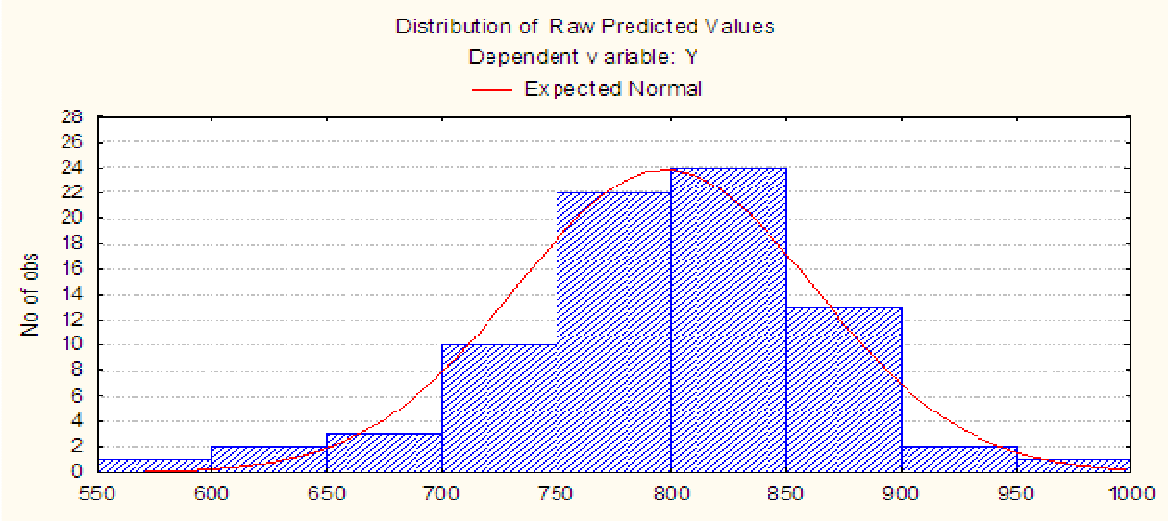


Рисунок 4.4 – Гистограмма распределения исходных признаков

Наряду с этим, наличие в матрице высоких значений парных коэффициентов корреляции между факторными признаками свидетельствует о присутствии мультиколлинеарности. Игнорирование данного обстоятельства при построении модели приведет к смещенности оценок и, как следствие - к неправильным выводам.

Correlations (Spreadsheet1)
Marked correlations are significant at p < ,05000
N=78 (Casewise deletion of missing data)

Variable	Means	Std.Dev.	Y	X1	X2	X3	X4	X5	X6	X7	X8
Y	794,5821	142,098	1,000000	-0,210334	-0,003324	-0,091043	0,123700	-0,231755	0,007377	0,076897	-0,282406
X1	218,5910	49,220	-0,210334	1,000000	-0,050566	0,071926	-0,141178	0,086641	-0,155964	-0,293059	-0,051264
X2	67,7949	46,406	-0,003324	-0,050566	1,000000	-0,293632	0,403687	0,471942	0,584427	0,151939	0,265317
X3	15,4872	4,759	-0,091043	0,071926	-0,293632	1,000000	-0,135106	-0,202920	-0,379662	-0,523763	-0,088541
X4	244,7692	487,758	0,123700	-0,141178	0,403687	-0,135106	1,000000	0,278833	0,268647	0,151491	0,222282
X5	756,0256	1047,753	-0,231755	0,086641	0,471942	-0,202920	0,278833	1,000000	0,405124	0,148629	-0,012156
X6	211,6833	278,733	0,007377	-0,155964	0,584427	-0,379662	0,268647	0,405124	1,000000	0,416018	0,133002
X7	68,9513	13,755	0,076897	-0,293059	0,151939	-0,523763	0,151491	0,148629	0,416018	1,000000	-0,020304
X8	106,3077	7,372	-0,282406	-0,051264	0,265317	-0,088541	0,222282	-0,012156	0,133002	-0,020304	1,000000

Рисунок 4.5 – Корреляционная матрица

Для недопущения такой ситуации необходимо устранить мультиколлинеарность на раннем этапе исследования с помощью специально предназначенных мето-

дов, например, шаговой регрессии. В ходе проведения шаговой регрессии были получены оценки регрессионной модели (рисунок 4.6).

Regression Summary for Dependent Variable: Y (Spreadsheet71
R= ,48087427 R?= ,23124006 Adjusted R?= ,18911623
F(4,73)=5,4895 p<,00064 Std.Error of estimate: 127,96

N=78	Beta	Std.Err. of Beta	B	Std.Err. of B	t(73)	p-level
Intercept			1580,293	219,3996	7,20281	0,000000
X2	0,187219	0,127018	0,573	0,3889	1,47395	0,144793
X4	0,244698	0,114156	0,071	0,0333	2,14354	0,035402
X5	-0,393098	0,119014	-0,053	0,0161	-3,30296	0,001484
X8	-0,391249	0,109106	-7,542	2,1031	-3,58594	0,000604

Рисунок 4.6 – Оценки множественной модели регрессии

Получено уравнение регрессии вида:

$$\tilde{y} = 1580{,}293 + 0{,}573x_2 + 0{,}071x_4 - 0{,}053x_5 - 7{,}542x_8$$
$$(0{,}12702)\quad (0{,}1142)\quad (0{,}119)\quad (0{,}1091)$$

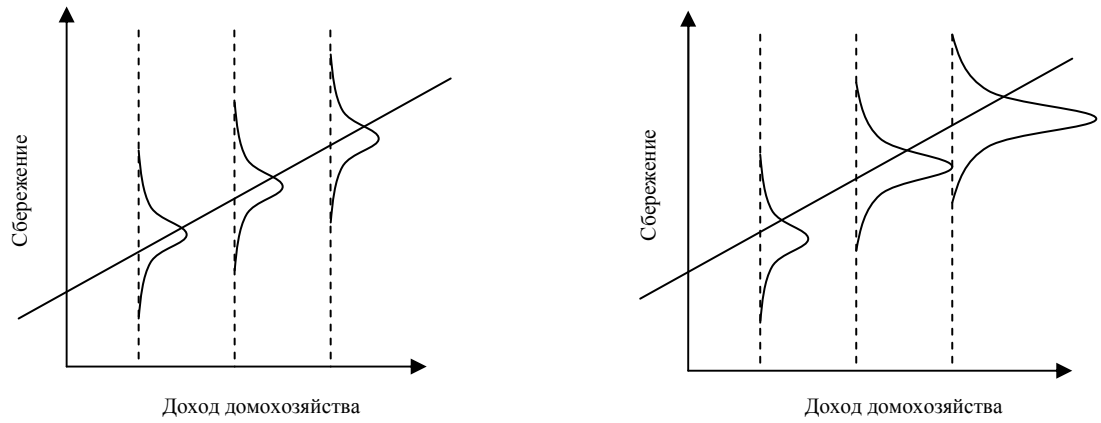
Фактор «Численность безработных» не был исключен из модели в связи с его достаточно высокой значимостью. Стандартизированный коэффициент регрессии отразил, что вариация заболеваемости населения на 0,19 % зависит от числа безработных в регионе. Значение *F*-критерия составило 5,49, табличное значение – 2,4. Следовательно, по *F*-критерию регрессионная модель в целом значима. Коэффициенты регрессии также значимы, т.к. расчетные значения по модулю превосходят по модулю табличное (2,0003). В ходе пошаговой регрессии, мультиколлинеарность была устранена, а полученная модель пригодна для экономического анализа и прогнозирования.

4.2 Гетероскедастичность

Одной из ключевых предпосылок МНК является условие постоянства дисперсий случайных отклонений ε_i для каждого значения x_i - *гомоскедастичность* (оди-

наковый разброс). Невыполнение данной предпосылки называется *гетероскедастичностью* (неодинаковый разброс) (рисунок 4.7).

В случае гомоскедастичности распределения остатков ε_i одинаковы, в отличие от распределений на графике гетероскедастичности, где диапазон вариации остатков изменяется при переходе от одного значения x_i к другому (соответственно дисперсия остатков неодинакова при разных значениях x_i).



гомоскедастичность гетероскедастичность

Рисунок 4.7 – Графический пример гомоскедастичного и гетероскедастичного распределения остатков

Поясним анализируемую предпосылку. В связи с тем, что случайное отклонение в каждом наблюдении имеет только одно значение, может возникнуть вопрос о том, что означает его дисперсия. Как указывает Доугерти К., имеется в виду его *возможное* поведение *до того*, как сделана выборка.

Фактические отклонения в выборке иногда будут положительными, иногда – отрицательными, иногда – относительно далекими от нуля, иногда – близкими к нулю, но у нас нет причин *a priori* ожидать появления особенно больших отклонений в любом данном наблюдении. Другими словами, вероятность того, что величина случайного члена примет какое-то данное положительное (или отрицательное) значение, будет одинаковой для всех наблюдений [6, с. 201]. Гетероскедастичность становится проблемой, когда значения переменных в уравнении регрессии значительно различаются в разных наблюдениях.

Часто в экономических данных встречаются «*выбросы*» (явления с аномально высокими или низкими значениями), которые в ряде случаев и становятся причиной гетероскедостичности. Естественные выбросы обусловлены различными видами экономической деятельности организаций, масштабом их деятельности и т.п. Искусственные выбросы – это ошибки, сделанные по вине регистратора.

Еще одной причиной возникновения гетероскедостичности стоит считать неверную спецификацию модели [*specification of a model*], о которой мы говорили во второй главе. Напомним, что спецификация модели - один из этапов построения экономико-математической модели, на котором на основании предварительного анализа рассматриваемого экономического объекта или процесса в математической форме выражаются обнаруженные связи и соотношения, а значит, параметры и переменные, которые на данном этапе представляются существенными для цели исследования [25].

Проблема гетероскедоскичности гораздо чаще встречается при работе с *пространственными данными* и довольно редко - при использовании *временных рядов*. Это связано с тем, что дисперсия возмущения ε_i, соответствующего значению X_i, по всей вероятности будет отличаться от дисперсии ε_j, если полученные эмпирически значения X_i и X_j независимой переменной X характеризуют элементы гетерогенной совокупности, например, объекты, значительно различающиеся по своим масштабам: фирмы с совершенно разным размахом операций, семьи, резко различающиеся между собой по уровню дохода, и т.п. Подобным источникам гетерогенности можно противопоставить наблюдение в разные моменты времени за одним и тем же экономическим объектом; такому наблюдению присущи устойчивые черты. Следует, однако, оговориться, что значительное изменение исследуемой величины во времени также может сопровождаться изменением дисперсии [17, с. 23].

При невыполнимости предпосылки о постоянстве дисперсий отклонений, т.е. при наличии при гетероскедастичности последствия применения МНК-оценок будут следующими:

1) оценки параметров по-прежнему останутся несмещенными и линейными;

2) оценки не будут эффективными (не будут иметь наименьшую дисперсию по сравнению с другими возможными оценками данного параметра);

3) дисперсии оценок будут рассчитываться со смещением;

4) вследствие вышесказанного все выводы, получаемые на основе соответствующих *t*- и *F*-статистик, а также интервальные оценки будут ненадежными, а статистические выводы, получаемые при стандартных проверках качества оценок, могут быть ошибочными и приводить к неверным заключениям по построенной модели.

В ряде случаев, зная характер данных, появление проблемы гетероскедостичности можно предвидеть и попытаться устранить этот недостаток еще на этапе спецификации.

В компьютерных пакетах реализованы некоторые процедуры коррекции на гетероскедастичность, например, тест Ньюи-Веста (*Newey-West-test)*, тест Уайта (*White-test) (Heteroscedastics Consistent Standard Error)*. Тест Уайта используется, если дисперсия ошибок зависит от времени, а ковариация рана нулю.

Однако значительно чаще рассматриваемую проблему приходится решать после построения уравнения регрессии. Для этого в настоящее время разработано довольно большое число тестов и критериев. Сущность их сводится к оценке различными способами взаимосвязи между:

- x_i и ε_i (или ε_i^2) – в случае парного регрессионного анализа;

- $\hat{y}_i$ и ε_i (или ε_i^2) – в случае множественного регрессионного анализа.

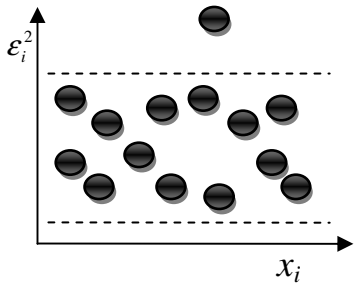
При этом в качестве нулевой гипотезой H_0 берется предположение о наличии гетероскедастичности (взаимосвязи между рассматриваемыми переменными).

Рассмотрим наиболее популярные и наглядные методы выявления гетероскедастичности.

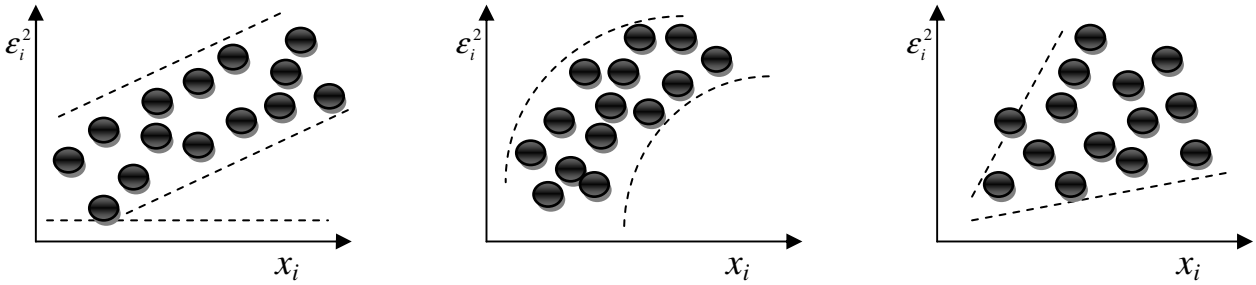
Графический анализ остатков.

В некоторых случаях гетероскедастичность очевидна визуально и ее можно обнаружить с помощью графического анализа отклонений. По оси абсцисс (*OX*) откладывается значения x_i (либо линейная комбинация объясняющих переменных

($\tilde{y}_i = b_0 + b_1 x_{1_i} + b_2 x_{2_i} + ... + b_k x_{k_i}$)), а по оси ординат ($OY$) либо отклонения ε_i, либо их квадраты ε_i^2 (с целью исключить отрицательные значения). Далее интерпретируют полученное изображение (рисунок 4.8).



гомоскедастичность



гетероскедастичность

Рисунок 4.8 – Графический анализ остатков регрессионной модели

Если все отклонения ε_i^2 находятся внутри полуполосы постоянной ширины, параллельной оси абсцисс, это говорит о независимости дисперсий ε_i^2 от значений переменной x_i и их постоянстве, т.е. выполняются условия гомоскедастичности. В противном случае речь идет о гетероскедастичности.

Достоинствами графического метода остатков являются простота применения, возможность реализации во всех статистических прикладных пакетах программ, способность выявить нелинейную взаимосвязь.

Тест ранговой корреляции Спирмена.

Коэффициент ранговой корреляции Спирмена является простой модификацией Коэффициента Пирсона, при которой величины x_i и y_i заменяются их рангами.

Поскольку ранги являются некоторой перестановкой чисел 1,2,…,*n* для каждой переменной, можно показать, что коэффициент ранговой корреляции Спирмена сводится к:

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}, \tag{4.15}$$

где $d_i = rank\ (x_i) - rank\left(e_i\right)\left(i = 1,2...,n\right)$.

Строго говоря в формуле (4.15) необходима коррекция, если имеются связки в двух ранжируемых множествах, но эффект коррекции пренебрежимо мал, если доля связок не слишком велика [26, с. 138].

Коэффициент изменяется в пределах от -1 до +1. Если мы получим значения, близкие к +1, значит, большим значениям независимой переменной отвечают большие значения результативной переменной. Если значение коэффициента близко к -1, то большие значения факторов отвечают меньшим значениям результата. Существуют достаточно полные таблицы распределений ρ [27, с. 400-406].

Этапы проведения теста Спирмена для выявления гетероскедастичности можно представить следующим образом:

1. Значения x_i и ε_i ранжируются.

2. Рассчитывают коэффициент Спирмена по формуле (4.15).

3. Для проверки значимости выдвигаются гипотезы:

- $H_0 : \rho = 0$ (нет гетероскедастичности);

- $H_1 : \rho \neq 0$ (есть гетероскедастичность).

Критерий проверки гипотезы основан на том факте, что

$$t_{факт} = \rho \frac{\sqrt{n-2}}{1-\rho^2} \tag{4.16}$$

имеет приближенно распределение Стьюдента с (n-2) степенями свободы.

4. $t_{факт}$ по модулю сравнивается с $t_{табл}$ (α/2; v=n-2). Если $t_{факт} > t_{табл}$, то гипотеза об отсутствии гетероскедастичности отклоняется.

В случае множественной регрессии проверка гипотезы может осуществляться с помощью *t*-статистики отдельно для каждой из объясняющих переменных.

Пример 4.3 - На примере регрессии, рассмотренной в параграфе 4.3 (исходные данные представлены в таблице Е.1 приложения Е) проведем проверку на гетероскедастичность одной из переменных, например, *X4* (выбросы загрязняющих веществ в атмосферный воздух, отходящих от стационарных источников).

Рассчитаем теоретические значения $\tilde{y}_x$ по уравнению регрессии и найдем остатки. Ранжируем совокупность по возрастанию (фрагмент расчетной таблицы приведен на рисунке 4.9).

	A	B	C	D	E	F	G	H	I
1	№ наблюдения	y	x4	теорет y	ε	модуль остатков	ранг x	ранг ε	d^2
2	1	756,7	132	790,51808	-33,818084	33,818084	47	14	1089
3	2	820,5	35	787,0225	33,477505	33,477505	20	13	49
4	3	925,8	35	787,0225	138,77751	138,777505	20	60	1600
5	4	548,8	77	788,53605	-239,73605	239,736049	35	70	1225
6	5	896,6	37	787,09457	109,50543	109,505431	22	52	900
7	6	757,6	12	786,19364	-28,593644	28,593644	6	12	36
8	7	799,9	54	787,7072	12,192802	12,192802	29	4	625
9	8	578,6	41	787,23872	-208,63872	208,638717	25	66	1681
10	9	682,6	368	799,02282	-116,42282	116,422816	67	53	196
11	10	659,3	205	793,14879	-133,84879	133,848785	56	57	1
...									
75	74	779,2	119	790,0496	-10,849603	10,849603	45	3	1764
76	75	809,8	25	786,66213	23,137875	23,137875	15	10	25
77	76	924,9	100	789,3649	135,5351	135,5351	40	59	361
78	77	707,9	23	786,59005	-78,690051	78,690051	12	36	576
79	78	1213,7	22	786,55401	427,14599	427,145986	9	78	4761
80	Итого	61977,4		61977,392					82568

Рисунок 4.9 – Фрагмент расчетной таблицы для проведения теста Спирмена

Сумма квадратов разностей рангов составила 82568. Тогда:

$$\rho_{x,\varepsilon} = 1 - 6 \times \frac{82568}{78(78^2 - 1)} = -0{,}04412 .$$

Рассчитаем статистику Стьюдента:

$$t_{расч} = \frac{-0{,}044120 \times \sqrt{78-2}}{\sqrt{1-(-0{,}04412)^2}} = -0{,}3854 .$$

Табличное значение статистики Стьюдента на уровне значимости $\alpha = 0{,}05$ с числом степеней свободы (78-2=76) составило $t_{0{,}05;76} = 1{,}984$. Расчетное значение по модулю меньше табличного, следовательно, гипотеза об отсутствии гетероскедастичности принимается на уровне значимости 5 % (с 95 %-ной вероятностью).

Аналогично проводится анализ для остальных факторных переменных.

Тест Глейзера [28].

Глейзер предложил процедуру, которая может быть использована для проверки гипотез о величине *n*.

Часто оказывается, что σ_i ведет себя несколько сложней, нежели в выражениях, где она ведет себя как линейная функция. Это обстоятельство побудило более точно описать σ_i. Глейзер предложил модель вида:

$$\sigma_i = \alpha + \beta \cdot |x_{il}|^{\gamma} + \delta_i . \qquad (4.17)$$

На это выражение смотрят как на регрессионную модель зависимости модулей отклонений от x_i , т.е. вместо значений результирующего признака будут оценки e_i (регрессионные остатки):

$$|e_i| = \alpha + \beta |x_{il}|^{\gamma} + \delta_i . \qquad (4.18)$$

γ – параметр, который подбирается исходя из следующих соображений: модель должна быть значима; значимо отличен от нуля β. Среди таких моделей выбирается та, в которой наибольшая величина коэффициента детерминации.

Глейзер обнаружил, что первые регрессии ($\gamma = -2; -1; -\frac{1}{2}; 1; 2$) в общем случае дают удовлетворительные результаты при обнаружении гетероскедастичности.

Тест Гольфельда-Квандта [29].

Данный тест применяется в случае, если ошибки регрессии можно считать нормально распределенными случайными величинами. Также предполагается, что дисперсия регрессионных остатков прямо или обратно пропорционально значению объясняющих переменных, вариацией которых и порождается гетероскедастичность. Выдвигаются гипотезы:

- $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_n^2$ (отсутствие гетероскедастичности);

- $H_1 : \exists i \neq j : \sigma_i^2 \neq \sigma_j^2$ (наличие гетероскедастичности).

Порядок проведения теста следующий:

1) ранжируются в порядке возрастания значения объясняющей переменной, которая подозревается на порождение гетероскедастичности; переставляются матрицы X и вектора Y в порядке возрастания той переменной, которая подозревается на порождение гетероскедастичности;

2) вся упорядоченная выборка объемом n разбивается на три подвыборки: первые n' наблюдений, последние n'' наблюдений, оставшиеся (n-$2n$) наблюдений. Обычно на середину берут четверть выборки объема, тогда $n' = n'' = (n-0{,}25n)/2$;

3) строятся уравнения регрессии y' по n' значениям первой подвыборки и y'' по n'' последней подвыборки;

4) оцениваются регрессионные остатки $\bar{e}'$ и $\bar{e}''$;

5) рассчитывается сумма квадратов отклонений $Q' = (\bar{e}')^T \cdot \bar{e}'$ и $Q'' = (\bar{e}'')^T \cdot \bar{e}''$;

6) строится статистика $F = \frac{\max(Q', Q'')}{\min(Q', Q'')}$, которая при справедливости нулевой гипотезы имеет закон распределения Фишера-Снедекора с числом степеней свободы $\nu_1 = n' - k - 1$, $\nu_2 = n'' - k - 1$. В случае если $Q'' > Q'$, делаем вывод о наличии прямой за-

висимости между объясняющей переменной и регрессионными остатками, если $Q' > Q''$, речь идет об обратной зависимости;

7) проверяется гипотеза о наличии/отсутствии гетероскедастичности.

Тест Уайта [30].

Данный тест применяется, если о форме гетероскедастичности ничего не известно и есть предположение, что дисперсии ошибок связаны с объясняющими переменными, а гетероскедастичность должна отражаться в остатках обычной регрессии исходной модели. Тестируется гипотеза H_0 : $\sigma_i^2 = \sigma^2$, $i = 1,2,...,n.$ Далее строим уравнение регрессии $y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + \varepsilon_i$. Находим $e_i^2 = (y_i - \tilde{y}_i)^2$. Строим вспомогательную регрессию квадратов остатков на все регрессоры *X*, их квадраты, попарные произведения и константу, если ее не было в составе исходных регрессоров: $\tilde{e}_i^2 = b_0 + b_1 x_{1i} + b_2 x_{2i} + b_3 x_{1i}^2 + b_4 x_{2i}^2 + b_5 x_{1i} x_{2i}$.

Для полученного уравнения рассчитывают R^2 и находят статистику Уайта: $W = nR^2$. По таблице χ^2 – распределения находят табличное значение с вероятностью α и степенями свободы $v = k(k+1)/2$. Сравнивают фактическое значение W-статистики и табличное значение χ^2. Если $W > \chi^2$, то гипотеза о гомоскедастичности может быть отвергнута.

На практике применение теста Уайта с включением и не включением попарных произведений дают, как правило, один и тот же результат.

Фактически, тест Уайта является очень общим. Хотя это является его достоинством, в то же время он имеет потенциально серьезный недостаток. Тестирование может обнаружить гетероскедастичность, но вместо этого может просто идентифицировать некоторую другую ошибку спецификации (как, например, некорректный функциональный вид уравнения регрессии). С другой стороны мощность теста Уайта может быть довольно низкой против некоторых определенных альтернативных гипотез, особенно если число наблюдений мало [31, с. 156].

Замечание: Если гипотеза о наличии гомоскедастичности отвергается, данный тест не дает указания на функциональную форму гетероскедастичности, и единст-

венным способом коррекции на гетероскедастичность является применение стандартных ошибок в форме Уайта [22, с. 88].

Предположив, что мы находимся в условиях модели с некоррелированными и гетероскедастичными остатками, а также предположив, что ковариационная матрица регрессионных остатков диагональная, Уайт показал, что оценка ковариационной матрицы, вычисленная по формуле:

$$\Sigma_b = n(X^T X)^{-1}(\frac{1}{n}\sum_{s=1}^{n} e_s^2 x_s x_s^T)\cdot(X^T X)^{-1}, \qquad (4.19)$$

где x_s^T, $s = 1,..n.$ $1\times k$ – векторы строки матрицы *X*, является состоятельной оценкой матрицы ковариаций оценок коэффициентов регрессии.

Стандартные отклонения, рассчитанные по этой формуле (4.19) называются стандартными ошибками в форме Уайта или состоятельными стандартными ошибками при наличии гетероскедастичности:

$$(\Sigma_b = n(X^T X)^{-1}(\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 \tilde{X}_i \tilde{X}_i^T)\cdot(X^T X)^{-1}, \varepsilon_i = y_i - \tilde{y}_i,\ \tilde{X}_i = (1, x_i^{(1)}, x_i^k)^T). \qquad (4.20)$$

Как было отмечено ранее, при несоблюдении условия Гаусса-Маркова о постоянстве дисперсии регрессионных остатков, оценки коэффициентов модели, вычисленные МНК, будут неэффективными (несмотря на их несмещенность). Поэтому при установлении гетероскедастичности возникает необходимость преобразования модели, а вид преобразования зависит от того, известны или нет дисперсии σ_i^2 отклонений ε_i, i= 1, 2, …, n. При известных для каждого наблюдения значениях σ_i^2 вместо МНК используют ВМНК - *метод взвешенных наименьших квадратов*.

В этом случае можно устранить гетероскедастичность, разделив каждое наблюдаемое значение на соответствующее ему значение дисперсии. Опишем ВМНК на приеме парной регрессии $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$.

Разделим обе части на σ_i :

$$\underbrace{\frac{y_i}{\sigma_i}}_{y^*} = \beta_0 \underbrace{\frac{1}{\sigma_i}}_{z_i} + \beta_1 \underbrace{\frac{x_i}{\sigma_i}}_{x_i^*} + \underbrace{\frac{\varepsilon_i}{\sigma_i}}_{v_i}. \qquad (4.21)$$

Тем самым наблюдениям с наименьшими дисперсиями придаются наибольшие «веса», а с максимальными дисперсиями - наименьшие «веса» (рисунок 4.10).

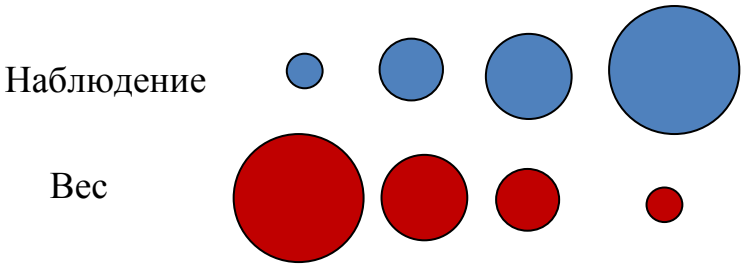


Рисунок 4.10 – Графическая иллюстрация результатов преобразования модели

Получим уравнение без свободного члена, но с дополнительной переменной *Z* и с преобразованным отклонением. Для преобразованной модели выполняются все условия Гаусса-Маркова. В этом случае оценки, полученные по МНК, будут несмещенными, состоятельными и эффективными.

В некоторых ситуациях априорно можно считать, что стандартное отклонение ошибки прямо пропорционально одной из независимых переменных, например, x_k: $\sigma^2{}_i = \sigma^2 x_i^2$. Тогда разделив уравнение на x_{ik} и вводя новые переменные, получим уравнение регрессии, которое удовлетворяет условиям Гаусса-Маркова. Допустим, что $\sigma^2{}_i = \sigma^2 x_1^2$, то тогда ВМНК предполагает оценку параметров следующего трансформированного уравнения:

$$\frac{y}{x_1} = b_0 \frac{1}{x_1} + b_1 + b_2 \frac{x_2}{x_1} + \ldots b_k \frac{x_k}{x_1}. \qquad (4.22)$$

Оценив для данного уравнения по МНК коэффициенты b_0, b_1, возвращаются к исходному уравнению регрессии.

Замечание: При наличии гетероскедастичности, когда применяется взвешенный метод наименьших квадратов коэффициент детерминации и скорректированный коэффициент детерминации *не являются* показателем качества регрессии.

В ряде случаев для устранения гетероскедастичности необходимо изменить спецификацию модели (например, линейную на лог-линейную, мультипликативную на аддитивную и т. п.).

4.3 Автокорреляция регрессионных остатков

Выше мы говорили о том, что получить несмещенную оценку с минимальной дисперсией в регрессионной модели МНК можно только в случае, если остатки (возмущения) в модели не зависимы друг от друга (не являются автокоррелированными). Если остатки автокоррелированы, то параметры регрессии не смещены, но стандартные ошибки недооценены и, в свою очередь, проверка значимости параметров ненадежна.

Автокорреляция (последовательная корреляция) определяется как корреляция между наблюдаемыми показателями, упорядоченными во времени (временные ряды) или в пространстве (перекрестные данные). Нарушения предпосылки Гаусса-Маркова о независимости остатков обусловлено, как правило, двумя причинами:

а) одна из объясняющих переменных, воздействие которой учитывается с помощью случайного члена, в действительности тесно связана с результативной переменной;

б) текущие наблюдения этой переменной коррелированы с ее прошлыми наблюдениями.

То же относится и к группе автокоррелированных переменных, которые в совокупности обнаруживают преимущественно линейную связь с результативной переменной.

Автокорреляция обычно имеет место, когда используются данные *временного ряда*. Чтобы подчеркнуть это, мы последуем за литературой и индексируем номер наблюдения индексом $t = 1,2,...,T$, а не индексом $i = 1,2,..., N$. Самое важное различие состоит в том, что теперь порядок наблюдений действительно имеет значение, и индекс отражает естественное упорядочивание. В общем, регрессионный остаток ε_i отражает влияние тех переменных, которые влияют на зависимую переменную, но которые не были включены в модель. Постоянство существования эффектов, не включенных в модель переменных, является частой причиной положительной автокоррелированности остатков. Если бы такие невключенные переменные наблюдались и могли бы быть включены в модель, то мы также могли бы интерпретировать полученную автокорреляцию как признак неправильно специфицированной модели. Этим объясняется, почему тесты на наличие автокорреляции очень часто интерпретируются как тесты на наличие неправильной спецификации. Некорректные функциональные формы, неучтенные переменные и неадекватная динамическая спецификация модели — все это может привести к наличию автокорреляции [31, с. 165].

Р. Винн, К. Холден отмечали в связи с этим, что «при работе с набором одновременных наблюдений явление автокорреляции или серийной корреляции остатков будет встречаться редко. Ведь вероятность того, что при обследовании выборки, состоящей из фирм, отраслей промышленности, стран и т.п., значение одной из переменных, зафиксированное для какого-либо объекта, окажется тесно связанным со значением, зафиксированным для другого объекта из той же выборки, невелика; с другой стороны, циклический характер движения многих экономических показателей приводит к тому, что примеры автокоррелированных остатков при работе с временными рядами встречаются часто, пожалуй, даже слишком часто» [17, с. 18].

Основным источником автокорреляции является природа рассматриваемых данных. Например, динамика такого показателя, как ВВП, характеризуется наличием достаточно устойчивых тенденций (рост или спад экономики). Таким образом, каждое текущее значение этого показателя обуславливает последующие значения, т.е. доход, полученный в текущем году, пойдет на развитие экономики в следующем периоде, тем самым увеличивая величину ВВП в этом периоде и так далее. Поэтому,

если в число регрессоров модели для описания динамики ВВП не включить лаги (лаг) зависимой переменной (ВВП), следует ожидать, что ошибки будут коррелированны. Постоянная направленность воздействия не включенных в уравнение переменных является наиболее частой причиной *положительной автокорреляции* - ее обычного для экономического анализа типа, а вероятность возникновения автокорреляции увеличивается при уменьшении интервала между рассматриваемыми уровнями.

В качестве основных причин автокорреляции можно назвать:

- ошибки измерения, возникающие в ходе статистического наблюдения;

- неверная спецификация модели (невключение значимых независимых переменных;

- невключение лаговых переменных, в том числе зависимой;

- неверно выбранный тип тренда (детерминированный или стохастический);

- неверно выбранная функциональная форма модели и т.д.

Рассмотрим, что будет с МНК-оценками в случае регрессионной модели с автокоррелированными остатками. МНК – оценка $\bar{b}$ является по-прежнему несмещенной оценкой вектора $\bar{\beta}$:

$$b = (X^T X)^{-1} X^T Y = (X^T X)^{-1} X^T (X\bar{\beta} + \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T \varepsilon, \tag{4.23}$$

$$M\bar{b} = M(\bar{\beta} + (X^T X)^{-1} X^T \bar{\varepsilon}) = \bar{\beta} + (X^T X)^{-1} X^T M\varepsilon = \bar{\beta}. \tag{4.24}$$

(согласно условиям Гаусса-Маркова).

Ковариационная матрица вектора случайного $\bar{b}$

$$\begin{aligned} &\sum_{\bar{b}} = M[(\bar{b} - M\bar{b})(\bar{b} - M\bar{b})^T] = M[(X^T X)^{-1} X^T \bar{\varepsilon}\bar{\varepsilon}^T X (X^T X)^{-1}] = \\ &= (X^T X)^{-1} X^T M(\bar{\varepsilon}\bar{\varepsilon}^T) X (X^T X)^{-1}. \end{aligned} \tag{4.25}$$

Σ_b является смещенной оценкой ковариационной матрицы ($M\varepsilon\varepsilon^T \neq \sigma^2 E_n$).

Делаем вывод, что использование МНК-оценок для регрессионной зависимости с автокоррелированными остатками приводит к следующим основным последствиям (в определенной степени сходным с последствиями гетероскедастичности):

1. Оценки параметров остаются несмещенными, но выборочные дисперсии этих оценок могут оказаться неоправданно большими по сравнению с дисперсиями достижимыми при применении несколько измененных методов оценивания.

2. С помощью обычных формул МНК для выборочных дисперсий параметров модели будет получена серьезная недооценка этих дисперсий (дисперсии оценены со смещением).

3. Получим неэффективные прогнозы, т.е. прогноз с чрезмерно большой выборочной дисперсией.

В силу вышесказанного выводы по *t*- и *F*-статистикам, определяющим значимость коэффициентов регрессии и коэффициента детерминации, возможно, будут неверными.

В практике эконометрических исследований разработаны ряд тестов и критериев, направленных на обнаружение автокорреляции. Во всех этих тестах в качестве нулевой гипотезы H_0 принимается предположение об отсутствии автокорреляции.

Графический анализа остатков.

Предполагается, что регулярная зависимость остатков от номера наблюдения или зависимость последующего значения остатка от предыдущего может быть свидетельством наличия автокорреляции.

По оси абсцисс (*OX*) откладывают либо время получения статистических данных, либо порядковые номера периодов наблюдения, а по оси ординат (*OY*) – отклонения ε_t (ε_t^2). При этом, если прослеживается определенная зависимость между переменными, то говорят о наличии автокорреляции.

Отсутствие зависимости в первом случае (рисунок 4.11) свидетельствует об отсутствии автокорреляции, в то время как на последующих графиках наблюдается связь между случайными отклонениями, т.е. имеет место автокорреляция.

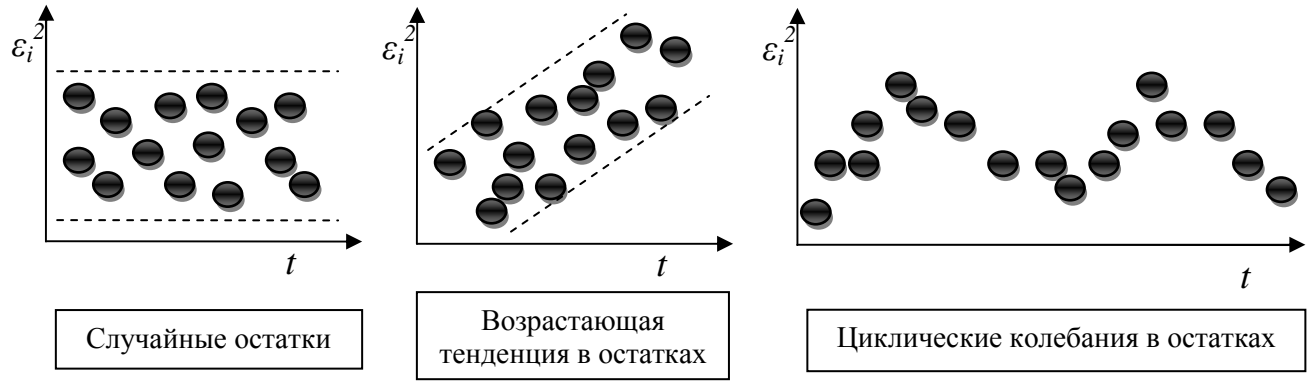


Рисунок 4.11 – Графический анализ остатков на автокорреляцию

Графический анализ позволяет только предположить наличие или отсутствие автокорреляции (заподозрить это явление), для ее выявления необходимо использовать статистические критерии.

В силу неизвестности значений параметров уравнения регрессии неизвестными будут также и истинные значения отклонений ε_t, $t = 1, 2, ..., T$. Поэтому выводы об их независимости осуществляются на основе оценок e_t, $t = 1, 2, ..., T$, полученных из эмпирического уравнения регрессии.

Метод рядов.

Строят уравнение регрессии и находят отклонения. Далее последовательно определяются знаки отклонений e_t, $t = 1, 2, ..., T$. Ряд определяется как непрерывная последовательность одинаковых знаков, а количество знаков в ряду называется длиной ряда. Визуальное распределение знаков свидетельствует о неслучайном характере связей между отклонениями. Если рядов слишком мало по сравнению с количеством уровней T, то вполне вероятна положительная автокорреляция. Если же рядов слишком много, то вероятна отрицательная автокорреляция. Для более детального анализа предлагается следующая процедура.

Пусть T – количество уровней ряда (объем выборки);

T_1 – общее количество знаков «+» при T наблюдениях (количество положительных отклонений e_t);

T_2 – общее количество знаков «-» при T наблюдениях (количество отрицательных отклонений e_t);

k - количество рядов.

Далее находим следующие величины:

$$M(k) = \frac{2T_1T_2}{T_1 + N_2} + 1, \tag{4.26}$$

$$D(k) = \frac{2T_1T_2\left(2T_1T_2 - T_1 - T_2\right)}{\left(T_1 + T_2\right)^2\left(T_1 + T_2 - 1\right)}. \tag{4.27}$$

Если при достаточно большом количестве наблюдений (T_1>10, T_2>10) количество рядов k лежит в пределах

$$M(k) - D(k) < k < M(k) + D(k), \tag{4.28}$$

то гипотеза об отсутствии автокорреляции не отклоняется.

Тест Дарбина-Уотсона [32].

Одним из самых популярных тестов в эконометрике является тест Дарбина - Уотсона. Два важных предположения, лежащие в основе этого теста, состоят в том, что мы можем рассматривать x_t-ые как детерминированные и что x_t содержит свободный член. Первое предположение является важным, поскольку оно требует, чтобы все регрессионные остатки были независимы от всех объясняющих переменных. Наиболее важно, что это исключает включение лагированных зависимых переменных в модель [31, с. 172].

Будем предполагать, что регрессионные остатки коррелированны и образуют наиболее простой процесс – авторегрессию первого порядка:

$$\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i. \tag{4.29}$$

где ρ - коэффициент корреляции между регрессионными остатками

δ_i - случайная величина, которая удовлетворяет требованиям, предъявляемым

к регрессионным остаткам КЛММР.

Нулевая гипотеза H_0: $\rho = 0$ (нет явления автокорреляции); альтернативная H_1: $\rho \neq 0$ (есть явление автокорреляции). Для проверки гипотезы используется статистика Дарбина-Уотсона:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}. \qquad (4.30)$$

Так как

$$\sum_{i=2}^{n}(e_i - e_{i-1})^2 = \sum_{i=2}^{n}(e_i^2 - 2e_i e_{i-1} + e_{i-1}^2) = \sum_{i=2}^{n} e_i^2 - 2\sum_{i=2}^{n} e_i e_{i-1} + \sum_{i=2}^{n} e_{i-1}^2 \approx 2\sum_{i=2}^{n} e_i^2 - 2\sum_{i-2}^{n} e_i e_{i-1}, \quad (4.31)$$

то

$$DW = \frac{2\left(\sum e_i^2 - \sum e_i e_{i-1}\right)}{\sum e_i^2} = 2\left(1 - r_{e_i e_{i-1}}\right). \qquad (4.32)$$

Учитывая тот факт, что $-1 \le \left(r_{e_i e_{i-1}}\right) \le 1$, можем указать, в каких пределах изменяется статистика DW:

- если $r_{e_i e_{i-1}} \approx 0$ (автокорреляция отсутствует), то $DW \approx 2$;

- если $r_{e_i e_{i-1}} \approx 1$ (положительная автокорреляция), то $DW \approx 0$;

- если $r_{e_i e_{i-1}} \approx -1$ (отрицательная автокорреляция), то $DW \approx 4$.

Следовательно, $0 \le DW \le 4$.

Рассчитав статистику DW, необходимо найти нижнюю $d_н$ и верхнюю $d_в$ границы на уровне значимости $\alpha = 0{,}05$ (таблица Ж.1 приложения Ж).

Если фактически наблюдаемое значение DW:

- $d_в < DW < 4\text{-}d_в$, то гипотеза об отсутствии автокорреляции принимается;

- $d_н < DW < d_в$ или $4\text{-}d_в < DW < 4\text{-}d_н$, область неопределенности критерия (вопрос об отвержении или принятии гипотезы остается открытым);

- $0 < DW < d_н$, то принимается альтернативная гипотеза о положительной автокорреляции;

- $4 - d_н < DW < 4$, то принимается альтернативная гипотеза об отрицательной автокорреляции.

В условиях справедливости нулевой гипотезы (H_0: $\rho = 0$) значения статистики Дарбина-Уотсона должны группироваться в некоторой окрестности своего среднего (т.е. в окрестности числа 2).

Следует отметить, что данный тест не лишен недостатков: наличие зоны неопределенности и ограниченность результата (выявляется лишь корреляция между соседними членами). Это приводит к необходимости использовать также и другие тесты на наличие автокорреляции.

Тест Дарбина.

Эта модификация теста Дарбина-Уотсона специально предназначена для случая, когда среди независимых переменных имеются запаздывающие значения зависимой переменной.

Строят *h*-статистику по формуле:

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{T}{1 - T\sigma^2_{y_{t-1}}}}, \qquad (4.33)$$

где DW – статистика Дарбина-Уотсона;

$\sigma^2_{y_{t-1}}$ - оценка дисперсии коэффициента при y_{t-1};

T – число уровней ряда.

Расчетное значение статистики сравнивается с критическим значением по таблице стандартного нормального распределения на заданном уровне значимости. Высокие значения *h* свидетельствуют против гипотезы об отсутствии автокорреляции. Недостаток теста в невозможности вычислить *h*, когда $\sigma^2_{y_{t-1}} > 1/T$.

Тест серий Бреуша-Годфри.

Преимущество теста Бреуша-Годфри по сравнению с тестом Дарбина-Уотсона заключается в первую очередь в том, что он проверяется с помощью статистического критерия, между тем как тест Дарбина-Уотсона содержит зону неопределенности для значений статистики DW. Другим преимуществом теста является возможность обобщения: в число регрессоров могут быть включены не только остатки с лагом 1, но и с лагом 2, 3 и т.д., что позволяет выявить корреляцию не только между соседними, но и между более отдаленными наблюдениями [16, с. 174-175].

Суть теста в следующем: если имеется корреляция между соседними наблюдениями, то в уравнении остатков регрессии (полученных обычным МНК) коэффициент ρ окажется значимо отличным от нуля:

$$e_t = \rho e_{t-1} + \varepsilon_t' . \qquad (4.34)$$

На основе *t*-критерия Стьюдента проверяют статистическую значимость параметра ρ. Если $t_{факт} > t_{табл}$ (параметр статистически значим), то в анализируемом ряду наблюдается автокорреляция.

Q - тест Бокса-Пирса.

На первом этапе рассчитывают эмпирическую статистику по формуле:

$$Q = T\sum_{j=1}^{p} r_j^2 . \qquad (4.35)$$

Выборочный коэффициент автокорреляции вычисляется по формуле:

$$r_j = \frac{\sum_{t=j+1}^{T} e_t e_{t-j}}{\sum_{t=1}^{T} e_t^2} . \qquad (4.36)$$

Далее находят табличное значение χ^2 при уровне значимости α и степенями свободы *p*. Если $Q > \chi^2$, то гипотеза об отсутствии автокорреляции отвергается.

Тест Льюинга-Бокса.

Рассчитывают эмпирическую статистику по формуле:

$$Q' = T(T-2)\sum_{j=1}^{p}\frac{r_j^2}{T-j}. \qquad (4.37)$$

Табличное значение χ^2 при уровне значимости α и степенями свободы *p* сравнивают полученным расчетным значением. Если $Q > \chi^2$, то гипотеза об отсутствии автокорреляции отвергается.

При наличии автокорреляции в модели требуются особые методы оценивания. В случае, если известны коэффициенты авторегрессии, можно воспользоваться обобщенным МНК, но подобная ситуация встречается крайне редко. При неизвестных коэффициентах существуют специальные процедуры оценивания модели, которые, как правило, имеют итеративный характер. Тем не менее, для устранения корреляции во времени чаще прибегают к изменению спецификации модели (исключают или добавляют регрессоры, включают лаги переменных), поскольку в значительном числе случаев именно неверная спецификация и является источником автокорреляции.

Однако если все разумные процедуры изменения спецификации модели, на ваш взгляд, исчерпаны, а автокорреляция имеет место, то можно предположить, что она обусловлена какими-то внутренними свойствами ряда ε_t. В этом случае можно воспользоваться обобщенным методом наименьших квадратов (ОМНК). Для его применения нужно специфицировать модель автокорреляции регрессионных остатков. В случае линейного регрессионного уравнения (либо в моделях, сводящихся к линейной), в качестве такой модели используется *авторегрессионный процесс первого порядка AR(1).*

Для простоты изложения *AR(1)* рассмотрим модель парной линейной регрессии. Тогда наблюдениям *t* и (*t* -1) соответствуют формулы:

$$y_t = b_0 + b_1 x_t + \varepsilon_t, \qquad (4.38)$$

$$y_{t-1} = b_0 + b_1 x_{t-1} + \varepsilon_t . \qquad (4.39)$$

Пусть случайные отклонения подвержены воздействию авторегрессии первого порядка:

$$\varepsilon_t = \rho \varepsilon_{t-1} + u_t , \qquad (4.40)$$

где u_t, $t = 2, 3, \ldots, T$ - случайные отклонения, удовлетворяющие всем предпосылкам МНК, а коэффициент ρ известен.

Вычтем из $y_t = b_0 + b_1 x_t + \varepsilon_t$ соотношение $y_{t-1} = b_0 + b_1 x_{t-1} + \varepsilon_t$, умноженное на ρ:

$$y_t - \rho y_{t-1} = b_0 (1 - \rho) + b_1 (x_t - \rho x_{t-1}) + (\varepsilon_t - \rho \varepsilon_{t-1}) . \qquad (4.41)$$

Последовательно заменяя

$$y_t' = y_t - \rho y_{t-1}; \quad x_t' = x_t - \rho x_{t-1}; \; b_0' = (1 - \rho); \varepsilon_t' = \varepsilon_t - \rho \varepsilon_{t-1}, \qquad (4.42)$$

получим:

$$y_t' = b_0' + b_1 x_t' + u_t . \qquad (4.43)$$

Так как по предположению коэффициент ρ известен, то очевидно, что y_t' , x_t' , u_t вычисляются достаточно просто. В силу того, что случайные отклонения u_t удовлетворяют предпосылкам МНК, оценки b_0' и b_1 будут обладать свойствами наилучших линейных несмещенных оценок.

Однако способ вычисления y_t' , x_t' приводит к потере первого наблюдения (если мы не обладаем предшествующим ему наблюдением). Число степеней свободы уменьшится на единицу, что при больших выборках не так существенно, но при малых выборках может привести к потере эффективности. Эта проблема обычно преодолевается с помощью *поправки Прайса-Винстена:*

$$x_t' = \sqrt{1-\rho^2} \times x_1 ;\ y_t' = \sqrt{1-\rho^2} \times y_1 . \qquad (4.44)$$

Авторегрессионное преобразование может быть обобщено на произвольное число независимых переменных, т.е. использовано для уравнения множественной регрессии.

Пример 4.4 - Исследуется взаимосвязь среднедушевых денежных доходов домохозяйств и среднедушевых денежных расходов на оплату услуг в Оренбургской области [33, с. 120-123].

Анализировалось 28 результатов наблюдений в поквартальной динамике за период с 2000 г. по 2006 г.

В ходе регрессионного анализа было получено следующее уравнение:

$$\tilde{y}_t = -45{,}3331 + 0{,}1207 x_t .$$

Результаты оценивания уравнения отразили значимость уравнения в целом и его параметра:

Число наблюдений	28
Стандартная ошибка коэффициента регрессии	0,00560
R - квадрат	0,94702
Значение *t*-критерия для коэффициента регрессии	21,557
Значение *F* - критерия	464,72
Критерий *DW*	2,28

Проверка модели на адекватность выявила, что гипотеза об автокорреляции в остатках не отвергается и не принимается (область неопределенности).

Дополнительно был проведен анализ автокорреляционной функции остатков модели, показавший наличие автокорреляции в остатках (рисунок 4.12).

Так как существует автокорреляция в остатках, найденные оценки параметров уравнения не являются эффективными вследствие нарушения предпосылок МНК.

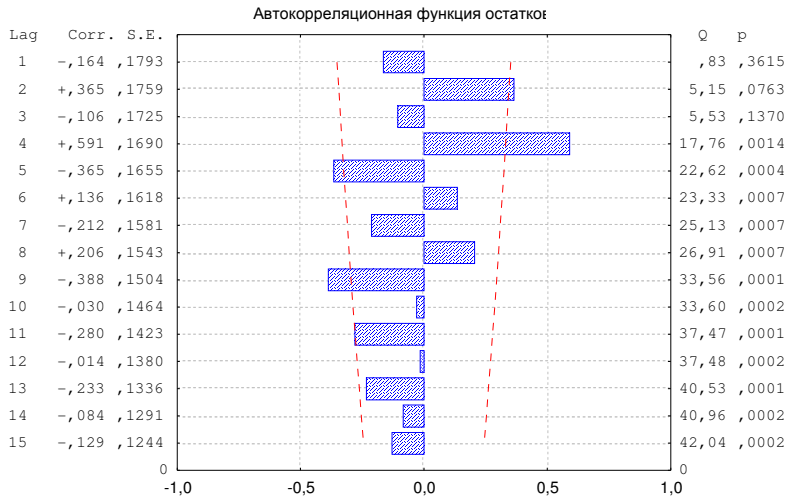


Рисунок 4.12 - Проверка остатков модели на автокоррелированность

Для получения эффективных оценок были рассчитаны параметры уравнения регрессии при наличии автокорреляции в остатках (используется обобщенный метод наименьших квадратов) в соответствии с формулами (4.42). Далее определялись параметры уравнения регрессии y_t' на x_t' обычным методом наименьших квадратов. В результате было получено следующее уравнение:

$$y_t' = -60{,}3596 + 0{,}1226 x_t' .$$

По формуле (4.45) пересчитан параметр β исходного уравнения:

$$\beta' = \beta(1-\rho) . \qquad (4.45)$$

В результате получено следующее уравнение зависимости среднедушевых расходов домохозяйств на оплату услуг от среднедушевых доходов:

$$\tilde{y}_t = -51{,}87293 + 0{,}1226 x_t$$

$$(0{,}00508)$$

Коэффициент детерминации уравнения составил 0,959, *t* – критерий для коэффициента регрессии равен 24,1. Уравнение значимо и значим коэффициент регрессии. График распределения остатков на нормальной вероятностной бумаге свидетельствует о том, что распределение остатков модели близко к нормальному (рисунок 4.13).

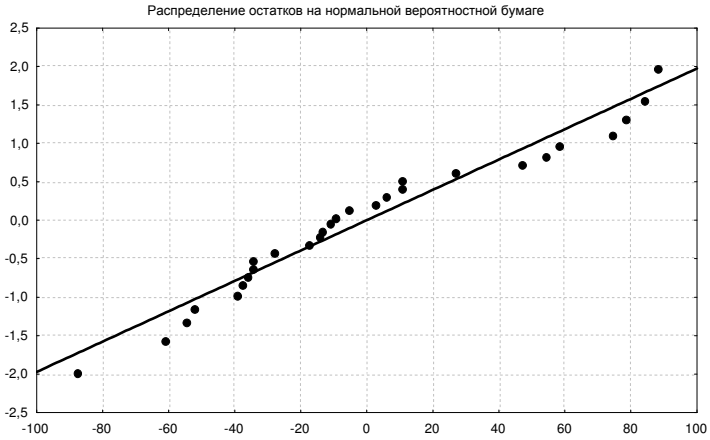


Рисунок 4.13 - Проверка остатков регрессионной модели на нормальность распределения

Регрессионная модель имеет следующую интерпретацию: с ростом среднедушевых доходов домохозяйств в месяц на 1000 р., среднедушевые расходы на оплату услуг в месяц вырастают на 123 р. С учетом того, что доля расходов на оплату услуг в среднедушевых доходах населения составляет примерно 11 %, полученные результаты вполне адекватны реальной ситуации.

На практике значение коэффициента ρ обычно неизвестно и его необходимо оценивать. Существует несколько методов оценивания.

Определение параметра ρ на основе статистики Дарбина-Уотсона.

Статистика Дарбина-Уотсона тесно связана с коэффициентом корреляции между соседними отклонениями через соотношение

$$DW \approx 2\left(1 - r_{e_i e_{i-1}}\right). \qquad (4.46)$$

Тогда в качестве оценки коэффициента ρ может быть взят коэффициент

$$r = 1 - \frac{DW}{2}. \tag{4.47}$$

Данный метод оценивания хорош при большом числе наблюдений.

Процедура Кохрейна-Оркатта.

Этот итерационный метод рассмотрим на примере парной регрессии $Y = \beta_0 + \beta_1 X + \varepsilon$ и авторегрессии первого порядка $\varepsilon_i = \rho\varepsilon_{i-1} + \delta_i$. Оценка

$$e_t = \rho e_{t-1} + \delta_t. \tag{4.48}$$

На выражение (4.48) будем смотреть как на линейную модель парной регрессии, где в роли результативного признака рассматривается e_t, а в роли объясняющей переменной e_{t-1}:

$$y = \begin{pmatrix} e_2 \\ \dots \\ e_n \end{pmatrix}, \quad x = \begin{pmatrix} e_1 \\ \dots \\ e_{n-1} \end{pmatrix}. \tag{4.49}$$

Методом наименьших квадратов оценивается уравнение регрессии $\tilde{y}_{i1} = b_0 + b_1 x_i$; для него определяются оценки $e_{i1} = (y_i - \tilde{y}_{i1})$.

Далее оценивается регрессионная зависимость

$$e_i = \tilde{\rho} e_{i-1} + \delta_i, \tag{4.50}$$

где $\tilde{\rho}$ - оценка коэффициента ρ.

Реализуя метод наименьших квадратов, найдем оценку коэффициента ρ и матрицу $\tilde{\Sigma}_0^{(1)}$. На следующем этапе:

$$b_{ОМНК} = (X^T \Sigma_0^{-1} X)^{-1} (X^T \Sigma_0^{-1} Y) \tag{4.51}$$

оценивается уравнение регрессии

$$\tilde{y}_{i2} = b_0 + b_1 x_i \,. \qquad (4.52)$$

Затем вновь вычисляются оценки e_{i2} отклонений и возвращаются к этапу реализации МНК. Процесс продолжается до тех пор, пока не будет достигнута требуемая точность (пока $\tilde{\rho}$ не стабилизируются), т.е. пока разность между предыдущей и последующей оценками ρ не станет меньше любого наперед заданного числа.

Метод Хилдрета-Лу.

Согласно данного метода в регрессии (4.42) b_0 и b_1 оцениваются для каждого возможного значения ρ из отрезка [-1,1] с любым шагом (например, 0,001; 0,01 и т. д.). Величина коэффициента ρ, дающая наименьшую стандартную ошибку регрессии, принимается в качестве оценки коэффициента ρ. И значения b_0' и b_1 оцениваются из уравнения регрессии именно с данным значением ρ. Этот итерационный метод широко используется в пакетах прикладных программ.

4.4 Спецификация модели множественной регрессии

Во второй главе мы уже останавливались на проблемах спецификации регрессионной модели. Рассмотрим подробнее некоторые особенности этого вопроса применительно к модели множественной регрессии.

Даже качественная модель является подгонкой спецификации модели под имеющийся набор данных. Поэтому вполне реальна картина, когда исследователи, обладающие разными наборами данных, строят разные модели для объяснения одной и той же переменной. Проблематичным является и использование модели для прогнозирования значений объясняемой переменной. Иногда хорошие с точки зрения диагностических тестов модели обладают весьма низкими прогнозными качествами.

Одно из главных направлений эконометрического анализа — постоянное совершенствование моделей. Здесь следует отметить, что какого-то универсального подхода, определяющего заранее возможные пути совершенствования, нет и, скорее всего, быть не может. Исследователь должен помнить, что совершенной модели не существует. В силу постоянно изменяющихся условий протекания экономических процессов не может быть и постоянно качественных моделей. Новые условия требуют пересмотра даже весьма устойчивых моделей.

До сих пор достаточно спорным является вопрос, как строить модели:

а) начинать с самой простой и постоянно усложнять ее;

б) начинать с максимально сложной модели и упрощать ее на основе проводимых исследований.

Оба подхода имеют как достоинства, так и недостатки. Так, если следовать схеме а), то происходит обыкновенная подгонка модели под эмпирические данные. При теоретически более оправданном подходе б) поиск возможных направлений совершенствования модели зачастую сводится к полному перебору, что делает проводимый анализ неэффективным. На этапах упрощения модели возможно также отбрасывание объясняющих переменных, которые были бы весьма полезны в упрощенной модели, поэтому построение модели является индивидуальным в каждой конкретной ситуации и опирается на серьезные знания экономической теории и статистического анализа.

Проверка всех возможных регрессий.

Процесс отбора существенных переменных можно рассматривать как процесс выбора *истинной модели* из множества возможных линейных моделей, которые могут быть построены с помощью набора предсказывающих переменных, и тогда полученные после отбора оценки коэффициентов можно рассматривать как несмещенные.

Решается следующая задача: для заданного значения k (k=1,2,…, p-1) путем полного перебора всех возможных комбинаций из k объясняющих переменных, отобранных их исходного (априорного) набора x_1, x_2,... x_p, состоящего из p предикторов, определить такие переменные, для которых коэффициент детерминации с результирующим показателем y был бы максимальным [34, с. 154].

Суть метода заключается в том, чтобы вычислить коэффициенты всех возможных регрессионных моделей и сравнить их характеристики. При сравнении рассчитанных моделей обычно принимают во внимание коэффициенты множественной детерминации R^2, остаточные дисперсии $\sigma^2_{ост}$ или C_k – статистика Маллоуза. При сравнении с использованием R^2 число предикторов, участвующих в модели, увеличивают до тех пор, пока прирост R^2 не станет слишком небольшим по сравнению с максимально возможным его значением R^2_{max}.

Если возможная самая полная структура модели регрессии:

$$y_i = b_0' + \sum_{j=2}^{m} b_j x_{ji} + \varepsilon_i, \qquad (4.54)$$

т. е. если модель содержит m коэффициентов, то процедура следующая:

1. Все уравнения делят на m подмножеств. Первое из них включает $y_i = b_0' + \varepsilon_i$, второе - все возможные уравнения с двумя коэффициентами, третье - все возможные уравнения с тремя коэффициентами и т.д.

2. Внутри каждого подмножества полученные оценки моделей упорядочивают по возрастанию величины R^2.

3. Исследуют в каждом подмножестве уравнение с максимальным значением R^2 и выясняют, нет ли какой-нибудь закономерности в последовательности появления регрессоров. Если при переходе от модели одного подмножества к модели следующего не наблюдается существенного прироста R^2, полагают, что нет особой нужды во включении в модель дополнительного регрессора (в таком случае можно воспользоваться корреляционной матрицей для выяснения вопроса, нет ли сильной корреляции между новым регрессором и каким-то из включенных ранее).

Обозначим через R^2_k - коэффициент множественной детерминации модели с самой полной возможной структурой, а через R^2_p - для модели с $p = k\text{-}q$ коэффициентами (включая и свободный член). Эйткен предложил критерий для проверки значимости различий между R^2_k и R^2_p. Различие считается незначимым, если выполнено условие:

$$\frac{R_k^2 - R_p^2}{1 - R_k^2 / n-k} = (k-1)F_{табл}\ , \qquad (4.55)$$

где $F_{табл}$ - табличное значение распределения Фишера при уровне значимости α и степенях свободы $v_1 = m\text{-}1$ и $v_2 = n\text{-}m$.

Если приведенное условие (4.55) не выполняется, модель с k коэффициентами лучше модели с меньшим числом (p) регрессоров.

Применение статистики $s^2_{ост}$ подробно описано выше. При очень малом числе регрессоров $s^2_{ост}$ велика, а с добавлением новых - уменьшается. Постепенно скорость уменьшения $s^2_{ост}$ тоже замедляется, и после включения определенного числа (скажем p) регрессоров величина $s^2_{ост}$ становится почти постоянной. Тогда можно считать, что модель с p регрессорами достаточно хороша, а $s^2_{ост}$ этой модели - оценка истинной дисперсии.

C_k – статистика Маллоуза.

Статистика Маллоуза рассчитывается по формуле:

$$C_k = \frac{RSS}{s^2 - (n-2k)} = \frac{\sum_{i=1}^{n}(\tilde{y}_i - \bar{y})^2}{s^2 - (n-2k)}, \qquad (4.56)$$

где n - число наблюдений;

k - число независимых переменных.

Если для выбора модели используется статистика Маллоуза, то наилучшей из построенных моделей признается модель с наименьшими значениями критерия.

Проверка всех возможных регрессий весьма трудоемкая процедура. Поскольку для каждого регрессора есть два пути (быть или не быть включенным в модель), число возможных комбинаций при условии, что свободный член всегда включен в модель, равно 2^{m-1}. Так, полный полином второй степени для 4 факторов содержит

15 коэффициентов (считая и свободный член). Число возможных регрессий в этом случае будет $2^{14} = 16384$.

Метод включения и метод исключения.

Эти два метода требуют гораздо меньше вычислений, чем проверка всех возможных регрессий. По методу включения сначала строят модель, состоящую только из свободного члена ($y_i = b_0' + \varepsilon_i$). Затем один за другим добавляют остальные регрессоры, а порядок их включения определяют по частным коэффициентам корреляции регрессоров x_j с откликом y. Регрессор, включаемый на данном этапе, должен иметь максимальный частный коэффициент корреляции. После введения нового регрессора в модель вычисляют множественный коэффициент детерминации R^2 и частный F-критерий для этого регрессора. Когда величина частного F-критерия станет меньше, чем табличное значение распределения Фишера $F_{\text{табл}}$, включение новых регрессоров в модель прекращается, поскольку считается, что это не уменьшит существенно вариации результативной переменной.

По поводу включения в модель дополнительных переменных Р. Винн и К. Холден отмечают следующее. Если в модель включена переменная, не имеющая отношения к делу, то оценки других параметров, рассчитанные по методу наименьших квадратов, останутся несмещенными, причем несмещенные дисперсии этих параметров также могут быть получены с помощью обычных процедур. И это вовсе не означает, что модель регрессии можно безнаказанно «засорять» переменными, выбранными «на авось». Во-первых, существует ненулевая вероятность того, что в результате использования выборочных данных переменная, которая вообще говоря, не имеет никакого отношения к модели, обнаружит существенную связь с зависимой переменной. Во-вторых, из того, что было сказано ранее по вопросу о мультиколлинеарности, следует, что в тех случаях, когда независимая переменная справедливо включена в модель, дисперсия оценки коэффициента при этой переменной – оценки, полученной по методу наименьших квадратов, неизбежно будет увеличиваться при включении в модель еще одной независимой переменной, которая, возможно, вообще не имеет отношения к делу; исключение из этого правила составляет лишь слу-

чай, когда выборочная корреляция между обеими переменными равна нулю [17, с. 29].

Метод исключения корректирует структуру модели в обратном порядке. Оценивают коэффициенты модели с самой полной возможной структурой и рассчитывают значения частных *F*-критериев для каждого регрессора при условии, что именно он исключается из модели.

Далее находят самое малое значение частного *F*-критерия (F_{min}) и сравнивают его с ранее выбранным при определенном уровне значимости табличным значением $F_{табл}$. В результате принимают одно из альтернативных решений:

а) если $F_{min} < F_{табл}$, регрессор, которому соответствует это наименьшее значение частного *F*-критерия, исключают из модели и переходят к следующему шагу исключения;

б) если $F_{min} > F_{табл}$, регрессионную модель далее не изменяют и процедуру исключения на этом заканчивают.

Из вышесказанного следует, что метод исключения обеспечивает лучшие результаты, чем метод включения. Последний не позволяет учесть влияние, оказываемое включаемым новым регрессором на вклады тех регрессоров, которые уже были включены в модель на предыдущих этапах.

Шаговая регрессия.

Шаговую регрессию можно рассматривать как промежуточный вариант между методами включения и исключения.

Задаются двумя табличными значениями *F*-распределения: F_{T1} для включения регрессора и F_{T2} для исключения регрессора. Метод начинает работать с включения одного регрессора, который выбирается из всех по наиболее подходящему частному коэффициенту корреляции. После каждого включения проверяют, нет ли среди ранее включенных какого-нибудь регрессора, теперь уже «ненужного». Это может случиться из-за коррелированности его с другими регрессорами.

Для проверки после включения данного члена в модель находят частный *F*-критерий как для метода исключения. Среди всех таких критериев выбирают наименьшее значение (F_{min}) и сравнивают его с F_{T2}. Если окажется, что $F_{min} < F_{T2}$, соот-

ветствующий регрессор исключается из модели, в противном случае модель остается без изменений. После этого переходят к новому включению, т. е. находят регрессор, имеющий максимальное значение частного коэффициента корреляции, вычисляют соответствующее ему значение частного F-критерия (F_{max}) и проверяют условие $F_{max} > F_{T1}$. Если оно выполнено, новый регрессор включают в модель, иначе структура модели не изменяется. Затем переходят к новому исключению. Процедура заканчивается, когда не удается более реализовать ни включения регрессора, ни исключения.

Для табличных значений F_{T1} и F_{T2} обычно выбирают одинаковые уровни значимости α. Чаще всего это $\alpha = 0{,}05$, но иногда и другие значения из интервала от 0,01 до 0,1. Бывает, что предпочитают взять уровень значимости для исключения больше, чем для включения, чтобы сохранить в модели больше включенных предикторов. Поступать наоборот нецелесообразно, поскольку тогда очень легко удалить важный предиктор из модели, и она может получиться неопределенной.

4.5 Вопросы для самоконтроля

1. Дайте определение мультиколлинеарности. Каковы последствия мультиколлинеарности при моделировании регрессии?

2. Перечислите способы выявления мультиколлинеарности. В чем их достоинства и недостатки?

3. Применение каких методов позволяет устранить или смягчить мультиколлинеарность?

4. Назовите причины возникновения гетероскедастичности. Какие тесты позволяют выявить гетероскедастичность?

5. Как проводится оценка методом взвешенных наименьших квадратов?

6. Охарактеризуйте причины возникновения и последствия автокорреляции остатков.

7. Какие процедуры позволяют выявить и устранить автокорреляцию?

8. Охарактеризуйте методы отбора переменных на этапе спецификации модели.

4.6 Тесты

1. Какое из следующих утверждений верно в случае гетероскедастичности остатков:

а) выводы по *t* и *F*- статистикам являются ненадежными;

б) гетероскедастичность проявляется через низкое значение статистики Дарбина-Уотсона;

в) при гетероскедастичности оценки остаются эффективными;

г) оценки параметров уравнения регрессии являются смещенными.

2. Как называется нарушение допущения о постоянстве дисперсии остатков?

а) мультиколлинеарность;

б) автокорреляция;

в) гетероскедастичность;

г) гомоскедастичность.

3. На чем основан тест Гольфельда-Квандта?

а) на использовании *t* – статистики;

б) на использовании *F* – статистики;

в) на использовании χ^2;

г) на графическом анализе остатков.

4. Если по *t*-критерию большинство коэффициентов регрессии статистически значимы, а модель в целом по *F*- критерию незначима то это может свидетельствовать

а) о наличии мультиколлинеарности;

б) об автокорреляции остатков;

в) о гетероскедастичности остатков;

г) такой вариант невозможен.

5. Автокорреляцией в статистике называется

а) зависимость вариации значений одного показателя от вариации значений другого;

б) зависимость между цепными уровнями;

в) отклонения от тенденции;

г) зависимость последующего уровня ряда от предыдущего.

6. Выбор списка переменных модели и типа взаимосвязи между ними выполняется на этапе

а) спецификации модели;

б) оценки параметров модели;

в) проведения статистического наблюдения;

г) проверки адекватности модели.

7. К основным ошибкам спецификации можно отнести:

а) добавление незначимой переменной;

б) удаление значимой переменной;

в) выбор неправильной формы модели;

г) низкое значение коэффициента детерминации.

8. Для выявления мультиколлинеарности применяется

а) тест Дарбина-Уотсона;

б) тест Бреуша-Годфри;

в) анализ матрицы парных коэффициентов корреляции;

г) показатель Хорла.

9. Обнаружить автокорреляцию можно с помощью теста

а) Дарбина-Уотсона;

б) Льюинга-Бокса;

в) Гольфельда-Квандта;

г) Уайта.

10. Гетерскедастичность выявляется с помощью

а) теста Льюинга-Бокса;

б) теста ранговой корреляции Спирмена;

в) Q - теста Бокса-Пирса;

г) теста Глейзера.

5 Нелинейные модели регрессии

Что необходимо знать из 5 главы:

1. Классы нелинейных регрессий и способы их оценивания.
2. Оценивание параметров регрессий, нелинейных по переменным.
3. Преобразование регрессионных моделей, нелинейных по оцениваемым параметрам.
4. Способы подбора линеаризующего преобразования.
5. Использование нелинейной регрессии в производственных функциях.

5.1. Понятие и способы оценивания нелинейной формы связи

При изучении взаимосвязи между социально-экономическими явлениями часть из них рассматривается в определенной области их существования, т.е. определяется регрессионное уравнение, которое пригодно не для всех возможных значений предикторов, а только для тех значений, которые заключены в некотором интервале. Если регрессионная зависимость определяется в сравнительно широкой области су-

ществования, то регрессии прироста (сокращения) факторного признака, как правило, соответствует неравномерный, непропорциональный прирост (сокращение) результативного признака и линейная форма связи неприменима. Если область вариации объясняющих переменных велика, это приводит к необходимости составления *нелинейных уравнений регрессии*. Встречаются и закономерности, форма связей которых значительной отличается от линейной даже в сравнительно узкой области вариации независимых переменных.

Выбор той или иной формы связи определяется следующими соображениями. Во-первых, избранный класс (или тип) уравнений регрессии должен отражать качественный характер экономических закономерностей, присущих изучаемым явлениям. Во-вторых, чтобы применить метод наименьших квадратов, необходимо использовать уравнения, которые по отношению к определяемым параметрам регрессии являются линейными или могут быть приведены к ним путем несложных преобразований. В-третьих, в уравнении регрессии следует ограничить количество определяемых параметров. Если их увеличить до числа единиц в совокупности, то линия регрессии на корреляционной диаграмме *пройдет через все точки*. Ясно, что в таком случае она отражает не основную закономерность связи, а случайные отступления от нее. Таким образом, для экономического анализа следует выбирать по возможности простые виды уравнений регрессии [13, с. 83].

Мы уже упоминали о том, что в эконометрике различают два основных класса нелинейных регрессий - регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам и регрессии, нелинейные по оцениваемым параметрам. Последние, в свою очередь, могут быть внутренне линейными и внутренне нелинейными.

Этап параметризации регрессионной модели, т.е. выбора параметрического класса функций $f(X,\beta)$ является одновременно наиболее важным и наименее формализованным и теоретически обоснованным этапом регрессионного анализа.

Если в результате реализации этого этапа исследователь пришел к выводу, что функция $f(X,\beta)$ нелинейная, то далее он может действовать следующим образом:

1. Попытаться подобрать такие преобразования к анализируемым переменным $y, x_1, x_2, ..., x_k$, которые позволили бы представить искомую зависимость в виде линейного соотношения между преобразованными переменными. Независимые переменные, имеющие степень, отличающуюся от первой, заменяются другими независимыми переменными в первой степени, и к новой системе переменных применяется обычный метод наименьших квадратов. После того, как получено уравнение с оцененными параметрами, введенные в него новые независимые переменные заменяются на первоначальные.

Другими словами, если $\varphi_0, \varphi_1, ...\varphi_p$ - те самые искомые функции, которые определяют переход к преобразованным переменным, т.е. $y^* = \varphi_0(y),\ x_1^* = \varphi_1(x_1),\ ...\ x_p^* = \varphi_p(x_p)$, то связь между y и $X = (x_1, x_2, ..., x_k)$ может быть представлена в виде линейной функции регрессии y^* от X^*, а именно:

$$y_i^* = \beta_0 + \beta_1 x_i^* + ... + \beta_p x_p^* + \varepsilon_i, \qquad i = 1, 2, ..., n. \tag{5.1}$$

Эту часть исследования обычно называют процедурой линеаризации модели.

Первым, кто предложил использовать критерий наименьших квадратов для оценивания линейных коэффициентов при подгонке кривой, был Лежандр (1805). Гаусс (1809) подвел статистическую базу под оценивание параметров, показав, что оценки наименьших квадратов максимизирует плотность нормального (Гауссова) закона распределения вероятности ошибок. Так Гаусс ввел метод максимума правдоподобия. Первые опыты приложения теории статистического оценивания к оценке модели были сделаны в области экономики Тьяллингом Чарльзом Купмансом (лауреатом Нобелевской премии по экономике 1975 года за вклад в теорию оптимального распределения ресурсов) и другими, начиная с 1930-го года. Их работа была опубликована в докладах Комиссии по охране окружающей среды [35].

Расчет оценок параметров нелинейных моделей требует обычно нахождения максимума или минимума нелинейной функции. Численные методы, носящие имена

Ньютона, Гаусса, Коши, известны уже очень давно, но их широкое применение для решения практических задач стало возможным лишь с появлением электронных компьютеров. Первую программу общего назначения для решения задач оценивания нелинейным методом наименьших квадратов создали Бут и Петерсон совместно с Боксом. В программе был реализован модифицированный метод Гаусса [36].

2. В ситуации, когда не представляется возможным линеаризация модели, искомую регрессионную зависимость исследуют в терминах исходных переменных, а именно: $y_i = f\left(X_i, \beta\right) + \varepsilon_i$.

Если спецификация регрессионных остатков ε_i соответствует условиям классической модели, то для вычисления МНК-оценок решается оптимизационная задача вида:

$$b_{МНК} = \arg \min_b \sum_{i=1}^{n} (y_i - f(X_i, \beta))^2 , \qquad (5.2)$$

т. е. используются итеративные методы нелинейной оптимизации на основе исходных переменных.

Измерение тесноты зависимости при любой форме связи осуществляется с помощью индекса корреляции или теоретического корреляционного отношения (при линейной зависимости теоретическое корреляционное отношение тождественно линейному коэффициенту корреляции).

На основании правила сложения дисперсий:

$$s_y^2 = s_x^2 + s_{y/x}^2 , \qquad (5.3)$$

где $s_y^2 = \dfrac{\sum (y_i - \bar{y})^2}{n}$ - общая вариация результативного признака,

учитывающая действие всех факторов;

$s_{y/x}^2 = \dfrac{\sum (y_i - \tilde{y}_i)^2}{n}$ - остаточная дисперсия;

$s_x^2 = s_y^2 - s_{y/x}^2$ - дисперсия, измеряющая вариацию признака *y*, возникающую

в результате вариации признака *x*.

Индекс корреляции рассчитывается по формуле:

$$\eta = \sqrt{\frac{s_y^2 - s_{y/x}^2}{s_y^2}} = \sqrt{1 - \frac{s_{y/x}^2}{s_y^2}} = 1 - \frac{\sum_{i=1}^{n}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (5.4)$$

Величина данного показателя находится в границах: $0 < \eta < 1$. Чем ближе значение корреляционного отношения к единице, тем теснее связь рассматриваемых признаков, тем более надежно найденное уравнение регрессии. Поскольку в расчете индекса корреляции используется соотношение факторной и общей суммы квадратов отклонений, то η^2 имеет тот же смысл, что и коэффициент детерминации (индекс детерминации).

Величина отклонений фактических и расчетных значений результативного признака ($y_i - \hat{y}_i$) по каждому наблюдению представляет собой ошибку аппроксимации. Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации как среднюю арифметическую простую.

5.2 Линеаризация уравнений регрессии

Рассмотрим оценивание параметров регрессий, нелинейных по переменным.

Полиномиальные модели второго порядка используются для характеристики процессов с монотонным развитием и отсутствием пределов роста. Данному условию отвечают, например, натуральные показатели промышленного производства. Зависимость параболического типа имеет вид:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon_i. \quad (5.5)$$

Заменяя переменные $x_1^* = x;\ \ x_2^* = x^2;$ получим двухфакторное уравнение линейной регрессии, для оценки параметров которого используется МНК:

$$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2^*. \tag{5.6}$$

Применение МНК для оценки параметров полинома второй степени приводит к следующей системе нормальных уравнений:

$$\begin{cases} nb_0 + b_1 \sum_{i=1}^{n} x_i + b_2 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i; \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 + b_2 \sum_{i=1}^{n} x_i^3 = \sum_{i=1}^{n} y_i x_i; \\ b_0 \sum_{i=1}^{n} x_i^2 + b_1 \sum_{i=1}^{n} x_i^3 + b_2 \sum_{i=1}^{n} x_i^4 = \sum_{i=1}^{n} y_i x_i^2. \end{cases} \tag{5.7}$$

Решение системы (5.7) возможно методом определителей:

$$b_0 = \frac{\Delta b}{\Delta}; b_1 = \frac{\Delta b_1}{\Delta};\ \ b_2 = \frac{\Delta b_2}{\Delta}, \tag{5.8}$$

где Δ - определитель системы;

$\Delta b_0;\ \ \Delta b_1;\ \ \Delta b_2$ - частные определители для каждого из параметров.

При $\beta_1 > 0$ и $\beta_2 < 0$ кривая симметрична относительно высшей точки, т.е. точки перелома кривой, изменяющей направление связи, а именно рост на падение.

При $\beta_1 < 0$ и $\beta_2 > 0$ кривая симметрична относительно своей низшей точки, что позволяет определять минимум функции в точке, меняющей направление связи, т. е. снижение на рост (рисунок 5.1).

В качестве примера параболической зависимости можно привести связь между производительностью труда работников (процент выполнения нормы выработки) и их возрастом (таблица 5.1).

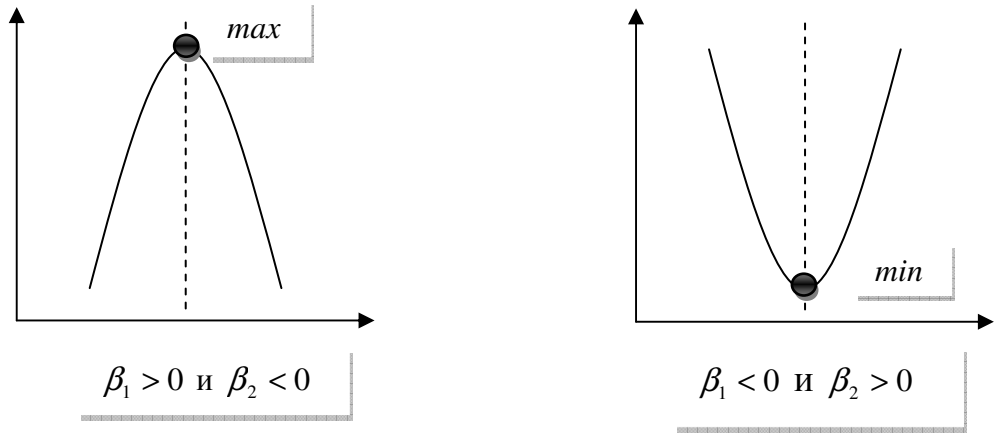


Рисунок 5.1 – Свойства параболы второго порядка

Таблица 5.1 – Данные по 20-ти работникам

№	Y	X	№	Y	X	№	Y	X	№	Y	X
1	84	19	6	89	21	11	110	47	16	115	27
2	92	23	7	113	35	12	102	49	17	105	45
3	80	21	8	118	31	13	108	48	18	116	43
4	85	23	9	111	25	14	112	46	19	108	40
5	94	25	10	102	25	15	113	28	20	122	35

Представим зависимость графически (рисунок 5.2).

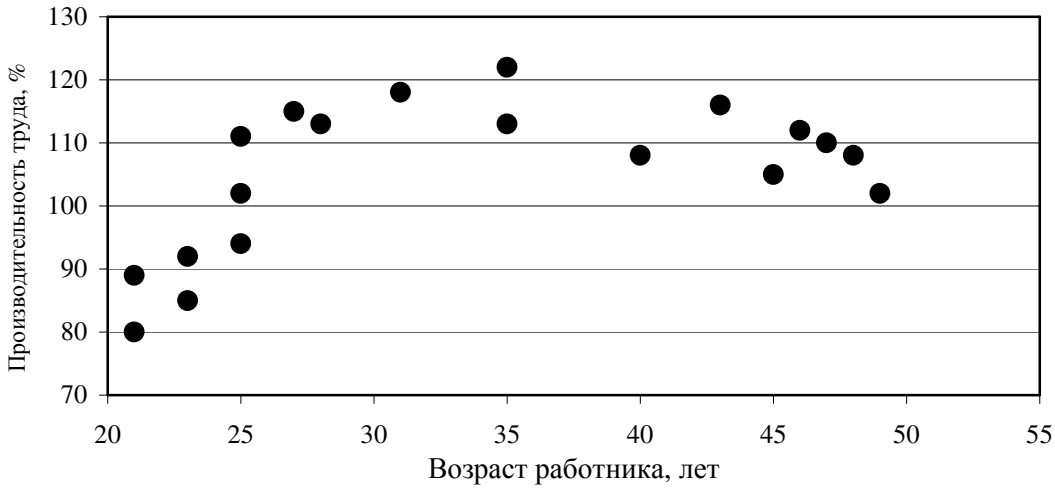


Рисунок 5.2 – Диаграмма рассеяния факторного и результативного признаков

Связь между переменными опишем в виде параболы второго порядка (используем ППП Statistica). Получим следующую оценку уравнения регрессии (рисунок 5.3).

	Level of confidence: 95.0% (alpha=0.050)					
	Estimate	Standard error	t-value df = 17	p-level	Lo. Conf Limit	Up. Conf Limit
b	-46,5106	21,88363	-2,12536	0,048513	-92,6811	-0,34022
b1	8,8879	1,37374	6,46984	0,000006	5,9895	11,78619
b2	-0,1195	0,01987	-6,01454	0,000014	-0,1615	-0,07760

Рисунок 5.3 – Результаты оценивания параболы второго порядка

Полученное уравнение зависимости со значимыми параметрами имеет вид:

$$\tilde{y} = -45{,}5106 + 8{,}8879x - 0{,}1195x^2.$$

Недостаток этого типа связей состоит в том, что кривая параболы по обе стороны от экстремума симметрична. В экономике таких связей, когда результативный признак убывает равномерно по мере увеличения отклонений объясняющей переменной в ту и другую сторону от экстремума, почти нет. Чаще исследователь имеет дело лишь с отдельными сегментами параболы, а не с полной параболической формой. Кроме того, параметры параболической связи не всегда могут быть логически истолкованы.

Если модель второго порядка не адекватна, то, может быть, подойдет модель третьего порядка. Однако вряд ли стоит механически добавлять в модель члены более высоких порядков. Часто оказывается продуктивным исследование возможностей каких-то иных преобразований предикторов, откликов или тех и других одновременно. То же замечание относится и к решению о переходе от первого порядка ко второму. Так, например, прямая, подобранная в координатах *log Y* от *X*, если она возможна, нередко предпочтительнее, чем квадратичная модель зависимости *Y* от *X*, если, конечно, поведение остатков делает оба эти выбора работоспособными [37, с. 276-277]. Поэтому, если график зависимости не отражает четко выраженного поли-

нома второго порядка (нет смены направленности связи признаков), то модель может быть заменена другой нелинейной функцией.

Зависимость гиперболического типа:

а) среди класса нелинейных функций, параметры которых оцениваются МНК, следует назвать широко распространенную в эконометрике равностороннюю гиперболу:

$$y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon \quad (0 < x < \infty). \tag{5.9}$$

Соответствующая кривая регрессии характеризуется двумя асимптотами (т.е. прямыми, к которым график функции неограниченно приближается, не достигая их) – горизонтальной $y = \beta_0$ и вертикальной $x = 0$ (рисунок 5.4).

С помощью преобразования объясняющей переменной $x^* = \frac{1}{x}$ эта зависимость приводится к линейному виду:

$$y = \beta_0 + \beta_1 x^* + \varepsilon. \tag{5.10}$$

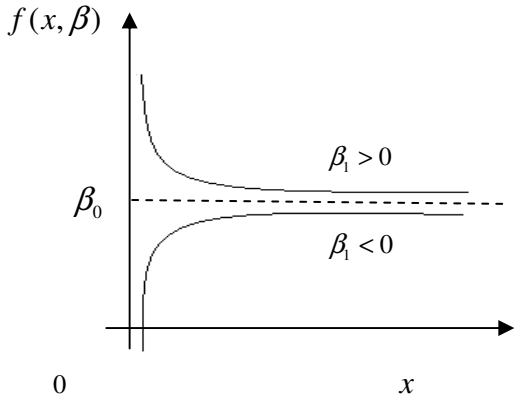


Рисунок 5.4 – Кривые регрессии равносторонней гиперболы

При вычислении МНК-оценок матрица X будет иметь вид:

$$X^* = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ \frac{1}{x_1} & \frac{1}{x_2} & \ldots. & \frac{1}{x_n} \end{pmatrix}^T. \tag{5.11}$$

Классическим примером равносторонней гиперболы является кривая английского экономиста А.В. Филлипса[1], которая отражает зависимость между уровнем безработицы и процентом прироста заработной платы. Анализируя данные более чем за 100-летний период, в конце 50-х годов XX в. Филлипс установил обратную зависимость между процентным приростом денежной заработной платы и уровнем безработицы. Следует отметить, что еще до Филлипса в 1920—30-е гг. подобную зависимость описывали И. Фишер, Я. Тинберген и Дж. Данлоп. Тем не менее, именно кривая Филлипса была использована кейнсианцами для обоснования своей теории инфляции и безработицы.

Показатель номинальной заработной платы был заменен на показатель уровня инфляции, в результате чего кривая демонстрировала, как с помощью инфляции возможно регулирование занятости и, следовательно, уровня производства (рисунок 5.5).

Рисунок 5.5 – Зависимость между уровнем безработицы и темпами инфляции (кривая Филлипса для США, данные для 60-х годов)

[1] Отношение современных экономистов к «кривой Филлипса» довольно противоречивое. Практически никто не отрицает ее существование, но ведутся споры о диапазоне действия выявленных Филлипсом зависимостей. Различные интерпретации кривой Филлипса давали Самуэльсон, Солоу, Фридмен. Многие оценивают Элбана Уильяма (Билла) Филлипса как человека, обогнавшего свое время, но не получившего в руки необходимых вычислительных мощностей для численной реализации своих эконометрических моделей. Так, его модель экономики Великобритании, составленная в 1961 году, была реализована на компьютере только к середине 70-х годов.

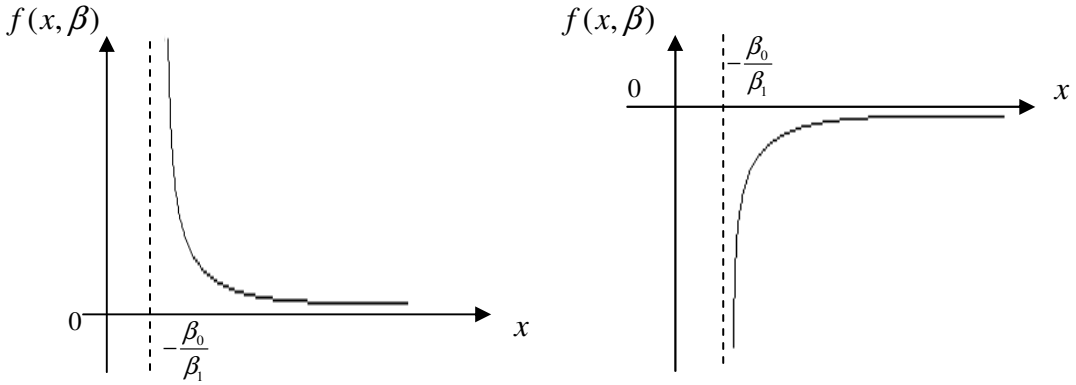
Каждая гиперболическая функция в каждой точке своего определения непрерывна и бесконечно дифференцируема.

При $\beta_1 > 0$ получим обратную зависимость, которая при $x \to \infty$ характеризуется нижней асимптотой (прямая, к которой график функции неограниченно приближается, но не пересекает ее), т.е. минимальным предельным значением y, оценкой которого служит параметр b_0.

При $\beta_1 < 0$ получим медленно возрастающую функцию с верхней асимптотой при $x \to \infty$, т. е. с максимальным предельным уровнем y, оценку которого в уравнении дает параметр b_0;

б) гиперболическая зависимость вида

$$y = \frac{1}{\beta_0 + \beta_1 x + \varepsilon} \quad (-\frac{\beta_0}{\beta_1} < x < \infty). \tag{5.12}$$



а) $\beta_0 < 0, \ \beta_1 > 0$ *б*) $\beta_0 > 0, \ \beta_1 < 0$

Рисунок 5.6 – Вид гиперболической зависимости $y = \dfrac{1}{\beta_0 + \beta_1 x + \varepsilon}$ при различных знаках параметров

С помощью преобразования результативной переменной $y^* = \dfrac{1}{y}$ эта зависимость приводится к линейному виду:

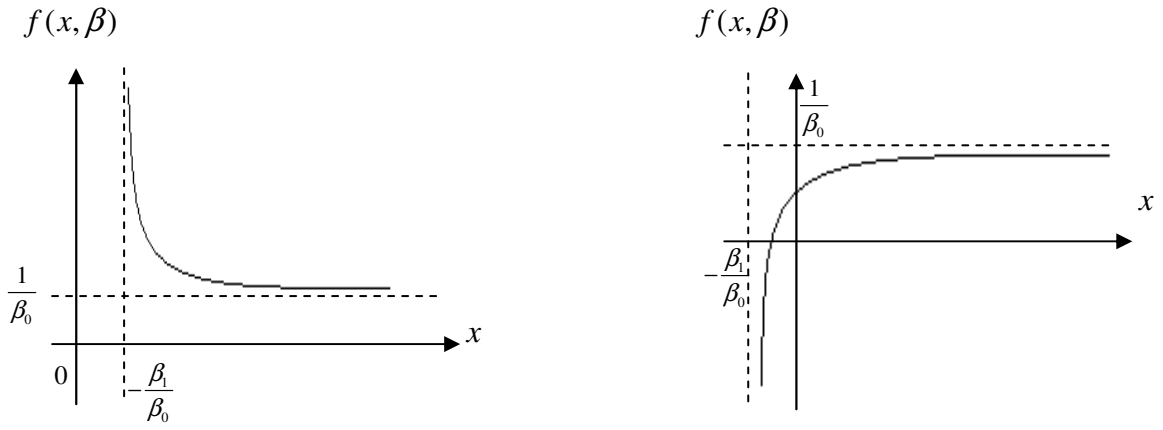
$$y^* = \beta_0 + \beta_1 x + \varepsilon\,. \qquad (5.13)$$

При вычислении МНК-оценок

$$Y^* = \left(\frac{1}{y_1} \quad \frac{1}{y_2} \quad \ldots \quad \frac{1}{y_n} \right)^T. \qquad (5.14)$$

в) гиперболическая зависимость вида:

$$y = \frac{x}{\beta_0 x + \beta_1 + x\varepsilon} \quad \left(-\frac{\beta_1}{\beta_0} < x < \infty\right). \qquad (5.15)$$



а) $\beta_0 > 0,\ \beta_1 < 0$ *б*) $\beta_0 > 0,\ \beta_1 > 0$

Рисунок 5.7 – Вид гиперболической зависимости $y = \dfrac{x}{\beta_0 + \beta_1 x + \varepsilon}$ при различных знаках параметров

Матрицы X^*, y^*, используемые в формулах МНК, должны формироваться не из наблюденных значений x_i, y_i, а из обратных к ним величин $x^* = \dfrac{1}{x},\ y^* = \dfrac{1}{y}$.

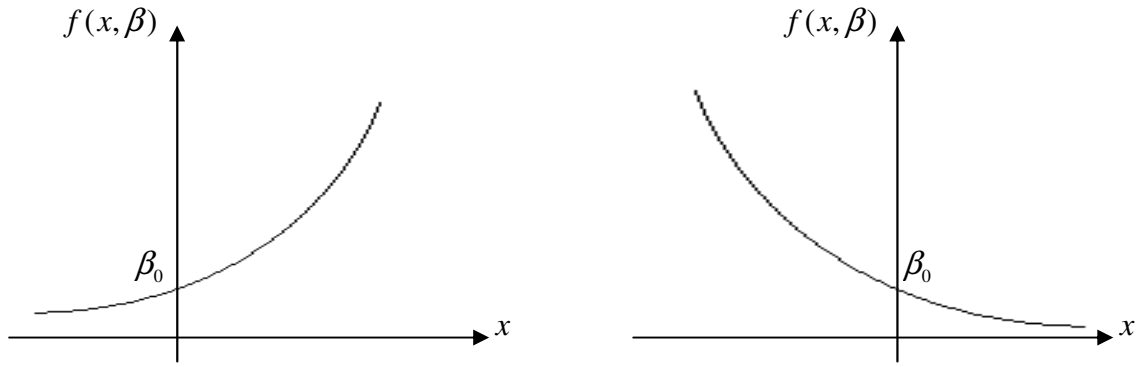
5.3 Регрессионные модели, нелинейные по оцениваемым параметрам

Иначе обстоит дело с регрессией, нелинейной по оцениваемым параметрам. Если нелинейная модель внутренне линейна, то она с помощью соответствующих преобразований может быть приведена к линейному виду. Если же нелинейная модель внутренне нелинейна (т.е. действительно нелинейна, периодична), то она не может быть сведена к линейной функции.

Показательная (экспоненциальная) зависимость. Достаточно широкий класс экономических показателей характеризуется приблизительно постоянным темпом относительного прироста во времени. Этому соответствует следующая форма зависимости:

а)

$$y = \beta_0 e^{\beta_1 x + \varepsilon}. \qquad (5.16)$$



а) $\beta_1 > 0$ *б*) $\beta_1 < 0$

Рисунок 5.8 – Экспоненциальная зависимость вида $y = \beta_0 e^{\beta_1 x + \varepsilon}$ при различных знаках параметров

Переход к новой переменной $y^* = \ln y$ позволяет свести исследуемому зависимость к линейному виду:

$$y^* = \beta_0^* + \beta_1 x + \varepsilon, \text{ где } \beta_0^* = \ln \beta_0 . \quad (5.18)$$

Для МНК-оценок используют

$$Y^* = \begin{pmatrix} \ln y_1 & \ln y_2 & \dots & \ln y_n \end{pmatrix}^T . \quad (5.19)$$

В качестве примера экспоненциальной регрессии можно привести зависимость слухового порога (при 4000 Гц в дБ) от возраста человека (рисунок 5.10).

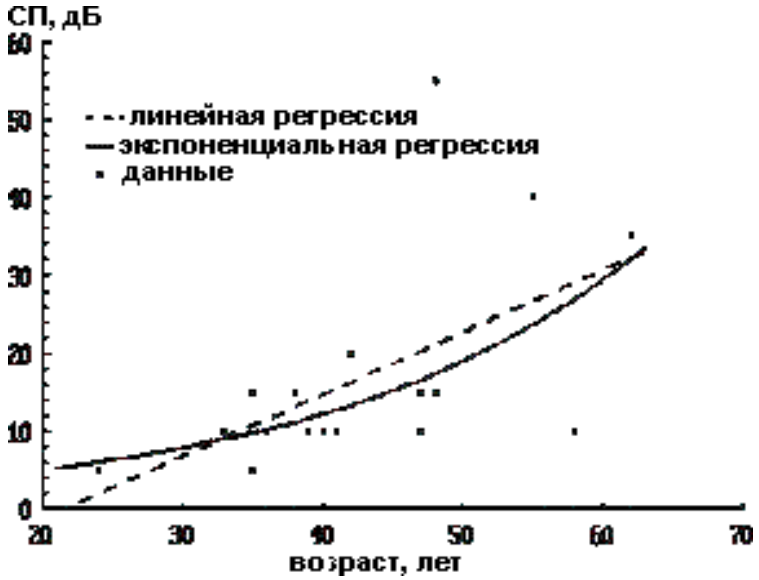


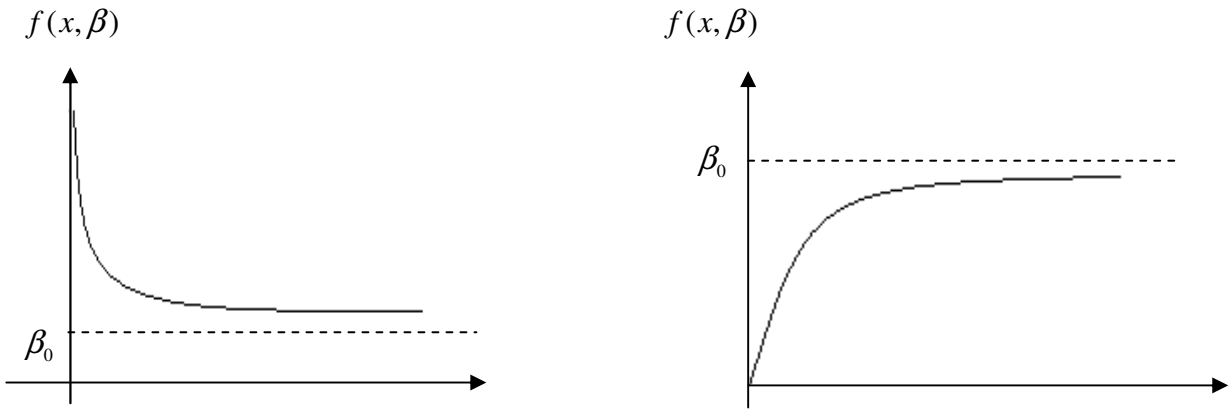
Рисунок 5.9 – Зависимость слухового порога от возраста человека [38]

С помощью функции (5.10) описываются процесс радиоактивного распада, затухающие колебания и т.п.

Экспонента также описывает содержание радиоактивного углерода-14 в зависимости от возраста органического объекта при значении коэффициента детерминации, близком к единице, что означает практически полное совпадение кривой с аппроксимируемыми данными.

б)

$$y = \beta_0 e^{\frac{\beta_1}{x} + \varepsilon} . \quad (5.20)$$



а) $\beta_1 > 0$ *б*) $\beta_1 < 0$

Рисунок 5.10 – Экспоненциальная зависимость вида $y = \beta_0 e^{\frac{\beta_1}{x}+\varepsilon}$ при различных знаках параметров

Линеаризация искомой зависимости достигается с помощью следующих преобразований переменных: $y^* = \ln y$, $x^* = \frac{1}{x}$, где $\beta_0^* = \ln \beta_0$.

Соответственно вектор-столбец

$$Y^* = \begin{pmatrix} \ln y_1 & \ln y_2 & \dots & \ln y_n \end{pmatrix}^T \tag{5.21}$$

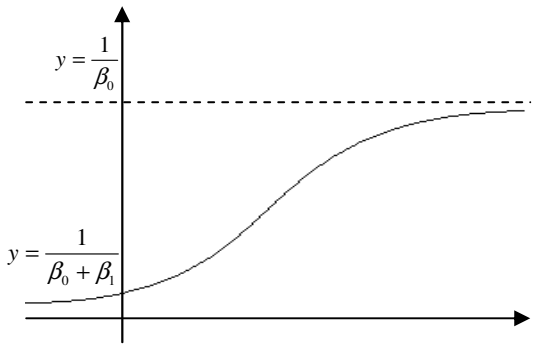
и матрица

$$X^* = \begin{pmatrix} 1 & 1 & \dots & 1 \\ \frac{1}{x_1} & \frac{1}{x_2} & \dots & \frac{1}{x_n} \end{pmatrix}^T. \tag{5.22}$$

в) возможно и одновременное использование логарифмирования, и преобразование в обратные величины. Примером подобной функции является *S*-образная (логистическая, сигмоидальная) кривая жизненного цикла:

$$y = \frac{1}{\beta_0 + \beta_1 e^{-x} + \varepsilon}, \quad 0 \le x < \infty. \tag{5.23}$$

Впервые такая кривая (рисунок 5.11) была применена А. Кетле для расчета численности населения. Она моделирует кривую роста вероятности некоего события, по мере изменения управляющих параметров (факторов риска).



$\beta_1 > 0$

Рисунок 5.11 – Логистическая кривая

Подобного типа функции используются при анализе статистических данных о бюджетах потребителей, где выдвигается гипотеза о существовании асимптотического уровня расходов, об изменении предельной склонности к потреблению товара, о существовании «порогового уровня дохода».

Кривая $f(x,\beta)$ имеет две горизонтальные асимптоты $y = 0$ *и* $y = \frac{1}{\beta_0}$ и точку перегиба ($x_0 = \ln(\frac{\beta_1}{\beta_0}), \quad y_0 = \frac{1}{2\beta_0}$).

Линеаризация этой зависимости производится с помощью перехода к переменным $y^* = \frac{1}{y}$, $x^* = e^{-x}$.

Соответственно вектор-столбец и матрица, участвующие в формулах МНК определяются следующим образом:

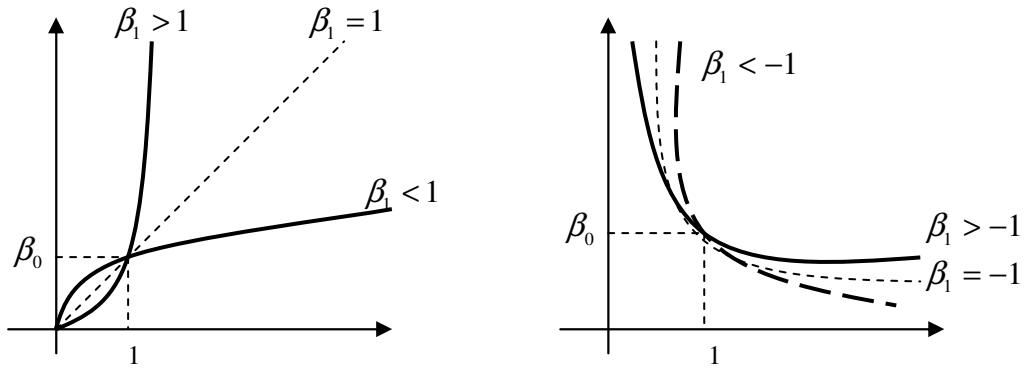
$$Y^* = \begin{pmatrix} \frac{1}{y_1} & \frac{1}{y_2} & \dots & \frac{1}{y_n} \end{pmatrix}^T \qquad X^* = \begin{pmatrix} 1 & 1 & \dots & 1 \\ e^{-x_1} & e^{-x_2} & .... & e^{-x_n} \end{pmatrix}^T . \qquad (5.24)$$

S-образная кривая с высокой вероятностью описывает развитие различных систем (зависимость показателей системы от вкладываемых в нее затрат).

Логистическую кривую используют при характеристике развития потенциала организации и ее положения во внешней среде: описания жизненных циклов спроса, технологии, товара и даже самой организации.

Зависимость степенного типа:

$$y = \beta_0 (x)^{\beta_1} . \qquad (5.25)$$



а) $\beta_1 > 0$ *б*) $\beta_1 < 0$

Рисунок 5.12 – Степенная зависимость при различных знаках параметров

Степенная модель нелинейная относительно оцениваемых параметров, т.к. включает параметры β_0 и β_1 мультипликативно. Однако ее можно считать внутренне линейной, т.к. логарифмирование данного уравнения по основанию *е* приводит его к линейному виду: $y^* = \ln y$, $x^* = \ln x$, где $\beta_0^* = \ln \beta_0$.

Важную роль зависимости степенного типа играют в задачах построения и анализа производственных функций, функций спроса.

Распространенность степенной функции в экономических исследованиях связана с тем, что параметр β_1 имеет четкое экономическое истолкование, т.е. он является коэффициентом эластичности**.** Это значит, что величина коэффициента β_1 по-

казывает, на сколько процентов изменится в среднем результат, если фактор изменится на 1 %.

Замечание: функции вида $y = \beta_0 + \beta_1 \frac{1}{x} + \varepsilon$, изображенные на рисунке 5.5 (случай $\beta_1 < 0$), вида $y = \frac{x}{\beta_0 + \beta_1 x + \varepsilon}$ (рисунок 5.9 б) и степенная функция $y = \beta_0 (x)^{\beta_1}$ считаются приемлемыми для описания кривых Энгеля, характеризующих соотношение между спросом на определенный товар и общей суммой дохода.

В XIX в. немецкий экономист и статистик Эрнст Энгель на основе данных о расходах семей с разным уровнем дохода установил, что с ростом дохода доля его, направляемая на продовольствие, снижается, так как продукты питания относятся к необходимым товарам (*necessary good*); доля, направляемая на жилье и связанные с ним расходы, а также на одежду, остается примерно неизменной[1].

Доля других расходов возрастает, но это увеличение не беспредельно, т.к. на все товары сумма долей не может быть больше единицы, или 100 %. На отдельные непродовольственные товары этот предел может характеризоваться величиной параметра β_0 (y - доля расходов на непродовольственные товары; x - доходы (или общая сумма расходов как индикатор дохода)).

Эти зависимости в микроэкономике получили статус закона Энгеля (*Engel's law*). Функция (5.25) может также применяться к кривым спроса, где y – это спрос на товар, x – цена товара, а β_1 - это эластичность спроса по цене. (На практике обычно такая функция объединяется с кривой Энгеля, в результате чего получается зависимость спроса одновременно от дохода и цены) [6, с. 115-116, 121] (рисунок 5.13).

[1] С. Г. Струмилин (Струмилло-Петрашкевич) (видный советский экономист, статистик, историк, социолог; под его руководством разработана первая в мире система материальных балансов; автор одного из методов построения индекса производительности труда, т.н. индекса Струмилина; один из авторов планов индустриализации СССР) на материалах пензенских бюджетов пришел к выводу, что доля расходов на питание находится в более тесной связи с размером семьи и возрастом ее членов.

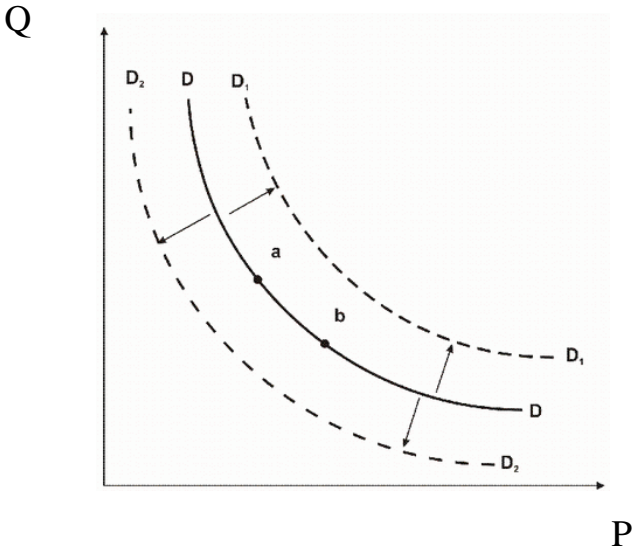


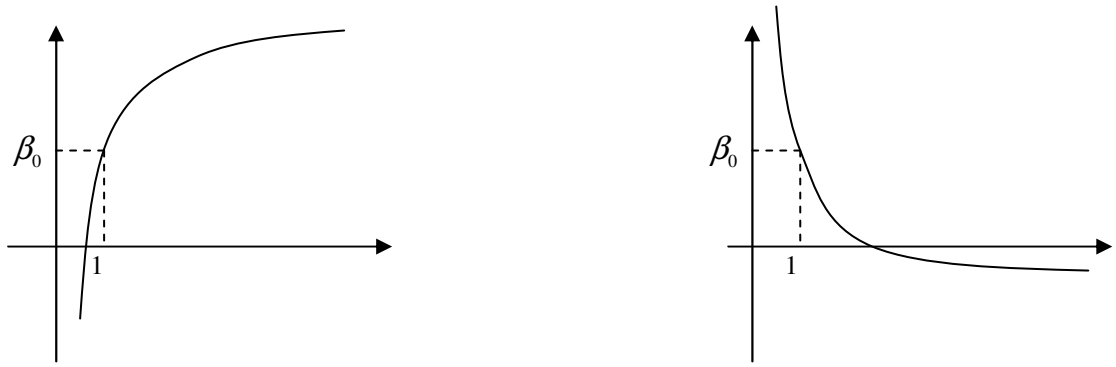
Рисунок 5.13 - Кривые спроса при разных состояниях дохода (D, D_1, D_2)

Зависимость логарифмического типа. Издавна обращено внимание исследователей на логарифмическую спираль как на модель эволюционирования сложных систем. Так, великий И.-В. Гёте считал ее символом жизни и духовного развития, математическим выражением соотношения формы и роста. Уже в наше время отечественный физик-теоретик А.Д. Панов и австралийский исследователь глобальной истории Г.Д. Снукс, причем независимо друг от друга, пришли к выводу, что эволюционные процессы в неорганической, биологической и социальной истории описываются единой логарифмической функцией. Логарифмические функции положены в основу собственно науковедческих изысканий [39].

Если показательная функция описывает изменение степени в зависимости от изменения ее показателя, то логарифмическая функция, наоборот, описывает изменение показателя степени в зависимости от изменения степени, поэтому логарифмическая функция является обратной к показательной:

$$y = \beta_0 + \beta_1 \ln x + \varepsilon \qquad (5.26)$$

Кривые на рисунке 5.14 проходят через точку $(1, \beta_0)$ и имеют в качестве вертикальной асимптоты ось y $(x = 0)$.



а) $\beta_1 > 0$ *б*) $\beta_1 < 0$

Рисунок 5.14 – Логарифмическая зависимость при различных знаках параметров

Переход к линейному виду зависимости осуществляется с помощью логарифмического преобразования объясняющей переменной $x^* = \ln x$.

В заключение остановимся еще раз на основных правилах выбора формы зависимости. В первую очередь необходимо исходить из экономической теории. После чего нужно оценить формальное качество полученной модели, а затем провести дополнительную проверку по нескольким содержательным критериям.

5.4 Подбор линеаризующего преобразования. Производственные функции

Выше мы рассмотрели набор зависимостей, которые поддаются линеаризации с помощью подходящих преобразований анализируемых переменных. Но решение вопроса о том, к какому именно из перечисленных линеаризуемых типов зависимостей следует отнести конкретный случай, является задачей не простой. Часто несколько разных нелинейных функций приблизительно соответствуют наблюдениям, если они лежат на некоторой кривой. Однако в случае множественного регрессионного анализа невозможно даже построить график. При рассмотрении альтернативных моделей с одним и тем же определением зависимой переменной процедура вы-

бора достаточно проста. Оценивают регрессию на основе всех вероятных функций и выбирают функцию, в наибольшей степени объясняющую изменения зависимой переменной.

В случае, когда разные модели используют разные функциональные формы, проблема выбора становится более сложной. Нельзя непосредственно сравнивать коэффициенты детерминации R^2 или суммы квадратов отклонений (например, нельзя сравнивать эти статистики для линейного и логарифмического вариантов модели регрессии). Если для одной модели R^2 значительно больше, чем для другой, то выбор такой модели оправдан, если же значения R^2 для двух моделей приблизительно равны, то проблема выбора значительно усложняется.

Если, например, стоит задача только сравнить модели с использованием y и $\log y$ в качестве зависимой переменной, то можно использовать преобразование Пола Зарембки [40]. Тест Зарембки предполагает такое преобразование масштаба наблюдений y, при котором обеспечивалась бы возможность непосредственного сравнения среднего квадратического отклонения в линейной и логарифмической моделях. Алгоритм теста включает следующие этапы:

1. Вычисляем среднее геометрического значений зависимой переменной по выборке и все ее значения делим на это среднее:

$$Y_i^* = Y_i / \sqrt[n]{Y_1 Y_2 ... Y_n} = Y_i / e^{\frac{1}{n}\left(\ln Y_1 + \ln Y_{2+...+\ln Y_n}\right)}. \qquad (5.27)$$

2. Оцениваются регрессии для линейной модели с использованием Y_i^* в качестве результативной переменной и для логарифмической модели с использованием $\ln Y_i^*$. Во всех других отношениях модели должны оставаться неизменными. Так как теперь значения среднего квадратического отклонения сравнимы, то модель с наименьшим значением SSR обеспечивает лучшее соответствие.

3. Для того чтобы проверить, не обеспечивает ли одна из моделей значимо лучшую аппроксимацию, вычисляют статистику χ^2 вида:

$$\chi^2 = \frac{n}{2} \times \left| \ln \frac{SSR1}{SSR2} \right| , \qquad (5.28)$$

которая сравнивается с критическим значением χ^2 - распределения с одной степенью свободы. Если расчетное значение статистики χ^2 превышает критическое при выбранном уровне значимости, то делается вывод о наличии значимых различий в качестве оценивания.

Английские статистики Г. Бокс и Д. Кокс [41] предложили более формализованную процедуру подбора линеаризующего преобразования. Их метод основан на предположении, что искомое преобразование принадлежит определенному однопараметрическому семейству преобразований вида:

$$y_i^*(\lambda) = \frac{y_i^{\lambda} - 1}{\lambda}, \quad x_i^*(\lambda) = \frac{x_{ij}^{\lambda} - 1}{\lambda}, \quad i = 1,2,...,n . \qquad (5.29)$$

Гипотезу можно сформулировать следующим образом: существует такое положительное или отрицательное число λ^*, что искомая регрессионная зависимость (5.30) или (5.31) будет удовлетворять всем требованиям КЛММР:

$$y_i^*(\lambda^*) = \beta_0 + \beta_1 x_{i1}(\lambda^*) + ... + \beta_k x_{ik}(\lambda^*) + \varepsilon_i , \qquad (5.30)$$

$$y_i^*(\lambda^*) = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i . \qquad (5.31)$$

Замечание 1: преобразования вида (5.29) применяются обычно к переменным, принимающим только положительные значения. В противном случае, вначале подбирают «сдвиговые» константы $c_0, c_1, ..., c_k$, которые обеспечивают положительность значений $y_i + c_0$ и $x_{ij} + c_j$ ($j = 1,2,...,k$), а затем к сдвинутым значениям переменных применяют данное преобразование, т.е.:

$$y_i^*(\lambda) = \frac{(y_i + c_0)^\lambda - 1}{\lambda},\ x_i^*(\lambda) = \frac{(x_{ij} + c_j)^\lambda - 1}{\lambda},\ i = 1,2,...,n. \qquad (5.32)$$

Замечание 2: семейство степенных преобразований вида (5.29 и 5.32) весьма широко. При $\lambda = 1$ модели (5.30) и (5.31) являются линейными относительно y_i и $x_{i1}, x_{i2}, ..., x_{ik}$. При $\lambda = 0$ имеем степенную зависимость между Y и X, поскольку $y_i^*(0) = \lim_{\lambda \to 0}(y_i^\lambda - 1)/\lambda = \ln y_i$ и $y_{ij}^*(0) = \lim_{\lambda \to 0}(x_{ij}^\lambda - 1)/\lambda = \ln x_{ij}$. При других значениях λ уравнения будут связывать между собой какие-то степени исходных переменных. Следовательно, подбор линеаризующего преобразования анализируемых переменных сводится к оценке параметра λ по имеющимся в нашем распоряжении исходным статистическим данным. Эта проблема решается с помощью метода максимального правдоподобия, который заключается в максимизации функции правдоподобия: $L(y_1^9, .., y_n^*, X, \lambda, \beta, \sigma^2) \to \max$.

Важную роль играют зависимости степенного типа в задачах построения и анализа производственных функций (ПФ).

Производственная функция – это экономико-математическая модель, позволяющая аппроксимировать зависимость результатов производственной деятельности фирмы, вида экономической деятельности или национальной экономики в целом от повлиявших на эти результаты факторов.

В основе понятия ПФ лежит представление об изучаемом экономическом объекте как об открытой динамической системе, выходом которой является производимая продукция, а входом - затраты различных видов производственных ресурсов. В качестве факторов производственных функций могут выступать следующие переменные: объем выпущенной продукции; объем основного капитала или основных фондов; объем трудовых ресурсов или трудовых затрат.

Простой разновидностью ПФ являются *однофакторные производственные функции* (ОПФ). Зависимой переменной в данных функциях является объем производства y, который зависит от единственной независимой переменной x - ресурсы.

Возможные способы использования ПФ:

1) определение объема выпуска при фиксированных заранее значениях показателей основных ресурсов;

2) определение влияния на объем выпуска изменения размеров одного или нескольких ресурсов;

3) определение характеристик производственного процесса, выражающихся через параметры ПФ.

Двухфакторные производственные функции (ДПФ) характеризуют зависимость объема производства от каких-либо факторов. Чаще всего это факторы объема основного капитала и трудовых ресурсов. К наиболее известным двухфакторным ПФ относятся функции Кобба-Дугласа[1]:

$$Y = A \cdot K^{\alpha} \cdot L^{\beta} \cdot e^{\varepsilon}, \quad 0 < \alpha < 1 \quad 0 < \beta < 1 \quad \alpha + \beta \approx 1, \tag{5.33}$$

где Y - объем выпуска продукции;

K - объем основного капитала;

L - затраты живого труда.

При использовании для построения производственной функции пространственной информации, т.е. данных нескольких фирм, относящихся к одному и тому же времени, предполагается, что поведение всех фирм может быть описано с помощью одной и той же функции. Для успешной экономической интерпретации полученной модели желательно, чтобы все эти фирмы принадлежали одной и той же отрасли. Кроме того, предполагается, что они располагают примерно одинаковыми производственными возможностями и уровнем административного управления. В каждое уравнение, параметры которого предстоит оценить, необходимо ввести еще случайную переменную ε, которая будет отражать воздействие на процесс производства

[1] Впервые модель была предложена Кнутом Уикселлом. В 1928 г. функция проверена на статистических данных Чарльзом Коббом и Полом Дугласом в работе «Теория производства». В этой статье была предпринята попытка эмпирическим путем определить влияние затрачиваемого капитала и труда на объем выпускаемой продукции в обрабатывающей промышленности США.

всех факторов, которые не вошли в состав производственной функции в явном виде [42, с. 39-40].

Мультипликативная модель (5.33) сводится к линейной путем логарифмирования обеих частей:

$$\ln Y = \ln A + \varepsilon \ln K + \beta \ln L + \varepsilon . \qquad (5.34)$$

Сумма коэффициентов α и β является *отдачей от масштаба.*

Если сумма показателей степени в ПФ Кобба-Дугласа равна 1, то ее можно записать в следующей форме:

$$\frac{Y}{L} = A \cdot \left(\frac{K}{L} \right)^{\alpha} \cdot e^{\varepsilon} . \qquad (5.35)$$

Получена зависимость производительности труда $\frac{Y}{L}$ от его капиталовооруженности $\frac{K}{L}$.

Функция Кобба-Дугласа с учетом технического прогресса имеет вид:

$$Y = A \cdot K^{\alpha} \cdot L^{\beta} \cdot e^{pt} , \qquad (5.36)$$

где t - время;

p - темп прироста объема производства благодаря техническому прогрессу.

Рассмотрим свойства двухфакторной производственной функции $f(x_1, x_2)$:

1) $f(0,0) = 0$ - без ресурсов нет выпуска;

2) $f(0, x_2) = f(x_1, 0) = 0$ - при отсутствии хотя бы одного из ресурсов нет выпуска;

3) $x_1 \geq x_2 \quad f(x_1) > f(x_2)$ - с ростом затрат хотя бы одного ресурса объем выпуска растет;

4) $x > 0 \Rightarrow \frac{\partial f(x)}{\partial x_i} > 0$ (первая частная производная положительная) – с ростом затрат одного ресурса при неизменном количестве другого ресурса объем выпуска растет;

5) $x > 0 \Rightarrow \frac{\partial^2 f(x)}{\partial x_i^2} \le 0$ (вторая частная производная не положительная) – с ростом затрат одного $i-го$ ресурса при неизменном количестве другого ресурса величина прироста выпуска на каждую дополнительную единицу $i-го$ ресурса не растет (закон убывающей эффективности);

6) $x > 0 \Rightarrow \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \ge 0$ - при росте одного ресурса предельная эффективность другого ресурса возрастает.

Еще один пример двухфакторной производственной функции - производственная функция Леонтьева. Это функция с фиксированными пропорциями факторов, предназначенная для моделирования строго детерминированных технологий, не допускающих отклонения от технологических норм использования ресурсов на единицу продукции (рисунок 5.15).

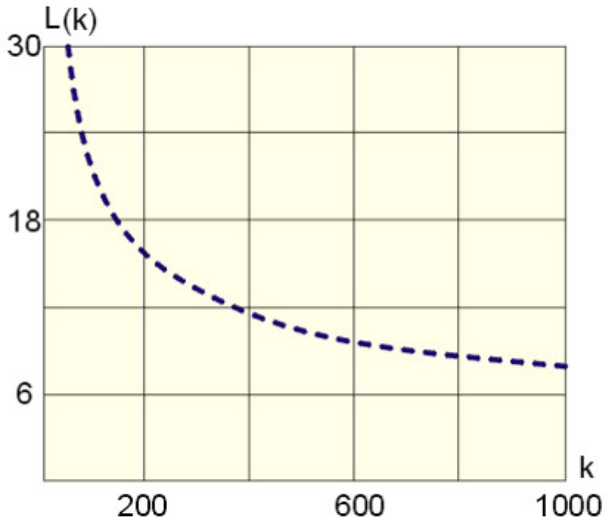


Рисунок 5.15 - Функция с фиксированными пропорциями факторов (Y = min (L/a1, K/a2) [43]

Производственная функция Леонтьева обычно используется для описания мелкомасштабных или полностью автоматизированных производственных объектов. Такая модель отражает тот факт, что ресурсы не могут заменять друг друга. Если один из ресурсов ограничен, можно рассчитать оптимальный (наименьший) требуемый объем второго ресурса [43].

Среди известных двухфакторных производственных функций можно назвать также производственные функции Солоу, Аллена, линейную, постоянной эластичности факторов.

5.5 Вопросы для самоконтроля

1. В каких случаях применяются нелинейные модели регрессии, и каким образом можно их оценивать?

2. Опишите порядок линеаризации зависимостей параболического и гиперболического типов.

3. Каковы способы приведения к линейному виду моделей, нелинейных по параметрам?

4. Как проводится подбор линеаризующего преобразования?

5. Приведите примеры производственных функций. С какой целью они составляются?

6. Каким образом описываются производственные функции?

5.6 Тесты

1. Отметьте правильную форму степенной функции:

а) $y = \beta + \frac{\beta_1}{x} + \varepsilon$;

б) $y = \beta_0 + \beta_1 \ln x + \varepsilon$;

в) $y = \beta_0 e^{\beta_1 x + \varepsilon}$;

г) $y = \beta_0 (x)^{\beta_1}$;

д) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon_i$.

2. К какому классу нелинейных регрессий относится равносторонняя гипербола:

а) регрессии, нелинейные относительно включенных в анализ переменных, но линейных по оцениваемым параметрам;

б) нелинейные регрессии по оцениваемым параметрам.

3. Для линеаризации экспоненциальной зависимости используют преобразование

а) $x^* = \ln x$;

б) $x_1^* = x; \quad x_2^* = x^2$;

в) $x^* = \frac{1}{x}$;

г) $y^* = \frac{1}{y}$;

д) $y^* = \ln y$.

4. Для моделирования зависимости спроса от цены товара можно использовать функцию

а) показательную;

б) логарифмическую;

в) гиперболическую;

г) параболическую;

д) линейную;

е) степенную.

5. Отметьте правильную форму показательной функции:

а) $y = \beta + \frac{\beta_1}{x} + \varepsilon$;

б) $y = \beta_0 + \beta_1 \ln x + \varepsilon$;

в) $y = \beta_0 e^{\beta_1 x + \varepsilon}$;

г) $y = \beta_0 (x)^{\beta_1}$;

д) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon_i$.

6. Для линеаризации степенной зависимости используют преобразование

а) $x^* = \ln x$;

б) $x_1^* = x; \quad x_2^* = x^2$;

в) $x^* = \frac{1}{x}$;

г) $y^* = \frac{1}{y}$;

д) $y^* = \ln y$.

7. Кривая Филлипса, отражающая зависимость между уровнем безработицы и процентом прироста заработной платы, является примером

а) параболы;

б) равносторонней гиперболы;

в) степенной зависимости;

г) логарифмической зависимости.

8. Отметьте правильную форму параболической функции:

а) $y = \beta + \frac{\beta_1}{x} + \varepsilon$;

б) $y = \beta_0 + \beta_1 \ln x + \varepsilon$;

в) $y = \beta_0 e^{\beta_1 x + \varepsilon}$;

г) $y = \beta_0 (x)^{\beta_1}$;

д) $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon_i$.

9. Зависимость, характеризующая развитие различных сторон потенциала организации и ее положения во внешней среде, описывается с помощью

а) экспоненты;

б) гиперболы;

в) логистической кривой;

г) параболы.

10. Модели $\ln Y = \beta 0 + \beta X + \varepsilon$, $Y = \beta 0 + \beta \ln X + \varepsilon$ называются:

а) полулогарифмическими;

б) логарифмическими;

в) линейными.

11. При тестировании регрессионных моделей с помощью преобразования П. Зарембски используется

а) *t*-критерий Стьюдента;

б) *F*-критерий Фишера;

в) критерий χ^2;

г) критерий Дарбина-Уотсона.

12. Какое свойство производственной функции свидетельствует о том, что с ростом затрат хотя бы одного ресурса, объем выпуска растет

а) $f(0,0) = 0$;

б) $f(0, x_2) = f(x_1, 0) = 0$;

в) $x_1 \geq x_2 \quad f(x_1) > f(x_2)$;

г) $x > 0 \Rightarrow \dfrac{\partial f(x)}{\partial x_i} > 0$.

д) $x > 0 \Rightarrow \dfrac{\partial^2 f(x)}{\partial x_i^2} \leq 0$;

е) $x > 0 \Rightarrow \dfrac{\partial^2 f(x)}{\partial x_i \partial x_j} \geq 0$.

13. Какой смысл у коэффициентов регрессии в логарифмических регрессионных моделях

а) показывают процентное изменение *Y* для данного абсолютного изменения *X*;

б) показывают процентное изменение *Y* для данного процентного изменения *X*;

в) показывают абсолютное изменение *Y* для данного процентного изменения *X*.

14. К какому классу нелинейных регрессий относится функция вида $y = \beta_0 + \beta_1 \ln x + \varepsilon$:

а) регрессии, нелинейные относительно включенных в анализ переменных, но линейных по оцениваемым параметрам;

б) нелинейные регрессии по оцениваемым параметрам.

15. Какое свойство производственной функции свидетельствует о том, что при росте одного ресурса предельная эффективность другого ресурса возрастает

а) $f(0,0) = 0$;

б) $f(0, x_2) = f(x_1, 0) = 0$;

в) $x_1 \geq x_2 \quad f(x_1) > f(x_2)$;

г) $x > 0 \Rightarrow \dfrac{\partial f(x)}{\partial x_i} > 0$.

д) $x > 0 \Rightarrow \dfrac{\partial^2 f(x)}{\partial x_i^2} \leq 0$;

е) $x > 0 \Rightarrow \dfrac{\partial^2 f(x)}{\partial x_i \partial x_j} \geq 0$.

16. При анализе производственной функции целесообразно использовать следующую модель:

а) линейную;

б) полиномиальную;

в) логарифмическую;

г) степенную;

д) экспоненциальную.

17. Модель вида $Y = A \cdot K^{\alpha} \cdot L^{\beta}$ носит название:

а) степенной модели;

б) функции Леонтьева;

в) функции Энгеля;

г) функции Кобба – Дугласа;

д) лог-линейной модели.

6 Модели регрессии с переменной структурой

Что необходимо знать из 6 главы:

1. Понятие фиктивных переменных, их виды и роль в эконометрических исследованиях.

2. Возможности использования бинарных фиктивных переменных в регрессионном анализе.

3. Регрессионные модели с фиктивными переменными, принимающими более двух значений.

4. Проведение регрессионного анализа с фиктивной результативной переменной.

5. Цель и этапы проведения теста Чоу.

6.1 Понятие и виды фиктивных переменных

Факторные переменные, применяемые в моделях регрессии, обычно могут принимать значения из какого-либо непрерывного интервала - инвестиции, размер заработной платы, уровни занятости и безработицы, потребительские цены, и т.п. Иногда может оказаться необходимым рассматривать в регрессионном анализе какие-либо качественные (атрибутивные) переменные, которые имеют два или более различных уровней. Особенно часто такая задача возникает при изучении данных выборочных обследований социологической, психологической или экономической направленности. К таким переменным можно отнести, например, профессию, пол, уровень образования, наличие вредных привычек, состояние в браке, наличие или отсутствие детей, сезонность и т.п.

В качестве примера можно привести изучение вторичного рынка автомобилей и построение регрессионной модели влияния факторов на цену продаваемого автомобиля. Наряду с числовыми переменными, такими как величина пробега или год выпуска, в выборке присутствуют переменные, характеризующие тип коробки передач (автоматическая, механическая, смешанная), наличие аудиосистемы (присутствует или отсутствует), участие автомобиля в ДТП (участвовал или не участвовал) и другие характеристики, которые оказывают влияние на величину стоимости. Ярким примером зависимости между нечисловыми и числовыми данными является влияние социально-экономических факторов на качество услуг (низкое или высокое), оказываемых коммерческими организациями и индивидуальными предпринимателями.

В англоязычной литературе по эконометрике переменные указанного выше типа называются *dummy variables*. что на русский язык часто переводится как *«фиктивные переменные»* (см., например, [4]). Следует, однако, ясно понимать, что *d* такая же «равноправная» переменная, как и любой из регрессоров x_j , $j = 1,2,...,k$. Ее «фиктивность» состоит только в том, что она количественным образом описывает качественный признак [22, с. 96].

Фиктивные переменные называют также структурными, искусственными, манекенными, двоичными, индикаторами. Для того чтобы ввести такие переменные в модель, мы можем приписать этим факторам некоторые уровни по порядку, учитывая тот факт, что различные качественные признаки могут иметь независимые детерминированные эффекты в результативной переменной.

Атрибутивные признаки могут существенно влиять на структуру линейных связей между переменными (поэтому в отечественной литературе для их обозначения используется термин «структурные переменные»). В этом случае говорят об исследовании регрессионных моделей с переменной структурой или построении регрессионных моделей по неоднородным данным.

При эконометрическом моделировании с включением качественных переменных можно использовать два подхода: регрессия строится для каждой качественно отличной группы единиц совокупности; регрессионная модель строится для совокупности в целом. В этом случае в регрессионную модель вводятся фиктивные переменные, т.е. строится регрессионная модель с переменной структурой, отражающей неоднородность данных. Второй подход обладает двумя важными преимуществами: во-первых, имеется простой способ проверки, является ли воздействие качественного фактора значимым; во-вторых, при условии выполнения определенных предположений регрессионные оценки оказываются более эффективными.

Таким образом, кроме моделей, содержащих только количественные объясняющие переменные (обозначаемые X_j), в эконометрике рассматривают содержащие лишь качественные переменные (обозначаемые D_i), либо те и другие одновременно. Фиктивные переменные могут вводиться не только в нелинейные, но и в нелинейные модели, приводимые путем преобразования к линейному виду.

В качестве фиктивных переменных обычно используют дихотомические (бинарные, булевы) переменные, принимающие два значения – «0» либо «1», так как при этом содержательная интерпретация полученных результатов достаточно проста. Обычно фиктивная переменная отражает два противоположных состояния качественного фактора и может выражаться в двоичной форме:

$$D_{ij} = \begin{cases} 1, \text{если объект обладает свойством;} \\ 0, \text{иначе.} \end{cases} \qquad i = \overline{1, p-1} \ . \qquad (6.1)$$

При этом, если качественная переменная имеет p градаций, то для отражения ее влияния на структуру искомой регрессионной связи необходимо ввести $(p-1)$ фиктивных переменных. Иначе для любого объекта наблюдения выполнялось бы тождество: $D_{i1} + D_{i2} + ... + D_{ip} = 1$, что означало бы линейную зависимость объясняющих переменных, и как следствие, невозможность получения МНК-оценок.

В связи с тем, что фиктивные переменные в регрессионных моделях могут располагаться как в левой, так и в правой части (быть зависимой переменой и не зависимой), а также входить как отдельно, так и совместно с объясняющими переменными, классификацию подобных моделей можно представить в виде следующей схемы (рисунок 6.1):

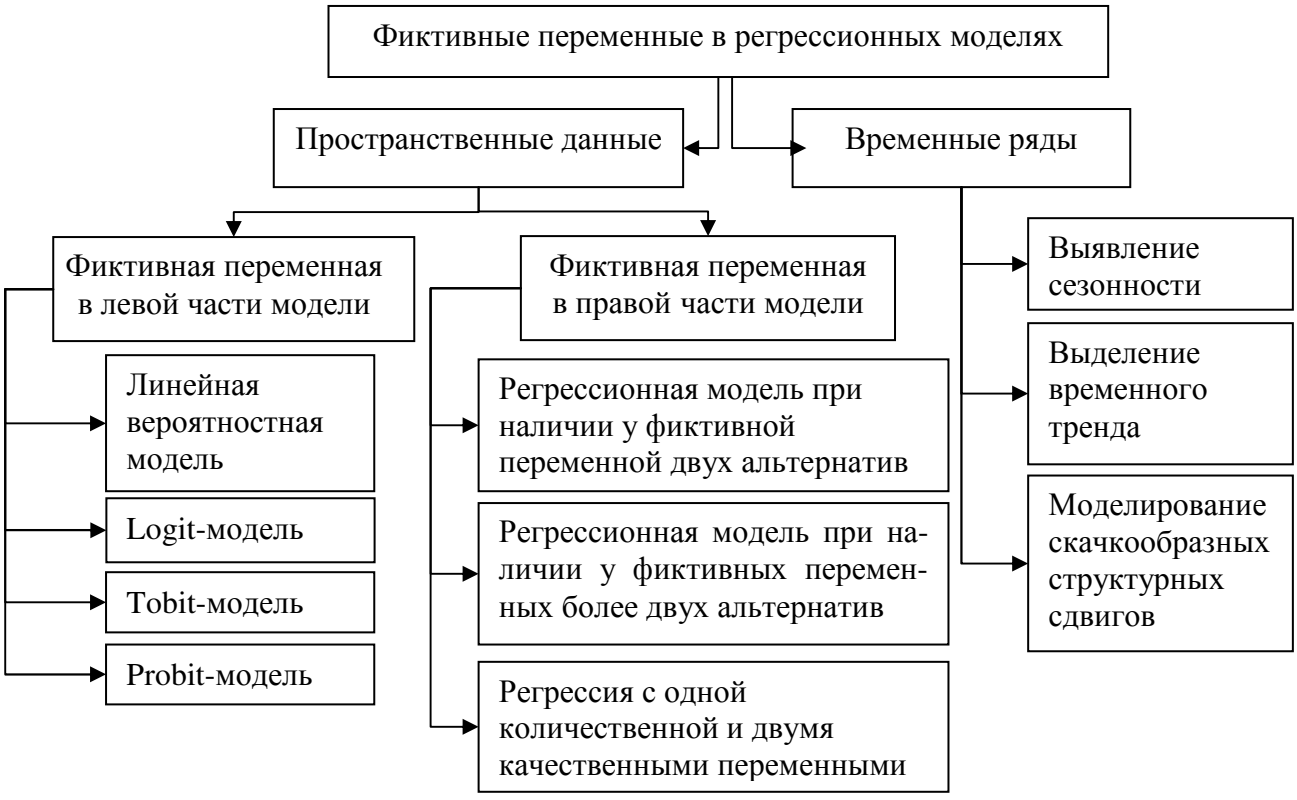


Рисунок 6.1 – Классификация регрессионных моделей с фиктивными переменными

Далее рассмотрим особенности применения фиктивных переменных при построении регрессионных моделей на основе пространственных данных.

6.2 Регрессионные модели с бинарными фиктивными переменными

Самым простым случаем использования фиктивных переменных является построение регрессионных моделей при наличии у нечисловой переменной только двух альтернатив, т.е. одной количественной и одной качественной переменной.

В этом случае первоначальная регрессионная модель изменится и примет вид:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \alpha_1 D_{i1} + \varepsilon_i,\ i = \overline{1,n}. \qquad (6.2)$$

Коэффициент α в приведенной модели *называется дифференциальным коэффициентом свободного члена*, так как он показывает, на какую величину отличается свободный член модели при значении фиктивной переменной, равном единице, от свободного члена модели при базовом значении фиктивной переменной.

Простота подхода, основанного на измерении влияния качественных переменных на количественную с помощью регрессионного анализа, объясняется тем, что для оценке неизвестных параметров используется обычный МНК, т.е. фиктивная переменная рассматривается как еще одна независимая переменная.

При рассмотрении фиктивных переменных показателен (и не только в плане эконометрического моделирования) пример, приведенный К. Доугерти [6, с. 263-267].

Пример 6.1 - Рассматривалась регрессионная зависимость веса новорожденного (y) от количества сигарет, выкуриваемых в день будущей матерью (x). Выборка составляла 964 наблюдения о родах. В качестве отправной точки была взята модель вида (в наших обозначениях):

$$y = b_0 + b_1 x + \varepsilon. \qquad (6.3)$$

Оценив регрессию по выборке, было получено следующее уравнение зависимости:

$$\tilde{y} = 3418 - 7{,}2x; \quad R = 0{,}012,$$

отражающее, что ребенок, рожденный некурящей матерью, будет иметь при рождении средний вес около 3400 г, а уменьшение веса новорожденного по причине курения составит чуть больше 7 г на каждую сигарету, выкуриваемую матерью.

Далее исследовалось воздействие на результативный признак качественного фактора: первенец ли родившийся ребенок. Данная переменная была представлена в виде фиктивной (искусственно введенной):

$$D = \begin{cases} 1, \text{ родившийся ребенок - первенец;} \\ 0, \text{ ребенок родился не первым.} \end{cases}.$$

Регрессионная модель с учетом нового фактора имеет вид:

$$y = b_0 + b_1 x + \alpha D, \tag{6.4}$$

где параметр D будет отражать в среднем разницу в весе новорожденного у первенцев и детей, родившихся не первыми.

Наблюдение	Первенец?	y	x	D	Наблюдение	Первенец?	y	x	D
1	Нет	3520	10	1	11	Нет	3210	29	1
2	Нет	3460	19	1	12	Нет	3290	15	1
3	Нет	3000	16	1	13	Да	3190	3	0
4	Нет	3320	26	1	14	Да	3060	12	0
5	Нет	3540	4	1	15	Да	3270	17	0
6	Нет	3310	14	1	16	Да	3170	14	0
7	Нет	3360	21	1	17	Да	3230	18	0
8	Нет	3650	10	1	18	Да	3700	11	0
9	Нет	3150	22	1	19	Да	3300	14	0
10	Нет	3440	8	1	20	Да	3460	9	0

Рисунок 6.2 – Выборка, содержащая 20 наблюдений [6]

По выборке, состоящей из 20-ти наблюдений (рисунок 6.2), результаты регрессионного анализа получились следующие:

$$\tilde{y} = 3444 + 103D - 11{,}9x; \quad R^2 = 0{,}19.$$

Параметр сдвига составил 103 грамма. Полученное уравнение можно представить как:

$\tilde{y} = 3444 - 11{,}9x$ (для первенца);

$\tilde{y} = 3547 - 11{,}9x$ (для ребенка, родившегося не первым).

Данные линии регрессии отражены на рисунке 6.3.

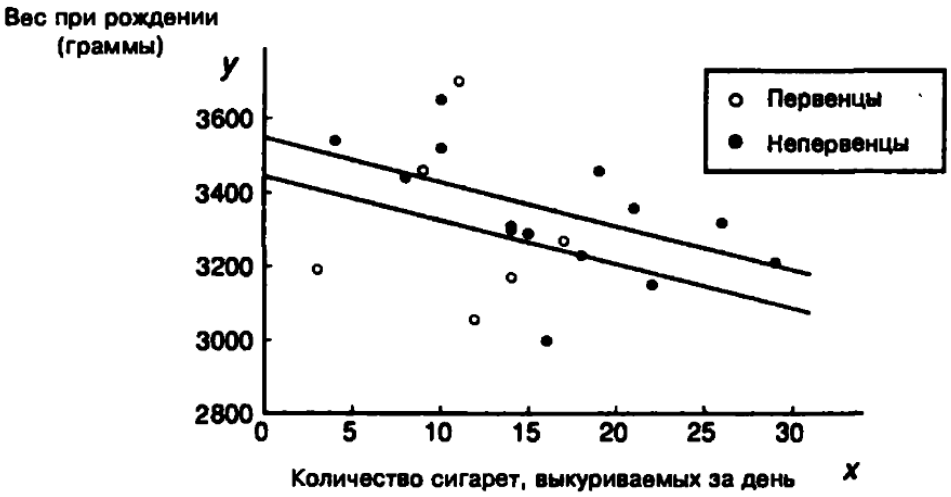


Рисунок 6.3 – Регрессия зависимости веса новорожденного от степени пристрастия будущей матери к курению [6]

Проверка значимости отличия параметров от нуля проводилась путем вычисления *t*-статистики Стьюдента (находили отношение величины коэффициента на его стандартную ошибку), которая при заданном уровне значимости сравнивалась с критическим значением *t*. В уравнении, составленном для 20-ти наблюдений значение *t* – статистики для фиктивной переменной составило 1,23 - сдвиг линии регрессии для первенцев и не первенцев незначим. Это объяснялось небольшим размером выборки, т.к. эффект, вызываемый тем, что ребенок первенец (или не первенец),

проявляется только как тенденция, он слишком невелик, чтобы оценить его по столь малому числу наблюдений. Оценивание регрессии по реальным данным о 964 родах дало следующий результат:

$$\tilde{y} = 3373 + 119D - 7{,}8x; \quad R^2 = 0{,}032.$$

При моделировании регрессии на реальных данных по 964 наблюдениям было получено значение *t* – статистики, равное 4,58, что свидетельствует о том, что в действительности сдвиг линии регрессии значим. Рассмотрим следующий пример.

Пример 6.2 - Имеются данные выборочного обследования вторичного рынка легковых автомобилей г. Оренбурга (март 2009 г.) (таблица 6.1).

Таблица 6.1 – Исходные данные для построения регрессионной модели влияния факторов на цену подержаного автомобиля

№ наблюдения	Пробег, тыс. км	*D*	Цена, тыс. р.	№ наблюдения	Пробег, тыс. км	*D*	Цена, тыс. р.
1	300	0	140	17	80	1	330
2	70	1	500	18	150	0	100
3	62	1	580	19	30	0	150
4	13	0	440	20	40	1	550
5	22	0	450	21	350	0	120
6	85	0	195	22	300	0	135
7	93	1	450	23	50	1	500
8	84	0	50	24	95	0	300
9	80	1	300	25	78	0	250
10	300	0	125	26	80	0	255
11	40	1	400	27	100	0	88
12	190	1	260	28	30	0	360
13	60	0	260	29	120	0	450
14	40	0	250	30	50	1	675
15	60	1	510	31	100	0	800
16	60	1	410	-	-	-	-
Примечание - В таблице приведены сведения по автомобилям иностранного производства							

С помощью регрессионной модели оценим влияние пробега автомобиля (x) и типа коробки передач (D) на цену продажи (y).

В данном случае фиктивная переменная будит иметь следующий вид:

$$D = \begin{cases} 1, \text{если автомобиль с автоматической коробкой передач;} \\ 0, \text{если автомобиль с механической коробкой передач.} \end{cases}$$

Тогда ожидаемая цена автомобиля при x км пробега будет:

- $f(y \mid x,\ D=0) = b_0 + b_1 x_i$ - для автомобилей с МКПП;

- $f(y \mid x,\ D=1) = b_0 + b_1 x + \alpha = (b_0 + \alpha) + b_1 x_i$ - для автомобилей с АКПП.

Цена автомобиля в данном случае является линейной функцией от пробега автомобиля. Причем и для автомобилей с МКПП и АКПП цена меняется с одним и тем же коэффициентом пропорциональности b_1, а свободные члены в приведенных моделях отличаются на величину α.

Проверив с помощью t-статистики статистические значимости коэффициентов b_0 и (b_0+α), можно определить, имеет ли место завышение (занижение) цены на вторичном рынке в зависимости от типа коробки передач. Если эти коэффициенты окажутся статистически значимыми, то влияние качественной переменной доказано. Более того, при $\alpha > 0$ «перевес» будет в сторону автомобилей с АКПП, при $\alpha < 0$ - в пользу автомобилей с МКПП.

Воспользовавшись возможности пакета программ Statistica (Анализ → Множественная регрессия) получим следующие результаты (таблица 6.2).

Из приведенной таблицы можно вывести уравнения:

- для автомобилей с МКПП - $\tilde{y}_i = 354{,}335 - 0{,}776 x_i$;

- для автомобилей с АКПП - $\tilde{y}_i = 354{,}335 - 0{,}776 x_i + 154{,}692 = 509{,}027 - 0{,}776 x_i$.

Так как параметр α статистически значим по t-критерию Стьюдента (p-уровень значимости не превышает 5 %) можно сделать вывод о существенности влияния автоматической коробки переключения передач на цену подержаных автомобилей иностранного производства на вторичном рынке г. Оренбурга.

Таблица 6.2 – Оценки регрессионной модели влияния факторов на цену подержаного автомобиля в г. Оренбурге

Параметры	β	Стандартная ошибка β	Искомые параметры	Стандартная ошибка искомых параметров	t-статистика Стьюдента	p-уровень значимости
Свободный член	-	-	354,335	51,620	6,864	0,000
x	-0,375	0,152	-0,776	0,315	-2,466	0,020
D	0,420	0,152	157,692	57,080	2,763	0,010

Полученные результаты можно представить графически. Для этого в частные регрессионные уравнения для авто с МКПП и АКПП подставляют значения независимой переменной x и полученные теоретические и фактические значения y наносят на поле корреляции.

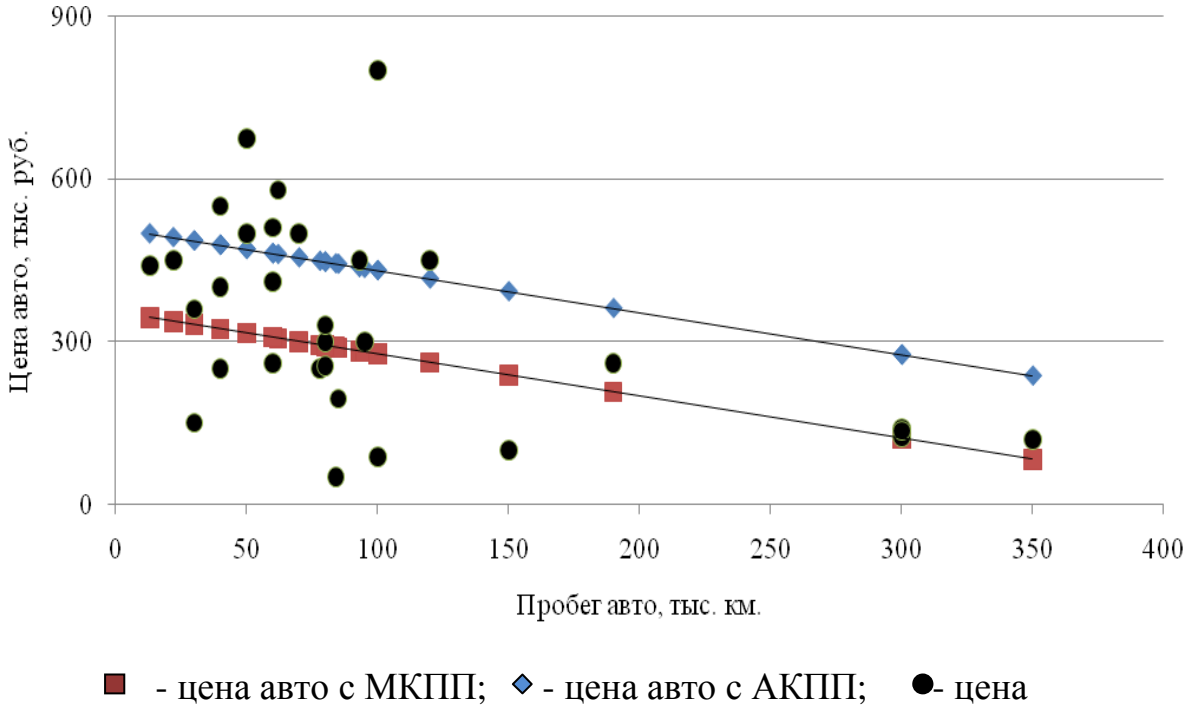


Рисунок 6.4 – Результаты регрессионного анализа влияния факторов на цену подержаного автомобиля в г. Оренбурге

Согласно приведенному рисунку, явно выделяются две группы автомобилей - с ценой, варьирующей относительно 300 тыс. р. и 500 тыс. р., что соответствует автомобилям с МКПП и АКПП.

На практике регрессионная модель 6.2 часто используется для подтверждения правильности разбиения совокупности на группы (например, в рамках кластерного анализа) и измерении влияния факторов в этих группах. В том случае, если параметр α статистически значим, исследователь убеждается в достоверности группировки и в действии в данных группах одних и тех же факторов. В противном случае необходимо рассматривать отдельные регрессионные модели по каждой группе, с выделением закономерностей, присущих каждой из выделенных групп.

6.3 Регрессионные модели с фиктивными переменными, принимающими более двух значений

В практике эконометрического моделирования распространен случай, когда зависимость выражается моделью с двумя объясняющими переменными, одна из которых количественная, а другая - качественная, при этом фиктивная переменная имеет более чем две альтернативы.

В этом случае регрессионное уравнение выглядит следующим образом:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik} + \alpha_1 D_{i1} + \alpha_2 D_{i2} + \varepsilon_i,\ i = \overline{1,n}. \qquad (6.5)$$

Здесь D - фиктивная переменная, имеющая следующие альтернативы:

$$D_1 = \begin{cases} 0, \text{если значение относится к первой категории;} \\ 1, \text{если значение не относится к первой категории.} \end{cases}$$

$$D_2 = \begin{cases} 0, \text{если значение относится ко второй категории;} \\ 1, \text{если значение не относится ко второй категории.} \end{cases} \qquad (6.6)$$

Соответственно регрессионные уравнения будут для каждой категории (альтернативы) будут иметь следующий вид:

- $f(y \mid x,\ D_1 = 0, D_2 = 0) = b_0 + b_1 x_i$ - для первой категории;

- $f(y \mid x,\ D_1 = 1, D_2 = 0) = (b_0 + \alpha_1) + b_1 x_i$ - для второй категории;

- $f(y \mid x,\ D_1 = 1, D_2 = 1) = (b_0 + \alpha_1 + \alpha_2) + b_1 x_i$ - для третьей категории.

Пример 6.3 - На основе выборочной совокупности, характеризующей вторичный рынок жилья г. Оренбурга (май 2010 г.) (таблица 6.3), оценим влияние площади квартиры (x) и типа стен дома (D) на цену продаваемой квартиры (y).

Таблица 6.3 – Исходные данные для построения регрессионной модели влияния факторов на цену кв. м жилья на вторичном рынке г. Оренбурга

№	Площадь, кв. м	Цена, млн. р.	D_1	D_2	№	Площадь, кв. м	Цена, млн. р.	D_1	D_2
1	47,0	1,200	0	1	18	30,0	1,100	0	1
2	49,0	1,150	1	1	19	33,5	1,600	1	0
3	42,5	1,250	1	0	20	29,0	0,950	1	1
4	36,0	1,370	0	1	21	40,0	1,250	0	1
5	38,0	1,420	0	1	22	70,0	3,600	1	0
6	35,0	1,360	1	0	23	44,0	1,900	1	0
7	33,0	1,550	1	1	24	49,0	1,450	0	1
8	30,0	1,250	1	1	25	35,3	0,950	1	0
9	39,0	1,850	0	1	26	48,0	1,220	1	0
10	30,0	1,100	1	0	27	49,5	1,850	1	0
11	30,7	1,050	1	0	28	60,0	1,630	1	0
12	33,0	1,380	1	1	29	41,0	1,080	1	1
13	37,0	1,650	1	1	30	48,0	1,180	1	1
14	37,7	1,430	0	1	31	41,0	1,430	1	0
15	36,5	1,100	0	1	32	39,0	1,300	0	1
16	35,8	1,170	1	1	33	45,0	1,450	1	0
17	52,3	1,150	0	1	-	-	-	-	-
Примечание - В таблице приведены сведения по однокомнатным квартирам.									

В данном примере фиктивная переменная будет принимать следующие значения:

$$D_1 = \begin{cases} 0, \text{если кирпичные стены;} \\ 1, \text{если другой тип стен.} \end{cases}$$

$$D_2 = \begin{cases} 0, \text{если панели;} \\ 1, \text{если другой тип стен.} \end{cases}$$

Оценка параметров регрессионного уравнения представлена в таблице 6.4.

Таблица 6.4 – Оценки регрессионной модели влияния факторов на цену кв. м жилья на вторичном рынке г. Оренбурга

Параметры	β	Стандартная ошибка β	Искомые параметры	Стандартная ошибка искомых параметров	t-статистика Стьюдента	p-уровень значимости
Свободный член	-	-	0,257	0,409	0,630	0,534
x	0,580	0,150	0,030	0,008	3,872	0,001
D_1	0,026	0,177	0,025	0,173	0,144	0,886
D_2	-0,133	0,182	-0,125	0,171	-0,730	0,471

Из таблицы следует, что статистически значимым является параметр b_1 (p-уровень значимости не превышает 5 %), коэффициенты при фиктивных переменных не значимы и указывают на отсутствие влияния типа стен многоэтажных домов на цену кв. метра жилья на вторичном рынке г. Оренбурга.

Рассмотренные выше регрессионные модели можно усложнить, построив множественное уравнение регрессии с двумя количественными (y, x) и двумя качественными (D_1, D_2) переменными, при этом фиктивные переменные измерены в дихотомической шкале. Стоит отметить, что при внешнем сходстве модели, рассмотренной выше с данной моделью, цели их построения различны. В первом случае модель строится для измерения влияния одного признака, принимающего несколько вариантов, во втором - для измерения влияния двух независимых друг от

друга признаков, каждый из которых принимает только два варианта – действует или не действует признак.

Пример 6.4 - Используя данные выборочного обследования вторичного рынка подержанных автомобилей г. Оренбурга (май 2010 г.), оценим влияние автопробега (x), типа двигателя (D_1) и мощности двигателя (D_2) на цену продаваемых автомобилей (y) (таблица 6.5).

Таблица 6.5 – Исходные данные для построения регрессионной модели влияния факторов на цену подержаного автомобиля в г.Оренбурге (выборка 2010 г.)

№	Цена, тыс. р.	Пробег, тыс. км	D_1	D_2	№	Цена, тыс. р.	Пробег, тыс. км	D_1	D_2
1	500	40	0	0	21	650	57	0	1
2	550	85	0	1	22	670	65	0	1
3	385	50	1	0	23	580	137	0	1
4	345	198	1	0	24	380	188	1	0
5	570	89	0	1	25	550	125	0	1
6	440	68	1	0	26	780	60	0	1
7	635	25	0	1	27	500	136	1	0
8	684	61	0	0	28	380	186	1	0
9	325	148	1	0	29	430	40	1	0
10	555	84	0	1	30	620	55	0	1
11	430	48	1	0	31	450	57	1	0
12	540	61	0	1	32	510	95	0	1
13	385	120	1	0	33	400	32	1	0
14	450	62	1	0	34	450	120	0	0
15	500	132	0	0	35	650	87	0	1
16	480	35	1	0	36	290	150	1	0
17	600	30	0	1	37	440	75	1	0
18	560	73	0	1	38	505	135	0	0
19	320	116	1	0	39	690	66	0	1
20	650	60	0	1	40	598	89	0	1
Примечание - Приведены сведения по автомобилям иностранного производства									

При этом фиктивные переменные имеют следующую кодировку:

$$D_1 = \begin{cases} 0, \text{если мотор работает на дизельном топливе;} \\ 1, \text{если мотор работает на бензине.} \end{cases}$$

$$D_2 = \begin{cases} 0, \text{если мощность мотора не превышает 100 л.с.;} \\ 1, \text{если мощность мотора больше 100 л.с.} \end{cases}$$

Дальнейшая процедура аналогична описанной выше, поэтому перейдем к рассмотрению результатов регрессионного анализа (таблица 6.6).

Таблица 6.6 – Оценки модели регрессии влияния факторов на цену подержанного автомобиля в г. Оренбурге

Параметры	β	Стандартная ошибка β	Искомые параметры	Стандартная ошибка искомых параметров	t-статистика Стьюдента	p-уровень значимости
Свободный член	-	-	588,328	34,700	16,955	0,000
x	-0,237	0,084	-0,620	0,221	-2,803	0,008
D_1	-0,543	0,129	-130,101	30,907	-4,209	0,000
D_2	0,280	0,131	66,721	31,123	2,144	0,039

Полученные параметры регрессионного уравнения статистически значимы по t-критерию Стьюдента, что доказывает влияние на цену нечисловых признаков.

Из представленных результатов получим следующие регрессионные зависимости:

- средняя цена авто с дизельным двигателем с мощностью менее 100 л. с.:

$$f(y \mid x,\ D_1 = 0, D_2 = 0) = b_0 + b_1 x_i,$$

$$\tilde{y}_i = 588{,}328 - 0{,}62 x_i.$$

- средняя цена авто с дизельным двигателем с более менее 100 л. с.:

$$f(y \mid x,\ D_1 = 0, D_2 = 1) = (b_0 + \alpha_2) + b_1 x_i,$$

$$\tilde{y}_i = (588{,}328 + 66{,}721) - 0{,}62x_i .$$

- средняя цена авто с бензиновым двигателем с мощностью менее 100 л. с.:

$$f(y \mid x,\ D_1 = 1, D_2 = 0) = (b_0 + \alpha_1) + b_1 x_i ,$$

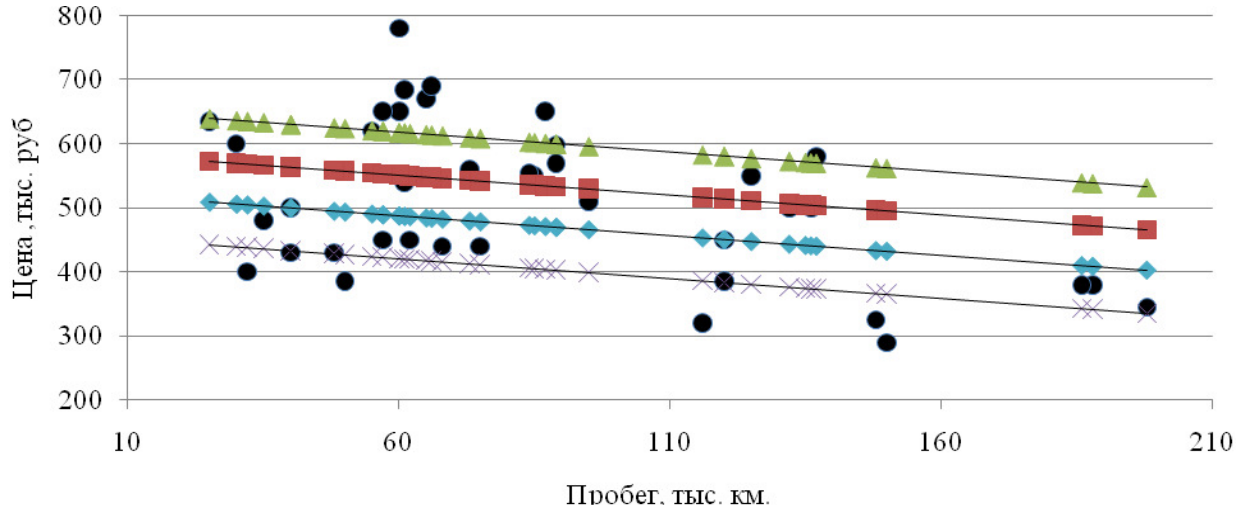
$$\tilde{y}_i = (588{,}328 - 130{,}101) - 0{,}62x_i .$$

- средняя цена авто с бензиновым двигателем с более менее 100 л. с.:

$$f(y \mid x,\ D_1 = 1, D_2 = 1) = (b_0 + \alpha_1 + \alpha_2) + b_1 x_i ,$$

$$\tilde{y}_i = (588{,}328 - 130{,}101 + 66{,}721) - 0{,}62x_i .$$

Отразим полученные результаты графически, для этого подставим значения независимой переменной *x* в полученные уравнения.



● - цена авто;
■ - средняя цена авто с дизельным двигателем менее 100 л.с.;
▲ - средняя цена авто с дизельным двигателем более 100 л.с.;
⊗ - средняя цена авто с бензиновым двигателем менее 100 л.с.;
◆ - средняя цена авто с бензиновым двигателем более 100 л.с.

Рисунок 6.4 – Результаты регрессионного анализа модели влияния факторов на цену подержаного автомобиля в г. Оренбурге

Все полученные регрессии отличаются лишь свободными членами, в результате чего они расположены на разных уровнях, но наклон у них одинаков.

Рассмотренные выше регрессные уравнения, в которых фиктивные переменные находились в правой части, являются самыми распространенными случаями в эконометрических исследованиях. Естественно, что предложенные выше схемы могут быть распространены на ситуации с произвольным числом количественных и качественных факторов, но не стоит забывать, что увеличение числа фиктивных переменных в модели приводит к усложнению интерпретации полученных результатов или делает их невозможными. Кроме того, в связи с нечисловой природой признаков, искомые параметры, как правило, получаются статистически незначимыми (хотя в индивидуальных регрессионных моделях могут оказаться значимыми).

6.4 Случай для фиктивной переменной в левой части уравнения

В “фиктивной” форме может быть выражена и зависимая переменная. Такая ситуация имеет место, например, при проведении социологических опросов, когда их результат может быть представлен двумя ответами “да”, “нет” (1 или 0) (предполагаемая покупка недвижимости, автомобиля; желание иметь ребенка в семье и т. п.), а влияющие на этот результат факторы выражаются в произвольной форме (количественные характеристики – уровень дохода, жилая площадь и т. п., качественные характеристики – уровень образования, состояние в браке и т. д.).

Тогда расчетные значения $\tilde{y}$, определенные по модели при различных комбинациях значений независимых переменных x_i, можно интерпретировать как оценку условий вероятности события y при фиксированных значениях x_i, i=1,2,..., n [44, с. 545].

Рассмотрим вариант регрессионных моделей с фиктивными переменными, когда зависимая переменная является структурной. Для решения подобной задачи можно прибегнуть двум группам моделей:

1. *Фиктивные переменные являются бинарными*, т.е. принимают два значения: «0» – истина и «1» – ложь. В качестве таковых можно назвать *логит (logit)-модель* и *пробит (probit)-модель*.

Пробит-модель основана на законе нормального распределения $N(0,1)$:

$$F(z) = f(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2 dt}. \tag{6.7}$$

Пробит-модель для бинарных данных:

$$P(y_i = 1 \mid X_i^T) = \int_{-\infty}^{X_i^T \beta} \varphi(t)dt = f(X_i^T \beta), \tag{6.8}$$

$$P(y_i = 0 \mid X_i^T) = 1 - f(X_i^T \beta).$$

Логит-модель основывается на логистическом законе распределения вероятностей. Функция распределения вероятностей логистического закона:

$$\Lambda(z) = \frac{e^z}{1 - e^z}. \tag{6.9}$$

Логит-модель для бинарных данных:

$$P(y_i = 1 \mid X_i^T) = \frac{e^{X_i^T \beta}}{1 - e^{X_i^T \beta}} = \Lambda(X_i^T \beta), \tag{6.10}$$

$$P(y_i = 0 \mid X_i^T) = \frac{1}{1 - e^{X_i^T \beta}} = 1 - \Lambda(X_i^T \beta).$$

Графики функции распределения нормального и логистического распределения при соответствующей нормировке достаточно близки. На интервале $z \in [-1{,}2;\ 1{,}2]$ они практически одинаковы. Однако логистическая функция мед-

леннее стремится к нулю или единице при $z \to \pm\infty$. В связи с этим обе рассмотренные модели дают похожий результат, если только изучаемая вероятность не слишком близка к нулю или единице.

2. *Фиктивная переменная принимает любое значение в заданном диапазоне*. В качестве подобных моделей можно назвать: модели множественного выбора, модели упорядоченного выбора, модели многовариантного бинарного выбора.

При рассмотрении регрессионных моделей, у которых фиктивная переменная находится в левой части, нельзя использовать обыкновенный МНК, т.к. полученные оценки не будут обладать свойствами наилучших линейных несмещенных оценок (BLUE). Поэтому для определения коэффициентов в этом случае используются другие методы [45].

Пример 6.5 - Используя данные выборочного обследования локального рынка услуг сотовой связи г. Оренбурга (июнь 2009 года) (таблица 6.7), оценим влияние на качество данных услуг (y) пола респондента (D) и средних ежемесячных трат на сотовую связь (x).

При этом зависимая переменная является дихотомической, принимает значение «0» в случае оценки качества как неудовлетворительное и «1», если качество удовлетворяет потребителя.

Что касается фиктивной переменной в правой части уравнения (пол респондента), то 0 присвоен женщинам, 1 – мужчинам.

Воспользуемся возможностями пакета Statistica, позволяющими оценивать параметры как logit-, так и probit-модели (Анализ → Углубленные методы анализа → Нелинейное оценивание).

Результатом оценивания является следующая probit-модель (метод оценивания Хука-Дживиса):

$$y_i = -2{,}54 + 0{,}115 D_i + 0{,}01 x_i, \quad \chi^2 = 29{,}999, \quad p^2 = 0{,}000.$$

Таблица 6.7 – Исходные данные для оценки влияния факторов на качество услуг сотовой связи на локальном рынке г. Оренбурга

№	Качество услуги	Пол респондента	Цена, р.	№	Качество услуги	Пол респондента	Цена, р.
1	0	1	150	26	1	1	450
2	0	0	120	27	1	1	350
3	0	0	150	28	1	1	350
4	0	0	150	29	1	1	450
5	0	1	500	30	1	1	450
6	1	1	350	31	1	1	500
7	1	1	350	32	1	1	500
8	1	1	350	33	0	1	150
9	1	1	350	34	0	0	250
10	1	1	450	35	0	0	150
11	1	0	250	36	0	0	150
12	1	1	250	37	0	0	150
13	1	1	500	38	0	1	150
14	0	0	150	39	1	0	250
15	0	1	250	40	1	0	350
16	0	0	150	41	1	0	350
17	0	1	150	42	1	1	350
18	0	0	250	43	1	1	450
19	0	1	250	44	0	0	150
20	1	0	250	45	1	0	250
21	1	1	250	46	1	1	250
22	1	1	350	47	1	1	250
23	1	1	350	48	1	1	450
24	1	1	450	49	1	1	450
25	1	1	450	50	1	1	500

На основе полученного уравнения регрессии можно сделать заключение, что влияние выделенных факторов существенно.

Для расширенной интерпретации полученных результатов можно прибегнуть к графику распределения (рисунок 6.5).

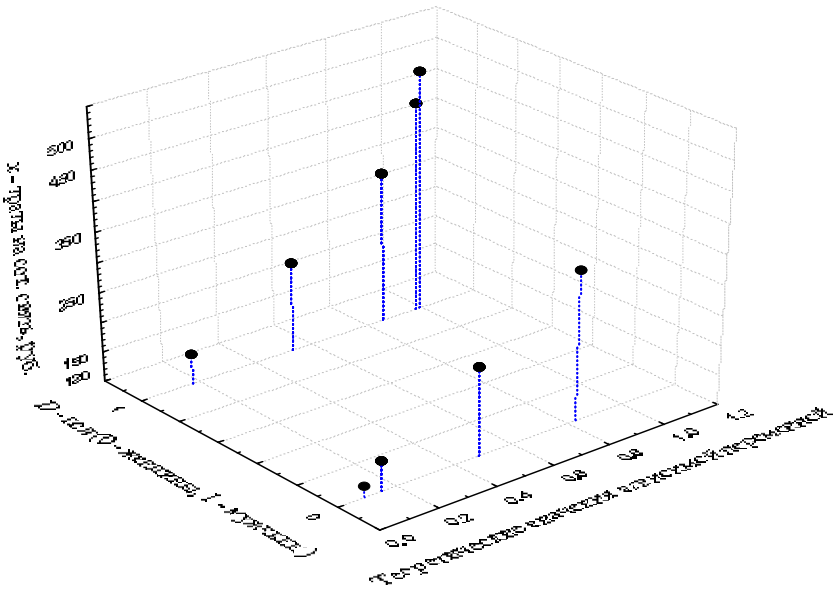


Рисунок 6.5 – Результаты моделирования регрессии факторов
на качество услуг сотовой связи в г. Оренбурге

Согласно приведенным на графике данным, женщины оценивают качество слуг связи ниже, при этом и тратят на связь они меньше.

6.5 Тест Чоу

В эконометрических исследованиях часто встречаются случаи, когда выборка наблюдений состоит из двух и более подвыборок, поэтому сложно установить, оценивать одну объединенную регрессию или проводить оценивание регрессии по каждой подвыборке. Прежде чем вводить фиктивные переменные, принято проверять выборочную совокупность на однородность.

Допустим, что наша выборка состоит из двух однородных частей, одна из которых объёмом n_1, а другая - n_2. Если n_1, n_2 значительно больше, чем $k+1$, то для каждой из таких подвыборок мы можем построить регрессионную модель. Если окажется, что оценки коэффициентов для одной однородной группы входят в доверительные интервалы для другой группы, то делается вывод о регрессионной одно-

родности выборочной совокупности и переходят к построению оценок на основе объединённой выборки объёмом $n_1 + n_2$.

Рассмотрим другую ситуацию, когда объём одной из подвыборок, например, второй, меньше или равен $k+1$. В этом случае вторая подвыборка не позволяет построить уравнение регрессии для однородной группы и для проверки гипотезы об однородности выборочной совокупности используется *тест Чоу*[1].

Алгоритм метода Чоу включает следующие шаги:

1) выдвигаются гипотезы вида:

$- H_0 : \overline{\beta}^{(1)} = \overline{\beta}^{(2)}; \sigma_\varepsilon^2(1) = \sigma_\varepsilon^2(2);$

$- H_1 : \overline{\beta}^{(1)} \neq \overline{\beta}^{(2)}; \sigma_\varepsilon^2(1) \neq \sigma_\varepsilon^2(2).$

1) строится общее уравнение регрессии по всем *n* наблюдениям:

$$y_i = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + \varepsilon_i, \quad i = 1,2,...,n; \tag{6.11}$$

2) рассчитывается сумма квадратов отклонений фактических значений от расчетных по полученному уравнению:

$$s_0 = \sum e_i^2 = \sum (y_i - \tilde{y}_i)^2 ; \tag{6.12}$$

3) общая выборка разбивается на две подвыборки объемами n_1 и n_2 соответственно ($n_1 + n_2 = n$). Для каждой из них строится уравнение регрессии:

$$y_{1i} = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + \varepsilon_i, \quad i = 1,2,...,n_1, \tag{6.13}$$

$$y_{2i} = b_0 + b_1 x_{1i} + b_2 x_{2i} + ... + b_k x_{ki} + \varepsilon_i, \quad i = n_1 + 1, n_1 + 2,...,n_2; \tag{6.14}$$

[1] Этот тест был назван так по имени своего создателя Г. Чоу (Chow, 1960), но приводимая здесь интерпретация теста была предложена в 1985 г. Х. Песараном, Р. Смитом и С. Ео.

4) рассчитываются суммы квадратов отклонений фактических значений y_i каждой из подвыборок от соответствующих уравнений регрессии:

$$s_1 = \sum_{i=1}^{n_1} e_i^2 = \sum_{i=1}^{n_1} (y_{1i} - \tilde{y}_{1i})^2 , \tag{6.15}$$

$$s_2 = \sum_{i=n_1+1}^{n_2} e_i^2 = \sum_{i=n_1+1}^{n_2} (y_{2i} - \tilde{y}_{2i})^2 . \tag{6.16}$$

Равенство $s_0 = s_1 + s_2$ возможно лишь при совпадении коэффициентов регрессии для всех уравнений. Чем сильнее различие в поведении Y для двух подвыборок, тем больше значение s_0 будет превосходить $s_1 + s_2$. Тогда разность

$$s_0 - (s_1 + s_2) \tag{6.17}$$

может быть интерпретирована как улучшение качества модели при разбиении интервала наблюдений на два подынтервала. Следовательно, дробь

$$\frac{s_0 - (s_1 + s_2)}{k+1} \tag{6.18}$$

определяет оценку уменьшения дисперсии регрессии за счет построения двух уравнений вместо одного. При этом число степеней свободы сократиться на $(k+1)$, т.к. вместо $(k+1)$ параметра объединенного уравнения теперь необходимо оценивать $(2k+2)$ параметра двух регрессий.

Дробь

$$\frac{(s_1 + s_2)}{n - 2k - 2} \tag{6.19}$$

- это необъясненная дисперсия зависимой переменной при использовании двух регрессий. Отсюда следует, что общую выборку целесообразно разбить на два подынтервала только в случае, если уменьшение дисперсии будет значимо больше оставшейся необъясненной дисперсии. Для определения, является ли значимым улучшение качества уравнения после разделения выборки, строится F – статистика, которая распределена с ($k+1$) и ($n-2k-2$) степенями свободы;

4) рассчитываем F – статистику по формуле

$$F = \frac{s_0 - s_1 - s_2}{s_1 + s_2} \cdot \frac{n - 2k - 2}{k + 1}; \qquad (6.20)$$

5) сравниваем расчетное значение F – статистики с ее критическим значением. Если $F_{набл} > F_{крит(\alpha;k+1;n-2k-2)}$, то основная гипотеза отклоняется, и качество частных регрессионных моделей превосходит качество общей модели регрессии (разбиение на подынтервалы имеет смысл). Здесь k - число количественных объясняющих переменных в уравнении регрессии (одинаково для всех трех уравнений). Если $F_{набл} < F_{крит(\alpha;k+1;n-2k-2)}$, то основная гипотеза принимается, и разбивать общую регрессию на подвыборки не имеет смысла.

Тест Чоу достаточен, если требуется только установить, что зависимости в подвыборках различаются. Оценивание регрессии с фиктивными переменными более информативно, т.к. позволяет рассмотреть вклад каждой фиктивной переменной, а также всей группы в целом.

6.6 Вопросы для самоконтроля

1. Дайте понятие фиктивной переменной. В каких случаях в эконометрических исследованиях используются фиктивные переменные?

2. Опишите схему регрессионного анализа пространственной совокупности с применением бинарных переменных.

3. Как называется и что характеризует параметр (коэффициент) при фиктивной переменной?

4. Каким образом проводится регрессионный анализ, если фиктивная переменная принимает более двух значений?

5. В чем особенность построения регрессионных моделей, у которых фиктивная переменная находится в левой части уравнения?

6. С какой целью проводится тест Чоу? Каковы этапы его проведения?

6.7 Тесты

1. В регрессионных моделях влияние качественного фактора выражается в виде:

а) фиктивной переменной;

б) эндогенной переменной;

в) лаговой переменной.

2. Фиктивные переменные в модели могут выступать в роли

а) только фактора;

б) только результата;

в) как фактора, так и результата.

3. Если качественная переменная имеет k альтернативных значений, то при моделировании используются:

а) ($k-1$) фиктивных переменных;

б) k фиктивных переменных;

в) ($k+1$) фиктивных переменных.

4. Оценка значимости параметров уравнения регрессии с фиктивными переменными осуществляется на основе:

а) *t* - критерия Стьюдента;

б) *F* - критерия Фишера – Снедекора;

в) средней квадратической ошибки;

г) средней ошибки аппроксимации.

5. В уравнении регрессии $\hat{y}_i = 56{,}6 - 21{,}6D_{1i} - 10{,}1D_{2i}$ (*y* – процент рабочих занятых ручным трудом в общей численности рабочих, D_1 = 1 для предприятий с высоким уровнем автоматизации производства, D_2 = 1 для предприятий со средним уровнем автоматизации производства) параметр при D_1 показывает, что

а) на предприятиях с низким уровнем автоматизации производства средний процент рабочих, занятых ручным трудом равен 21,6;

б) на предприятиях с высоким уровнем автоматизации производства распространенность ручного труда ниже на 21,6 п.п. по сравнению с предприятиями с низким уровнем автоматизации производства;

в) на предприятиях с высоким уровнем автоматизации производства средний процент рабочих, занятых ручным трудом равен 21,6.

6. Фиктивные переменные вводятся в:

а) только в линейные модели;

б) только в модели множественной нелинейной регрессии;

в) только в нелинейные модели;

г) как в линейные, так и в нелинейные модели, приводимые к линейному виду.

7. Тест Чоу основан на сравнении:

а) дисперсий;

б) коэффициентов детерминации;

в) математических ожиданий;

г) средних.

8. Если в тесте Чоу $F_{набл} > F_{крит}$, то считается, что

а) разбиение на подынтервалы целесообразно с точки зрения улучшения качества модели;

б) модель является статистически незначимой;

в) модель является статистически значимой;

г) нет смысла разбивать выборку на части.

7 Системы эконометрических регрессионных уравнений

Что необходимо знать из 7 главы:

1. Понятие, общий вид и классы систем эконометрических регрессионных уравнений.

2. Задачи исследования и структурная форма системы одновременных уравнений.

3. Преобразования, используемые для получения приведенной формы СОУ.

4. Понятие, необходимые и достаточные условия идентификации структурной формы СОУ.

5. Методы оценивания параметров структурной модели СОУ.

7.1 Понятие и анализ проблемы решения системы регрессионных уравнений

В реальных явлениях и процессах экономики так называемые «результативные» признаки (производительность труда, фондоотдача, себестоимость, прибыль, рентабельность и т.д.) не изолированы, а также взаимосвязаны друг с другом, поэтому нужен *системный подход* к экономике, учет и эконометрическое моделирование не отдельных показателей, а целых систем взаимосвязанных показателей.

При этом в одних уравнениях факторная переменная рассматривается как объясняющая (независимая), но в тоже время она входит в другое уравнение как зави-

симая, объясняемая переменная. Другими словами значения объясняемых и объясняющих переменных формируются одновременно под воздействием некоторых внешних факторов. Поэтому система таких уравнений получила название *система одновременных уравнений* (СОУ).

Предположим, изучается модель спроса как соотношение цен и количества потребляемых товаров. Одновременно для прогнозирования спроса необходима модель предложения товаров, в которой рассматривается также взаимосвязь между количеством и ценой предлагаемых благ. Это позволяет достичь равновесия между спросом и предложением.

Другой пример: при оценке эффективности производительности нельзя руководствоваться только моделью рентабельности. Она должна быть дополнена моделью производительности труда, а также моделью себестоимости единицы продукции. Потребность в использовании системы одновременных уравнений возрастает, если мы переходим от исследований на микроуровне к макроэкономическим расчетам.

При рассмотрении СОУ переменные делятся на два больших класса: эндогенные (X) и экзогенные переменные (Y). Э. Ферстер и Б. Рейнц дают следующие определения переменным этих классов [21, с. 245-246].

Эндогенными (зависимыми, внутренними) переменными являются экономические величины, которые объясняются эконометрической моделью. Значения эндогенных переменных формируются в результате одновременного взаимодействия переменных, образующих модель. Эндогенные переменные зависят от экзогенных и возмущающих переменных. То есть значения эндогенных переменных формируются в процессе функционирования анализируемой системы под воздействием экзогенных переменных и во взаимодействии друг с другом

Значения экзогенных (независимых, внешних) переменных в каждый период времени t определяются вне модели. Экзогенные переменные являются внешними, наперед заданными экономическими величинами. Они, следовательно, объясняются не моделью, а экономическими факторами и закономерностями, лежащими за границами этой модели. Экзогенные переменные определяют эндогенные переменные,

но сами не находятся под их влиянием. Таким образом, между эндогенными и экзогенными переменными существуют только односторонние стохастические причинные отношения.

Вопрос, какие переменные следует рассматривать как экзогенные, решается, прежде всего на основе детального анализа экономического явления. Экзогенными переменными могут быть природные, технические, демографические и некоторые социальные факторы. В связи с тем, что регрессионной моделью нельзя охватить весь причинно-следственный комплекс явлений в экономике, исследователь вынужден выделять только определенную часть связей, отдавая предпочтение наиболее существенным. Неучтенными остаются некоторые влияющие величины, которые не объясняются моделью, или сила их взаимосвязей так мала, что ими пренебрегают. Такие переменные можно также отнести к экзогенным. Деление переменных на экзогенные и эндогенные относительно. Оно зависит от природы изучаемого явления, а также от цели, с которой эта модель строится.

Кроме перечисленных, в СОУ выделяют лаговые (запаздывающие) и предопределенные переменные. Лаговые переменные – это экзогенные и эндогенные переменные, датированные предыдущими моментами времени (X_{t-1}, Y_{t-1}). В связи с тем, что лаговые переменные в момент времени *t* также не объясняются эконометрической моделью, мы можем отнести их к заранее заданным экзогенным.

Предопределенными переменными выступают:

– текущие экзогенные переменные (они объясняются не эконометрической моделью, а факторами вне этой модели);

- лаговые экзогенные переменные (X_t, X_{t-1}) (их значения принадлежат предыдущим периодам и не объясняются данной моделью);

- лаговые эндогенные переменные (Y_{t-1}) (их предопределенность следует из предшествующего объяснения в эконометрической модели).

Порядок оценивания параметров системы эконометрических уравнений имеет свои особенности. Это связано с тем, что в регрессионных уравнениях системы независимые переменные и случайные погрешности оказываются коррелированы между собой.

Рассмотрим ситуацию, где исследование одного объекта характеризуется m эндогенными признаками и k предопределенными признаками в разные моменты времени t.

В общем виде система одновременных уравнений выглядит следующим образом:

$$\begin{cases} \beta_{11} y_{1t} + \beta_{12} y_{2t} + ... + \beta_{1m} y_{mt} + C_{11} x_{1t} + C_{12} x_{2t} + ... + C_{1k} x_{kt} = \delta_{1t}; \\ \beta_{21} y_{1t} + \beta_{22} y_{2t} + ... + \beta_{2m} y_{mt} + C_{21} x_{1t} + C_{22} x_{2t} + ... + C_{2k} x_{kt} = \delta_{2t}; \\ ..... \\ \beta_{m1} y_{1t} + \beta_{m2} y_{2t} + ... + \beta_{mm} y_{mt} + C_{m1} x_{1t} + C_{m2} x_{2t} + ... + C_{mk} x_{kt} = \delta_{mt}. \end{cases} \quad (7.1)$$

где: $t = 1,2...n$;

$y_{1t}, ... y_{mt}$ - эндогенные переменные, измеренные в момент времени t;

$x_{1t}, ... x_{kt}$ - предопределенные переменные, отнесенные к моменту времени t;

β_{ij} $_{\substack{i=\overline{1,m} \\ j=\overline{1,m}}}$ - неизвестные коэффициенты при эндогенных переменных структурной формы СОУ;

c_{il} $_{\substack{i=\overline{1,m} \\ l=\overline{1,k}}}$ - неизвестные коэффициенты при предопределенных переменных структурной формы СОУ;

$\delta_{1t} ... \delta_{mt}$ - регрессионные остатки в момент времени t компоненты которого некоррелированы между собой и для разных t, гомоскедастичны при каждом t.

Задачей исследования системы уравнений является определение оценок параметров уравнений на основе исходных данных обо всех признаках по совокупности (выборке) предприятий, фирм, регионов и других экономических объектов, которые (оценки) были бы надежными, несмещенными и состоятельными. Итак, можно дать следующее определение: система эконометрических регрессионных уравнений – это такое множество уравнений, в котором одни и те же эндогенные переменные в одних уравнениях являются результативными признаками, а в других уравнениях играют роль факторов наряду с факторными эндогенными переменными.

Введем следующие обозначения:

- $y_t = (y_{1t}, ... y_{mt})^T$ - вектор эндогенных переменных, измеренных в момент времени *t*;

- $x_t = (x_{1t}, ... x_{kt})^T$ - вектор предопределенных переменных, отнесенных к моменту времени *t*;

- $\underset{m\times m}{B} = \begin{pmatrix} \beta_{11} & \beta_{12} & ... & \beta_{1m} \\ \beta_{21} & \beta_{22} & ... & \beta_{2m} \\ ... & ... & ... & ... \\ \beta_{m1} & \beta_{m2} & ... & \beta_{mm} \end{pmatrix}$ - матрица неизвестных коэффициентов при эндогенных переменных структурной формы СОУ, причем $\det B \neq 0$;

- $\underset{m\times k}{C} = \begin{pmatrix} c_{11} & c_{12} & ... & c_{1k} \\ c_{21} & c_{22} & ... & c_{2k} \\ ... & ... & ... & ... \\ c_{m1} & c_{m2} & ... & c_{mk} \end{pmatrix}$ - матрица неизвестных коэффициентов при предопределенных переменных структурной формы СОУ;

- $\delta_t = (\delta_{1t} ... \delta_{mt})^T$ - вектор регрессионных остатков в момент времени *t*, компоненты которого некоррелированы между собой и для разных *t*, гомоскедастичны при каждом *t*.

Тогда СОУ в матричном виде:

$$BY_t + CX_t = \delta_t . \qquad (7.2)$$

Будем предполагать, что коэффициент при $i-ой$ эндогенной переменной равен 1, т.е. β_{ii}=1. Это позволяет каждое уравнение СОУ представить в виде:

$$y_{it} = -\beta_{i1} y_{1t} ... - \beta_{ii-1} y_{i-1t} - \beta_{ii+1} y_{i+1t} - ... - \beta_{im} y_{mt} - C_{i1} x_{1t} - ... - C_{ik} x_{kt} + \varepsilon_{it}, \quad t = \overline{1,n} . \qquad (7.3)$$

Система одновременных уравнений (7.1) и (7.2) называется структурной формой СОУ.

Рассмотрим граф связей для трех эндогенных переменных, наглядно показывающий направление влияния вариации признака-причины к признаку-следствию в виде стрелки (рисунок 7.1).

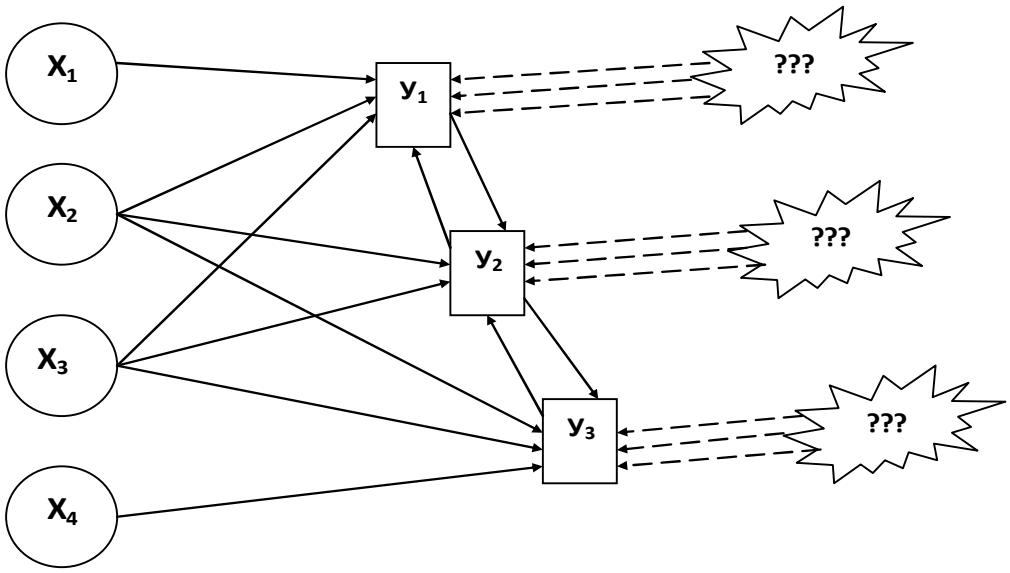


Рисунок 7.1 – Граф связей

На результативный признак y_1 влияют экзогенные факторы x_1; x_2; x_3, а также эндогенный признак y_2. На y_2, в свою очередь, влияют экзогенные переменные x_2; x_3, а также эндогенные признаки y_1 и y_3. На y_3 влияют экзогенные факторы x_2, x_3, x_4, а также эндогенный признак y_2. Кроме того, на каждый эндогенный признак влияет ряд *неизвестных* факторов, изображенных в виде «облака» с пунктирными стрелками [10].

Та же система связей может быть записана и в форме структурных уравнений, выражающих структуру связей:

$$\begin{cases} y_1 = a_1 + b_{11}x_1 + b_{12}x_2 + b_{13}x_3 + c_{12}y_2 + \varepsilon_1; \\ y_2 = a_2 + b_{22}x_2 + b_{23}x_3 + c_{21}y_1 + c_{23}y_3 + \varepsilon_2; \\ y_3 = a_3 + b_{32}x_2 + b_{33}x_3 + b_{34}x_4 + c_{32}y_2 + \varepsilon_3. \end{cases} \quad (7.4)$$

Первый индекс при коэффициентах регрессии - это номер уравнения или эндогенной переменной, стоящей в левой части уравнения, второй индекс - номер фактора. Коэффициенты при экзогенных переменных обозначены «*b*», а при эндогенных переменных – «*c*».

Очевидно, что проблема решения заключена в тех эндогенных переменных, которые входят в правую часть уравнений, то есть как бы «занимают чужое место» в том «порочном круге», который мы видим на графе связей (рисунок 7.1): чтобы определить y_1, нужно знать y_2, но чтобы определить y_2 нужно знать y_1.

Для определения параметров первого структурного уравнения можно было бы ввести значения $x_1; x_2; x_3; y_2$ для каждой единицы совокупности и решить, как обычное регрессионное уравнение. Но проблема в неизвестных факторах («облаке»), влияющих на вариацию y_2 . Если мы, решая первое уравнение, будем исходить, кроме известных экзогенных переменных $x_1; x_2; x_3$ и из фактических значений y_{2i} для каждой единицы совокупности, то получим оценки параметров $a_1; b_{11}; b_{12}; b_{13}$ и c_{12}, *зависящие также и от неизвестных факторов* y_{2i}. Представим себе, что y_1 - это народнохозяйственная производительность труда (ВВП на одного занятого в экономике), а y_2 - среднедушевой доход. На этот среднедушевой доход влияет, в числе многих факторов и такой, как среднее число детей в семье. Однако недопустимо, чтобы этот фактор влиял на вариацию производительности труда, это экономически абсурдно. То, что y_2 влияет на вариацию y_1, еще не значит, что *любой* фактор вариации y_2 должен влиять на вариацию y_1 и наоборот. Такова логическая сторона проблемы.

Математическая ее сторона связана с тем, что параметры уравнений регрессии оцениваются с помощью метода наименьших квадратов (МНК). Данный метод, как мы уже рассматривали, дает несмещенные и состоятельные оценки лишь при соблюдении ряда условий, в том числе условия гомоскедастичности - отклонения фактических значений результативного признака теоретических должны быть постоянными, независимыми от расчетных величин и от величин факторов. Вспомним определение корреляционной зависимости: от значений фактора должны зависеть средние значения результативного признака, но не его вариация (σ_{y_j}). Если возникнет корреляция, связь между отклонениями ($y_i - \tilde{y}_i$) и неизвестными факторами,

влияющими на другую (другие) эндогенную переменную, входящую в правую часть уравнения, МНК применять нельзя, некорректно. Поскольку значения факторов «из облака» неизвестны, нельзя исключить, что такая связь возникнет, или, как часто пишут в учебниках эконометрики, появится «корреляция с ошибками», нарушающая предпосылки применения МНК.

Каким образом можно избавиться от «корреляции с ошибками»? Очевидно, в левой части уравнения должна стоять эндогенная переменная в ее фактических значениях, а в правой части либо нужно вообще исключить эндогенные переменные, либо они должны присутствовать в «очищенном виде», *как расчетные значения из уравнений только с известными экзогенными переменными* [10].

7.2 Приведенная форма системы одновременных уравнений

Приведенными называются уравнения, полученные из структурных путем подстановки взамен эндогенной переменной в правую часть уравнения ее выражения из другого структурного уравнения, в котором эта эндогенная переменная находится в левой части. После такой подстановки производят преобразования, при которых члены уравнения, содержащие эндогенную переменную, собирают в левую часть, а в правой части остаются только экзогенные переменные.

Пусть для изучения структурной формы СОУ проведены наблюдения в момент времени $t = \overline{1,n}$. Матрица наблюденных значений эндогенных переменных будет иметь вид:

$$\underset{m\times n}{Y} = \begin{pmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ y_{m1} & y_{m2} & \cdots & y_{mn} \end{pmatrix}. \qquad (7.5)$$

Значения предопределенных переменных:

$$\underset{k\times n}{X}=\begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \dots & \dots & \dots & \dots \\ x_{k1} & x_{k2} & \dots & x_{kn} \end{pmatrix}. \tag{7.6}$$

Значения регрессионных остатков:

$$\underset{m\times n}{\Delta}=\begin{pmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1n} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2n} \\ \dots & \dots & \dots & \dots \\ \delta_{m1} & \delta_{m2} & \dots & \delta_{mn} \end{pmatrix}. \tag{7.7}$$

Тогда модель (7.1) и (7.2) для всех моментов времени будет иметь вид:

$$BY + CX = \Delta\,. \tag{7.8}$$

Если предположить, что матрица B невырожденная $\det B \neq 0;$ то, умножив обе части системы (7.1) слева на B^{-1}, получим:

$$BY_t + CX_t = \delta_t\,, \tag{7.9}$$

$$B^{-1}BY_t + B^{-1}CX_t = B^{-1}\delta_t\,,\quad Y_t = -B^{-1}CX_t + B^{-1}\delta_t\,. \tag{7.10}$$

Обозначим $\pi = \left\{\pi_{ij}\right\}_{\substack{i=1,m\\ j=1,k}} \equiv -B^{-1}C$ или $B\pi = -C$; $\varepsilon_t = B^{-1}\delta_t$ - вектор регрессионных остатков приведенной формы. Тогда:

$$Y_t = \pi X_t + \varepsilon_t\,. \tag{7.11}$$

Модель (7.11) называется приведенной формой СОУ.

Приведенная форма СОУ представляет собой систему линейных функций эндогенных переменных от экзогенных:

$$\begin{cases} y_{1t} = \pi_{11} x_{1t} + \pi_{12} x_{2t} + ... + \pi_{1k} x_{kt} + \varepsilon_{1t}; \\ y_{2t} = \pi_{21} x_{1t} + \pi_{22} x_{2t} + ... + \pi_{2k} x_{kt} + \varepsilon_{2t}; \\ ..... \\ y_{mt} = \pi_{m1} x_{1t} + \pi_{m2} x_{2t} + ... + \pi_{mk} x_{kt} + \varepsilon_{mt}. \end{cases}, \qquad (7.12)$$

где π - неизвестные коэффициенты приведенной формы СОУ.

Оценив построчно матрицу приведенной формы из ЛММР, получим:

$$y_{it} = \pi_{i1} x_{1t} + \pi_{i2} x_{2t} + ... + \pi_{ik} x_{kt} + \varepsilon_{it}, \; t = \overline{1,n}; \; i = \overline{1,m}. \qquad (7.13)$$

В приведенной форме СОУ неизвестным является параметр π. Уравнение представляет собой обычную модель множественной регрессии, где вместо i стоит t, а вместо β стоят π. Применив метод наименьших квадратов или обобщенный метод наименьших квадратов, мы можем найти элементы матрицы.

После оценки матрицы π, нам необходимо найти матрицы B *и* C. Здесь мы сталкиваемся с проблемой: для оценки матриц B *и* C имеем систему $\pi = -B^{-1}C$ или $B\pi = -C$, в которой число неизвестных (m^2+mk) больше, чем число уравнений (mk). Следовательно, данная система (т.е. задача, связанная с нахождением B *и* C по элементам матрицы π) неразрешима. Однако практика эконометрического моделирования показывает, что матрицы B и C имеют достаточно разреженную структуру (большое количество нулевых элементов) и поэтому возможны ситуации, в которых система будет иметь решение.

Преобразование структурных уравнений в приведенные рассмотрим на простейшем примере (будем считать, что переменные Y центрированы, т.е. a=0).

Пример 7.1 - Модель, описывающая зависимость спроса и предложения (y_{1t}) некоторого товара от его цены и дохода в условиях равновесия:

$$y_{1t} = \beta_1 y_{2t} + \varepsilon_{1t} \text{ (предложение)},$$

$$y_{1t} = \beta_2 y_{2t} + c_1 x_{1t} + \varepsilon_{2t} \text{ (спрос)}, \qquad (7.14)$$

где y_{1t} – спрос (предложение);

y_{2t} – цена переменные;

x_{1t} – доход;

β_1, β_2, c_1 - неизвестные коэффициенты, которые подлежат определению.

Перейдем к приведенной форме. Для этого умножим первое уравнение на β_2, а второе уравнение на β_1.

$$\begin{cases} \beta_2 y_{1t} = \beta_2 \beta_1 y_{2t} + \beta_2 \varepsilon_{1t}; \\ \beta_1 y_{1t} = \beta_1 \beta_2 y_{2t} + \beta_1 c_1 x_{1t} + \beta_1. \end{cases} \qquad (7.15)$$

Далее из второго уравнения вычтем первое:

$$y_{1t}(\beta_1 - \beta_2) = \beta_1 \beta_2 y_{2t} - \beta_2 \beta_1 y_{2t} + \beta_1 c_1 x_{1t} - \beta_2 \varepsilon_{1t} + \beta_1 \varepsilon_{2t},$$

$$y_{1t}(\beta_1 - \beta_2) = \beta_1 c_1 x_{1t} - \beta_2 \varepsilon_{1t} + \beta_1 \varepsilon_{2t}. \qquad (7.16)$$

Выразим

$$y_{1t} = \frac{\beta_1 c_1}{\beta_1 - \beta_2} x_{1t} + \frac{\beta_1 \varepsilon_{2t} - \beta_2 \varepsilon_{1t}}{\beta_1 - \beta_2}. \qquad (7.17)$$

Обозначив $\frac{\beta_1 \varepsilon_{2t} - \beta_2 \varepsilon_{1t}}{\beta_1 - \beta_2}$ через δ_{1t}, получим первое уравнение приведенной формы:

$$y_{1t} = \pi_{11} x_{1t} + \delta_{1t}. \qquad (7.18)$$

Аналогичным образом проведем преобразования для первого уравнения. Получим:

$$y_{1t} - y_{1t} = \beta_1 y_{2t} - \beta_2 y_{2t} + \varepsilon_{1t} - c_1 x_{1t} - \varepsilon_{2t}, \tag{7.19}$$

$$\beta_1 y_{2t} - \beta_2 y_{2t} + \varepsilon_{1t} - c_1 x_{1t} - \varepsilon_{2t} = 0, \tag{7.20}$$

$$y_{2t}(\beta_1 - \beta_2) + \varepsilon_{1t} - c_1 x_{1t} - \varepsilon_{2t} = 0, \tag{7.21}$$

$$y_{2t}(\beta_1 - \beta_2) = -\varepsilon_{1t} + \varepsilon_{2t} + c_1 x_{1t}, \tag{7.22}$$

$$y_{2t} = \frac{c_1}{\beta_1 - \beta_2} x_{1t} + \frac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \beta_2}. \tag{7.23}$$

Обозначив $\frac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \beta_2}$ через δ_{2t}, получим второе уравнение приведенной формы:

$$y_{2t} = \pi_{21} x_{1t} + \delta_{2t}. \tag{7.24}$$

Приведенная форма системы одновременных уравнений имеет вид:

$$\begin{cases} y_{1t} = \pi_{11} x_{1t} + \delta_{1t}; \\ y_{2t} = \pi_{21} x_{1t} + \delta_{2t}. \end{cases} \tag{7.25}$$

где $\pi_{11} = \frac{\beta_1 c_1}{\beta_1 \beta_2}$;

$\pi_{21} = \frac{c_1}{\beta_1 - \beta_2}$;

$\delta_{1t} = \frac{\beta_1 \varepsilon_{2t} - \beta_2 \varepsilon_{1t}}{\beta_1 - \beta_2}$;

$$\delta_{2t} = \frac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \beta_2}.$$

Уравнения системы (7.25) приведены «к решаемому с помощью МНК виду»: не содержат в правой части эндогенных переменных вместе с их ошибками. Любая система приведенных уравнений может быть решена при соблюдении условий, общих для регрессионно - корреляционного анализа.

Однако решение приведенных уравнений не является конечной целью изучения системы эконометрических уравнений. Целью является решение структурных уравнений, отображающих реальную систему связи признаков в экономике. Каким образом от решения приведенных уравнений перейти к структурным уравнениям и всегда ли возможен такой переход, рассмотрим на следующем этапе анализа - идентификации системы уравнений.

7.3 Идентификация системы уравнений

Слово «идентификация» хотя и неточно, можно выразить русскими словами: узнавание, опознание, установление единства. На этом этапе устанавливается: едины или нет приведенные уравнения со структурными; можно ли по коэффициентам приведенных уравнений опознать, вычислить коэффициенты структурных уравнений [10]. Другими словами, под *идентификацией*[1] понимается возможность численной оценки параметров структурной формы по оценкам коэффициентов приведенной формы.

Из курса математики известно, что не любая система уравнений имеет решение. Невозможно, например, решить систему двух уравнений с тремя и более неизвестными.

[1] Более подробно с проблемами идентификации можно ознакомиться в главе 16 Green W.H. (1997). Econometric Analysis, 3rd edition. Prentice-Hall, Upper Saddle River, New Jersey.

С другой стороны, система трех уравнений с двумя переменными может быть решена, однако имеет не одно, а, по крайней мере, три решения, то есть опять же определенного ответа не имеет. Сходная ситуация имеет место и в вопросе об идентификации системы структурных уравнений по приведенным. Однозначное решение, то есть *точную идентификацию* имеет такая система, в которой число коэффициентов регрессии приведенных уравнений точно равно числу коэффициентов регрессии структурных уравнений.

Условие точной идентификации может быть выражено разными способами на разных стадиях решения задачи, не обязательно требуется уже иметь коэффициенты приведенных уравнений. Так, уже по виду графа связей можно провести идентификацию, сравнив число входящих связей от эндогенных переменных с числом отсутствующих связей от тех экзогенных переменных, которые входят в другие эндогенные, влияющие на данную. Если эти числа равны, то уравнение данной эндогенной точно идентифицируемое.

На рисунке 7.2 мы видим все три возможных результата идентификации.

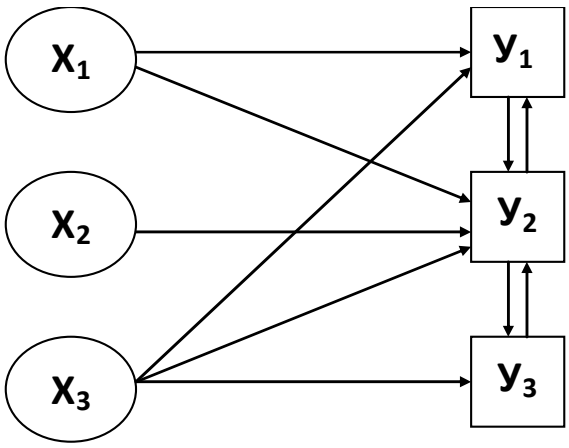


Рисунок 7.2 - Три результата идентификации

Поскольку все эндогенные переменные взаимосвязаны, в приведенных уравнениях каждой эндогенной будет три экзогенных. На y_1 влияет непосредственно две экзогенных и одна эндогенная переменные.

Условие точной идентификации соблюдено. На y_2 влияют непосредственно все три экзогенных и две эндогенные (в структурном уравнении $\tilde{y}_2$ будет пять коэффициентов регрессии). Из трех коэффициентов приведенного уравнения невозможно определить пять неизвестных коэффициентов структурного уравнения – оно *неидентифицируемо, неразрешимо.* На y_3 непосредственно влияет одна экзогенная и одна эндогенная, в структурном уравнении будет два коэффициента. Из трех коэффициентов приведенного уравнения можно получить три разных оценки коэффициентов структурного уравнения – оно *сверхидентифицируемое*. Его однозначное решение возможно другим методом [10].

Система уравнений в целом идентифицируется по «худшему» из уравнений – в данном примере *система неидентифицируемая*. Если в системе есть точно- и сверхидентифицируемые уравнения, то система признается сверхидентифицируемой и решается как таковая.

Исходя из вышесказанного, выведем основные определения.

Определение 1. Уравнение структурной формы СОУ называется точно идентифицируемым, если его коэффициенты однозначно определяются по оценкам коэффициентов приведенной формы СОУ.

Определение 2. Уравнение структурной формы СОУ называется неидентифицируемым, если его коэффициенты нельзя определить по оценкам коэффициентов приведенной формы СОУ.

Определение 3. Уравнение структурной формы СОУ называется сверхидентифицируемым, если его коэффициенты оцениваются по оценкам коэффициентов приведенной формы СОУ не единственным образом.

Рассмотрим необходимые и достаточные условия идентификации (применяются только к структурной форме СОУ).

Введем обозначения:

- m - количество эндогенных переменных в системе;

- k - количество экзогенных переменных в системе;

- m_i - количество эндогенных переменных в $i-ом$ уравнении, проверяемом на идентифицируемость (причем $m_i \le m$) ;

- k_i - количество экзогенных переменных всего в $i-ом$ уравнении, проверяемом на идентифицируемость (причем $k_i \le k$).

1. Уравнение системы идентифицируемо в том случае, если число эндогенных переменных равно числу регрессионных уравнений, т.е. матрица B - квадратная (причем $|B| \neq 0$).

2. Ранг матрицы X равен количеству экзогенных переменных в системе ($rangX = k$).

3. Каждому уравнению структурной формы ставится в соответствие вектор-строка из ($m+k$) элементов (вектор исключающих априорных ограничений).

$$I\ элемент\begin{cases}1, если\ первая\ эндогенная\ переменная\ присутствует\ в\ уравнении;\\0, если\ она\ отсутствует.\end{cases}$$

$$II\ элемент\begin{cases}1, если\ вторая\ эндогенная\ переменная\ присутствует\ в\ уравнении;\\0, если\ она\ отсутствует.\end{cases}$$

…….……..……

$$"m"\ элемент\begin{cases}1, если\ m-я\ эндогенная\ переменная\ присутствует\ в\ уравнении;\\0, если\ она\ отсутствует.\end{cases}$$

$$"m+1"\ элемент\begin{cases}1, если\ первая\ экзогенная\ переменная\ присутствует\ в\ уравнении;\\0, если\ она\ отсутствует.\end{cases}$$

……...

$$"m+k"\ элемент\begin{cases}1, если\ k-я\ экзогенная\ переменная\ присутствует\ в\ уравнении;\\0, если\ она\ отсутствует.\end{cases}$$

Если среди векторов исключающих априорных ограничений нет одинаковых, то это является необходимым условием идентифицируемости системы.

4. Данное условие относится не ко всей системе, а к каждому уравнению системы в отдельности.

Всего в системе m эндогенных и k экзогенных переменных, в $i-ом$ уравнении присутствуют m_i эндогенных переменных и k_i экзогенных переменных. Перенумеруем их таким образом, чтобы в первых m_i позициях стояли эндогенные пере-

менные, в первых k_i позициях – экзогенные переменные. Будем считать, что в $i-ом$ уравнении у нас присутствуют именно первые m_i эндогенные переменные и первые k_i экзогенные переменные. Введем в рассмотрение вектор-строку:

$$\underset{1\times m}{B(i)} = (B_{i1}, B_{i2}, ..., B_{im}, 0, ..., 0), \tag{7.26}$$

$$\underset{1\times k}{C(i)} = (C_{i1}, C_{i2}, ..., C_{ik}, 0, ..., 0), \tag{7.27}$$

$$B(i)Y_t + C(i)X_t = \Delta. \tag{7.28}$$

В соответствии с перенумерацией, представим матрицу π в блочном виде:

$$\underset{m\times k}{\pi} = \begin{pmatrix} \underset{m_i\times k_i}{\pi(i)} & \underset{m_i\times k-k_i}{\pi(i)} \\ \underset{m-m_i\times k_i}{\pi_y(i)} & \underset{m-m_i\times k_i}{\pi_{xy}(i)} \end{pmatrix}. \tag{7.29}$$

Вспомним, как связаны матрицы B и π: $\pi = -B^{-1}C$ или $B\pi = -C$. Тогда:

$$(B_{i1}, B_{i2}, ..., B_{im}, 0, ..., 0 \times \begin{pmatrix} \underset{m_i\times k_i}{\pi(i)} & \underset{m_i\times k-k_i}{\pi(i)} \\ \underset{m-m_i\times k_i}{\pi_y(i)} & \underset{m-m_i\times k_i}{\pi_{xy}(i)} \end{pmatrix} = -(C_{i1}, C_{i2}..., C_{ik}, 0, ..., 0). \tag{7.30}$$

$$\begin{cases} B(i)\times\pi(i) = -C(i); \\ B(i)\times\pi_x(i) = 0. \end{cases} \tag{7.31}$$

Получили систему линейных уравнений, где $(k-k_i)$ – число уравнений, (m_i-1) – число неизвестных.

Необходимое условие идентифицируемости $i-ого$ уравнения системы: $k-k_i \geq m_i - 1$ (количество исключенных из уравнения экзогенных переменных должно

быть не меньше количества эндогенных переменных в этом уравнении, уменьшенного на единицу) – в этом случае система (7.31) будет иметь решение.

5. Необходимое и достаточное условие идентифицируемости: $rang\pi_x(i) = m_i - 1$ - существование единственного решения.

Дополнительные условия идентификации.

1. Если по условиям задачи один из коэффициентов регрессии заранее известен (например, равен единице), то из проверки идентификации он исключается, не учитывается.

2. Не подлежит идентификации уравнение, являющееся *тождеством*, то есть верным при любых значениях коэффициентов.

3. Не подлежит идентификации рекуррентная система уравнений, при которой каждая эндогенная переменная зависит от предыдущей по графу связей, но не зависит от последующих эндогенных, так как рекуррентная система может быть решена без преобразования структурных уравнений в приведенные.

4. Если все экзогенные переменные входят в уравнения всех эндогенных переменных, и последние связаны друг с другом, то система заведомо неидентифицируемая [10].

В заключение отметим, что делать, если система неидентифицируемая. Необходимо уменьшить число коэффициентов регрессии в структурных уравнениях, то есть исключить один (или более) экзогенный фактор. Какой из них – следует решить, принимая в расчет и содержательное значение фактора и тесноту его связи с результативным, эндогенным признаком. Если без какого-то фактора система вообще теряет смысл, нужно искать другие эндогенные и экзогенные переменные, то есть другой путь исследования объекта.

Попробуем ответить на вопрос об идентифицируемости параметров β_1, β_2, c_1 структурной формы СОУ (пример 7.1), т.е. о возможности их выражения через коэффициенты приведенной формы СОУ π_{11}, π_{21}.

Поделив π_{11} на π_{21}, получим:

$$\frac{\pi_{11}}{\pi_{21}} = \frac{\beta_1 c_1}{\beta_1 - \beta_2} \times \frac{\beta_1 - \beta_2}{c_1} = \beta_1 .$$

Используя коэффициенты приведенной формы СОУ можно найти коэффициенты первого уравнения структурной формы, но не сможем найти коэффициенты второго уравнения структурной формы. Следовательно, в приведенном выше примере (7.1), первое уравнение СОУ является точно идентифицируемым, а второе неидентифицируемым.

Пример 7.2 - Рассмотрим более сложную модель, описывающую предложение и спрос в условиях равновесия, включив в модель для спроса процентную ставку x_{2t}. В итоге модель спроса-предложения будет иметь вид:

$$y_{1t} = \beta_1 y_{2t} + \varepsilon_{1t};$$
$$y_{1t} = \beta_2 y_{2t} + c_1 x_{1t} + c_2 x_{2t} + \varepsilon_{2t}.$$

Перейдем к приведенной форме, умножив первое уравнение на β_2, а второе уравнение - на β_1:

$$\beta_2 y_{1t} = \beta_2 \beta_1 y_{2t} + \beta_2 \varepsilon_{1t};$$
$$\beta_1 y_{1t} = \beta_1 \beta_2 y_{2t} + \beta_1 c_1 x_{1t} + \beta_1 c_2 x_{2t} + \beta_1 \varepsilon_{2t}.$$

Из второго уравнения вычтем первое:

$$\beta_1 y_{1t} - \beta_2 y_{1t} = \beta_1 \beta_2 y_{2t} - \beta_2 \beta_1 y_{2t} + \beta_1 c_1 x_{1t} + \beta_1 c_2 x_{2t} + \beta_1 \varepsilon_{2t} - \beta_2 \varepsilon_{1t},$$
$$y_{1t}(\beta_1 - \beta_2) = \beta_1 c_1 x_{1t} + \beta_1 c_2 x_{2t} + \beta_1 \varepsilon_{2t} - \beta_2 \varepsilon_{1t},$$
$$y_{1t} = \frac{\beta_1 c_1}{\beta_1 - \beta_2} x_{1t} + \frac{\beta_1 c_2}{\beta_1 - \beta_2} x_{2t} + \frac{\beta_1 \varepsilon_{2t} - \beta_2 \varepsilon_{1t}}{\beta_1 - \beta_2}.$$

Обозначив $\frac{\beta_1 \varepsilon_{2t} - \beta_2 \varepsilon_{1t}}{\beta_1 - \beta_2}$ через δ_{1t}, получим первое уравнение приведенной формы:

$$y_{1t} = \pi_{11} x_{1t} + \pi_{12} x_{2t} + \delta_{1t}.$$

Аналогично проведем преобразования для второго уравнения:

$$\begin{aligned}
&y_{1t} - y_{1t} = \beta_1 y_{2t} - \beta_2 y_{2t} - c_1 x_{1t} - c_2 x_{2t} + \varepsilon_{1t} - \varepsilon_{2t},\\
&\beta_1 y_{2t} - \beta_2 y_{2t} - c_1 x_{1t} - c_2 x_{2t} + \varepsilon_{1t} - \varepsilon_{2t} = 0,\\
&y_{2t}(\beta_1 - \beta_2) = c_1 x_{1t} + c_2 x_{2t} + \varepsilon_{2t} - \varepsilon_{1t},\\
&y_{2t} = \frac{c_1}{\beta_1 - \beta_2} x_{1t} + \frac{c_2}{\beta_1 - \beta_2} x_{2t} + \frac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \beta_2}.
\end{aligned}$$

Обозначив $\frac{\varepsilon_{2t} - \varepsilon_{1t}}{\beta_1 - \beta_2}$ через δ_{2t}, получим второе уравнение приведенной формы:

$$y_{2t} = \pi_{21} x_{1t} + \pi_{22} x_{2t} + \delta_{2t}.$$

Приведенная форма СОУ имеет вид:

$$\begin{cases} y_{1t} = \pi_{11} x_{1t} + \pi_{12} x_{2t} + \delta_{1t}; \\ y_{2t} = \pi_{21} x_{1t} + \pi_{22} x_{2t} + \delta_{2t}. \end{cases},$$

$$\pi = \begin{pmatrix} \frac{\beta_1 c_1}{\beta_1 - \beta_2} & \frac{\beta_1 c_2}{\beta_1 - \beta_2} \\ \frac{c_1}{\beta_1 - \beta_2} & \frac{\beta_1 c_2}{\beta_1 - \beta_2} \end{pmatrix},$$

$$\frac{\pi_{11}}{\pi_{21}} = \beta_1,$$

$$\frac{\pi_{11}}{\pi_{22}} = \beta_2.$$

Получаем, что коэффициент β_1 структурной формы СОУ может быть определен по коэффициентам приведенной формы СОУ двумя различными способами, которые, вообще говоря, дают два разных результата.

В приведенном примере (7.2), первое уравнение СОУ является сверхидентифицируемым, а второе - неидентифицируемым.

7.4 Оценивание параметров структурной модели

Каждое уравнение системы одновременных уравнений не может рассматриваться как самостоятельная часть системы, поэтому применение традиционного метода наименьших квадратов для определения его параметров невозможно, т.к. нарушаются условия МНК: проблема мультиколлинеарности, случайные ошибки уравнения коррелируют с результативными переменными.

Методы оценивания систем одновременных уравнений можно разделить на методы, позволяющие оценивать каждое уравнение поочередно, и методы, предназначенные для оценивания всех уравнений сразу, т.е. всей модели в целом. Примерами первой группы методов служат двухшаговый метод наименьших квадратов и метод ограниченной информации для одного уравнения, а примерами второй группы методов – трехшаговый метод наименьших квадратов и метод максимального правдоподобия полной информации [4, с. 380].

В 1961 г. Г. Тейлор разработал семейство оценок коэффициентов структурной модели[1], позволившим развить метод ДМНК, практически вытеснивший более трудоемкий метод максимального правдоподобия при ограниченной информации, который использовался достаточно широко.

Трехшаговый МНК был предложен в 1962 г. А. Зельнером и Г. Тейлом, и он пригоден для всех видов уравнений структурной модели, хотя при некоторых ограничениях на параметры более эффективным может оказаться двухшаговый МНК.

[1] См. Тейл, Г. Экономические прогнозы и принятия решений : пер. с англ. / Г. Тейл. – М. : Статистика, 1971. – С. 281-282.

Традиционным методом оценивания для точно идентифицируемой системы одновременных уравнений является косвенный метод наименьших квадратов (КМНК). Алгоритм КМНК включает в себя следующие шаги:

- на основе структурной формы модели составляется ее приведенная форма;

- приведенные коэффициенты каждого уравнения оцениваются обычным методом наименьших квадратов;

- коэффициенты приведенной формы модели трансформируются в параметры структурной модели (при точно идентифицируемой СОУ по элементам матрицы π можно единственным образом найти элементы матриц B и C).

Пример 7.3 - По имеющимся данным (таблица 7.1) построим модель вида

$$\begin{cases} y_1 = f(y_2, x_1); \\ y_2 = f(y_1, x_2). \end{cases},$$

рассчитав соответствующие структурные коэффициенты.

Таблица 7.1 – Исходные данные для построения системы уравнений

Год	Годовое потребление продукта А на душу населения, кг	Оптовая цена за 1 кг продукта А, р.	Доход на душу населения, р.	Расходы по обработке продукта А, % к цене
	y_1	y_2	x_1	x_2
1	60	5,0	1300	60
2	62	4,0	1300	56
3	65	4,2	1500	56
4	62	5,0	1600	63
5	66	3,8	1800	50
Итого	315	22,0	7500	285

Составим систему структурных уравнений:

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

Для выбора метода оценки параметров проверим систему на идентифицируемость. В модели x_1, x_2 - предопределенные переменные и такое же количество эндогенных переменных - y_1 и y_2; число эндогенных переменных равно числу регрессионных уравнений.

Проверим необходимое условие идентифицируемости каждого уравнения системы - количество исключенных из уравнения экзогенных переменных должно быть не меньше количества эндогенных переменных в этом уравнении, уменьшенного на единицу. Имеем:

- k_1=2, m_1=1 – для первого уравнения;

- k_2=2, m_2=1 – для второго уравнения.

В обоих случаях $k - k_i \geq m_i - 1$, следовательно, оба уравнения точно идентифицируемы, откуда следует, что система в целом тоже точно идентифицирована.

Необходимое и достаточное условие идентифицируемости - $rang\,\pi_x(i) = m_i - 1$ - существование единственного решения. Так как в нашем примере система состоит только из двух уравнений, то данное условие не проверяется.

Для определения параметров точно идентифицируемой модели используем. КМНК. На первом этапе структурную форму преобразуем в приведенную форму. Параметры модели определяются с помощью традиционного МНК. Найдем данные параметры используя функцию Excel Сервис– Анализ данных– Регрессия (при этом необходимо учесть, что в уравнениях отсутствует свободный член).

Результаты регрессионного анализа приведенной формы представлены на рисунке 7.3. Следовательно, приведенная форма примет вид:

$$\begin{cases} y_1 = 0{,}02151 x_1 + 0{,}5363 x_2; \\ y_2 = -0{,}00025 x_1 + 0{,}08385 x_2. \end{cases}$$

На следующем этапе определим коэффициенты структурной модели.

ВЫВОД ИТОГОВ	y_1								ВЫВОД ИТОГОВ	y_2							
Регрессионная статистика									*Регрессионная статистика*								
Множественный R	0,9984								Множественный R	0,9987							
R-квадрат	0,9969								R-квадрат	0,9974							
Нормированный R-	0,6625								Нормированный R-ква	0,66319							
Стандартная ошибк	4,5479								Стандартная ошибка	0,29178							
Наблюдения	5								Наблюдения	5							
Дисперсионный анализ									Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*					*df*	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	2	19807	9903,5	478,82	0,0021				Регрессия	2	97,825	48,912	574,5	0,0017			
Остаток	3	62,049	20,683						Остаток	3	0,2554	0,0851					
Итого	5	19869							Итого	5	98,08						
Коэффициенартная опатисти		*)-Значени*	*ижние 95*	*Верхние 95%*	*жние 95,0*	*»жние 95,0%*			*Коэффициенартная опатисти*		*)-Значени*	*ижние 95*	*рхние 95*	*жние 95,0*	*»*		
	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д		0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
A11	0,0215	0,0077	2,7792	0,069	-0,0031	0,046141351	-0,0031	0,0461	A21	-0,0002	0,0005	-0,5015	0,6505	-0,0018	0,0013	-0,0018	
A12	0,5363	0,2047	2,6199	0,079	-0,1152	1,187696673	-0,1152	1,1877	A22	0,08385	0,0131	6,3847	0,0078	0,0421	0,1256	0,0421	

Рисунок 7.3 – Результаты регрессионного анализа уравнений приведенной формы

В первом уравнении структурной формы в правой части присутствуют переменные y_2 и x_1, следовательно, необходимо из второго уравнения выразить переменную x_2 через переменные y_2 и x_1:

$$x_2 = \frac{y_2 + 0{,}00025x_1}{0{,}08385}.$$

Подставим полученное выражение в первое уравнение и приведем подобные слагаемые:

$$y_1 = 0{,}0215x_1 + 0{,}5363 \times \left(\frac{y_2 + 0{,}00025x_1}{0{,}08385} \right) = 0{,}02151x_1 + 6{,}39595y_2 + 0{,}001599x_1 =$$

$$= 0{,}0231x_1 + 6{,}39595y_2.$$

Во втором уравнении структурной формы в правой части присутствуют переменные y_1 и x_2. Необходимо из первого уравнения выразить переменную x_1 через переменные y_1 и x_2:

$$x_1 = \frac{y_1 - 0{,}5363x_2}{0{,}02151}.$$

Подставим полученное выражение в первое уравнение и приведем подобные слагаемые:

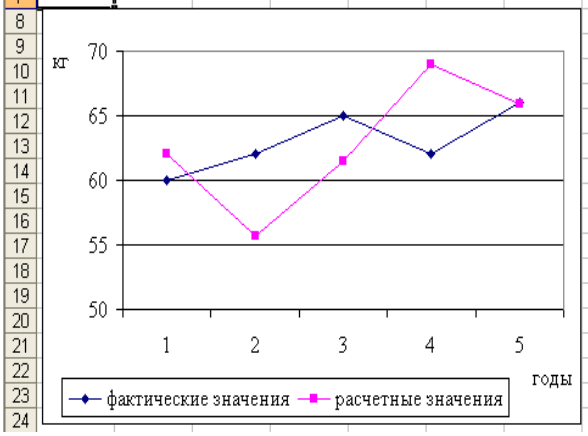
$$y_2 = -0{,}00025 \times \left(\frac{y_1 - 0{,}5363x_2}{0{,}02151} \right) + 0{,}08385x_2 = -0{,}01158y_1 + 0{,}0062x_2 + 0{,}08385x_2 = .$$
$$= -0{,}0116y_1 + 0{,}0901x_2.$$

Структурная форма модели примет вид:

$$\begin{cases} y_1 = 0{,}0231x_1 + 6{,}39595y_2; \\ y_2 = -0{,}0116y_1 + 0{,}0901x_2. \end{cases}$$

Рассчитаем по полученным уравнениям теоретические значения $\tilde{y}_1$ и $\tilde{y}_2$. Результаты расчетов представлены на рисунке 7.4.

	A	B	C	D	E	F	G
1	годы	y_1	y_2	x_1	x_2	$\hat{y}_1$	$\hat{y}_2$
2	1	60	5,0	1300	60	62	4,7
3	2	62	4,0	1300	56	56	4,3
4	3	65	4,2	1500	56	62	4,3
5	4	62	5,0	1600	63	69	5,0
6	5	66	3,8	1800	50	66	3,7



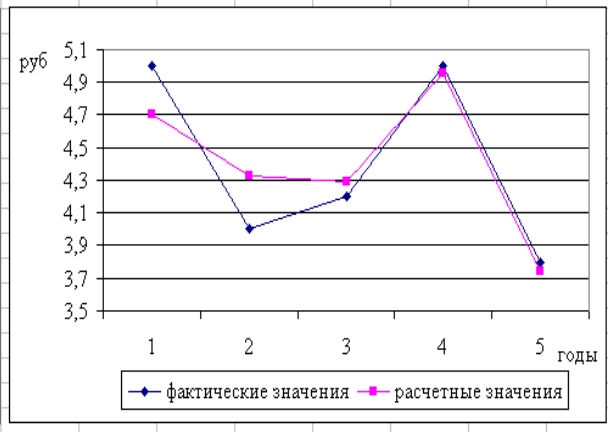


Рисунок 7.4 – Фактические и расчетные значения переменных y_1 и y_2

Если система сверхидентифицируема, то КМНК не используется, т.к. он не дает однозначных оценок для параметров структурной модели. В таких случаях применяется ДМНК. Данный метод называется двухшаговым в связи с тем, что МНК используется дважды: на первом шаге при определении приведенной формы модели и нахождении на ее основе оценок теоретических значений эндогенных переменных, и на втором шаге применительно к структурному сверхидентифицируемому уравнению при определении структурных коэффициентов.

Двухшаговый метод наименьших квадратов реализуется в несколько этапов:

- на основе структурной формы модели составляется ее приведенная форма;

- с помощью обычного МНК определяются оценки коэффициентов приведенных уравнений;

- рассчитываются значения тех эндогенных переменных, которые выступают в качестве факторных в сверхидентифицируемом уравнении;

- подставив эти значения вместо фактических в структурную форму, обычным МНК оценивают структурные коэффициенты модели.

Выделим две главные особенности двухшагового МНК.

1. ДМНК может применяться для оценки не только сверхидентифицируемых, но и точно идентифицируемых уравнений. В этом случае оценки, полученные ДМНК и КМНК, совпадут.

2. В случае если значения коэффициентов детерминации по уравнениям приведенной формы велики и превышают 0,8 ($R^2 > 0{,}8$), то оценки структурных параметров, полученные ДМНК и обычным МНК, будут близки. Это связано с тем, что при высоком значении R^2 расчетные значения инструментальных переменных не будут сильно отличаться от фактического значения соответствующих эндогенных переменных.

3. Если коэффициент детерминации R^2 для приведенного уравнения низкий, то расчетные значения эндогенной переменной будут плохой аппроксимацией ее фактических значений и применение ДМНК может оказаться неэффективным.

Согласно алгоритму трехшагового метода наименьших квадратов первоначально с целью оценки коэффициентов каждого структурного уравнения применяют

двухшаговый метод наименьших квадратов, а затем определяют оценку для ковариационной матрицы случайных возмущений. После этого с целью оценивания коэффициентов всей системы применяется обобщенный метод наименьших квадратов [42, с. 43].

7.5 Вопросы для самоконтроля

1. Сформулируйте основные цели использования системы одновременных уравнений.

2. Назовите возможные способы построения систем уравнений.

3. В чем различие между структурной и приведенной формами СОУ?

4. Сформулируйте проблемы и условия идентификации системы уравнений.

5. Опишите этапы реализации косвенного и двухшагового методов наименьших квадратов.

7.6 Тесты

1. Проблема идентификации модели системы уравнений состоит

а) в получении однозначно определенных параметров модели, заданной системой одновременных уравнений;

б) в выборе и реализации методов статистического оценивания неизвестных параметров модели по исходным статистическим данным;

в) в проверке адекватности модели.

2. Приведенная форма модели представляет собой:

а) систему нелинейных функций экзогенных переменных от эндогенных;

б) систему линейных функций эндогенных переменных от экзогенных;

в) систему линейных функций экзогенных переменных от эндогенных;

г) систему нормальных уравнений.

3. Для оценивания параметров сверхидентифицируемого уравнения применяется

в) МНК;

б) КМНК;

в) ДМНК.

4. Экзогенные переменные - это

а) зависимые переменные;

б) независимые переменные;

в) переменные, датированные предыдущими моментами времени.

5. В матричном виде структурная форма системы одновременных эконометрических уравнений имеет вид

а) $BY_t + CX_t = \delta_t$;

б) $Y_t = \pi X_t + \varepsilon_t$;

в) $Y = X\beta + \varepsilon$.

6. Модель считается идентифицируемой, если

а) каждое уравнение системы идентифицируемо;

б) среди уравнений модели есть хотя бы одно идентифицируемое;

в) среди уравнений модели есть хотя бы одно сверхидентифицируемое.

7. Если $k - k_i \geq m_i - 1$ и ранг матрицы $rang\pi_x(i) = m_i - 1$**,** то уравнение

а) сверхидентифицируемо;

б) неидентифицируемо;

в) точно идентифицируемо;

г) ситуация не определена.

8 Моделирование одномерного временного ряда

Что необходимо знать из главы 8:

8.1 Понятие и основные элементы временного ряда.

8.2 Автокорреляция уровней временных рядов и выявление его структуры. Стационарные временные ряды и их основные характеристики.

8.3 Моделирование тенденции временных рядов. Оценка параметров уравнения тренда.

8.4 Моделирование сезонных и циклических колебаний.

8.1 Понятие и основные элементы временного ряда

Временной ряд (ВР) или ряд динамики, динамический ряд – это последовательность упорядоченных во времени числовых показателей, характеризующих уровень состояния и изменения изучаемого явления.

ВР состоят из двух элементов:

1) периода времени, за который или по состоянию на который приводятся числовые значения (t);

2) числовых значений того или иного показателя, называемых уровнями ряда (y).

В практике исследования динамики явлений и прогнозирования принято считать, что значения уровней временных рядов могут содержать следующие компоненты:

- тренд (u_t);
- сезонную компоненту (S_t);
- циклическую компоненту (V_t);
- случайную компоненту (ε_t).

Под *трендом* понимают изменение, определяющее общее направление развития, основную тенденцию ВР. Это систематическая составляющая долговременного действия.

Наряду с долговременными тенденциями во ВР часто возникают более или менее регулярные колебания – периодические составляющие рядов динамики.

Если период колебаний не превышает одного года, то их называют *сезонными* (например, колебания цен на сельскохозяйственную продукцию).

При большем периоде колебания считают, что во временных рядах имеет место *циклическая* составляющая (циклы деловой активности Кондратьева).

Если из ВР удалить тренд и периодические составляющие, то останется *случайная компонента.*

Факторы, под действием которых формируется нерегулярная компонента, разделяют на два вида:

– факторы резкого, внезапного действия - вызывают более значительные отклонения – катастрофические колебания;

– текущие факторы – вызывают случайные колебания и являются результатом действия большого числа побочных причин.

Если ВР представляется в виде суммы соответствующих компонент, то полученная модель носит название аддитивной:

$$Y_t = u_t + S_t + v_t + \varepsilon_t . \qquad (8.1)$$

Если в виде произведения – мультипликативной:

$$Y_t = u_t \cdot S_t \cdot v_t \cdot \varepsilon_t . \qquad (8.2)$$

Также выделяют модели смешанного типа:

$$Y_t = u_t \cdot S_t \cdot v_t + \varepsilon_t . \qquad (8.3)$$

Отличительная особенность аддитивной модели заключается в том, что амплитуда сезонных колебаний, отражающих отклонения от тренда или среднего, остается примерно постоянной, неизменной во времени [46, с. 24-26].

В моделях временных рядов результативный признак является функцией переменной времени или переменных, относящихся к другим моментам времени.

К моделям временных рядов, представляющих собой зависимость результативного признака от времени, относятся модели:

– тренда (зависимости результативного признака от трендовой компоненты);

– сезонности (зависимости результативного признака от сезонной компоненты);

– тренда и сезонности.

К моделям временных рядов, представляющих собой зависимость результативного признака от переменных, датированных другими моментами времени, относятся модели:

– с распределенным лагом (объясняющие поведение результативного признака в зависимости от предыдущих значений факторных переменных);

– авторегрессии (объясняющие поведение результативного признака в зависимости от предыдущих значений результативных переменных);

– ожиданий (объясняющие поведение результативного признака в зависимости от будущих значений факторных или результативных переменных).

Модели временных рядов подразделяют также на модели, построенные по стационарным и нестационарным временным рядам. Стационарные временные ряды – ряды, имеющие постоянное среднее значение и колеблющиеся вокруг него с постоянной дисперсией. В таких рядах распределение показателя – уровня ряда не зависят от времени, т.е. стационарный временной ряд не содержит трендовой или сезонной компонент. В нестационарных временных рядах распределение уровня ряда зависят от переменной времени.

8.2 Автокорреляция уровней временного ряда и выявление его структуры. Стационарные временные ряды и их основные характеристики

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

Степень тесноты связи между последовательностями наблюдений ВР

$$y_1, y_2, ..., y_{n-\tau} \quad \text{и} \quad y_{1+\tau}, y_{2+\tau}, ..., y_n$$

(сдвинутых относительно друг друга на τ единиц, или, с лагом τ) может быть определена с помощью коэффициента корреляции

$$r(\tau) = \frac{\overline{y_t \cdot y_{t-\tau}} - \overline{y}_t \cdot \overline{y}_{t-\tau}}{\sigma_t \cdot \sigma_{t-\tau}}, \qquad (8.4)$$

где $\overline{y_t \cdot y_{t-\tau}} = \dfrac{\sum_{i=1+\tau}^{n} y_i \cdot y_{i-\tau}}{n-\tau}$;

$\overline{y}_t = \dfrac{\sum_{i=1+\tau}^{n} y_i}{n-\tau}$ - средний уровень ряда $y_{1+\tau}, y_{2+\tau}, ..., y_n$;

$\overline{y}_{t-\tau} = \dfrac{\sum_{i=1+\tau}^{n} y_{i-\tau}}{n-\tau}$ - средний уровень ряда $y_1, y_2, ..., y_{n-\tau}$;

$\sigma_t, \sigma_{t-\tau}$ - средние квадратические отклонения для рядов $y_{1+\tau}, y_{2+\tau}, ..., y_n$ и $y_1, y_2, ..., y_{n-\tau}$ соответственно.

Так как коэффициент $r(\tau)$ измеряет корреляцию между членами одного и того же ряда, его называют коэффициентом автокорреляции. Лаг определяет порядок коэффициента автокорреляции. Если $\tau = 1$, то имеем коэффициент автокорреляции 1-го порядка, если $\tau = 2$ - второго порядка и т.д. Следует учитывать, что с увеличением лага на единицу число пар значений, по которым рассчитывается коэффициент ав-

токорреляции, уменьшается на единицу. Поэтому обычно рекомендуют максимальный порядок коэффициента автокорреляции, равный n/4.

Рассчитав несколько коэффициентов автокорреляции, можно определить лаг τ, при котором автокорреляция $r(\tau)$ наиболее высокая, выявив тем самым структуру ВР. Если наиболее высоким оказывается значение $r(1)$, то исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался $r(\tau)$, то ряд содержит помимо тенденции колебания периодом τ. Если ни один из коэффициентов не является статистически значимым, можно сделать одно из предположений:

– либо ряд не содержит тенденции и циклических колебаний;

– либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужен дополнительный анализ.

Последовательность коэффициентов автокорреляции 1-го, 2-го и т.д. порядков называют автокорреляционной функцией (рисунок 8.1). График зависимости значений коэффициентов автокорреляции от величины лага – коррелограммой (рисунок 8.2) [47].

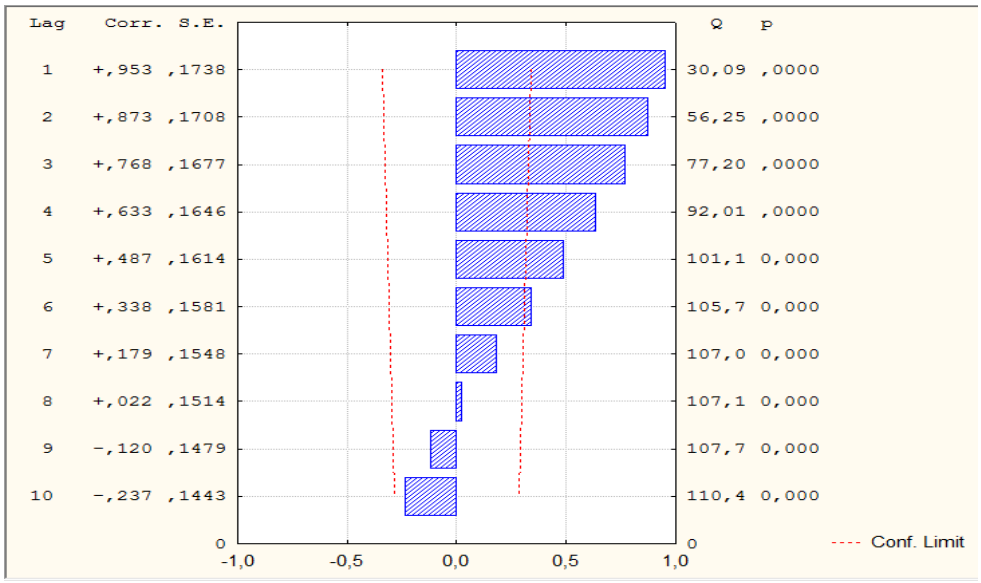


Рисунок 8.1 - Автокорреляционная функция числа родившихся на 1000 человек населения в РФ (1980г. - 2009г.)[1]

[1] Построено в ППП Statistica 6.0 по данным сайта http://www. gks.ru

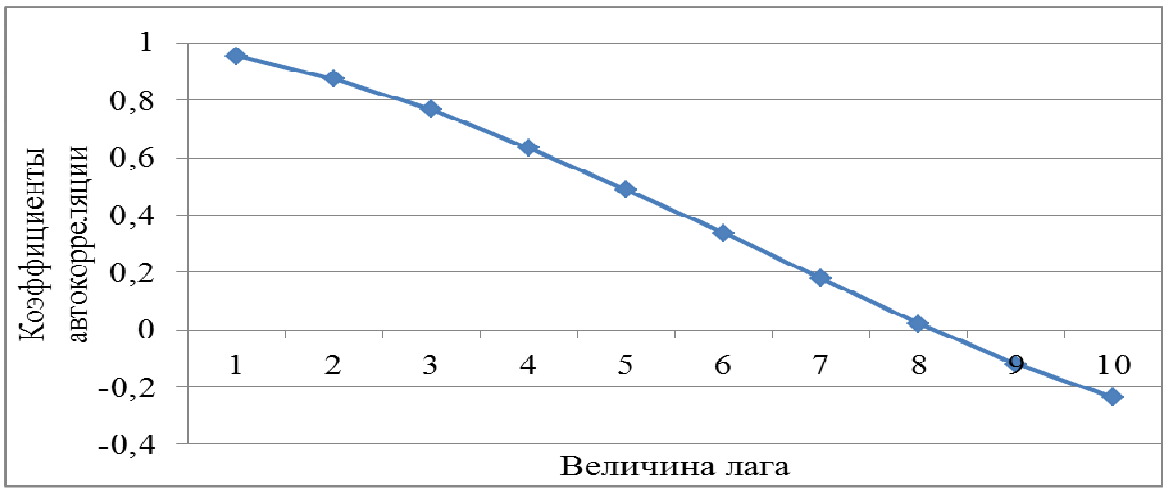


Рисунок 8.2 – Коррелограмма автокорреляционной функции числа родившихся на 1000 человек населения в РФ

Полученные значения автокорреляционной функции свидетельствуют о смене тенденции во временном ряду числа родившихся на 1000 человек населения в РФ: значения коэффициентов стремительно уменьшаются с увеличением величины лага и меняют знак. Это подтверждает и графическое изображение анализируемого временного ряда (рисунок 8.3).

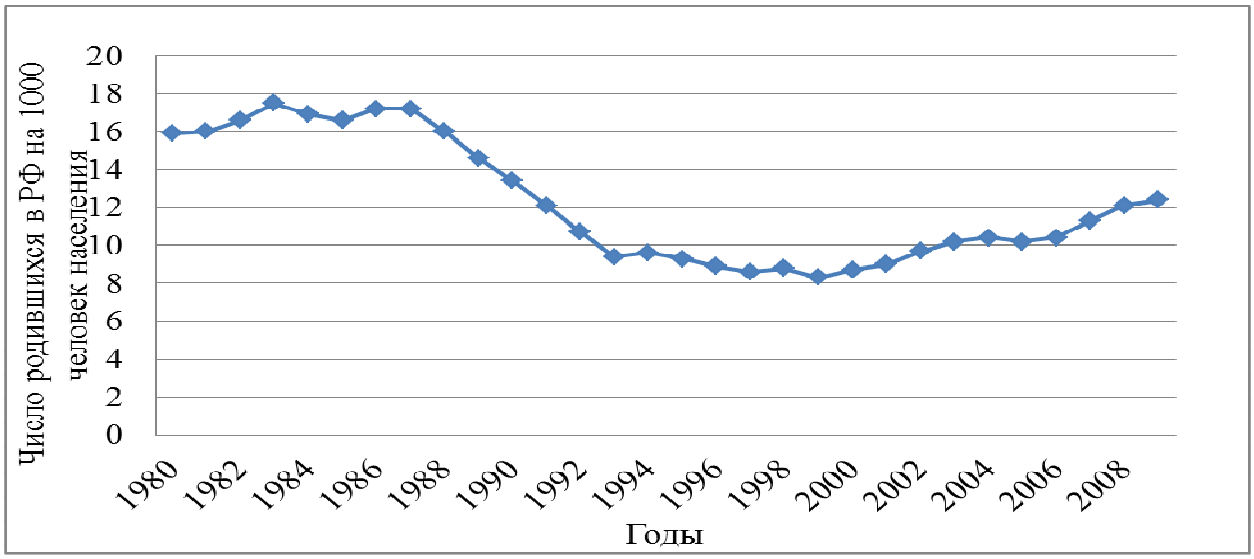


Рисунок 8.3 - Динамика числа родившихся на 1000 человек населения в РФ

Если во временном ряду наблюдается монотонная тенденция, то автокорреляционая функция имеет значения, близкие к +1, которые медленно снижаются с увеличением лага (рисунок 8.4).

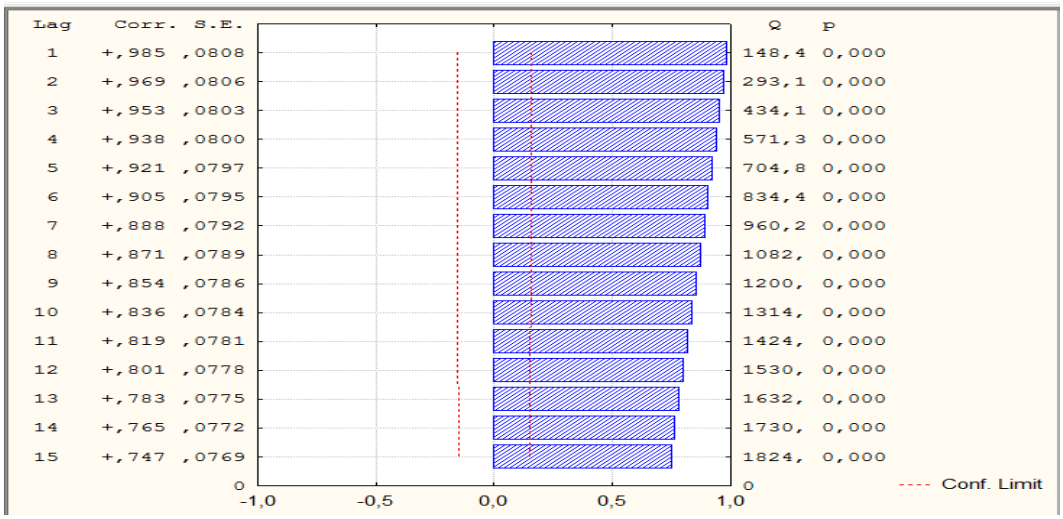


Рисунок 8.4 - Автокорреляционная функция реальных денежных доходов населения в РФ с поправкой на сезонность (январь 2000г. – июнь 2012г.)[1]

Для стационарного временного ряда автокорреляционная функция имеет статистически не значимые коэффициенты с монотонным их убыванием (рисунок 8.5).

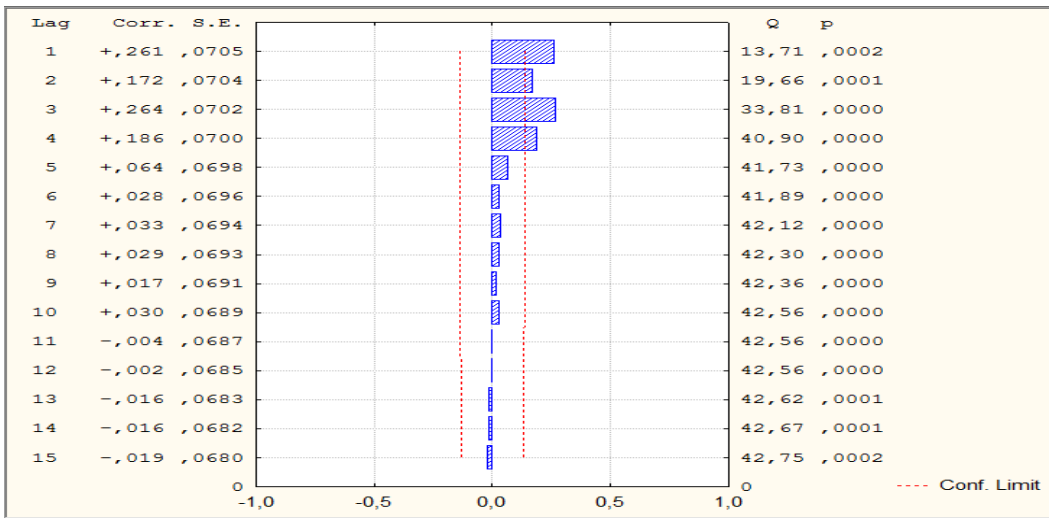


Рисунок 8.5 - Автокорреляционная функция индекса потребительских цен в РФ (январь 1996 г. – июнь 2012г.)

[1] Построено в ППП Statistica 6.0 по данным сайта http://www. hse.ru

Рассмотрим формальное определение стационарности.

Стохастический процесс Y_t называется *стационарным в сильном смысле (строго стационарным или стационарным в узком смысле)*, если совместное распределение вероятностей всех переменных $y_{t1}, y_{t2},..., y_{tm}$ точно то же самое, что и для переменных $y_{t1+\tau}, y_{t2+\tau},..., y_{tm+\tau}$.

Под стационарным процессом в слабом смысле (в широком смысле) понимается стохастический процесс, для которого среднее и дисперсия независимо от рассматриваемого периода времени имеют постоянное значение, а автоковариация зависит только от длины лага между рассматриваемыми переменными:

$$\begin{aligned} &\mu(y_t) = \mu(y_{t+\tau}) = \mu; \\ &D(y_t) = \mu(y_t - \mu)^2 = \mu(y_{t+\tau} - \mu)^2 = \gamma(0) = const; \\ &\operatorname{cov}(y_t, y_{t+\tau}) = \mu[(y_t - \mu)(y_{t+\tau} - \mu)] = \gamma(\tau). \end{aligned} \qquad (8.5)$$

Из этого следует, что автокорреляция будет зависеть только от сдвига по времени τ и не будет зависеть от t [46, с. 27-29].

В практической аналитической работе стационарность временного ряда означает отсутствие:

– тренда;

– систематических изменений дисперсии;

– строго периодичных флуктуаций;

– систематически изменяющихся взаимосвязей между элементами временного ряда [48, с. 337].

8.3 Моделирование тенденции временных рядов. Оценка параметров уравнения тренда

Прежде чем перейти к определению тенденции и выделению тренда, нужно выяснить, существует ли вообще тенденция в исследуемом процессе. Для этой цели

разработано множество критериев: критерий серий, метод проверки разностей средних уровней, метод Фостера-Стюарта. Основные подходы к решению этой задачи основаны на статистической проверке гипотез о случайности ряда: $\mathrm{H}_0 = \mathrm{M}_Y(t) = a = \mathrm{const}$.

Рассмотрим *критерий серий*, который имеет две модификации [46, с. 51-54]:

– критерий серий, основанный на медиане выборки;

– критерий «восходящих и нисходящих» серий.

Алгоритм первой модификации включает следующие шаги:

1. Из исходного ряда с уровнями $y_1, y_2, ..., y_n$ образуется ранжированный ряд $y_1', y_2', ..., y_n'$ (где y_1' – наименьшее значение из уровней исходного ряда).

2. Определяется медиана ($\mathrm{M_e}$) этого вариационного ряда. В случае нечетного значения длины ряда n (n=2m+1) $M_e = y'_{m+1}$, в противном случае (n=2m) $M_e = (y'_m + y'_{m+1})/2$.

3. Образуется последовательность δ_i из плюсов и минусов по следующему правилу:

$$\delta_i = \begin{cases} +, \text{если } y_t > M_e, t = 1,2,...,n; \\ \\ -, \text{если } y_t < M_e, t = 1,2,...,n. \end{cases} \qquad (8.6)$$

Если значение уровня исходного ряда y_t равно медиане, то это значение пропускается.

4. Подсчитывается $\nu(n)$ – число серий в совокупности δ_i, где под серией понимается последовательность подряд идущих плюсов и минусов. Один плюс или один минус тоже будут считаться серией. Определяется $\tau_{\max}(n)$ – протяженность самой длиной серии.

5. Проверка гипотезы основывается на том, что при условии случайности ряда (при отсутствии систематической составляющей) протяженность самой длиной серии не должна быть слишком большой, а общее число серий – слишком маленьким.

Поэтому, для того чтобы не была отвергнута гипотеза о случайности исходного ряда, должны выполняться следующие неравенства:

$$\upsilon(n) > \left[\frac{1}{2}\left(n+1-1{,}96\sqrt{n-1}\right)\right],$$
$$\tau_{\max}(n) < [1{,}43\ln(n+1)], \tag{8.7}$$

где n – длина временного ряда.

Если хотя бы одно из неравенств нарушается, то гипотеза отвергается с вероятностью ошибки α, заключенной между 0,05 и 0,0975 (следовательно, подтверждается наличие зависящей от времени неслучайной составляющей).

Пример 8.1 - Применение критерия серий, основанного на медиане выборки.

1. Из исходного временного ряда числа родившихся на 1000 человек населения РФ (y_t), образуем ранжированный ряд (y_t') (таблица 8.1).

Таблица 8.1 - Формирование серий

Год	y_t, $^0/_{00}$	y_t'	δ_i
1	2	3	4
1999	8,3	8,3	-
2000	8,7	8,7	-
2001	9,0	9,0	-
2002	9,7	9,7	-
2003	10,2	10,2	
2004	10,4	10,2	+
2005	10,2	10,4	
2006	10,4	10,4	+
2007	11,3	11,3	+
2008	12,1	12,1	+
2009	12,4	12,4	+

2. Определяем медиану (M_e) ранжированного временного ряда. Так как значение длины ряда нечетное, то $M_e = y'_{m+1} = 10{,}2$.

3. Образуем последовательность δ_i из плюсов и минусов по правилу:

$$\delta_i = \begin{cases} +, \text{ если } y_t > M_e, t = 1,2,...,n; \\ \\ -, \text{ если } y_t < M_e, t = 1,2,...,n. \end{cases}$$

Если значение уровня исходного ряда y_t равно медиане, то это значение пропускается (столбец 4 таблицы 8.1).

4. Подсчитывается $\nu(n)$ – число серий в совокупности δ_i, где под серией понимается последовательность подряд идущих плюсов и минусов. Один плюс или один минус тоже будут считаться серией. Определяется $\tau_{\max}(n)$ – протяженность самой длиной серии. Получаем: $\nu(n)=2$, $\tau_{\max}(n)=5$.

5. Чтобы не была отвергнута гипотеза о случайности исходного ряда, должны выполняться следующие неравенства:

$$\upsilon(n) > \left[\frac{1}{2}\left(n+1-1{,}96\sqrt{n-1}\right)\right],$$

$$\tau_{\max}(n) < [1{,}43\ln(n+1)],$$

где n – длина временного ряда.

Рассчитаем правые части неравенств:

$$\frac{1}{2}\left(11+1-1{,}96\sqrt{11-1}\right) = 2{,}9,$$

$$1{,}43 \cdot \ln(11+1) = 3{,}6.$$

Так как в правой части неравенства стоят квадратные скобки, означающие целую часть числа, то сравнения будем проводить с целыми числами, соответственно с 2 и 3.

Получим: 2>2, 5<3.

Оба неравенства нарушаются, следовательно, гипотеза отвергается с вероятностью ошибки α, заключенной между 0,05 и 0,0975 (следовательно, подтверждается наличие зависящей от времени неслучайной составляющей).

Алгоритм критерия «восходящих и нисходящих» серий.

1. Образуется последовательность плюсов и минусов, но по другому правилу.

Для временного ряда с уровнями $y_1, y_2, ..., y_n$ определяется вспомогательная последовательность, исходя из условий:

$$\delta_i = \begin{cases} +, \text{если } y_{t+1} - y_t > 0, \text{для } \ t = 1,2,...,n; \\ \\ -, \text{если } y_{t+1} - y_t < 0, \text{для } \ t = 1,2,...,n. \end{cases} \tag{8.8}$$

В случае, когда последующее наблюдение окажется равным предыдущему, учитывается только одно наблюдение.

2. Подсчитывается общее число серий $\nu(n)$ и протяженность самой длиной серии $\tau_{\max}(n)$ аналогично. Серия, состоящая из «+», – «восходящая серия», из «-» – нисходящая.

3. Для того чтобы не была отвергнута гипотеза о случайности исходного ряда, должны выполняться следующие неравенства:

$$\begin{aligned} &\upsilon(n) > \left[\frac{1}{3}\left((2n-1) - 1{,}96\sqrt{\frac{16n-29}{90}}\right)\right], \\ &\tau_{\max}(n) < \tau_0(n) \end{aligned} \tag{8.9}$$

где $\tau_0(n)$ - табличное значение, зависящее от *n* (таблица 8.2).

Таблица 8.2

n	$n \le 26$	$26 \le n \le 153$	$153 \le n \le 1170$
$\tau_0(n)$	5	6	7

Если хотя бы одно из неравенств нарушается, то нулевая гипотеза отвергается.

Пример 8.2 - Рассмотрим пример применения критерия «восходящих» и «нисходящих» серий.

1. Последовательность плюсов и минусов образуется по другому правилу. Для временного ряда с уровнями $y_1, y_2, \ldots, y_n$ (таблица 8.3) определяется вспомогательная последовательность, исходя из условий:

$$\delta_i = \begin{cases} +, \text{если } y_{t+1} - y_t > 0, \text{для } \ t = 1,2,...,n; \\ \\ -, \text{если } y_{t+1} - y_t < 0, \text{для } \ t = 1,2,...,n. \end{cases}$$

Таблица 8.3 - Формирование серий

Год	y_t, $^0/_{00}$	δ_i
1	2	3
1999	8,3	
2000	8,7	+
2001	9,0	+
2002	9,7	+
2003	10,2	+
2004	10,4	+
2005	10,2	-
2006	10,4	+
2007	11,3	+
2008	12,1	+
2009	12,4	+

В случае, когда последующее наблюдение окажется равным предыдущему, учитывается только одно наблюдение.

2. Подсчитывается общее число серий $\nu(n)$ и протяженность самой длиной серии $\tau_{\max}(n)$ аналогично. Серия, состоящая из «+» – «восходящая серия», из «–» – «нисходящая серия». Получим: $\nu(n)=3$; $\tau_{\max}(n)=5$ (таблица 8.2).

3. Для того чтобы не была отвергнута гипотеза о случайности исходного ряда, должны выполняться следующие неравенства:

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

$$\tau_{\max}(n) < \tau_0(n),$$

где $\tau_0(n)$ – табличное значение, зависящее от n (таблица 8.4).

Таблица 8.4

n	$n \le 26$	$26 \le n \le 153$	$153 \le n \le 1170$
$\tau_0(n)$	5	6	7

Рассчитаем значения правой части первого неравенства:

$$\left[\frac{1}{3}(2\cdot 11-1)-1{,}96\cdot\sqrt{\frac{16\cdot 11-29}{90}}\right]=3.$$

Табличное значение $\tau_0(n)=5$.

Проверка выполнения условий показывает, что оба неравенства не выполняются. Следовательно, нулевая гипотеза отвергается, динамика временного ряда характеризуется наличием систематической составляющей – в изменении числа родившихся на 1000 человек населения в РФ за 1999-2009 гг. присутствует тенденция.

Алгоритм метода разности средних уровней имеет следующую последовательность [49, с. 330-331]:

1. Анализируемый ряд разбивается на две примерно равные по числу членов части n_1 и n_2, каждая из которых рассматривается как самостоятельная (частная) выборка:

$$y^{(1)} = \left(y_1, y_2, ..., y_{n_1}\right), \tag{8.10}$$

$$y^{(2)} = \left(y_{n_1+1}, y_{n_1+2}, ..., y_n\right), \tag{8.11}$$

где $n = n_1 + n_2$.

2. По каждой из частных выборок выполняется оценка средних:

$$\overline{y^{(1)}} = \frac{1}{n_1} \cdot \sum_{t=1}^{n_1} y_t \,, \tag{8.12}$$

$$\overline{y^{(2)}} = \frac{1}{n_2} \cdot \sum_{t=n_1+1}^{n} y_t \,. \tag{8.13}$$

3. Вычисляется разность средних:

$$R = \overline{y^{(1)}} - \overline{y^{(2)}} \,. \tag{8.14}$$

4. Проверяется статистическая значимость разности средних – гипотеза $H_0 : \overline{y^{(1)}} = \overline{y^{(2)}}$ при помощи t – статистики Стьюдента:

$$t_R = \frac{R}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \,, \tag{8.15}$$

где s – несмещенная выборочная оценка дисперсии уровней ряда:

$$s=\sqrt{\frac{\sum_{t=1}^{n_1}\left(y_t-\overline{y^{(1)}}\right)^2+\sum_{t=n_1+1}^{n}\left(y_t-\overline{y^{(2)}}\right)^2}{(n_1-1)+(n_2-1)}}=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n-2}}, \tag{8.16}$$

где

$$s_1^2=\frac{1}{n_1-1}\sum_{t=1}^{n_1}\left(y_t-\overline{y^{(1)}}\right)^2, \tag{8.17}$$

$$s_2^2=\frac{1}{n_2-1}\sum_{t=n_1+1}^{n}\left(y_t-\overline{y^{(2)}}\right)^2. \tag{8.18}$$

Если $|t_{\text{набл}}|<t_{\text{табл}(\alpha;n-2)}$, то гипотеза H_0 принимается, во временном ряду тенденция отсутствует.

В основе формулы (8.15) лежит предположение о несущественном различии дисперсий частных выборок и отсутствии зависимости между частными выборками. Поэтому, перед расчетом t-статистики Стьюдента необходимо проверить гипотезу о несущественном различии значений дисперсий уровней ряда в частных выборках. Проверка осуществляется при помощи F-критерия Фишера: формируется статистика

$$F=\frac{s_1^2}{s_2^2}. \tag{8.19}$$

Вычисленное значение статистики сравнивается с ее критическим (табличным) значением. Если $F > F_{\text{крит}(\alpha;n_1-1;n_2-1)}$, то гипотеза о несущественном различии значений дисперсий уровней ряда в частных подвыборках отклоняется, и метод разности средних уровней не может быть применен.

Пример 8.3 - Рассмотрим пример применения метода разности средних уровней. Данные временного ряда (таблица 8.1) разобьем на две частные выборки объемами $n_1=6, n_2=5$ (таблицы 8.5, 8.6).

Таблица 8.5 - Расчет дисперсии первой частной выборки

Год	y_t, ‰	$\left(y_t - \overline{y^{(1)}}\right)^2$
1	2	3
1999	8,3	1,174
2000	8,7	0,467
2001	9,0	0,147
2002	9,7	0,100
2003	10,2	0,667
2004	10,4	1,034
Итого	56,3	3,588
В среднем	9,4	0,718

По формулам (8.12) и (8.13) вычислим средние по частным выборкам:

$$\overline{y^{(1)}} = \frac{1}{n_1} \cdot \sum_{t=1}^{n_1} y_t = \frac{1}{6} \cdot 56{,}3 = 9{,}4 ,$$

$$\overline{y^{(2)}} = \frac{1}{n_2} \cdot \sum_{t=n_1+1}^{n} y_t = \frac{1}{5} \cdot 56{,}4 = 11{,}3 .$$

Таблица 8.6 - Расчет дисперсии второй частной выборки

Год	y_t, ‰	$\left(y_t - \overline{y^{(2)}}\right)^2$
1	2	3
2005	10,2	1,166
2006	10,4	0,774
2007	11,3	0,000
2008	12,1	0,672
2009	12,4	1,254
Итого	56,4	3,868
В среднем	11,3	0,967

Разность средних составит:

$$R = \overline{y^{(1)}} - \overline{y^{(2)}} = 9{,}4 - 11{,}3 = -1{,}9\,.$$

Оценки дисперсий частных выборок равны:

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{t=1}^{n_1}\left(y_t - \overline{y^{(1)}}\right)^2 = \frac{1}{6-1}\cdot 3{,}588 = 0{,}718\,,$$

$$s_2^2 = \frac{1}{n_2 - 1}\sum_{t=n_1+1}^{n}\left(y_t - \overline{y^{(2)}}\right)^2 = \frac{1}{5-1}\cdot 3{,}868 = 0{,}967\,.$$

Несмещенную выборочную оценку дисперсии уровней ряда вычислим по формуле:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n - 2}} = \sqrt{\frac{5\cdot 0{,}718 + 4\cdot 0{,}967}{11 - 2}} = 0{,}910.$$

Проверку предпосылки осуществим при помощи F-критерия Фишера:

$$F = \frac{s_1^2}{s_2^2} = \frac{0{,}718}{0{,}967} = 0{,}742.$$

Табличное значение *F*-критерия для уровня значимости 0,05 и числе степеней свободы (5;4) равно 6,256. Таким образом, $F < F_{\text{табл}}$, гипотеза о несущественности различий дисперсий уровней ряда в частных подвыборках не отклоняется, следовательно, может быть применен метод разности средних уровней.

t-критерий Стьюдента составит:

$$t_R = \frac{R}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{-1{,}9}{0{,}910\cdot\sqrt{\frac{1}{5} + \frac{1}{4}}} = -3{,}106\,.$$

Критическое значение *t*-статистики для уровня значимости 0,05 и 9 степеням свободы равно 2,262, т.е. $|t_R| > t_{\text{крит}}$, нулевая гипотеза отвергается, во временном ряду присутствует тенденция.

Одним из наиболее распространенных методов проверки временных рядов на стационарность является метод Фостера – Стюарта. Алгоритм метода состоит в следующем.

1. Каждый уровень ряда сравнивается со всеми предшествующими. При этом определяются вспомогательные характеристики:

$$m_t = \begin{cases} 1, \text{ если } x_t > x_k, \quad k = 1,2,...,t-1; \\ 0, \text{ иначе.} \end{cases} \tag{8.20}$$

$$l_t = \begin{cases} 1, \text{ если } x_t < x_k, \quad k = 1,2,...,t-1; \\ 0, \text{ иначе.} \end{cases} \tag{8.21}$$

2. Вычисляется значения величин

$$d_t = m_t - l_t, t = \overline{2,n}, \tag{8.22}$$

$$S_t = m_t + l_t, t = \overline{2,n}. \tag{8.23}$$

Таким образом, величина d_t может принимать значения:

а) минус 1 – если уровень ряда наименьший;

б) 0 – если уровень ряда не является ни наибольшим, ни наименьшим;

в) 1 – уровень ряда наибольший.

Величина S_t может принимать значения:

а) 0 – если уровень ряда не является ни наибольшим, ни наименьшим;

б) 1 – в противном случае .

3. Вычисляются суммы:

$$D = \sum_{t=2}^{n} d_t \,, \tag{8.24}$$

$$S = \sum_{t=2}^{n} S_t \,. \tag{8.25}$$

Показатель D изменяется от минус (n-1) до (n-1), и применяется для обнаружения тенденции изменения средней величины уровней ряда.

Показатель S изменяется от 0 до (n-1) и применяется для обнаружения тенденции изменения дисперсии уровней ряда.

4. С помощью критерия Стьюдента проверяется гипотеза об отсутствии тенденции в средней и дисперсии. Для этого определяется

$$t_D = \frac{D}{\sigma_D}, \tag{8.26}$$

$$\text{где } \sigma_D = \sqrt{2\sum_{t=2}^{n} \frac{1}{t}} \approx \sqrt{2\ln(n) - 0{,}8456}\,, \tag{8.27}$$

$$t_S = \frac{S - \mu}{\sigma_S}\,, \tag{8.28}$$

$$\text{где } \sigma_S = \sqrt{2\sum_{t=2}^{n} \frac{1}{t} - 4\sum_{t=2}^{n} \frac{1}{t^2}} \approx \sqrt{2\ln n - 3{,}4253}\,, \tag{8.29}$$

$$\mu = 2\sum_{t=2}^{n} \frac{1}{t}\,. \tag{8.30}$$

Если $|t_{\text{набл}}| > t_{\text{кр}}$, то H_0 отвергается, следовательно, тренд есть.

Пример 8.4 - В качестве примера применения метода Фостера – Стюарта рассмотрим проверку на стационарность уровней динамического ряда числа родившихся на 1000 человек населения в РФ (y_t), представленных в таблице 8.7.

Таблица 8.7 - Вычисление характеристик ряда

Год	t	y_t	m_t	l_t	d_t	S_t	$1/t$	$1/t^2$
1	2	3	4	5	6	7	8	9
1999	1	8,3	-	-	-	-	-	-
2000	2	8,7	1	0	1	1	0,500	0,250
2001	3	9,0	1	0	1	1	0,333	0,111
2002	4	9,7	1	0	1	1	0,250	0,063
2003	5	10,2	1	0	1	1	0,200	0,040
2004	6	10,4	1	0	1	1	0,167	0,028
2005	7	10,2	0	0	0	0	0,143	0,020
2006	8	10,4	0	0	0	0	0,125	0,016
2007	9	11,3	1	0	1	1	0,111	0,012
2008	10	12,1	1	0	1	1	0,100	0,010
2009	11	12,4	1	0	1	1	0,091	0,008
итого	-	104,4	-	-	8	8	2,020	0,558

Значения величин m_t, l_t, d_t, и S_t, вычисленные по формулам 8.20 – 8.23, представлены в столбцах 4 – 7 таблицы 8.7.

Показатели D и S (8.24), (8.25) – итоги столбцов 6 и 7 таблицы 8.7 соответственно.

Рассчитаем *t*-критерий Стьюдента. Проверяется гипотеза об отсутствии тенденции в средней. Для этого воспользуемся формулами (8.26) и (8.27).

$$\sigma_D = \sqrt{2 \cdot 2{,}02} = 2{,}010,$$

$$t_D = \frac{8}{2{,}010} = 3{,}980 .$$

Табличное значение $t_{\text{крит}(0,05;10)} = 2{,}228$. Таким образом, неравенство $|t| < t_{\text{крит}}$ нарушается, следовательно, нулевая гипотеза об отсутствии тенденции в средней отвергается.

Для проверки гипотезы об отсутствии тенденции в дисперсии воспользуемся формулами (8.28), (8.29) и (8.30).

$$\sigma_S = \sqrt{2 \cdot 2{,}02 - 4 \cdot 0{,}558} = 1{,}344 ,$$

$$\mu = 2 \cdot 2{,}020 = 4{,}04 ,$$

$$t_S = \frac{8 - 4{,}04}{1{,}344} = 2{,}946 .$$

Так как $|t_{\text{набл}}| > t_{кр}$, то H_0 отклоняется. Следовательно, гипотеза об отсутствии тенденции в дисперсии отклоняется.

В целом применение четырех критериев (двух модификаций критерия серий, метода разности средних уровней, метода Фостера – Стюарта) позволяет сделать вывод, что с вероятностью 0,95 тренд во временном ряду присутствует.

При наличии тенденции в ряду динамики его уровни можно рассматривать как функцию времени (кривые роста). Кривые роста условно разделяют на 3 класса:

Первый класс включает функции, используемые для описания процессов с монотонным характером развития и отсутствием пределов роста (класс полиномов, экспоненциальная (показательная) кривая, логарифмическая парабола);

$$\tilde{y}_i = a + b \cdot t_i \text{ - линейный} \quad , \tag{8.31}$$

$$\tilde{y}_i = a + b \cdot t + c \cdot t^2 \text{ - параболический,} \tag{8.32}$$

$$\tilde{y}_i = a \cdot k^{t_i} \text{ - экспоненциальный.} \tag{8.33}$$

Ко второму классу относятся кривые, описывающие процесс, который имеет предел роста в исследуемом периоде (кривые насыщения) – потребление каких либо продуктов, расход удобрений на единицу площади (модифицированная экспонента, гиперболические кривые):

$$\tilde{y} = a + \frac{b}{t} \text{ - гипербола,} \tag{8.34}$$

$$\tilde{y} = c - ab^t \text{ - модифицированная экспонента.} \tag{8.35}$$

Третий класс включает кривые насыщения, имеющие точку перегиба (*S* - образные кривые). Эти кривые описывают как бы два последовательных лавинообразных процесса: один с ускорением развития, другой – с замедлением. Применяют в демографических исследованиях, страховых расчетах, определении спроса на новый вид продукции (кривая Гомперца, логистическая кривая):

$$\tilde{y} = \frac{1}{c + ab^t} \text{ или } \tilde{y} = \frac{c}{1 + be^{-at}} \text{ - логистическая кривая,} \tag{8.36}$$

$$\tilde{y} = ca^{b^t} \text{ - кривая Гомперца.} \tag{8.37}$$

Существует несколько практических подходов, облегчающих процесс выбора формы кривой роста.

Наиболее простой путь – визуальный, опирающийся на графическое изображение временного ряда. Если на графике исходного ряда тенденция развития не четко просматривается, то можно преобразовать ряд (например, сгладить).

Для выбора степени полинома применяют метод последовательных разностей, который предполагает вычисление первых, вторых и т.д. разностей уровней ряда:

$$\Delta y_t = y_t - y_{t-1}; \;\; \Delta^2 y_t = \Delta y_t - \Delta y_{t-1} \text{ и т.д.} \tag{8.38}$$

Расчет ведется до тех пор, пока разности не будут примерно равными. Порядок разностей принимается за степень выравнивающего полинома.

Однако чаще всего на практике форму кривой выбирают по наименьшей сумме квадратов отклонений фактических уровней от расчетных. Используя этот подход, следует иметь, что к ряду, состоящему из m точек можно подобрать многочлен степени (m-1), проходящей через все m точек, однако, такая кривая не слишком пригодна как для выделения тенденции, так и для прогнозирования. Иногда в качестве критерия выбирается средняя квадратическая ошибка

$$S = \sqrt{\frac{\sum (y_t - \tilde{y}_t)^2}{n-k}},$$

где n - длина ряда;

k - число оцениваемых коэффициентов в модели.

Использование этого подхода проходит в два этапа: на первом происходит ограничение приемлемых функций, исходя из содержательного анализа задачи, на втором - осуществляется расчет критерия и выбор по нему функции.

Оценка параметров линейного, параболического и гиперболического трендов.

Основой методики оценки параметров служит метод наименьших квадратов, который дает оценки, отвечающие принципу максимального правдоподобия: сумма квадратов отклонений фактических уровней от тренда (от выравненных по уравнению тренда уровней) должна быть минимальной для данного типа уравнения.

Эта методика близка к методике корреляционно-регрессионного анализа связей – парной регрессии. Однако между ними есть и принципиальные различия: выступающий при расчете уравнения тренда в качестве независимой переменной ряд номеров периодов или моментов времени не является случайной варьирующей переменной X регрессионного анализа. Ряд значений времени – это жестко упорядоченный ряд величин, и, следовательно, не может быть речи о корреляции между ним и значениями зависимой переменной – варьирующих уровней показателя, изме-

няющегося во времени. Нередко применяемые в литературе и в программах для ПК коэффициенты корреляции со временем или фактических уровней с выравненными (т.е. тоже упорядоченными) уровнями тренда таковыми на самом деле не являются и не могут измерять какой-либо «тесноты связи». Чем длиннее период, охватываемый рядом, тем автоматически становятся больше так называемые коэффициенты корреляции при той же самой скорости роста уровней и той же рамой силе колебаний. Таким образом, эти лжекоэффициенты не могут характеризовать соотношение между ролью факторов тенденции и ролью факторов колеблемости [10, С. 71].

Уравнение прямой линии тренда.

Уравнение имеет вид:

$$\tilde{y}_i = a + bt_i ,$$

где $\tilde{y}_i$ – уровень тренда для периода или момента с номером t_i;

a – свободный член уравнения, равный среднему уровню тренда для периода (момента) с нулевым номером t_i;

b – главный параметр линейного тренда – его константа – среднее абсолютное изменение за принятую в ряду единицу времени.

Величина параметров a и b определяется по методу наименьших квадратов путем приравнивания частных первых производных функции $f(a,b) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bt_i)^2$ к нулю.

Имеем:

$$\frac{\partial f}{\partial a} = 2\sum_{i=1}^{n} (y_i - a - bt_i) \cdot (-1) = 0 , \qquad (8.39)$$

$$\frac{\partial f}{\partial b} = 2\sum_{i=1}^{n} (y_i - a - bt_i) \cdot (-t) = 0 . \qquad (8.40)$$

После алгебраических преобразований получаем два «нормальных уравнения» МНК для прямой:

$$na + b\sum_{i=1}^{n} t_i = \sum_{i=1}^{n} y_i \quad , \tag{8.41}$$

$$a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} (y_i t_i) \quad . \tag{8.42}$$

Решая эти уравнения с двумя неизвестными по данным фактического временного ряда y_i (i=1–n), получаем значения a и b. Если номера периодов (моментов) времени отсчитываются от начала ряда так, что первый период (момент) обозначен номером t=1, то свободный член a есть уровень тренда для предыдущего периода (момента), а не первого в ряду, как часто ошибочно полагают. Для первого периода уровень тренда $\tilde{y}_1$ равен $a+b$, для второго $\tilde{y}_2 = a+2b$ и т.д.

Однако рациональнее начало отсчета времени перенести в середину ряда, т.е. при нечетном n – на период (момент) с номером (n+1)/2, а при четном числе уровней ряда – на середину между периодом номером n/2 и (n/2)+1. В последнем случае все номера периодов t_i будут дробными. При нумерации периодов времени точно от середины ряда, половина номеров t_i будут отрицательными числами (аналогично годам до нашей эры), а половина – положительными, сумма их, т.е. $\sum_{i=1}^{n} t_i = 0$. В таком случае система нормальных уравнений МНК распадается на два уравнения с одним неизвестным в каждом [10, с. 71-73]:

$$na = \sum_{i=1}^{n} y_i \quad , \tag{8.43}$$

$$b\sum_{i=1}^{n} t^2 = \sum_{i=1}^{n} y_i t_i \quad . \tag{8.44}$$

Откуда имеем:

$$a = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y} \ , \tag{8.45}$$

$$b = \frac{\sum_{i=1}^{n} (y_i t_i)}{\sum_{i=1}^{n} t_i^2} \ . \tag{8.46}$$

К сожалению, многие компьютерные программы не предусматривают такого упрощения, и нумерация периодов (моментов) в них производится с начала ряда, с номера $t = 1$, причем пользователь никак об этом не предупреждается. При расчетах без компьютера, конечно, следует применять упрощенный прием. Знаменатель в формуле (8.46) при нумерации периодов от середины ряда вычисляется устно при $n \leq 10$, или по формуле [10, с. 73-74]:

$$\sum_{i=-\frac{n+1}{2}}^{+\frac{n+1}{2}} t_i^2 = \frac{n^3 - n}{12} \ .$$

Пример 8.5 - Приведем расчет линейного тренда по временному ряду реальных денежных доходов населения Оренбургской области. Динамика данного показателя с 1998 по 2006 г. представлена в таблице 8.8.

По приведенным данным параметры линейного тренда (при расчете используем формулы (8.45) и (8.46)) составят:

$$a = \bar{y} = \frac{952}{9} = 105{,}8 \ \%,$$

$$b = \frac{222}{60} = 3{,}7 \ \text{п.п. в год.}$$

Уравнение тренда:

$$\tilde{y}_i = 105{,}8 + 3{,}7 \cdot t_i \ ,$$

где $t_i = 0$ в 2002 г.

В среднем реальные денежные доходы населения увеличивались на 3,7 п.п. в год.

Сумма уровней тренда должна равняться сумме фактических уровней [50, с. 78-79].

Таблица 8.8 - Динамика реальных денежных доходов населения Оренбургской области, в процентах к предыдущему году

Год	Уровень, y_i ,%	Номер года, t_i	$y_i\, t_i$	t^2	Тренд $\tilde{y}_i$, %
1	2	3	4	5	6
1998	78	-4	-312	16	91
1999	89	-3	-267	9	95
2000	110	-2	-220	4	98
2001	110	-1	-110	1	102
2002	113	0	0	0	106
2003	113	1	113	1	109
2004	112	2	224	4	113
2005	114	3	342	9	117
2006	113	4	452	16	121
Итого	952	0	222	60	952
Источник: Оренбургская область, 2007: статистический ежегодник / Территориальный орган федеральной службы государственной статистики по Оренбургской области. – Оренбург, 2007. - - 428с.					

Уравнение параболического (II порядка) тренда. Уравнение имеет вид: $\tilde{y}_i = a + bt_i + ct_i^2$. Для вычисления параметров a, b, c по методу наименьших квадратов три частных производных функции: $f(a,b,c) = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2$ приравниваются к нулю и после преобразований получаем систему трех уравнений с тремя неизвестными:

$$na + b\sum_{i=1}^{n} t_i + c\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i \quad (8.47)$$

$$a\sum_{i=1}^{n} t_i + b\sum_{i=1}^{n} t_i^2 + c\sum_{i=1}^{n} t_i^3 = \sum_{i=1}^{n} y_i t_i \quad (8.48)$$

$$a\sum_{i=1}^{n} t_i^2 + b\sum_{i=1}^{n} t_i^3 + c\sum_{i=1}^{n} t_i^4 = \sum_{i=1}^{n} y_i t_i^2 \quad (8.49)$$

При переносе начала отсчета периодов (моментов) времени в середину ряда суммы нечетных степеней номеров этих периодов $\sum t_i$ и $\sum t_i^3$ обращаются в нуль. При этом второе уравнение обращается в уравнение с одним неизвестным, откуда [50]:

$$b = \frac{\sum_{i=1}^{n} y_i t_i}{\sum_{i=1}^{n} t_i^2}.$$

Уравнения (8.47) и (8.49) образуют систему двух уравнений с двумя неизвестными:

$$na + c\sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} t_i^2 + c\sum_{i=1}^{n} t_i^4 = \sum_{i=1}^{n} y_i t_i^2$$

где, напомним,

$$\sum_{i=1}^{n} t_i^2 = \frac{n^3 - n}{12},$$

$$\sum_{i=1}^{n} t_i^4 = \frac{3n^5 - 10n^3 + 7n}{240}.$$

Пример 8.6 - Приведем пример расчета параболического тренда по данным таблицы И.1 приложения И.

Вычисляем параметры параболы:

$$b = \frac{74268252}{96460} \cong 769{,}94,$$

$$105a + 96460c = 11943282,$$
$$96460a + 159486964c = 10971116366.$$

$$a + 918{,}67c = 113745{,}5,$$
$$a + 1653{,}4c = 113737{,}5.$$

$$734{,}7c = -8{,}07;\ c = -0{,}011;\ a = 113755{,}6.$$

Уравнение тренда:

$$\tilde{y}_i = 113755{,}6 + 769{,}94t_i - 0{,}011t_i^2,$$

где t=0 в 1952 г.

Интерпретация параметров тренда такова: численность населения России в 1900-2004 гг. возрастала в среднем за год на 769938 человек с замедлением роста уровней на $2 \cdot 0{,}011 = 0{,}022$ тыс. чел. Средняя численность населения на середину периода была равна 113755,6 тыс. человек.

Если бы параболический тренд вычислялся на ПК по программе, предусматривающей нумерацию лет от начала с номера t = 1, то уравнение имело бы вид [50, с. 79-81]:

$$\tilde{y}_i = 72918 + 771{,}1t_i - 0{,}011t_i^2,$$

где t_i = 1 в 1900 г.

Гиперболическое уравнение тренда. Уравнение имеет вид: $\tilde{y}_i = a + \frac{b}{t_i}$,

т.е. отличается от линейного уравнения тем, что вместо t_i в первой степени включает номера периодов времени (моментов) в минус первой степени: $\frac{1}{t_i}$. Соответственно, нормальные уравнения метода наименьших квадратов получат вид:

$$na + b\sum_{i=1}^{n}\frac{1}{t_i} = \sum_{i=1}^{n} y_i \quad (8.50)$$

,

$$a\sum_{i=1}^{n}\frac{1}{t_i} + b\sum_{i=1}^{n}\frac{1}{t_i^2} = \sum_{i=1}^{n}\frac{y_i}{t_i} \quad (8.51)$$

.

Однако при этом нельзя, в отличие от линейного тренда, переносить начало отсчета периодов времени в середину, так как гипербола не имеет постоянного параметра изменения уровней на протяжении всего периода, и все величины $\frac{1}{t_i}$ должны быть положительными.

Рассмотрим расчет гиперболического уравнения тренда (таблица 8.9) по данным числа прибывших в Оренбургскую область (человек).

Таблица 8.9 - Расчет гиперболического уравнения тренда

Год	y_i, человек	t_i	$\frac{1}{t_i}$	$\frac{1}{t_i^2}$	$\frac{y_i}{t_i}$	Тренд, $\tilde{y}_i$
1	2	3	4	5	6	7
2000	3623,7	1,0	1,0	1,0	3623,7	3510,4
2001	2738,7	2,0	0,5	0,3	1369,3	2878,6
2002	2435,7	3,0	0,3	0,1	811,9	2668,0
2003	2478,3	4,0	0,3	0,1	619,6	2562,8
2004	2475,1	5,0	0,2	0,0	495,0	2499,6
2005	2776,8	6,0	0,2	0,0	462,8	2457,5
2006	2476,0	7,0	0,1	0,0	353,7	2427,4
Итого	19004,3	28,0	2,6	1,5	7736,0	19004,3
Источник: Оренбургская область, 2007: статистический ежегодник / Территориальный орган федеральной службы государственной статистики по Оренбургской области. – Оренбург, 2007. – 428с.						

Нормальные уравнения МНК:

$$7a + 2{,}6b = 19004{,}3,$$

$$2{,}6a + 1{,}5b = 7736.$$

Решая систему уравнений, получаем:

$$a = 2246{,}9;\ b = 1263{,}5.$$

Уравнение гиперболического тренда числа прибывших в Оренбургскую область имеет вид:

$$\tilde{y}_i = 2246{,}9 + \frac{1263{,}5}{t_i},$$

где $t_i = 1$ в 2000 г.

Величина численности прибывших 2247 человек – это предел, к которому стремится сокращение численности прибывших на территорию Оренбургской области [50, с. 81-89].

Оценка параметров экспоненциального, логарифмического и логистического уравнений тренда. Данные типы трендов объединены в одну группу в связи с необходимостью при оценке их параметров прибегать к логарифмированию. При расчете логарифмического уравнения тренда логарифмируют номера периодов (моментов) времени, а при расчете параметров экспоненциального и логистического трендов – сами уровни. Поскольку отрицательные числа не имеют действительных логарифмов, если нужно логарифмировать номера периодов времени, то нельзя переносить начало их отсчета в середину ряда. Если же сами уровни могут принимать отрицательные значения, например, уровни финансового результата от реализации, уровни температуры воздуха или почвы, то необходимо перенести начало отсчета уровней на величину, алгебраически меньшую наименьшего реального уровня. Например, температуру следует выразить не в градусах Цельсия, а в Кельвинах, финансовый результат при наибольшем убытке 83 млн. р., отсчитывать от –100 млн. р., чтобы самый низкий уровень выразился как 17 млн. р. По окончании расчета тренда нетрудно восстановить обычные единицы измерения. Так, получив тренд финансового результата при отсчете от –100 млн. р. как:

$$\tilde{y}_i = 27 \cdot 1{,}028^{t_i},$$

нужно по нему рассчитать все уровни тренда, а затем прибавить к ним величину –100 млн. р. Начиная с t = 48, уровни тренда станут положительными числами в обычном смысле: 47<[ln(100:27): ln 1,028]< 48.

Экспоненциальное уравнение тренда.

Формула уравнения имеет вид:

$$\tilde{y}_i = a \cdot k^{t_i}.$$

Для нахождения параметров a и k уравнение логарифмируем:

$$\ln \tilde{y}_i = \ln a + t_i \ln k.$$

В такой форме, т.е. для логарифмов, уравнение соответствует линейному, и, следовательно, метод наименьших квадратов дает для логарифмов a и k нормальные уравнения, аналогичные таковым для параметров a и b линейного тренда:

$$n \ln a + \ln k \sum_{i=1}^{n} t_i = \sum_{i=1}^{n} \ln y_i, \tag{8.52}$$

$$\ln a \sum_{i=1}^{n} t_i + \ln k \sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} t_i \ln y_i. \tag{8.53}$$

Так как номера периодов времени не логарифмируются, можно перенести начало их отсчета в середину ряда и упростить систему:

$$n \ln a = \sum_{i=1}^{n} \ln y_i, \text{ откуда } \ln a = \overline{\ln y_i},$$

$$\ln k \sum_{i=1}^{n} t_i^2 = \sum_{i=1}^{n} t_i \ln y_i, \text{ откуда } \ln k = \frac{\sum_{i=1}^{n} t_i \ln y_i}{\sum_{i=1}^{n} t_i^2}.$$

Пример 8.7 - По данным таблицы К.1 приложения К получим:

$$\ln a = \frac{670{,}18}{92} = 7{,}2846;\ \ a = 1457{,}7\ ;$$

$$\ln k = \frac{3564{,}6}{64883} = 0{,}0549;\ \ k = 1{,}0565\ .$$

Уравнение тренда примет вид:

$$\tilde{y}_i = 1457{,}7 \cdot 1{,}0565^{t_i}\ ,$$

где t = 0,5 в 1946 г.

Итак, национальное богатство в период с 1900 по 1991г. возрастало со среднегодовым темпом роста, равным корню девяносто первой степени из среднего темпа за десятилетие, найденного по данным табл. 6.9, т.е. $\sqrt[91]{1{,}0565} = 1{,}0006$, или 0,06 % прироста в год [50, с. 89-95].

*Логарифмическое уравнение тренда.*Особенность этого типа тренда заключается в том, что логарифмировать необходимо номера периодов (моментов) времени: $\tilde{y} = a + b\ln t.$ Следовательно, все номера должны быть положительными числами. Однако это вовсе не означает, что нумерацию следует начинать с числа 1. Дело в том, что величина логарифма быстро возрастает при переходе от единицы к двум: натуральный логарифм единицы равен нулю, а логарифм двух равен 0,693, имеем рост на 0,693; в то же время логарифм четырех равен 1,386, а логарифм пяти равен 1,609, имеем прирост лишь на 0,223 и т.д. Если уровень изучаемого ряда в начале возрастает втрое быстрее, чем между четвертым и пятым периодом, тогда нумерация от единицы допустима. Если же уменьшение прироста уровней происходит значительно медленнее, нумерацию периодов (моментов) следует начинать не с единицы, а с большего числа.

Пример 8.8 - Покажем методику расчета логарифмического уравнения тренда на примере динамики естественной убыли населения в Оренбургской области за 1985 – 2006 гг. (таблица 8.10). Система нормальных уравнений для оценки параметров тренда имеет вид:

$$\begin{cases} a \cdot n + b \cdot \sum \ln t = \sum y; \\ a \cdot \sum \ln t + b \cdot \sum (\ln t)^2 = \sum y \cdot \ln t. \end{cases}$$

Таблица 8.10 - Расчет логарифмического тренда коэффициента естественной убыли населения в Оренбургской области

Год	y_i	t_i	$\ln t_i$	$(\ln t)^2$	$y \cdot \ln t$	$\tilde{y}_i$
1	2	3	4	5	6	7
1985	8,4	1	0	0	0	13,7
1986	10,4	2	0,693	0,480	7,209	9,4
1987	9,5	3	1,099	1,207	10,437	6,9
1988	8,1	4	1,386	1,922	11,229	5,1
1989	7	5	1,609	2,590	11,266	3,7
1990	5,8	6	1,792	3,210	10,392	2,6
1991	3,5	7	1,946	3,787	6,811	1,6
1992	1,5	8	2,079	4,324	3,119	0,8
1993	-2	9	2,197	4,828	-4,394	0,1
1994	-3,1	10	2,303	5,302	-7,138	-0,6
1995	-3,3	11	2,398	5,750	-7,913	-1,2
1996	-3,2	12	2,485	6,175	-7,952	-1,7
1997	-3,3	13	2,565	6,579	-8,464	-2,2
1998	-2,9	14	2,639	6,965	-7,653	-2,7
1999	-4,6	15	2,708	7,334	-12,457	-3,1
2000	-4,7	16	2,773	7,687	-13,031	-3,5
2001	-4,7	17	2,833	8,027	-13,316	-3,9
2002	-4,3	18	2,890	8,354	-12,429	-4,2
2003	-4,4	19	2,944	8,670	-12,956	-4,6
2004	-4,1	20	2,996	8,974	-12,283	-4,9
2005	-5	21	3,045	9,269	-15,223	-5,2
2006	-3,9	22	3,091	9,555	-12,055	-5,5
Итого	0,7	-	48,471	120,988	-86,801	0,7

Источник: Оренбургская область, 2007: статистический ежегодник / Территориальный орган федеральной службы государственной статистики по Оренбургской области. – Оренбург, 2007. – 428 с.

По данным таблицы 8.10 составим систему уравнений:

$$\begin{cases} 22 \cdot a + 48{,}471 \cdot b = 0{,}7; \\ 48{,}471 \cdot a + 120{,}99 \cdot b = -86{,}8. \end{cases}$$

Решив систему, получим: $a = 13{,}744$, $b = -6{,}224$.

Тогда уравнение логарифмического тренда имеет вид:

$$\tilde{y} = 13{,}744 - 6{,}224 \cdot \ln t,$$

где $t = 1$ в 1985 году.

По этому уравнению рассчитаны уровни тренда $\tilde{y}_i$ в таблице 8.10. Суммы фактических и теоретических уровней полностью совпали. Кривая хорошо отражает тенденцию.

Логистическое уравнение тренда. Уравнение имеет вид:

$$\tilde{y}_i = \frac{y_{\max} - y_{\min}}{e^{a_0 + a_1 t_i} + 1} + y_{\min}.$$

При расчете этого уравнения логарифмируют величину, производную от уровней ряда, но не номера периодов (моментов) времени. Поэтому рационально производить эту нумерацию от середины ряда. Особенностью логистического тренда является этап обоснования значений максимального и минимального уровней временного ряда. Это обоснование производится на основе, во-первых, уровней фактического ряда, а во-вторых, теоретических соображений, т.е. внешних по отношению к статистике, относящихся к содержанию изучаемого процесса [50, с. 96-98].

Уравнение логистического тренда в общем виде непосредственно логарифмировать невозможно. Преобразуем его в форму:

$$\frac{\tilde{y}_{\max} - \tilde{y}_{\min}}{\tilde{y}_i - \tilde{y}_{\min}} - 1 = e^{a_0 + a_1 t_i}$$

и обозначим его левую часть, т.е.

$$\frac{\tilde{y}_{\max}-\tilde{y}_{\min}}{\tilde{y}_i-\tilde{y}_{\min}}-1=\tilde{\zeta}_i,\ \text{т.е. } \tilde{\zeta}_i=e^{a_0+a_1t_i};\ \ln\tilde{\zeta}_i=a_0+a_1t_i.$$

Условие метода наименьших квадратов:

$$\sum_{i=1}^{n}\left(\ln\zeta_i-\ln\tilde{\zeta}_i\right)^2\to\min.$$

Подставляя значение $\ln\tilde{\zeta}_i$, имеем:

$$\sum_{i=1}^{n}\left(\ln\zeta_i-a_0-a_1t_i\right)^2\to\min.$$

После вычисления частных производных по a_0 и по a_1, получаем нормальные уравнения МНК для логистической кривой, аналогичные таковым для прямой линии, т.к. заменой на ζ фактически проведена линеаризация функции логистической кривой:

$$na_0+a_1\sum_{i=1}^{n}t_i=\sum_{i=1}^{n}\ln\zeta_i \tag{8.54}$$

$$a_0\sum_{i=1}^{n}t_i+a_1\sum_{i=1}^{n}t_i^2=\sum_{i=1}^{n}t_i\ln\zeta_i \tag{8.55}$$

При переносе начала отсчета периодов (моментов) времени в середину ряда система упрощается до двух уравнений с одним неизвестным в каждом из них:

$$na_0=\sum_{i=1}^{n}\ln\zeta_i,\ \text{откуда}\ a_0=\overline{\ln\zeta_i},$$

$$a_1\sum_{i=1}^{n}t_i^2=\sum_{i=1}^{n}t_i\ln\zeta_i,\ \text{откуда}\ a_1=\frac{\sum_{i=1}^{n}t_i\ln\zeta_i}{\sum_{i=1}^{n}t_i^2}.$$

Итак, алгоритм расчета логистической кривой состоит из десяти этапов:

1) обоснование величин $\tilde{y}_{\max}$ и $\tilde{y}_{\min}$;

2) вычисление по фактическому временному ряду значений $\zeta_i = \frac{\tilde{y}_{\max} - \tilde{y}_{\min}}{y_i - \tilde{y}_{\min}} - 1$;

3) вычисление $\ln \zeta_i$;

4) нумерация периодов или моментов времени от середины ряда;

5) умножение $\ln \zeta_i$ на t_i;

6) подсчет итоговых сумм $\sum_{i=1}^{n} \ln \zeta_i; \sum_{i=1}^{n} t_i \ln \zeta_i$.

7) вычисление a_0 и a_1;

8) вычисление $\ln \tilde{\zeta}_i = a_0 + a_1 t_i$;

9) вычисление $\tilde{\zeta}_i = \exp(a_0 + a_1 t_i)$ для всех периодов;

10) вычисление уровней тренда $\tilde{y}_i = \frac{\tilde{y}_{\max} - \tilde{y}_{\min}}{\tilde{\zeta}_i + 1} + \tilde{y}_{\min}$.

Проведем расчет логистического тренда по данным рисунка 8.6 и таблицы 8.11.

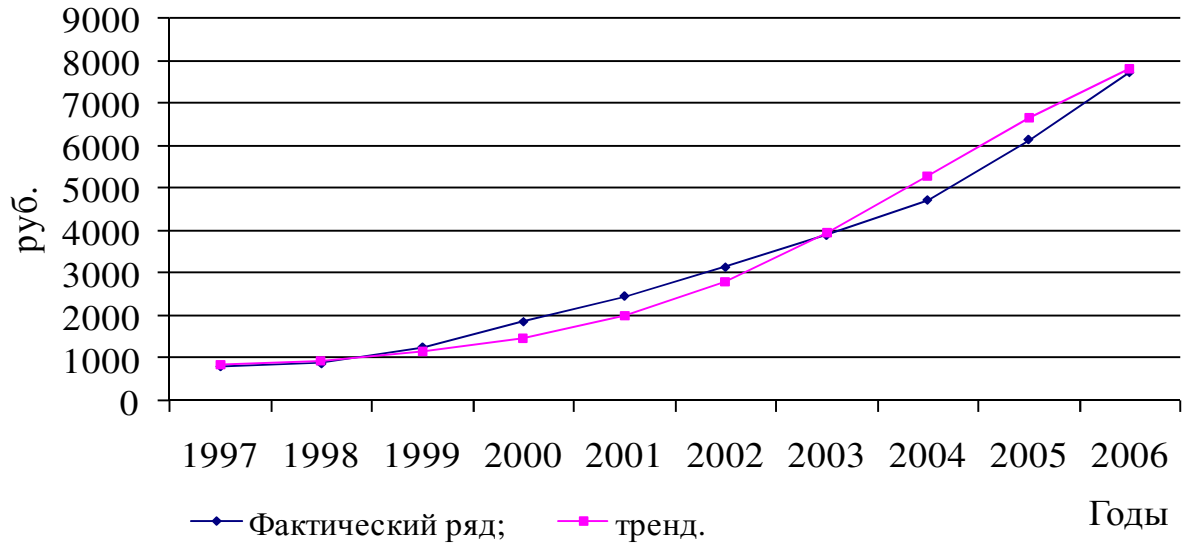


Рисунок 8.6 - Логистическая тенденция динамики среднемесячной номинальной начисленной заработной платы работающих в экономике, р. (до 1998 г. - тыс.р.)

Таблица 8.11 - Расчет логистического тренда среднемесячной номинальной начисленной заработной платы работающих в экономике, р. (до 1998 г. - тыс.р.)

Год	y_i,	$\zeta_i = \frac{10000-700}{y_i-700}-1$	$\ln \zeta_i$	t_i	$t_i \ln \zeta_i$	t^2	$\tilde{\zeta}_i$	Тренд $\tilde{y}_i = \frac{10000-700}{\tilde{\zeta}_i+1}+700$
1	2	3	4	5	6	7	8	9
1997	788	104,68	4,7	-4,5	-20,9	20,3	71,1	829
1998	855	59,00	4,1	-3,5	-14,3	12,3	38,8	934
1999	1249	15,94	2,8	-2,5	-6,9	6,3	21,2	1120
2000	1849	7,09	2,0	-1,5	-2,9	2,3	11,5	1442
2001	2460	4,28	1,5	-0,5	-0,7	0,3	6,3	1976
2002	3142	2,81	1,0	0,5	0,5	0,3	3,4	2799
2003	3898	1,91	0,6	1,5	1,0	2,3	1,9	3940
2004	4735	1,30	0,3	2,5	0,7	6,3	1,0	5304
2005	6164	0,70	-0,4	3,5	-1,2	12,3	0,6	6676
2006	7753	0,32	-1,1	4,5	-5,1	20,3	0,3	7835
Итого	32893	-	15,4	-	-50,0	82,5	-	32853

Источник: Оренбургская область, 2007: статистический ежегодник / Территориальный орган федеральной службы государственной статистики по Оренбургской области. – Оренбург, 2007. – 428 с.

Исходя из границ периода времени и фактических уровней ряда, получаем:

$$\tilde{y}_{\min} = 700;\ \tilde{y}_{\max} = 10000;$$

$$a_0 = \frac{15{,}4}{10} = 1{,}154\ ;\ a_1 = \frac{-50{,}0}{82{,}5} = -0{,}606$$

$$\tilde{\zeta}_i = \exp[1{,}154 + t_i(-0{,}606)]\,.$$

Уравнение логистического тренда среднемесячной номинальной начисленной заработной платы работающих в экономике имеет вид:

$$\tilde{y}_i = \frac{10000 - 700}{e^{1{,}154 - 0{,}606 \cdot t_i} + 1} + 700\,.$$

Рисунок 8.6 показывает достаточно близкое приближение логистической кривой к исходным данным. Напомним, что, в отличие от прямой и параболы, алгоритм расчета других кривых не предусматривает автоматического равенства сумм выравненных и фактических уровней, они совпадают только при идеальном выражении тенденции ряда данным уравнением тренда [50, с. 98-101].

8.4 Моделирование сезонных и циклических колебаний

Известно несколько подходов к анализу структуры временных рядов, содержащих сезонные и циклические колебания (моделирование циклических колебаний в целом осуществляется аналогично моделированию сезонных колебаний, поэтому мы рассмотрим только методы моделирования последних).

Простейший подход – расчет значений сезонной компоненты методом скользящей средней и построение аддитивной или мультипликативной модели ВР.

Алгоритм построения тренд – сезонной аддитивной модели:

1) сглаживание временного ряда с помощью простой скользящей средней. Период скольжения должен быть равен 1 году (если период четный, то проводится центрирование скользящей средней);

2) рассчитывают абсолютные показатели сезонности:

$$S_i = y_i - \tilde{y}\,, \tag{8.56}$$

где $\tilde{y}$ - выровненные скользящие средние;

3) рассчитываются средние показатели сезонности для одноименных кварталов (месяцев):

$$\overline{S}_i = \frac{1}{n}\sum S_i\,; \tag{8.57}$$

4) если $\sum \overline{S}_i \neq 0$, проводится корректировка сезонной компоненты:

$$\hat{S}_i = \overline{S}_i - \frac{1}{n}\sum_{i=1}^{n} \overline{S}_i\,; \tag{8.58}$$

5) проводим десезоналирование ВР: из исходных уровней вычитаем скорректированную сезонную компоненту:

$$y_i - \hat{S}_i\,; \tag{8.59}$$

6) по десезоналированному ВР проводим аналитическое выравнивание;

7) рассчитываем тренд с учетом сезонности:

$$y_S = \hat{y}_t + \hat{S}_i\,. \tag{8.60}$$

Рассмотрим пример построения аддитивной тренд - сезонной модели.

Графический анализ исходного временного ряда (рисунок 8.7) свидетельствует о наличии трендовой компоненты, характер которой близок к линейному развитию: имеется устойчивая, ярко выраженная тенденция роста объемов продаж.

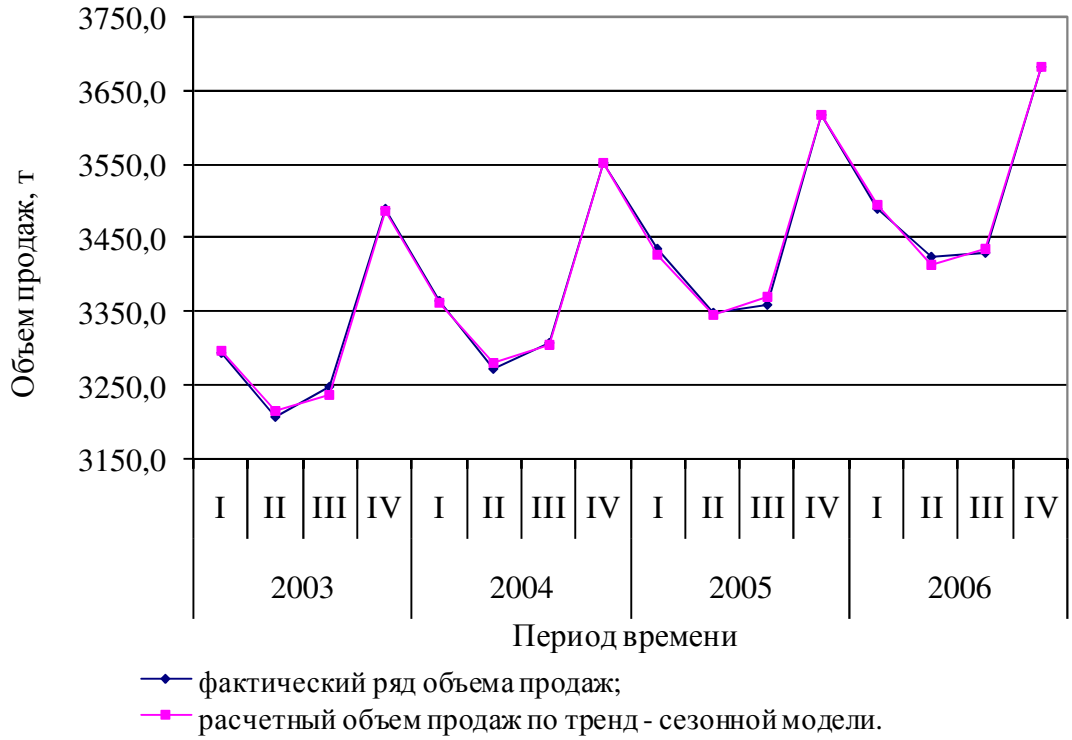


Рисунок 8.7 - Квартальная динамика объема продаж

Также отчетливо видны сезонные колебания (период которых равен одному году). Наиболее существенные «всплески» в динамике показателя просматриваются в IV квартале. Так как амплитуда сезонных колебаний остается примерно постоянной, то для описания и прогнозирования динамики временного ряда можно использовать аддитивную модель.

Проведем сглаживание временного ряда с помощью центрированной скользящей средней по формуле (период скольжения равен одному году, т.е. для нашего примера он равен 4):

$$\tilde{y}_i = \frac{1/2 \cdot y_{i-2} + y_{i-1} + y_i + y_{i+1} + 1/2 \cdot y_{i+2}}{4};$$

$$\tilde{y}_3 = \frac{1/2 \cdot 3294{,}1 + 3205{,}9 + 3247{,}1 + 3488{,}2 + 1/2 \cdot 3364{,}7}{4} = 3317{,}6$$ и т.д.

Рассчитаем абсолютные показатели сезонности по формуле $S_i = y_i - \tilde{y}_i$. Результаты расчетов скользящей средней и показателя сезонности представлены в таблице 8.12.

Таблица 8.12 - Динамика объема продаж продукции

Год	Номер квартала	y_i	$\tilde{y}_i$	S_i
2003	I	3294,1	–	–
	II	3205,9	–	–
	III	3247,1	3317,6	-70,6
	IV	3488,2	3334,6	153,7
2004	I	3364,7	3350,0	14,7
	II	3270,6	3365,4	-94,9
	III	3305,9	3382,4	-76,5
	IV	3552,9	3400,7	152,2
2005	I	3435,3	3416,9	18,4
	II	3347,1	3431,6	-84,6
	III	3358,8	3446,3	-87,5
	IV	3617,6	3462,5	155,1
2006	I	3488,2	3480,9	7,4
	II	3423,5	3497,8	-74,3
	III	3429,4	–	–
	IV	3682,4	–	–

Определим средние показатели сезонности по формуле**:** $\overline{S}_j = \frac{1}{n}\sum S_i$ **,** т.е. для I квартала средний показатель сезонности составит:

$$\overline{S}_1 = \frac{14{,}7 + 18{,}4 + 7{,}4}{3} = 13{,}4804 .$$

Аналогично рассчитывают показатели для других кварталов.

Так как сумма средних показателей сезонности не равна нулю, проведем их корректировку по формуле**:**

$$\tilde{S}_j = \bar{S}_j - \frac{1}{n}\sum_{j=1}^{n}\bar{S}_j \text{ .}$$

Скорректированный показатель сезонности для I квартала составит:

$\tilde{S}_1 = 13{,}4804 - \frac{1}{4}\cdot 4{,}4118 = 12{,}3775$ и т.д.

Результаты расчетов средних и скорректированных показателей сезонности заносим в таблице 8.13.

Таблица 8.13 - Оценивание сезонной компоненты в аддитивной модели

Номер квартала	$\bar{S}_j$	$\tilde{S}_j$
I	13,4804	12,3775
II	–84,5588	–85,6618
III	–78,1863	–79,2892
IV	153,6765	152,5735
Итого	4,4118	0

На следующем этапе определим десезонализированный ряд объема продаж: из исходных уровней вычитаем скорректированную сезонную компоненту: $y_i - \tilde{S}_i$**.** По десезонализированному временному ряду проводим аналитическое выравнивание по линейному тренду и рассчитываем тренд с учетом сезонности: $y_S = \tilde{y}_t + \tilde{S}_i$.

Уравнение тренда имеет вид:

$$\tilde{y}_i = 3267{,}2 + 16{,}442t \ (R^2 = 0{,}993).$$

Результаты расчетов представлены в таблице 8.14.

Таблица 8.14 - Прогнозирование объема продаж с помощью аддитивной тренд – сезонной модели

Год	Номер квартала	t	y_i	$\tilde{S}_j$	$y_i - \tilde{S}_j$	$\tilde{y}_i$	y_S
2003	I	1	3294,1	12,3775	3281,7	3283,9	3296,3
	II	2	3205,9	-85,662	3291,5	3300,4	3214,7
	III	3	3247,1	-79,289	3326,3	3316,8	3237,5
	IV	4	3488,2	152,574	3335,7	3333,3	3485,8
2004	I	5	3364,7	12,3775	3352,3	3349,7	3362,1
	II	6	3270,6	-85,662	3356,3	3366,2	3280,5
	III	7	3305,9	-79,289	3385,2	3382,6	3303,3
	IV	8	3552,9	152,574	3400,4	3399,0	3551,6
2005	I	9	3435,3	12,3775	3422,9	3415,5	3427,9
	II	10	3347,1	-85,662	3432,7	3431,9	3346,3
	III	11	3358,8	-79,289	3438,1	3448,4	3369,1
	IV	12	3617,6	152,574	3465,1	3464,8	3617,4
2006	I	13	3488,2	12,3775	3475,9	3481,2	3493,6
	II	14	3423,5	-85,662	3509,2	3497,7	3412,0
	III	15	3429,4	-79,289	3508,7	3514,1	3434,8
	IV	16	3682,4	152,574	3529,8	3530,6	3683,1
2007*	I	17*	–	12,3775	–	3547,0	3559,4
	II	18*	–	-85,662	–	3563,5	3477,8
* Прогнозируемый уровень							

Ожидаемый объем продаж в первом полугодии составит:

$$\tilde{y}_{Sp} = 3559{,}4 + 3477{,}8 = 7037{,}2 \text{ т.}$$

При мультипликативной модели уровень ВР можно представить в виде сомножителей:

$$y_i = \hat{y}_t \cdot K_S \cdot E, \tag{8.61}$$

где K_S - коэффициент сезонности;

E – коэффициент влияния случайности $\left(\frac{y_i}{y_S}\right)$.

Алгоритм построения тренд – сезонной мультипликативной модели:

1) сглаживание ВР с помощью скользящей средней;

2) рассчитываем коэффициент сезонности

$$K_S = \frac{y_i}{\tilde{y}_i}; \tag{8.62}$$

3) определяем средние показатели сезонности для одноименных кварталов (месяцев)

$$\overline{K}_j = \frac{1}{n}\sum K_{Si}; \tag{8.63}$$

4) если при поквартальном наблюдении $\sum \overline{K} \neq 4$, а при помесячном $\sum \overline{K} \neq 12$, то выполняется корректировка коэффициента сезонности

$$\hat{K}_j = \overline{K}_j \cdot \frac{4(12)}{\sum \overline{K}_j}; \tag{8.64}$$

5) исключаем сезонность из уровней ряда

$$\frac{y_i}{\hat{K}_j}; \tag{8.65}$$

6) проводится аналитическое выравнивание десезоналированного ряда;

7) рассчитываются уровни ВР, обусловленные влиянием тенденции и сезонности

$$y_S = \hat{y}_t \cdot \hat{K}_j. \tag{8.66}$$

Аддитивная модель целесообразна, если размах сезонных колебаний изменяется слабо.

Рассмотрим пример построения мультипликативной тренд сезонной модели по условным данным об объеме производства.

Графический анализ исходного временного ряда (рисунок 8.8) свидетельствует о наличии трендовой компоненты, характер которой близок к линейному развитию: имеется устойчивая, ярко выраженная тенденция снижения объема производства.

Рисунок 8.8 - Квартальная динамика объема производства

Также отчетливо видны сезонные колебания (период которых равен 1 году). Наиболее существенные «всплески» в динамике показателя просматриваются в I квартале. Так как амплитуда сезонных колебаний постепенно уменьшается, то для описания и прогнозирования динамики временного ряда можно использовать мультипликативную модель.

Проведем сглаживание временного ряда с помощью центрированной скользящей средней по формуле (период скольжения равен 1 году, т.е. для нашего примера он равен 4):

$$\tilde{y}_i = \frac{1/2 \cdot y_{i-2} + y_{i-1} + y_i + y_{i+1} + 1/2 \cdot y_{i+2}}{4}.$$

$$\tilde{y}_3 = \frac{1/2 \cdot 92{,}0 + 73{,}6 + 58{,}9 + 80{,}9 + 1/2 \cdot 88{,}3}{4} = 75{,}9 \text{ и т.д.}$$

Рассчитаем коэффициенты сезонности по формуле

$$K_S = \frac{y_i}{\tilde{y}_i}.$$

Результаты расчетов скользящей средней и коэффициента сезонности представлены в таблице 8.15.

Определяем средние показатели сезонности для одноименных кварталов (месяцев):

$$\overline{K}_j = \frac{1}{n} \sum K_{Si}.$$

т.е. для I квартала средний коэффициент сезонности составит:

$$\overline{K}_1 = \frac{1{,}193 + +1{,}197 + 1{,}212}{3} = 1{,}201.$$

Аналогично рассчитывают средние коэффициенты сезонности и для других кварталов.

Таблица 8.15 - Разложение уровней ряда по мультипликативной модели

Год	Номер квартала	t	y_i	$\tilde{y}_i$	K_S	$\tilde{K}_j$	$\frac{y_i}{\tilde{K}_j}$	$\tilde{y}_i$	y_S
2003	I	1	92,0	-	-	1,200	76,6	79,3	95,2
	II	2	73,6	-	-	0,948	77,7	77,8	73,7
	III	3	58,9	75,9	0,776	0,785	75,0	76,2	59,8
	IV	4	80,9	75,0	1,080	1,068	75,8	74,7	79,8
2004	I	5	88,3	74,0	1,193	1,200	73,6	73,1	87,8
	II	6	69,9	72,7	0,962	0,948	73,8	71,6	67,8
	III	7	55,2	70,8	0,779	0,785	70,3	70,1	55,0
	IV	8	73,6	69,0	1,067	1,068	68,9	68,5	73,2
2005	I	9	80,9	67,6	1,197	1,200	67,4	67,0	80,4
	II	10	62,5	66,2	0,944	0,948	66,0	65,4	62,0
	III	11	51,5	64,4	0,800	0,785	65,6	63,9	50,1
	IV	12	66,2	62,5	1,059	1,068	62,0	62,3	66,6
2006	I	13	73,6	60,7	1,212	1,200	61,3	60,8	72,9
	II	14	55,2	58,9	0,938	0,948	58,2	59,2	56,1
	III	15	44,1	-	-	0,785	56,3	57,7	45,3
	IV	16	58,9	-	-	1,068	55,1	56,1	59,9
2007*	I	17	-	-	-	-	-	54,6	65,5
	II	18	-	-	-	-	-	53,0	50,3
* Прогнозируемый уровень									

Так как сумма средних коэффициентов сезонности не равна 4, проведем их корректировку по формуле:

$$\tilde{K}_j = \overline{K}_j \cdot \frac{4}{\sum \overline{K}_j}.$$

Так, скорректированный коэффициент сезонности для I квартала составит:

$$\tilde{K}_j = 1{,}201 \cdot \frac{4}{4{,}002} = 1{,}200 \text{ и т.д.}$$

Результаты расчетов средних и скорректированных показателей сезонности заносим в таблице 8.16.

Таблица 8.16 - Оценивание сезонной компоненты в мультипликативной модели

Номер квартала	$\overline{K}_j$	$\tilde{K}_j$
1	1,201	1,200
2	0,948	0,948
3	0,785	0,785
4	1,068	1,067
Итого	4,002	4,000

На следующем этапе определим десезоналированный ряд объема производства:

$$\frac{y_i}{\tilde{K}_j}.$$

По десезонализированному временному ряду проводим аналитическое выравнивание по линейному тренду и рассчитываем тренд с учетом сезонности:

$$y_S = \tilde{y}_t \cdot \tilde{K}_j.$$

Уравнение тренда имеет вид:

$$\tilde{y}_t = 80{,}881 - 1{,}5467 \cdot t \ \ (R^2 = 0{,}9723).$$

Ожидаемый объем производства в первом полугодии составит:

$$\tilde{y}_{Sp} = 65{,}5 + 50{,}3 = 115{,}8 \text{ тыс. т.}$$

Моделирование сезонных колебаний с помощью фиктивных переменных.

Рассмотрим ещё один метод моделирования ВР, содержащего сезонные колебания, - построение модели регрессии с включением фактора времени и фиктивных переменных.

Количество фиктивных переменных в такой модели должно быть на единицу меньше числа моментов (периодов) времени внутри одного цикла колебаний. Каждая фиктивная переменная отражает сезонную (циклическую) компоненту ВР для какого – либо одного периода. Она равна 1 для данного периода и нулю для всех остальных.

Пусть имеется ВР, содержащий циклические колебания периодичностью К. Модель регрессии с фиктивными переменными для этого ряда:

$$y_t = a + bt + c_1 x_1 + ... + c_j x_j + ... + c_{k-1} x_{k-1} + \varepsilon_t, \qquad (8.67)$$

где $x_j = \begin{cases} 1 \text{ для каждого j внутри каждого цикла;} \\ 0 \text{ во всех остальных случаях.} \end{cases}$.

Например, при моделировании сезонных колебаний на основе поквартальных данных за несколько лет число кварталов внутри одного года K=4, а общий вид модели:

$$y_t = a + bt + c_1 x_1 + c_2 x_2 + c_3 x_3 + \varepsilon_{t,} \qquad (8.68)$$

где $x_1 = \begin{cases} 1 \text{ для 1 квартала} \\ 0 \text{ во всех остальных случаях} \end{cases}$;

$x_2 = \begin{cases} 1 \text{ для 2 квартала} \\ 0 \text{ во всех остальных случаях} \end{cases}$;

$x_3 = \begin{cases} 1 \text{ для 3 квартала} \\ 0 \text{ во всех остальных случаях} \end{cases}$.

Уравнение тренда для каждого квартала будет иметь следующий вид:

- для 1 квартала: $y_t = a + bt + c_1 + \varepsilon_t$; (8.69)

- для 2 квартала: $y_t = a + bt + c_2 + \varepsilon_t$; (8.70)

- для 3 квартала: $y_t = a + bt + c_3 + \varepsilon_t$; (8.71)

- для 4 квартала: $y_t = a + bt + \varepsilon_t$. (8.72)

Таким образом, фиктивные переменные позволяют дифференцировать величину свободного члена уравнения регрессии для каждого квартала. Она составит:

- для 1 квартала ($a+c_1$);

- для 2 квартала ($a+c_2$);

- для 3 квартала ($a+c_3$);

- для 4 квартала a.

Параметр b в этой модели характеризует среднее абсолютное изменение уровней ряда под воздействием тенденции.

Рассмотрим пример построения уравнения регрессии с включением фактора времени и фиктивных переменных по условным данным об объеме производства. Модель для квартальной динамики имеет вид:

$$y_t = a + bt + c_1 x_1 + c_2 x_2 + c_3 x_3 + \varepsilon_{t,}$$

где $x_1 = \begin{cases} 1 \text{ для I квартала;} \\ 0 \text{ во всех остальных случаях.} \end{cases}$

$x_2 = \begin{cases} 1 \text{ для II квартала;} \\ 0 \text{ во всех остальных случаях.} \end{cases}$

$x_3 = \begin{cases} 1 \text{ для III квартала;} \\ 0 \text{ во всех остальных случаях.} \end{cases}$

Оценим параметры уравнения традиционным МНК с помощью табличного редактора Excel (таблица 8.17).

Таблица 8.17 - Исходные данные для расчета параметров уравнения регрессии с фиктивными переменными во временном ряду объема производства, т

Год	Номер квар-тала	y	t	x_1	x_2	x_3	$\tilde{y}_t$			
							I кв.	II кв.	III кв.	IV кв.
2003	I	92,0	1	1	0	0	92,9			
	II	73,6	2	0	1	0		74,5		
	III	58,9	3	0	0	1			61,7	
	IV	80,9	4	0	0	0				79,1
2004	I	88,3	5	1	0	0	86,8			
	II	69,9	6	0	1	0		68,4		
	III	55,2	7	0	0	1			55,5	
	IV	73,6	8	0	0	0				73,0
2005	I	80,9	9	1	0	0	80,6			
	II	62,5	10	0	1	0		62,2		
	III	51,5	11	0	0	1			49,3	
	IV	66,2	12	0	0	0				66,8
2006	I	73,6	13	1	0	0	74,5			
	II	55,2	14	0	1	0		56,1		
	III	44,1	15	0	0	1			43,2	
	IV	58,9	16	0	0	0				60,7
2007*	I		17				59,1			
	II		18					57,6		
* Прогноз										

Уравнение регрессии примет вид:

$$\tilde{y}_t = 85{,}3 - 1{,}54 \cdot t + 9{,}17 \cdot x_1 - 7{,}68 \cdot x_2 - 19{,}02 \cdot x_3 .$$

Параметры c_1, c_2, c_3 характеризуют отклонения уровней временного ряда от уровней, учитывающих сезонные воздействия в IV квартале. Величина параметра $b = -1{,}54$ говорит о том, что в среднем за квартал происходит снижение объема производства на 1,54 тонны (рисунок 8.9).

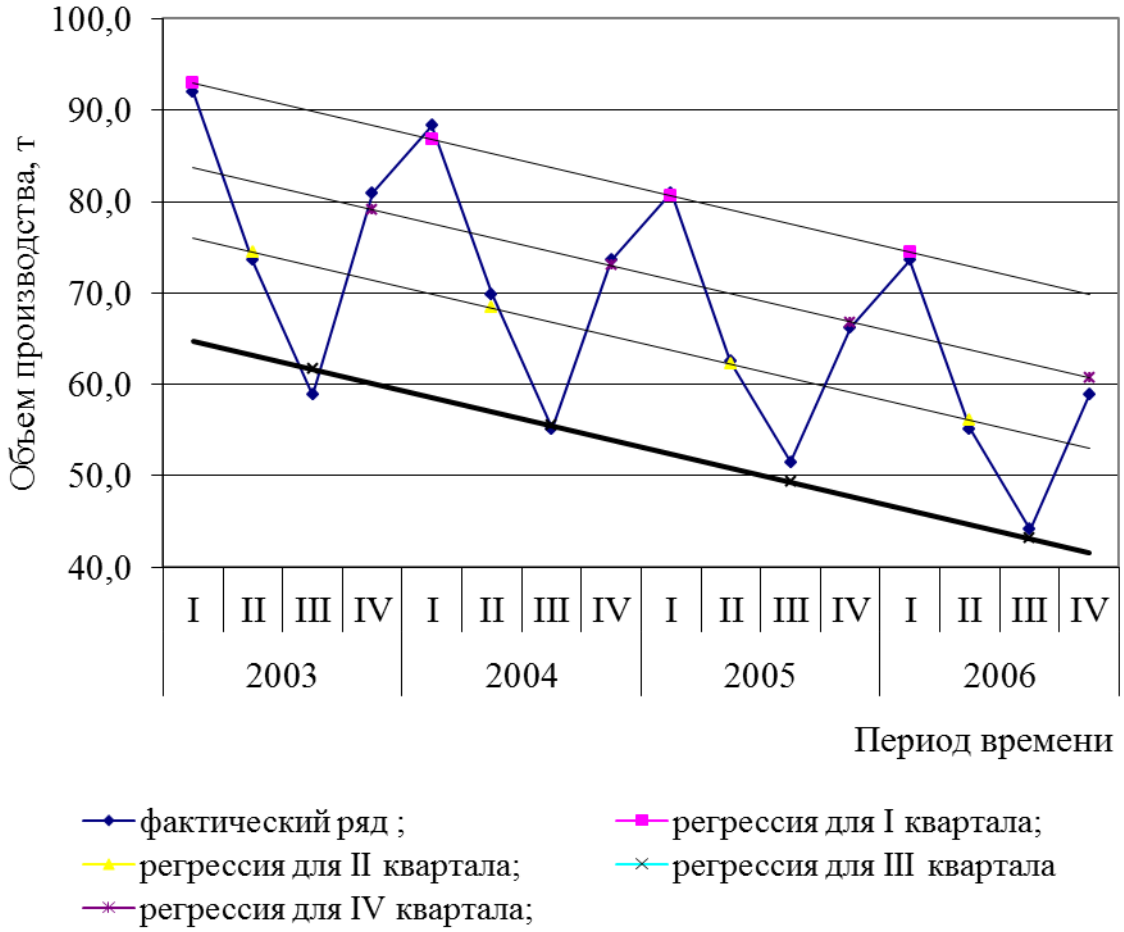


Рисунок 8.9 - Моделирование сезонных колебаний объема производства с помощью фиктивных переменных

Уравнение тренда для каждого квартала будет иметь следующий вид:

- для I квартала: $\tilde{y}_t = 94{,}48 - 1{,}54t$;
- для II квартала: $\tilde{y}_t = 77{,}62 - 1{,}54t$;
- для III квартала: $\tilde{y}_t = 66{,}29 - 1{,}54t$;
- для IV квартала: $\tilde{y}_t = 85{,}3 - 1{,}54t$.

Результаты оценивания модели представлены в таблицах 8.18, 8.19, 8.20.

Таблица 8.18 - Регрессионная статистика

Показатель	Значение
Множественный *R*	0,99
R-квадрат	0,99
Нормированный *R*-квадрат	0,99
Стандартная ошибка	1,62
Наблюдения	16

Таблица 8.19 - Дисперсионный анализ

Показатель	*df*	*SS*	*MS*	*F*	Значимость *F*
Регрессия	4	2758,4	689,6	262,2	0,000
Остаток	11	28,93	2,6		
Итого	15	2787,2			

Таблица 8.20 - Параметры уравнения регрессии

Параметр	Коэффициент	Стандартная ошибка	*t*-статистика	*P*-значение	Нижние 95 %	Верхние 95 %
a	85,30	1,22	70,14	0,00	82,63	87,98
b	–1,54	0,09	–16,99	0,00	–1,74	–1,34
c_1	9,17	1,18	7,78	0,00	6,58	11,77
c_2	–7,68	1,16	–6,61	0,00	–10,23	–5,12
c_3	–19,02	1,15	–16,53	0,00	–21,55	–16,48

Чтобы получить теоретические значения объема производства на I и II кварталы 2007 г. необходимо в соответствующее уравнение регрессии подставить следующие значения фактора времени *t*.

Так, прогноз на I квартал составит:

$$\tilde{y}_t = 94{,}48 - 1{,}54 \cdot 17 = 59{,}1 \text{ т},$$

на II квартал –

$$\tilde{y}_t = 77{,}62 - 1{,}54 \cdot 18 = 57{,}6 \text{ т}.$$ [47, с. 124-145].

8.5 Вопросы для самоконтроля

1. Какие виды временных рядов вы знаете? Приведите примеры.

2. Поясните, в чем состоят характерные отличия временных рядов от пространственных выборок.

3. Как на стадии графического анализа динамики временного ряда можно определить характер сезонности (аддитивный или мультипликативный)?

4. Охарактеризуйте основные типы кривых роста, наиболее часто используемые на практике при построении трендовых моделей.

5. Какие методы проверки ряда на стационарность вы знаете?

8.6 Тесты

1. Модели временных рядов - это:

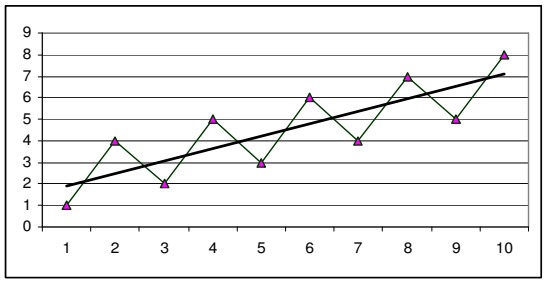
а) модели, построенные по данным, характеризующим один показатель за ряд последовательных моментов времени ;

б) модели, построенные по данным, характеризующим несколько взаимосвязанных показатель за ряд последовательных моментов времени ;

в) модели, построенные по данным, характеризующим совокупность различных объектов в определенный момент времени;

г) модели, построенные по данным, характеризующим совокупность различных объектов в определенный момент времени;

2. На рисунке изображена модель



а) мультипликативная;

б) кратная;

в) смешанная;

г) аддитивная.

3. Аддитивной моделью временного ряда называется модель:

а) в которой временной ряд представлен как сумма трендовой, циклической или случайной компонент;

б) в которой временной ряд представлен как произведение трендовой, циклической или случайной компонент;

в) в которой временной ряд представлен как отношение трендовой компоненты к циклической;

г) в которой временной ряд представлен как разность трендовой, циклической или случайной компонент.

4. Если уровни временного ряда изменяются в арифметической прогрессии, т.е. когда первые разности уровней (абсолютные приросты) более или менее постоянны, то для описания лучшим образом подойдет:

а) линейная функция;

б) парабола второго порядка;

в) гипербола;

г) степенная функция.

5. Уравнение $\tilde{y}_i = a + b \cdot t_i$ называется:

а) линейным трендом;

б) параболическим трендом;

в) гиперболическим трендом;

г) экспоненциальным трендом.

9 Динамические эконометрические модели

Что необходимо знать из главы 9:

1. Авторегрессионные модели.
2. Модели с распределенным лагом.
3. Модели адаптивных ожиданий и неполной корректировки.

9.1 Авторегрессионные процессы

Часто экономические показатели, представленные временными рядами, имеют настолько сложную структуру, что моделирование таких рядов путем построения моделей тренда, сезонности и применения традиционных подходов не приводит к удовлетворительным результатам. Во временных рядах остатков прослеживаются статистические зависимости, которые можно моделировать.

В последнее время большое внимание уделяется моделированию стационарных временных рядов, так как многие временные ряды могут быть приведены к стационарному виду после операции выделения тренда, фильтрации сезонной компоненты или взятия разности. Как правило, ряд остатков – это стационарный ряд. Наиболее распространенные модели стационарных рядов – модели авторегрессии и модели скользящего среднего [46, с. 141].

*Авторегрессионные модели***.** В авторегрессии каждое значение ряда находится в линейной зависимости от предыдущих значений. Если анализируемый динамиче-

ский процесс зависит от значений, отстоящих на p временных лагов назад, то авторегрессионный процесс порядка p, т.е. AR (p):

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} + \varepsilon_t, \quad (9.1)$$

где ε_t – «белый шум» с $\mu_\varepsilon = 0$;

α_0 – свободный член (часто приравнивается к нулю (опускается)).

Используя функцию оператора лага, можно представить авторегрессионную модель в виде:

$$\left(1 - \alpha_0 - \alpha_1 B - \alpha_2 B^2 - ... - \alpha_p B^p\right) y_t = \Phi(B) y_t = \varepsilon_t, \quad (9.2)$$

где B – оператор сдвига, т.е. преобразование ряда, смещающего его на один временной такт;

$\Phi(B)$ – оператор авторегрессии.

Для выполнения условия стационарности все корни многочлена $\Phi(B)$ должны лежать вне единичного круга, т.е. все корни характеристического уравнения $1 - \alpha_1 z - \alpha_2 z^2 - ... - \alpha_p z^p = 0$ должны быть по модулю больше 1 и различны, т.е. $|z| > 1$. Если $|z| = 1$, процесс называется *процессом единичного корня* и является нестационарным.

Рассмотрим простейший вариант линейного авторегрессионного процесса – модель авторегрессии 1-го порядка – AR(1), или марковский процесс [46, с. 142].

Эта модель может быть представлена в виде:

$$y_t = \alpha y_{t-1} + \varepsilon_t, \quad (9.3)$$

где α – числовой коэффициент, $|\alpha| < 1$;

ε_t – последовательность случайных величин, образующих «белый шум».

Основные свойства Марковского процесса:

$$\begin{aligned} &\mu y_t = 0, \\ &D(y_t) = \frac{\sigma_0^2}{1-\alpha^2}, \\ &\operatorname{cov}(y_t, y_{t\pm k}) = \alpha^k D(y_t), \\ &\rho(y_t, y_{t\pm k}) = \alpha^k. \end{aligned} \tag{9.4}$$

Значения частной автокорреляционной функции равны нулю для всех лагов $k \geq 2$, что может быть использовано при подборе модели. Этот результат для теоретической ЧАКФ и может не выполняться для выборочной АКФ. Однако если выборочные частные корреляции статистически незначимо отличаются от нуля при $k \geq 2$, то использование модели $AR(1)$ не противоречит исходным данным.

Условие стационарности ряда для $AR(1)$ определяется требованием к коэффициенту α: $|\alpha| < 1$.

Из авторегрессионных процессов выше 1-го порядка в экономической практике часто встречаются так называемые *процессы Юла.* Они описываются с помощью модели $AR(2)$:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t. \tag{9.5}$$

Выражение для вычисления любого значения АКФ $\rho(k)$:

$$\rho(k) = \alpha_1 \rho(k-1) + \alpha_2 \rho(k-2). \tag{9.6}$$

Подставим в данное выражение значение $k = 1,2$. С учетом того, что $\rho(0) = 1$, а $\rho(-1) = \rho(1)$, получим:

$$\begin{cases} \rho(1) = \alpha_1 + \alpha_2 \rho(1); \\ \rho(2) = \alpha_1 \rho(1) + \alpha_2. \end{cases} \quad (9.7)$$

Эта система называется *системой Юла - Уокера для AR(2).* Из нее можно получить выражения для определения параметров α_1 и α_2:

$$\alpha_1 = \frac{\rho(1)[1-\rho(2)]}{1-\rho^2(1)}, \quad (9.8)$$

$$\alpha_2 = \frac{\rho(2)-\rho^2(1)}{1-\rho^2(1)}. \quad (9.9)$$

Условия стационарности процесса *AR*(2):

$$\begin{array}{l} |\alpha_2| < 1; \\ \alpha_1 + \alpha_2 < 1; \\ \alpha_2 - \alpha_1 < 1. \end{array} \quad (9.10)$$

ЧАКФ для процесса *AR*(*p*) будет иметь ненулевые значения лишь при $k \le p$, а начиная с лага $k = p+1$ теоретическая ЧАКФ равна нулю. Это свойство становится ключевым при подборе порядка *p* авторегрессионной модели для конкретных экономических временных рядов [46, с. 148].

При прогнозировании на практике реальные параметры ARMA-процесса α_k и β_j заменяются своими оценками $\tilde{\alpha}_k$ и $\tilde{\beta}_j$, а случайные шоки ε_t - на остатки $\tilde{\varepsilon}_t$, полученные при оценивании модели, или на ошибки предыдущих прогнозов.

Прогнозирование значения y_t на период $(t+h)$ по авторегрессионной модели производят следующим образом.

Сначала вычисляют значения

$$\tilde{y}_{t+1} = \alpha_0 + \alpha_1 y_t + \alpha_2 y_{t-1} + ... + \alpha_p y_{t-p} + \varepsilon_t. \quad (9.11)$$

Затем в модель

$$\tilde{y}_{t+2} = \alpha_0 + \alpha_1 \tilde{y}_{t+1} + \alpha_2 y_t + ... + \alpha_p y_{t-p+1} + \varepsilon_t \qquad (9.12)$$

подставляют вычисленное значение $\tilde{y}_{t+1}$ и определяют величину $\tilde{y}_{t+2}$ и т.д.

Рассмотрим пример построения авторегрессионной модели 1-го порядка доходов бюджета Оренбургской области (таблица 9.1). Для оценивания параметров модели AR (*p*) применим *МНК* y_t на $y_{t-1}, y_{t-2}, ..., y_{t-p}$, хотя некоторые свойства тестовых статистик будут искажены ввиду присутствия среди регрессоров лагов зависимой переменной.

Таблица 9.1 - Значения доходов консолидированного бюджета Оренбургской области, млн. р.

Период	2001	2002	2003	2004	2005	2006
Январь	1119,3	865,5	968,8	1196,8	944,1	1573,0
Февраль	352,2	998,4	900,0	1091,1	1317,3	1521,5
Март	1006,9	1145,1	1402,0	1629,4	2893,2	3215,2
Апрель	1177,8	1585,6	1898,8	2620,2	2234,3	2872,5
Май	1084,4	1301	1538,8	1603,7	2393,7	3792,4
Июнь	891,4	980,3	1232,7	1692,8	1834,2	2721,7
Июль	928,2	1403,5	1650,1	2267,5	2205,4	3097,2
Август	1178,4	1455,7	1486,9	1804,6	3051,7	4229,2
Сентябрь	989,4	1163,5	1364,3	1782,8	2035,7	2119,6
Октябрь	932,2	1532,0	1974,6	1921,0	2241,3	3756,5
Ноябрь	1080,4	1299,9	1551,1	2802,3	4245,3	3416,1
Декабрь	1243,5	1549,1	1795,6	2639,6	3699,7	3478,7

Так как данный вид модели применяется только для стационарных временных рядов, необходимо проверить гипотезу о наличии тенденции либо применить гра-

фический анализ. По виду графика анализируемого временного ряда можно сделать вывод о его нестационарности (рисунок 9.1).

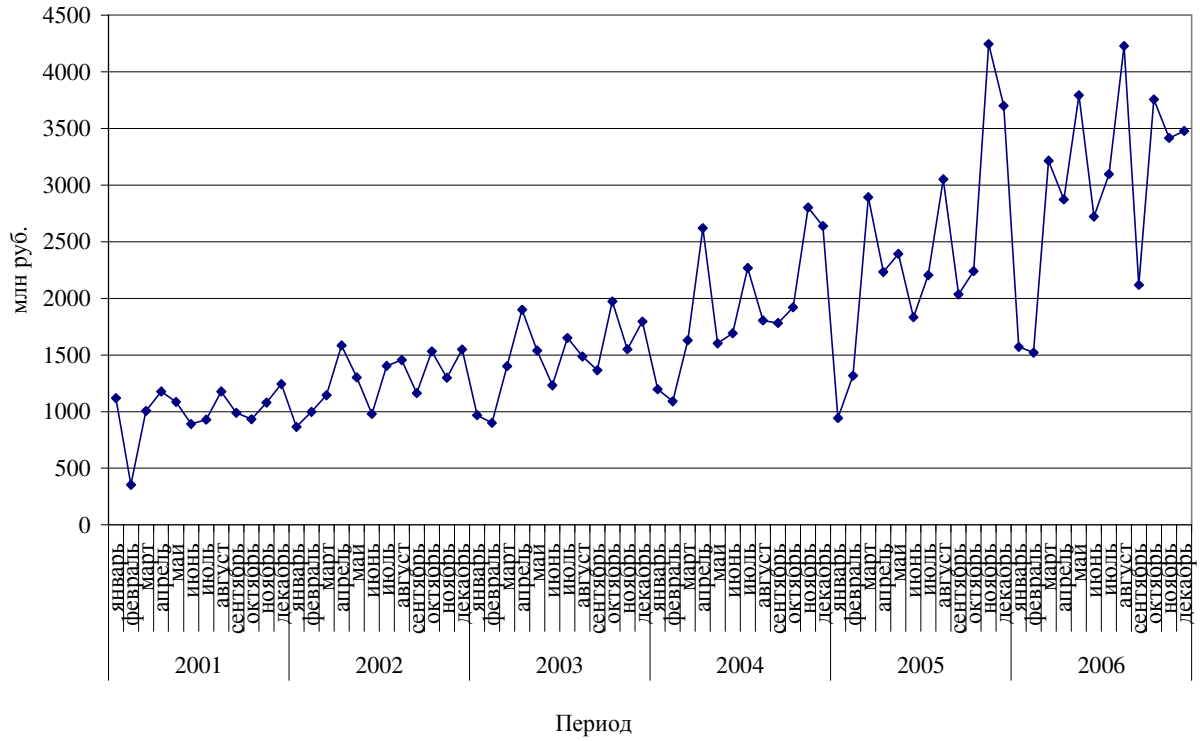


Рисунок 9.1 - Динамика доходов консолидированного бюджета Оренбургской области

Следовательно, для моделирования с помощью авторегрессионной модели ряд необходимо привести к стационарному виду. Для этого необходимо исключить тенденцию – так как она близка к линейной, найдем отклонения от прямолинейного тренда и построим авторегрессионную модель для ряда остатков. Уравнение тренда имеет вид:

$$\tilde{y}_t = 586{,}3 + 34{,}54 \cdot t \, .$$

Для выбора порядка авторегрессионной модели необходимо изучить поведение автокорреляционной и частной автокорреляционной функций (таблица 9.2).

Как видно из таблицы 9.2, АКФ и ЧАКФ экспоненциально затухает, меняя знак. Следовательно, можно предположить, что для описания временного ряда целе-

сообразно применить модель авторегрессии со скользящими средними в остатках 1-го порядка ARMA (1,1).

Таблица 9.2 – АКФ и ЧАКФ временного ряда доходов бюджета

Лаг	АКФ	ЧАКФ	Лаг	АКФ	ЧАКФ
1	-0,065	-0,065	10	-0,479	-0,400
2	-0,456	-0,462	11	0,063	-0,064
3	-0,070	-0,183	12	0,474	0,190
4	0,105	0,182	13	0,030	0,026
5	0,111	-0,022	14	-0,207	0,097
6	-0,187	-0,280	15	-0,121	0,047
7	0,014	-0,011	16	0,072	-0,079
8	0,309	0,191	17	0,105	0,055
9	-0,077	-0,002	18	-0,183	0,029

Построение модели AR (1): $\tilde{y}_t = \alpha_1 y_{t-1} + \varepsilon_t$. Параметры данной модели определим с помощью МНК в ППП STATISTICA. Результаты оценивания представлены на рисунке 9.2.

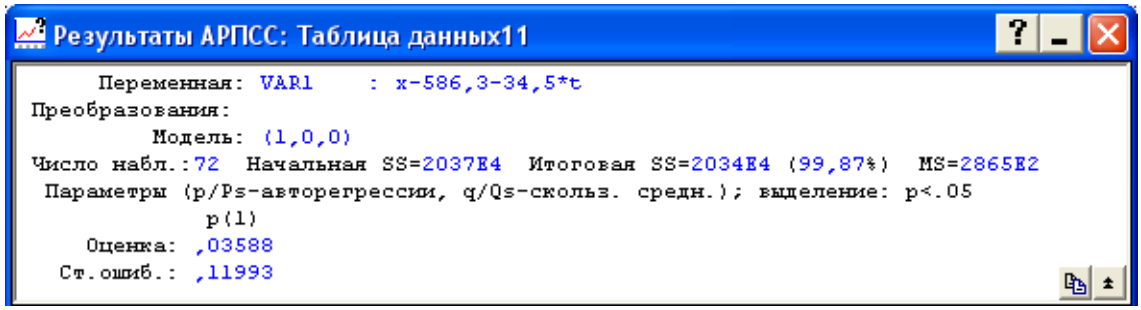


Рисунок 9.2 – Вывод итогов МНК-оценивания модели AR (1) по отклонениям от тренда

Модель AR(1) примет вид: $\varepsilon_t = 0{,}03588 \cdot \varepsilon_{t-1}$. Случайный компонент может быть выражен как $\varepsilon_t = y_t - \tilde{y}_t = y_t - 586{,}26 - 34{,}535 \cdot t$, поэтому подставив это выражение в модель *AR* (1), получим:

$$y_t - 586{,}26 - 34{,}535 \cdot t = 0{,}03588 \cdot \left(y_{t-1} - 586{,}26 - 34{,}535 \cdot (t-1)\right).$$

В результате соответствующих преобразований получим следующую модель доходов бюджета:

$$\tilde{y}_t = 0{,}03588 \cdot y_{t-1} + 33{,}296 \cdot t + 516{,}464 .$$

Теоретические значения y_t на период *(t+l)* по авторегрессионной модели находят следующим образом.

Сначала вычисляют значение $\dot{y}_{t+1}$ по формуле

$$\dot{y}_{t+1} = \alpha_1 y_t + \alpha_2 y_{t-1} + ... + \alpha_p y_{t-p} .$$

Затем в модель $\dot{y}_{t+2} = \alpha_1 \dot{y}_{t+1} + \alpha_2 y_t + ... + \alpha_p y_{t-p+1}$ подставляют вычисленное значение $\dot{y}_{t+1}$ и определяют величину $\dot{y}_{t+2}$ и т.д.

Рассчитаем доходы бюджета на январь и февраль 2007 г.

Точечные теоретические значения составят:

- на январь:

$\dot{y}_{t+1} = 0{,}03588 \cdot 2986{,}39 + 33{,}296 \cdot 73 + 516{,}464 = 3054{,}22$ млн. р.,

- на февраль:

$\dot{y}_{t+2} = 0{,}03588 \cdot 3054{,}22 + 33{,}296 \cdot 74 + 516{,}464 = 3089{,}95$ млн. р.

Доверительный интервал для теоретических значений определяется по формуле:

$$\tilde{y}_{t+l} \pm S_y \cdot t(\alpha; k).$$

Среднее квадратическое отклонение равно:

$$S_y = \sqrt{\frac{\sum (y_t - \tilde{y}_t)^2}{n}} = \sqrt{\frac{20556319{,}18}{72}} = 534{,}326 .$$

Табличное значение *t*-критерия Стьюдента: *t*(0,05; 69)= 1,995.

Тогда доверительные границы на январь составят: $3054{,}22 \pm 1065{,}95$, т.е. с вероятностью 95 % в январе 2007 г. доходы бюджета могли составить от 1988,27 до 4120,18 млн. р. В феврале 2007 г. доходы бюджета с заданной вероятностью могли составить от 2024 млн. р. до 4155,9 млн. р.

*Модели скользящего среднего***.** Модель скользящего среднего предполагает, что в ошибках модели в предшествующие периоды сосредоточена информация обо всей предыстории ряда. В этой модели каждое новое значение – среднее между текущей флуктуацией и несколькими (в частности, одной) предыдущими ошибками.

Процесс скользящего среднего порядка q, обозначаемого MA(q), имеет вид:

$$y_t = \varepsilon_t - \beta_1 \varepsilon_{t-1} - \beta_2 \varepsilon_{t-2} - ... - \beta_q \varepsilon_{t-q}, \qquad (9.13)$$

где ε_t - «белый шум» (импульс, шок) с $\mu = 0$.

В моделях скользящего среднего для обеспечения стационарности ряда не требуется накладывать никаких ограничений на параметры $\beta_1, \beta_2, ..., \beta_q$. Однако если в модели MA(1) параметр $|\beta| \geq 1$, то текущее значение y_t будет зависеть от своих прошлых значений, берущихся с весами, бесконечно растущими по мере удаления в прошлое:

$$\begin{aligned} &\varepsilon_t = y_t + \beta\varepsilon_{t-1} = y_t + \beta(y_{t-1} + \beta\varepsilon_{t-2}) = y_t + \beta y_{t-1} + \beta^2(y_{t-2} + \beta\varepsilon_{t-3}) = ... \\ &= y_t + \beta y_{t-1} + \beta^2 y_{t-2} + \beta^3 y_{t-3} + ... \qquad \Rightarrow y_t = \varepsilon_t - \sum_{k=1}^{\infty} \beta^k y_{t-k}. \end{aligned} \qquad (9.14)$$

Чтобы избежать этого, надо, чтобы веса образовывали сходящийся ряд, т.е. $|\beta| < 1$.

Широко распространены в статистической практике модели скользящего среднего 1-го и 2-го порядков:

$$MA(1):\ y_t = \varepsilon_t - \beta\varepsilon_{t-1}, \tag{9.15}$$

$$MA(2):\ y_t = \varepsilon_t - \beta_1\varepsilon_{t-1} - \beta_2\varepsilon_{t-2}. \tag{9.16}$$

Для модели *MA (1)* формулы для прогнозирования имеют вид:

$$\begin{aligned} &\tilde{Y}_{t+1} = \beta \cdot \varepsilon_t, \\ &\tilde{Y}_{t+h} = 0 \text{ для } h \geq 2. \end{aligned} \tag{9.17}$$

Для процесса *MA (2)* формулы для прогнозирования:

$$\begin{aligned} &\tilde{Y}_{t+1} = -\beta_1 \cdot \varepsilon_t - \beta_2\varepsilon_{t-1}, \\ &\tilde{Y}_{t+2} = -\beta_2\varepsilon_t, \\ &\tilde{Y}_{t+h} = 0 \text{ для } \ h \geq 3. \end{aligned} \tag{9.18}$$

Пример построения модели MA (q). В случае чистого *MA (q)*- процесса обычно используются нелинейные методы наименьших квадратов для оценивания параметров. Наиболее распространен метод условной суммы квадратов (*CSS*), в котором отсутствующие данные значений «белого шума» ε_t генерируются как ex-post ошибки прогноза при условии минимума суммы квадратов ошибок.

Так, для процесса *MA (1)*: $y_t = \varepsilon_t - \beta \cdot \varepsilon_{t-1}$ генерируются значения $\tilde{\varepsilon}_t = y_t + \tilde{\beta} \cdot \tilde{\varepsilon}_{t-1}$ с некоторым коэффициентом $\tilde{\beta}$, который оценивается при условии $S(\tilde{\beta}) = \sum_1^T \tilde{\varepsilon}_t^2 = \min$. Поскольку это нелинейная функция относительно параметра β, минимизация происходит в результате итеративного процесса.

Оценим параметры модели *MA (1)*: $y_t = \varepsilon_t - \beta \cdot \varepsilon_{t-1}$ по данным таблицы 9.1 в ППП Statistica.

Так как ряд нестационарный, рассмотрим два способа приведения к стационарному виду.

Первый способ – *взятие первых разностей.* Результаты оценивания параметров представлены на рисунке 9.3. Модель примет вид:

$$\Delta^1_{y_t} = \varepsilon_t + 0{,}79188 \cdot \varepsilon_{t-1}.$$

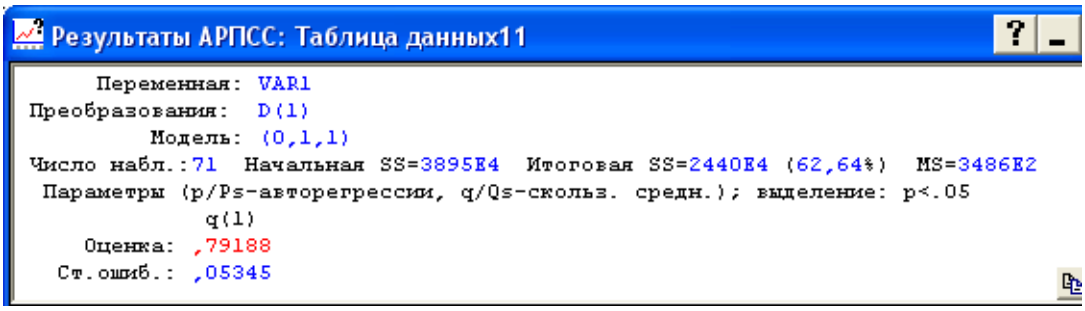


Рисунок 9.3 - Результаты оценивания *MA(1)* по первым разностям

От этой модели можно перейти к модели для уровней ряда:

$$y_t = y_{t-1} + \varepsilon_t + 0{,}79188 \cdot \varepsilon_{t-1}.$$

Второй способ - *отклонение от линейного тренда.* Результаты оценивания параметров представлены на рисунке 9.4.

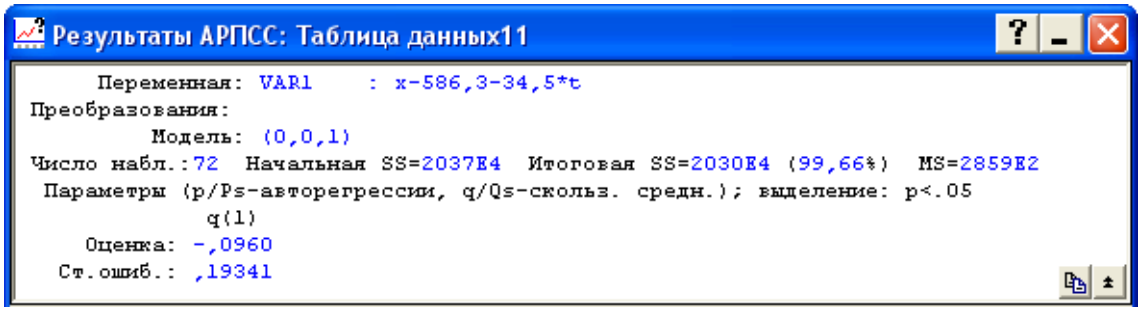


Рисунок 9.4 - Результаты оценивания *MA(1)* по отклонениям от тренда

Полученная модель имеет вид:

$$y_t - \tilde{y}_t = \varepsilon_t - 0{,}0960\varepsilon_{t-1}.$$

Откуда

$$y_t = \tilde{y}_t + \varepsilon_t - 0{,}0960 \cdot \varepsilon_{t-1} = 586{,}3 + 34{,}5 \cdot t + \varepsilon_t - 0{,}0960 \cdot \varepsilon_{t-1},$$

где $\tilde{y}_t$ - расчетные значения по линейному тренду.

При прогнозировании на практике реальные параметры β_j заменяются своими оценками $\tilde{\beta}_j$, а случайные шоки ε_t - на остатки $\tilde{\varepsilon}_t$, полученные при оценивании модели, или на ошибки предыдущих прогнозов.

Для модели *MA (1)* формулы для прогнозирования имеют вид:

$$\tilde{Y}_{t+1} = \beta \cdot \varepsilon_t,$$
$$\tilde{Y}_{t+h} = 0 \text{ для } h \geq 2.$$

В ППП Ststistica прогнозные значения по полученным моделям можно получить в табличной форме и графической соответственно рисунки 9.5 и 9.6.

Прогнозы; Модель:(0,0,1) Сезонный лаг: 12 (Таблица данных11)
Исход.:VAR1 : x-586,3-34,5*t
Начало исходных: 1 Конец исходн.: 72

Набл. N	Прогноз	Нижний 90,0000%	Верхний 90,0000%	Ст.ошиб.
73	36,23661	-854,879	927,3527	534,6911
74	0,00000	-895,213	895,2134	537,1496
75	0,00000	-895,213	895,2134	537,1496
76	0,00000	-895,213	895,2134	537,1496
77	0,00000	-895,213	895,2134	537,1496
78	0,00000	-895,213	895,2134	537,1496
79	0,00000	-895,213	895,2134	537,1496
80	0,00000	-895,213	895,2134	537,1496

Рисунок 9.5 **-** Прогноз по MA(1) для первых разностей

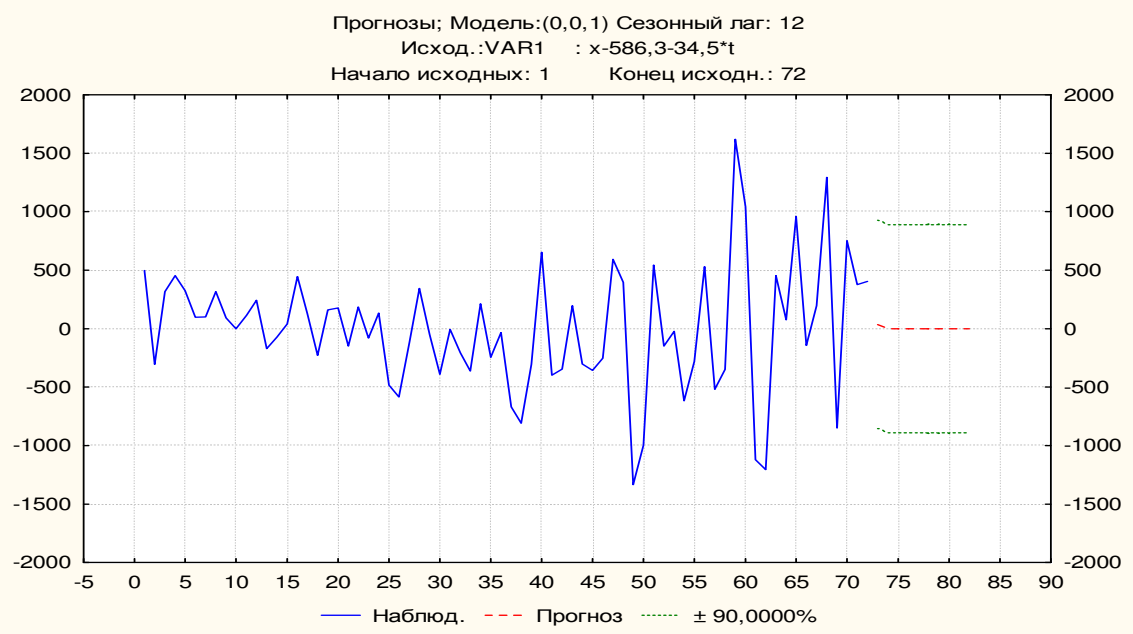


Рисунок 9.6 **-** Фактические, прогнозные значения и доверительные границы прогноза доходов бюджета Оренбургской области

Авторегрессионные модели со скользящими средними в остатках. На практике в целях экономичного описания анализируемого процесса в модель могут быть включены как члены, описывающие авторегрессионные составляющие, так и члены, моделирующие остаток в виде процесса скользящих средних. Такой процесс называется *процессом авторегрессии скользящего среднего – ARMA (p,q)*:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - ... - \beta_q \varepsilon_{t-q} \qquad (9.19)$$

или

$$y_t - \alpha_0 - \alpha_1 y_{t-1} - \alpha_2 y_{t-2} + ... + \alpha_p y_{t-p} = \varepsilon_t - \beta_1 \varepsilon_{t-1} - ... - \beta_q \varepsilon_{t-q}\,. \qquad (9.20)$$

Здесь единственное слагаемое ошибки ε_t AR-процесса заменяется на процесс MA (q).

Такая модель может интерпретироваться как линейная модель множественной регрессии, в которой в качестве объясняющих переменных выступают прошлые

значения самой зависимой переменной, а в качестве регрессионного остатка – скользящие средние из элементов «белого шума».

Стационарность процесса ARMA обеспечивается условием $|\alpha|<1$, а обратимость, в свою очередь, гарантируется выполнением условия $|\beta|<1$.

Одним из наиболее важных этапов построения моделей стационарных временных рядов является определение ее порядка. Предварительная оценка производится на основе экономического анализа. Чрезмерное повышение порядка модели может и не повысить ее точность. Одновременно расчет большего числа коэффициентов модели при неизменной выборке снижает достоверность оценки каждого из коэффициентов. В то же время недостаточное число коэффициентов модели не позволяет отразить в должной мере динамику процесса и оценить его дальнейшие изменения.

Для определения порядка процесса модели исследуются такие характеристики, как автокорреляционная функция и частная автокорреляционная функция.

На практике, как правило, используют следующие виды моделей, идентифицировать которые можно с помощью анализа АКФ и ЧАКФ (таблица 9.3) [46, с. 206].

ARMA-процессы имеют более сложную структуру по сравнению со схожими по поведению AR- или MA- процессами в чистом виде, но при этом ARMA- процессы характеризуются меньшим количеством параметров, что является одним из их преимуществ.

Для модели *ARMA (1,1)* формулы для прогнозирования имеют вид:

$$\begin{aligned} &\tilde{Y}_{t+1} = \alpha_0 + \alpha_1 y_t - \beta \cdot \varepsilon_t, \\ &\tilde{Y}_{t+h} = \alpha_0 + \alpha_1 y_{t+h-1} \quad \text{для } h \geq 2. \end{aligned} \tag{9.21}$$

Доверительный интервал прогноза в предположении, что ε_t имеет характеристики «белого шума», вычисляется по формуле:

Таблица 9.3 - Свойства АКФ и ЧАКФ

Функция	*ARMA (1,0)*	*ARMA (2,0)*	*ARMA (0,1)*	*ARMA (0,2)*	*ARMA (1,1)*
АКФ	Экспоненциально затухает (монотонно или знакопеременно)	Экспоненциально затухает или имеет форму синусоидальной волны	Выброс (пик) на лаге 1	Выбросы (пики) на лагах 1,2	Экспоненциально затухает от значения $\rho(1)$ (монотонно или знакопеременно)
ЧАКФ	Выброс (пик) на лаге 1	Выбросы (пики) на лагах 1,2	Экспоненциально затухает (монотонно или знакопеременно)	Экспоненциально затухает или имеет форму синусоидальной волны	Экспоненциально затухает от значения $\rho_{ч}(1)$ (монотонно или знакопеременно)

$$\tilde{y}_t - t_\alpha \tilde{\sigma}_{\varepsilon_t} \le y_t \le \tilde{y}_t + t_\alpha \tilde{\sigma}_{\varepsilon_t}, \qquad (9.22)$$

где y_t – истинное значение исследуемого параметра;

$\tilde{y}_t$ – предсказываемое значение исследуемого параметра;

$\tilde{\sigma}_{\varepsilon_t} = \dfrac{\sum_{t=1}^{n} \varepsilon_t^2}{n-p}$ – оценка дисперсии случайной величины ε_t;

ε_t – остатки в уравнении авторегрессии;

n – число наблюдений;

p – порядок авторегрессии;

t_α – табличное значение t-критерия Стьюдента.

Пример построения модели ARMA (p,q). Если модель ARMA содержит скользящие средние, то МНК – оценивание, как и в случае с MA-процессами, уже не является возможным. В связи с этим оценивание параметров моделей ARMA в основном проводится по тем же принципам, что и оценивание параметров для MA-процессов, но становится намного сложнее. Например, появляется проблема выбора первоначальных значений y_t из-за наличия регрессоров – лагов зависимой переменной. Наиболее распространенными методами оценивания параметров являются: нелинейный МНК и метод максимального правдоподобия.

Оценим параметры модели ARMA по отклонениям от линейного тренда, используя ППП Statistica. Результаты оценивания представлены на рисунке 9.7.

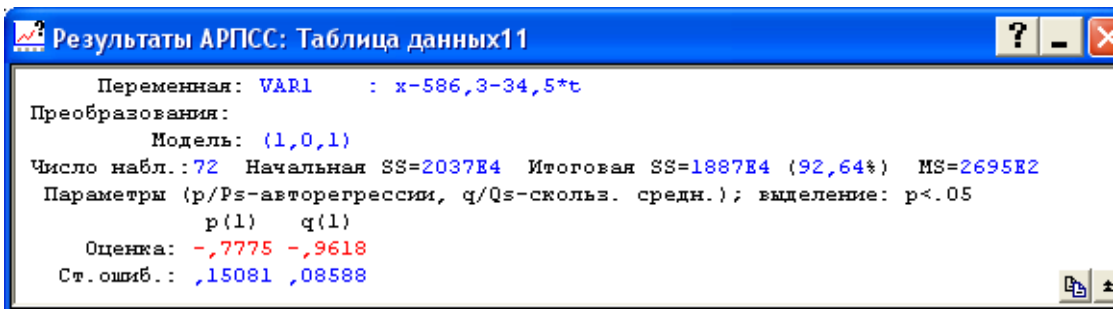


Рисунок 9.7 - Результаты оценивания модели ARMA

Модель примет вид:

$$y_t - \tilde{y}_t = -0{,}7775 \cdot (y_{t-1} - \tilde{y}_{t-1}) + \varepsilon_t - 0{,}9618\varepsilon_{t-1}.$$

Подставим в полученную модель вместо $\tilde{y}$ уравнение тренда и после раскрытия скобок и приведения подобных слагаемых перейдем к модели вида:

$$y_t = 1046{,}15 - 61{,}324 \cdot t - 0{,}7775 \cdot y_{t-1} + \varepsilon_t - 0{,}9618 \cdot \varepsilon_{t-1}.$$

Прогнозные значения отклонений от тренда представлены на рисунке 9.8, их графическое изображение – на рисунке 9.9.

Набл. N	Прогнозы; Модель:(1,0,1) Сезонный лаг: 12 (Таблица данных11) Исход.:VAR1 : x-586,3-34,5*t Начало исходных: 1 Конец исходн.: 72 Прогноз	Нижний 90,0000%	Верхний 90,0000%	Ст.ошиб.
73	218,670	-646,76	1084,102	519,1820
74	-170,015	-1050,02	709,995	527,9275
75	132,186	-756,52	1020,892	533,1445
76	-102,774	-996,70	791,149	536,2736
77	79,906	-817,15	976,967	538,1563
78	-62,127	-961,08	836,826	539,2913
79	48,303	-851,79	948,397	539,9761
80	-37,555	-938,34	863,228	540,3897
81	29,199	-872,00	930,399	540,6396
82	-22,702	-924,15	878,750	540,7906

Рисунок 9.8 **-** Прогноз по модели *ARMA(1,1)*

Модели ARIMA. Экономические временные ряды за редким исключением нестационарны. Нестационарность чаще всего проявляется в наличии зависящей от времени неслучайной составляющей *f(t)*. Для описания таких рядов используется модель авторегрессии – проинтегрированного скользящего среднего *ARIMA* (*p,d,q*) (модель Бокса – Дженкинса).

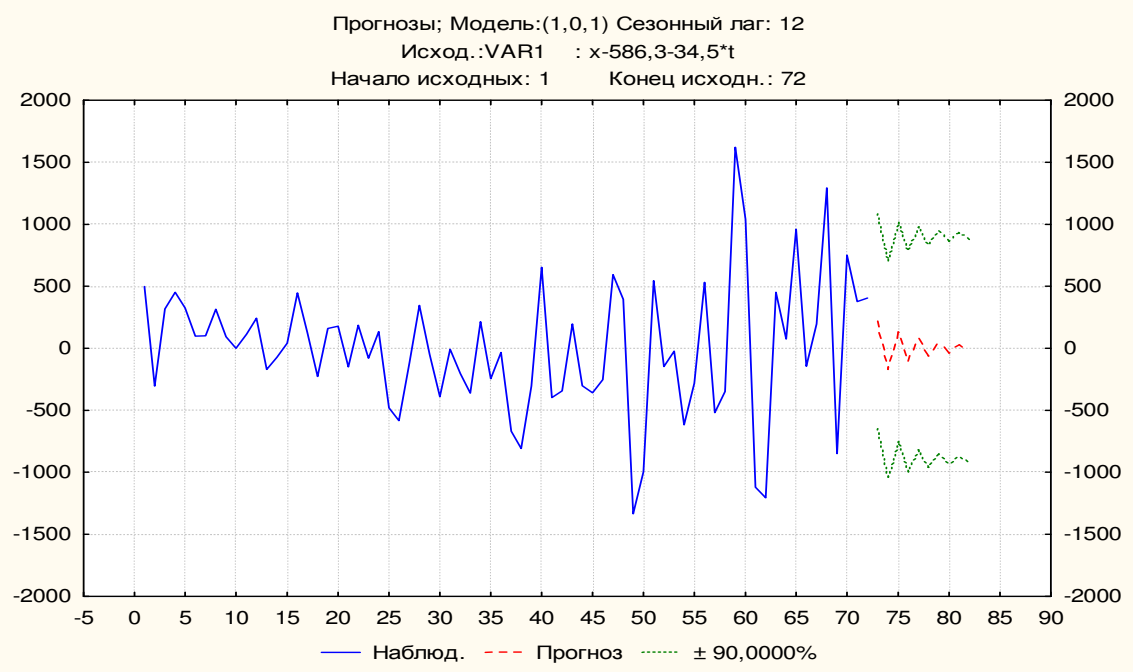


Рисунок 9.9 **-** Графическое изображение прогноза по модели *ARMA(1,1)*

Модель ARIMA используется для описания временных рядов, обладающих свойствами:

1) ряд включает аддитивно составляющую *f(t),* имеющую вид алгебраического полинома;

2) ряд, получившийся после применения к нему процедур последовательных разностей, может быть описан моделью *ARMA* (*p,q*).

Пусть X_t – нестационарный процесс со стационарными разностями *d*-го порядка, т.е. $Y_t = \Delta^d X_t$ – стационарный процесс, а $\Delta^{d-1} X_t$ – нестационарный. Это означает, что X_t интегрируем *d*-го порядка.

Если Y_t – процесс *ARMA* (*p,q*), т.е.

$$Y_t = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{t-2} + ... + \alpha_p Y_{t-p} + \varepsilon_t - \beta_1 \varepsilon_{t-1} - ... - \beta_q \varepsilon_{t-q}, \qquad (9.23)$$

тогда X_t называется процессом ARIMA (*p,d,q*). Часто среднее или свободный член приравниваются к нулю (опускаются) [48, с. 388].

Построение модели ARIMA по реализации случайного процесса Бокс и Дженкинс предложили разбить на несколько этапов:

1. Устанавливается порядок интеграции *d*, т.е. нужно добиться стационарности ряда, взяв достаточное количество последовательных разностей. Для определения значения *d* может быть применен эвристический критерий. Использование данного критерия основано на определении оценки

$$\tilde{\sigma}^2(k) = \frac{\frac{1}{n-k}\sum_{t=1}^{n-k}\left(\Delta^k y_t\right)^2}{C_{2k}^k}, \qquad (9.24)$$

где $\Delta^k y_t$ – последовательные разности исходного ряда $y_1, y_2, ..., y_n$;

k – порядок разностей, $k = 1,2,\ldots$

Начиная с некоторого значения $k = k_0$ величина $\tilde{\sigma}^2(k)$ стабилизируется, оставаясь примерно на одном и том же уровне при росте *k*. Тогда порядок разности (*d*) следует принять равным k_0.

О том, что необходимая для стационарности ряда степень разности достигнута, будет свидетельствовать быстрое затухание АКФ.

2. Для полученного стационарного временного ряда строятся АКФ и ЧАКФ. Исследуя характер их поведения, выдвигаются гипотезы о значениях параметров *p* и *q*, т.е. подбирается модель *ARMA*(*p,q*). На данном этапе формируется базовый набор моделей, включающий 1, 2 или даже большее количество моделей.

3. Для всех моделей, отобранных на 2 этапе, оцениваются коэффициенты $\alpha_0, \alpha_1, ..., \alpha_p, \beta_1, \beta_2, ..., \beta_q$, используя следующие методы:

- традиционный метод наименьших квадратов (МНК);
- метод максимального правдоподобия;
- нелинейный МНК;
- алгоритм Марквардта.

Все эти оценки при больших объемах выборок асимптотически эквивалентны.

4. Выбирается наиболее подходящая модель среди оцененных:

а) проверяется адекватность модели на основе анализа остатков (у адекватной модели остатки должны быть похожи на «белый шум»). Для этого проводится проверка значимости коэффициентов автокорреляции, используя следующие подходы:

- если выборочный коэффициент автокорреляции r_k выходит за интервал $\pm t_\alpha / \sqrt{n}$, то гипотеза H_0 о равенстве нулю коэффициента автокорреляции ρ_k отвергается;
- проверяется равенство нулю сразу τ первых значений АКФ на основе Q – статистики Бокса – Пирса:

$$Q = n\sum_{k=1}^{\tau} r_k^2 \tag{9.25}$$

или на основе теста Бокса – Льюнга:

$$\tilde{Q} = n(n+2)\sum_{k=1}^{\tau} \frac{r_k^2}{n-k}. \tag{9.26}$$

Если Q $(\text{или } \tilde{Q}) > \chi^2_{табл}$ с $\nu = \tau - p - q$ степенями свободы, то как группа первые τ коэффициентов автокорреляции значимы (рекомендуется рассматривать $\tau \approx n/4$);

б) Отбирается оптимальная модель по наивысшему качеству с меньшим числом параметров с использованием информационного критерия Акайка и Шварца:

- информационный критерий Акайка:

$$AIC = \frac{p+q}{n} + \ln\left(\frac{\sum_{t=1}^{n} e_t^2}{n}\right); \tag{9.27}$$

- критерий Шварца:

$$SIK = \frac{(p+q)\ln n}{n} + \ln\left(\frac{\sum_{t=1}^{n} e_t^2}{n}\right). \qquad (9.28)$$

Предпочтение следует отдать модели с меньшим значением критерия.

Прогнозирование ARIMA-процессов Y_t может быть представлено в виде двухшаговой процедуры [48, с. 416– 421]:

1) экстраполируется стационарный ARMA-процесс;

2) вместо взятия разностей проводится обратная операция интегрируемости, т.е. суммирование спрогнозированных на шаге 1 приращений $\tilde{Y}_{T(+h)} = \Delta^d \tilde{X}_{T(+h)}$, чтобы получить сначала $\Delta^{d-1}\tilde{X}_{T(+h)}$, а затем по аналогии $\Delta^{d-2}\tilde{X}_{T(+h)}$ и, наконец, $\tilde{X}_{T(+h)}$. Оценка дисперсии ошибки прогноза, а следовательно, и ширины доверительного интервала прогноза проводится аналогичным образом – повторным суммированием дисперсий ошибок прогноза ARMA-процесса X_t.

Другим возможным вариантом получения прогноза является построение индивидуальной одношаговой формулы.

С этой целью в уравнение вместо Y_t подставляют разности

$$\Delta^d X_T = (1-L)^d X_t. \qquad (9.29)$$

Решив полученное уравнение относительно X_t, получим формулу, которая может быть экстраполирована для $t = T + h$ и таким образом преобразована в формулу для прогнозирования на h шагов вперед величин $\tilde{X}_{T(+h)}$ с началом отсчета в момент времени *T*.

Пример построения модели ARIMA (p,d,q). Оценим параметры модели в ППП STATISTICA. Результаты оценивания представлены на рисунке 9.10.

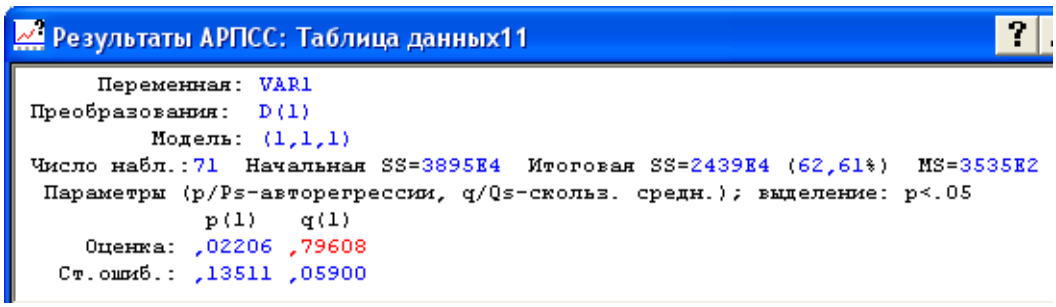


Рисунок 9.10 - Результаты оценивания модели ARIMA

Таким образом, мы получили модель:

$$\Delta_t^1 = 0{,}02206 \cdot \Delta_{t-1}^1 + \varepsilon_t - 0{,}79608 \cdot \varepsilon_{t-1}.$$

Подставив в модель вместо Δ_t^1 – $(y_t - y_{t-1})$ и вместо Δ_{t-1}^1 – $(y_{t-1} - y_{t-2})$, раскрыв скобки и приведя подобные слагаемые, получим модель:

$$\tilde{y}_t = 1{,}02206\, y_{t-1} - 0{,}02206\, y_{t-2} + \varepsilon_t - 0{,}79608 \varepsilon_{t-1}.$$

Теоретическое значение для модели *ARIMA (1,1,1)* на один шаг вперед определяется по формуле

$$\tilde{y}_{t+1} = (1+\alpha) \cdot y_t - \alpha \cdot y_{t-1} - \beta \varepsilon_t,$$

т.е. на январь 2007 г. расчетное значение составит:

$(1+0{,}02206) \cdot 3478{,}7 - 0{,}02206 \cdot 3416{,}1 - 0{,}79608 \cdot (52{,}0) = 3438{,}7$ млн. р.

Формула прогноза на два шага вперед:

$$\tilde{y}_{t+2} = (1+\alpha) \cdot y_{t+1} - \alpha \cdot y_t.$$

Тогда на февраль 2007 г. расчетное значение составит:

$(1+0{,}02206) \cdot 3438{,}7 - 0{,}02206 \cdot 3478{,}7 = 3437{,}8$ млн. р.

Интервальный прогноз при среднем квадратическом отклонении 763,95 и статистике Стьюдента на 5%-ном уровне значимости для 68 степеней свободы, составившей 1,995 для января 2007 г. будет находиться в границах от 1914,7 млн. р. до 4202,7 млн. р., а в феврале 2007 г. доходы бюджета с вероятностью 95 % могли составить от 1913,8 млн. р. до 4961,8 млн. р.

9.2 Модели с распределенным лагом

Модели с распределенными лагами бывают двух типов:

- с конечным числом лагов:

$$y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + ... + b_k \cdot x_{t-k} + \varepsilon_t; \quad (9.30)$$

- с бесконечным числом лагов:

$$y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + b_2 \cdot x_{t-2} + ... + \varepsilon_t. \quad (9.31)$$

Наиболее распространены на практике модели с конечным числом лагов. Для оценки параметров таких моделей, при правильной ее спецификации, может быть применен обычный МНК. В этом случае переменные $x_t, x_{t-1}, ..., x_{t-k}$ рассматриваются как объясняющие переменные множественной регрессии.

Вместе с тем применение обычного МНК к таким моделям в большинстве случаев затруднительно по следующим причинам:

1) при наличии тенденции объясняющие переменные тесно связаны между собой, что вызывает мультиколлинеарность;

2) если в анализируемые временные ряды нестационарные, то возможна автокорреляция остатков.

Вышеуказанные обстоятельства приводят к значительной неопределенности относительно оценок параметров модели, снижению их точности и получению не-

эффективных оценок. Чистое влияние факторов на результат в таких условиях выявить невозможно. Поэтому на практике параметры моделей с распределенным лагом учитывают определенные ограничения на коэффициенты регрессии и условия выбранной структуры лага.

Лаги, структуру которых можно описать с помощью полиномов, называют также лагами Алмон, по имени Ш. Алмон, впервые обратившей внимание на такое представление лагов.

полиномиальная структура лага, т. е. зависимость коэффициентов регрессии b_i от величины лага описывается полиномом *k*-й степени. Частным случаем полиномиальной структуры лага является линейная модель

Формально модель зависимости коэффициентов b_j от величины лага у в форме полинома можно записать так:

- для полинома первой степени $b_j = c_0 + c_1 j$;
- для полинома второй степени $b_j = c_0 + c_1 j + c_2 j^2$;
- для полинома третьей степени $b_j = c_0 + c_1 j + c_2 j^2 + c_3 j^3$ и т. д.

В наиболее общем виде для полинома *k*-й степени имеем:

$$b_j = c_0 + c_1 j + c_2 j^2 + ... + c_k j^k .$$

Тогда каждый из коэффициентов b_j модели (9.30) можно выразить следующим образом:

$$\begin{aligned}
b_0 &= c_0, \\
b_1 &= c_0 + c_1 + c_2 + ... + c_k, \\
b_2 &= c_0 + 2c_1 + 4c_2 + ... + 2^k c_k, \\
b_3 &= c_0 + 3c_1 + 9c_2 + ... + 3^k c_k, \\
&\dots \\
b_l &= c_0 + lc_1 + l^2 c_2 + ... + l^k c_k .
\end{aligned} \qquad (9.32)$$

Подставив в (9.30) найденные соотношения для b_j (9.32) получим:

$$y_t = a + c_0 x_t + (c_0 + c_1 + ... + c_k) \cdot x_{t-1} + (c_0 + 2c_1 + 4c_2 + ... + 2^k c_k) \cdot x_{t-2} + ... +$$
$$+ (c_0 + 3c_1 + 9c_2 + ... + 3^k c_k) \cdot x_{t-3} + (c_0 + l \cdot c_1 + l^2 \cdot c_2 + ... + l^k \cdot c_k) \cdot x_{t-l} + \varepsilon_t. \quad (9.33)$$

Перегруппируем слагаемые:

$$y_t = a + c_0 (x_t + x_{t-1} + ... + x_{t-l}) + c_1 (x_{t-1} + 2x_{t-2} + 3x_{t-3} + ... + l \cdot x_{t-l}) + c_2 \times$$
$$\times (x_{t-1} + 4x_{t-2} + 9x_{t-3} + ... + l^2 \cdot x_{t-1}) + ... + c_k \cdot (x_{t-1} + 2^k x_{t-2} + 3^k x_{t-3} + ... + l^k x_{t-l}) + \varepsilon_t. \quad (9.34)$$

Обозначим слагаемые в скобках при c_i как новые переменные:

$$z_0 = x_t + x_{t-1} + x_{t-2} + ... + x_{t-l} = \sum_{j=0}^{l} x_{t-j} \ ,$$

$$z_1 = x_{t-1} + 2x_{t-2} + 3x_{t-3} + ... + l \cdot x_{t-l} = \sum_{j=1}^{l} j \cdot x_{t-j},$$

$$z_2 = x_{t-1} + 4x_{t-2} + 9x_{t-3} + ... + l^2 \cdot x_{t-l} = \sum_{j=1}^{l} j^2 \cdot x_{t-j}, \quad (9.35)$$

………………………………………………..

$$z_k = x_{t-1} + 2^k x_{t-2} + 3^k x_{t-3} + ... + l^k \cdot x_{t-l} = \sum_{j=1}^{l} j^k \cdot x_{t-j}.$$

Перепишем модель (9.34) с учетом соотношений (9.35):

$$y_t = a + c_0 z_0 + c_1 z_1 + c_2 z_2 + ... + c_k z_k + \varepsilon_t. \quad (9.36)$$

Процедура применения метода Алмон для расчета параметров модели с распределенным лагом выглядит следующим образом:

1) определяется максимальная величина лага *l*;

2) определяется степень полинома к, описывающего структуру лага;

3) по соотношениям (9.35) рассчитываются значения переменных $z_0,..., z_k$;

4) определяются параметры уравнения линейной регрессии (9.36);

5) с помощью соотношений (9.32) рассчитываются параметры исходной модели с распределенным лагом.

Применение метода Алмон сопряжено с рядом проблем.

Во-первых, величина лага *l* должна быть известна заранее. При ее определении лучше исходить из максимально возможного лага, чем ограничиваться лагами небольшой длины. Выбор меньшей величины лага по сравнению с его реальным значением приведет к тому, что в модели регрессии не будет учтен фактор, оказывающий значительное влияние на результат, т. е. к неверной спецификации модели. Влияние этого фактора в такой модели будет выражено в остатках. Тем самым в модели не будут соблюдаться предпосылки МНК о случайности остатков, а полученные оценки ее параметров окажутся неэффективными и смещенными. Выбор большей величины лага по сравнению с ее реальным значением будет означать включение в модель статистически незначимого фактора и снижение эффективности полученных оценок, однако эти оценки все же будут несмещенными.

Известно несколько практических подходов к определению реальной величины лага, например построение нескольких уравнений регрессии и выбор наилучшего из этих уравнений или применение формальных критериев, например критерия Шварца. Однако наиболее простым способом является измерение тесноты связи между результатом и лаговыми значениями фактора. Кроме того, оптимальную величину лага можно приближенно определить на основе априорной информации экономической теории или проведенных ранее эмпирических исследований.

Во-вторых, необходимо установить степень полинома к. Обычно на практике ограничиваются рассмотрением полиномов второй и третьей степени, применяя следующее простое правило: выбранная степень полинома к должна быть на единицу больше числа экстремумов в структуре лага. Если априорную информацию о структуре лага получить невозможно, величину к проще всего определить путем сравнения моделей, построенных для различных значений к, и выбора наилучшей модели.

В-третьих, переменные z, которые рассчитываются как линейные комбинации исходных переменных x, будут коррелировать между собой в случаях, когда наблюдается высокая связь между самими исходными переменными. Поэтому оценку параметров модели (9.36) приходится проводить в условиях мультиколлинеарности факторов. Однако мультиколлинеарность факторов Z_0,.., Z_k в модели (9.36) сказывается на оценках параметров $b_0,...,b_l$ в несколько меньшей степени, чем если бы эти оценки были получены путем применения обычного МНК непосредственно к исходной модели в условиях мультиколлинеарности факторов $x,..., x_{t-l}$. Это связано с тем, что в модели (9.36) мультиколлинеарность ведет к снижению эффективности оценок $c_0,\ldots, c_k$ поэтому каждый из параметров $b_0,...,b_l$ которые определяются как линейные комбинации оценок $c_0,..., c_k$ будет представлять собой более точную оценку, а стандартные ошибки этих параметров не будут превышать стандартные ошибки параметров, полученных по исходной модели обычным МНК.

Метод Алмон имеет два неоспоримых преимущества:

- он достаточно универсален и может быть применен для моделирования процессов, которые характеризуются разнообразными структурами лагов;
- при относительно небольшом количестве переменных в (9.36) (обычно выбирают к = 2 или к = 3), которое не приводит к потере значительного числа степеней свободы, с помощью метода Алмон можно построить модели с распределенным лагом любой длины.

Рассмотрим пример оценки параметров модели с распределенным лагом по данным об индексах выпуска по базовым видам экономической деятельности, с поправкой на сезонность (Y) и инвестиций в основной капитал (X). Максимальную величину лага примем равной трем. Спецификация модели с распределёнными лагами для данной задачи имеет вид:

$$y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + b_2 \cdot x_{t-2} + b_3 \cdot x_{t-3} + \varepsilon_t .$$

Исходные и расчетные данные для оценки параметров представлены в таблице Л.1 приложения Л. Результаты оценивания обычным МНК представлены на рисунке 9.11.

Регрессионная статистика	
Множественный R	0,709
R-квадрат	0,503
Нормированный R-квадрат	0,478
Стандартная ошибка	7,851
Наблюдения	86

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	4	5049,440	1262,360	20,478	0,000
Остаток	81	4993,270	61,645		
Итого	85	10042,710			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
a	134,514	2,313	58,154	0,000	129,912	139,116
b_0	0,009	0,003	3,118	0,003	0,003	0,014
b_1	0,006	0,003	2,188	0,032	0,001	0,012
b_2	0,006	0,003	2,138	0,036	0,000	0,012
b_3	0,008	0,003	2,876	0,005	0,002	0,013

Рисунок 9.11 – Вывод итогов регрессионного анализа

Модель с распределенным лагом примет вид:

$$\tilde{y}_t = 134{,}514 + 0{,}009 \cdot x_t + 0{,}006 \cdot x_{t-1} + 0{,}006 \cdot x_{t-2} + 0{,}008 \cdot x_{t-3}.$$

Вычисленные значения t- критерия Стьюдента и F – критерия Фишера свидетельствуют о значимости как модели в целом, так и ее параметров.

Для нахождения параметров методом Алмон рассчитаем значения переменных Z_0, Z_1, Z_2 (столбцы 6-8 таблицы Л.1 приложения Л). Оценки параметров МНК найдем с помощью стандартной функции MS Excel. Результаты представлены на рисунке 9.12.

Регрессионная статистика	
Множественный R	0,709
R-квадрат	0,503
Нормированный R-квадрат	0,485
Стандартная ошибка	7,803
Наблюдения	86

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	3	5049,418	1683,139	27,641	0,000
Остаток	82	4993,292	60,894		
Итого	85	10042,710			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%
α	134,514	2,299	58,512	0,000	129,941	139,087
a_0	0,009	0,003	3,271	0,002	0,003	0,014
a_1	-0,003	0,005	-0,656	0,514	-0,013	0,007
a_2	0,001	0,002	0,626	0,533	-0,002	0,004

Рисунок 9.12 - Вывод итогов регрессионного анализа методом Алмон

В результате уравнение регрессии примет вид:

$$\tilde{y}_t = 134{,}514 + 0{,}009 \cdot Z_0 - 0{,}003 \cdot Z_1 + 0{,}001 \cdot Z_2 .$$

Вычислим оценки параметров:

$$\tilde{b}_0 = \tilde{a}_0 = 0{,}009,$$
$$\tilde{b}_1 = \tilde{a}_0 + \tilde{a}_1 + \tilde{a}_2 = 0{,}009 - 0{,}003 + 0{,}001 = 0{,}006,$$
$$\tilde{b}_2 = \tilde{a}_0 + 2\tilde{a}_1 + 4\tilde{a}_2 = 0{,}009 - 2 \cdot 0{,}003 + 4 \cdot 0{,}001 = 0{,}006,$$
$$\tilde{b}_3 = \tilde{a}_0 + 3\tilde{a}_1 + 9\tilde{a}_2 = 0{,}009 - 3 \cdot 0{,}003 + 9 \cdot 0{,}001 = 0{,}008.$$

Как видим, оценки параметров, полученные по методу Алмон, совпали с МНК – оценками. Таким образом, исходная спецификация с оцененными параметрами имеет такой же вид, как и полученная с помощью обычного МНК.

Рассмотрим интерпретацию параметров модели с распределенными лагами.

Коэффициент регрессии b_0 при переменной X_t характеризует среднее абсолютное изменение Y_t при изменении X_t на 1 единицу своего измерения в некоторый фиксированный момент времени *t*, без учета воздействия лаговых значений фактора *X*. Этот коэффициент называют *краткосрочным мультипликатором.*

В момент $t + 1$ совокупное воздействие факторной переменной X_t на результат у составит (b_0+b_1) условных единиц, в момент t+2 это воздействие можно охарактеризовать суммой $(b_0+b_1+b_2)$ и т. д. Полученные таким образом суммы называют *промежуточными мультипликаторами.*

С учетом конечной величины лага можно сказать, что изменение переменной x_t в момент t на 1 у. е. приведет к общему изменению результата через k моментов времени на $(b_0+b_1+\ldots+b_k)$ абсолютных единиц.

Введем следующее обозначение:

$$b_0+b_1+\ldots+b_k = b.$$

Величину *b* называют *долгосрочным мультипликатором*, который показывает абсолютное изменение в долгосрочном периоде $t + k$ результата *Y* под влиянием изменения на 1 ед. фактора *X*.

Предположим,

$$\beta_j = \frac{b_j}{b},\ j = 0:1.$$

Назовем полученные величины относительными коэффициентами модели с распределенным лагом. Если все коэффициенты b_j имеют одинаковые знаки, то для любого *j:*

$$0<\beta_j<1 \text{ и } \sum_{j=0}^{l}\beta_j = 1.$$

В этом случае относительные коэффициенты β_j являются весами для соответствующих коэффициентов b_j. Каждый из них измеряет долю общего изменения результативного признака в момент времени $t + j$.

Зная величины β_j, с помощью стандартных формул можно определить еще две важные характеристики модели множественной регрессии: *величину среднего и медианного лагов.*

Средний лаг рассчитывается по формуле средней арифметической взвешенной:

$$\bar{l} = \sum_{j=0}^{l} j \cdot \beta_j .$$

и представляет собой средний период, в течение которого будет происходить изменение результата под воздействием изменения фактора в момент времени *t*. Небольшая величина среднего лага свидетельствует об относительно быстром реагировании результата на изменение фактора, тогда как высокое его значение говорит о том, что воздействие фактора на результат будет сказываться в течение длительного периода времени.

Медианный лаг – это величина лага, для которого $\sum_{j=0}^{l_{ме}} \beta_j \approx 0{,}5$. Это тот период времени, в течение которого с момента времени *t* будет реализована половила общего воздействия фактора на результат.

Для нашего примера краткосрочный мультипликатор равен 0,009. Это означает, что увеличение индекса инвестиций в основной капитал на 1 % ведет в среднем к росту индекса выпуска по базовым видам экономической деятельности, с поправкой на сезонность на 0,009 %. в том же периоде. Под влиянием увеличения индекса инвестиций в основной капитал на 1 % индекс выпуска по базовым видам экономической деятельности, с поправкой на сезонность возрастет в момент времени $t + 1$ — на 0,009 + 0,006 = 0,015 %, $t + 2$ - на 0,015 + 0,006 = 0,022 %. Наконец, долгосрочный мультипликатор для данной модели составит: b = 0,009 + 0,006 + 0,006 + 0,008 = 0,029 %.

Относительные коэффициенты регрессии в этой модели равны:

$$\beta_1 = 0{,}009/0{,}029 = 0{,}307, \qquad \beta_2 = 0{,}006/0{,}029 = 0{,}205,$$
$$\beta_3 = 0{,}006/0{,}029 = 0{,}205, \qquad \beta_4 = 0{,}008/0{,}029 = 0{,}273.$$

Следовательно, 30,7 % общего увеличения индекс выпуска по базовым видам экономической деятельности, с поправкой на сезонность, вызванного ростом индекса инвестиций в основной капитал, происходит в текущем моменте времени; 20,5 % - в момент *t*+1; 20,5 % - в момент *t*+2 и 27,3 % этого увеличения приходится на момент времени *t*+3.

Средний лаг в данной модели определяется как

$$\bar{l} = 0 \cdot 0{,}307 + 1 \cdot 0{,}205 + 2 \cdot 0{,}205 + 3 \cdot 0{,}273 = 1{,}433 \text{ мес.}$$

Небольшая величина лага (около 1,5 мес.) подтверждает, что большая часть эффекта роста индекса инвестиций в основной капитал проявляется сразу же. Медианный лаг в данном примере составляет чуть менее 2 месяцев.

Параметры модели с бесконечной величиной лага обычным МНК или с помощью иных стандартных статистических методов определить нельзя, поскольку модель включает бесконечное число факторных переменных. Однако, приняв определенные допущения относительно структуры лага, оценки ее параметров все же можно получить. Эти допущения состоят в наличии геометрической структуры лага, т. е. такой структуры, когда воздействия лаговых значений фактора на результат уменьшаются с увеличением величины лага в геометрической прогрессии.

Изложенный в этом разделе подход к оценке параметров моделей с распределенным лагом впервые был предложен Л.М. Койком. Он предположил, что существует некоторый постоянный темп λ ($0 < \lambda < 1$) уменьшения во времени лаговых воздействий фактора на результат. Если, например, в период t результат изменялся под воздействием изменения фактора в этот же период времени на b_0 ед., то под воздействием изменения фактора, имевшего место в период *t* - 1, результат изменится

на $b_0 \cdot \lambda$ ед.; в период t - 2 - на $b_0 \bullet \lambda \bullet \lambda = b_0 \bullet b_0 \cdot \lambda^2$ ед. и т. д. Для некоторого периода t - l это изменение результата составит: $b_0 \bullet \lambda^l$ ед. В более общем виде можно записать:

$$b_j = b_0 \cdot \lambda^j ; j = 0,1,2,..., \ 0 < \lambda < 1. \qquad (9.37)$$

Ограничение на значения $X > 0$ обеспечивает одинаковые знаки для всех коэффициентов $b_j > 0$, а ограничение $\lambda < 1$ означает, что с увеличением лага значения параметров модели

$$y_t = a + b_0 \cdot x_t + b1 \cdot x_{t-1} + b_2 \cdot x_{t-2} + ... + \varepsilon_t . \qquad (9.38)$$

убывают в геометрической прогрессии. Чем ближе λ к 0, тем выше темп снижения воздействия фактора на результат во времени и тем большая доля воздействия на результат приходится на текущие значения фактора x_t.

Выразим с помощью формулы (9.37) все коэффициенты b_j в модели (9.38) через b_0 и λ:

$$y_t = a + b_0 \cdot x_t + b_0 \cdot \lambda \cdot x_{t-1} + b_0 \cdot \lambda^2 \cdot x_{t-2} + ... + \varepsilon_t . \qquad (9.39)$$

Тогда для периода t-1 модель (9.39) можно записать следующим образом:

$$y_{t-1} = a + b_0 \cdot x_{t-1} + b_0 \cdot \lambda \cdot x_{t-2} + b_0 \cdot \lambda^2 \cdot x_{t-3} + ... + \varepsilon_{t-1} . \qquad (9.40)$$

Умножим обе части модели (9.40) на λ, получим:

$$\lambda \cdot y_{t-1} = \lambda \cdot a + b_0 \cdot \lambda \cdot x_{t-1} + b_0 \cdot \lambda^2 \cdot x_{t-2} + b_0 \cdot \lambda^3 \cdot x_{t-3} + ... + \lambda \cdot \varepsilon_{t-1} . \qquad (9.41)$$

Вычтем найденное соотношение (9.41) из соотношения (9.39):

$$y_t - \lambda \cdot y_{t-1} = a - \lambda \cdot a + b_0 \cdot x_t + \varepsilon_{t-1} - \lambda \cdot \varepsilon_{t-1}. \quad (9.42)$$

В результате преобразований (9.42) мы получаем модель Койка:

$$y_t = a \cdot (1-\lambda) + b_0 \cdot x_t + \lambda \cdot y_{t-1} + u_t. \quad (9.43)$$

Полученная модель - это модель двухфакторной линейной регрессии (точнее — авторегрессии). Определив ее параметры, мы найдем λ и оценки параметров a и b_0 исходной модели. Далее с помощью соотношений (9.37) несложно определить параметры b_1, $b_{2,}$... модели (9.38). Применив обычный МНК к оценке параметров модели (9.43), получим смещенные оценки параметров ввиду наличия в этой модели в качестве фактора лаговой результативной переменной y_{t-1}.

Описанный выше алгоритм получил название «преобразования Койка». Это преобразование позволяет перейти от модели с бесконечными распределенными лагами к модели авторегрессии, содержащей две независимые переменные x_t и y_{t-1}.

Несмотря на бесконечное число лаговых переменных в модели (9.38), геометрическая структура лага позволяет определить величины среднего и медианного лагов в модели Койка. Поскольку сумма коэффициентов регрессии в модели (9.38) - это сумма геометрической прогрессии, то средний лаг определяется как $\bar{l} = \frac{\lambda}{1-\lambda}$.

Нетрудно заметить, что при $\lambda = 0{,}5$ средний лаг $\bar{l} = 1$, а при $\lambda < 0{,}5$ средний лаг $\bar{l} < 1$, т. е. воздействие фактора на результат в среднем занимает менее одного периода времени. Величину (1-λ) интерпретируют обычно как скорость, с которой происходит адаптация результата во времени к изменению факторного признака. Для расчета медианного лага необходимо выполнение следующего условия:

$$\sum_{j=0}^{l_{me}-1} \beta_j = 0{,}5.$$

Поэтому медианный лаг в модели Койка равен:

$$l_{Me} = \frac{\ln 0{,}5}{\ln \lambda}.$$

Рассмотрим пример оценки спецификации модели (9.43) по данным таблицы Л.1 приложения Л.

Оцененная форма спецификации 9.43 имеет вид:

$$y_t = 2{,}94 - 0{,}0003 \cdot x_t + 0{,}985 \cdot y_{t-1},$$

где $\tilde{a} = 197{,}79;\ \tilde{b}_0 = -0{,}0003;\ \tilde{\lambda} = 0{,}985$

Средний лаг определяется как $\bar{l} = \frac{0{,}985}{1 - 0{,}985} = 66{,}197$.

Медианный лаг $l_{Me} = \frac{\ln 0{,}5}{\ln 0{,}985} = \frac{-0{,}693}{-0{,}015} = 46{,}23$.

Таким образом, в среднем воздействие индекса инвестиций в основной капитал на индекс выпуска по базовым видам экономической деятельности, с поправкой на сезонность проявляется в течение 66 месяцев, причем половина этого воздействия реализуется в течение первых 46 месяцев, с момента изменения индекса инвестиций в основной капитал.

9.3 Модели адаптивных ожиданий и неполной корректировки

Если в модели не фактическое значение переменной, а ее желаемое (ожидаемое) значение, то такие модели относят ко второму типу ДЭМ - моделям адаптивных ожиданий, либо к моделям частичной (неполной) корректировки.

Модель адаптивных ожиданий учитывает желаемое (ожидаемое) значение факторного признака x^*_{t+1}. Например, ожидаемое в будущем (в период *t*+1) значение курса доллара x^*_{t+1} влияет на инвестиции в текущем периоде y_t.

В общем виде модель адаптивных ожиданий записывается так:

$$y_t = a + b_0 \cdot x^*_{t+1} + u_t . \qquad (9.44)$$

Желаемое (ожидаемое) значение переменных определяется по значению реальных (фактических) переменных в предыдущий период (t).

Механизм формирования ожиданий в модели адаптивных ожиданий следующий:

$$x^*_{t+1} - x^*_t = \lambda \cdot \left(x_t - x^*_t\right) \qquad (9.45)$$

или

$$x^*_{t+1} = \lambda \cdot x_t + (1-\lambda) \cdot x^*_t \qquad (9.46)$$

где $0 \le \lambda \le 1$.

Т.е. значение переменной, ожидаемое в следующий период x^*_{t+1}, формируется как среднее арифметическое взвешенное ее реального и ожидаемого значения в текущем периоде. Чем больше величина λ, тем быстрее ожидаемое значение адаптируется предыдущим реальным значениям. Чем меньше λ, тем ожидаемое значение в будущем ближе к ожидаемому значению предыдущего периода x^*_t, т.е. тенденции в ожиданиях сохраняются.

Для оценки параметров модели адаптивных ожиданий обычный МНК применить невозможно, т.к. модель включает ожидаемые значения факторной переменной, которые нельзя получить эмпирическим путем. Поэтому для оценки параметров исходную модель преобразуют. Используя выражение (9.46), преобразуем модель адаптивных ожиданий к виду:

$$y_t = a + b_0 \cdot (\lambda \cdot x_t + (1-\lambda) \cdot x_t^*) + u_t = a + \lambda b_0 x_t + (1-\lambda) b_0 x_t^* + u_t . \quad (9.47)$$

Умножим модель 1 для периода *t*-1 на (1-λ) и получим:

$$(1-\lambda) \cdot y_{t-1} = (1-\lambda) \cdot a + (1-\lambda) b_0 \cdot x_t^* + (1-\lambda) \cdot u_{t-1} . \quad (9.48)$$

Вычтем почленно полученное выражение из преобразованной модели:

$$y_t - (1-\lambda) \cdot y_{t-1} = a - (1-\lambda) \cdot a + \lambda \cdot b_0 \cdot x_t + (1-\lambda) \cdot u_{t-1} . \quad (9.49)$$

или

$$y_t = \lambda a + \lambda b_0 x_t + (1-\lambda) y_{t-1} + u_t^* , \quad (9.50)$$

где $u_t^* = u_t - (1-\lambda) \cdot u_{t-1}$.

Мы получили модель авторегрессии, определив параметры которой, можно легко перейти к параметрам исходной модели.

Полученная модель включает только фактические значения переменных, поэтому ее параметры можно определить с помощью стандартных статистических методов.

Исходная модель адаптивных ожиданий (9.44), характеризующая зависимость результативного признака от ожидаемых значений факторного признака, называется долгосрочной функцией модели адаптивных ожиданий.

Модель 9.50 называется краткосрочной функцией модели адаптивных ожиданий.

Модель частичной корректировки относится ко 2 типу ДЭМ и учитывает желаемое (ожидаемое) значение результативного признака y_t^*. В общем виде такую модель можно записать как:

$$y_t^* = a + b_0 x_t + u_t. \qquad (9.51)$$

Ожидаемое значение переменных определяется по значению реальных (фактических) переменных в предыдущий момент времени t-1.

В таких моделях предполагается, что фактическое приращение зависимой переменной $y_t - y_{t-1}$ пропорционально разнице между желаемым уровнем и фактическим значением в предыдущий период

$$y_t - y_{t-1} = \lambda\left(y_t^* - y_{t-1}\right) + v_t. \qquad (9.52)$$

или

$$y_t = \lambda \cdot y_t^* + (1-\lambda) \cdot y_{t-1} + v_t, \qquad (9.53)$$

где $0 \le \lambda \le 1$.

Из этого следует, что y_t получается как среднее арифметическое взвешенное желаемого уровня y_t^* и фактического значения этой переменной в предыдущем периоде y_{t-1}. Чем больше величина λ, тем быстрее происходит процесс корректировки.

Если $\lambda = 1$ то $y_t = y_t^*$ и полная корректировка происходит за 1 период.

Если $\lambda = 0$ то корректировка y_t не происходит.

Подставив (9.51) в (9.53) получим:

$$y_t = a \cdot \lambda + b_0 \cdot \lambda \cdot x_t + (1-\lambda) \cdot y_{t-1} + v_t + \lambda \cdot u_t = a \cdot \lambda + b_0 \cdot \lambda \cdot x_t + (1-\lambda) \cdot y_{t-1} + \varepsilon_t \qquad (9.54)$$

Параметры преобразованного уравнения a, λ, b_0 могут быть оценены с помощью МНК. Соотношение (9.54) называют краткосрочной функцией МЧК, а уравнение (9.51) – долгосрочной функцией МЧК [48,с. 483-487].

9.4 Вопросы для самоконтроля

1. Дайте определение стационарного временного ряда в узком и широком смысле.
2. Назовите виды моделей стационарных временных рядов.
3. Перечислите основные свойства марковского процесса и процесса Юла.
4. Как рассчитываются краткосрочный, промежуточный и долгосрочный-мультипликаторы в моделях с конечным числом лагов?
5. В чем отличия моделей адаптивных ожиданий и частичной корректировки?

9.5 Тесты

1. Модели с распределительным лагом это:

а) модели, в которых содержится не только текущие, но и лаговые значения факторных переменных

б) модели, в которых в качестве факторных переменных содержится лаговые значения результативной переменной;

в) модели, в которых в качестве факторных переменных используются фиктивные переменные;

г) модели, в которых в качестве факторных переменных используются качественные переменные.

2. Авторегрессионные модели – это

а) модели, в которых содержится не только текущие, но и лаговые значения факторных переменных

б) модели, в которых в качестве факторных переменных содержится лаговые значения результативной переменной;

в) модели, в которых в качестве факторных переменных используются фиктивные переменные;

г) модели, в которых в качестве факторных переменных используются качественные переменные.

3. Распределение весов в модели Алмон:

а) геометрическое;

б) арифметическое;

в) полиномиальное.

4. Веса в модели Койка с увеличением лага:

а) убывают;

б) возрастают;

в) не меняются;

г) меняются в зависимости от влияния лаговых переменных на эндогенную.

5. Полиномиальное распределение весов в моделях с распределенным лагом имеет распределение:

а) Стьюдента;

б) Алмон;

в) Койка.

10 Корреляция и регрессия по временным рядам

Что необходимо знать из главы 10:

1. Корреляция между временными рядами: сущность, ограничения.
2. Методы измерения корреляции по временным рядам.
3. Регрессия по вмененным рядам и прогнозирование на ее основе.

10.1 Корреляция между временными рядами: сущность, ограничения

Предполагается, что читатель знаком с теорией корреляции в пространственных совокупностях и ее показателями, которые здесь используются. Корреляция временных рядов применяется для решения следующих задач:

1. Взамен пространственной корреляции ввиду отсутствия однородной совокупности или данных о таковой. Например, при изучении связи между средним душевым доходом в стране и душевым потреблением картофеля. Совокупность стран явно неоднородна, не везде потребляется картофель, единственная возможность измерить связь – по данным той же страны за ряд лет.

2. При изучении взаимодействующих процессов, например, при изучении связи между урожайностью и колебаниями солнечной активности. Изучать эту связь по пространственной совокупности вообще невозможно: для всех регионов на Земле показатели солнечной активности одинаковы.

Корреляция между двумя (для простоты возьмем два) признаками означает, что если величина одного из них больше средней по совокупности, то и величина другого, в основном, тоже больше его средней (прямая связь) или же, в основном, меньше его средней (обратная связь). Но если оба признака имеют одинаково направленные тренды, то уровни лет после середины периода, как правило, больше средних величин, или, при трендах к снижению, оба признака имеют уровни меньше средних. Выходит, что в динамике между любыми признаками, имеющими тенденцию изменения, всегда есть связь: либо прямая (оба тренда в одном направлении),

либо обратная (тренды в разных направлениях). Результат абсурдный. В любой развитой стране в 1970–1990-х годах рос уровень производства компьютеров. Одновременно, росло число инфицированных ВИЧ-инфекцией и больных СПИД. Но при очень высокой корреляции уровней обоих рядов, никакой реальной связи процессов нет. Это один из видов "ложной корреляции". Как же отличить ложную корреляцию от истинной? Конечно, прежде всего, как и при изучении связей в пространственной совокупности, нужно обосновать связь по существу, объяснить ее причинный механизм. Эта задача не статистическая, в данном учебнике не рассматривается. Решается специалистом в той сфере знаний, которая изучает объект, процесс: агрономом, инженером, экономистом, социологом, биохимиком, астрономом и т.д. Без причинного обоснования лучше не начинать измерение связи в динамике.

Но даже и после такого обоснования остается открытым вопрос: при наличии одинаково направленных трендов двух причинно-связанных признаков не преувеличится ли теснота связи за счет наличия трендов? Если, например, в стране растет производство и применение минеральных удобрений, растет и урожайность сельхозкультур, но ведь она растет не только по причине роста применения удобрений, а также и за счет других факторов: селекции новых сортов, мелиорации, орошения, механизации производства, роста экономической заинтересованности фермеров и еще ряда факторов. А при коррелировании уровней урожайности и доз удобрений за 20–25 лет, прогресс всех факторов урожайности будет "списан" на дозу удобрений. Получится коэффициент детерминации, превышающий 50 или даже 70 %, и где гарантия, что к истинной корреляции и здесь не примешана ложная? Такой гарантии нет.

Могут возразить: - А разве не может так случиться, что и в пространственной совокупности предприятий, у тех из них, которые вносят больше дозы минеральных удобрений, одновременно и семена лучше, и сельхозмашины, и кадры более подготовлены, и экономика сильнее? Да, это возможно, но именно лишь возможно, как возможно и несовпадение факторов, влияющих на урожайность. А параллельная тенденция динамики факторов во времени - это не просто возможность, а в 90 % стран и регионов – достоверный факт. Так что "примесь ложной корреляции" в про-

странственных совокупностях намного меньше, чем при коррелировании временных рядов. И, следовательно, если есть возможность изучить, измерить, моделировать связь результативного признака с его факторами не по рядам динамики, а в пространственной совокупности – так и следует поступать.

Проблема ложной корреляции почти целиком снимается, если причинная связь обоснована не столько между тенденциями динамики, сколько между колебаниями факторного и результативного признаков. Например, колебания урожайности во влагонедостаточных регионах, как Оренбургская область, причинно связана не с какой-либо тенденцией изменения суммы осадков, а с её колебаниями в отдельные годы. К тенденции же роста урожайности осадки никакого отношения (причинной связи) не имеют. Снимается ложная корреляция тем, что колебания других факторов, влияющих на урожайность – экономических, организационных – не связаны или слабо связаны с колебаниями осадков. Тенденции факторов связаны часто, колебания – почти никогда. Поэтому связь между колебаниями одного фактора с результативным показателем (его колебаниями) почти всегда свободна от ложной корреляции, наведенной другими факторами.

В последующих разделах данной главы речь и будет идти, в основном, о корреляции между колебаниями признаков, о методиках ее измерения и моделирования. Что же касается проблемы измерения связи между тенденциями, между самими уровнями временных рядов, включающих тенденцию, а не только колебания, то эта проблема не может считаться решенной. Излагаемые здесь же методики решают только ограниченный класс задач – измерение связи между колебаниями факторного (факторных) признака и колебаниями результативного признака.

Строго говоря, это жесткое ограничение не вполне новое, оно относится и к пространственной корреляции, в том смысле, что и в ней измеряется связь вариации результативного признака с вариацией фактора. За счет вариации дозы минеральных удобрений объясняется 38 % вариации урожайности пшеницы между хозяйствами области (r^2=0,38), а не 38 % уровня урожайности, как иногда неверно говорят [10, с. 83-185].

10.2 Методы измерения корреляции по временным рядам

Как было изложено выше, наличие ложной корреляции связано с тенденцией каждого из рядов динамики, с автокорреляцией их уровней. Поэтому даже если корреляция рядов динамики экономически оправдана, при построении регрессионной модели для последующего прогноза требуется их предварительная специальная обработка.

Чтобы иметь возможность использовать корреляционные методы для изучения связей по динамическим рядам, нужно исключить влияние автокорреляции и сделать уровни каждого из взаимосвязанных рядов статистически независимыми. Если ряды динамики характеризуются не только тенденцией, но и периодическими колебаниями, то при исследовании корреляции по рядам динамики следует учесть оба фактора, т.е. из первоначальных данных должна быть исключена как тенденция, так и периодическая составляющая и лишь затем измерена корреляция рядов динамики. По изменениям случайной компоненты одного ряда в зависимости от колеблемости случайной компоненты другого ряда можно судить и о тесноте связи между исследуемыми рядами динамики. Однако и остаточные величины (отклонения уровней от тренда) могут оказаться автокоррелированными в силу неправильно выбранного вида тренда. Поэтому следует проверять наличие автокорреляции в остатках (основные подходы нами были рассмотрены ранее).

Если выдвигается гипотеза, что остаточные величины связаны между собой нелинейными соотношениями, то обобщенная оценка тесноты связи может быть дана через индекс корреляции:

$$R = \sqrt{1 - \frac{S_{yx}^2}{\sigma_{d_y}^2}}, \qquad (10.1)$$

где $\sigma_{d_y}^2 = \frac{\sum \left(d_y - \bar{d}_y\right)^2}{n}$ – дисперсия остаточных величин результативного признака;

$S^2_{d_{yx}} = \frac{\sum\left(d_y - \hat{d}_{yx}\right)^2}{n-p}$ – дисперсия, характеризующая отклонения фактических значений остатков результативного признака от теоретических, рассчитанных на основе уравнения регрессии.

Абсолютная величина индекса корреляции находится в пределах: $0 \le R \le 1$. Чем ближе R к 1, тем теснее связь.

Если предполагается линейная связь между остаточными величинами рядов, то теснота связи между двумя динамическими рядами измеряется линейным коэффициентом корреляции. Он может быть определен по отклонениям от тренда или по последовательным разностям:

- по отклонениям от тренда

$$r_{d_y d_x} = \frac{\sum d_y d_x}{\sqrt{\sum d_y^2 \sum d_x^2}}, \qquad (10.2)$$

где $d_y = y_t - \tilde{y}_t$; $d_x = x_t - \tilde{x}_t$ – остаточные величины;

$\tilde{y}_t, \tilde{x}_t$ – теоретические значения по уравнениям тренда.

- по последовательным разностям

$$r_{\Delta_y \Delta_x} = \frac{\left(\overline{\Delta_y^{(k)} \Delta_x^{(k)}} - \overline{\Delta_y^{(k)}} \cdot \overline{\Delta_x^{(k)}}\right)}{\sigma_{\Delta_y^{(k)}} \cdot \sigma_{\Delta_x^{(k)}}}, \qquad (10.3)$$

где $\Delta_y^{(k)}, \Delta_x^{(k)}$ – последовательные разности (формула (8.38));

k - порядок разностей.

Коэффициент корреляции принимает значения в интервале $-1 \le r \le 1$. Отрицательная величина его указывает на обратную связь между динамикой явлений. Чем коэффициент корреляции ближе по абсолютной величине к 1, тем теснее рассматриваемая связь.

Метод отклонений от тренда является более точным, т.к. позволяет учесть любой тип тенденции описываемым уравнением тренда. При использовании метода последовательных разностей может быть исключена только тенденция, описываемая полиномами различных степеней (при линейной тенденции берутся первые разности, при параболической – вторые разности и т.д.)

В качестве примера по данным таблице 10.1 оценим влияние расходов на конечное потребление (в текущих рыночных ценах; миллионов рублей) на ВВП в расчете на душу населения, р. (1992-1997 гг.- тыс. р.)[1].

Если коррелировать исходные уровни, то коэффициент корреляции составит $r_{xy} = 0{,}997$, однако из-за наличия в каждом из рядов четкой тенденции можно предположить, что это значение имеет смещение. Применим метод устранения тенденции и оценим тесноту связи между рассматриваемыми показателями, используя рассмотренные коэффициенты корреляции.

Для расчета коэффициента корреляции по отклонениям от тренда необходимо для временных рядов факторного и результативного признаков провести аналитическое выравнивание. Каждый из рядов имеет ускоренно повышающуюся тенденцию. Поэтому в качестве аппроксимирующей модели целесообразно принять полином второго порядка. В результате аналитического выравнивания нами получены уравнения трендов:

$$\tilde{x} = 2000000 - 1000000 \cdot t + 13071 \cdot t^2 \quad \left(R^2 = 0{,}9926\right)$$

$$\tilde{y} = 16717 - 9429{,}4 \cdot t + 1250{,}3 \cdot t^2 \quad \left(R^2 = 0{,}9865\right)$$

Вспомогательные расчеты для вычисления коэффициента корреляции представлены в таблице 10.2.

[1] По данным сайта http://www. gks.ru

Таблица 10.1 - Исходные данные для проведения корреляционного и регрессионного анализа по временным рядам

Год	ВВП в расчете на душу населения, р. (1992-1997 гг.- тыс. р.) *y*	Расходы на конечное потребление (в текущих рыночных ценах, млн. р.) *x*	Фактор времени *t*
1	2	3	4
1991	9,4	855,4	1
1992	128,0	9183,6	2
1993	1155,3	106755,4	3
1994	4115,3	422052,7	4
1995	9627,7	1016594,3	5
1996	13551,7	1435869,8	6
1997	15836,9	1776137,6	7
1998	17807,3	2003790,1	8
1999	32763,2	3285678,1	9
2000	49834,9	4476850,9	10
2001	61267,3	5886860,6	11
2002	74457,9	7484115,5	12
2003	91364,8	9058687,6	13
2004	118391,4	11477849,6	14
2005	150997,0	14438149,2	15
2006	188909,5	17809740,7	16
2007	233948,1	21968579,5	17
2008	290771,3	27543511,4	18
2009	273318,2	29351191,6	19
2010	314395,5	32070250,9	20

Таблица 10.2 - Расчет коэффициента корреляции по отклонениям от тренда

t	y	x	$\tilde{y}$	$\tilde{x}$	d_y	d_x	$d_y \cdot d_x$	d_y^2	d_x^2
A	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1991	9,4	855,4	8537,9	1130716	-8528,5	-1129861	9636019539	72735312,3	1276585879321,0
1992	128,0	9183,6	2859,4	522864	-2731,4	-513680	1403065552	7460546,0	263867142400,0
1993	1155,3	106755,4	-318,5	176444	1473,8	-69688,6	-102707058,7	2172086,4	4856500970,0
1994	4115,3	422052,7	-995,8	91456	5111,1	330596,7	1689712793	26123343,2	109294178050,9
1995	9627,7	1016594,3	827,5	267900	8800,2	748694,3	6588659579	77443520,0	560543154852,5
1996	13551,7	1435869,8	5151,4	705776	8400,3	730093,8	6133006948	70565040,1	533036956798,4
1997	15836,9	1776137,6	11975,9	1405084	3861,0	371053,6	1432637950	14907321,0	137680774073,0
1998	17807,3	2003790,1	21301	2365824	-3493,7	-362034	1264838186	12205939,7	131068617156,0
1999	32763,2	3285678,1	33126,7	3587996	-363,5	-302317,9	109892556,7	132132,3	91396112660,4
2000	49834,9	4476850,9	47453	5071600	2381,9	-594749,1	-1416632881	5673447,6	353726491950,8
2001	61267,3	5886860,6	64279,9	6816636	-3012,6	-929775	2801040165	9075758,8	864481550625,0
2002	74457,9	7484115,5	83607,4	8823104	-9149,5	-1338989	12251079856	83713350,3	1792891542121,0
2003	91364,8	9058687,6	105436	11091004	-14070,7	-2032316	28596108741	197984598,5	4130308323856,0
2004	118391,4	11477849,6	129764	13620336	-11372,8	-2142486	24366064781	129340579,8	4590246260196,0
2005	150997,0	14438149,2	156594	16411100	-5596,5	-1972951	11041620272	31320812,3	3892535648401,0
2006	188909,5	17809740,7	185923	19463296	2986,1	-1653555	-4937680586	8916793,2	2734244138025,0
2007	233948,1	21968579,5	217754	22776924	16194,2	-808345	-13090500599	262252113,6	653421639025,0
2008	290771,3	27543511,4	252085	26351984	38686,3	1191527	46095770980	1496629807,7	1419736591729,0
2009	273318,2	29351191,6	288917	30188476	-15598,5	-837284	13060374474	243313202,3	701044496656,0
2010	314395,5	32070250,9	328249	34286400	-13853,5	-2216149	30701420172	191919462,3	4911316390201,0
2011	-	-	370082	38645756	-	-	-	-	-
Итого	1942650,7	191622705	1942527	205154920	123,7	-13532215	177623791418,05	2943885167,17	29152282389068,00
В среднем	97132,54	9581135,2	97126,4	10257746	6,185	-676610,8	8881189570,90	147194258,36	1457614119453,40

Таблица 10.3 - Расчет коэффициента корреляции по последовательным разностям

t	y	x	Δy	Δx	$\Delta'' y$	$\Delta'' x$	$\Delta'' y \cdot \Delta'' x$	$\left(\Delta'' y - \overline{\Delta'' y}\right)^2$	$\left(\Delta'' x - \overline{\Delta'' x}\right)^2$
A	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
1991	9,4	855,4	-	-	-	-	-	-	-
1992	128,0	9183,6	118,6	8328,2	-	-	-	-	-
1993	1155,3	106755,4	1027,3	97571,8	908,7	89243,6	81095659,3	1868096,7	3764138118,3
1994	4115,3	422052,7	2960,0	315297,3	1932,7	217725,5	420798073,9	117500,4	4506346647,9
1995	9627,7	1016594,3	5512,4	594541,6	2552,4	279244,3	712743151,3	76682,8	16550340780,7
1996	13551,7	1435869,8	3924,0	419275,5	-1588,4	-175266,1	278392673,2	14929594,4	106186220457,8
1997	15836,9	1776137,6	2285,2	340267,8	-1638,8	-79007,7	129477818,8	15321614,0	52717938139,4
1998	17807,3	2003790,1	1970,4	227652,5	-314,8	-112615,3	35451296,4	6709567,7	69280279109,4
1999	32763,2	3285678,1	14955,9	1281888,0	12985,5	1054235,5	13689775085,3	114704457,0	816564034706,7
2000	49834,9	4476850,9	17071,7	1191172,8	2115,8	-90715,2	-191935220,2	25498,8	58231178363,8
2001	61267,3	5886860,6	11432,4	1410009,7	-5639,3	218836,9	-1234086930,2	62643795,2	4656796927,6
2002	74457,9	7484115,5	13190,6	1597254,9	1758,2	187245,2	329214510,6	267582,0	1343151237,1
2003	91364,8	9058687,6	16906,9	1574572,1	3716,3	-22682,8	-84296089,6	2075952,7	30025602214,4
2004	118391,4	11477849,6	27026,6	2419162,0	10119,7	844589,9	8546996411,0	61531735,1	481627294194,9
2005	150997,0	14438149,2	32605,6	2960299,6	5579,0	541137,6	3019006670,4	10913222,4	152522606810,7
2006	188909,5	17809740,7	37912,5	3371591,5	5306,9	411291,9	2182684984,1	9189487,0	67962262481,6
2007	233948,1	21968579,5	45038,6	4158838,8	7126,1	787247,3	5610002984,5	23528482,0	405324658500,7
2008	290771,3	27543511,4	56823,2	5574931,9	11784,6	1416093,1	16688090746,3	90423299,8	1601482474215,0
2009	273318,2	29351191,6	-17453,1	1807680,2	-74276,3	-3767251,7	279817517444,7	5860175531,5	15349531949876,2
2010	314395,5	32070250,9	41077,3	2719059,3	58530,4	911379,1	53343383274,6	3164615649,2	578790663198,1
Итого	1942650,7	191622705	314386,1	32069395,5	40958,7	2710731,1	383374312544,5	9439117748,8	19801067935980,3
В среднем	97132,54	9581135,23	16546,64	1687862,92	2275,48	150596,17	21298572919,1	524395430,5	1100059329776,7

Коэффициент корреляции рядов *x* и *y* по отклонениям от тренда составит:

$$r_{d_y d_x} = \frac{177623791418{,}05}{\sqrt{294385167{,}17 \cdot 29152282389068{,}00}} = 0{,}606 .$$

Полученное значение коэффициента свидетельствует о наличии прямой связи средней силы между колебаниями расходов на конечное потребление и колебаниями ВВП в расчете на душу населения.

Для расчета коэффициента корреляции по методу последовательных разностей построим вспомогательную таблицу 10.3.

Подставляя в формулу расчетные данные, получим:

$$r_{\Delta_y \Delta_x} = \frac{(212985729{19{,}1} - 2275{,}48 \cdot 150596{,}17)}{\sqrt{524395430{,}5} \cdot \sqrt{1100059329776{,}7}} = 0{,}873 .$$

Следовательно, можно сделать вывод о наличии прямой тесной связи скорости ряда расходов на конечное потребление и ВВП в расчете на душу населения.

10.3 Регрессия по вмененным рядам и прогнозирование на ее основе

Уравнение регрессии по рядам динамики можно построить тремя способами [51]:

1) регрессия по последовательным разностям;

2) регрессия по отклонениям от тренда;

3) регрессия по уровням ряда с включением в нее фактора времени.

В каждом из этих способов оценка параметров регрессии дается традиционным методом наименьших квадратов, т.е. как и при построении уравнения регрессии в статике и при построении уравнения трендов. Рассмотрим интерпретацию параметров регрессии и ее использование при прогнозировании.

Математически доказано, что при наличии во временном ряду линейной тенденции ее можно устранить, перейдя к первым разностям, т.е. к цепным абсолютным приростам (Δ_t); при тенденции в виде параболы второй степени для ее устранения берутся вторые разности, т.е. абсолютные ускорения (Δ_t''); если тенденция описывается полиномом третьей степени, то рассчитываются третьи разности и т.д.

Модели регрессии, построенные по разностям второго порядка и выше мало информативны. Поэтому ограничимся рассмотрением модели регрессии по первым разностям.

В уравнении *регрессии по первым разностям*

$$\Delta_y = a + b\Delta_{x_t} \qquad (10.4)$$

параметр b показывает на сколько изменится скорость роста результативного признака с изменением скорости роста факторного признака на единицу своего измерения. Чтобы использовать данное уравнение в прогнозировании, необходимо определить на перспективу скорость развития факторного признака, тогда рост скорости результативного признака составит: $\Delta y_p = a + b\Delta x_p$. От данного уравнения можно перейти к уравнению, в котором прогнозируется уровень ряда, а не его скорость. Для этого необходимо раскрыть содержание абсолютного прироста, выразив его через соответствующие значения уровней ряда:

$$(y_p - y_n) = a + b(x_p - x_n), \qquad (10.5)$$

где y_p – прогнозное значение результативного признака;

y_n – конечный уровень динамического ряда;

x_p – прогнозное значение факторного признака результативного признака;

x_n – конечный уровень факторного признака.

Соответственно прогнозное значение для ряда y составит:

$$y_p = y_n + a + b(x_p - x_n). \qquad (10.6)$$

Пример 10.1 - По данным таблицы 10.3, построим уравнение регрессии по первым разностям. Используя МНК, получим уравнение регрессии:

$$\Delta y = -1788{,}21 + 0{,}01\Delta x \left(R^2 = 0{,}76;\ DW = 2{,}36\right).$$

Коэффициент регрессии показывает, что увеличение абсолютного прироста расходов на конечное потребление на 1 млн. р. приводит в среднем к увеличению абсолютного прироста ВВП в расчете на душу населения на 0,01 р.

Расчетное значение ВВП в расчете на душу населения на 2011 г. при увеличении расходов на конечное потребление на 1 млн. р. относительно 2010 г., составит:

$$y_{2011} = 314395{,}5 - 1788{,}21 + 0{,}01(33070250{,}9 - 32070250{,}9) = 323470{,}05 \text{ р.}$$

Регрессия по отклонениям от тренда имеет вид $d_y = a + b \cdot d_x + \varepsilon_t$. Коэффициент регрессии b показывает, на сколько в среднем изменяется величина отклонений от тренда по ряду y с изменением случайных колебаний ряда x на одну единицу.

Если в анализируемых временных рядах наблюдается тенденция, описываемая линейным трендом, то уравнение регрессии принимает вид: $d_y = b \cdot d_x + \varepsilon_t$. Коэффициент регрессии в данном случае означает, что случайные отклонения по ряду y в среднем в b раз выше случайных колебаний по ряду x.

Для прогноза удобно от уравнения в отклонениях от тренда перейти к уравнению, связывающему между собой конкретные уровни временных рядов. Подставим в уравнение регрессии по отклонениям от тренда значения dy и dx:

$$(y_t - \tilde{y}_t) = a + b \cdot (x_t - \tilde{x}_t), \qquad (10.7)$$

откуда

$$y_t = \tilde{y}_t + a + b \cdot (x_t - \tilde{x}_t). \qquad (10.8)$$

Данную модель можно использовать для прогноза:

$$y_p = \tilde{y}_{t=p} + a + b\left(x_p - \tilde{x}_{t=p}\right), \qquad (10.9)$$

где y_p – прогнозное значение результативного признака;

$\tilde{y}_{t=p}$ – прогноз по тренду результативного признака;

x_p – прогнозное значение факторного признака;

$\tilde{x}_{t=p}$ – прогноз по тренду факторного признака.

Результат прогноза зависит от качества прогноза фактора x, от качества трендовых моделей, используемых для прогнозирования.

В качестве примера рассмотрим оценку параметров модели с помощью МНК, по данным таблицы 10.3. Уравнение регрессии примет вид:

$$d_y = 6019{,}26 + 0{,}009 \cdot d_x; \left(R = 0{,}54; DW = 1{,}23\right)$$

Коэффициент регрессии показывает, что в среднем на 0,009 р. изменяется величина отклонений от тренда по ряду ВВП в расчете на душу населения с изменением случайных колебаний ряда расходов на конечное потребление на 1 млн.р.

Подставив соответствующие значения в модель, получим расчетное значение ВВП в расчете на душу населения на 2011 г. если расходы на конечное потребление составят x_p = 33070250,9 млн.р.:

$$y_p = 370082 + 6019{,}26 + 0{,}009 \cdot \left(33070250{,}9 - 38645756\right) = 325921{,}71 \text{ р.}$$

Математически доказано, что если при измерении связи по динамическим рядам непосредственно ввести в уравнение регрессии фактор времени *t* и определять параметры уравнения по исходным уровням, то автокорреляция в рядах динамики

будет устранена. Это значит, что при изучении связи между двумя признаками по динамическим рядам следует при линейной их зависимости искать уравнение вида:

$$\tilde{y}_t = a + bx + ct\,. \qquad (10.10)$$

Параметры такого уравнения также находятся МНК. Параметр b фиксирует силу связи y с x, т.е. от показывает среднее изменение y с изменением x на единицу своего измерения при неизменной тенденции.

Параметр c характеризует среднегодовой абсолютный прирост результативного признака, при закреплении фактора x на постоянном уровне.

Оценим параметры спецификации модели (10.10) по данным таблицы 10.1, используя МНК. Уравнение регрессии примет вид:

$$\tilde{y}_t = -3054{,}07 + 0{,}01 \cdot x + 743{,}99 \cdot t; \left(R^2 = 0{,}995; DW = 1{,}57\right)$$

Параметр b показывает, что с ростом расходов на конечное потребление на 1 млн.р. ВВП в расчете на душу населения увеличится на 0,01 р. при неизменной тенденции. Среднегодовой абсолютный прирост ВВП в расчете на душу населения составил 743,99 р. при условии неизменности расходов на конечное потребление.

Теоретическое значение ВВП в расчете на душу населения для 2011 г. при расходах на конечное потребление x_p = 33070250,9 млн.р. составит:

$$\tilde{y}_t = -3054{,}07 + 0{,}01 \cdot 33070250{,}9 + 743{,}99 \cdot (20+1) = 331410{,}25 \text{ р.}$$

В настоящее время чаще всего строится регрессия по рядам динамики с *введением в модель фактора времени*. Это связано с тем, что при таком подходе упрощается обработка материала: не нужно определять тренды по всем рядам динамики, искать отклонения по ним, строить модель по отклонениям от трендов и переходить далее обратно к уровням.

Фактор времени чаще всего вводится в модель в виде линейного члена, даже если другие факторы подвергаются логарифмированию или иному преобразованию.

Если во временных рядах наблюдается тенденция, описываемая полиномом второго и более высоких порядков, то для случая многофакторной зависимости строится регрессия вида:

$$\tilde{y}_t = a + b_1 \cdot x_1 + b_2 \cdot x_2 + ... + b_p \cdot x_p + c_1 \cdot t + c_2 \cdot t^2 + ... + c_k \cdot t^k . \qquad (10.11)$$

Вместе с тем, если во временных рядах наблюдается четкая тенденция и значения коэффициентов корреляции факторов x с фактором времени t превышают значения коэффициентов корреляции факторов x с зависимой переменной y, т.е. $r_{xt} > r_{xy}$, то коэффициенты регрессии оказываются ненадежными и экономически неинтерпретируемыми.

При увеличении числа факторов, включаемых в регрессию, рассмотренные проблемы устранения автокорреляции уровней рядов динамики остаются, но появляются новые, связанные с построением множественной регрессии: мультиколлинеарность факторов, отбор их и др. [51, с. 214-216].

10.4 Вопросы для самоконтроля

1. В чем специфика построения регрессионной модели по рядам динамики?
2. Перечислите основные способы построения регрессионных моделей по рядам динамики. Какой способ применяется на практике чаще?
3. Назовите основные способы оценки тесноты и направления связи по рядам динамики.
4. В чем суть построения модели регрессии первых разностей?
5. В чем суть построения модели регрессии по отклонениям от тренда?
6. В чем суть построения модели регрессии с включением фактора времени?

10.5 Тесты

1. В уравнении регрессии по первым разностям $\Delta y = a + b\Delta x$ параметр b:

а) показывает, на сколько изменится скорость роста результативного признака с изменением скорости роста факторного признака на единицу своего измерения;

б) показывает, во сколько раз изменится скорость роста результативного признака с изменением скорости роста факторного признака на единицу своего измерения;

в) означает, что случайные отклонения по ряду y в среднем в b раз выше случайных колебаний по ряду x;

г) фиксирует силу связи y с x, т.е. он показывает среднее изменение y с изменением x на единицу.

2. В уравнении регрессии по отклонениям от тренда $dy = b \cdot dx$ параметр b:

а) показывает, на сколько изменится скорость роста результативного признака с изменением скорости роста факторного признака на единицу своего измерения;

б) показывает, во сколько раз изменится скорость роста результативного признака с изменением скорости роста факторного признака на единицу своего измерения;

в) означает, что случайные отклонения по ряду y в среднем в b раз выше случайных колебаний по ряду x;

г) фиксирует силу связи y с x, т.е. он показывает среднее изменение y с изменением x на единицу.

3. В уравнении регрессии по уровням ряда с включением фактора времени $\tilde{y}_t = a + bx + ct$ параметр b:

а) показывает, на сколько изменится скорость роста результативного признака с изменением скорости роста факторного признака на единицу своего измерения;

б) показывает, во сколько раз изменится скорость роста результативного признака с изменением скорости роста факторного признака на единицу своего измерения;

в) означает, что случайные отклонения по ряду y в среднем в b раз выше случайных колебаний по ряду x;

г) фиксирует силу связи y с x, т.е. показывает среднее изменение y с изменением x на единицу.

4. Прогнозная модель по уравнению регрессии отклонений от тренда имеет вид:

а) $y_p = \tilde{y}_{t=p} + b\left(x_p - \tilde{x}_{t=p}\right)$;

б) $y_p = y_n + a + b\left(x_p - x_n\right)$;

в) $\tilde{y}_t = a \cdot bx \cdot ct$.

г) $\tilde{y}_t = a + bx + ct$.

5. Получено уравнение регрессии по первым разностям $\Delta y = 3{,}56 + 10{,}71\Delta x$. Коэффициент регрессии показывает:

а) что случайные отклонения по ряду y в среднем в 3,56 раза выше случайных колебаний по ряду x;

б) что случайные отклонения по ряду x в среднем в 10,71 раза выше случайных колебаний по ряду y;

в) рост скорости x на 3,56 единиц своего измерения способствует скорости роста y на 10,71 единиц;

г) рост скорости x на 1 единицу своего измерения способствует скорости роста y на 10,71 единиц.

11 Регрессионные модели для панельных данных

Что необходимо знать из главы 11:

1. Понятие и преимущества использования панельных данных.
2. Проблемы использования панельных данных.
3. Виды регрессионных моделей, применяемых к панельным данным.
4. Статистические тесты, призванные решить проблему выбора модели на основе проверки гипотез.

11.1 Понятие и преимущества использования панельных данных

В классическом курсе эконометрики рассматривается два *типа выборочных данных*: пространственные и временные.

Под *пространственной выборкой* понимается набор показателей экономических переменных, полученный в данный момент времени. Например, котировки акций на различных фондовых биржах, набор сведений по разным фирмам (объем производства, себестоимости продукции и т.д.).

Временными данными является набор сведений, характеризующий один и тот же объект, но за различные периоды или моменты времени. Например, ежедневный курс доллара или евро на ММВБ.

Вместе с тем, внедрение новых информационных технологий в научные исследования позволило применять в практике статистического анализа более совершенные методы оценки происходящих в стране и регионах социально-экономических явлений и процессов. Одним из таких методов является метод панельных данных. Панельные или пространственные данные, по определению И.И. Елисеевой [48] - это множество данных, состоящих из наблюдений за однотипными статистическими объектами в течение нескольких временных периодов.

Если периодов времени наблюдений больше числа наблюдаемых объектов, то говорят об объединенном временном ряде. Модели, применяемые к панельным данным, больше ориентированы на определение различий между объектами, нежели на временные аспекты, хотя и содержат информацию относительно развития однотипных явлений во времени. В ходе анализа данных по каждой единице совокупности за несколько лет, повышается устойчивость определяемых моделей и их надежность при решении различных экономических задач.

Работа с панельными данными получает все большее распространение в мире, несмотря на значительные затраты, связанные с проведением панельных опросов. Панельные данные дают исследователям больший простор для маневра в условиях ограниченной информации. Во-первых, за счет большого количества наблюдений увеличивается количество степеней свободы, сокращается мультиколлинеарность переменных и, следовательно, растет эффективность оценок. Во-вторых, наличие данных о межвременной и пространственной вариации переменных позволяет с большей легкостью справиться с проблемой пропущенных переменных, которые коррелируют с объясняющими переменными. В-третьих, возникают огромные возможности для анализа неоднородных данных[1].

Официальные статистические публикации содержат показатели, которые характеризуют одни и те же объекты в заданные периоды или моменты времени, поэтому методы анализа пространственных данных могут найти применение в различных разделах социально- экономической статистики. Данный метод можно назвать логическим продолжением метода «заводо-лет». Данный метод предполагает, что по каждой единице совокупности фиксируются данные за ряд периодов времени (лет). При сводке они рассматриваются как равноправные, независимо от того, к какому периоду времени относятся. Объединенный временно-пространственный массив данных обрабатывается как однородный, с использованием общих формул регрессионного (корреляционного) анализа. [13, с.250]. Применение данного метода обосновывалось стабильными условиями деятельности предприятий в условиях плано-

[1] См. Hsiao C. Analysis of Panel Data / C. Hsiao. – Cambridge: Cambridge University Press, 2004. – 366 с.

вой экономики, в то время как в условиях резких изменений экономической конъюнктуры, масштаба и структуры цен данные становятся несопоставимыми, поэтому необходимо использовать модели, учитывающие эти особенности.

Из определения следует, что панельные данные сочетают в себе как данные пространственного типа, так и данные типа временных рядов. Благодаря специальной структуре панельные данные позволяют строить более гибкие и содержательные модели. Преимущества панельных данных следующие:

1) большее число наблюдений обеспечивает большую эффективность оценивания параметров экономической модели;

2) возникает возможность учитывать и анализировать индивидуальные отличия между экономическими единицами;

3) появляется возможность контроля над неоднородностью объектов;

4) возможность идентифицировать эффекты, недоступные в анализе пространственных данных.

Панельные данные можно представить в виде таблицы, в которой признаки располагаются по столбцам, по строкам – данные о первом объекте за T периодов (строки 1,2,3,…,T), затем о втором объекте (строки T+1,T+2, T+3,…,2T) и т.д. Всего NT строк (таблица 11.1) [52, с.5].

Возможны различные модификации панельных данных. Наибольшее распространение получили сбалансированные и несбалансированные панели.

Если данные присутствуют по всем объектам за все периоды времени, то панель называется *сбалансированной*.

Достаточно часто из-за технических, организационных или иных причин в некоторые периоды времени не удается собрать сведения для всех объектов, включенных в выборку первоначально. Чтобы сохранить репрезентативность, отсутствующие объекты приходится заменять другими. В результате получается *несбалансированная панель*.

Таблица 11.1 - Схема представления панельных данных

Объекты	Признаки				
Объект 1	t =1	X_{11}	Y_{11}	Z_{11}	…
	t =2	X_{12}	Y_{12}	Z_{12}	…
	…	…	…	…	…
	t =T	X_{1T}	Y_{1T}	Z_{2T}	…
Объект 2	t =1	X_{21}	Y_{21}	Z_{21}	…
	t =2	X_{22}	Y_{22}	Z_{22}	…
	…	…	…	…	…
	t =T	X_{2T}	Y_{2T}	Z_{2T}	…
…	…	…	…	…	…
Объект N	t =1	X_{N1}	Y_{N1}	Z_{N1}	…
	t =2	X_{N2}	Y_{N2}	Z_{N2}	…
	…	…	…	…	…
	t =T	X_{NT}	Y_{NT}	Z_{NT}	…

11.2 Проблемы использования панельных данных

Гетерогенное смещение. Привлекательность панельных данных проистекает из теоретической возможности элиминировать в регрессионной модели влияние некоторых специфических трудно измеряемых факторов, например политики.

Если данные генерируются простым контролируемым экспериментом, то могут быть применены стандартные статистические методы. К несчастью, большая часть панельных данных поступает из очень сложных процессов повседневной экономической жизни. Типичное предположение, что *результативный признак* генерируется параметрической функцией распределения вероятностей, может быть нереальным. Игнорирование таких гетерогенных параметров может привести к несостоятельности оценок.

Рассмотрим следующую модель:

$$Y_{it} = a_i + b_i X_{it} + \varepsilon_{it}, \qquad (11.1)$$

где X - единственная экзогенная переменная;

ε_{it} - случайная ошибка подчиняется обычным предположениям теоремы Гаусса - Маркова.

Параметры a_i и b_i могут быть различны для различных индивидуумов, хотя и оставаться постоянными во времени.

Следовательно, будут встречаться различные выборочные распределения, которые могут серьезно смещать регрессию y_{it} на X_{it}, оцененную по всем NT-наблюдениям и игнорирующую индивидуальную неоднородность коэффициентов модели (11.1).

Вышесказанное можно проиллюстрировать следующими примерами:

1) гетерогенный (неодинаковый) для различных индивидуумов свободный член и гомогенный (одинаковый) наклон (рисунок 11.1): $a_i \neq a_j$, $b_i = b_j$ для $\forall i, j$.

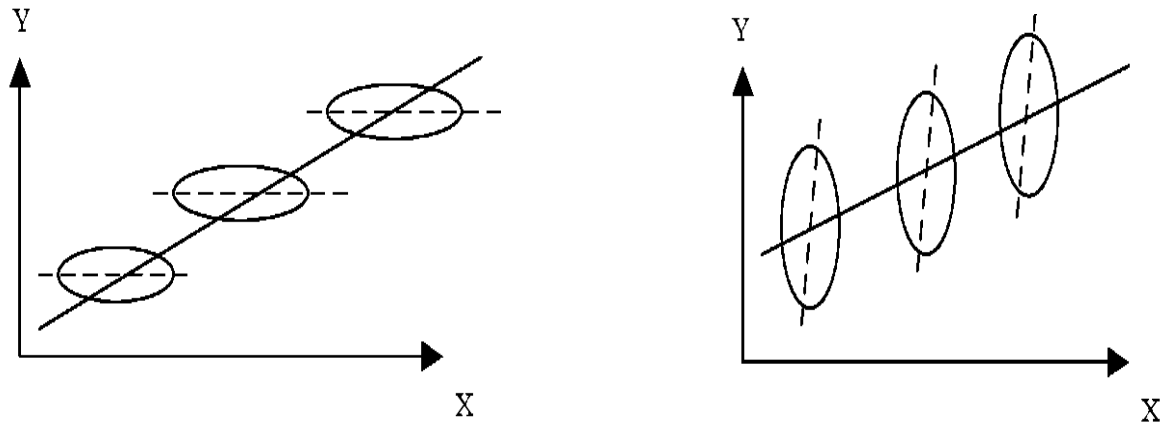


Рисунок 11.1 – Случай гетерогенного для различных индивидуумов свободного члена и гомогенного наклона

Здесь:

 - диаграммы рассеяния для отдельных индивидуумов во времени;

———— - индивидуальные регрессии;

---------- - регрессия по всем NT наблюдениям.

Во всех этих ситуациях сквозная регрессия, игнорирующая гетерогенность константы, является смещенной, причем направление смещения не может быть диагностировано априорно;

2) и свободный член, и наклон гетерогенны: существуют такие i, j, для которых $a_i \neq a_j$, $b_i \neq b_j$ (рисунок 11.2).

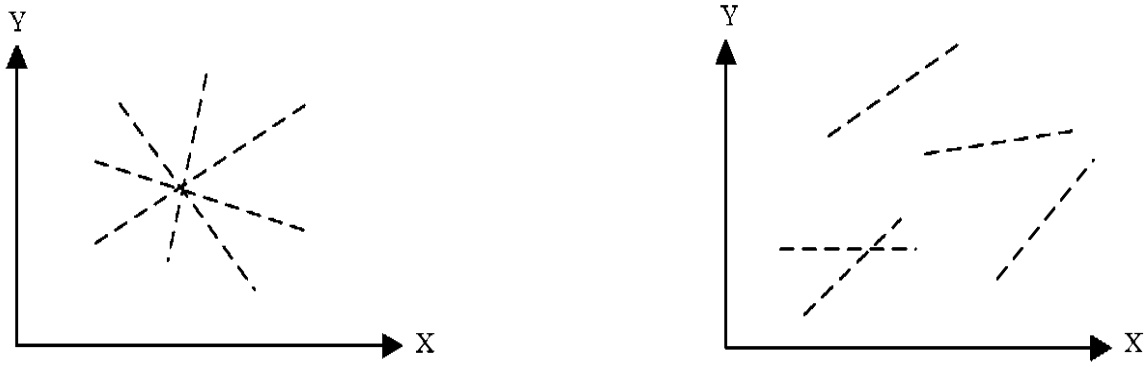


Рисунок 11.2 – Случай гетерогенного для различных индивидуумов свободного члена и наклона

На первом рисунке изображена ситуация, когда сквозная регрессия приводит к бессмысленному результату, так как индивидуальные направления (коэффициенты наклона) существенно различаются. Во втором случае некий смысл сквозной регрессии имеется, но приводит к ложным результатам о криволинейности сквозного соотношения.

Аналогичные примеры можно привести в случае, когда свободный член и наклон изменяются со временем и одинаковы для индивидуумов.

Смещение самоотбора. Другой распространенный источник смещения - неслучайная выборка. Например, известный факт, что в данных РМЭЗ (Российский мониторинг экономического положения и здоровья населения) практически нет наблюдений, относящихся к индивидуумам из высокодоходных групп населения. Ко-

гда такие неполные данные используются в качестве зависимой (объясняемой) переменной, это может повлечь за собой смещение самоотбора. Чтобы это продемонстрировать, рассмотрим пример с пространственными данными. Зависимость ищется в виде:

$$Y_i = X_i'b + \varepsilon_i, \quad i = 1,...,N, \quad E(\varepsilon_i) = 0,\ D(\varepsilon_i) = \sigma_\varepsilon^2 I, \tag{11.2}$$

где Y_i - заработная плата,

X - набор экзогенных переменных, включая образование,
интеллект и т.д.;

I - единичная диагональная матрица.

Причем при $Y_i = b'X_i + \varepsilon_i \le L$ - индивидуумы включаются в выборку; при $Y_i > L$ - исключаются.

Для простоты теперь предположим, что все экзогенные переменные принимают одни и те же значения для всех наблюдений, кроме образования (которое измеряется как продолжительность обучения).

Линия регрессии, построенная по усеченным данным, будет иметь меньший угол наклона, чем ее аналог, который мог бы быть получен по полной выборке. Таким образом, влияние образования оказывается недооцененным. Это происходит оттого, что в данных выборок такого типа появляется корреляция между объясняемой переменной Y_i и случайной ошибкой ε_i, что ведет к недооценке или переоценке влияния экзогенных переменных.

Смещение самоотбора при анализе панельных данных часто является следствием истощения выборки, т.е. постепенного убывания числа объектов наблюдения. Истощение панели - это типичное явление. Панели домохозяйств могут истощаться из-за перемещений, распадов семей, а также из-за отказов участвовать в опросах в дальнейшем. Если выбытие происходит по случайным причинам, смещения самоотбора может и не быть, но если существуют некие скрытые закономерности, то смещение неизбежно. Например, при повышении уровня доходов у домохозяйства мо-

гут пропасть стимулы участвовать в опросе, и тогда в выборке будут оставаться низкодоходные слои населения, что сделает выборку нерепрезентативной.

Перечисленные проблемы могут быть разрешены с помощью некоторых специальных приемов. Это может быть переход или к несбалансированным панелям, где разные индивидуумы наблюдаются в течение различного числа тактов времени, или к панелям с замещением, где выбывшие объекты заменяются новыми, или использованием псевдопанелей, где в качестве объектов наблюдения выступают не отдельные индивидуумы, а группы индивидуумов со схожими (в некотором смысле) характеристиками. Хотя, конечно, это осложняет процесс оценивания.

Для решения проблемы самоотбора при исследовании пространственных выборок используют модель Хекмана. В настоящее время появились разработки, обобщающие эту модель для анализа панельных данных.

К часто встречающимся недостаткам панелей можно отнести также немногочисленность наблюдений, составляющих временные ряды для отдельных индивидуумов [53, с. 274-277].

11.3 Виды регрессионных моделей, применяемых к панельным данным. Статистические тесты, призванные решить проблему выбора модели на основе проверки гипотез

При работе с реальными панельными данными всегда возникает проблема выбора модели. На содержательном уровне разницу между моделями можно интерпретировать следующим образом. Обычная модель предполагает, что у экономических единиц нет индивидуальных различий, и в некоторых простых ситуациях такое предположение оправдано. В модели с фиксированным эффектом считается, что каждая экономическая единица «уникальна» и не может рассматриваться как результат случайного выбора из некоторой генеральной совокупности. Такой подход вполне справедлив, когда речь идет о странах, крупных регионах, отраслях промышленности, больших предприятиях. Если же объекты попали в панель «случай-

но» в результате выборки из большой совокупности, то приемлемой является модель со случайным эффектом.

Для стандартной модели регрессии качество подгонки обычно измеряет коэффициент детерминации или скорректированный коэффициент детерминации. Однако в моделях с панельными данными нецелесообразно использовать коэффициент детерминации для того, чтобы определить какой метод оценивания лучше. Так если одну и ту же модель оценить, например, обычным методом наименьших квадратов и с помощью случайного эффекта, то объединенный коэффициент детерминации в первом случае всегда будет больше соответствующего объединенного коэффициента для второго метода, даже если более адекватным является использование случайного эффекта. Тем не менее, коэффициенты детерминации можно применять для сравнения моделей, отличающихся набором регрессоров и оцениваемых одним и тем же методом.

Помимо теоретических соображений существуют статистические тесты призванные решить проблему выбора модели на основе проверки гипотез.

Модель с фиксированными эффектами имеет вид:

$$y_{it} = \alpha_i + \beta_1 x_{1,it} + \beta_2 x_{2,it} + \beta_3 x_{3,it} + ...\beta_k x_{k,it} + \varepsilon_{it}, \qquad (11.3)$$

где α_i - индивидуальные эффекты;

$\varepsilon_{it} \sim$ iid $(0,\sigma^2)$, $\mathrm{cov}(\varepsilon_{it}, \varepsilon_{js}) = 0$, i ≠ j, t ≠ s; i = $\overline{1,N}$; t = 1,.,T.

Тестом Вальда проверяется гипотеза о равенстве нулю всех индивидуальных эффектов. Если они равны между собой (нулевая гипотеза не отвергается), то модели с фиксированными эффектами следует предпочесть обычную регрессию. Если же нулевая гипотеза отвергается ($F_{набл} > F_{крит}$), то модель с фиксированными эффектами лучше подходит для описания данных, чем модель простой регрессии:

$$H_0 : \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = ... = \alpha_N = 0;\ H_1 : \exists \alpha_j,\ j \in \{1,...,k\} : \alpha_j \neq 0.$$

Для проверки гипотезы используется статистика:

$$F(N-1, NT-N-K) = \frac{(RSS - RSS^{FE})/(N-1)}{RSS^{FE}/(NT-N-K)}, \quad (11.4)$$

где *RSS* – сумма квадратов остатков обычной модели регрессии.

В условиях справедливости нулевой гипотезы статистика (11.4) распределена по закону Фишера-Снедекора с $(N-1)$ и $(NT-N-K)$ степенями свободы.

Если нулевая гипотеза о равенстве всех индивидуальных эффектов отвергается, то необходимо их рассчитать и учитывать при прогнозировании:

$$\alpha_i = \bar{y}_i - b_1\bar{x}_{1i} - b_2\bar{x}_{2i} - b_3\bar{x}_{3i} - b_4\bar{x}_{4i} - ... - b_k\bar{x}_{ki},$$

где α_i - индивидуальные эффекты;

$b_1, b_2, b_3, ..., b_k$ - оценки соответствующих коэффициентов модели;

$\bar{y}_i, \bar{x}_{1i}, \bar{x}_{2i}, \bar{x}_{3i}, ..., \bar{x}_{ki}$ - средние значения соответствующих признаков в модели для *i*-объекта за все периоды времени.

Для проверки значимости оцененного уравнения регрессии с фиксированными эффектами, то есть для проверки нулевой гипотезы $H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$ (альтернативная гипотеза имеет вид $H_1: \exists\ j: \beta_j \neq 0, \quad j = 1,2,3,...$) используется критерий:

$$F(K, NT-N-K) = \frac{(TSS^{FE} - RSS^{FE})/K}{RSS^{FE}/(NT-N-K)}, \quad (11.5)$$

где $TSS^{FE} = \sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it} - \bar{y}_i)^2$ - смещенная оценка общей дисперсии зависимой переменной;

RSS^{FE} – сумма квадратов остатков модели с фиксированными эффектами.

В условиях справедливости нулевой гипотезы статистика (11.5) распределена по закону Фишера-Снедекора с K и ($NT – N – K$) степенями свободы.

Для проверки значимости случайных эффектов используется тест Бреуша-Пагана (тест Множителей Лагранжа).

Выдвигается нулевая гипотеза $H_0 : \sigma_u^2 = 0,\ H_1 : \sigma_u^2 \neq 0$. Используется статистика

$$\frac{\tilde{\sigma}_B^2}{\tilde{\sigma}_w^2} \in F(N-K, NT-N-K),$$

где $\tilde{\sigma}_B^2$ - оценка дисперсии ошибок в between – регрессии;

$\tilde{\sigma}_w^2$ - оценка дисперсии ошибок в within – регрессии.

Для больших выборок в качестве статистики используется множитель Лагранжа:

$$LM = \frac{NT}{2\cdot(T-1)}\left[\frac{\sum_{i=1}^{N}(T\cdot e_i)}{\sum_{i=1}^{N}\sum_{t=1}^{T}e_{it}^2} - 1\right]^2 \in \chi^2(1), \qquad (11.6)$$

где N – количество объектов;

T – количество периодов времени;

e_{it} - регрессионные остатки в обычной модели регрессии.

Если нулевая гипотеза отвергается, то модель со случайными эффектами лучше описывает данные, чем обычная регрессия [52, с.47-48].

В качестве примера практического приложения рассмотренных моделей оценим влияние факторов на показатель среднедушевого объема потребления платных услуг населением на примере Оренбургской области (в сопоставимых ценах) за период 1998-2005 гг. [33, с. 13]. В связи с тем, что существует значительная диффе-

ренциация в потреблении платных услуг в городах и районах области, разные факторы с различной силой влияли на объем потребления услуг во времени, построение модели линейной регрессии по панельным данным проведем отдельно для городов области и отдельно для районов.

Поскольку исследуемые территориально-административные единицы области вряд ли можно считать выборками из большой популяции: каждый район и город уникальны в своем роде, имеют свои собственные особенности, влияние которых учитывается с помощью параметров α_i, логично предположить, что в данной ситуации необходимо постулировать модель с фиксированными эффектами.

Предварительно построим разведочную модель, включив в нее все имеющиеся независимые показатели, при этом согласно методике, предложенной выше, строим модель с фиксированными эффектами для 35-ти районов области и тестируем ее с помощью *F*-критерия. Получаем следующий результат: наблюдаемое значение *F*-критерия – 5,27 с вероятностью p=0,0000.

Табличное значение *F*-критерия с 18 и 228 степенями свободы составляет 1,52, следовательно, нулевая гипотеза о равенстве всех параметров уравнения отвергается, т.е. можно утверждать, что модель с фиксированными эффектами предпочтительнее, чем обычная регрессия. Исключив незначимые по *t*-статистике переменные, получим следующие результаты (таблица 11.2).

Табличные значения *F* – критерия составили 1,03 и 2,37 соответственно. Наблюдаемые значения выше табличных, параметры полученной модели статистически значимы согласно *t*-статистике Стьюдента, следовательно, модель с фиксированными эффектами предпочтительнее обычной регрессии, а гипотеза о равенстве индивидуальных эффектов отвергается.

Результаты построенной модели со случайными эффектами представлены в таблице 11.3. Параметры обоих уравнений практически совпадают, однако, в модели со случайными эффектами добавляется еще одна значимая переменная X_9.

Таблица 11.2 - Результаты моделирования регрессии с фиксированными эффектами (районы области)

Показатели	Значения коэффициентов	Стандартная ошибка	*t* - статистика Стьюдента	*p*-уровень	Нижняя доверительная граница	Верхняя доверительная граница
X_1	2,358	1,098	2,15	0,033	0,1958	4,521
X_4	-4,917	1,8348	-2,68	0,008	-8,531	-1,304
X_5	0,271	0,0348	7,88	0,000	0,204	0,339
X_8	0,033	0,0148	2,25	0,025	0,004	0,062
Константа	-253,841	475,0858	-0,53	0,594	-1189,692	682,009
F(34, 241) = 11.10 Prob > F = 0.0000 F(4,241) = 43,06 Prob > F = 0,0000						

Таблица 11.3 - Результаты моделирования регрессии со случайными эффектами (районы области)

Показатели	Значения коэффициентов	Стандартная ошибка	*t*	*p*	Нижняя доверительная граница	Верхняя доверительная граница
X_1	2,507	1,025	2,45	0,014	0,499	4,516
X_4	-4,096	1,659	-2,47	0,014	-7,348	-0,844
X_5	0,284	0,033	8,61	0,000	0,219	0,348
X_8	0,039	0,014	2,78	0,005	0,012	0,066
X_9	4,349	2,248	1,96	0,050	-0,056	8,756
Константа	-1384,233	627,213	-2,21	0,027	-2613,548	-154,919

В результате проведения теста множителей Лагранжа получаем, что $\chi^2(1)$, = 234,71 (*p*-уровень равен 0,000), т.е. нулевая гипотеза отвергается и модель со случайным эффектом предпочтительнее обычной регрессии.

Оценки модели со случайными эффектами являются более эффективными, чем оценки модели с фиксированными эффектами. Однако для «законного» использования модели со случайными эффектами мы должны быть уверены в некоррелированности индивидуальных эффектов и регрессоров, ибо если это не так, то мы получим вообще несостоятельные оценки. Поэтому для сравнения моделей и выбора наиболее подходящей спецификации используется специальная проверка – тест Хаусмана.

Нулевая гипотеза заключается в предположении об отсутствии корреляции между случайными эффектами и регрессорами (если это требование не выполняется, оценки модели со случайными эффектами не будут состоятельными):

$H_0 : corr(u_i . X_{it}) = 0$ - u_i могут быть рассмотрены как случайные эффекты;

$H_1 : corr(u_i . X_{it}) \neq 0$ - u_i следует рассматривать как фиксированные эффекты.

Этот тест построен на разности оценок модели с фиксированными эффектами (они состоятельны как в случае основной, так и альтернативной гипотезы) и оценок модели со случайными эффектами (они состоятельны только при основной гипотезе).

Для проверки нулевой гипотезы строится статистика:

$$W = [b_{FE} - b_{RE}]^T [\mathrm{cov}(b_{FE}) - \mathrm{cov}(b_{RE})]^{-1} \left| b_{FE} - b_{RE} \right| \in \chi^2(k), \qquad (11.7)$$

где b_{FE} – оценки параметров модели с фиксированными эффектами;

b_{RE} – оценки параметров модели со случайными эффектами;

cov (b_{FE}) и *cov* (b_{RE}) – оценки ковариационных матриц для параметров моделей с фиксированными и случайными эффектами [52, с. 50-51].

Если нулевая гипотеза не отвергается, то можно выбрать модель со случайными эффектами, оценки которой будут эффективными. В противном случае следует выбрать модель с фиксированными эффектами.

В нашем примере получаем $\chi^2_{крит}$ = 1,145; $W_{набл}$ = 7,79; p = 0,1683. Следовательно, делаем выбор в пользу модели со случайными эффектами:

$$\tilde{y}_{it} = -1384{,}233 + 2{,}5074x_1 - 4{,}0960x_4 + 0{,}2838x_5 + 0{,}0389x_8 + 4{,}350x_9 + \varepsilon_{it}\,.$$

На основе полученной модели можно сделать следующие выводы. Величина и динамика уровня потребления платных услуг населением районов Оренбургской области в значительной мере связана с вариацией и динамикой таких показателей как среднесписочная численность занятых на 1000 человек в трудоспособном возрасте (X_1), среднегодовая численность работников, занятых в сельскохозяйственном производстве на 1000 человек в трудоспособном возрасте (X_4), оборот розничной торговли на душу населения в сопоставимых ценах, р. (X_5), инвестиции в основной капитал на душу населения, р. (X_8), обеспеченность населения врачебными амбулаторно-поликлиническими учреждениями (число посещений в смену на 10000 населения) (X_9).

Наблюдается положительная связь со всеми признаками, кроме X_4. Увеличение данного фактора на единицу приводит к снижению среднедушевого объема потребления услуг на 4 единицы. Как уже отмечалось, крайне низкий уровень заработной платы работающих в сельскохозяйственном производстве, негативно отражается на объеме потребления платных услуг населением районов. С увеличением показателей X_1, X_5, X_8, X_9 на единицу, среднедушевой объем потребления услуг вырастет соответственно на 2,51; 0,28; 0,04 и 4,35 единиц соответственно. Основные факторы, влияющие на результативный признак – это общие социально-экономические индикаторы развития территориального образования, То есть влияние факторов, отражающих уровень развития сферы услуг как таковой, незначимо, следовательно, и сама сфера услуг в районах области недостаточно развита.

Аналогичным образом проведем моделирование среднедушевого объема потребления платных услуг в 12-ти городах области. В результате оценки приходим к следующей модели с фиксированными эффектами (таблице 11.4):

Таблица 11.4 - Результаты моделирования регрессии с фиксированными эффектами (города области)

Показатели	Значения коэффициентов	Стандартная ошибка	*t*-статистика	*p*-уровень	Нижняя доверительная граница	Верхняя доверительная граница
X_3	510,515	123,893	4,12	0,000	263,912	757,119
X_5	0,112	0,022	5,17	0,000	0,069	0,156
X_{11}	66,886	17,188	3,89	0,000	32,675	101,097
X_{12}	10,510	4,5368	2,32	0,002	1,480	19,539
X_{13}	36,202	7,162	5,06	0,000	21,947	50,456
Константа	-20077,76	2630,134	-7,63	0,000	-25312,91	-14842,61
F(11, 79) = 8,20 Prob > F = 0,0000 F(5,79) = 83,16 Prob > F = 0,0000						

Параметры уравнения значимы. Табличные значения *F*- критерия составили 1,75 и 2,21 соответственно. Наблюдаемые значения выше табличных, следовательно, нулевые гипотезы о не значимости индивидуальных эффектов и не значимости уравнения регрессии с фиксированными переменными, отвергаются.

Аналогично была получена модель со случайными эффектами, значение $\chi^2_{\alpha}(1)$, = 11,76 (*p*-уровень равен 0,0006), т.е. нулевая гипотеза отвергается и модель со случайным эффектом предпочтительнее обычной регрессии.

Тест Хаусмана показал, что модель с фиксированными эффектами предпочтительнее модели со случайными эффектами (χ^2_{α}(k), = 108,43, *p*=0,0000; $\chi^2_{крит} = 1{,}145$). Оценки индивидуальных эффектов найдем по формуле:

$$\alpha_i = \bar{y}_i - \sum_{k=1}^{5} \beta_k \bar{x}_{ki}\,, \qquad (11.8)$$

где $\alpha_i, i = \overline{1,12}$ - индивидуальные эффекты;

$\beta_k, k = \overline{1,5}$ - оценки соответствующих коэффициентов модели;

$\bar{y}_i, \bar{x}_{ki}, i = \overline{1,12}$ - средние значения соответствующих признаков в модели для i объекта за все периоды времени.

Оценки индивидуальных эффектов приведены в таблице 11.5.

Таблица 11.5 - Оценки индивидуальных эффектов

Город	α	Город	α
г. Абдулино	-20673,8	г. Новотроицк	-21746,4
г. Бугуруслан	-18523	г. Оренбург	-21850
г. Бузулук	-20350,4	г. Орск	-20778,7
г. Гай	-22525,9	г. Соль-Илецк	-17166
г. Кувандык	-18831,2	г. Сорочинск	-19671,8
г. Медногорск	-19951	г. Ясный	-18864,9

В результате получена следующая модель зависимости среднедушевого объема потребления платных услуг (в сопоставимых ценах) от определяющих факторов:

$$\tilde{y}_{it} = \alpha_i + 510{,}515x_{3t} + 0{,}112x_{5t} + 66{,}886x_{11t} + 10{,}510x_{12t} + 36{,}20184x_{13t} + \varepsilon_{it}\,.$$

При увеличении площади жилищ, приходящихся в среднем на одного жителя городов на единицу, среднедушевой объем потребления платных услуг вырастает на 510,5 ед.; при увеличении объема розничной торговли на душу населения на единицу потребление платных услуг вырастет в среднем на 0,11 ед.; при росте показателя «благоустройство жилищного фонда водопроводом» на единицу объем потребляе-

мых населением услуг вырастет на 66,9 ед.; с увеличением обеспеченности населения квартирными телефонными аппаратами на единицу объем услуг на душу населения вырастет на 10,5 ед. и с ростом показателя обеспеченности населения личными автомобилями на единицу среднедушевой объем потребляемых платных услуг возрастет на 36,2 ед.

Таким образом, на величину и динамику среднедушевого объема потребления платных услуг в городах Оренбургской области влияют, в основном, факторы, отражающие развитие сферы платных услуг в целом.

11.4 Вопросы для самоконтроля

1. Какие данные называют панельными? В чем преимущества их использования?
2. Какие панели называют незакрытыми?
3. Назовите достоинства и недостатки моделей со случайными и фиксированными эффектами.
4. Как проводится проверка гипотезы о значимости групповых эффектов?
5. Как проверяется гипотеза о значимости случайных эффектов?

11.5 Тесты

1. Какой тест используют для проверки значимости случайных эффектов:

а) Чоу;

б) Вайта;

в) Фишера – Снедекора;

г) Бреуша-Пагана.

2. Какой тест используется для сравнения моделей с фиксированными и случайными эффектами:

а) Бреуша-Пагана;

б) Вайта;

в) Фишера – Снедекора;

г) Хаусмана.

3. Какому закону распределения подчиняется статистика в тесте Хаусмана:

а) нормальному;

б) Хи-квадрат;

в) Фишера;

г) Стьюдента.

4. Расчеты, выполненные по панельным данным для 5 объектов и 3 периодов времени, дали следующие результаты. Сумма квадратов остатков для множественной регрессии с двумя независимыми переменными составила 104,42. Сумма квадратов их средних по группам 193,41. Величина тестовой статистики LM, теста Бреуша – Пагана, равна:

а) 17,85;

б) 2,72;

в) 0,85;

г) 7,22.

5. МНК – оценки для модели со случайными эффектами:

а) несмещенные, состоятельные, эффективные;

б) смещенные, состоятельные, эффективные;

в) несмещенные, несостоятельные, эффективные

г) несмещенные, состоятельные, неэффективные

Список использованных источников

1 Пасхавер, И.С. Общая теория статистики : для программированного обучения: учеб. пособие / И.С. Пасхавер, А.Л. Яблочник; под ред. проф. М.М. Юзбашева. – М. : Финансы и статистика, 1983. – 432 с., ил.

2 Ефимова, М.Р. Общая теория статистики: учебник / М.Р. Ефимова, Е.В. Петрова, В.Н. Румянцев. – 2 - е изд., испр. и доп. – М. : ИНФРА-М, 2007. – 416 с. – (Высшее образование). - ISBN 5-16-002179-5.

3 Салин, В. Н. Курс теории статистики для подготовки специалистов финансово-экономического профиля : учебник для студентов по специальностям "Финансы и кредит", "Бухгалтерский учет, анализ и аудит", "Мировая экономика", "Налоги и налогообложение" / В. Н. Салин, Э. Ю. Чурилова. – М. : Финансы и статистика, 2007 . – 480 с. - ISBN 978-5-279-03063-7.

4 Джонстнон, Дж. Эконометрические методы : пер. с англ. / Дж. Джонстон. – М. : Статистика, 1980. – 444 с.

5 Снедекор, Дж. У. Статистические методы в применении к исследованиям в сельском хозяйстве и биологии / Дж.У. Снедекор. – М. : Сельхозиздат, 1961. – 503 с.

6 Доугерти, К. Введение в эконометрику : пер. с англ. / К. Доугерти. – М. : ИНФРА-М, 1999. – 402 с. - ISBN 5-86225-458-7.

7 Новак, Э. Введение в методы эконометрики: сборник задач : пер. с польск. / Э. Новак; под ред. И.И. Елисеевой. – М. : Финансы и статистика, 2004. – 248 с. - ISBN 5-279-02927-0.

8 Гладилин, А.В. Эконометрика : учеб. пособие / А.В. Гладилин, А.Н. Герасимов, Е.И. Громов. – М. : КНОРУС, 2006. – 232 с. - ISBN 5-85971-118-2.

9 Дружинин, Н.К. Математическая статистика в экономике / Н.К. Дружинин. – М. : Статистика, 1971. – 264 с.

10 Афанасьев, В.Н. Эконометрика : учебник / В.Н Афанасьев, М.М. Юзбашев, Т.И. Гуляева; под общ. ред. М.М. Юзбашева. – М.: Финансы и статистика, 2005. - 256 с. - ISBN 5-279-02738-3.

11 Миллс, Ф. Статистические методы : пер. с англ. / Ф. Миллс. – М. : Госстатиздат, 1958. – 589 с.

12 Айвазян, С.А. Прикладная статистика. Основы эконометрики : учебник для вузов : в 2 т.– Т. 1. Теория вероятностей и прикладная статистика / С.А. Айвазян, В.С. Мхитарян. - 2-е изд., испр. – М. : ЮНИТИ-ДАНА, 2001. – 656 с. - ISBN 5-238-00304-8.

13 Крастинь, О.П. Разработка и интерпретация моделей корреляционных связей в экономике / О.П. Крастинь. – Рига : Зинате, 1983. – 302 с.

14 Четыркин, Е.М. Вероятность и статистика / Е.М. Четыркин, И.Л. Калихман. – М. : Финансы и статистика, 1982. – 319 с.

15 Кейн, Э. Экономическая статистика и эконометрия. Введение в количественный экономический анализ. Вып. 2 : пер. с англ. / Э. Кейн – М. : Статистика, 1977. – 232 с. с ил.

16 Кремер, Н.Ш. Эконометрика : учебник для вузов / Н.Ш. Кремер, Б.А. Путко; под ред. проф. Н.Ш. Кремера. – М. : ЮНИТИ-ДАНА, 2007. – 311 с. – ISBN 5-238-00333-1.

17 Винн, Р. Введение в прикладной эконометрический анализ : пер. с англ. / Р. Винн, К. Холден – М. : Финансы и статистика, 1981. – 294 с.

18 Эренберг, А. Анализ и интепретация статистических данных : пер. с англ. / А. Эренберг – М. : Финансы и статистика, 1981. – 406 с. : ил. – (Библиотечка иностранных книг для экономистов и статистиков).

19 Афифи, А. Статистический анализ : подход с использованием ЭВМ : пер. с англ. / А. Афифи, С. Эйзен. – М. : Мир, 1982. – 488 с.

20 Многомерный статистический анализ в экономике : учеб. пособие для вузов / Л.А. Сошникова, В.Н. Тамашевич, Г. Уебе, М. Шефер ; под ред. проф. В.Н. Тамашевича. – М. : ЮНИТИ-ДАНА, 1999. – 598 с. - ISBN 5-238-00099-5.

21 Ферстер, Э. Методы корреляционного и регрессионного анализа : Руководство для экономистов : пер. с нем. / Э. Ферстер, Б. Ренц. – М. : Финансы и статистика, 1983. – 302 с.

22 Магнус, Я.Р. Эконометрика. Начальный курс : учебник / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий. – 5-е изд., испр. – М. : Дело, 2001. – 400 с. - ISBN 5-7749-0055-X.

23 Hoerl, A.E. Application of ridge analysis to regression problems / A.E. Hoerl // Chemical Engineering Progress, vol. 58. № 3. - March 1962. - P. 54-59.

24 Farrar D. E. Multicollinearity in Regression Analysis : The Problem Revisited / D.E. Farrar, R.R. Glauber // The Review of Economics and Statistics, vol. 49. - № 1. - February, 1967. - P. 92-107.

25 Лопатников, Л. И.Экономико-математический словарь : словарь современной экономической науки / Л.И. Лопатников. - 5-е изд., перераб. и доп. - М. : Дело, 2003. - 520 с. - ISBN 5-7749-0275-7.

26 Справочник по прикладной статистике: в 2-х т. / под ред. Э. Ллойда, У. Ледермана, С.А. Айвазяна, Ю.Н. Тюрина : пер. с англ. – М. : Финансы и статистика, 1990. – Т. 2. – 525 с. - ISBN 5-279-00244-5.

27 Owen, D.B. Handbook of Statistical Tables / D.B. Owen. - Pergamon Press, and Addison-Wesley, 1962. – 580 p.

28 Glejser, H. A New Test for Heteroskedasticity / H. Glejser // Journal of the American Statistical Association, vol. 64. - 1969. - P. 316-323.

29 Goldfeld, S.M. Some Tests for Homoscedasticity / S.M. Goldfeld, R.E. Quandt // Journal of the American Statistical Association, 60. - 1965. - P. 539–547.

30 White H.F. Heteroscedasticity – Consistent Covariance Matrix Estimator and a Direct Test for Heteroscedasticity / H.F. White. – Econometrica, vol. 48. - 1980. - P. 817-838.

31 Вербик, Марно. Путеводитель по современной эконометрике : пер. с англ. / М. Вербик; научн. ред. и предисл. С. А. Айвазяна — М : Научная книга, 2008. – 616 с. – ISBN 978-5-913-035-4.

32 Durbin J. Testing for Serial Correlation in Least-Squares Regression / J. Durbin, G.S. Watson. - Biometrica, vol. 37. - 1950. - P. 409-428; and vol. 38. - 1951. - P. 159-178.

33 Афанасьев, В.Н. Статистические методы в исследовании потребления платных услуг домашними хозяйствами : учеб. пособие для вузов / В.Н. Афанасьев, Т.В. Леушина. – Оренбург : ОГУ, 2011. – 156 с. – ISBN 978-5-7410-1138-6.

34 Айвазян, С.А. Методы эконометрики : учебник / С.А. Айвазян. – М. : Магистр : ИНФРА-М, 2010. – 512 с. – ISBN 978-5-9776-0153-5 (в пер.).

35 Hood, W.C. Studies in Econometric Method / W.C. Hood, T.C. Koopmans. - Cowles Commission Monograph, 1953. - № 14.

36 Бард, Й. Нелинейное оценивание параметров : пер. с англ. / Й. Бард; под ред. и с предисл. В.Г. Горского – М. : Статистика, 1979. – 349 с.

37 Дрейпер, Н. Прикладной регрессионный анализ : в 2-х кн. / Н. Дрейпер, Г. Смит : пер. с англ. – 2- е изд., перераб. и доп.. – М : Финансы и статистика, 1986. – Кн. 1. - 366 с. : ил. - (Математико-статистические методы за рубежом).

38 Проблемы определения биовозраста : сравнение эффективности методов линейной и нелинейной регрессии / Т.М. Смирнова [и др.] // Профилактика старения, 1999. – Выпуск 2. – Режим доступа : http://medi.ru/.

39 Сахарова, Ю.В. Самоорганизация социальных систем : основания и интерпретационые возможности использования логарифмических моделей / Ю.В. Сахарова. – Режим доступа : http://www.teoria-practica.ru/-7-2012/sociology/sakharova.pdf.

40 Zarembka, P. Functional Form in the Demand for Money / P. Zarembka // Journal of the American Statistical Association, vol. 63. – 1968. - P. 502-511.

41 Box, G.E.P. An analysis of transformations / G.E.P. Box, D.R. Cox // J. Roy. Statist. Soc., vol. 26. – 1964. - P. 211-327.

42 Мхитарян, В.С. Эконометрика : учебно-методический комплекс / В.С. Мхитарян, М.Ю. Архипова, В.П. Сиротин. – М. : Изд. центр ЕАОИ, 2008. – 144 с.

43 Производственные функции в управлении проектами. Научные и учебно - методические разработки Института инноватики. – Режим доступа : http://www.ii.spb.ru.

44 Тихомиров, Н. П. Эконометрика : учеб. для вузов / Н. П. Тихомиров, Е. Ю. Дорохина; Рос. экон. акад. им. Г. В. Плеханова. - М. : Экзамен, 2003. - 512 с. - ISBN 5-94692-438-9.

45 Бородич, С.А. Эконометрика : учебное пособие / С.А. Бородич. – 3-е изд., - Минск : Новое знание, 2006. – 408 с. – ISBN 985-475-206-2.

46 Дуброва, Т.А. Статистические методы прогнозирования: учеб. пособие для вузов / Т.А. Дуброва - М.: ЮНИТИ-ДАНА, 2003. – 206 с. - ISBN 5-238-00497-4.

47 Афанасьев В.Н. Моделирование и прогнозирование временных рядов: учеб.-метод. пособие для вузов / В.Н. Афанасьев, Т.В. Лебедева. – М.: Финансы и статистика, 2009. – 292 с. - ISBN: 978-5-279-03402-4.

48 Эконометрика : учебник для студентов вузов, обучающихся по специальности 061700 «Статистика» / под ред. И. И. Елисеевой . - 2-е изд., перераб. и доп. - М. : Финансы и статистика, 2008. - 576 с. - ISBN 978-5-279-02786-6.

49 Бабешко Л.О. Основы эконометрического моделирования: учеб. пособие / Л.О. Бабешко - М. : КомКнига, 2006. – 432 с. - ISBN 978-5-484-00757-8.

50 Афанасьев В.Н. Анализ временных рядов и прогнозирование: учебник / В.Н. Афанасьев, М.М. Юзбашев. – 2-е изд., перераб. и доп. - М. : Финансы и статистика; ИНФРА-М, 2010. – 320 с. - ISBN 978-5-279-03400-0.

51 Статистика **:** учебник / И.И. Елисеева [и др.]; под ред. проф. И.И. Елисеевой. – М. : КНОРУС, 2006.– С.552. – ISBN 5-85971-294-4.

52 Балаш, В.А. Модели линейной регрессии для панельных данных: учеб. пособие / В.А. Балаш., О.С. Балаш – М., 2002. – 65 с.

53 Ратникова, Т.А. Введение в эконометрический анализ панельных данных / Т.А. Ратникова // Экономический журнал ВШЭ. -2006. - № 2. – С. 267-316.

Приложение А
(справочное)

Квантили распределения $\chi^2(\nu)$

Таблица А.1

ν	α									
	0,005	0,010	0,025	0,050	0,100	0,900	0,950	0,975	0,990	0,995
1	0,000039	0,00016	0,00098	0,0039	0,0158	2,71	3,84	5,02	6,63	7,88
2	0,0100	0,0201	0,0506	0,1026	0,2107	4,61	5,99	7,38	9,21	10,60
3	0,0717	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,34	12,84
4	0,207	0,297	0,484	0,711	1,064	7,78	9,49	11,14	13,28	14,86
5	0,412	0,554	0,831	1,15	1,61	9,24	11,07	12,83	15,09	16,75
6	0,676	0,872	1,24	1,64	2,20	10,64	12,59	14,45	16,81	18,55
7	0,989	1,24	1,69	2,17	2,83	12,02	14,07	16,01	18,48	20,28
8	1,34	1,65	2,18	2,73	3,49	13,36	15,51	17,53	20,09	21,96
9	1,73	2,09	2,70	3,33	4,17	14,68	16,92	19,02	21,67	23,59
10	2,16	2,56	3,25	3,94	4,87	15,99	18,31	20,48	23,21	25,19
11	0,60	3,05	3,82	4,57	5,58	17,28	19,68	21,92	24,73	26,76
12	3,07	3,57	4,40	5,23	6,30	18,55	21,03	23,34	26,22	28,30
13	3,57	4,11	5,01	5,89	7,04	19,81	22,36	24,74	27,69	29,82
14	4,07	4,66	5,63	6,57	7,79	21,06	23,68	26,12	29,14	31,32
15	4,60	5,23	6,26	7,26	8,55	22,31	25,00	27,49	30,58	32,80
16	5,14	5,81	6,91	7,96	9,31	23,54	26,30	28,85	32,00	34,27
18	6,26	7,01	8,23	9,39	10,86	25,99	2887	31,53	34,81	37,16
20	7,43	8,26	9,59	10,85	12,44	28,41	31,41	34,17	37,57	40,00
24	9,89	10,86	12,40	13,85	15,66	33,20	36,42	39,36	42,98	45,56
30	13,79	14,95	16,79	18,49	20,60	40,26	43,77	46,98	50,89	63,67
40	20,71	22,16	24,43	26,51	29,05	51,81	55,76	59,34	63,69	66,77
60	35,53	37,48	40,48	43,19	46,46	74,40	79,08	83,30	88,38	91,95
80	51,17	53,54	57,15	60,39	64,28	96,58	101,88	106,6	112,3	116,3
100	67,33	70,06	74,22	77,93	82,36	118,50	124,34	129,6	135,8	140,2
120	83,85	86,92	91,58	95,70	100,62	140,2	146,57	152,2	159,0	163,6

Приложение Б
(справочное)

Критические значения коэффициента корреляции для уровней значимости 0,05; 0,01

Таблица Б.1

d. f.	α=0,05	α=0,01
1	0,996917	0,9998766
2	0,95000	0,990000
3	0,8783	0,95873
4	0,8114	0,91720
5	0,7545	0,8745
6	0,7067	0,8343
7	0,6664	0,7977
8	0,6319	0,7646
9	0,6021	0,7348
10	0,5760	0,7079
11	0,5529	0,6835
12	0,5324	0,6614
13	0,5139	0,6411
14	0,4973	0,6226
15	0,4821	0,6055
16	0,4683	0,5897
17	0,4555	0,5751
18	0,4438	0,5614
19	0,4329	0,5487
20	0,4227	0,5368
25	0,3809	0,4869
30	0,3494	0,4487
35	0,3246	0,4182
40	0,3044	0,3932
45	0,2875	0,3721
50	0,2732	0,3541
60	0,2500	0,3248
70	0,2919	0,3017
80	0,2172	0,2830
90	0,2050	0,2673
100	0,1946	0,2540

Для простой корреляции d. f. на 2 меньше, чем число пар вариантов; в случае частной корреляции необходимо также вычесть число исключаемых переменных.

Приложение В
(справочное)

Значения F-критерия Фишера на уровне значимости $\alpha = 0,05$

Таблица В.1

k2	k1									
	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,5	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	307	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2,74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	1,76	1,48
50	4,03	3,18	2,79	2,56	2,4	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,89	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1,88	1,65	1,31
90	3,95	310	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3,07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Приложение Г
(справочное)

Критические значения t-критерия Стьюдента на уровне значимости 0,10; 0,05; 0,01

Таблица Г.1

Число степеней свободы d.f.	P			d. f.	P		
	0,10	0,05	0,01		0,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,6041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	∞	1,6449	1,9600	2,5758

Приложение Д
(справочное)

z - преобразование. Значение величины *z* для значений *R*

Таблица Д.1

r	0	1	2	3	4	5	6	7	8	9
0,0	0,0000	0,0100	0,0200	0,0300	0,0400	0,0501	0,0601	0,0701	0,0802	0,0902
0,1	0,1003	0,1105	0,1206	0,1308	0,1409	0,1511	0,1614	0,1717	0,1820	0,1923
0,2	0,2027	0,2132	0,2237	0,2342	0,2448	0,2554	0,2661	0,2769	0,2877	0,2986
0,3	0,3095	0,3206	0,3317	0,3428	0,3541	0,3654	0,3769	0,3884	0,4001	0,4118
0,4	0,4236	0,4356	0,4477	0,4599	0,4722	0,4847	0,4973	0,5101	0,5230	0,5361
0,5	0,5493	0,5627	0,5763	0,5901	0,6042	0,6184	0,6328	0,6475	0,6625	0,677
0,6	0,6931	0,7089	0,7250	0,7414	0,7582	0,753	0,7928	0,8107	0,8291	0,8480
0,7	0,8673	0,8872	0,9076	0,9287	0,9505	0,9730	0,9962	1,0203	1,0454	1,0714
0,8	1,0986	1,1270	1,1568	1,1881	1,2212	1,2562	1,2933	1,3331	1,3758	1,4219
0,9	1,4722	1,5275	1,5890	1,6584	1,7380	1,8318	1,9459	2,0923	2,2976	2,6467

Приложение Е
(справочное)

Исходные данные для многомерного анализа

Таблица Е.1

Субъекты РФ	Y	X1	X2	X3	X4	X5	X6	X7	X8
1	2	3	4	5	6	7	8	9	10
Белгородская область	756,7	247,6	41	8,4	132	237	77	69,5	106
Брянская область	820,5	276	52	13,6	35	117	78	61	110
Владимирская область	925,8	302,5	46	18,3	35	156	129	74,7	102
Воронежская область	548,8	190,3	85	19,1	77	460	134	68,1	111
Ивановская область	896,6	192,8	43	20,5	37	168	102	66,1	101
Калужская область	757,6	253	36	11,4	12	129	92	68,1	115
Костромская область	799,9	288,8	23	17,3	54	1800	47	62,1	107
Курская область	578,6	172,3	48	10,8	41	248	37	62,7	109
Липецкая область	682,6	247,6	28	10	368	175	87	80,6	121
Московская область	659,3	279	125	10,3	205	2482	1309	79,5	102
Орловская область	874,6	250,6	37	14,9	23	84	50	68,3	99
Рязанская область	702	176,1	48	15,3	134	186	89	70,9	110
Смоленская область	765	167,6	40	14,7	48	169	73	61,2	103
Тамбовская область	706	294	43	10,9	46	100	13	64	100
Тверская область	834,5	188,3	47	13,1	60	1494	99	58,7	92
Тульская область	726,5	301,8	47	11	167	283	196	79,9	95
Ярославская область	907,7	159	53	12,8	81	278	240	75,5	104
г. Москва	693,7	128,8	105	10	63	1496	909	99,8	106
Республика Карелия	1078,1	198,8	36	15,2	108	203	190	67,5	97
Республика Коми	1037	210,6	56	14,7	595	506	117	72,8	98
Архангельская область	1049,1	176,9	48	14,3	545	662	416	57,4	106
Вологодская область	887,1	281,9	52	17	474	602	152	63,2	106
Калининградская область	799,2	289,8	55	12,6	29	112	88	90,9	107
Ленинградская область	547,4	290,1	48	12,8	226	6623	291	70,8	91
Мурманская область	890,4	174,5	46	13	288	1502	339	97,6	101
Новгородская область	882,3	242,3	20	14,9	46	108	97	55,5	104
Псковская область	677,8	289	35	15,8	22	335	52	57,9	105
г. Санкт-Петербург	860,1	117,2	70	8,7	57	1017	1346	98,7	109
Республика Адыгея	694	264	19	16,1	4	119	29	62	121
Республика Калмыкия	679,8	214	21	37,3	3	371	29	39,7	107
Краснодарский край	575,8	238,4	174	15,6	139	3142	863	74,3	120
Астраханская область	744,3	146,7	44	14,2	125	772	69	68,4	108
Волгоградская область	709,1	198,1	106	13,4	201	731	186	69,9	99
Ростовская область	789,7	259,4	169	15,1	176	2330	270	69,7	106
Республика Дагестан	766,4	260,8	182	9,2	18	3030	77	44,4	115
Кабардино-Балкарская Республика	400	227,1	51	15,8	3	443	33	78	113

Продолжение таблицы Е.1

1	2	3	4	5	6	7	8	9	10
Карачаево-Черкесская Республика	446,6	303,7	21	18,8	20	43	51	58,3	118
Республика Северная Осетия - Алания	646	141,3	37	10,4	6	264	82	96,5	120
Ставропольский край	527,4	253,4	94	18,8	66	3372	144	76,3	104
Республика Башкортостан	894,5	236,8	183	12	388	745	341	62,1	117
Республика Марий Эл	864,5	298,8	40	24,6	33	91	60	65,3	103
Республика Мордовия	725,1	193,2	25	19	34	70	46	58,4	101
Республика Татарстан	846,3	226,4	127	8	263	639	490	79,6	110
Удмуртская Республика	957,7	171	77	14,6	101	301	105	68,7	105
Чувашская Республика	937,9	210,2	63	18,7	31	122	86	57,9	99
Пермский край	932,6	184,4	123	13,8	325	2472	313	73	110
Кировская область	805,6	204,7	67	14,1	102	237	205	56,9	102
Нижегородская область	855,2	210,2	140	12,5	156	1112	472	74,3	110
Оренбургская область	858,4	191,5	82	14,2	617	1653	122	73,5	104
Пензенская область	760,4	254,8	45	15,5	22	235	111	63,8	100
Самарская область	944,7	209,6	99	15,1	308	863	397	84,5	112
Саратовская область	750,2	192,2	89	16,9	95	532	24	67,1	88
Ульяновская область	895,9	276,4	60	17	39	182	111	70,1	101
Курганская область	866,3	336,6	58	16,8	55	62	50	49,5	100
Свердловская область	728,7	218,8	202	10,1	1169	990	763	78,9	103
Тюменская область	849,9	185,4	131	12,5	3132	1825	202	81,4	118
Челябинская область	870,8	232,9	144	10,4	749	868	845	79,2	114
Республика Алтай	861,8	230	12	17,9	6	8	0,3	28,3	109
Республика Бурятия	667,4	242,1	49	19,8	95	499	41	48,8	113
Республика Тыва	585,1	219,5	28	30	23	19	9	36,9	106
Республика Хакасия	803,5	262,4	25	18,5	96	119	38	64,8	99
Алтайский край	1036,6	205,4	115	24,3	207	340	15	64,1	104
Забайкальский край	741,6	178,2	61	19,3	138	234	78	48,2	111
Красноярский край	814,2	186,4	101	18,4	2491	2296	444	70,2	110
Иркутская область	908,8	200,5	133	18,4	597	1008	594	67,4	109
Кемеровская область	796,9	209,5	131	10,9	1411	1751	700	70,9	113
Новосибирская область	718,7	171,5	109	16,7	228	676	107	73,7	114
Омская область	839,9	185,4	87	14,3	230	258	177	63,7	116
Томская область	679,5	144,4	44	17,4	345	531	14	69,9	118
Республика Саха (Якутия)	1023,5	175	43	19,1	161	164	86	52,3	109
Камчатский край	863,2	202,6	15	19,8	37	165	46	93	100
Приморский край	796	189,8	105	16,3	233	708	371	74.0	107
Хабаровский край	736,5	163,8	70	15,8	117	369	191	80,5	99
Амурская область	779,2	159,1	31	24,3	119	87	82	62	103
Магаданская область	809,8	179,9	6	13,8	25	77	27	91,4	97
Сахалинская область	924,9	203,6	27	10,9	100	263	46	85,9	115
Еврейская автономная область	707,9	282,6	9	19,7	23	25	15	58,7	105
Чукотский автономный округ	1213,7	130,4	1	10,5	22	25	5	88,8	90

Приложение Ж
(справочное)

Распределение критерия Дарбина-Уотсона для положительной автокорреляции на уровне значимости 0,05

Таблица Ж.1

n	V=1		V=2		V=3		V=4		V=5	
	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2	d_1	d_2
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,1	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,9	1,71	0,78	1,9	0,67	2,1
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,4	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,2	1,41	1,10	1,54	1	1,68	0,9	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,8	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,9	1,92
24	1,27	1,45	1,19	1,55	1,1	1,66	1,01	1,78	0,93	1,9
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,3	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,89
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,1	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,2	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
31	1,36	1,50	1,3	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,63	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,8
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,8
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,8
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,5	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,6	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,5	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
85	1,62	1,67	1,6	1,70	1,57	1,72	1,55	1,75	1,52	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
95	1,64	1,69	1,62	1,71	1,6	1,73	1,58	1,75	1,56	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Приложение И
(справочное)

Расчет параболического тренда численности населения России

Таблица И.1

Год	Уровень, y_i тыс.чел.	t_i	$y_i\, t_i$	t^2	t^4	$y_i t_i^2$	Тренд, $\tilde{y}_i$
1	2	3	4	5	6	7	8
1900	70200	-52	-3650400	2704	7311616	189820800	73689
1901	71476	-51	-3645276	2601	6765201	185909076	74460
1902	72748	-50	-3637400	2500	6250000	181870000	75231
1903	74043	-49	-3628107	2401	5764801	177777243	76002
1904	75361	-48	-3617328	2304	5308416	173631744	76773
1905	76702	-47	-3604994	2209	4879681	169434718	77544
1906	78067	-46	-3591082	2116	4477456	165189772	78315
1907	79457	-45	-3575565	2025	4100625	160900425	79086
1908	80871	-44	-3558324	1936	3748096	156566256	79857
1909	82311	-43	-3539373	1849	3418801	152193039	80628
1910	83776	-42	-3518592	1764	3111696	147780864	81399
1911	85267	-41	-3495947	1681	2825761	143333827	82170
1912	86785	-40	-3471400	1600	2560000	138856000	82940
1913	88330	-39	-3444870	1521	2313441	134349930	83711
1914	89902	-38	-3416276	1444	2085136	129818488	84482
1915	90262	-37	-3339694	1369	1874161	123568678	85253
1916	90623	-36	-3262428	1296	1679616	117447408	86024

Продолжение таблицы И.1

1	2	3	4	5	6	7	8
1917	91009	-35	-3185315	1225	1500625	111486025	86794
1918	90099	-34	-3063366	1156	1336336	104154444	87565
1919	89198	-33	-2943534	1089	1185921	97136622	88336
1920	88247	-32	-2823904	1024	1048576	90364928	89106
1921	88079	-31	-2730449	961	923521	84643919	89877
1922	87912	-30	-2637360	900	810000	79120800	90648
1923	87755	-29	-2544895	841	707281	73801955	91418
1924	89379	-28	-2502612	784	614656	70073136	92189
1925	91042	-27	-2458134	729	531441	66369618	92959
1926	92735	-26	-2411110	676	456976	62688860	93730
1927	93786	-25	-2344650	625	390625	58616250	94500
1928	94848	-24	-2276352	576	331776	54632448	95271
1929	95923	-23	-2206229	529	279841	50743267	96041
1930	97010	-22	-2134220	484	234256	46952840	96812
1931	98109	-21	-2060289	441	194481	43266069	97582
1932	99220	-20	-1984400	400	160000	39688000	98352
1933	100345	-19	-1906555	361	130321	36224545	99123
1934	101481	-18	-1826658	324	104976	32879844	99893
1935	102631	-17	-1744727	289	83521	29660359	100663
1936	103794	-16	-1660704	256	65536	26571264	101434
1937	104932	-15	-1573980	225	50625	23609700	102204
1938	106639	-14	-1492946	196	38416	20901244	102974

Продолжение таблицы И.1

1	2	3	4	5	6	7	8
1939	108377	-13	-1408901	169	28561	18315713	103745
1940	110098	-12	-1321176	144	20736	15854112	104515
1941	109217	-11	-1201387	121	14641	13215257	105285
1942	108343	-10	-1083430	100	10000	10834300	106055
1943	106617	-9	-959553	81	6561	8635977	106825
1944	105764	-8	-846112	64	4096	6768896	107595
1945	104918	-7	-734426	49	2401	5140982	108365
1946	104079	-6	-624474	36	1296	3746844	109136
1947	103246	-5	-516230	25	625	2581150	109906
1948	102420	-4	-409680	16	256	1638720	110676
1949	101601	-3	-304803	9	81	914409	111446
1950	101438	-2	-202876	4	16	405752	112216
1951	103112	-1	-103112	1	1	103112	112986
1952	104813	0	0	0	0	0	113756
1953	106542	1	106542	1	1	106542	114526
1954	108300	2	216600	4	16	433200	115295
1955	110087	3	330261	9	81	990783	116065
1956	111903	4	447612	16	256	1790448	116835
1957	113749	5	568745	25	625	2843725	117605
1958	115626	6	693756	36	1296	4162536	118375
1959	117534	7	822738	49	2401	5759166	119145
1960	119046	8	952368	64	4096	7618944	119914

Продолжение таблицы И.1

1	2	3	4	5	6	7	8
1961	120766	9	1086894	81	6561	9782046	120684
1962	122407	10	1224070	100	10000	12240700	121454
1963	123848	11	1362328	121	14641	14985608	122224
1964	125179	12	1502148	144	20736	18025776	122993
1965	126309	13	1642017	169	28561	21346221	123763
1966	127189	14	1780646	196	38416	24929044	124533
1967	128026	15	1920390	225	50625	28805850	125302
1968	128696	16	2059136	256	65536	32946176	126072
1969	129379	17	2199443	289	83521	37390531	126841
1970	130079	18	2341422	324	104976	42145596	127611
1971	130704	19	2483376	361	130321	47184144	128380
1972	131445	20	2628900	400	160000	52578000	129150
1973	132210	21	2776410	441	194481	58304610	129919
1974	132941	22	2924702	484	234256	64343444	130689
1975	133775	23	3076825	529	279841	70766975	131458
1976	134690	24	3232560	576	331776	77581440	132228
1977	135645	25	3391125	625	390625	84778125	132997
1978	136596	26	3551496	676	456976	92338896	133767
1979	137551	27	3713877	729	531441	100274679	134536
1980	138291	28	3872148	784	614656	108420144	135305
1981	139028	29	4031812	841	707281	116922548	136075
1982	139816	30	4194480	900	810000	125834400	136844

Продолжение таблицы И.1

1	2	3	4	5	6	7	8
1983	140766	31	4363746	961	923521	135276126	137613
1984	141842	32	4538944	1024	1048576	145246208	138382
1985	142823	33	4713159	1089	1185921	155534247	139152
1986	143835	34	4890390	1156	1336336	166273260	139921
1987	145115	35	5079025	1225	1500625	177765875	140690
1988	146343	36	5268348	1296	1679616	189660528	141459
1989	147400	37	5453800	1369	1874161	201790600	142228
1990	148041	38	5625558	1444	2085136	213771204	142997
1991	148543	39	5793177	1521	2313441	225933903	143767
1992	148704	40	5948160	1600	2560000	237926400	144536
1993	148673	41	6095593	1681	2825761	249919313	145305
1994	148366	42	6231372	1764	3111696	261717624	146074
1995	148306	43	6377158	1849	3418801	274217794	146843
1996	147976	44	6510944	1936	3748096	286481536	147612
1997	147502	45	6637590	2025	4100625	298691550	148381
1998	147105	46	6766830	2116	4477456	311274180	149150
1999	146388	47	6880236	2209	4879681	323371092	149918
2000	145300	48	6974400	2304	5308416	334771200	150687
2001	145600	49	7134400	2401	5764801	349585600	151456
2002	145200	50	7260000	2500	6250000	363000000	152225
2003	144200	51	7354200	2601	6765201	375064200	152994
2004	143500	52	7462000	2704	7311616	388024000	153763
Итого	11943282	0	74268252	96460	159486964	10971116366	11943277
Источник: Симчера В.М. Развитие экономики России за 100 лет: 1900-2000. Исторические ряды, вековые тренды, периодические циклы / В.М. Симчера. – М.: ЗАО «Издательство Экономика», 2007. – 683 с. - ISBN 5-282-02627-9. Расчеты автора.							

Приложение К
(справочное)

Расчет экспоненциального тренда национального богатства РФ в сопоставимых ценах

Таблица К.1

Год	y_i , млрд. р.	t_i	$y_i t_i$	$\ln y_i$	$t_i \ln y_i$	Тренд, $\tilde{y}_i$, млрд. р.
1	2	3	4	5	6	7
1900	248	-45,5	5,51	-250,86	2070,25	120
1901	248	-44,5	5,51	-245,35	1980,25	126
1902	249	-43,5	5,52	-240,01	1892,25	134
1903	250	-42,5	5,52	-234,66	1806,25	141
1904	261	-41,5	5,56	-230,93	1722,25	149
1905	262	-40,5	5,57	-225,52	1640,25	158
1906	262	-39,5	5,57	-219,95	1560,25	166
1907	263	-38,5	5,57	-214,53	1482,25	176
1908	264	-37,5	5,58	-209,10	1406,25	186
1909	265	-36,5	5,58	-203,66	1332,25	196
1910	277	-35,5	5,62	-199,65	1260,25	207
1911	278	-34,5	5,63	-194,15	1190,25	219
1912	279	-33,5	5,63	-188,65	1122,25	231
1913	280	-32,5	5,63	-183,13	1056,25	244
1914	271	-31,5	5,60	-176,47	992,25	258
1915	273	-30,5	5,61	-171,09	930,25	273
1916	264	-29,5	5,58	-164,49	870,25	288
1917	268	-28,5	5,59	-159,34	812,25	305
1918	277	-27,5	5,62	-154,66	756,25	322
1919	287	-26,5	5,66	-149,98	702,25	340

Продолжение таблицы К.1

1	2	3	4	5	6	7
1920	296	-25,5	5,69	-145,10	650,25	359
1921	306	-24,5	5,72	-140,23	600,25	379
1922	327	-23,5	5,79	-136,06	552,25	401
1923	341	-22,5	5,83	-131,22	506,25	423
1924	355	-21,5	5,87	-126,25	462,25	447
1925	369	-20,5	5,91	-121,17	420,25	473
1926	395	-19,5	5,98	-116,59	380,25	499
1927	411	-18,5	6,02	-111,34	342,25	528
1928	438	-17,5	6,08	-106,44	306,25	557
1929	455	-16,5	6,12	-100,98	272,25	589
1930	484	-15,5	6,18	-95,82	240,25	622
1931	513	-14,5	6,24	-90,48	210,25	657
1932	545	-13,5	6,30	-85,06	182,25	694
1933	566	-12,5	6,34	-79,23	156,25	734
1934	599	-11,5	6,40	-73,55	132,25	775
1935	635	-10,5	6,45	-67,76	110,25	819
1936	683	-9,5	6,53	-62,00	90,25	865
1937	723	-8,5	6,58	-55,96	72,25	914
1938	765	-7,5	6,64	-49,80	56,25	965
1939	810	-6,5	6,70	-43,53	42,25	1020
1940	868	-5,5	6,77	-37,21	30,25	1078
1941	878	-4,5	6,78	-30,50	20,25	1138
1942	899	-3,5	6,80	-23,80	12,25	1203
1943	932	-2,5	6,84	-17,09	6,25	1271
1944	988	-1,5	6,90	-10,34	2,25	1342
1945	1012	-0,5	6,92	-3,46	0,25	1418
1946	1047	0,5	6,95	3,48	0,25	1498
1947	1071	1,5	6,98	10,46	2,25	1583
1948	1108	2,5	7,01	17,53	6,25	1672
1949	1134	3,5	7,03	24,62	12,25	1767

Продолжение таблицы К.1

1	2	3	4	5	6	7
1950	1191	4,5	7,08	31,87	20,25	1867
1951	1307	5,5	7,18	39,47	30,25	1972
1952	1450	6,5	7,28	47,32	42,25	2083
1953	1597	7,5	7,38	55,32	56,25	2201
1954	1760	8,5	7,47	63,52	72,25	2325
1955	1929	9,5	7,56	71,87	90,25	2457
1956	2127	10,5	7,66	80,46	110,25	2595
1957	2353	11,5	7,76	89,28	132,25	2742
1958	2588	12,5	7,86	98,23	156,25	2897
1959	2855	13,5	7,96	107,42	182,25	3060
1960	3093	14,5	8,04	116,54	210,25	3233
1961	3341	15,5	8,11	125,77	240,25	3416
1962	3610	16,5	8,19	135,16	272,25	3609
1963	3905	17,5	8,27	144,73	306,25	3813
1964	4223	18,5	8,35	154,44	342,25	4028
1965	4561	19,5	8,43	164,29	380,25	4255
1966	4874	20,5	8,49	174,08	420,25	4496
1967	5209	21,5	8,56	184,00	462,25	4750
1968	5582	22,5	8,63	194,11	506,25	5018
1969	5988	23,5	8,70	204,39	552,25	5301
1970	6416	24,5	8,77	214,78	600,25	5601
1971	6866	25,5	8,83	225,28	650,25	5917
1972	7360	26,5	8,90	235,95	702,25	6251
1973	7879	27,5	8,97	246,73	756,25	6604
1974	8437	28,5	9,04	257,65	812,25	6977
1975	9030	29,5	9,11	268,70	870,25	7371
1976	9624	30,5	9,17	279,75	930,25	7787
1977	10238	31,5	9,23	290,87	992,25	8227
1978	10916	32,5	9,30	302,18	1056,25	8692
1979	11598	33,5	9,36	313,51	1122,25	9183

Продолжение таблицы К.1

1	2	3	4	5	6	7
1980	12283	34,5	9,42	324,85	1190,25	9701
1981	13 002	35,5	9,47	336,29	1260,25	10249
1982	13 764	36,5	9,53	347,84	1332,25	10828
1983	14 573	37,5	9,59	359,51	1406,25	11439
1984	15 433	38,5	9,64	371,30	1482,25	12085
1985	16 302	39,5	9,70	383,11	1560,25	12768
1986	17 152	40,5	9,75	394,87	1640,25	13489
1987	17 958	41,5	9,80	406,53	1722,25	14251
1988	19 065	42,5	9,86	418,86	1806,25	15056
1989	19 766	43,5	9,89	430,29	1892,25	15906
1990	20 579	44,5	9,93	441,98	1980,25	16804
1991	20 902	45,5	9,95	452,62	2070,25	17753
Итого	377965	0	670,18	3564,60	64883,00	329987
Источник: Симчера В.М. Развитие экономики России за 100 лет: 1900-2000. Исторические ряды, вековые тренды, периодические циклы / В.М. Симчера. – М.: ЗАО «Издательство Экономика», 2007. – 683 с. - ISBN 5-282-02627-9. Расчеты автора.						

Приложение Л
(справочное)

Данные для построения модели с распределенным лагом

Таблица Л.1

Год	Месяц	Y	X_t	X_{t-1}	X_{t-2}	X_{t-3}	Z_0	Z_1	Z_2
1	2	3	4	5	6	7	8	9	10
2005	1	128,2	157,9						
	2	128,9	196,7						
	3	129,6	221,6						
	4	130,4	226,7	221,6	196,7	157,9	802,9	1088,7	2429,5
	5	131,3	267,2	226,7	221,6	196,7	912,2	1260	2883,4
	6	132,2	333,8	267,2	226,7	221,6	1049,3	1385,4	3168,4
	7	133,1	326,8	333,8	267,2	226,7	1154,5	1548,3	3442,9
	8	133,8	359,1	326,8	333,8	267,2	1286,9	1796	4066,8
	9	134,4	373,5	359,1	326,8	333,8	1393,2	2014,1	4670,5
	10	135,2	354,4	373,5	359,1	326,8	1413,8	2072,1	4751,1
	11	136,3	407,8	354,4	373,5	359,1	1494,8	2178,7	5080,3
	12	137,7	624,5	407,8	354,4	373,5	1760,2	2237,1	5186,9
2006	1	139,1	188,0	624,5	407,8	354,4	1574,7	2503,3	5445,3
	2	140,4	231,9	188,0	624,5	407,8	1452,2	2660,4	6356,2
	3	141,6	282,1	231,9	188,0	624,5	1326,5	2481,4	6604,4
	4	142,6	285,5	282,1	231,9	188,0	987,5	1309,9	2901,7

Продолжение таблицы Л.1

1	2	3	4	5	6	7	8	9	10
	5	143,1	363,1	285,5	282,1	231,9	1162,6	1545,4	3501,0
	6	143,5	436,2	363,1	285,5	282,1	1366,9	1780,4	4044,0
	7	144,0	406,7	436,2	363,1	285,5	1491,5	2018,9	4458,1
	8	144,8	466,5	406,7	436,2	363,1	1672,5	2368,4	5419,4
	9	146,0	499,1	466,5	406,7	436,2	1808,5	2588,5	6019,1
	10	147,5	493,5	499,1	466,5	406,7	1865,8	2652,2	6025,4
	11	148,7	542,2	493,5	499,1	466,5	2001,3	2891,2	6688,4
	12	149,5	847,8	542,2	493,5	499,1	2382,6	3026,5	7008,1
2007	1	150,1	272,2	847,8	542,2	493,5	2155,7	3412,7	7458,1
	2	150,7	317,7	272,2	847,8	542,2	1979,9	3594,4	8543,2
	3	151,7	367,1	317,7	272,2	847,8	1804,8	3405,5	9036,7
	4	153,2	388,6	367,1	317,7	272,2	1345,6	1819,1	4087,7
	5	154,8	503,4	388,6	367,1	317,7	1576,8	2075,9	4716,3
	6	155,9	615,9	503,4	388,6	367,1	1875	2381,9	5361,7
	7	156,2	579,0	615,9	503,4	388,6	2086,9	2788,5	6126,9
	8	156,4	622,8	579,0	615,9	503,4	2321,1	3321,0	7573,2
	9	156,8	657,6	622,8	579,0	615,9	2475,3	3628,5	8481,9
	10	158,0	730,0	657,6	622,8	579,0	2589,4	3640,2	8359,8
	11	160,0	789,3	730,0	657,6	622,8	2799,7	3913,6	8965,6
	12	162,4	1316,6	789,3	730,0	657,6	3493,5	4222,1	9627,7

Продолжение таблицы Л.1

1	2	3	4	5	6	7	8	9	10
2008	1	164,7	388,4	1316,6	789,3	730,0	3224,3	5085,2	11044,0
	2	166,4	476,9	388,4	1316,6	789,3	2971,2	5389,5	12759,0
	3	167,5	536,2	476,9	388,4	1316,6	2718,1	5203,5	13880,0
	4	167,8	580,1	536,2	476,9	388,4	1981,6	2655,2	5939,4
	5	167,6	716,6	580,1	536,2	476,9	2309,8	3083,2	7017,0
	6	167,1	826,4	716,6	580,1	536,2	2659,3	3485,4	7862,8
	7	166,0	773,5	826,4	716,6	580,1	2896,6	3999,9	8913,7
	8	164,3	833,7	773,5	826,4	716,6	3150,2	4576,1	10529,0
	9	161,5	918,4	833,7	773,5	826,4	3352	4859,9	11365,0
	10	157,8	937,1	918,4	833,7	773,5	3462,7	4906,3	11215,0
	11	153,6	935,4	937,1	918,4	833,7	3624,6	5275,0	12114,0
	12	149,6	1439,3	935,4	937,1	918,4	4230,2	5564,8	12949,0
2009	1	146,7	362,7	1439,3	935,4	937,1	3674,5	6121,4	13615,0
	2	144,9	454,4	362,7	1439,3	935,4	3191,8	6047,5	14539,0
	3	144,2	488,2	454,4	362,7	1439,3	2744,6	5497,7	14859,0
	4	144,3	517,1	488,2	454,4	362,7	1822,4	2485,1	5570,1
	5	145,1	599,7	517,1	488,2	454,4	2059,4	2856,7	6559,5
	6	146,5	719,2	599,7	517,1	488,2	2324,2	3098,5	7061,9
	7	148,2	675,3	719,2	599,7	517,1	2511,3	3469,9	7771,9
	8	149,6	720,5	675,3	719,2	599,7	2714,7	3912,8	8949,4

Продолжение таблицы Л.1

1	2	3	4	5	6	7	8	9	10
2009	9	150,5	801,5	720,5	675,3	719,2	2916,5	4228,7	9894,5
	10	151,1	822,9	801,5	720,5	675,3	3020,2	4268,4	9761,2
	11	151,7	886,2	822,9	801,5	720,5	3231,1	4587,4	10513,0
	12	152,3	1455,7	886,2	822,9	801,5	3966,3	4936,5	11391,0
2010	1	153,2	353,6	1455,7	886,2	822,9	3518,4	5696,8	12407,0
	2	154,3	449,0	353,6	1455,7	886,2	3144,5	5923,6	14152,0
	3	155,2	522,1	449,0	353,6	1455,7	2780,4	5523,3	14965,0
	4	155,8	562,0	522,1	449,0	353,6	1886,7	2480,9	5500,5
	5	156,1	684,6	562,0	522,1	449,0	2217,7	2953,2	6691,4
	6	155,9	845,5	684,6	562,0	522,1	2614,2	3374,9	7631,5
	7	155,9	728,7	845,5	684,6	562,0	2820,8	3900,7	8641,9
	8	156,2	847,2	728,7	845,5	684,6	3106,0	4473,5	10272,0
	9	157,3	941,3	847,2	728,7	845,5	3362,7	4841,1	11372,0
	10	158,7	992,8	941,3	847,2	728,7	3510,0	4821,8	10888,0
	11	160,2	1042,3	992,8	941,3	847,2	3823,6	5417,0	12383,0
	12	161,3	1787,1	1042,3	992,8	941,3	4763,5	5851,8	13485,0
2011	1	161,9	379,3	1787,1	1042,3	992,8	4201,5	6850,1	14892,0
	2	162,1	491,8	379,3	1787,1	1042,3	3700,5	7080,4	16908,0
	3	162,2	571,3	491,8	379,3	1787,1	3229,5	6611,7	18093,0
	4	162,6	629,1	571,3	491,8	379,3	2071,5	2692,8	5952,2
	5	163,2	802,7	629,1	571,3	491,8	2494,9	3247,1	7340,5

Продолжение таблицы Л.1

1	2	3	4	5	6	7	8	9	10
2011	6	164,1	961,6	802,7	629,1	571,3	2964,7	3774,8	8460,8
	7	165,2	855,0	961,6	802,7	629,1	3248,4	4454,3	9834,3
	8	166,6	986,4	855,0	961,6	802,7	3605,7	5186,3	11926,0
	9	167,8	1115,0	986,4	855,0	961,6	3918,0	5581,2	13061,0
	10	168,7	1225,4	1115,0	986,4	855	4181,8	5652,8	12756,0
	11	169,4	1271,1	1225,4	1115,0	986,4	4597,9	6414,6	14563,0
	12	169,7	2200,4	1271,1	1225,4	1115	5811,9	7066,9	16208,0
2012	1	169,6	469,5	2200,4	1271,1	1225,4	5166,4	8418,8	18313,0
	2	169,4	610,9	469,5	2200,4	1271,1	4551,9	8683,6	20711,0
	3	169,5	714,7	610,9	469,5	2200,4	3995,5	8151,1	22293,0
	4	169,8	727,9	714,7	610,9	469,5	2523,0	3345,0	7383,8
	5	170,4	923,6	727,9	714,7	610,9	2977,1	3990,0	9084,8
Источник: http://www. hse.ru									