Министерство образования и науки Российской Федерации
Байкальский государственный университет экономики и права

Л.Н. Ежова, Р.З. Абдуллин, В.Р. Абдуллин

ЭКОНОМЕТРИЧЕСКИЕ МЕТОДЫ И МОДЕЛИ

Учебное пособие для магистрантов,
обучающихся по направлению «Экономика»

Иркутск
Издательство БГУЭП
2012

УДК 519.862.6(075.8)
ББК 22.17я7
Е41

Печатается по решению редакционно-издательского совета
Байкальского государственного университета экономики и права

Рецензенты канд. физ.-мат. наук, доц. Т.Г. Тюрнева
канд. физ.-мат. наук, доц. О.Г. Леонова

Ежова Л.Н.
Е41 Эконометрические методы и модели : учеб. пособие / Л.Н. Ежова, Р.З. Абдуллин, В.Р. Абдуллин. – Иркутск : Изд-во БГУЭП, 2012. – 93 с.

ISBN 978-5-7253-2525-6

Приводятся основные теоретические сведения курса «Эконометрика», примеры построения и анализа эконометрических моделей. Каждая глава сопровождается контрольными вопросами, задачами и упражнениями, которые позволяют осуществить самоконтроль и приобрести практические навыки построения и анализа эконометрических моделей социально-экономических процессов.

Для магистрантов, обучающихся по направлению «Экономика».

ББК 22.17я7

ISBN 978-5-7253-2525-6

Оглавление

Предисловие

Современные процессы и явления в экономике предъявляют все более высокие требования к качеству их анализа и прогнозирования, что невозможно без использования различных методов моделирования этих процессов и явлений. Это повышает требования к уровню подготовки экономистов в области методов и инструментов экономических исследований в условиях многоступенчатого университетского образования. Методы построения математических моделей взаимосвязей экономических показателей, основанные на использовании реальной экономической информации, дает эконометрика, назначение которой в количественном анализе закономерностей, присущих реальным процессам в экономике. Основная образовательная программа бакалавров направления «экономика» не дает представления об эконометрике как о дисциплине, объединяющей результаты и методы экономической теории, теории статистики и математики.

Учебное пособие адресовано магистрантам, впервые приступающим к изучению эконометрики. При этом предполагается, что они знакомы с основами микро- и макроэкономики, экономической статистики, теории вероятностей и математической статистики, линейной алгебры и математического анализа, информатики.

Овладение методами эконометрики позволит будущим магистрам экономики проводить исследования и оценивать связи между экономическими показателями, моделируя их с помощью разнообразных математических соотношений – эконометрических моделей, на основе данных статистических наблюдений. Эконометрический подход предусматривает также проверку пригодности или соответствия выбранной модели изучаемому объекту, что в большинстве случаев позволяет осуществить проверку справедливости положений экономической теории.

Пособие содержит краткие теоретические сведения, традиционно относящиеся к основам эконометрики: построение моделей парной линейной и нелинейной регрессии; множественная регрессия и возникающие при этом особенности, системы одновременных регрессионных уравнений.

Эконометрические методы реализованы в различных пакетах прикладных программ. Практикум на компьютере по данному курсу эконометрики использует Microsoft Excel, как наиболее распространенный и доступный инструментарий построения эконометрических моделей и анализа их пригодности.

Введение

Эконометрическое моделирование социально-экономических процессов

В этой главе мы обсудим предмет и основные задачи эконометрики, приведем общую схему эконометрических исследований, опишем этапы построения эконометрических моделей, рассмотрим их классификацию и тех данных, которые используются в практике эконометрического моделирования.

В.1. Предмет и основные задачи эконометрики

Эконометрика является сравнительно молодой отраслью науки, известной под таким названием (или названием «эконометрия») только с 1930 г. Введя термин «эконометрика» для обозначения самостоятельной отрасли научных исследований, крупнейший норвежский экономист и статистик Рагнар Фриш провозгласил в качестве основной задачи «развитие экономической теории в ее связи со статистикой и математикой».

Зарождение эконометрики является следствием междисциплинарного подхода к изучению экономики. Эта наука возникла в результате взаимодействия трех компонент: экономической теории, статистических и математических методов. Впоследствии к ним присоединилось развитие вычислительной техники и программного обеспечения, как условие развития эконометрики и возможности ее использования в реальных задачах. Существуют различные варианты определения эконометрики. Буквально термин «эконометрия» (мы будем придерживаться названия «эконометрика»), обозначает измерение в экономике, и измерение действительно является важной частью эконометрики. Оценка национального дохода или разработка индекса розничных цен – важные проблемы измерения, однако это не эконометрические проблемы.

Эконометрика – это наука, в которой с помощью статистических методов устанавливаются количественные взаимосвязи между экономическими переменными. То есть под эконометрикой следует понимать определенный набор математико-статистических средств, позволяющих проверять или верифицировать модельные соотношения между анализируемыми экономическими показателями и оценивать неизвестные значения параметров в этих соотношениях на основе исходных экономических данных.

Эконометрику можно определить как специальный вид экономического анализа, в котором объединены два аспекта: общий теоретический метод, часто формулируемый математически, и эмпирическое измерение экономических показателей. Таким образом, один из ответов на вопрос о том, что же такое эконометрика, может звучать так: это наука, связанная с эмпирическим обоснованием и подтверждением экономических законов. Как правило, основные результаты экономической теории носят не количественный, а качественный характер. Так, из теории следует, что при прочих равных условиях повышение

цены товара ведет к уменьшению спроса на него. Однако вопрос о том, на сколько снизится спрос при увеличении цены конкретного товара в конкретных условиях, уже выходит за рамки экономической теории. Ответ на него можно дать, используя методы эконометрики, которые позволяют эмпирически, то есть на основе данных об экономических показателях, принять или опровергнуть положения теории. Для того чтобы получить количественные зависимости для экономических показателей, используются данные или наблюдения, которые, как правило, не являются экспериментальными. В экономике, в отличие от любой другой науки (химии, биологии, медицины и т. д.), мы не в состоянии проводить многократные эксперименты и «вмешиваться» в постановку и организацию таких экспериментов.

Можно выделить с одной стороны – эконометрические методы, с другой – их приложения к конкретным экономическим задачам. Применяемые в эконометрике методы базируются на разделах регрессионного, дисперсионного и корреляционного анализов. Однако специфичность задач, с которыми здесь сталкиваются, вызывает необходимость особых изменений в принятых подходах и разработке специальных приемов. Взаимосвязи, которые исследуются с помощью этих методов, например, функции спроса или производственные функции, являются сердцевиной экономической теории, в то же время конкретная их форма, принятая в конкретном исследовании, может быть совершенно новой.

С точки зрения теоретиков эконометрическое исследование начинается после того, как

1) выбрана математическая модель объекта с фиксированной формой всех зависимостей и с неизвестными параметрами при входящих в модель переменных;

2) получено множество наблюдений над входящими в модель переменными в соответствующие моменты времени;

3) поставлена задача отыскания значений неизвестных параметров, обеспечивающих наилучшее (с точки зрения фиксированного критерия) приближение модельных значений переменных к их значениям, наблюдавшимся в действительности, проверки по отношению к ним разнообразных гипотез и верификации полученной модели, то есть проверить ее пригодность или адекватность. Построенная и верифицированная модель может использоваться в задачах прогноза и управления объектом исследования.

В соответствии с этим можно определить источники, на которых базируется эконометрическая наука:

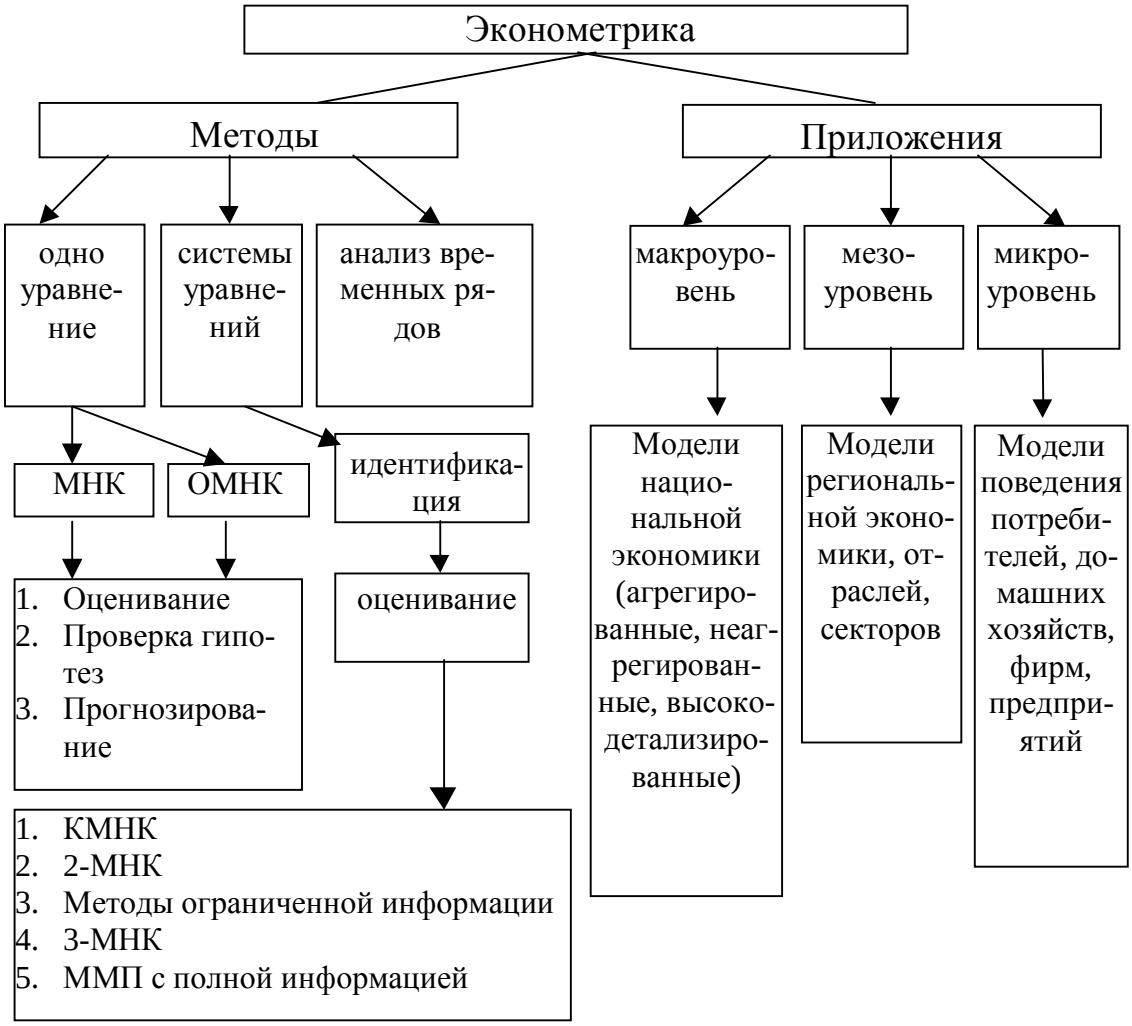
– экономическая теория (макро- и микроэкономика, математическая экономика);

– социально-экономическая статистика (включая информационное обеспечение экономических исследований);

– основы теории вероятностей и математической статистики;

– математический анализ и линейная алгебра.

Ниже представлена структура эконометрических исследований. Эта схема, конечно, условна. Однако она поможет лучше понять существующую точку зрения на эконометрику и ее задачи.



В.2. Эконометрические модели

Каждый изучающий экономику сталкивается с принципиальной идеей о взаимосвязях между экономическими показателями и необходимостью описания этих взаимосвязей с помощью математических соотношений. Например, формирующийся на рынке спрос на некоторый товар есть функция его цены; затраты, связанные с изготовлением какого-либо продукта, зависят от объема производства; потребительские расходы могут быть функцией дохода и т. д. Все это примеры связей между двумя переменными, одна из которых (спрос на товар, производственные затраты, потребительские расходы) играет роль объясняемой переменной (или результирующего показателя), а другие интерпретируются как объясняющие переменные (факторы или регрессоры). Однако реально в каждое такое соотношение приходится вводить несколько объясняющих переменных и случайную составляющую, отражающую влияние на ре-

зультирующий показатель всех неучтенных факторов или обусловленную другими причинами. Спрос на товар можно рассматривать как функцию его цены, потребительского дохода и цен на конкурирующие и дополняющие товары, производственные затраты будут зависеть от объема производства, от его динамики и от цен на основные производственные ресурсы; потребительские расходы можно определить как функцию дохода, ликвидных активов и предыдущего уровня потребления. При этом участвующая в каждом из этих соотношений случайная составляющая обуславливает стохастический или статистический характер зависимости. Эта зависимость выражается в том, что если мы зафиксируем на определенных уровнях значения объясняющих переменных, допустим, цены на сам товар и на конкурирующие с ним или дополняющие товары, а также потребительский доход, то не можем ожидать, что тем самым однозначно определяется спрос на этот товар. Иными словами, в реальной ситуации мы имеем случайное варьирование величины спроса относительно некоторого уровня даже при неизменных значениях всех объясняющих переменных.

Большая часть традиционных экономических теорий, в которых связи между экономическими показателями отражаются с помощью диаграмм и алгебраических формул, имеет дело с точными функциональными соотношениями – экономическими моделями. Количество связей, включаемых в экономическую модель, зависит от условий, при которых эта модель конструируется, и от того, насколько подробно стремимся мы объяснить то или иное явление. Например, традиционная модель спроса и предложения должна объяснять соотношения между ценой и объемом выпуска, характерные для некоторого определенного рынка. Она содержит три уравнения, а именно уравнение спроса, уравнение предложения и уравнение реакции рынка (см. пример В.2).

Все экономические модели, независимо от того, относятся они ко всему хозяйству или к его элементам (т. е. к макроэкономике, отрасли, фирме или рынку), имеют некоторые общие особенности. Во-первых, они основаны на предположении, что поведение экономических переменных определяется с помощью совместных и одновременных операций с некоторым числом экономических соотношений. Во-вторых, принимается гипотеза, в силу которой модель, допуская упрощение сложной действительности, тем не менее, улавливает главные характеристики изучаемого объекта. В-третьих, создатель модели полагает, что на основе достигнутого с ее помощью понимания реальной системы удастся предсказать ее будущее движение и, возможно, управлять им в целях улучшения экономического благосостояния.

Чтобы проиллюстрировать сказанное, рассмотрим пример достаточно общей и приближенной макроэкономической модели.

Пример В.1. Предположим, что экономист-теоретик сформулировал следующие положения:

- объем потребления C_t в период времени t есть возрастающая функция от имеющегося в этот период национального дохода Y_t за вычетом подоходного налога T_t, но возрастающая, видимо, медленнее, чем рост дохода;

- объем инвестиций I_t периода t есть возрастающая функция национального дохода Y_t и убывающая функция характеристики государственного регулирования (например, нормы процента R_t);
- национальный доход Y_t за период времени t есть сумма потребительских C_t, инвестиционных I_t и государственных G_t закупок товаров и услуг.

Наша первая задача – перевести эти положения на математический язык. Возникает вопрос: какие соотношения выбрать между переменными – линейные или нелинейные (логарифмические, полиномиальные и т. д.). Даже определив форму конкретного соотношения, мы оставляем еще нерешенной проблему выбора для различных уравнений запаздываний по времени. Будут ли, например, инвестиции текущего периода реагировать на национальный доход, произведенный в последнем периоде, или же на них скажется динамика нескольких предыдущих периодов? Обычный выход из этих трудностей состоит в выборе при первоначальном анализе наиболее простой из возможных форм этих зависимостей. Тогда появляется возможность записать на основе указанных выше положений следующие соотношения:

$$C_t = a_0 + a_1\left(Y_t - T_t\right), \qquad \text{(B.1)}$$

$$I_t = b_1 Y_{t-1} + b_2 R_t, \qquad \text{(B.2)}$$

$$Y_t = C_t + I_t + G_t, \qquad \text{(B.3)}$$

где априорные ограничения выражены неравенствами $0 < a_1 < 1,\ b_1 > 0,\ b_2 < 0.$ Эти три соотношения вместе с ограничениями образуют модель.

Модель сформулирована (два уравнения, объясняющие поведение потребителей и инвесторов, и одно тождество) для дискретных моментов времени, и выбрано запаздывание (лаг) в один период для отражения воздействия национального дохода предыдущего периода на инвестиции.

Уравнения поведения имеют здесь форму точных функциональных зависимостей, однако, как мы увидим позднее, это нереалистично, и нельзя приступать к эконометрическим разработкам, не пользуясь некоторыми дополнительными стохастическими спецификациями. То есть уравнения (B.1) и (B.2) должны содержать аддитивные случайные составляющие e_t и d_t, обусловленные необходимостью учесть влияние соответственно на C_t и I_t ряда неучтенных факторов. Действительно, нереалистично ожидать, что величина потребления C_t будет однозначно определяться уровнями национального дохода Y_t и подоходного налога T_t; аналогично величина инвестиций I_t зависит, очевидно, не только от достигнутого в предыдущий год уровня национального дохода Y_{t-1} и от величины нормы процента R_t, но и от ряда других, не учтенных в уравнении (B.2), факторов. Таким образом, для реальной ситуации мы имеем линейную относительно анализируемых переменных и аддитивную относительно случайных составляющих e_t и d_t модель:

$$C_t = a_0 + a_1\left(Y_t - T_t\right) + e_t; \quad \text{(B.4)}$$

$$I_t = b_1 Y_{t-1} + b_2 R_t + d_t; \quad \text{(B.5)}$$

$$Y_t = C_t + I_t + G_t, \quad \text{(B.6)}$$

где $0 < a_1 < 1,\ b_1 > 0,\ b_2 < 0.$ Здесь коэффициенты или параметры a_0, a_1, b_1, b_2 и характеристики случайных составляющих e_t и d_t неизвестны до получения «наблюдений» над экономическими переменными.

Если мы поставим задачу нахождения оценок этих параметров по результатам наблюдений (исходным статистическим данным), найдем эти оценки и верифицируем полученную модель, то мы получим математическую модель, описывающую конкретное (а не гипотетическое) экономическое явление. Такая модель будет эконометрической.

Мы привели здесь этот пример, чтобы пояснить общие черты одного из важнейших этапов эконометрического моделирования, в процессе которого исследователь математически формализует отдельные положения экономической теории (этап «выбора» модели). В дальнейшем мы используем этот пример для пояснения некоторых основных понятий эконометрического моделирования.

B.3. Типы моделей

Основным этапом эконометрического моделирования является выбор модели. Модель, построенная и верифицированная на основе данных наблюдений над объясняемыми и объясняющими переменными, может быть использована для прогноза значений зависимых переменных в будущем или для других наборов значений объясняющих переменных, а также для анализа взаимосвязей изучаемых экономических переменных.

Можно выделить три основных класса моделей, которые применяются для анализа и / или прогноза явлений и процессов в экономике.

1. Регрессионные модели с одним уравнением. В таких моделях зависимая (объясняемая) переменная y представляется в виде функции

$$y = f(x_1, \mathbf{K}, x_k; b_1, \mathbf{K}, b_p),$$

где $x_1, \mathbf{K}, x_k$ – независимые (объясняющие) переменные-факторы, а $b_1, \mathbf{K}, b_p$ – параметры. В зависимости от вида такой функции модели делятся на линейные и нелинейные (как по независимым переменным, так и по неизвестным параметрам). Например, можно исследовать спрос на мороженое как линейную функцию от времени, температуры воздуха, среднего уровня доходов. Зависимость же заработной платы от возраста, пола, уровня образования, стажа работы может и не быть линейной.

Область применения моделей в виде одного регрессионного уравнения обширна. Проблемам теории оценивания неизвестных параметров модели, ее верификации, отбора значимых факторов и другим посвящен огромный объем

литературы [1 – 10]. Эта тема является, пожалуй, стержневой в эконометрике и основной в данном пособии.

2. Системы одновременных уравнений. Эти модели описываются системами уравнений. Системы могут состоять из тождеств и регрессионных уравнений, каждое из которых может, кроме объясняющих переменных, включать в себя также объясняемые переменные из других уравнений системы. Таким образом, мы имеем здесь набор объясняемых переменных, связанных через уравнения системы. Переменные, значения которых определяются из уравнений системы, называются эндогенными (внутренними), а переменные, значения которых определяются вне модели, называются экзогенными (предопределенными).

Примером может служить модель, представленная соотношениями (В.4) – (В.6), в которой национальный доход Y_t, измеренный в момент времени t, играет роль объясняющей переменной в уравнении (В.4) и объясняемой переменной в тождестве (В.6). В этой модели эндогенными переменными являются C_t, I_t, Y_t, а предопределенными переменными T_t, Y_{t-1}, R_t, G_t. Другим примером системы одновременных уравнений является, приведенная ниже, модель равновесия спроса и предложения на рынке некоторого товара.

Пример В.2. Модель равновесия спроса и предложения:

$$Q_t^S = a_1 + a_2 P_t + a_3 P_{t-1} + e_t \qquad \text{(предложение)},$$

$$Q_t^D = b_1 + b_2 P_t + b_3 Y_t + d_t \qquad \text{(спрос)},$$

$$Q_t^S = Q_t^D \qquad \text{(равновесие)}.$$

Здесь Q_t^D спрос на товар в момент времени t, Q_t^S – предложение товара в момент времени t, P_t – цена товара в момент времени t, Y_t – доход потребителей в момент времени t. Цена товара P_t и спрос на товар $Q_t = Q_t^D = Q_t^S$ определяются из уравнений модели, т. е. являются эндогенными переменными. Предопределенными переменными в данной модели являются доход Y_t и значение цены товара в предыдущий момент времени P_{t-1}.

Системы одновременных уравнений требуют относительно более сложный математический аппарат. Они могут использоваться для построения моделей макро- и микроэкономики, моделей страновой экономики и др.

3. Модели временных рядов. К этому классу относятся модели, построенные по данным, характеризующим изучаемый объект за ряд последовательных моментов или промежутков времени. В этих моделях объясняющим фактором в явной или неявной форме выступает время t. К этому классу относятся модели:

тренда: $y(t) = T(t) + e_t$,

где $T(t)$ – временной тренд заданного параметрического вида (например, линейный $T(t) = a + bt$, параболический $T(t) = a + bt + ct^2$) описывает основную тенденцию изменения во времени изучаемого признака, e_t – случайная (стохастическая) компонента;

сезонности: $y(t) = S(t) + e_t$,

где $S(t)$ – периодическая (сезонная) компонента характеризует периодические (сезонные) отклонения значений изучаемого признака от основной тенденции, e_t – случайная компонента;

тренда и сезонности:

$y(t) = T(t) + S(t) + e_t$ (аддитивная) или $y(t) = T(t) \cdot S(t) + e_t$ (мультипликативная), где $T(t)$ – временной тренд заданного параметрического вида, $S(t)$ – периодическая (сезонная) компонента, e_t – случайная компонента.

К моделям временных рядов относится множество более сложных моделей, таких как модели адаптивного прогноза, модели авторегрессии и скользящего среднего и др. Их общей чертой является то, что они объясняют поведение временного ряда, исходя только из его предыдущих значений. Такие модели могут применяться, например, для изучения и прогнозирования объема продаж авиабилетов, спроса на продукты питания, краткосрочного прогноза процентных ставок и т. п.

В.4. Типы данных и измерения в экономике

При построении эконометрических моделей могут использоваться следующие типы данных.

Пространственные данные представляют собой набор значений одних и тех же экономических показателей (объем производства, количество работников, доходы и расходы населения и др.) по однородным экономическим объектам (фирмам, предприятиям, домашним хозяйствам) в один и тот же момент времени (пространственный срез). К ним также относятся данные по курсам покупки / продажи наличной валюты в какой-либо день по обменным пунктам данного города, и т. д.

Временные ряды представляют наборы значений экономических показателей за последовательные моменты или промежутки времени. Примерами их могут быть ежеквартальные данные по инфляции, средней заработной плате, национальному доходу, денежной эмиссии за последние годы и т. д. Отличительной чертой временных данных является то, что они естественным образом упорядочены во времени, и наблюдения в близкие моменты времени часто бывают зависимыми.

Основной базой данных для эконометрических исследований служат данные официальной статистики либо данные бухгалтерского учета. Отсюда, проблемы измерений в экономике это проблемы статистики и учета. Именно в общей теории статистики обсуждаются вопросы о том, какие показатели применяются для измерения результатов работы промышленного предприятия, фирмы, отрасли; как оценить остатки оборотных средств и т.д. Подробное изложение подобных и других вопросов, связанных со шкалами измерений, точностью измерения и пр., выходит за рамки программы курса и данного пособия.

Контрольные вопросы

1. Охарактеризуйте предмет и основные задачи эконометрики.
2. Что является «фундаментом» эконометрики?
3. Каковы этапы эконометрического исследования? Какие проблемы приходится решать эконометристу?
4. В чем разница между экономическими и эконометрическими моделями?
5. Перечислите типы моделей, используемых в практических исследованиях.
6. Какие данные используются в эконометрическом моделировании?
7. Приведите собственные примеры постановок задач эконометрического моделирования.

1. Модели парной регрессии

В этой главе мы рассмотрим основные принципы построения моделей для двух переменных (линейных и нелинейных). Будут представлены основной метод оценивания неизвестных параметров – метод наименьших квадратов – с характеристикой свойств оценок и интерпретацией полученных результатов, а также способы верификации модели и прогноз на ее основе.

1.1. Модель парной линейной регрессии

Рассмотрим простейший случай, когда в исследование включены два экономических показателя, две переменных величины, одну из которых мы определим как зависимую, объясняемую величину и обозначим через y, другую – как независимую, объясняющую, оказывающую влияние на y. Эту переменную назовем фактором (или регрессором) и обозначим через x. Мы предполагаем также наличие связи между ними, простейшей формой которой будет линейная вида

$$y = a + bx, \tag{1.1}$$

где a и b – неизвестные параметры.

Действительно, взаимосвязи экономических переменных часто близки к линейным. Даже если некоторая зависимость, вообще говоря, не является линейной, часто она может быть приближенно описана как линейная в основном диапазоне наблюдаемых значений своих переменных.

Возможны и другие формы связи между переменными x и y:

$$y = ae^{bx}, \quad y = ax^{b}, \quad y = a + b\frac{1}{x}.$$

Третье из этих соотношений линейно относительно a и b (линейно относительно y и $\frac{1}{x}$), а первое и второе могут быть сведены к линейной форме для преобразованных переменных, если взять логарифмы от обеих частей

$$\ln y = \ln a + bx \quad \text{и} \quad \ln y = \ln a + b\ln x.$$

Если ввести $y' = \ln y$ и $x' = \ln x$, то мы получим линейную зависимость вида (1.1). Подробнее вопрос о построении таких моделей мы рассмотрим в п. 1.7.

Таким образом, в модели (1.1) a и b – постоянные, а x и y могут непосредственно или после логарифмических или иных преобразований представлять экономические переменные, например такие, как цены или спрос. Очевидно, что при таком подходе охватывается широкая область функциональных взаимосвязей между исходными экономическим переменными.

Задача построения модели (1.1) состоит в определении значений неизвестных параметров a и b – их оценок – по имеющимся в нашем распоряже-

нии данным так, чтобы полученное соотношение «наилучшим» образом описывало зависимость y от x. В каком смысле будет пониматься «наилучшее» приближение реально наблюдаемых данных к их теоретическим ожидаемым значениям мы рассмотрим в п. 1.2. Здесь же отметим, что в действительности, имея набор значений двух переменных x_i, y_i, $i=1,\mathbf{K},n$; и изображая пары (x_i, y_i) точками на координатной плоскости $X\,\mathrm{O}Y$ (рис. 1.1), мы имеем разброс этих точек относительно реальной линии связи.

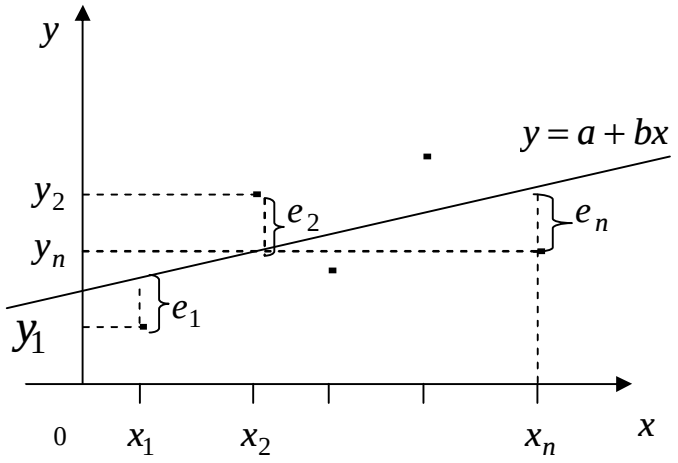


Рис. 1.1. Диаграмма рассеяния и теоретическая линия связи

Предположим, например, что мы изучаем зависимость между расходами на питание и доходом семей, используя данные о семейных бюджетах, относящиеся к некоторому фиксированному промежутку времени. Обозначим через y общую величину расходов на питание, а через x – объем распределяемого дохода. Соберем данные о бюджетах, допустим, $n=10000$ семей, и образуем пары соответствующих измерений для величин $x_i, y_i, i=1,\mathbf{K},10000$. Предположим, что мы уже разделили семьи на группы по их размеру и составу и рассматриваем интересующую нас связь между y и x внутри конкретной группы (условие «однородности» наблюдений). Естественно, мы не ожидаем, что у всех семей этой группы, имеющих одинаковый доход, будут и одинаковые потребительские расходы. Одни потратят больше других, а некоторые, наоборот, меньше. Однако можно надеяться, что величины расходов сгруппируются вокруг некоторого значения, соответствующего тому объему дохода, о котором шла речь. Эта идея находит свое формальное воплощение в новой гипотезе о характере линейной зависимости:

$$y = a + bx + e, \tag{1.2}$$

где e – случайная (или стохастическая) переменная, способная принимать и положительные, и отрицательные значения.

Таким образом, если мы рассмотрим подгруппу семей, располагающих доходом x_k, то средним значением их потребительских расходов окажется величина $a + bx_k$, в то время как реальные объемы потребления для семей в под-

группе будут $a + bx_k + e_k$, где случайная величина e_k измеряет отклонения потребительских расходов каждой отдельной семьи от среднего значения.

Запишем уравнение зависимости (1.2) для n наблюдений x_i, y_i:

$$y_i = a + bx_i + e_i \qquad i = 1, \mathbf{K}, n. \tag{1.3}$$

Здесь x_i – неслучайная (детерминированная) величина, а y_i, e_i – случайные величины; y_i – объясняемая (зависимая) переменная, x_i – объясняющая (независимая) переменная, фактор или регрессор. Уравнение (1.3) называется также регрессионным уравнением или линейной регрессионной моделью с двумя переменными (моделью парной линейной регрессии).

Какова природа случайной составляющей или ошибки e_i? Источниками ошибок могут быть разные причины:

1. Пропущенные объясняющие переменные. Соотношение между y и x почти наверняка является очень большим упрощением. В действительности существуют другие факторы, также влияющие на y, которые не учтены в формуле (1.1). Влияние этих факторов приводит к тому, что наблюдаемые точки лежат вне прямой (см. рис. 1.1). Часто возникают ситуации, когда мы не включаем в регрессионное уравнение переменные, только потому, что не знаем, как их измерить, например психологические факторы. Либо существуют также другие факторы, которые мы можем измерить, но которые оказывают такое слабое влияние, что их не стоит учитывать. Объединив все эти составляющие, мы и получаем то, что обозначено через e.

2. Агрегирование переменных. Во многих случаях рассматриваемая зависимость – это попытка объединить вместе некоторое число микроэкономических соотношений. Например, функция суммарного потребления – это попытка общего выражения решений многих отдельных семей о расходах. Так как отдельные соотношения, вероятно, имеют разные параметры, любая попытка определить соотношение между совокупными потребительскими расходами и доходом является лишь аппроксимацией, наблюдаемое расхождение при этом приписывается наличию случайной составляющей.

3. Неправильное описание структуры модели. Структура модели может быть описана неправильно или не вполне правильно. Например, если зависимость относится к данным о временном ряде, то значение y может зависеть не от фактического значения x, а от значения, которое ожидалось в предыдущем периоде. Если ожидаемое и фактическое значения тесно связаны, то будет казаться, что между y и x существует зависимость, но это будет лишь аппроксимация, и расхождение вновь будет связано с наличием случайной величины e.

4. Неправильная функциональная спецификация. Функциональное соотношение между y и x математически может быть определено неправильно, т. е. сам вид функциональной зависимости выбран неверно. Например, мы рассматриваем зависимость между потребительскими расходами и доходом семей,

используя линейную функцию, а истинная зависимость может быть более сложной, нелинейной.

5. Ошибки измерения. Ошибки могут сопровождать любые наблюдения или измерения экономических показателей. Например, данные по расходам семьи на питание составляются на основании записей участников опросов, которые, как предполагается, тщательно фиксируют свои ежедневные расходы. Разумеется, при этом возможны ошибки. В данном случае источниками ошибок являются особенности собранного материала (присущ элемент случайности).

Таким образом, можно считать, что случайные величины e_i являются суммарным проявлением всех этих факторов.

Сформулируем теперь те основные предпосылки или гипотезы, которые лежат в основе линейной регрессионной модели с двумя переменными.

Основные гипотезы:

1. $y_i = a + bx_i + e_i, i = 1, \mathbf{K}, n,\ n > 2$, – спецификация модели.

2. $x_1, \mathbf{K}, x_n$ – детерминированные величины, линейно не связанные между собой т.е. вектор $(x_1, \mathbf{K}, x_n)^T$ не коллинеарен вектору $(1, \mathbf{K}, 1)^T$.

3. $e_1, \mathbf{K}, e_n$ – случайные величины, для которых

3a. $Me_i = 0,\ M\left(e_i^2\right) = D(e_i) = s^2$ – не зависит от i.

3b. $M\left(e_i e_j\right) = 0$ при $i \neq j$, т.е. $e_1, e_2, \mathbf{K}, e_n$ – некоррелированны для разных наблюдений.

Часто добавляется условие

3c. $e_i \sim N\left(0, s^2\right)$, т. е. e_i – нормально распределенные случайные величины с математическим ожиданием или средним значением, равным нулю, и дисперсией s^2.

Гипотезы 1-3c определяют нормальную линейную модель парной регрессии. Для такой модели условие 3b. эквивалентно условию статистической независимости ошибок e_i, e_j при $i \neq j$.

Обсудим предпосылки или гипотезы, лежащие в основе построения такой модели.

1. Спецификация модели отражает наше представление о механизме зависимости y_i от x_i и сам выбор объясняющей переменной x; на линейный характер связи может указывать и разброс точек на диаграмме рассеивания.

2. Величины $x_1, \mathbf{K}, x_n$ являются неслучайными или детерминированными, линейно не связанными между собой. Если же в реальной ситуации их значения также представляются результатами измерений, то предполагается, что ошибки таких измерений пренебрежимо малы.

3a. Условие $M(e_i) = 0$ означает отсутствие систематических ошибок, ошибки носят только случайный характер. Условие независимости дисперсий ошибок от номера наблюдений $M\left(e_i^2\right) = D(e_i) = s^2,\ i = 1, \mathbf{K}, n$, или однородности

наблюдений называется также гомоскедастичностью; случай, когда $M\left(e_i^2\right)= s_i^2$, т. е. условие гомоскедастичности не выполняется. называется гетероскедастичностью. Ниже на рис. 1.2 приведен пример типичного разброса точек для случая гомоскедастичности ошибок; на рис. 1.3 – пример данных с гетероскедастичными ошибками.

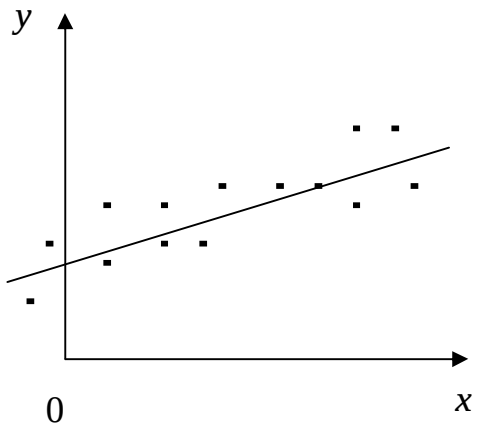


Рис. 1.2. Однородные наблюдения ($Me_i^2 = s^2,\ i = 1,\mathbf{K},n$)

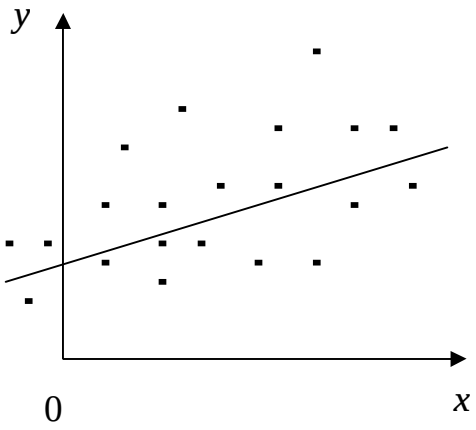


Рис. 1.3. Неоднородные наблюдения ($Me_i^2 = s_i^2, i = 1,\mathbf{K},n$)

3b. Условие $M\left(e_i e_j\right)=0,\ i \neq j$, указывает на некоррелированность ошибок, а в случае нормальной модели, и на независимость для разных наблюдений. Это требование оказывается вполне естественным в широком классе реальных ситуаций, особенно, если речь идет о пространственных данных (значения анализируемых переменных регистрируются на различных объектах: индивидуумах, семьях, предприятиях, банках, регионах и т. п.). Однако условие часто нарушается, когда наши данные являются временными рядами. В случае, когда это условие не выполняется, говорят об автокорреляции ошибок.

3с. Так как можно считать, что случайная составляющая e_i в различных наблюдениях обусловлена суммарным аддитивным эффектом большого числа независимых случайных факторов, ни один из которых не является доминирующим, то обращение к центральной предельной теореме теории вероятностей служит достаточным обоснованием выбора нормального распределения для нее.

1.2. Оценивание неизвестных параметров модели: метод наименьших квадратов

Рассмотрим задачу «наилучшей» аппроксимации набора наблюдений (x_i, y_i), $i = 1, \mathbf{K}, n$, линейной функцией $y = a + bx$ в смысле минимизации величины

$$R = \sum_{i=1}^{n} \left(y_i - a - bx_i \right)^2 . \tag{1.4}$$

Нахождение оценок $\hat{a}$ и $\hat{b}$ в соответствии с этим условием называется методом наименьших квадратов (МНК). Запишем необходимые условия экстремума:

$$\begin{cases} \dfrac{\partial R}{\partial a} = -2 \sum_{i=1}^{n} \left(y_i - a - bx_i \right) = 0 \\ \dfrac{\partial R}{\partial b} = -2 \sum_{i=1}^{n} \left(y_i - a - bx_i \right) x_i = 0. \end{cases}$$

Решение этой системы нормальных уравнений дает нам явный вид оценок

$$\begin{aligned} \hat{a} &= \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2} \\ \hat{b} &= \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i \right)^2} \end{aligned} \tag{1.5}$$

(для краткости индексы суммирования у знака суммы $\sum$ опущены).

Если $\hat{b}$ найдено по формуле (1.5), то $\hat{a} = \bar{y} - \hat{b}\bar{x}$, где $\bar{x} = \frac{1}{n} \sum x_i$, $\bar{y} = \frac{1}{n} \sum y_i$.

Уравнение прямой линии $y = \hat{a} + \hat{b}x$, полученное в результате минимизации величины (1.4), проходит через точку $(\bar{x}, \bar{y})$. Единственность МНК-оценок (1.5) обеспечивается предпосылкой 2.

Из общей теории МНК при сделанных выше предпосылках 3a, 3b следуют свойства МНК-оценок: 1) линейная зависимость от y, 2) несмещенность, 3) эффективность, поскольку в классе линейных несмещенных оценок МНК-оценки обладают наименьшей возможной дисперсией (теорема Гаусса-Маркова [1; 4]).

Несмещенные оценки дисперсий и ковариаций оценок $\hat{a}$ и $\hat{b}$ определяются по формулам

$$\hat{D}(\hat{a}) = \frac{\sum x_i^2}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2}; \tag{1.6}$$

$$\hat{D}(\hat{b}) = \frac{n}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2}; \tag{1.7}$$

$$\operatorname{cov}(\hat{a},\hat{b}) = \frac{-\sum x_i}{n\sum x_i^2 - \left(\sum x_i\right)^2} \cdot \frac{R_{\min}}{n-2}, \tag{1.8}$$

где $R_{\min} = \sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right)^2$ – остаточная сумма квадратов и под $\hat{a}$, $\hat{b}$ понимаются их значения, найденные по формулам (1.5).

Несмещенной оценкой дисперсии ошибок наблюдений будет $S^2 = \hat{s}^2 = \frac{R_{\min}}{n-2}$.

Остатки регрессии e_i определяются из уравнения

$$y_i = \hat{y}_i + e_i = \hat{a} + \hat{b}x_i + e_i.$$

Не следует путать остатки регрессии с ошибками регрессии в уравнении модели $y_i = a + bx_i + e_i$. Разница состоит в том, что остатки e_i в отличие от ошибок e_i вычисляются. С учетом введенного обозначения для остатков можно записать несмещенную оценку дисперсии s^2:

$$S^2 = \hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2.$$

Пример использования формул (1.5) – (1.7) мы рассмотрим ниже в п. 1.6 с тем, чтобы проиллюстрировать все этапы построения и анализа линейной модели и задачу прогнозирования на ее основе.

1.3. Доверительные интервалы для коэффициентов регрессии. Проверка гипотез

С помощью формул (1.5) мы можем получить по данным наблюдений над величинами x, y лишь оценки неизвестных параметров линейной модели. Поэтому возникает вопрос о точности и надежности найденных оценок. В математической статистике этот вопрос решается построением доверительных интервалов для истинных значений параметров, которые по сути представляют собой

множество всех возможных гипотетических значений, не противоречащих результатам экспериментов.

Если выполнено условие 3с. нормальной линейной регрессионной модели, т. е. $e_i \sim N(0, s^2)$, $i = 1, \mathbf{K}, n$, то y_i будут также нормально распределены. Отсюда и МНК-оценки коэффициентов регрессии $\hat{a}$ и $\hat{b}$ имеют совместное нормальное распределение как линейные функции от y_i.

Если гипотеза нормальности ошибок не выполняется, то нормальность оценок, вообще говоря, неверна. Однако при некоторых условиях регулярности на поведение x_i при росте n, оценки $\hat{a}$ и $\hat{b}$ имеют асимптотически нормальное распределение, т. е. $\hat{a} \sim N(a, \hat{D}(\hat{a}))$, $\hat{b} \sim N(b, \hat{D}(\hat{b}))$ при $n \to \infty$.

В этих условиях справедливы формулы интервальных оценок или доверительных интервалов:

$$\hat{a} - t_g \sqrt{\hat{D}(\hat{a})} < a < \hat{a} + t_g \sqrt{\hat{D}(\hat{a})}, \tag{1.9}$$

$$\hat{b} - t_g \sqrt{\hat{D}(\hat{b})} < b < \hat{b} + t_g \sqrt{\hat{D}(\hat{b})}, \tag{1.10}$$

где $t_g = t\left(\frac{1+g}{2}, n-2\right)$ – квантиль t-распределения (распределения Стьюдента) уровня $\frac{1+g}{2}$ и числа степеней свободы $n-2$. Здесь g – доверительная вероятность или надежность:

$$P\left(\hat{a} - t_g \sqrt{\hat{D}(\hat{a})} < a < \hat{a} + t_g \sqrt{\hat{D}(\hat{a})}\right) = g,$$

это вероятность того, что построенный нами доверительный интервал покроет истинное значение параметра a. Аналогично можно определить g и для параметра b. Обычно значения доверительной вероятности стандартизованы и принимаются равными 0,9; 0,95; 0,99; 0,999.

Доверительный интервал для неизвестной дисперсии ошибок наблюдений s^2:

$$\frac{(n-2)S^2}{u_2} < s^2 < \frac{(n-2)S^2}{u_1}, \tag{1.11}$$

где $u_1 = c^2\left(\frac{1-g}{2}, n-2\right)$ и $u_2 = c^2\left(\frac{1+g}{2}, n-2\right)$ – квантили c^2-распределения.

При статистическом исследовании реальной ситуации возникает необходимость не только оценить неизвестные параметры модели, но и проверить по отношению к ним некоторые гипотезы. Например, можно ли считать потребление пропорционально зависящим от дохода $(a = 0)$? Будет ли предельная склонность к потреблению больше половины $\left(b > \frac{1}{2}\right)$? И, наконец, служит ли линейная зависимость адекватным отражением эмпирических данных?

Статистики, которые использовались для построения доверительных интервалов, могут использоваться и для проверки гипотез о параметрах модели.

Так, для проверки гипотезы $H_0 : a = a_0$ против альтернативной гипотезы $H_1 : a \neq a_0$ используется статистика

$$t = \frac{\hat{a} - a_0}{\sqrt{\hat{D}(\hat{a})}} = \frac{\hat{a} - a_0}{S_{\hat{a}}} \sim t(n-2), \qquad (1.12)$$

распределенная по закону Стьюдента с $(n-2)$ степенями свободы.

Аналогично для гипотезы $H_0 : b = b_0$ и $H_1 : b \neq b_0$ используется критерий, статистика которого

$$t = \frac{\hat{b} - b_0}{\sqrt{\hat{D}(\hat{b})}} = \frac{\hat{b} - b_0}{S_{\hat{b}}} \sim t(n-2). \qquad (1.13)$$

Мы отвергаем гипотезу H_0 (и принимаем H_1) с уровнем значимости $a = 1 - g$, если $|t_0| > t_{\frac{1+g}{2}}$ (или $|t_0| > t_{1-\frac{a}{2}}$), t_0 – наблюдаемое или экспериментальное значение t-статистики, в противном случае гипотезу H_0 следует принять, т. е. считать, что результаты наблюдений согласуются с гипотезой H_0, не противоречат ей.

Для такого вида альтернативной гипотезы H_1 область принятия H_0 совпадает с доверительным интервалом для соответствующего неизвестного параметра: гипотеза H_0 принимается на уровне значимости a, если построенный доверительный интервал для a (или b) в форме (1.9) (или (1.10)) покрывает гипотетическое значение параметра a_0 (или b_0).

Если $a_0 = 0$ и $b_0 = 0$, речь идет о проверке гипотез о статистической незначимости параметров модели. Фактор, коэффициент при котором в линейной модели статистически незначим, оказывает несущественное влияние на y и должен быть исключен из модели.

Для проверки гипотезы $H_0 : s^2 = s_0^2$, против $H_1 : s^2 \neq s_0^2$ может использоваться доверительный интервал (1.11). Гипотезу H_0 принимаем с уровнем $a = 1 - g$, если интервал покрывает значение s_0^2.

При использовании современных статистических пакетов программ не требуется искать нужные квантили t-распределения (или c^2-распределения), поскольку в них (пакетах) рассчитывается уровень ошибки, с которой можно отвергнуть нулевую гипотезу и , если он меньше желаемого значения, либо равен ему, то нулевая гипотеза отвергается.

1.4. Верификация модели

Пригодность построенной модели $\hat{y} = \hat{a} + \hat{b}x$ или ее верификация, а также качество оценивания регрессии может быть проверено двумя равноценными

способами: дисперсионным анализом в регрессии и с использованием элементов теории корреляции.

1. Дисперсионный анализ в регрессии

Суть метода заключается в разложении общей суммарной дисперсии выходной величины y на составляющие, обусловленные действием входных переменных-факторов, и остаточную дисперсию, обусловленную ошибкой или всеми неучтенными в данной модели переменными. Фактор оказывает несущественное влияние на y, если соответствующая ему дисперсия и дисперсия ошибок статистически незначимы. Для проверки гипотезы о равенстве таких дисперсий используется критерий Фишера (F-критерий). Поскольку для оценок дисперсий используются суммы квадратов SS (от англ. sum of squares) отклонений значений данной переменной от ее средней величины, то можно говорить о разложении общей суммы квадратов $SS_{общ.}$ на составляющие. Этой идеи мы и будем придерживаться далее.

Рассмотрим $SS_{общ.} = \Sigma(y_i - \bar{y})^2$ – величину, характеризующую разброс значений y_i относительно среднего значения $\bar{y}$. Разобьем эту сумму на две части: объясненную регрессионным уравнением и не объясненную (т. е. связанную с ошибками e_i).

Обозначим через $\hat{y}_i = \hat{a} + \hat{b}x_i$ предсказанное по модели значение y_i, тогда $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$ (см. рис. 1.4).

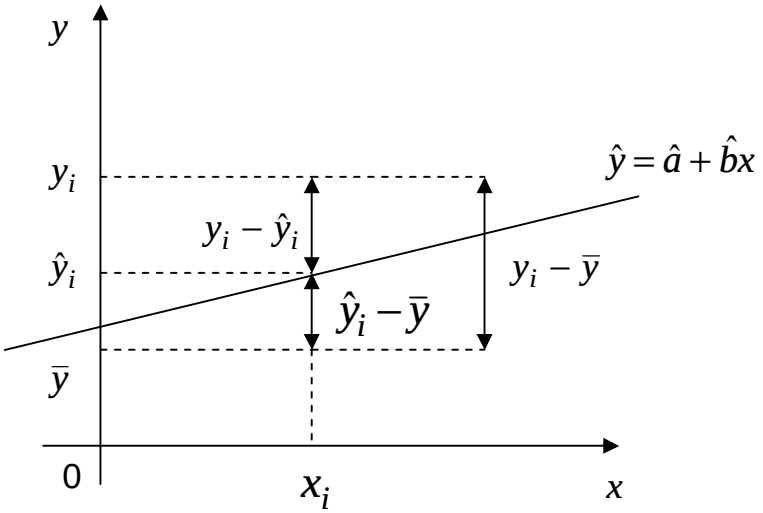


Рис. 1.4

Тогда $SS_{общ.}$ представляется в виде суммы трех слагаемых: $SS_{общ.} = \Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2 + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$. Здесь $\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$, так как $\Sigma e_i = \Sigma(y_i - \hat{a} - \hat{b}x_i) = 0$, $\Sigma(y_i - \hat{a} - \hat{b}x_i)x_i = \Sigma e_i x_i = 0$.

Действительно, $\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \Sigma e_i\left(\hat{a} + \hat{b}x_i - \bar{y}\right) = (\hat{a} - \bar{y})\Sigma e_i + \hat{b}\Sigma e_i x_i = 0$. Поэтому справедливо равенство

$$\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2;$$

$$SS_{общ.} = SS_R + SS_{ост.}. \qquad (1.14)$$

Здесь через $SS_R = \Sigma(\hat{y}_i - \bar{y})^2$ обозначена сумма квадратов, объясненная регрессией, и $SS_{ост.} = \Sigma(y_i - \hat{y}_i)^2$ – остаточная сумма квадратов, обусловленная ошибкой.

Коэффициентом детерминации, или долей объясненной дисперсии y, называется

$$R^2 = 1 - \frac{SS_{ост.}}{SS_{общ.}} = \frac{SS_R}{SS_{общ.}}. \qquad (1.15)$$

В силу определения $0 \le R^2 \le 1$. Если $R^2 = 0$, то это значит, что регрессия ничего не дает, т. е. фактор x не улучшает качество предсказания y_i по сравнению с тривиальным предсказанием $\hat{y}_i = \bar{y}$.

Другой крайний случай $R^2 = 1$ означает точную подгонку: все наблюдаемые значения (x_i, y_i) лежат на регрессионной прямой (все остатки $e_i = 0$).

Чем ближе к 1 значение R^2, тем лучше качество подгонки или качество регрессии, $\hat{y}$ более точно аппроксимирует y.

Гипотеза об отсутствии линейной функциональной связи между x и y может быть записана как $H_0: b = 0$. Критерий, статистика которого (1.13) распределена по закону Стьюдента, эквивалентен здесь критерию, статистика которого

$$F = \frac{MS_R}{MS_{ост.}} = \frac{SS_R/1}{SS_{ост.}/(n-2)} \sim F(1, n-2) \qquad (1.16)$$

распределена по закону Фишера со степенями свободы $(1, n-2)$.

То есть проверка гипотезы $H_0: b = 0$ с использованием t- и F-статистик дает в данном случае (для одномерной регрессионной модели) тождественные результаты.

Здесь через MS_R и $MS_{ост.}$ обозначены средние квадраты (от англ. mean of squares), которые дают несмещенные оценки соответствующих теоретических дисперсий.

Используя коэффициент детерминации (1.15), получим для F-статистики

$$F = (n-2)\frac{R^2}{1-R^2}. \qquad (1.17)$$

Вычисления, необходимые для дисперсионного анализа уравнения регрессии, обычно сводят в таблицу (табл. 1.1).

Таблица 1.1

Дисперсионный анализ парной регрессии

Источник дисперсии	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F	Критическая точка $F_{кр.} = F(a;1,n-2)$	Гипотеза $H_0: b=0$
Регрессор x	1	SS_R	$MS_R = \frac{SS_R}{1}$	$F = \frac{MS_R}{MS_{ост.}}$		
Ошибка	$n-2$	$SS_{ост.} = SS_{общ} - SS_R$	$MS_{ост.} = \frac{SS_{ост.}}{n-2}$	–	–	–
Общая дисперсия (итог)	$n-1$	$SS_{общ.}$	–	–	–	–

Если при заданном уровне значимости a наблюдаемое значение F-статистики больше критической точки $F_0 > F(a;1,n-2)$, то гипотеза $H_0: b=0$ отвергается, то есть связь между x и y есть, и результаты наблюдений не противоречат предположению о ее линейности. В противном случае $H_0: b=0$ принимается и постулируется отсутствие значимой линейной функциональной связи между x и y. Исходя из соотношения (1.16), малым значениям F-статистики будут соответствовать и малые значения коэффициента детерминации R^2 (плохая аппроксимация данных).

2. Использование элементов теории корреляции

Другой способ верификации линейной модели состоит в использовании элементов теории корреляции. Мерой линейной связи двух величин является коэффициент корреляции, выборочное значение которого

$$r_B = \hat{r} = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{[n\Sigma x_i^2 - (\Sigma x_i)^2][n\Sigma y_i^2 - (\Sigma y_i)^2]}} \qquad (1.18)$$

будет его несмещенной оценкой.

Значения коэффициента корреляции принадлежат промежутку $[-1;1]$. Чем больше его абсолютное значение к 1, тем теснее связь между признаками. Положительная величина коэффициента корреляции свидетельствует о прямой связи между ними, отрицательная – о наличии обратной связи между признаками.

Гипотеза об отсутствии линейной функциональной связи между x и y может быть записана как $H_0 : r = 0$. Для проверки H_0 используется критерий, статистика которого

$$t = \frac{r_B \sqrt{n-2}}{\sqrt{1-r_B^2}} \sim t(n-2) \qquad (1.19)$$

распределена по закону Стьюдента с $(n-2)$ степенями свободы.

Вывод о значимости корреляции между x и y может быть сделан, если $|t_0| > t_{1-\frac{a}{2}}$, где $t_{1-\frac{a}{2}} = t\left(1-\frac{a}{2}, n-2\right)$ – квантиль t – распределения, a – уровень значимости.

Здесь также вычисляется коэффициент детерминации $R^2 = r_B^2$ (чаще всего выражаемый в %). Он равен, как уже отмечалось, той доле дисперсии y, которая объяснена линейной зависимостью от x. Если $r_B = 0{,}9$,то это значит, что линейная регрессия y на x объясняет 81% дисперсии y. Остальные 19% приходятся на долю прочих факторов, не учтенных в уравнении регрессии.

1.5. Интерпретация уравнения регрессии

Существуют два этапа интерпретации уравнения регрессии. Первый этап состоит в словесном толковании уравнения так, чтобы это было понятно человеку, не являющемуся специалистом в области статистики или эконометрики. На втором этапе необходимо решить, следует ли ограничиться этим или провести более детальное исследование зависимости, например, проверить по отношению к исследуемым переменным некоторые статистические гипотезы, либо улучшить качество и предсказательные свойства модели.

Представим простой способ интерпретации коэффициентов линейного уравнения регрессии $\hat{y} = \hat{a} + \hat{b}x$, когда y и x – переменные с простыми, естественными единицами измерения.

Во-первых, можно сказать, что увеличение x на одну единицу измерения приведет к увеличению y в среднем на $\hat{b}$ единиц (в единицах измерения и переменной x и переменной y). Здесь коэффициент регрессии $\hat{b}$ есть абсолютный показатель силы связи, характеризующий среднее абсолютное изменение результата y при изменении фактора x на единицу своего измерения. Вторым шагом является проверка, каковы действительно единицы измерения x и y, и замена слова «единица» фактическим количеством.

Постоянная $\hat{a}$ дает прогнозируемое значение y (в единицах y), если $x = 0$. Это может иметь или не иметь реального смысла в зависимости от конкретной ситуации. Если $x = 0$ находится достаточно далеко от выборочных значений переменной x, то буквальная интерпретация может привести к не-

верным результатам; даже если линия регрессии довольно точно описывает значения наблюдаемой выборки, мы не можем гарантировать, что это ее свойство сохранится при экстраполяции влево или вправо. В случае, когда интерпретация $\hat{a}$ не имеет никакого смысла, эта константа выполняет единственную функцию: она позволяет определить положение линии регрессии на графике.

При интерпретации уравнения регрессии важно помнить о трех вещах. Во-первых, $\hat{a}$ является лишь оценкой a, а $\hat{b}$ – оценкой параметра b. Поэтому вся интерпретация в действительности представляет собой лишь оценку. Во-вторых, уравнение регрессии отражает только общую тенденцию для выборки. При этом каждое отдельное наблюдение подвержено воздействию случайностей. В-третьих, верность интерпретации зависит от правильности спецификации уравнения.

В заключение, обратим внимание на то, что для линейного уравнения $y = a + bx$ эластичность $E = f'(x)\dfrac{x}{y} = \dfrac{bx}{y}$. Поэтому при интерпретации уравнения регрессии значение эластичности в любой точке будет зависеть не только от значения $\hat{b}$, но также и от значений y и x в данной точке.

1.6. Прогноз на основе линейной модели

Построенная адекватная модель может использоваться для прогнозирования. Оценка прогнозируемых величин в регрессионном анализе получается подстановкой в регрессию значений независимых переменных. Таким образом, прогноз на основе уравнения регрессии является условным типа: «если независимые переменные равны таким-то величинам, то зависимая переменная составит такую-то величину».

Рассмотрим подробнее задачу прогноза на основе линейной модели. Предположим, что мы хотим распространить нашу модель, содержащую две переменные, на другие значения независимой переменной и поставить проблему прогнозирования среднего значения y, соответствующего некоторому данному значению x_0, которое может лежать как между выборочными наблюдениями от x_1 до x_n, так и вне соответствующего интервала. Наш прогноз может быть точечным или интервальным.

В случае точечного прогноза мы определяем

$$y_0 = a + bx_0. \qquad (1.20)$$

Мы не останавливаемся здесь на доказательстве того, что наилучшей несмещенной линейной оценкой для (1.20) будет $\hat{y}_0 = \hat{a} + \hat{b}x_0$, где $\hat{a}$ и $\hat{b}$ – МНК-оценки (1.5). Обоснование этого факта можно найти, например, в [1].

Итак, $M(\hat{y}_0 / x_0) = y_0 = a + bx_0$.

Можно показать, что дисперсия величины $\hat{y}_0$ определится как

$$D(\hat{y}_0) = s^2\left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\Sigma(x_i - \bar{x})^2}\right]. \qquad (1.21)$$

Отсюда видно, что дисперсия прогноза возрастает по мере удаления значения x_0 от среднего $\bar{x}$, использованного для расчета $\hat{a}$ и $\hat{b}$.

Подставляя в (1.21) вместо s^2 ее несмещенную оценку $\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$, мы получим оценку дисперсии прогнозируемого значения $\hat{D}(\hat{y}_0)$. Тогда доверительный интервал для прогностического значения y_0 можно записать в виде

$$\left(\hat{a} + \hat{b}x_0\right) \pm t_g\,\hat{s}\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\Sigma(x_i - \bar{x})^2}}, \qquad (1.22)$$

где $t_g = t\left(\frac{1+g}{2}, n-2\right)$, g – доверительная вероятность.

Построением доверительного интервала решается задача интервального прогноза.

Рассмотрим пример построения линейной регрессии, статистического анализа полученных результатов и прогноза по модели.

Пример 1.1. Исследуем зависимость розничного товарооборота (млн. руб.) магазинов от среднесписочного числа работников. Товарооборот как результирующую объясняемую переменную обозначим через y, а среднесписочное число работников (чел.) как независимую объясняющую переменную (фактор или регрессор) – через x. На объем товарооборота оказывают также влияние такие факторы, как объем основных фондов, их структура, площади торговых залов и подсобных помещений, расположение магазинов по отношению к потокам покупателей, сезонность и др. Предположим, что в исследуемой группе магазинов значения этих последних факторов примерно одинаковы, поэтому влияние различия их значений на изменение объема товарооборота оказывается незначительно. То есть можно считать, что анализ зависимости производится в условиях гомоскедастичности ошибок наблюдений.

В табл. 1.2. во втором и третьем столбцах приведены значения соответственно объемов розничного товарооборота и среднесписочного числа работников, а в следующих столбцах – значения расчетных величин, необходимых для определения оценок параметров линейной регрессии, их дисперсий и дисперсии случайной составляющей, а также статистик, необходимых для верификации модели.

Используем итоги столбцов 4-6 для определения оценок параметров модели по формулам (1.5):

$$\hat{a} = \frac{106508 \cdot 9{,}6 - 904 \cdot 1168{,}6}{8 \cdot 106508 - 904^2} = -0{,}974;$$
$$\hat{b} = \frac{8 \cdot 1168{,}6 - 904 \cdot 9{,}6}{8 \cdot 106508 - 904^2} = 0{,}01924. \qquad (1.23)$$

Таким образом, уравнение линии регрессии, найденное по результатам наблюдений, можно записать в виде

$$\hat{y} = -0{,}974 + 0{,}01924x.$$

Подставляя в это уравнение значения x_i, определим $\hat{y}_i$ (выровненные или вычисленные по модели значения результирующего признака). Заполнив столбцы 7-9 и определив итоги по этим столбцам, найдем оценку дисперсии случайной составляющей или ошибки s^2:

$$\hat{s}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{0{,}047}{6} \approx 0{,}008.$$

Таблица 1.2

Результаты наблюдений и необходимые расчеты для построения модели парной линейной регрессии

Порядковый номер магазина	x_i	y_i	x_i^2	y_i^2	$x_i y_i$	$\hat{y}_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	2	3	4	5	6	7	8	9
1	73	0,5	5 329	0,25	36,5	0,43	0,07	0,0049
2	85	0,7	7 225	0,49	59,5	0,661	0,039	0,0015
3	102	0,9	10 404	0,81	91,8	0,998	-0,088	0,0077
4	115	1,1	13 225	1,21	126,5	1,239	-0,139	0,0193
5	122	1,4	14 884	1,96	170,8	1,373	0,027	0,0007
6	126	1,4	15 876	1,96	176,4	1,45	-0,05	0,0025
7	134	1,7	17 956	2,89	227,8	1,604	0,096	0,0092
8	147	1,9	21 609	3,61	279,3	1,854	0,046	0,0021
Итог	904	9,6	106 508	13,18	1168,6	9,609	0,001	0,0479

Для построения интервальных оценок параметров регрессии найдем по формулам (1.6), (1.7) оценки дисперсий точечных оценок этих параметров:

$$\hat{D}(\hat{a}) = \frac{106508}{34846} \cdot 0{,}008 \approx 0{,}0244; \quad \sqrt{\hat{D}(\hat{a})} \approx 0{,}156.$$

$$\hat{D}(\hat{b}) = \frac{8}{34848} \cdot 0{,}008 \approx 0{,}0000018;\ \sqrt{\hat{D}(\hat{b})} \approx 0{,}0013.$$

При доверительной вероятности $g = 0{,}95$ получаем $t_{0{,}95} = t(0{,}975;6) = 2{,}447$ (см. приложение 2). Согласно формулам (1.9), (1.10) имеем

$$-0{,}974 - 2{,}447 \cdot 0{,}156 < a < -0{,}974 + 2{,}447 \cdot 0{,}156\ ,$$
$$0{,}01924 - 2{,}447 \cdot 0{,}0013 < b < 0{,}01924 + 2{,}447 \cdot 0{,}0013$$

или

$$\begin{gathered} -1{,}3557 < a < -0{,}5923 \\ 0{,}016 < b < 0{,}022. \end{gathered} \qquad (1.24)$$

Итак, наилучшие точечные оценки неизвестных параметров, соответствующие результатам наблюдений, определяются (1.23), и все их возможные значения, выходящие за пределы интервалов (1.24), маловероятны.

Проверка значимости коэффициентов регрессии – это проверка гипотез $H_0 : a = 0$ и $H_0 : b = 0$ при альтернативных $H_1 : a \neq 0$ и $H_1 : b \neq 0$. Наблюдаемые значения t-статистик, вычисленные по формулам (1.12), (1.13), для этих гипотез равны соответственно $t_0 = \dfrac{-0{,}974}{0{,}156} = -6{,}244$ и $t_0 = \dfrac{0{,}01924}{0{,}0013} = 14{,}8$. Критическая точка для 5% уровня значимости и числа степеней свободы $n - 2 = 6$ равна $t_{кр.} = 2{,}447$. Так как в обоих случаях $|t_0| > t_{кр.}$, то гипотезы о незначимости коэффициентов регрессии следует отвергнуть, т. е. считать, что как среднесписочное число работников, так и другие не учтенные в модели факторы существенно влияют на объем розничного товарооборота.

Проверку гипотез H_0 здесь можно было бы провести и с использованием построенных доверительных интервалов для параметров модели: интервальные оценки этих параметров есть области принятия нулевых гипотез. Так как интервальные оценки теоретических коэффициентов регрессии не содержат гипотетических значений, равных нулю, то гипотезы H_0 в том и в другом случаях следует отвергнуть.

Верификацию модели осуществим вначале дисперсионным анализом в регрессии.

Для расчета сумм квадратов $SS_{общ.}$, SS_R и $SS_{ост.}$, составим вспомогательную таблицу (табл. 1.3), имея в виду, что $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i = \dfrac{1}{8} \cdot 9{,}6 = 1{,}2$.

Итак, $SS_{общ.} = \sum_{i=1}^{n} (y_i - \bar{y})^2 = 1{,}66$, $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = 1{,}6091$ и $SS_{ост.} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = 0{,}0479$. Очевидно, здесь $1{,}66 \approx 1{,}6091 + 0{,}0479$ (мы выну-

ждены поставить знак приближенного равенства из-за погрешностей округления).

Вычисления, необходимые для дисперсионного анализа, сведем в таблицу (табл. 1.4).

Таблица 1.3

Расчет сумм квадратов

Порядковый номер магазина	y_i	$\hat{y}_i$	$y_i - \bar{y}$	$\hat{y}_i - \bar{y}$	$y_i - \hat{y}_i$	$(y_i - \bar{y})^2$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \hat{y}_i)^2$
1	2	3	4	5	6	7	8	9
1	0,5	0,43	-0,7	-0,77	0,07	0,49	0,5929	0,0049
2	0,7	0,661	-0,5	-0,539	0,039	0,25	0,2905	0,0015
3	0,9	0,998	-0,3	-0,202	-0,088	0,09	0,0408	0,0077
4	1,1	1,239	-0,1	0,039	-0,139	0,01	0,0015	0,0193
5	1,4	1,373	0,2	0,173	0,027	0,04	0,030	0,0007
6	1,4	1,45	0,2	0,25	-0,05	0,04	0,0625	0,0025
7	1,7	1,604	0,5	0,404	0,096	0,25	0,1632	0,0092
8	1,9	1,854	0,7	0,654	0,046	0,49	0,4277	0,0021
Итог	9,6	9,609	0,0	0,009	0,001	1,66	1,6091	0,0479

Так как $F_0 > F(0{,}05;1{,}6)$ $(201{,}13 > 5{,}99)$, то гипотеза $H_0 : b = 0$ отвергается, т. е. результаты наблюдений не противоречат предположению о наличии связи и о ее линейности.

Коэффициент детерминации $R^2 = \dfrac{SS_R}{SS_{общ.}} = 0{,}97$ показывает, что в исследуемой ситуации 97% общей дисперсии розничного товарооборота объясняется изменениями числа работников, в то время как на все остальные факторы приходится лишь 3% изменчивости товарооборота.

Найдем коэффициент корреляции. Используя формулу (1.18), получим

$$r_в = \frac{8 \cdot 1168{,}6 - 904 \cdot 9{,}6}{\sqrt{\left[8 \cdot 106508 - 904^2\right]\left[8 \cdot 13{,}18 - 9{,}6^2\right]}} \approx 0{,}985.$$

Таблица 1.4

Дисперсионный анализ в регрессии

Источник дисперсии	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F	Критическая точка	Гипотеза H_0: $b=0$
Регрессор x	1	1,6091	1,6091	$F=\frac{1,6091}{0,008}=$ $=201,13$	$F(0,05;1,6)=$ $=5,99$	H_1: $b\neq 0$
Ошибка (остаток)	6	0,0479	0,008	–	–	–
Общая дисперсия (итог)	7	1,66	–	–	–	–

Коэффициент детерминации здесь также равен $R^2=r^2=0,97$. Высокое значение как коэффициента корреляции, так и коэффициента детерминации свидетельствует о том, что данные наблюдений хорошо согласуются с представлением их в виде линейной регрессионной модели.

Дадим интерпретацию коэффициентам регрессии. Если не учитывать, что мы имеем не теоретическую, а эмпирическую линию регрессии, то коэффициент $\hat{b}=0,01924$ показывает, что увеличение среднесписочной численности на одного человека приводит к увеличению объема товарооборота в среднем на 19,24 тыс. руб. Это своего рода эмпирический норматив приростной эффективности использования работников данной группы магазинов. Если увеличение численности на одного работника приводит к меньшему росту объема товарооборота, то прием его на работу не обоснован.

Отрицательное значение коэффициента $\hat{a}=-0,974$ означает, что если мы рассмотрим магазины без работников, то объем товарооборота будет снижаться; хотя сама ситуация может показаться парадоксальной. Здесь константа определяет положение линии регрессии на графике. Найдем эластичность E. Так как $\bar{x}=\frac{1}{n}\sum x_i=\frac{1}{8}\cdot 904=113$, $\bar{y}=\frac{1}{n}\sum y_i=\frac{1}{8}\cdot 9,6=1,2$, то $E=0,01924\cdot\frac{113}{1,2}\approx 1,8$. Если среднесписочное число работников увеличить на 1% от среднего значения, то розничный товарооборот возрастет на 1,8% от среднего.

Полученное уравнение регрессии может быть использовано для прогноза. В частности, пусть намечается открытие магазина такого же типа с численностью работников $x_0=140$ чел., тогда достаточно обоснованный объем товарооборота следует установить по уравнению регрессии

$$\hat{y}_0=-0,974+0,01924\cdot 140=1,72 \text{ млн. руб.}$$

Доверительный интервал с надежностью $g = 0{,}95$ для теоретического значения прогноза определим по формуле (1.22):

$$1{,}72 - 2{,}447 \cdot 0{,}039 < y_0 < 1{,}72 + 2{,}447 \cdot 0{,}039$$

или $$1{,}625 < y_0 < 1{,}815.$$

То есть мы на 95% уверены в том, что объем товарооборота для магазинов с численностью работников 140 чел. будет в указанных пределах.

1.7. Нелинейная регрессия

Многие экономические процессы наилучшим образом описываются нелинейными соотношениями, например, нелинейными функциями спроса и производственными функциями. Здесь мы рассмотрим нелинейные модели, которые с помощью преобразования переменных, сводятся к линейным, и потому для их построения могут использоваться описанные выше приемы.

В случае простого регрессионного анализа (линейного однофакторного) речь идет об уравнениях вида

$$y = a + bx, \tag{1.25}$$

состоящих из постоянной величины (которая может и отсутствовать), независимой переменной, умноженной на некоторый коэффициент, и случайной составляющей (ошибки), которой мы можем временно пренебречь. В общем случае линейное уравнение выглядит так

$$y = a + b_1 x_1 + b_2 x_2 + \mathbf{K} \quad . \tag{1.26}$$

Уравнения вида

$$y = a + \frac{b}{x}, \tag{1.27}$$

$$y = ax^b \tag{1.28}$$

являются нелинейными. Их графические изображения для выбранных значений a и b будут представлены кривыми.

Зависимости (1.27) и (1.28) считаются приемлемыми для описания кривых Энгеля, характеризующих соотношение между спросом на определенный товар (y) и общей суммой дохода (x). Как можно определить параметры a и b в каждом уравнении, зная значения y и x?

Заметим, что уравнение (1.27) является линейным по неизвестным параметрам a и b и нелинейным по переменной x. Поэтому оценки параметров могут быть найдены по формулам (1.5) (с заменой $z_i = \frac{1}{x_i}$). Уравнение (1.27) примет вид $y = a + bz$.

Нелинейность по переменным всегда можно обойти путем использования соответствующих определений. Например, для модели вида

$$y = a + b_1 x_1^2 + b_2 \sqrt{x_2} + \mathbf{K}$$

можно определить $z_1 = x_1^2$, $z_2 = \sqrt{x_2}$ и т. д., тогда модель или соотношение примет вид

$$y = a + b_1 z_1 + b_2 z_2 + \mathbf{K}$$

и теперь оно является линейным как по переменным, так и по параметрам. Такой тип преобразований является лишь косметическим, он не меняет свойств оценок, полученных для линейных моделей, и обычно уравнения регрессии записываются с нелинейными выражениями относительно переменных. Это позволяет избежать лишних обозначений.

Уравнение (1.28) является нелинейным как по параметрам, так и по переменной x. Такое соотношение может быть преобразовано в линейное уравнение путем логарифмирования:

$$\ln y = \ln a + b \ln x. \qquad (1.29)$$

Если обозначить $y' = \ln y$, $z = \ln x$ и $a' = \ln a$, то уравнение (1.29) можно переписать в следующем виде

$$y' = a' + bz. \qquad (1.30)$$

Процедура оценивания регрессии теперь будет следующей. Сначала вычислим y' и z для каждого наблюдения путем взятия логарифмов от исходных значений. Затем оценим регрессионную зависимость y' от z. Коэффициент при z будет представлять собой непосредственно оценку $\hat{b}$. Постоянный член является оценкой $\hat{a}'$, т. е. $\ln \hat{a}$. Для получения оценки a необходимо взять антилогарифм, т. е. вычислить $\exp(a')$.

Функции вида (1.28) часто встречаются в эконометрическом моделировании. Для таких функций эластичность y по x равна b. Действительно, если соотношение между y и x имеет вид (1.28), то эластичность

$$E = f'(x)\frac{x}{y} = abx^{b-1}\frac{x}{ax^b} = b.$$

Оценка этого коэффициента по результатам наблюдений будет показывать, на сколько процентов в среднем изменится значение y при изменении x на 1% от своего среднего значения. Например, если имеется кривая Энгеля вида $y = 0{,}01x^{0{,}3}$ (y – спрос, а x – доход), то это означает, что эластичность спроса по доходу равна 0,3, т. е. изменение x на 1% от среднего уровня $\bar{x}$ вызывает изменение y на 0,3% от среднего уровня $\bar{y}$.

Функция вида (1.28) может также применяться к кривым спроса, где y – спрос на товар, x – цена товара, а b – эластичность спроса по цене. (На практике обычно такая функция спроса объединяется с кривой Энгеля, в результате чего получается зависимость спроса одновременно от дохода и цены.)

При моделировании процессов в экономике могут использоваться и показательные (или экспоненциальные) функции вида

$$y = ae^{bx}. \qquad (1.31)$$

Наиболее общим их приложением является случай, когда предполагается, что переменная y имеет постоянный темп прироста во времени, в этом случае вместо x обычно используется время (t), а вместо b – постоянный темп прироста (r):

$$y = ae^{rt}. \qquad (1.32)$$

Если зависимость y от t задана уравнением (1.32), то абсолютный прирост y за единицу времени $\left(\frac{dy}{dt}\right)$ определяется как

$$\frac{dy}{dt} = rae^{rt} = ry.$$

Следовательно, относительный прирост y за единицу времени можно записать так

$$\frac{dy/dt}{y} = \frac{ry}{y} = r.$$

Следует помнить, что оценка $\hat{r}$, которую мы получаем при оценивании регрессии (1.32), представляет собой оценку темпа прироста в абсолютном выражении. Обычно говорят о процентных темпах прироста, это значит, что полученную оценку нужно умножить на 100. Следовательно, если оценка составляет 0,053, это означает, что темп прироста в процентах будет 5,3% за период.

Как же найти оценки неизвестных параметров модели (1.32)? Если имеются значения y для нескольких временных периодов $(1, \mathbf{K}, T)$, то параметры a и r можно оценить, если прологарифмировать (по основанию e) обе части уравнения (1.32):

$$\ln y = \ln a + rt. \qquad (1.33)$$

Если определить $y' = \ln y$ и $a' = \ln a$, то из соотношения (1.33) получим:

$$y' = a' + rt.$$

Таким образом, оценивая регрессию между $\ln y$ и t, мы непосредственно получаем по формулам (1.5) оценку темпа прироста $\hat{r}$ и $\hat{a}'$. Обычно оценка параметра a имеет второстепенное значение, но если она представляет интерес, то можно получить $\hat{a}$, потенцируя $\hat{a}'$.

Пример 1.2. [21]. Предположим, что по результатам наблюдений за расходами на питание в США за период с 1959 по 1983 г. была построена кривая Энгеля в виде соотношения (1.28). Преобразованное в результате логарифмирования и оцененное выражение имело вид:

$$\hat{\ln y} = 1{,}20 + 0{,}55 \ln x.$$

Выполнив обратные преобразования, получим

$$\hat{y} = e^{1{,}20} x^{0{,}55} = 3{,}32 x^{0{,}55}.$$

Если уравнение (1.28) представляет собой правильную формулу зависимости, т. е. модель адекватна, то полученный результат предполагает, что эластичность спроса на продукты питания по доходу составляет 0,55, что означает,

что увеличение личного располагаемого дохода на 1% от среднего уровня $\bar{x}$ приведет к увеличению расходов на питание на 0,55% от среднего уровня $\bar{y}$. Коэффициент 3,32 не имеет простого толкования. Он помогает прогнозировать значения y при заданных значениях x, приводя их к единому масштабу.

Те же данные о расходах на питание были использованы для оценивания экспоненциального временного тренда типа (1.32), также приведенного к линейному виду путем логарифмирования [см. уравнение (1.33)]. Оцененная зависимость имеет вид:

$$\hat{\ln y} = 4{,}58 + 0{,}02t .$$

Выполнив обратные преобразования, получим:

$$y = e^{4{,}58} e^{0{,}02t} = 97{,}5 e^{0{,}02t} .$$

Уравнение показывает, что расходы на продукты питания в течение выборочного периода росли с темпом 2% в год. В этом случае постоянный множитель имеет интерпретацию, так как он «прогнозирует», что в момент $t = 0$, т. е. в 1958 г. общие расходы на питание составили 97,5 млрд. долл. Такой прогноз, безусловно, не имеет важного значения, так как легко можно найти в справочниках действительные расходы на питание в 1958 г.

До сих пор мы ничего не говорили о том, как осуществленные преобразования модели (например, логарифмирование) повлияют на случайную составляющую e. Основное требование здесь состоит в том, чтобы случайная составляющая в преобразованном уравнении присутствовала в виде слагаемого и удовлетворяла условиям 3а – 3с (см. п. 1.1). В противном случае коэффициенты регрессии, полученные по методу наименьших квадратов, не будут обладать обычными свойствами и проводимые для них выводы на основе проверки гипотез окажутся недостоверными.

В случае нелинейных регрессий степень концентрации распределения наблюдаемых точек вблизи линии регрессии показывает корреляционное отношение или индекс корреляции

$$h = \sqrt{\frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}}, \qquad (1.34)$$

где $\hat{y}_i$ – рассчитанные по модели значения переменной y, y_i – фактические или наблюдаемые значения этой переменной, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – среднее значение y, найденное по n наблюдениям, $i = 1, \mathbf{K}, n$.

Из определения индекса корреляции следует, что $0 \le h \le 1$. Если $h = 1$, имеет место функциональная зависимость (все точки сосредоточены на кривой регрессии), если $h = 0$, оцененная модель непригодна.

В отличие от линейного коэффициента корреляции индекс корреляции характеризует тесноту нелинейной связи между переменными в соответствии с той функциональной зависимостью, по которой рассчитаны значения $\hat{y}_i$. Он не характеризует направление связи. Очевидно, что если значения $\hat{y}_i$ рассчитаны по уравнению парной линейной регрессии, значения индекса корреляции и линейного коэффициента корреляции по абсолютной величине совпадут.

Здесь также определяется коэффициент детерминации $R^2 = h^2$, интерпретация которого дается в процентах. Как и в случае линейной регрессии, коэффициент детерминации показывает ту долю вариации переменной y, которая объяснена вариацией фактора x, включенного в уравнение регрессии.

Пример 1.3. По результатам наблюдений над розничным товарооборотом (y) и среднесписочным числом работников (x) (см. пример 1.1) нелинейную модель $y = ax^b$. Сравнить линейную и нелинейную аппроксимацию данных.

Для того чтобы использовать формулы МНК-оценок неизвестных параметров, линеаризуем предложенную модель путем логарифмирования исходных данных:

$$lny = lna + blnx.$$

В табл. 1.5 во втором и третьем столбцах приведены значения логарифмов объемов розничного товарооборота и среднесписочного числа работников соответственно, а в следующих столбцах – значения расчетных величин, необходимых для определения оценок параметров модели, а также статистик для ее верификации.

Таблица 1.5

Результаты преобразованных наблюдений и необходимые расчеты для построения модели

Порядковый номер	lnx_i	lny_i	$(lnx_i)^2$	$lnx_i \cdot lny_i$	$\hat{y}_i$	$\hat{y}_i - \bar{y}$	$(\hat{y}_i - \bar{y})^2$
1	2	3	4	5	6	7	8
1	4,290	-0,693	18,4041	-2,9730	0,4915	-0,7085	0,5020
2	4,443	-0,357	19,7402	-1,5865	0,6583	-0,5417	0,2934
3	4,625	-0,105	21,3906	-0,4856	0,9342	-0,2658	0,0706
4	4,745	0,095	22,5150	0,4508	1,1762	-0,0238	0,0006
5	4,804	0,336	23,0784	1,6141	1,3175	0,1175	0,0138
6	4,836	0,336	23,3869	1,6249	1,4017	0,2017	0,0407
7	4,898	0,351	23,9904	2,6008	1,5776	0,3776	0,1426
8	4,990	0,642	24,9001	3,2036	1,8845	0,6845	0,4685
Итог	37,631	0,785	177,4057	4,4494	–	–	1,5322

Используем итоги столбцов 2-5 и находим оценки $\widehat{lna}$ и $\hat{b}$ по формулам (1.5):

$$ln\hat{a} = \frac{177{,}4057 \cdot 0{,}785 - 37{,}631 \cdot 4{,}4494}{8 \cdot 177{,}4057 - (37{,}631)^2} = -8{,}9338,$$

$$\hat{b} = \frac{8 \cdot 4{,}4494 - 37{,}631 \cdot 0{,}785}{8 \cdot 177{,}4057 - (37{,}631)^2} = 1{,}9201.$$

Отсюда $\hat{a} = e^{-8{,}9338} \approx 0{,}00013$, и оцененная модель запишется в виде

$$\hat{y} = 0{,}00013x^{1{,}92}.$$

Подставляя в это уравнение значения x_i из таблицы 1.2, определим $\hat{y}_i$ (столбец 6). Так как среднее $\bar{y} = \frac{1}{8} \cdot 9{,}6 = 1{,}2$, то заполнив столбцы 7, 8, получим индекс корреляции по формуле (1.34)

$$\eta = \sqrt{\frac{1{,}5322}{1{,}66}} = 0{,}9607.$$

Коэффициент детерминации $R^2 = \eta^2 = (0{,}9607)^2 = 0{,}923$.

Напомним, что для линейной модели $R^2 = 0{,}97$ и коэффициент корреляции $r_{\text{В}} = 0{,}985$. Таким образом, для данного набора значений x_i, y_i линейная модель более пригодна для описания зависимости товарооборота от среднесписочного числа работников. В оцененной нелинейной модели оценка параметра $\hat{b} = 1{,}92$ означает, что если среднесписочное число работников увеличится на 1% от среднего, то объем товарооборота возрастет на 1,92% от среднего.

Контрольные вопросы, задачи и упражнения

1.1. Перечислите основные гипотезы регрессионного анализа и обсудите их.

1.2. Каковы источники случайной составляющей регрессионной модели.

1.3. Выведите формулу для МНК-оценки параметра b уравнения $y = bx$, т. е. методом наименьших квадратов по n наблюдениям (x_i, y_i) получите оценку коэффициента наклона в регрессии без свободного члена.

1.4. Наблюдения 16 пар (x_i, y_i) дали следующие результаты: $\sum y_i^2 = 526$, $\sum x_i^2 = 657$, $\sum x_i y_i = 492$, $\sum y_i = 64$, $\sum x_i = 96$. Оцените регрессию $y_i = a + bx_i + e_i$ и проверьте гипотезу, что коэффициент b равен 1,0.

1.5. Уравнения регрессии между расходами на коммунальные услуги (y) и располагаемым личным доходом (x) временем (t) имеют вид (в скобках указаны стандартные ошибки):

$$\hat{y} = -27{,}6 + 0{,}178x, \quad \hat{y} = 48{,}9 + 4{,}84t.$$
$$(3{,}4) \quad (0{,}004) \qquad (1{,}5) \quad (0{,}10)$$

Постройте доверительные интервалы теоретических коэффициентов моделей, проверьте значимость коэффициентов, полагая $n = 10$.

1.6. Предположим, что по принятой гипотезе 10% предельного дохода расходуется на питание. Проверьте эту гипотезу, используя результат оценивания регрессии, представленной в уравнении

$$y = \underset{(2,4)}{55,3} + \underset{(0,003)}{0,093}x.$$

1.7. Как определить качество модели $\hat{y} = \hat{a} + \hat{b}x$?

1.8. Используя данные наблюдений:

Наблюдения	x_i	y_i	$\hat{y}_i$
1	1	3	3,1667
2	2	5	4,6667
3	3	6	6,1667

вычислите коэффициент корреляции и коэффициент детерминации. Какой вывод можно сделать на основании полученных значений?

1.9. В упражнении 1.5 значение коэффициента R^2 в модели регрессии между расходами на коммунальные услуги и располагаемым личным доходом составило 0,9875. Вычислите соответствующую F-статистику и проверьте гипотезу об адекватности этой модели при уровне значимости 0,05.

1.10. Интерпретируйте результаты оценивания регрессий в упражнении 1.5, дайте им экономическое толкование.

1.11. Используя уравнение регрессии (упр. 1.5), получите точечный и интервальный прогнозы предполагаемых расходов на оплату жилья, если располагаемый личный доход составит 700 (у.е.).

1.12. Могут ли следующие нелинейные уравнения быть преобразованы в уравнения, линейные по параметрам?

а) $y = a/(b - x)$,

б) $y = e^{a+bx}$,

в) $y = \dfrac{x}{a + bx}$,

г) $y = ae^{-bx}$.

1.13. Логарифмические регрессии между а) расходами на продукты питания или б) на оплату жилья и личным располагаемым доходом имели следующий вид (в скобках приведены стандартные ошибки):

а) $\widehat{\ln y} = \underset{(0,11)}{1,20} + \underset{(0,02)}{0,55}\ln x$; $R^2 = 0,98$,

б) $\widehat{\ln y} = \underset{(0,16)}{-3,48} + \underset{(0,02)}{1,23}\ln x$; $R^2 = 0,99$.

Проверьте соответствующие статистические гипотезы и определите 95%-доверительный интервал для эластичности по доходу в каждом случае.

2. Модели множественной регрессии

Естественным обобщением регрессионной модели с двумя переменными является многомерная регрессионная модель или модель множественной регрессии. В этой главе регрессионный анализ по методу наименьших квадратов обобщается для случая, когда в модели вместо одной независимой переменной-фактора используется несколько независимых переменных-факторов количественной и качественной природы.

2.1. Линейная модель множественной регрессии

Рассмотрим общую линейную модель с k переменными. Пусть существует линейное соотношение между объясняемой переменной y, $(k-1)$ объясняющими переменными-регрессорами $x_2, x_3 \mathbf{K}, x_k$, и случайным возмущением (ошибкой) e. Если мы имеем выборку n наблюдений над этими переменными, то можно записать

$$y_i = b_1 + b_2 x_{i2} + \mathbf{K} + b_k x_{ik} + e_i . \qquad (2.1)$$

Коэффициенты b_i и параметры распределения случайной величины e неизвестны. Наша задача состоит в получении наилучших их оценок.

Гипотезы, лежащие в основе модели множественной регрессии, являются естественным обобщением соответствующих гипотез для модели парной регрессии:

1. $y_i = b_1 + b_2 x_{i2} + \mathbf{K} + b_k x_{ik} + e_i$, $i = 1, \mathbf{K}$, n, $n > k$, – спецификация модели, или

$$y_i = b_1 x_{i1} + b_2 x_{i2} + \mathbf{K} + b_k x_{ik} + e_i \qquad (2.2)$$

(то есть можно различать модели со свободным членом вида (2.1) или без свободного члена; очевидно, в модели (2.1) переменная $x_{i1} = 1$ для всех $i = 1, \mathbf{K}, n$).

2. $x_{i1}, \mathbf{K}, x_{ik}$ – детерминированные величины; векторы $x_1 = (1, \mathbf{K}, 1)^T$, $x_2 = (x_{21}, \mathbf{K}, x_{2n})^T$, $\mathbf{K}$, $x_k = (x_{k1}, \mathbf{K}, x_{kn})^T$ – линейно независимы в R^n.

3. $e_1, \mathbf{K}, e_n$ – случайные величины, для которых

3a. $Me_i = 0$, $M\left(e_i^2\right) = D(e_i) = s^2$ для всех $i = 1, \mathbf{K}, n$.

3b. $M(e_i e_j) = 0$ при $i \neq j$ – статистическая независимость (некоррелированность) ошибок для разных наблюдений.

3c. $e_i \sim N(0, s^2)$, т. е. e_i – нормально распределенная случайная величина со средним 0 и дисперсией s^2.

В дальнейшем, стремясь к наибольшей компактности изложения, будем использовать матричные обозначения.

Обозначим через $Y = (y_1, \mathbf{K}, y_n)^T$ $(n \times 1)$ матрицу (вектор-столбец) наблюдений над объясняемой переменной y, $B = (b_1, \mathbf{K}, b_k)^T$ – $(k \times 1)$ вектор неизвестных параметров модели; $e = (e_1, \mathbf{K}, e_n)^T$ – $(n \times 1)$ вектор ошибок;

$X = \begin{bmatrix} x_{11} \mathbf{K} x_{1k} \\ \mathbf{L}\ \mathbf{L}\ \mathbf{L} \\ x_{n1} \mathbf{K} x_{nk} \end{bmatrix}$ – $(n \times k)$ матрица значений объясняющих переменных.

Условия 1–3 в матричной записи выглядят следующим образом:

1. $Y = XB + e$ – спецификация модели;

2. X – детерминированная матрица, имеющая максимальный ранг k, $rankX = k$.

3a.b. $M(e) = 0$, $V(e) = M(ee^T) = s^2 I_n$

(здесь матрица $V(e)$ называется матрицей вариаций или матрицей дисперсий-ковариаций: диагональные элементы этой матрицы равны дисперсиям ошибок s^2, внедиагональные элементы характеризуют корреляционные связи; через I_n обозначена $n \times n$ единичная матрица).

Дополнительное условие

3c. $e \sim N(0, s^2 I_n)$.

В случае выполнения предпосылок 1–3c имеем нормальную линейную модель множественной регрессии.

2.2. Оценивание неизвестных параметров модели

Как и в случае регрессионного уравнения с одной переменной (см. п. 1.2) оценки неизвестных параметров $\hat{b}_i$ модели (2.2) находятся по методу наименьших квадратов из условия минимума суммы квадратов ошибок наблюдений:

$$R = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} b_j x_{ij} \right)^2 \to \min .$$

В матричных обозначениях:

$$R = (Y - XB)^T (Y - XB) \to \min . \qquad (2.3)$$

Необходимые условия экстремума дают систему нормальных уравнений:

$$\frac{\partial R}{\partial b_r} = -2 \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{k} b_j x_{ij} \right) x_{ir} = 0 , \; r = 1, \mathbf{K}, k .$$

Или в матричных обозначениях:

$$X^T Y - X^T X B = 0 .$$

Откуда, учитывая существование матрицы $(X^T X)^{-1}$ в силу условия 2 ($\det(X^T X) \neq 0$), находим МНК-оценку для вектора неизвестных параметров

$$\hat{B} = \left(X^T X\right)^{-1} X^T Y. \qquad (2.4)$$

(Сравните с аналогичной формулой (1.5), полученной для регрессионного уравнения с одной независимой переменной, и попытайтесь получить ее, используя общее решение в матричном виде).

Матрица вариаций оценок $\hat{B}$ равна

$$V\left(\hat{B}\right) = s^2\left(X^T X\right)^{-1}, \qquad (2.5)$$

где несмещенной оценкой дисперсии ошибок наблюдений будет

$$\hat{s}^2 = S^2 = \frac{R_{\min}}{n-k} = \frac{1}{n-k} e^T e. \qquad (2.6)$$

Формула (2.6) позволяет записать оценку матрицы вариаций (2.5) и тем самым оценку дисперсий МНК-оценок неизвестных параметров модели:

$$\hat{V}\left(\hat{B}\right) = \left(X^T X\right)^{-1} \frac{R_{\min}}{n-k};$$

$$\hat{D}\left(\hat{b}_i\right) = \hat{V}_{ii} = \left(X^T X\right)^{-1}_{ii} \frac{R_{\min}}{n-k}. \qquad (2.7)$$

Для $R_{\min}$ можно также получить выражение

$$R_{\min} = e^T e = Y^T Y - \hat{B}^T X^T Y. \qquad (2.8)$$

МНК-оценки (2.4) являются несмещенными и обладают наименьшей дисперсией в классе линейных несмещенных оценок, т. е. являются наиболее эффективными (теорема Гаусса-Маркова) [1; 4; 5].

2.3. Доверительные интервалы и проверка статистических гипотез

Статистический анализ множественной линейной регрессии для нормальной модели производится по аналогии с тем, как это делалось в случае двумерной модели.

Проверка гипотезы $H_0 : b_i = b_{i0}$ по t-критерию, статистика которого

$$t = \frac{\hat{b}_i - b_{i0}}{\sqrt{\hat{D}\left(\hat{b}_i\right)}} \sim t(n-k), \qquad (2.9)$$

выполняется для коэффициентов множественной регрессии так же, как это делается в парном регрессионном анализе (см. п. 1.3). Отметим, что критическая точка $t_{кр.}$ при любом уровне значимости a зависит от числа степеней свободы, которое равно $(n-k)$, где n – число наблюдений, k – число оцененных параметров модели.

Доверительные интервалы определяются точно так же, как и в случае двумерной регрессионной модели, с учетом замечания относительно числа степеней свободы. Так, доверительный интервал вида

$$\hat{b}_i - t_g \sqrt{\hat{D}\left(\hat{b}_i\right)} < b_i < \hat{b}_i + t_g \sqrt{\hat{D}\left(\hat{b}_i\right)}$$

покроет истинное неизвестное значение параметра b_i с доверительной вероятностью или надежностью $g = 1 - a$.

Очевидно гипотеза $H_0 : b_i = b_{i0}$ будет принята с уровнем значимости a , если соответствующий доверительный интервал содержит гипотетическое значение b_{i0}.

Отметим, что проверка значимости коэффициентов регрессии или значимости влияния регрессоров – это проверка гипотез $H_0 : b_i = 0$. Регрессор принимается статистически незначимым, если доверительный интервал для соответствующего коэффициента регрессии покрывает нуль.

2.4. Качество модели: дисперсионный анализ и коэффициент R^2

Качество оценивания многомерной регрессии, как и в случае регрессионной модели с одной независимой переменной, можно определить дисперсионным анализом в модели и с использованием коэффициента детерминации R^2.

Общая сумма квадратов $SS_{общ.} = \sum_{i=1}^{n} (y_i - \bar{y})^2$ разбивается здесь на две части: объясненную регрессионным уравнением и не объясненную (т. е. связанную с ошибками e_i):

$$SS_{общ.} = SS_R + SS_{ост.},$$

где $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$, $SS_{ост.} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$.

Гипотеза об отсутствии линейной функциональной связи между объясняемой переменной y и регрессорами $x_2, \mathbf{K}, x_k$ может быть записана как $H_0 : b_2 = \mathbf{K} = b_k = 0$ (мы предполагаем, что в число регрессоров включена константа – свободный член), т. е. нулевая гипотеза состоит в том, что коэффициенты при всех регрессорах равны нулю.

Для проверки этой гипотезы используется критерий, статистика которого

$$F = \frac{\sum (\hat{y}_i - \bar{y})^2 / (k-1)}{\sum (y_i - \hat{y}_i)^2 / (n-k)} = \frac{MS_R}{MS_{ост.}} \sim F(k-1, n-k) \qquad (2.10)$$

имеет распределение Фишера с соответствующими числами степеней свободы.

Если $F_0 > F_{кр.}(a; k-1, n-k)$, гипотеза H_0 отвергается на уровне значимости a ; уравнение в целом значимо и оцененная линейная множественная регрессия

$$\hat{y} = \hat{b}_1 + \hat{b}_2 x_2 + \mathbf{K} + \hat{b}_k x_k$$

пригодна для описания зависимости между y и $x_2, \mathbf{K}, x_k$.

Вычисления, необходимые для дисперсионного анализа множественной регрессии, обычно сводят в таблицу (табл. 2.1).

Таблица 2.1

Дисперсионный анализ множественной регрессии

Источник дисперсии	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F	Критическая точка $F_{кр.}(a, k-1, n-k)$	Гипотеза $H_0: b_2 = \mathbf{K} = b_k = 0$
Модель (регрессоры $x_2, \mathbf{K}, x_k$)	$k-1$	SS_R	$MS_R = \frac{SS_R}{k-1}$	$F = \frac{MS_R}{MS_{ост.}}$		
Ошибка	$n-k$	$SS_{ост.} = SS_{общ} - SS_R$	$MS_{ост.} = \frac{SS_{ост.}}{n-k}$	–	–	–
Общая дисперсия (итог)	$n-1$	$SS_{общ.}$	–	–	–	–

Как и ранее в (1.15), определим коэффициент детерминации

$$R^2 = 1 - \frac{SS_{ост.}}{SS_{общ.}} = \frac{SS_R}{SS_{общ.}}. \qquad (2.11)$$

Коэффициент $R^2 \in [0,1]$ показывает качество подгонки регрессионной модели к наблюдаемым значениям y_i.

Если $R^2 = 0$, то регрессия не улучшает качество предсказания y_i по сравнению с тривиальным предсказанием $\hat{y}_i = \bar{y}$. Другой крайний случай $R^2 = 1$ означает точную подгонку: все точки наблюдений лежат на регрессионной плоскости.

Определенная в (2.10) F-статистика с учетом коэффициента детерминации R^2 определится как

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

Заметим, что при добавлении еще одного регрессора или еще одной объясняющей переменной к уравнению регрессии коэффициент R^2, вообще говоря, возрастает. Если взять число регрессоров, равным числу наблюдений, всегда можно добиться того, что $R^2 = 1$, но это вовсе не будет означать, что суще-

ствует содержательная, имеющая экономический смысл зависимость y от регрессоров. Для того, чтобы устранить эффект, связанный с ростом R^2 при возрастании числа регрессоров, вводится скорректированный коэффициент детерминации $\overline{R}^2$:

$$\overline{R}^2 = 1 - \frac{SS_{ост.}/(n-k)}{SS_{общ.}/(n-1)} = 1 - \frac{MS_{ост.}}{MS_{общ.}}. \qquad (2.12)$$

Корректировка R^2 на число регрессоров оправдана тем, что числитель дроби в (2.12) есть несмещенная оценка дисперсии ошибок, а знаменатель – несмещенная оценка дисперсии y.

Для скорректированного коэффициента детерминации $\overline{R}^2$ справедливо

$$\overline{R}^2 = 1 - \left(1 - R^2\right)\frac{n-1}{n-k} = \frac{n-1}{n-k}R^2 - \frac{k-1}{n-k} = R^2 - \frac{k-1}{n-k}\left(1 - R^2\right).$$

Отсюда, по мере роста k увеличивается отношение $\frac{k-1}{n-k}$, и, следовательно, возрастает размер корректировки коэффициента R^2 в сторону уменьшения, т. е. $R^2 \geq \overline{R}^2$ для $k > 1$.

Использование скорректированного коэффициента детерминации $\overline{R}^2$ более корректно для сравнения регрессий при изменении числа регрессоров. Однако следует иметь в виду, что иногда даже плохо определенная модель регрессии может дать высокий коэффициент детерминации R^2, и признание этого факта привело к снижению значимости R^2. Теперь он рассматривается лишь как один из показателей, который должен быть проверен при построении модели регрессии. Следовательно, и корректировка этого коэффициента мало что дает.

2.5. Интерпретация коэффициентов множественной регрессии и прогнозирование на ее основе

Множественный регрессионный анализ позволяет разграничить влияние независимых переменных, допуская при этом возможность их коррелированности (проблема наличия связи между регрессорами или их мультиколлинеарность будет обсуждаться в п. 3.1.). Коэффициент регрессии при каждой переменной x дает оценку ее влияния на величину y в случае неизменности влияния на нее всех остальных переменных x. Так, например, в оцененной регрессии

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_2 x_2$$

коэффициенты $\hat{b}_1$ и $\hat{b}_2$ являются показателями силы связи, характеризующими абсолютное (в натуральных единицах измерения) изменение объясняемой пе-

ременной y при изменении каждого из x_1 и x_2 соответственно на единицу своего измерения при фиксированном влиянии второй переменной.

Относительными показателями силы связи в уравнении множественной регрессии являются частные коэффициенты эластичности:

$$E_{yx_j} = \hat{b}_j \frac{\bar{x}_j}{\bar{y}},$$

где $\bar{x}_j$ и $\bar{y}$ – выборочные средние величины объясняющей переменной x_j и результирующего показателя y соответственно, значения которых подсчитаны в ходе статистического анализа рассматриваемой регрессионной модели.

Эластичность E_{yx_j} показателя y по переменной x_j приблизительно определяет на сколько процентов изменится значение y от своего среднего уровня при изменении объясняющей переменной x_j на 1% от ее среднего уровня.

Прогноз на основе модели множественной регрессии может быть, так же как и в случае модели парной регрессии, точечным и интервальным. Если задан дополнительный набор объясняющих переменных – вектор $x^0 = \left(x_1^0, x_2^0, \mathbf{K}, x_k^0\right)$, то точечный прогноз получается подстановкой прогнозных значений регрессоров в уравнение модели. Для получения интервального прогноза вначале рассчитывается оценка дисперсии оценки прогнозируемой величины

$$\hat{D}\left(\hat{y}_0\right) = \hat{s}^2 \left[1 + x^0 \left(X^T X\right)^{-1} x^{0^T}\right],$$

где $\hat{s}^2 = S^2 = \frac{R_{\min}}{n-k}$.

С надежностью g можно утверждать, что истинное значение прогнозируемой величины y_0 покрывается интервалом

$$\hat{y}_0 - t_g \sqrt{\hat{D}\left(\hat{y}_0\right)} < y_0 < \hat{y}_0 + t_g \sqrt{\hat{D}\left(\hat{y}_0\right)}.$$

Здесь $t_g = t\left(\frac{1+g}{2},\ n-k\right)$ – квантиль распределения Стьюдента.

Пример 2.1. На предприятиях Российской Федерации изучалась зависимость объема производства (y) от капитальных вложений (x_1) и выполнения нормы выработки (x_2). Исходные данные для 14 предприятий приведены в табл. 2.2.

В данном примере мы располагаем пространственной выборкой объема $n = 14$; число объясняющих переменных $k = 2$. Специальный анализ технологий сбора исходных статистических данных показал, что гипотеза о взаимной некоррелированности и гомоскедастичности ошибок наблюдений может быть принята. Поэтому мы можем записать уравнения статистической связи между y_i и x_{i1}, x_{i2} в виде

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + e_{i,}\ \ i = 1, \mathbf{K}, 14$$

с выполнением условий 2–3с (см. п. 2.1).

Матрица X будет составлена из трех столбцов размерности 14 каждый; в качестве первого столбца используется вектор, состоящий из единиц, а столбцы 2 и 3 представлены соответственно 3 и 4 столбцами табл. 2.2. Вектор-столбец Y определяется 2-м столбцом табл. 2.2.

Таблица 2.2

Данные об объеме производства (y – млн. руб.) капитальных вложениях (x_1 – млн. руб.) и выполнении нормы выработки (x_2 – %)

Номер предприятия	y_i	x_1	x_2
1	2	3	4
1	52,8	16,3	99,5
2	48,4	16,8	98,9
3	52,4	18,5	99,2
4	50,0	16,3	99,3
5	54,9	17,9	99,8
6	53,9	17,4	99,6
7	53,8	17,5	99,5
8	53,1	16,1	99,8
9	52,4	16,2	99,7
10	53,0	17,0	99,8
11	52,9	16,7	99,9
12	53,1	17,5	100,0
13	60,1	19,1	100,2
14	60,0	19,0	100,1
Итог	750,8	242,3	1395,3
Средний итог	53,63	17,31	99,66

Применение формулы (2.4) к исходным данным позволяет получить следующие МНК-оценки для параметров модели:

$$\hat{b}_0 = -426{,}993,\ \hat{b}_1 = 1{,}911,\ \hat{b}_2 = 4{,}492.$$

Таким образом, оценка множественной регрессии в данном случае имеет вид

$$\hat{y} = -426{,}993 + 1{,}911 x_1 + 4{,}492 x_2. \qquad (2.13)$$

Сумма квадратов остатков, вычисленная по формуле (2.8) с использованием результатов оценивания уравнения (2.13), равна $R_{\min} = 14{,}918$. Отсюда несмещенная оценка дисперсии ошибок наблюдений получится равной $\hat{s}^2 = R_{\min}/(14-3) = 1{,}356$. С учетом этого можно записать оценку матрицы вариаций МНК-оценок коэффициентов регрессии:

$$\hat{V} = \begin{bmatrix} 8869{,}49 & 10{,}36 & -90{,}79 \\ 10{,}36 & 0{,}12 & -0{,}12 \\ -90{,}79 & -0{,}12 & 0{,}93 \end{bmatrix}$$

(диагональные элементы этой матрицы равны оценкам дисперсий МНК-оценок $\hat{b}_0, \hat{b}_1, \hat{b}_2$; внедиагональные – их ковариациям).

Стандартная форма записи оцененной модели, объединяющая информацию о значениях оценок регрессии $\hat{b}_i$ и их средних квадратических ошибках S_i, как правило, имеет следующий вид:

$$\hat{y} = \underset{(94{,}18)}{-426{,}993} + \underset{(0{,}35)}{1{,}911 x_1} + \underset{(0{,}96)}{4{,}492 x_2}.$$

В скобках под значениями оцененных коэффициентов регрессии $\hat{b}_i$ указаны оценки их средних квадратических отклонений S_i.

Соответствующие t-статистики для проверки гипотез $H_{0i}: \; b_i = 0$, $i = 0{,}1{,}2$, равны 4,5339; 5,561 и 4,6515 соответственно. Критическая точка $t_{kp} = t(0{,}975; 14-3) = t(0{,}975; 11) = 2{,}201$ (см. приложение 2). Сравнение полученных значений t-статистик с критической точкой показывает, что нулевые гипотезы о статистической незначимости коэффициентов регрессии должны быть отвергнуты.

Качество всей модели в целом определим дисперсионным анализом модели. Вычисления, необходимые для этого, сведем в табл. 2.3. Так как значение F –статистики, найденное по модели, больше критической точки, гипотеза об отсутствии линейной функциональной связи между объемом производства (y), капитальными вложениями (x_1) и выполнением нормы выработки (x_2) отвергается. Коэффициент детерминации $R^2 = 0{,}884$.

Зависимость y от x_1, x_2 характеризуется как тесная, в которой 88,4% вариации объема производства определяются вариацией учтенных в модели факторов.

Для характеристики силы влияния x_1 на y и x_2 на y рассчитываем частные коэффициенты эластичности:

$$E_{yx_1} = 1{,}911 \cdot \frac{17{,}31}{53{,}63} = 0{,}617\%; \;\; E_{yx_2} = 4{,}492 \cdot \frac{99{,}66}{53{,}63} = 8{,}347\%.$$

С увеличением капитальных вложений (x_1) на 1% от их среднего уровня объем производства (y) возрастает на 0,617% от своего среднего уровня; при

увеличении выполнения нормы выработки (x_2) на 1% от своего среднего уровня объем производства (y) возрастает на 8,347%.

Таблица 2.3

Дисперсионный анализ множественной регрессии

Источник дисперсии	Число степеней свободы	Сумма квадратов SS	Средний квадрат MS	Критерий Фишера F	Крит. точка $F(0,05;2,11)$	Гипотеза H_0
Модель (регрессоры x_1, x_2)	2	113,56	56,778	41,8673	3,98	$b_1 \neq 0$ $b_2 \neq 0$
Ошибка	11	14,918	1,3561	–	–	–
Общая дисперсия (итог)	13	128,47	–	–	–	–

Очевидно, что сила влияния выполнения нормы выработки x_2 на объем производства оказалась больше, чем сила влияния капитальных вложений x_1. На этот же факт указывает и сравнение оценок коэффициентов регрессии: $\hat{b}_2 > \hat{b}_1$.

2.6. Множественная регрессия в нелинейных моделях

В предыдущей главе было показано, что линейные модели регрессии могут быть описаны как линейные в двух отношениях: как линейные по переменным и как линейные по параметрам или коэффициентам регрессии. Для линейного регрессионного анализа требуется линейность только по параметрам, так как нелинейность по переменным может быть устранена либо изменением определений, либо, если это возможно, логарифмированием.

Например, зависимость

$$y = b_0 + b_1 x_1^2 + b_2 \sqrt{x_2}$$

является линейной по неизвестным параметрам b_0, b_1, b_2 и нелинейной по переменным x_1, x_2. Для определения МНК-оценок этих параметров можно воспользоваться формулой (2.4), имея в виду, что первый столбец матрицы X будет состоять из единиц, второй – из квадратов наблюдений над переменной x_1, а третий – из корней квадратных из данных для переменной x_2. Если случайная составляющая (не показана явно в уравнении) удовлетворяла условиям 2–3с (см. п. 2.1), то свойства МНК-оценок, полученных при этом, будут совпадать со свойствами МНК-оценок параметров модели (2.1).

Нелинейность по параметрам является более серьезной проблемой. Если, однако, правая часть модели состоит из членов вида x^b или e^{bx}, умноженных друг на друга, а случайная составляющая мультипликативна, то модель может быть линеаризована логарифмированием обеих ее частей.

Пример 2.2. Функция спроса

$$y = ax^{b_1} p^{b_2} e,$$

где y – расходы на товар, x – доход, p – относительная цена, а e – случайная составляющая, может быть преобразована в линейную по параметрам модель:

$$\ln y = \ln a + b_1 \ln x + b_2 \ln p + \ln e.$$

Если вы оцениваете регрессию между данными для $\ln y, \ln x$ и $\ln p$, то коэффициент при $\ln x$ будет непосредственной оценкой b_1 – эластичности спроса по доходу, а коэффициент при $\ln p$ будет оценкой b_2 – эластичности спроса по цене.

Пример 2.3. [21] Производственная функция Кобба-Дугласа.

В 1927 г. Пол Дуглас, экономист по образованию, обнаружил, что если нанести на одну и ту же диаграмму графики логарифмов показателей реального объема выпуска (Y), капитальных затрат (K) и затрат труда (L), то расстояния от точек графика показателей выпуска до точек графиков показателей затрат труда и капитала будут составлять постоянную пропорцию. Затем он обратился к математику Чарльзу Коббу с просьбой найти математическую зависимость, обладающую такой особенностью, и Кобб предложил следующую функцию:

$$Y = AK^a L^{1-a}. \qquad (2.14)$$

Эта функция была предложена примерно 30 годами раньше Филипом Уикстидом, как было указано Ч. Коббом и П. Дугласом в их классической работе (Cobb, Douglas, 1929), но они были первыми, кто использовал для ее построения эмпирические данные.

Если провести линеаризацию уравнения (2.14) путем логарифмирования обеих его частей, то, используя МНК, мы получим две различные оценки a. Коэффициент при $\ln K$ даст нам одну оценку, а коэффициент при $\ln L$, который является оценкой $(1-a)$, позволит нам вычислить другую оценку. Вместо этого разделим обе части уравнения (2.14) на L и перепишем его следующим образом:

$$Y/L = A(K/L)^a e \qquad (2.15)$$

(включая случайную составляющую e). В этой форме функция может быть интерпретирована как соотношение выпуска на одного работника и капитальным затратам на одного работника (зависимость производительности труда от его капиталовооруженности). Теперь логарифмируя ее, получим

$$\ln(Y/L) = \ln A + a \ln(K/L) + \ln e.$$

При использовании для оценивания этого уравнения данных реального объема производства, реальных капитальных затрат и реальных затрат труда

промышленности США в 1899-1922 гг. получены следующие результаты (в скобках указаны стандартные ошибки оценок коэффициентов модели):

$$\hat{\ln}(Y/L) = 0{,}02 + 0{,}25\ln(K/L),\ R^2 = 0{,}63$$
$$(0{,}02)\ (0{,}04)\qquad F = 38{,}0\quad .$$

Формула Кобба-Дугласа, конечно, является частным случаем более общей формулы:

$$Y = AK^a L^b e, \tag{2.16}$$

где показатели эластичности a, b выпуска по затратам капитала и труда соответственно не связаны между собой. При линеаризации (2.16) путем логарифмирования и использовании тех же данных, что и для модели (2.15), получено

$$\hat{\ln} Y = -0{,}18 + 0{,}23\ln K + 0{,}81\ln L,\ R^2 = 0{,}96$$
$$(0{,}43)\ (0{,}06)\qquad (0{,}15)\qquad F = 236{,}1\quad .$$

Здесь эластичность выпуска продукции по затратам капитала составляет 0,23, что очень близко к предыдущей оценке, а эластичность по затратам труда составляет 0,81, что несколько выше предыдущей оценки, равной 1 – 0,25=0,75.

Если в модели (2.16) $a + b = 1$ (т.е. модель такова, что при расширении масштаба производства – увеличении затрат капитала K и труда L в некоторое число раз – объем производства возрастет в то же число раз) функция Кобба-Дугласа представится соотношением (2.14).

Функция Кобба-Дугласа с учетом технического прогресса имеет вид

$$Y = AK^a L^b e^{qt} e, \tag{2.17}$$

где t – время, параметр q определяет темп прироста объема производства благодаря техническому прогрессу.

Модель (2.17) приводится к линейному виду путем логарифмирования.

Контрольные вопросы, задачи и упражнения

2.1. Перечислите свойства МНК-оценок линейной множественной регрессии и прокомментируйте их.

2.2. Что является основной характеристикой качества модели, ее прогностической силы?

2.3. Пятифакторное уравнение линейной регрессии для переменной y оценено по 31 наблюдениям. При этом объясненная регрессией и остаточная дисперсия соответственно равны 8 и 2. Вычислите коэффициент детерминации и расчетное или наблюдаемое значение F –статистики.

2.4. Оценка множественной регрессии между расходами на коммунальные услуги, располагаемым личным доходом и индексом относительных цен получена в виде

$$\hat{y} = -43{,}4 + 0{,}181x + 0{,}137\,p.$$

Дайте экономическую интерпретацию этого результата.

2.5. Оценка логарифмической регрессии между расходами на коммунальные услуги, располагаемым личным доходом и относительной ценой этих услуг получена по тем же данным, что и модель в упр. 2.4, и имеет вид

$$\hat{\ln y} = -1{,}60 + 1{,}18 \ln x - 0{,}34 \ln p.$$

Дайте интерпретацию этого уравнения. Сравните ее с интерпретацией, данной для упр. 2.4. В каком смысле она лучше?

2.6. На основе $n = 30$ наблюдений получено следующее уравнение регрессии зависимой переменной y на три независимые переменные x_1, x_2, x_3:

$$y = 25{,}1 + 1{,}2x_1 + 1{,}0x_2 - 0{,}50x_3$$

Стандартные ошибки	(2,1)	(1,5)	(1,3)	(0,06)
t-значения	(11,9)	( )	( )	( )
95%-доверительные границы	($\pm$4,3)	( )	( )	( )

а) Заполните пропуски.

б) Оцените значимость коэффициентов регрессии.

2.7. Бюджетное обследование пяти случайно выбранных семей дало следующие результаты (в тыс. р.):

Семья	Накопления, S	Доход, Y	Имущество, W
1	3	40	60
2	6	55	36
3	5	45	36
4	3,5	30	15
5	1,5	30	90

а) Оцените регрессию $S = b_1 + b_2 Y + b_3 W + e$.

б) Постройте 95%-доверительные интервалы для коэффициентов регрессии.

в) Проверьте с уровнем значимости $a = 0{,}05$ следующие гипотезы:

1) $b_3 = 0$ (стоимость имущества несущественна);

2) $b_2 = 0$ (величина дохода несущественна);

3) $b_2 = 1{,}57$ (такое значение коэффициента b_2 могло быть с высокой степенью надежности установлено для другой страны и вас интересует вопрос, верно ли это для вашей страны).

г) Пусть некоторая семья имеет доход $Y = 30$ тыс. руб. и имущество стоимостью $W = 52{,}5$ тыс. руб. Чему равна прогнозная величина ее накоплений?

д) Дайте оценку полученного уравнения на основе коэффициента детерминации и общего F-критерия Фишера.

3. Некоторые особенности при построении моделей множественной регрессии

В предыдущей главе рассматривались основные теоретические вопросы построения и анализа моделей множественной регресси. Здесь мы рассмотрим некоторые особенности или проблемы, которые возникают при их практическом использовании.

3.1. Мультиколлинеарность

На практике исследователю нередко приходится сталкиваться с ситуацией, когда полученная им регрессия является «плохой», т.е. $t-$ статистики большинства оценок малы, что свидетельствует о незначимости соответствующих независимых переменных (регрессоров). В то же время $F-$ статистика (2.10) может быть достаточно большой, что говорит о значимости регрессии в целом. Одна из возможных причин такого явления носит название мультиколлинеарности и возникает при наличии высокой корреляции между регрессорами. Эта проблема является обычной для регрессий временных рядов, т.е. когда данные состоят из ряда наблюдений в течение какого-то периода времени. Если две или более независимые переменные имеют ярко выраженный тренд, то они будут тесно коррелированы, и это может привести к мультиколлинеарности.

Одним из условий классической регрессионной модели является предположение о линейной независимости объясняющих переменных, что означает линейную независимость столбцов матрицы исходных данных X или (эквивалентно), что матрица X^TX имеет полный ранг k ($rankX = k$, где k – число оцениваемых параметров модели). Это предположение потребовалось, чтобы обеспечить обратимость матрицы X^TX, необходимую для вычисления МНК-оценки $\hat{B} = \left(X^TX\right)^{-1}X^TY$. (Если ранг матрицы X меньше k, то и ранг X^TX меньше k, т. е. матрица X^TX оказывается вырожденной, ее определитель равен нулю, а значит, не существует обратная матрица $\left(X^TX\right)^{-1}$). При нарушении этого условия, т.е. когда один из столбцов матрицы X есть линейная комбинация остальных столбцов, говорят, что имеет место полная коллинеарность. В этой ситуации нельзя построить единственную МНК-оценку $\hat{B}$, что формально следует из вырожденности матрицы X^TX и невозможности решить систему нормальных уравнений.

Нетрудно также понять и содержательный смысл этого явления. Рассмотрим следующий простой пример регрессии:

$$C = b_1 + b_2S + b_3N + b_4T + e,$$

где C – потребление, S – зарплата, N – доход, получаемый вне работы, T – полный доход. Поскольку выполнено равенство $T = S + N$, то для произвольного числа h исходную регрессию можно переписать в следующем виде

$$C = b_1 + b_2' S + b_3' N + b_4' T + e,$$

где $b_2' = b_2 + h,\ b_3' = b_3 + h,\ b_4' = b_4 - h.$

Таким образом, одни и те же наблюдения могут быть объяснены различными наборами коэффициентов b, т. е. мы наблюдаем неединственность МНК-оценок. Кроме того, если с учетом равенства $T = S + N$ переписать исходное уравнение в виде

$$C = b_1 + (b_2 + b_4)S + (b_3 + b_4)N + e,$$

то ясно, что оценить можно три параметра b_1, $(b_2 + b_4)$, $(b_3 + b_4)$, а не четыре исходных.

В общем случае можно показать, что если $rank(X^T X) = l < k$, то оценить можно только l линейных комбинаций исходных коэффициентов (так называемых параметрических функций). Если есть полная коллинеарность, то можно выделить в матрице X максимальную линейно независимую систему столбцов и, удалив остальные столбцы, провести новую регрессию.

На практике полная коллинеарность (т.е. когда все или некоторые из объясняющих переменных подчиняются точной (функциональной) линейной связи) встречается исключительно редко (так как ее несложно избежать уже на предварительной стадии анализа и отбора множества объясняющих переменных). Гораздо чаще приходится сталкиваться с ситуацией, когда матрица X имеет полный ранг, но между регрессорами имеется высокая степень корреляции, т. е. когда матрица $X^T X$, говоря нестрого, близка к вырожденной, $\det(X^T X) \approx 0$. Тогда говорят о наличии мультиколлинеарности. В этом случае МНК-оценки формально существуют, но обладают «плохими» свойствами. Это нетрудно объяснить, используя геометрическую интерпретацию МНК.

Регрессию можно рассматривать как проекцию в пространстве R^n вектора Y на векторы, образованные столбцами матрицы X. Если между этими векторами существует приблизительная линейная зависимость, то операция проектирования становится неустойчивой: небольшое изменение в исходных данных может привести к существенному изменению оценок.

На рис. 3.1 представлено разложение вектора наблюдений Y на оси, соответствующие двум независимым переменным x_1 и x_2.

Векторы Y и Y' мало отличаются друг от друга, но в силу того, что угол между регрессорами x_1 и x_2 мал, разложения проекций этих векторов по x_1 и x_2 отличаются значительно.

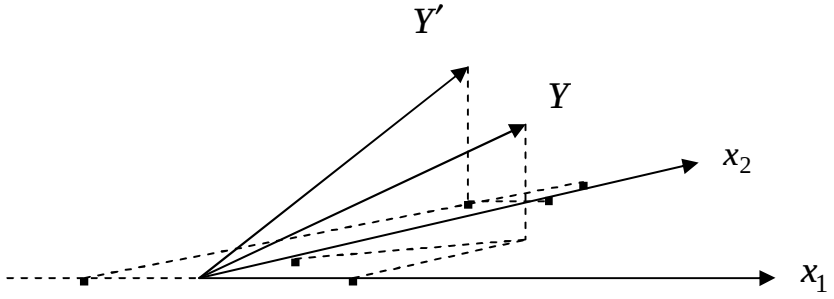


Рис. 3.1

У проекции вектора Y оба коэффициента разложения по x_1 и x_2 (отметим, что это и есть МНК-оценки) положительны и относительно невелики. У проекции вектора Y' коэффициент при x_1 принимает отрицательное значение, а коэффициент при x_2 значительно больше. В силу этого обстоятельства интерпретация коэффициентов регрессии становится весьма проблематичной.

Реальная (или частичная) мультиколлинеарность возникает в случаях, когда между объясняющими переменными существуют достаточно тесные линейные статистические связи. Точных количественных критериев для определения наличия или отсутствия мультиколлинеарности не существует. Тем не менее, возможны некоторые рекомендации по ее выявлению.

1. В первую очередь анализируют матрицу парных коэффициентов корреляции, точнее, ту ее часть, которая относится к объясняющим переменным. Считается, что если значения коэффициентов корреляции по абсолютной величине больше 0,75 – 0,80, то это свидетельствует о присутствии мультиколлинеарности.

2. Если $\det\left(X^T X\right)$ оказывается близким к нулю (допустим, одного порядка с накапливающимися ошибками вычислений), то это тоже свидетельствует о наличии мультиколлинеарности.

3. Важную роль в анализе мультиколлинеарности играет и минимальное собственное число $l_{\min}$ матрицы $X^T X$. Если $l_{\min}$ близко к нулю, то и $\det\left(X^T X\right)$ близок к нулю и наоборот. Поэтому, наряду с величиной $\det\left(X^T X\right)$ (или вместо нее), вычисляют и сравнивают с накапливающимися ошибками от округлений значение $l_{\min}$, т. е. минимальный корень уравнения

$$\left|X^T X - lI\right| = 0.$$

4. Наконец, о присутствии явления мультиколлинеарности говорят и некоторые внешние признаки построенной модели, которые являются его следствием. К ним в первую очередь следует отнести такие:

а) небольшое изменение исходных статистических данных (добавление или изъятие небольшой порции наблюдений) приводит к существенному изменению оценок параметров модели, вплоть до изменения их знаков;

б) оценки имеют большие стадартные ошибки, малую значимость, в то время, как модель в целом является значимой (высокое значение коэффициента детерминации R^2 и соответствующей F -статистики);

в) оценки параметров имеют неправильные с точки зрения экономической теории знаки или неоправданно большие значения.

Для устранения или уменьшения мультиколлинеарности используется ряд методов. Самый простой из них состоит в том, что из двух объясняющих переменных, имеющих высокий коэффициент корреляции, одну переменную исключают из рассмотрения. При этом, какую переменную оставить, а какую удалить, решают в первую очередь на основании экономических соображений. Если с экономической точки зрения ни одной из переменных нельзя отдать предпочтение, то оставляют ту из двух переменных, которая имеет больший кооэффициент корреляции с зависимой переменной.

Другой метод устранения или уменьшения мультиколлинеарности заключается в переходе от несмещенных оценок неизвестных параметров модели, найденных по методу наименьших квадратов, к смещенным оценкам, обладающим, однако, меньшим разбросом относительно истиных значений этих параметров, т.е. меньшими стандартными ошибками. Это достигается использованием «ридж-регрессии» (или «гребневой регрессии») [2; 3; 6].

Еще одним из возможных методов устранения или уменьшения мультиколлинеарности является использование пошаговых процедур отбора наиболее информативных переменных.

На первом шаге рассматривается лишь одна переменная, имеющая с y (зависимой переменной) наибольший коэффициент детерминации, равный квадрату коэффициента корреляции. На втором шаге включается в регрессию новая объясняющая переменная, которая вместе с первоначально отобранной образует пару объясняющих переменных, имеющих с y наиболее высокий (скорректированный) коэффициент детерминации. На третьем шаге вводится в регрессию еще одна объясняющая переменная, которая вместе с двумя первоначально отобранными образует тройку объясняющих переменных, имеющих с y наибольший (скорректированный) коэффициент детерминации и т.д.

Процедура введения новых переменных продолжается до тех пор, пока будет увеличиваться соответствующий (скорректированный) коэффициент детерминации.

Более подробно методы устранения мультиколлинеарности и примеры описаны в [6].

3.2. Фиктивные переменные

Независимые переменные в регрессионных моделях, как правило, имеют «непрерывные» области изменения (национальный доход, уровень безработицы, размер зарплаты и т. д.). Однако теория не накладывает никаких ограничений на характер регрессоров, в частности, некоторые переменные могут принимать всего два значения или, в более общей ситуации, дискретное множество

значений. Необходимость рассматривать такие переменные возникает довольно часто в тех случаях, когда требуется принимать во внимание какой-либо качественный признак. Например, при исследовании зависимости зарплаты от различных факторов может возникнуть вопрос, влияет ли на ее размер и, если да, то в какой степени, наличие у работника высшего образования. Также можно задать вопрос, существует ли дискриминация в оплате труда между мужчинами и женщинами. В принципе можно оценивать соответствующие уравнения внутри каждой категории, а затем изучать различия между ними, но введение дискретных переменных позволяет оценивать одно уравнение сразу по всем категориям.

Рассмотрим пример с заработной платой. Пусть y (руб.) – заработная плата работника, $x = (x_1, x_2, \mathbf{K}, x_k)^T$ – набор объясняющих (независимых) переменных или количественных признаков, от которых может зависеть величина y (трудовой стаж, категория оплаты и т. д.). В действительности, y и x_j – это логарифмы соответствующих характеристик, так как связь между заработной платой и определяющими ее признаками имеет мультипликативный (степенной) характер. Логарифмирование степенной зависимости позволяет перейти к линейной аддитивной модели:

$$y_i = b_1 x_{i1} + b_2 x_{i2} + \mathbf{K} + b_k x_{ik} + e_i, \ i = 1, \mathbf{K}, n, \qquad (3.1)$$

где y_i – размер зарплаты i–го работника.

Теперь нам интересно включить в рассмотрение такой фактор, как наличие или отсутствие у работника высшего образования. Введем новую, бинарную, переменную d, полагая

$$d_i = \begin{cases} 1, & \text{если в i-том наблюдении индивидуум} \\ & \text{имеет высшее образование;} \\ 0, & \text{в противном случае.} \end{cases}$$

Рассмотрим новую систему

$$y_i = b_1 x_{i1} + b_2 x_{i2} + \mathbf{K} + b_k x_{ik} + c d_i + e_i, \ i = 1, \mathbf{K}, n \qquad (3.2)$$

Иными словами, принимая модель (3.2), мы считаем, что средняя зарплата есть $X^T B$ (в матричном обозначении) при отсутствии высшего образования и $X^T B + c$ – при его наличии. Таким образом, величина c интерпретируется как среднее изменение зарплаты при переходе из одной категории (без высшего образования) в другую (с высшим образованием) при неизменных значениях остальных параметров. К модели (3.2) можно применить МНК и получить оценки соответствующих коэффициентов. Тестируя гипотезу $H_0 : c = 0$, мы проверяем предположение о несущественном различии в зарплате между категориями.

В англоязычной литературе по эконометрике переменные указанного выше типа называются dummy variables («фиктивные» переменные). Следует,

однако, ясно понимать, что d такая же «равноправная» переменная, как и любой из регрессоров x_j $(j=1,\mathbf{K},k)$. Ее «фиктивность» состоит только в том, что она количественным образом описывает качественный признак.

Качественное различие можно формализовать с помощью любой переменной, принимающей два значения, а не обязательно значения 0 или 1. Однако в эконометрической практике почти всегда используют лишь фиктивные переменные типа «0 – 1», поскольку в этом случае интерпретация выглядит наиболее просто. Если бы в рассмотренном выше примере переменная d принимала значение, скажем, 5 для работника с высшим образованием и 2 для работника без высшего образования, то коэффициент при этом регрессоре равнялся бы трети среднего изменения зарплаты при получении высшего образования.

Если включаемый в рассмотрение качественный признак имеет не два, а несколько значений, то в принципе можно было бы ввести дискретную переменную, принимающую такое же количество значений. Но этого фактически никогда не делают, так как тогда трудно дать содержательную интерпретацию соответствующему коэффициенту. В этих случаях целесообразно использовать несколько бинарных или фиктивных переменных.

Типичным примером подобной ситуации является исследование сезонных колебаний.

Пусть, например, y_i – объем потребления некоторого продукта в i–ый месяц, например, мороженого, и есть все основания считать, что потребление зависит от времени года. Для выявления влияния сезонности можно ввести три фиктивные переменные d_1, d_2, d_3:

$$d_{i1} = \begin{cases} 1, & \text{если месяц i является зимним} \\ 0, & \text{в остальных случаях;} \end{cases}$$

$$d_{i2} = \begin{cases} 1, & \text{если месяц i является весенним} \\ 0, & \text{в остальных случаях;} \end{cases}$$

$$d_{i3} = \begin{cases} 1, & \text{если месяц i является летним} \\ 0, & \text{в остальных случаях} \end{cases}$$

и оценивать уравнение

$$y_i = b_0 + b_1 d_{i1} + b_2 d_{i2} + b_3 d_{i3} + e_i . \qquad (3.3)$$

Отметим, что мы не вводим четвертую переменную d_4, относящуюся к осени, иначе тогда для любого месяца i выполнялось бы тождество $d_{i1} + d_{i2} + d_{i3} + d_{i4} = 1$, что означало бы линейную зависимость регрессоров в (3.3) и, как следствие, невозможность получения МНК-оценок. Интерпретация коэффициентов в (3.3) будет такой:

среднемесячный объем потребления для осенних месяцев – $\hat{b}_0$,

для зимних – $(\hat{b}_0 + \hat{b}_1)$,

для весенних – $(\hat{b}_0 + \hat{b}_2)$,

для летних – $\left(\hat{b}_0 + \hat{b}_3\right)$.

Таким образом, оценки коэффициентов $\hat{b}_i$, $i = 1,2,3$, показывают средние сезонные отклонения в объеме потребления по отношению к осенним месяцам. Тестируя, например, стандартную гипотезу $H_0 : b_3 = 0$, мы проверяем предположение о несущественном различии в объеме потребления между летним и осенним сезоном. Гипотеза $H_0 : b_1 = b_2$ эквивалентна предположению об отсутствии различия в потреблении между зимой и весной и т. д.

Фиктивные переменные, несмотря на свою внешнюю простоту, являются весьма гибким инструментом при исследовании влияния качественных признаков. Кроме этого фиктивные переменные позволяют строить и оценивать так называемые кусочно-линейные модели, которые можно применять для исследования структурных изменений.

Рассмотрим пример. Пусть y – зависимая переменная и пусть для простоты есть только две независимые переменные: x и постоянный (свободный) член. Предположим, что x и y представлены в виде временных рядов $\{(x_t, y_t)$, $t = 1, \mathbf{K}, n\}$. Например, x_t – размер основного фонда некоторого предприятия в период t, y_t – объем продукции, выпущенной в этот же период.

Из некоторых априорных соображений исследователь считает, что в момент t_0 произошла структурная перестройка и линия регрессии будет отличаться от той, что была до момента t_0, но общая линия остается непрерывной (см. рис. 3.2).

Чтобы оценить такую модель введем бинарную переменную R_t, полагая $R_t = \begin{cases} 0, & t \le t_0 \\ 1, & t > t_0 \end{cases}$, и запишем следующее регрессионное уравнение

$$y_t = b_1 + b_2 x_t + b_3 \left(x_t - x_{t_0}\right) R_t + e_t . \qquad (3.4)$$

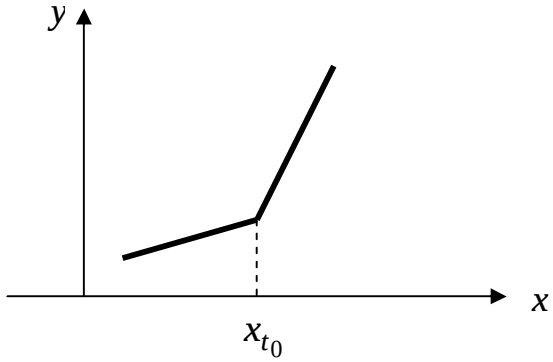


Рис. 3.2

Нетрудно проверить, что линия регрессии, соответствующая уравнению (3.4), имеет коэффициент наклона b_2 для $t \le t_0$ и $b_2 + b_3$ для $t > t_0$, и разрыва в точке x_{t_0} не происходит.

Действительно, для $t > t_0$ имеем

$$y_t = b_1 + b_2 x_t + b_3 x_t - b_3 x_{t_0} + e_t$$

или

$$y_t = b_1 + (b_2 + b_3) x_t - b_3 x_{t_0} + e_t,$$

т. е. угловой коэффициент равен $b_2 + b_3$. Таким образом, тестируя гипотезу $H_0 : b_3 = 0$, мы проверяем предположение о том, что фактически структурного изменения не произошло.

В заключение отметим, что с помощью фиктивных переменных можно исследовать влияние разных качественных признаков (например, уровень образования и наличие или отсутствие детей), а также их взаимное влияние (эффект взаимодействия). Следует только быть внимательным, чтобы при включении нескольких бинарных переменных не нарушить линейную независимость регрессоров (см. пример с сезонными колебаниями).

Пример 3.1. ([4]). Рынок квартир в Москве (данные для этого исследования собраны студентами РЭШ в 1994 и 1996 гг.).

После проведенного анализа по $n = 464$ наблюдениям была выбрана логарифмическая форма модели:

$$\ln y = 7{,}106 + 0{,}670 \ln x_1 + 0{,}431 \ln x_2 + 0{,}147 \ln x_3 - 0{,}114 \ln x_4 - 0{,}0686 d_1 +$$
$$0{,}134 d_2 + 0{,}042 d_3 + 0{,}114 d_4 + 0{,}214 d_5 + 0{,}140 d_6 + 0{,}164 d_7 + 0{,}169 d_8,$$

где

y – цена квартиры (в долларах США),

x_1 – жилая площадь (в кв.м.),

x_2 – площадь нежилых помещений (в кв.м.),

x_3 – площадь кухни (в кв.м.),

x_4 – расстояние от центра Москвы (в км).

Фиктивные переменные:

$$d_1 = \begin{cases} 1, & \text{если квартира на 1-ом или последнем этаже,} \\ 0, & \text{в противном случае,} \end{cases}$$

$$d_2 = \begin{cases} 1, & \text{если квартира в кирпичном доме,} \\ 0, & \text{в противном случае,} \end{cases}$$

$$d_3 = \begin{cases} 1, & \text{если в квартире есть балкон,} \\ 0, & \text{в противном случае,} \end{cases}$$

$$d_4 = \begin{cases} 1, & \text{если в доме есть лифт,} \\ 0, & \text{в противном случае,} \end{cases}$$

$$d_5 = \begin{cases} 1, & \text{для однокомнатных квартир,} \\ 0, & \text{для всех остальных,} \end{cases}$$

$$d_6 = \begin{cases} 1, & \text{для двухкомнатных квартир,} \\ 0, & \text{для всех остальных,} \end{cases}$$

$$d_7 = \begin{cases} 1, & \text{для трехкомнатных квартир,} \\ 0, & \text{для всех остальных,} \end{cases}$$

$$d_8 = \begin{cases} 1, & \text{для четырехкомнатных квартир,} \\ 0, & \text{для всех остальных.} \end{cases}$$

Из анализа t–статистик получено,что все коэффициенты регрессии, кроме коэффициентов при d_5 и d_6, значимы при доверительной вероятности $g = 0{,}95$.

Коэффициент при $\ln x_1$, равный 0,67, означает, что увеличение жилой площади квартиры на 1% увеличивает ее цену на 0,67%. Иначе говоря, эластичность цены квартиры по жилой площади равна 0,67.

Отрицательное значение коэффициента при x_4 (-0,114) означает, что увеличение расстояния от центра города на 1% уменьшает цену квартиры на 0,11%.

Рассмотрим интерпретацию фиктивных переменных $d_1, \mathbf{K}, d_8$.

Отрицательный коэффициент при d_1 означает, что квартира на 1-ом или последнем этаже стоит на 6,9% дешевле аналогичной квартиры на средних этажах. Квартира в кирпичном доме стоит на 13,4% дороже аналогичной квартиры в панельном доме, присутствие лифта увеличивает стоимость на 11,4%, а наличие балкона – на 4,2%.

Переменные d_5, d_6, d_7, d_8 были включены в регрессию, чтобы учесть возможные различия в структуре рынка жилья для квартир с разным количеством комнат. Отмечается, что в выборке были 5-ти,6-ти и даже 8-ми комнатные квартиры, поэтому переменные $d_5 + d_6 + d_7 + d_8 \neq 1$ (т. е. в сумме не дают константу, что означает отсутствие полной коллинеарности факторов).

Было показано, что коэффициенты при d_6, d_7, d_8 можно считать равными. Из уравнения регрессии видно, что квартиры с числом комнат от 2 до 4 стоят дороже многокомнатных, а однокомнатные – еще дороже (при прочих равных условиях).

3.3. Частная корреляция

В том случае, когда имеется одна независимая переменная x и одна зависимая y, естественной мерой их линейной связи является (выборочный) коэффициент корреляции r_B (1.18) или парный коэффициент корреляции r_{yx}. Для многомерной регрессии мы можем найти значения таких коэффициентов для y и каждой из независимых переменных $x_1, \mathbf{K}, x_k$. Из парных коэффициентов корреляции можно составить матрицу парных коэффициентов корреляции и

сделать вывод о наличии или отсутствии в построенной модели мультиколлинеарности факторов.

Высокое значение коэффициента корреляции между исследуемой зависимой и какой-либо независимой переменной может, как и раньше, означать высокую степень зависимости, но может быть обусловлено и другой причиной. А именно, есть третья переменная, которая оказывает сильное влияние на две первые, что и служит в конечном счете причиной их высокой коррелированности. Поэтому возникает естественная задача найти «чистую» корреляцию между двумя переменными, исключая (линейное) влияние других факторов. Это можно сделать с помощью коэффициента частной корреляции.

Для простоты предположим, что имеется модель парной регрессии

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e,$$

где Y – $(n\times 1)$ вектор наблюдений зависимой переменной, X_1, X_2 – $(n\times 1)$ векторы независимых переменных, b_0, b_1, b_2 – (скалярные) параметры, e – $(n\times 1)$ вектор ошибок. Наша цель – определить корреляцию между y и, например, первым регрессором x_1 после исключения влияния x_2.

Соответствующая процедура устроена следующим образом:

1) Осуществим регрессию Y на X_2 и константу и получим прогнозные значения

$$\hat{Y} = \hat{a}_1 + \hat{a}_2 X_2 \ ;$$

2) Осуществим регрессию X_1 на X_2 и константу и получим прогнозные значения

$$\hat{X}_1 = \hat{g}_1 + \hat{g}_2 X_2 \ ;$$

3) Удалим влияние X_2, взяв остатки $e_Y = Y - \hat{Y}$ и $e_{X_1} = X_1 - \hat{X}_1$;

4) Определим (выборочный) коэффициент частной корреляции между y и x_1 при исключении влияния x_2 как (выборочный) коэффициент корреляции между e_Y и e_{X_1}:

$$r_{yx_1/x_2} = r_{e_Y, e_{X_1}}. \qquad (3.5)$$

Из свойств МНК следует, что остатки e_Y и e_{X_1} не коррелированы с X_2. Именно в этом смысле указанная процедура соответствует интуитивному представлению об «исключении» (линейного) влияния переменной x_2.

Прямыми вычислениями можно показать, что справедлива следующая формула, связывающая коэффициенты частной и обычной корреляции:

$$r_{yx_1/x_2} = \frac{r_{yx_1} - r_{yx_2} r_{x_1x_2}}{\sqrt{\left(1 - r_{yx_2}^2\right)\left(1 - r_{x_1x_2}^2\right)}}. \qquad (3.6)$$

Здесь значения частного коэффициента корреляции r_{yx_1/x_2} лежат в интервале $[-1;1]$ как у обычного коэффициента корреляции. Если $r_{yx_1/x_2} = 0$, то гово-

ря нестрого, это означает отсутствие прямого (линейного) влияния переменной x_1 на y.

Существует тесная связь между коэффициентом частной корреляции r_{yx_1/x_2} и коэффициентом детерминации R^2, а именно:

$$1-R^2=\left(1-r_{yx_2}^2\right)\cdot\left(1-r_{yx_1/x_2}^2\right).$$

Описанная выше процедура очевидным образом обобщается на случай, когда исключается влияние не одной, а нескольких переменных: достаточно переменную x_2 заменить на набор переменных $x_2, \mathbf{K}$, сохраняя определение (3.5). Формула (3.6) естественно усложнится. Подробнее об этом можно прочесть в книге [3; 6].

Проиллюстрируем приведенное выше понятие частных коэффициентов корреляции и их отличие от обычных коэффициентов корреляции на следующем примере.

Пример 3.2. Изучается зависимость выработки продукции на одного работника (y – млн. руб.) от ввода в действие новых основных фондов (в % от стоимости фондов на конец года, x_1 – коэффициент обновления основных фондов) и от удельного веса рабочих высокой квалификации в общей численности рабочих (x_2 – %). По результатам наблюдений с использованием ППП Статграф были обоснованы гипотезы, лежащие в основе множественного регрессионного анализа. В результате получено уравнение

$$\hat{y}=1{,}8353+0{,}9459x_1+0{,}0856x_2.$$

Здесь $\hat{b}_0=1{,}8353$ оценивает агрегированное влияние прочих (кроме x_1, x_2) факторов на объясняемую переменную y; $\hat{b}_1$ и $\hat{b}_2$ указывают, что с увеличением x_1 и x_2 на единицу их значений, результат увеличивается, соответственно, на 0,9459 млн. руб. и на 0,0856 млн. руб. Сравнивать эти значения не следует, т.к. они зависят от единиц измерения каждого признака и потому несопоставимы между собой.

Соответствующие t – статистики: $t_{\hat{b}_0}=3{,}9$, $t_{\hat{b}_1}=4{,}45$, $t_{\hat{b}_2}=1{,}42$. Так как $t_{крит}\approx 2-3$, то b_2 – статистически незначим, т.е. x_2 можно исключить из модели как несущественно влияющий или неинформативный.

Значения линейных коэффициентов парной корреляции, представленные ниже в матрице парных коэффициентов, определяют тесноту парных зависимостей переменных, указанных в данном уравнении множественной регрессии.

Парная корреляция

	y	x_1	x_2
y	1	0,9699	0,9408
x_1	0,9699	1	0,9428
x_2	0,9408	0,9428	1

$r_{yx_1} = 0{,}9699$ и $r_{yx_2} = 0{,}9408$ говорит о весьма тесной связи выработки «y» как с коэффициентом обновления основных фондов – x_1, так и с долей рабочих высокой квалификации – x_2.

Межфакторная связь $r_{x_1x_2} = 0{,}9428$ весьма тесная и превышает тесноту связи x_2 с y, $r_{yx_2} = 0{,}9408$. Связь между x_1, x_2: $r_{x_1x_2} = 0{,}9428$, т.е. имеет место мультиколлинеарность факторов.

Ниже в матрице приведены линейные коэффициенты частной корреляции, которые оценивают тесноту связи значений двух переменных, исключая влияние всех других переменных, представленных в уравнении множественной регрессии:

Частная корреляция

	y	x_1	x_2
y	1	0,7335	0,3247
x_1	0,7335	1	0,3679
x_2	0,3247	0,3679	1

Коэффициенты частной корреляции дают более точную характеристику тесноты зависимости двух признаков, чем коэффициенты парной корреляции, так как «очищают» парную зависимость от взаимодействия данной пары признаков с другими признаками, представленными в модели.

Наиболее тесно связаны y и x_1, $r_{yx_1/x_2} = 0{,}7335$, связь y с x_2 гораздо слабее, т. к. $r_{yx_2/x_1} = 0{,}3247$, а межфакторная зависимость x_1 и x_2 выше, чем парная частная y и x_2, $r_{x_1x_2/y} = 0{,}3679 > r_{yx_2/x_1} = 0{,}3247$. Все это приводит к выводу о необходимости исключить фактор x_2 – доля высококвалифицированных рабочих – из правой части уравнения множественной регрессии.

Если сравнить коэффициенты парной и частной корреляции, то можно увидеть, что из-за высокой межфакторной зависимости коэффициенты парной корреляции дают завышенные оценки тесноты связи:

$$r_{yx_1} = 0{,}9699 \sim r_{yx_1/x_2} = 0{,}7335$$
$$r_{yx_2} = 0{,}9408 \sim r_{yx_2/x_1} = 0{,}3247.$$

Именно по этой причине рекомендуется при наличии сильной коллинеарности (мультиколлинеарности) факторов исключать из исследования тот фактор, у которого теснота парной зависимости меньше, чем теснота межфакторной связи $\begin{pmatrix} 0{,}9408 < 0{,}9428 \\ 0{,}3247 < 0{,}3679 \end{pmatrix}$.

Контрольные вопросы, задачи и упражнения

3.1. Что такое полная коллинеарность и мультиколлинеарность факторов? Перечислите характерные признаки мультиколлинеарности.

3.2. Какие из перечисленных факторов учитываются в регрессии с помощью фиктивных переменных: 1) профессия, 2) курс доллара, 3) численность населения, 4) размер среднемесячных потребительских расходов, 5) местоположение пункта продажи?

3.3. С помощью фиктивных переменных напишите уравнение, соответствующее наличию двух структурных изменений в моменты времени t_0 и t_1, $t_0 < t_1$.

3.4. Предположим, что вы оцениваете регрессионную зависимость расходов на мороженое от располагаемого личного дохода, используя наблюдения по месяцам. Объясните, как вы введете фиктивные переменные для оценки сезонных колебаний? Какую интерпретацию дадите коэффициентам регрессии? Какие гипотезы сможете протестировать?

3.5. Рассчитайте парные и частные коэффициенты корреляции для данных примера 2.1. Сделайте вывод о наличии или отсутствии в модели мультиколлинеарности факторов.

Глава 4. Системы эконометрических уравнений

При моделировании достаточно сложных экономических объектов часто приходится использовать не одно, а несколько уравнений, чаще всего связанных между собой. В таких случаях модель объекта описывается системой эконометрических уравнений, которую необходимо оценить при проведении регрессионного анализа. Проблема оценивания систем уравнений требует введения новых понятий и разработки новых методов. Эти вопросы и будут обсуждаться в данной главе. Вначале мы рассмотрим простую задачу оценивания системы, в которой уравнения связаны потому, что ошибки в разных уравнениях коррелированы между собой, – это так называемая система внешне не связанных уравнений. Затем мы исследуем общие системы, которые в эконометрике называются системами одновременных уравнений, и частный случай таких систем – рекурсивные системы.

4.1. Внешне не связанные уравнения

Для того, чтобы понять постановку задачи и суть проблемы, рассмотрим следующий пример. Предположим, что исследуется зависимость инвестиций y, осуществляемых некоторым предприятием (например, Иркутским алюминиевым заводом), от его дохода x_1 и размера основного фонда x_2:

$$y_i = a_0 + a_1 x_{1i} + a_2 x_{2i} + e_i,\ i = 1, \mathbf{K}, n. \qquad (4.1)$$

Представим теперь, что имеется ряд наблюдений другого аналогичного предприятия (например, Братского алюминиевого завода):

$$z_i = b_0 + b_1 t_{1i} + b_2 t_{2i} + h_i,\ i = 1, \mathbf{K}, n. \qquad (4.2)$$

Уравнения (4.1) и (4.2) можно оценивать по отдельности. Внешне они выглядят как не связанные друг с другом. Но ясно, что в данной ситуации ошибки e_i и h_i коррелированы, так как для каждого $i = 1, \mathbf{K}, n$ (или $t = 1, \mathbf{K}, T$) предприятия действуют в «одной экономической среде». Поэтому целесообразно объединить уравнения (4.1) и (4.2) и оценивать их совместно, используя доступный обобщенный метод наименьших квадратов [4; 6].

Общую задачу можно сформулировать следующим образом. Даны M регрессионных уравнений (в матричном виде)

$$\begin{aligned} Y_1 &= X_1 B_1 + e_1 \\ Y_2 &= X_2 B_2 + e_2 \\ &\mathbf{KKKKK} \\ Y_M &= X_M B_M + e_M, \end{aligned} \qquad (4.3)$$

где Y_i – $(n \times 1)$ вектор зависимых переменных, X_i – $(n \times k_i)$ матрица независимых переменных, B_i – $(k_i \times 1)$ вектор неизвестных параметров, e_i – $(n \times 1)$ вектор ошибок, $i = 1, \mathbf{K}, M$. Будем предполагать, что $Me_i = 0$ и $M(e_{is} e_{jt}) = s_{ij}$ при $s = t$ и $M(e_{is} e_{it}) = 0$ при $s \neq t$. Последнее условие можно представить так:

$$M\left(e_i e_j^T\right) = s_{ij} I_n,\ i, j = 1, \mathbf{K}, M\ . \qquad (4.4)$$

Иными словами, заданы M регрессионных уравнений, по каждому из которых имеется n наблюдений. (Или T наблюдений в случае временных рядов). Если данные имеют структуру временных рядов, то считается, что ошибки во всех уравнениях коррелированы в один и тот же момент времени и некоррелированы для других моментов.

Каждое отдельное уравнение в системе (4.3) удовлетворяет условиям классической регрессионной модели и может быть оценено обычным МНК.Однако, если объединить эти уравнения и применить ОМНК, то можно повысить эффективность оценивания.

Обозначим

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \text{M} \\ Y_M \end{pmatrix}, \qquad X = \begin{pmatrix} X_1 & 0 & \mathbf{K} & 0 \\ 0 & X_2 & \mathbf{K} & 0 \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ 0 & 0 & \mathbf{K} & X_M \end{pmatrix}, \qquad B = \begin{pmatrix} B_1 \\ B_2 \\ \text{M} \\ B_M \end{pmatrix}, \qquad e = \begin{pmatrix} e_1 \\ e_2 \\ \text{M} \\ e_M \end{pmatrix},$$

$$\sum = \left(s_{ij}\right), \quad i, j = 1, \mathbf{K}, M\ .$$

Тогда система (4.3) перепишется в виде

$$Y = XB + e\ .$$

Используя понятие произведения Кронекера двух матриц, ковариационную матрицу вектора ошибок можно представить так:

$$M\left(ee^T\right) = \Omega = \Sigma \otimes I_n\ .$$

В качестве справки приведем пример произведения Кронекера двух матриц:

$$\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix} \otimes \begin{bmatrix} 2 & 0 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 1\begin{bmatrix} 2 & 0 \\ 1 & 4 \end{bmatrix} & 2\begin{bmatrix} 2 & 0 \\ 1 & 4 \end{bmatrix} \\ 0\begin{bmatrix} 2 & 0 \\ 1 & 4 \end{bmatrix} & 3\begin{bmatrix} 2 & 0 \\ 1 & 4 \end{bmatrix} \end{bmatrix} = \begin{bmatrix} 2 & 0 & 4 & 0 \\ 1 & 4 & 2 & 8 \\ 0 & 0 & 6 & 0 \\ 0 & 0 & 3 & 12 \end{bmatrix}.$$

Предположим,что матрица Σ не вырождена.

Для построения оценки $\hat{B}$ применим ОМНК:

$$\hat{B} = \left(X^T \Omega^{-1} X\right)^{-1} X^T \Omega^{-1} Y \text{ или}$$

$$\hat{B} = \left[X^T \left(\Sigma^{-1} \otimes I_n\right) X\right]^{-1} X^T \left(\Sigma^{-1} \otimes I_n\right) Y\ . \qquad (4.5)$$

(здесь мы воспользовались известным свойством произведения Кронекера: для двух квадратных невырожденных матриц A и B справедливо $\left(A \otimes B\right)^{-1} = A^{-1} \otimes B^{-1}$).

Нетрудно понять, что в общем случае оценка (4.5) отличается от оценки, полученной в результате применения обычного МНК к каждому уравнению в системе (4.3). Есть, однако, две ситуации, когда эти оценки совпадают.

1) Уравнения в (4.3) действительно не связаны друг с другом, т. е. $s_{ij} = 0$ при $i \neq j$.

2) Все уравнения в (4.3) имеют один и тот же набор независимых переменных, т. е. $X_1 = X_2 = \mathbf{K} = X_M$.

Для использования доступного ОМНК нужно оценить матрицу Σ. Это можно сделать, применяя к каждому уравнению системы (4.3) обычный МНК, получая векторы остатков e_i, $i = 1, \mathbf{K}, M$, и беря в качестве оценок ковариаций s_{ij} величины $\left(e_i^T e_j\right)/n$, т. е. $\hat{s}_{ij} = \left(e_i^T e_j\right)/n$. Можно проверить, что эти оценки являются состоятельными.

Отметим в заключение, что эффективность $\hat{B}$, полученной таким способом, тем выше, чем сильнее корреляция между ошибками.

4.2. Системы одновременных уравнений

В теории экономико-статистического моделирования систему взаимосвязанных регрессионных уравнений и тождеств, в которой одни и те же переменные в различных регрессионных уравнениях могут одновременно выступать и в роли результирующих показателей (эндогенных переменных) и в роли объясняющих (экзогенных) переменных, принято называть системой одновременных (эконометрических) уравнений.

Как мы уже сказали, эконометрическая модель содержит так называемые эндогенные и экзогенные переменные. Эндогенными являются те переменные, которые в силу принятых концепций определяются внутренней структурой изучаемого явления, иначе говоря, их значения выясняются на основе модели. В свою очередь, экзогенные переменные по определению независимы от структуры явления и их значения (в том числе прогностические) устанавливаются вне модели. Модель содержит также различного рода параметры (коэффициенты), которые определяются в ходе статистического оценивания путем обработки имеющейся информации.

То, как классифицированы переменные (эндогенные или экзогенные) зависит от теоретической схемы или принятой модели. Внеэкономические переменные, например, климатические условия, постоянно бывают экзогенными. В то же время экономические переменные, такие как экспорт и правительственные расходы, могут в одной модели рассматриваться как эндогенные, а в другой – как экзогенные. При этом в соотношения могут входить переменные, относящиеся не только к периоду t, но и к предшествующим периодам, называемые лаговыми («запаздывающими») переменными.

Для экономистов большой интерес представляет количественный анализ модели, т. е. нахождение оценок параметров на основании имеющейся в распоряжении исследователя информации о значениях переменных. Первая из возникающих здесь проблем: можно ли в предложенной модели однозначно восстановить значение некоторого параметра или же его определение принципи-

ально невозможно на основе рассматриваемой модели? Это так называемая проблема идентифицируемости – первоочередная на этапе формирования модели, поскольку прежде, чем переходить к процедурам оценивания необходимо быть уверенным, что их применение имеет смысл.

Проблема оценивания здесь также имеет свои особенности. Основная трудность состоит в том, что в эконометрических моделях переменная, играющая роль независимой (объясняющей – экзогенной) переменной в одном соотношении, может быть зависимой в другом. Это приводит к тому, что в регрессионных уравнениях системы экзогенные переменные и случайные возмущения оказываются, вообще говоря, коррелированными. Наконец, в современной практике встречаются модели, имеющие десятки и даже сотни уравнений (в том числе и нелинейных), в связи с чем возникают и вычислительные трудности.

Все это обусловило необходимость построения специальной теории, изучающей статистический аспект таких моделей. К настоящему времени довольно хорошо разработан ее раздел, относящийся к моделям, описываемым системами линейных уравнений. Основные положения этой теории мы и изучим с вами.

Прежде чем перейти к формулировке общей линейной модели, рассмотрим вначале два примера простой классической макромодели. В первом отсутствует случайное возмущение (мы его опустим для упрощения выкладок).

Пример 4.1. Рассмотрим простую макромодель, которую мы уже обсуждали во введении (см. (В.1) – (В.3)), но которую мы приводим здесь для того, чтобы проиллюстрировать основные понятия, характерные для систем одновременных уравнений.

Итак, предположим, что потребление C есть возрастающая функция от имеющегося в наличии дохода Y, но возрастающая медленнее, чем рост дохода

$$C_t = a_0 + a_1(Y_t - T_t),\ 0 < a_1 < 1. \qquad (4.6)$$

Объем инвестиций есть возрастающая функция национального дохода и убывающая функция характеристики государственного регулирования (например, нормы процента), т. е.

$$I_t = b_1 Y_{t-1} + b_2 R_t,\ b_1 > 0,\ b_2 < 0. \qquad (4.7)$$

И, наконец, национальный доход есть сумма потребительских, инвестиционных и государственных закупок товаров и услуг (условие макроэкономического равновесия):

$$Y_t = C_t + I_t + G_t. \qquad (4.8)$$

Здесь T_t – подоходный налог в момент t, R_t – инструмент государственного регулирования в момент t, G_t – государственные закупки товаров и услуг в момент времени t.

Соотношения (4.6)–(4.8) следует рассматривать как систему одновременных уравнений, так как одна и та же переменная, например, национальный до-

ход Y_t в момент t играет роль объясняемой переменной в (4.8) и объясняющей – в (4.6).

Проведем классификацию переменных модели:

C_t, Y_t, I_t – текущие эндогенные переменные;

T_t, R_t, G_t – текущие экзогенные переменные;

Y_{t-1} – лаговая эндогенная переменная.

Модель предназначена для объяснения значений эндогенных переменных в текущем периоде времени t на основе значений, принимаемых экзогенными и лаговыми эндогенными переменными. В более общих ситуациях в модели могут появиться и лаговые значения экзогенных переменных. Оба множества экзогенных (текущих и лаговых) и лаговые эндогенные переменные называют предопределенными переменными. Схематически работа модели в последовательные моменты времени может быть описана с помощью диаграммы:

Соотношения (4.6)–(4.8) описывают структурную форму модели. Приведенная форма получится, если каждая из текущих эндогенных переменных выразится в виде функции только предопределенных переменных.

Подставляя (4.7) и (4.8) в (4.6), получим

$$C_t = a_0 + a_1(C_t + I_t + G_t - T_t) = a_0 + a_1(C_t + b_1 Y_{t-1} + b_2 R_t + G_t - T_t), \text{ т. е.}$$

$$C_t = \frac{a_0}{1-a_1} + \frac{a_1 b_1}{1-a_1} Y_{t-1} + \frac{a_1 b_2}{1-a_1} R_t + \frac{a_1}{1-a_1}(G_t - T_t), \qquad (4.9)$$

$$I_t = b_1 Y_{t-1} + b_2 R_t \qquad (4.10)$$

(инвестиционное уравнение в своем первоначальном виде имеет приведенную форму, так как в нем нет других текущих эндогенных переменных, кроме I_t).

Затем, используя (4.8), (4.9) и (4.10), получим

$$Y_t = \frac{a_0}{1-a_1} + \frac{a_1 b_1}{1-a_1} Y_{t-1} + \frac{a_1 b_2}{1-a_1} R_t + \frac{a_1}{1-a_1}(G_t - T_t) + b_1 Y_{t-1} + b_2 R_t + G_t =$$

$$= \frac{a_0}{1-a_1} + \left(\frac{a_1 b_1}{1-a_1} + b_1\right) Y_{t-1} + \left(\frac{a_1 b_2}{1-a_1} + b_2\right) R_t + \left(\frac{a_1}{1-a_1} + 1\right) G_t - \frac{a_1}{1-a_1} T_t, \text{ отсюда}$$

$$Y_t = \frac{a_0}{1-a_1} + \frac{b_1}{1-a_1} Y_{t-1} + \frac{b_2}{1-a_1} R_t + \frac{1}{1-a_1} G_t - \frac{a_1}{1-a_1} T_t \ . \qquad (4.11)$$

Уравнения (4.9)–(4.11) образуют приведенную форму модели. Все коэффициенты в приведенной форме модели представляют собой функции первона-

чальных коэффициентов ее структурной формы. При этом особое значение придается коэффициентам при экзогенных переменных. Эти коэффициенты часто интерпретируют как импульсные мультипликаторы, поскольку они показывают реакцию в текущем периоде каждой эндогенной переменной на изменение текущего значения любой экзогенной переменной. Например, увеличение на единицу значения переменной, отражающей государственное регулирование, вызовет изменение C_t на $\frac{a_1 b_2}{1-a_1}$, а I_t на b_2. Поскольку модель линейная, эффект от одновременного изменения экзогенных переменных будет равен сумме частных эффектов. Так, одновременное увеличение на единицу объема государственных закупок G_t и налога T_t оставит потребление C_t и инвестиции I_t неизменными, так как $\frac{a_1}{1-a_1} - \frac{a_1}{1-a_1} = 0$, и инвестиции I_t вообще не зависят от G_t и T_t, а соответствующий прирост национального дохода будет равен единице, так как $\frac{1}{1-a_1} - \frac{a_1}{1-a_1} = \frac{1-a_1}{1-a_1} = 1$.

Пример 4.2. В этом примере мы введем в модель случайную составляющую для того, чтобы проиллюстрировать те особенности, которые возникают при реализации известных процедур оценивания неизвестных параметров модели.

Пусть модель содержит функцию спроса и тождество, определяющее доход:

$$C_t = a + bY_t + e_t \qquad (4.12)$$

$$Y_t = C_t + Z_t . \qquad (4.13)$$

Содержательный смысл модели спроса состоит в утверждении, что потребительские расходы, т. е. спрос, пропорционален доходу. В свою очередь доход есть сумма потребительских и непотребительских расходов.

Математическую формулировку модели представляют соотношения (4.12), (4.13), где C_t – потребительские расходы, Y_t – доход, Z_t – непотребительские расходы, e_t – случайная составляющая (учитывающая неполноту информации, незамкнутость системы и т. п.), рассматриваемые в момент времени t. Предполагается, что Z принимает множество значений, определяемых вне модели. Например, Z может определяться руководителями общества каким-либо способом, не зависящим от C и Y. Будем считать C и Y эндогенными переменными, т. е. переменными, значения которых определяются в результате одновременного взаимодействия образующих модель соотношений, а Z – экзогенной переменной, значения которой определяются вне модели. Случайные величины e_t, $t = 1, \mathbf{K}, n$, некоррелированы, имеют нулевые средние и одинаковые дисперсии s^2, т. е. $M(e_t) = 0$ для всех $t = 1, \mathbf{K}, n$;

$$M(\varepsilon_t \varepsilon_{t+s}) = \begin{cases} 0, & \text{для } s \neq 0 \text{ и всех } t \quad (4.14a) \\ \sigma^2, & \text{для } s = 0 \text{ и всех } t \quad (4.14b) \end{cases}$$

Предполагается также, что Z и e независимы (это свойство удовлетворяется как для переменной Z, принимающей множество фиксированных значений, так и для переменной Z, принимающей произвольные значения, распределенные случайным и независимым от e образом).

Требуется оценить параметры модели a, b и s^2. Если наша задача состоит в получении «хороших» оценок параметров модели (4.12), то мы прежде всего можем рассмотреть применение обычного метода наименьших квадратов. Условия (4.14a) и (4.14b) означают отсутствие как гетероскедастичности, так и автокорреляции. Поэтому для обоснования применения МНК остается только решить вопрос о независимости между e и Y. Подставляя (4.12) в (4.13), получим

$$Y_t = a + bY_t + Z_t + e_t \text{ или}$$

$$Y_t = \frac{a}{1-b} + \frac{1}{1-b} Z_t + \frac{e_t}{1-b}.$$

Так как $MY_t = \frac{a}{1-b} + \frac{1}{1-b} Z_t$, то $Y_t - MY_t = \frac{e_t}{1-b}$ и

$$\operatorname{cov}(e_t, Y_t) = M\{e_t(Y_t - MY_t)\} = \frac{1}{1-b} M\left(e_t^2\right) \neq 0.$$

Таким образом, входящие в уравнение случайная составляющая и объясняющая переменная оказываются коррелированными, а потому непосредственное применение к (4.12) метода наименьших квадратов приведет к смещенным оценкам параметров a и b. Это смещение возникает в случае конечных выборок, однако, оценки, найденные обычным МНК, будут к тому же и несостоятельными, т. е. смещение в оценках сохранится для бесконечно больших выборок.

Так как корреляция между e и Y в уравнении (4.12) приводит к нежелательным последствиям, естественно рассмотреть альтернативные методы оценивания, которые позволяют их избежать. Такие методы, как косвенный МНК, двух-, трехшаговые МНК, пригодные для решения подобных задач, и будут рассмотрены нами далее.

Прежде чем обсуждать проблему оценивания одновременных уравнений, рассмотрим матричную спецификацию общей линейной модели и ее идентифицируемость.

Системы одновременных уравнений в матричной форме. Проблема идентификации.

Предположим, что имеется следующая система уравнений для момента времени t:

$$b_{11} y_{1t} + b_{12} y_{2t} + \mathbf{K} + b_{1m} y_{mt} + g_{11} x_{1t} + g_{12} x_{2t} + \mathbf{K} + g_{1k} x_{kt} = e_{1t}$$

$$b_{21}y_{1t} + b_{22}y_{2t} + \mathbf{K} + b_{2m}y_{mt} + g_{21}x_{1t} + g_{22}x_{2t} + \mathbf{K} + g_{2k}x_{kt} = e_{2t}$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \qquad (4.15)$$
$$b_{m1}y_{1t} + b_{m2}y_{2t} + \mathbf{K} + b_{mm}y_{mt} + g_{m1}x_{1t} + g_{m2}x_{2t} + \mathbf{K} + g_{mk}x_{kt} = e_{mt},$$

где через y_{it} обозначены значения эндогенных переменных в момент t, а через x_{jt} – как значения экзогенных переменных, так и лаговые значения эндогенных переменных, e_{it} – случайные возмущения, $i = 1, \mathbf{K}, m$, $j = 1, \mathbf{K}, k$, $t = 1, \mathbf{K}, n$. Эти две последние группы переменных объединены и образуют вместе класс предопределенных переменных.

Совокупность равенств (4.15) и будет системой одновременных уравнений в структурной форме. Структурная форма модели – это система уравнений, отражающая связь между переменными в соответствии с положениями экономической теории и характеризующая структуру экономики или ее сектора. Параметры структурной формы модели называют структурными параметрами. Если модель содержит тождества, то без потери общности их можно назвать уравнениями, в которых структурные параметры при переменных равны 1.

Приведенная форма модели – это система уравнений, в которой каждая эндогенная переменная есть линейная функция от всех предопределенных переменных модели. Для экономической интерпретации применяются структурные уравнения, для прогнозирования – приведенная форма.

Будем считать, что в каждом уравнении один из коэффициентов b при какой-либо эндогенной переменной равен единице – это естественное условие нормировки. Оно позволяет каждое уравнение системы разрешить относительно одной эндогенной переменной.

Введем обозначения

$$Y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \\ \mathrm{M} \\ y_{mt} \end{pmatrix}, \; X_t = \begin{pmatrix} x_{1t} \\ x_{2t} \\ \mathrm{M} \\ x_{kt} \end{pmatrix}, \; e_t = \begin{pmatrix} e_{1t} \\ e_{2t} \\ \mathrm{M} \\ e_{mt} \end{pmatrix},$$

$$B = \begin{pmatrix} b_{11} & b_{12} & \mathbf{K} & b_{1m} \\ b_{21} & b_{22} & \mathbf{K} & b_{2m} \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ b_{m1} & b_{m2} & \mathbf{K} & b_{mm} \end{pmatrix}, \; \Gamma = \begin{pmatrix} g_{11} & g_{12} & \mathbf{K} & g_{1k} \\ g_{21} & g_{22} & \mathbf{K} & g_{2k} \\ \mathbf{K} & \mathbf{K} & \mathbf{K} & \mathbf{K} \\ g_{m1} & g_{m2} & \mathbf{K} & g_{mk} \end{pmatrix}.$$

Тогда система (4.15) перепишется в виде

$$BY_t + \Gamma X_t = e_t. \qquad (4.16)$$

Здесь B – $(m \times m)$ матрица, состоящая из коэффициентов при текущих значениях эндогенных переменных, Γ – $(m \times k)$ матрица из коэффициентов при предопределенных переменных, Y_t, X_t, e_t – вектор-столбцы.

Подчеркнем, что деление переменных на экзогенные и эндогенные должно проводиться вне модели. Одним из основных требований к экзогенным переменным является условие их некоррелируемости с ошибками в каждом наблюдении t. Будем предполагать, что

1) $M(e_t)=0$;

2) $M(e_t e_t^T)=\Sigma$, матрица Σ не зависит от t и положительно определена;

3) векторы e_t и e_s при $t \neq s$ некоррелированы;

4) матрица B невырождена, т. е. $\det B \neq 0$.

Используя условие 4), умножим обе части равенства (4.16) слева на B^{-1}:

$$Y_t = -B^{-1}\Gamma X_t + B^{-1}e_t = \Pi X_t + h_t, \qquad (4.17)$$

где $\Pi = -B^{-1}\Gamma$, $h_t = B^{-1}e_t$.

Полученная система (4.17) будет приведенной формой модели. Элементами матриц B и Γ являются структурные коэффициенты, а элементами матрицы Π в (4.17) – коэффициенты приведенной формы.

Нетрудно понять, что в общем случае эндогенные переменные и ошибки в структурной системе коррелированы, поэтому, как уже отмечалось, применение к какому-либо из уравнений обычного метода наименьших квадратов даст смещенные и несостоятельные оценки структурных коэффициентов. В то же время коэффициенты приведенной формы могут быть состоятельно оценены, поскольку переменные x_t некоррелированы со структурными ошибками e_t и, следовательно, с ошибками приведенной формы модели h_t.

Проблема идентификации или, правильнее сказать, идентифицируемости относится к структурным параметрам, а не к параметрам приведенной формы. Она может быть сформулирована следующим образом: можно ли в предположении, что элементы матрицы Π в (4.17) известны, однозначно определить некоторые или все элементы матриц B и Γ.

Мы не будем здесь давать строгое формальное определение идентифицируемости структурной модели. Подробное изложение этого вопроса можно найти, например, в [1; 4; 6]. Подчеркнем лишь, что тот или иной структурный коэффициент идентифицируем, если он может быть вычислен на основе коэффициентов приведенной формы. Соответственно какое-либо уравнение в структурной форме модели будем называть идентифицируемым, если идентифицируемы все его коэффициенты. Модель считается идентифицируемой, если каждое уравнение системы идентифицируемо. Если хотя бы одно из уравнений системы неидентифицируемо, то и вся модель считается неидентифицируемой. Следует иметь в виду, что проблема идентифицируемости логически предшествует задаче оценивания. Если система не идентифицируема, то это означает, что с имеющимися в нашем распоряжении наблюдениями, независимо от их числа, совместимы многие модели.

Данное уравнение системы точно идентифицировано, если его структурные параметры однозначно определяются по приведенным коэффициентам.

Структурные параметры такого уравнения можно найти косвенным методом наименьших квадратов (см. п. 4.3). Если из приведенной формы модели можно получить несколько оценок структурных параметров, то уравнение сверхидентифицировано. Структурные параметры такого уравнения определяются двухшаговым методом наименьших квадратов. Сверхиндетифицируемая модель содержит хотя бы одно сверхидентифицируемое уравнение. Если структурные параметры уравнения модели нельзя найти через приведенные коэффициенты, то такое структурное уравнение называется неидентифицируемым, и численные оценки его параметров найти нельзя.

Для того чтобы определить, идентифицировано ли структурное уравнение модели, по каждому уравнению и модели в целом подсчитывают: K – число предопределенных переменных модели, k – число предопределенных переменных в каждом уравнении, m – число эндогенных переменных в каждом уравнении. Далее для каждого уравнения в отдельности проверяют следующее соотношение:

$$K - k \geq m - 1. \qquad (4.18)$$

Если число предопределенных переменных, не входящих в уравнение, строго больше числа эндогенных переменных, входящих в уравнение, минус 1 ($K - k > m - 1$), уравнение сверхидентифицировано.

Если число предопределенных переменных, не входящих в уравнение, равно числу эндогенных переменных, входящих в уравнение, минус 1 ($K - k = m - 1$), уравнение точно идентифицировано.

Если число предопределенных переменных, не входящих в уравнение, строго меньше числа эндогенных переменных, входящих в уравнение, минус 1 ($K - k < m - 1$), уравнение неидентифицировано.

Примите во внимание, что нет необходимости исследовать на идентификацию тождества модели, поскольку их структурные параметры известны и равны 1. Однако переменные, входящие в тождества учитываются при подсчете числа эндогенных и предопределенных переменных модели.

Счетное правило отражает необходимое, но недостаточное условие идентификации. Более точно условия идентифицируемости определяются, если накладывать ограничения на коэффициенты матриц параметров структурной модели.

Уравнение идентифицируемо, если по отсутствующим в нем переменным (эндогенным и экзогенным) можно из коэффициентов при них в других уравнениях системы получить матрицу, определитель которой не равен нулю, а ранг матрицы не меньше, чем число эндогенных переменных в системе минус 1.

Для того чтобы проверить достаточное условие идентификации, составляется матрица коэффициентов при переменных модели. В соответствии с достаточным условием идентификации ранг матрицы коэффициентов при переменных, не входящих в уравнение, для которого проверяется достаточное условие, должен быть равен числу эндогенных переменных модели минус единица.

Обсуждая проблему идентификации, следует иметь в виду, что при неполной идентификации невозможно получить оценки некоторых или даже всех параметров. В случае точной идентификации все методы оценивания дают одинаковые результаты.

Пример 4.3. Исследуем на индетифицируемость простую макромодель (4.6)–(4.8). Приведем классификацию переменных модели:

C_t, Y_t, I_t – текущие эндогенные переменные;

T_t, R_t, G_t – текущие экзогенные переменные;

Y_{t-1} – лаговая эндогенная переменная.

Здесь $K = 4$ – число предопределенных переменных модели. Используем необходимое условие идентификации – счетное правило (4.18):

Уравнение 1: $K = 4$, $k = 1$ (T_t), $m = 2$ (Y_t, C_t). Так как $4-1 > 2-1$, уравнение сверхидентифицируемо.

Уравнение 2: $K = 4$, $k = 2$ (Y_{t-1}, R_t), $m = 1$ (I_t). Так как $4-2 > 1-1$ уравнение сверхидентифицируемо.

Тождество (4.8) на идентификацию, как мы уже отмечали, не проверяется. Таким образом, по необходимому условию оба структурных уравнения модели сверхидентифицируемы.

Проверяем для каждого из уравнений достаточное условие. Для этого составим матрицу коэффициентов при переменных модели:

	C_t	I_t	Y_t	T_t	R_t	G_t	Y_{t-1}
Уравнение 1	-1	0	a_1	$-a_1$	0	0	0
Уравнение 2	0	-1	0	0	b_2	0	b_1
Тождество	1	1	-1	0	0	1	0

В соответствии с достаточным условием идентификации ранг матрицы коэффициентов при переменных, не входящих в уравнение, для которого проверяется достаточное условие, должен быть равен числу эндогенных переменных модели минус 1, то есть 3-1=2.

Уравнение 1: матрица коэффициентов при переменных, не входящих в уравнение, имеет вид $A = \begin{pmatrix} -1 & b_2 & 0 & b_1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$.

Ее ранг равен 2, так как $\det A^* = \begin{vmatrix} -1 & b_2 \\ 1 & 0 \end{vmatrix} \neq 0$. Достаточное условие идентификации для уравнения 1 выполняется.

Уравнение 2: выпишем матрицу коэффициентов при переменных, не входящих в это уравнение, $A = \begin{pmatrix} -1 & a_1 & -a_1 & 0 \\ 1 & -1 & 0 & 1 \end{pmatrix}$.

Ранг этой матрицы равен 2, так как $\det A^* = \begin{vmatrix} -1 & a_1 \\ 1 & -1 \end{vmatrix} \neq 0$. Достаточное условие также выполняется.

Таким образом, модель в целом сверхидентифицируема, так как оба ее структурных уравнений сверхидентифицируемы по необходимому и достаточному условиям.

4.3. Методы оценивания систем одновременных уравнений

Как уже отмечалось, независимо от того, хотим ли мы оценить только одно из уравнений системы (4.15) или каждое уравнение этой модели, мы оказываемся в ситуации, когда ни обыкновенный метод наименьших квадратов, ни различные версии обобщенного МНК в общем случае не обеспечивают удовлетворительную процедуру оценивания. Если обыкновенный МНК применить к уравнению модели, в котором присутствуют несколько текущих значений эндогенных переменных, то придется одну из них выбрать в качестве «зависимой» переменной для данного уравнения. Тогда оставшиеся (одно или несколько) текущие значения эндогенных переменных, участвующие в этом соотношении, будут, вообще говоря, коррелировать с ошибками, и потому МНК-оценки параметров модели окажутся смещенными и несостоятельными. Только в случае рекурсивных моделей обыкновенный МНК, как мы увидим ниже, дает оптимальный способ оценивания.

В более общем случае, когда модель состоит из одновременных уравнений, не удовлетворяющих специальным предположениям о рекурсивности, существует простой метод оценивания – косвенный метод наименьших квадратов, но он применим лишь к точно идентифицируемым уравнениям. Этот метод состоит в оценивании обычным МНК параметров приведенной формы и подстановке оценок в выражения для коэффициентов структурной формы через коэффициенты приведенной формы, что приводит к смещенным, но состоятельным оценкам. В случае сверхидентифицируемости косвенный МНК не применим.

Для оценивания произвольных систем одновременных уравнений в настоящее время имеется довольно значительное количество методов, которые делятся на две группы. К первой группе относятся методы, применимые к каждому уравнению в отдельности, т. е. позволяющие оценивать каждое из уравнений поочередно; и вторая группа содержит методы, предназначенные для оценивания всей системы в целом, т. е. всех уравнений сразу.

Примерами первой группы являются двухшаговый метод наименьших квадратов (2МНК), метод максимума правдоподобия с ограниченной информацией, т. е. для одного уравнения, называемый также методом наименьшего дисперсионного соотношения или методом Комиссии Коулса и некоторые другие. Примерами методов второй группы являются трехшаговый метод наименьших квадратов (3МНК) и метод максимального правдоподобия полной информации. Несколько особняком стоят итерационные методы или методы неподвижной точки, которые обладают определенными вычислительными достоинствами, что немаловажно при исследовании систем большой размерности, однако статистические их свойства изучены в недостаточной степени [1; 5].

Существует специальный тип систем одновременных уравнений – так называемые рекурсивные системы, – для которых при определенном выборе порядка и взаимосвязей оцениваемых отдельных уравнений системы процедура МНК приводит к оцениванию всех ее уравнений. С точки зрения задач статистического оценивания этот тип систем одновременных уравнений является простейшим, поэтому мы с него и начнем.

1. Рекурсивные системы

Система одновременных уравнений удовлетворяет свойству рекурсивности, если она построена следующим образом. В качестве 1-го уравнения системы определяют соотношение, в котором присутствует только одна эндогенная переменная y_1 (соответственно, и индексирует ее первым номером). Так что первое уравнение системы содержит одну эндогенную переменную и какое-то количество предопределенных переменных. Второе уравнение системы может содержать не более двух эндогенных переменных; это, если необходимо, y_1 («участница» 1-го уравнения) и y_2. В третье уравнение, кроме уже участвовавших во 2-м уравнении y_1 и y_2, можно включить опять только одну эндогенную переменную y_3 и т. д. В результате мы получим модель вида (4.16), в которой матрица B является нижней треугольной матрицей, т. е. $b_{ij}=0$ при $j>i$ для всех $i=1,\mathbf{K},m$ (при сохранении условия нормировки $b_{ii}=1$). Если для систем такого вида дополнительно потребовать взаимную некоррелируемость случайных ошибок (диагональность ковариационной матрицы: $M\left(ee^T\right)=diag(s_{11},s_{22},\mathbf{K},s_{mm})$) и независимость ее от t, то оценки структурных параметров в каждом отдельном уравнении системы с помощью прямого метода наименьших квадратов будут состоятельными, а при нормальности ошибок – и асимптотически эффективными. Под прямым МНК понимается следующая процедура, последовательно примененная к i–му уравнению системы ($i=1,\mathbf{K},m$): с помощью обычного МНК строятся оценки коэффициентов регрессии y_i по всем включенным в это уравнение эндогенным и предопределенным переменным.

Пример 4.4. Рекурсивная система с тремя эндогенными и одной экзогенной переменной может быть записана в виде

$$y_{1t}+g_{1t}x_t=e_{1t}$$
$$b_{21}y_{1t}+y_{2t}+g_{21}x_t=e_{2t}$$
$$b_{31}y_{1t}+b_{32}y_{2t}+y_{3t}+g_{31}x_t=e_{3t}.$$

Так как e_2 не коррелирует с y_1, а e_3 не коррелирует с y_1 и y_2, то второе и третье уравнения этой системы могут оцениваться путем непосредственного применения обыкновенного МНК, как, впрочем, и первое уравнение, которое содержит одну эндогенную и одну экзогенную переменные. Для применения МНК система переписывается в виде

$$y_{1t}=-g_{1t}x_t+e_{1t}$$

$$y_{2t} = -b_{21}y_{1t} - g_{21}x_t + e_{2t}$$
$$y_{3t} = -b_{31}y_{1t} - b_{32}y_{2t} - g_{31}x_t + e_{3t}.$$

Указанные выше привлекательные свойства рекурсивных систем вызывают желание использовать именно их в эконометрических исследованиях, так как считается, что большинство реальных механизмов формирования рассматриваемых в модели экономических показателей функционируют в рекурсивном (а не одновременном режиме).

Рассмотрим пример спецификации модели в виде рекурсивной системы одновременных уравнений при описании процесса формирования равновесных цен и количеств предлагаемых на рынке товаров.

Пример 4.5. [1] Пусть y_{1t} – цена некоторого товара в момент времени t, а y_{2t} – объем продаж этого товара в тот же момент времени. Естественно предположить, что объем продаж y_{2t} зависит от цены y_{1t} и от объема продаж в предыдущий момент времени y_{2t-1}. В свою очередь, цена товара y_{1t} зависит от объема его продаж в предыдущий момент времени (т. е. y_{2t-1}). В данной схеме цена y_{1t} и объем продаж y_{2t} играют роль эндогенных переменных, а лаговая переменная y_{2t-1} играет роль единственной предопределенной переменной, которую мы обозначим через x_t (т.е. $y_{2t-1} = x_t$). Таким образом, анализируемая ситуация будет описана рекурсивной системой

$$\begin{array}{l} y_{1t} + g_{11}x_t = e_{1t} \\ b_{21}y_{1t} + y_{2t} + g_{21}x_t = e_{2t} \end{array}, \; t = 1, \mathbf{K}, n. \qquad (4.19)$$

Очень важным моментом правильной спецификации этой модели является выбор продолжительности рассматриваемого периода времени. Действительно, продавец устанавливает цены, а покупатель на них реагирует. При этом торговые запасы будут либо накапливаться, либо рассасываться. Продавец среагирует на эту динамику и т. д. Если выбрать в качестве периода один день, то сделанные в модели допущения выглядят естественными, так как последовательность причинных связей $y_{2t-1} \to y_{1t} \to y_{2t}$ является линейной цепью и не содержит никаких петель обратной связи. Это позволяет нам предположить, что ошибки или возмущения, влияющие на спрос (e_{2t}) и предложение (e_{1t}), являются независимыми.

Однако в действительности приходится рассматривать системы, отличные от рекурсивных типа (4.19), в связи с тем, что исследователь обычно располагает некоторыми усредненными (агрегированными) данными. Например, данные о рыночной конъюнктуре могут быть усреднены по недельным или месячным периодам. Предположим, что публикуются не дневные, а только недельные данные о средней недельной цене $\bar{y}_{1t}$ и среднем объеме дневных продаж $\bar{y}_{2t}$. Тогда вынужденное агрегирование соответствующих ошибок $\bar{e}_{1t}$ и $\bar{e}_{2t}$ в системе (4.19) делает их взаимно коррелированными, а саму модель – неиден-

тифицированной. В этой ситуации модель спроса и предложения («крест» Маршалла) представляется более естественной:

$$P_t = a_0 + a_1 Q_t + e_t$$
$$Q_t = b_0 + b_1 P_t + h_t .$$

Здесь использованы привычные для экономистов обозначения: $P_t = \bar{y}_{1t}$ – средняя цена за неделю t, $Q_t = \bar{y}_{2t}$ – средний объем ежедневных продаж за неделю t.

Без введения дополнительных переменных эта модель оказывается теперь даже неидентифицируемой. Однако если бы идентифицирующие ее переменные и существовали, то, как правило, введение их в модель и вынужденное агрегирование по временным периодам может превратить рекурсивную модель в обычную систему одновременных уравнений со всеми вытекающими отсюда проблемами ее оценивания.

2. Косвенный метод наименьших квадратов

Косвенный метод наименьших квадратов (КМНК) (или метод приведенной формы) предназначен для оценивания структурных параметров отдельного уравнения системы и может дать результат (без сочетания с другими методами, например, с двухшаговым методом наименьших квадратов) только в применении к точно идентифицируемому уравнению.

Суть КМНК состоит в следующем. Сначала структурная форма преобразуется в приведенную, затем с помощью МНК оцениваются параметры каждого уравнения приведенной формы модели в отдельности. Наконец, параметры приведенной формы трансформируются в параметры структурной формы модели. Иначе говоря, на этом этапе осуществляется обратный переход от системы с численными параметрами приведенной формы к системе структурной формы. Оценки структурных параметров, полученные КМНК, получаются состоятельными.

Пример 4.6. Для иллюстрации КМНК рассмотрим простую структурную форму

$$y_1 = b_{12} y_2 + g_{11} x_1 + e_1$$
$$y_2 = b_{21} y_1 + g_{22} x_2 + e_2 .$$

Оба уравнения точно идентифицируемы, по необходимому условию

(1) $K = 2 \quad (x_1, x_2), \quad k = 1 \quad (x_1), \quad m = 2 \quad (y_1, y_2)$

$$K - k = 2 - 1 = m - 1 = 2 - 1.$$

(2) $K = 2 \quad (x_1, x_2), \quad k = 1 \quad (x_2), \quad m = 2 \quad (y_1, y_2)$

$$K - k = 2 - 1 = m - 1 = 2 - 1.$$

Достаточное условие легко проверить самостоятельно в качестве упражнения.

Приведенная форма имеет вид

$$y_1 = a_1 x_1 + a_2 x_2 + h_1$$
$$y_2 = b_1 x_1 + b_2 x_2 + h_2 .$$

Пусть в результате статистического наблюдения собраны данные об эндогенных переменных y_1, y_2 и экзогенных переменных x_1 и x_2. На основе этой информации с помощью МНК оценим неизвестные параметры приведенной формы, т. е. получим $\hat{a}_1$, $\hat{a}_2$ и $\hat{b}_1$, $\hat{b}_2$. Это первый этап косвенного метода наименьших квадратов. На втором этапе необходимо по найденным оценкам $\hat{a}_i$, $\hat{b}_i$, $i=1,2$ определить значения структурных параметров b и g. Для этого используем соотношения, связывающие структурные параметры каждого уравнения, с параметрами приведенной формы:

$$a_1 = \frac{g_{11}}{1-b_{12}b_{21}},\ a_2 = \frac{b_{12}g_{22}}{1-b_{12}b_{21}};$$

$$b_1 = \frac{b_{21}g_{11}}{1-b_{12}b_{21}},\ b_2 = \frac{g_{22}}{1-b_{12}b_{21}}.$$

Заменим в этих выражениях неизвестные значения коэффициентов их оценками, из полученной системы четырех уравнений с четырьмя неизвестными найдем оценки структурных коэффициентов $\hat{b}_{12}$, $\hat{b}_{21}$, $\hat{g}_{11}$, $\hat{g}_{22}$.

В этом случае МНК-оценки параметров приведенной формы получаются несмещенными и состоятельными, однако оценки структурных коэффициентов, найденные из этой системы, будут только состоятельными.

Если система сверхидентифицируема, то один и тот же структурный коэффициент допускает разные выражения через параметры приведенной формы, так как в системе, связывающей эти коэффициенты число уравнений превышает число неизвестных. В этом случае наиболее простым и в то же время надежным является двухшаговый метод наименьших квадратов (2МНК).

3. Двухшаговый метод наименьших квадратов

Опишем в общих чертах суть вычислений по двухшаговому методу, которым оцениваются коэффициенты лишь одного уравнения сверхидентифицированной системы.

К процедуре оценивания параметров при применении 2МНК прибегают дважды. На первом шаге производится оценивание обычным МНК параметров приведенной формы. Это дает возможность получить оценки систематической и случайной составляющей эндогенной переменной y, т. е. предполагается, что $y_i = \hat{y}_i + h_i$, где $\hat{y}_i$ – оценки значений этой переменной, полученные по приведенной форме.

На втором шаге эндогенные переменные, находящиеся в правой части структурных уравнений, заменяются их оценками $\hat{y}_i$. К преобразованому таким путем структурному уравнению применяется обычный МНК.

Оценки структурных параметров, полученные 2МНК, получаются, вообще говоря, смещенными, но состоятельными и эффективными.

Отметим, что в большинстве эконометрических компьютерных пакетов для оценивания систем одновременных уравнений реализован именно двухша-

говый метод наименьших квадратов, при использовании которого фактически каждое уравнение оценивается независимо от других.

4. Трехшаговый метод наименьших квадратов

Метод применяется для оценки параметров системы одновременных уравнений в целом. Сначала к каждому уравнению применяется двухшаговый метод для определения оценок коэффициентов и оценок дисперсий случайных ошибок. Затем с использованием найденных оценок дисперсий возмущений строится оценка ковариационной матрицы. После этого для оценивания коэффициентов всей системы применяется обобщенный метод наименьших квадратов. Трехшаговый метод в случае, когда возмущения, входящие в различные структурные уравнения, коррелируют друг с другом, оказывается асимптотически эффективнее двухшагового метода.

При практическом использовании ЗМНК требуется иметь в виду, что:

1) каждое уравнение, являющееся определением (т. е. все тождества), необходимо исключить из системы прежде, чем приступать к вычислениям;

2) каждое неидентифицируемое уравнение также исключается;

3) в системе остаются только точно идентифицируемые и сверхидентифицируемые уравнения, причем с вычислительной точки зрения целесообразно применять трехшаговую процедуру к каждой из этих групп уравнений отдельно;

4) если матрица ковариаций для структурных возмущений блочно-диагональная, то вся процедура трехшагового оценивания может быть применена отдельно к каждой группе уравнений, соответствующих одному блоку.

Завершим эту главу описанием классической макроэкономической модели Клейна и результатов ее оценивания с помощью обычного и двухшагового метода наименьших квадратов [4; 6].

Пример 4.7. Модель Клейна 1. В 1950 году Л. Клейн предложил динамическую модель макроэкономики, получившую название модель Клейна 1. Она описывается следующей системой уравнений.

$$C_t = a_0 + a_1 P_t + a_2 P_{t-1} + a_3 \left(W_t^P - W_t^G\right) + e_{1t} \quad \text{(потребление)},$$

$$I_t = b_0 + b_1 P_t + b_2 P_{t-1} + b_3 K_{t-1} + e_{2t} \quad \text{(инвестиции)},$$

$$W_t^P = g_0 + g_1 X_t + g_2 X_{t-1} + g_3 A_t + e_{3t} \quad \text{(зарплата в частном секторе)},$$

$$X_t = C_t + I_t + G_t \quad \text{(совокупный спрос в равновесии)},$$

$$P_t = X_t - T_t - W_t^P \quad \text{(доход частного сектора)},$$

$$K_t = K_{t-1} + I_t \quad \text{(капитал)}.$$

Переменные, стоящие в левых частях уравнений, являются эндогенными. Экзогенными переменными в данной модели являются: G – государственные расходы, не включающие зарплату, T – непрямые налоги плюс чистый доход от экспорта, W^G – зарплата в государственном секторе, A_t – временной тренд (в годах, начиная с 1931 года). Кроме того, включены три лаговые переменные.

Модель содержит три поведенческих уравнения, одно уравнение равновесия и два тождества.

Приведем результаты оценивания первых трех уравнений на основе ежегодных данных для экономики США за период с 1921 по 1941 г. с помощью обычного МНК и двухшагового МНК (в скобках указаны оценки стандартных ошибок).

Обычный метод наименьших квадратов:

$$C_t = 16{,}2 + 0{,}193P_t + 0{,}090P_{t-1} + 0{,}796\left(W_t^P - W_t^G\right),$$
$$(1{,}30)\ (0{,}091)\quad (0{,}091)\quad (0{,}040)$$
$$I_t = 10{,}1 + 0{,}480P_t + 0{,}333P_{t-1} - 0{,}112K_{t-1},$$
$$(5{,}47)\ (0{,}097)\quad (0{,}101)\quad (0{,}027)$$
$$W_t^P = 1{,}48 + 0{,}439X_t + 0{,}146X_{t-1} + 0{,}130A_t.$$
$$(1{,}27)\ (0{,}032)\quad (0{,}037)\quad (0{,}032)$$

Двухшаговый метод наименьших квадратов:

$$C_t = 16{,}6 + 0{,}017P_t + 0{,}216P_{t-1} + 0{,}810\left(W_t^P - W_t^G\right),$$
$$(1{,}32)\ (0{,}118)\quad (0{,}107)\quad (0{,}040)$$
$$I_t = 20{,}3 + 0{,}150P_t + 0{,}616P_{t-1} - 0{,}158K_{t-1},$$
$$(7{,}54)\ (0{,}173)\quad (0{,}162)\quad (0{,}036)$$
$$W_t^P = 1{,}50 + 0{,}439X_t + 0{,}147X_{t-1} + 0{,}130A_t.$$
$$(1{,}15)\ (0{,}036)\quad (0{,}039)\quad (0{,}029)$$

Контрольные вопросы, задачи и упражнения

4.1. Как классифицируются переменные в системах одновременных уравнений?

4.2. Что такое идентифицируемость модели? Запишите порядковое условие идентификации.

4.3. Для модели спроса и предложения:

$$Q_t^s = a_0 + a_1P_t + a_2P_{t-1} + e_t \qquad \text{(предложение)}$$
$$Q_t^D = b_0 + b_1Y_t + b_2P_t + h_t \qquad \text{(спрос)}$$
$$Q_t^s = Q_t^D \qquad \text{(равновесие)}$$

укажите, какие переменные являются эндогенными, а какие – экзогенными.

4.4. Исследуйте на идентифицируемость модель, приведенную в примере 4.7.

4.5. Опишите процедуру оценивания параметров модели в примере 13.3.

4.6. Для модели:

$$C_t = a + bY_t + e_t$$
$$Y_t = C_t + I_t + G_t$$
$$I_t = g + dY_t + h_t$$

запишите приведенную форму; с помощью порядкового и достаточного условий идентификации проверьте, идентифицирована ли данная модель. Укажите,

каким методом вы будете определять структурные параметры каждого уравнения. В предположении, что имеются все необходимые исходные данные, кратко опишите методику расчетов.

4.7. Рассматривается статическая модель экономики страны

$$C = a_0 + a_1 Y + e$$

$$Y = C + I,$$

где C – личное потребление в постоянных ценах, Y – национальный доход в постоянных ценах, I – инвестиции в отрасли экономики страны в постоянных ценах.

Система приведенных уравнений оказалась следующей:

$$C = 44{,}6 + 3{,}2I \qquad R^2 = 0{,}975;$$

$$Y = 44{,}6 + 4{,}2I \qquad R^2 = 0{,}985.$$

Дайте интерпретацию коэффициентов приведенной формы модели. Определите параметры структурной формы модели и дайте их интерпретацию. Укажите, какая форма модели используется для прогноза.

Приложения

Приложение 1

Квантили распределения “Хи-квадрат” $c_p^2(k)$

k \ p	**0,010**	**0,025**	**0,05**	**0,10**	**0,90**	**0,95**	**0,975**	**0,990**
1	0,0157	0,0982	0,0393	0,0158	2,71	3,84	5,02	6,63
2	0,0201	0,0506	0,103	0,211	4,61	5,99	7,38	9,21
3	0,115	0,216	0,352	0,584	6,25	7,81	9,35	11,3
4	0,297	0,484	0,711	1,06	7,78	9,49	11,1	13,3
5	0,554	0,831	1,15	1,61	9,24	11,1	12,8	15,1
6	0,872	1,24	1,64	2,20	10,6	12,6	14,4	16,8
7	1,24	1,69	2,17	2,83	12,0	14,1	16,0	18,5
8	1,65	2,18	2,73	3,49	13,4	15,5	17,5	20,1
9	2,09	2,70	3,33	4,17	14,7	16,9	19,0	21,7
10	2,56	3,25	3,94	4,87	16,0	18,3	20,5	23,2
11	3,05	3,82	4,57	5,58	17,3	19,7	21,9	24,7
12	3,57	4,40	5,23	6,30	18,5	21,0	23,3	26,2
13	4,11	5,01	5,89	7,04	19,8	22,4	24,7	27,7
14	4,66	5,63	6,57	7,79	21,1	23,7	26,1	29,1
15	5,23	6,26	7,26	8,55	22,3	25,0	27,5	30,6
16	5,81	6,91	7,96	9,31	23,5	26,3	28,8	32,0
17	6,41	7,56	8,67	10,1	24,8	27,6	30,2	33,4
18	7,01	8,23	9,39	10,9	26,0	28,9	31,5	34,8
19	7,63	8,91	10,1	11,7	27,2	30,1	32,9	36,2
20	8,26	9,59	10,9	12,4	28,4	31,4	34,2	37,6
21	8,90	10,3	11,6	13,2	29,6	32,7	35,5	38,9
22	9,54	11,0	12,3	14,0	30,8	33,9	36,8	40,3
23	10,2	11,7	13,1	14,8	32,0	35,2	38,1	41,6
24	10,9	12,4	13,8	15,7	33,2	36,4	39,4	43,0
25	11,5	13,1	14,6	16,5	34,4	37,7	40,6	44,3
26	12,2	13,8	15,4	17,3	35,6	38,9	41,9	45,6
27	12,9	14,6	16,2	18,1	36,7	40,1	43,2	47,0
28	13,6	15,3	16,9	18,9	37,9	41,3	44,5	48,3
29	14,3	16,0	17,7	19,8	39,1	42,6	45,7	49,6
30	15,0	16,8	18,5	20,6	40,3	43,8	47,0	50,9
35	18,5	20,6	22,5	24,8	46,1	49,8	53,2	57,3
40	22,2	24,4	26,5	29,1	51,8	55,8	59,3	63,7
45	25,9	28,4	30,6	33,4	57,5	61,7	65,4	70,0
50	29,7	32,4	34,8	37,7	63,2	67,5	71,4	76,2
75	49,5	52,9	56,1	59,8	91,1	96,2	100,8	106,4
100	70,1	74,2	77,9	82,4	118,5	124,3	129,6	135,6

Приложение 2

Квантили распределения Стьюдента $t_p(k)$

k \ p	**0,900**	**0,950**	**0,975**
1	3,078	6,314	12,706
2	1,886	2,920	4,303
3	1,638	2,353	3,182
4	1,533	2,132	2,776
5	1,476	2,015	2,571
6	1,440	1,943	2,447
7	1,415	1,895	2,365
8	1,397	1,860	2,306
9	1,383	1,833	2,262
10	1,372	1,812	2,228
11	1,363	1,796	2,201
12	1,356	1,782	2,179
13	1,350	1,771	2,160
14	1,345	1,761	2,145
15	1,341	1,753	2,131
16	1,337	1,746	2,120
17	1,333	1,740	2,110
18	1,330	1,734	2,101
19	1,328	1,729	2,093
20	1,325	1,725	2,086
21	1,323	1,721	2,080
22	1,321	1,717	2,074
23	1,319	1,714	2,069
24	1,318	1,711	2,064
25	1,316	1,708	2,060
26	1,315	1,706	2,056
27	1,314	1,703	2,052
28	1,313	1,701	2,048
29	1,311	1,699	2,045
30	1,310	1,697	2,042
40	1,303	1,684	2,021
60	1,296	1,671	2,000
120	1,289	1,658	1,980
∞	1,282	1,645	1,960

Приложение 3

Квантили распределения Фишера $F_p(k_1, k_2)$

k_2 \ k_1	$p = 0{,}95$							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73
8	5,32	4,66	4,07	3,84	3,69	3,58	3,50	3,44
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70
15	4,54	3,68	3,29	3,05	2,90	2,79	2,71	2,64
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10
120	3,92	3,07	2,68	2,45	2,29	2,17	2,09	2,02
∞	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94

k_2 \ k_1	$p = 0{,}95$							
	9	**10**	**12**	**15**	**20**	**24**	**30**	**40**
1	240,5	241,9	243,9	245,9	248,0	249,1	250,1	251,1
2	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47
3	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59
4	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72
5	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46
6	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77
7	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34
8	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04
9	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83
10	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66
11	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53
12	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43
13	2,71	2,67	2,60	2,63	2,46	2,42	2,38	2,34
14	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27
15	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20
16	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15
17	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10
18	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06
19	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03
20	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99
21	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96
22	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94
23	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91
24	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89
25	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87
26	2,27	2,22	2,15	2,07	1,99	1,95	1,90	1,85
27	2,25	2,20	2,13	2,06	1,97	1,93	1,88	1,84
28	2,24	2,19	2,12	2,04	1,96	1,941	1,87	1,82
29	2,22	2,18	2,10	2,03	1,94	1,90	1,85	1,81
30	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79
40	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69
60	1,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59
120	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50
∞	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,39

Список рекомендуемой литературы

1. Джонстон Дж. Эконометрические методы / Дж. Джонстон. – М. : Статистика, 1980.

2. Доугерти К. Введение в эконометрику / К. Доугерти. – М. : ИНФРА-М, 1999.

3. Кремер Н. Ш. Эконометрика / Н. Ш. Кремер, Б. А. Путко. – М. :ЮНИТИ-DANA, 2003.

4. Лизер С. Эконометрические методы и задачи / С. Лизер. – М. : Статистика, 1971.

5. Магнус Я. Р Эконометрика. Начальный курс / Я. Р. Магнус, П. К. Катышев, А. А. Пересецкий. – М. : Дело, 1997.

6. Маленво Э. Статистические методы эконометрии / Э. Маленво. – М. : Статистика, 1975.

7. Эконометрика / под ред. В. С. Мхитаряна. – М. : Проспект, 2008.

8. Эконометрика / под ред. И. И. Елисеевой. – М. : Финансы и статистика, 2001.

9. Greene W. N. Econometric Analysis / W. N. Greene. – Second edition. – New York : Macmillan Publishing Company, 1993.

10. Goldberger A. A course in Econometrics / A. Goldberger. – Cambridge : MA: Harvard University Press, 1990.

11. Maddala G. S. Introduction to econometrics / G. S. Maddala. – New York : Macmillan Publishing Company, 1988.

Учебное издание

Ежова Людмила Николаевна
Абдуллин Рафаэль Зинатович
Абдуллин Владимир Рафаэлевич

ЭКОНОМЕТРИЧЕСКИЕ МЕТОДЫ И МОДЕЛИ

Учебное пособие для магистрантов,
обучающихся по направлению «Экономика»

Издается в авторской редакции

ИД № 06318 от 26.11.01.
Подписано в печать 20.06.12. Формат 60x90 1/16. Бумага офсетная. Печать трафаретная. Усл. печ. л. 5,8. Тираж 200 экз. Заказ

Издательство Байкальского государственного университета
экономики и права.
664015, г. Иркутск, ул. Ленина, 11.
Отпечатано в ИПО БГУЭП.