Введение

Эконометрия – это самостоятельная научная дисциплина, объединяющая совокупность теоретических результатов, приемов, методов и моделей, предназначенных для того, чтобы на базе

i **экономической теории**,

ii **экономической статистики**,

iii **математико-статистического инструментария**

придавать конкретное количественное выражение общим (качественным) закономерностям, обусловленным экономической теорией.

Источники формирования составляющих частей эконометрии представлены на рис. 1.

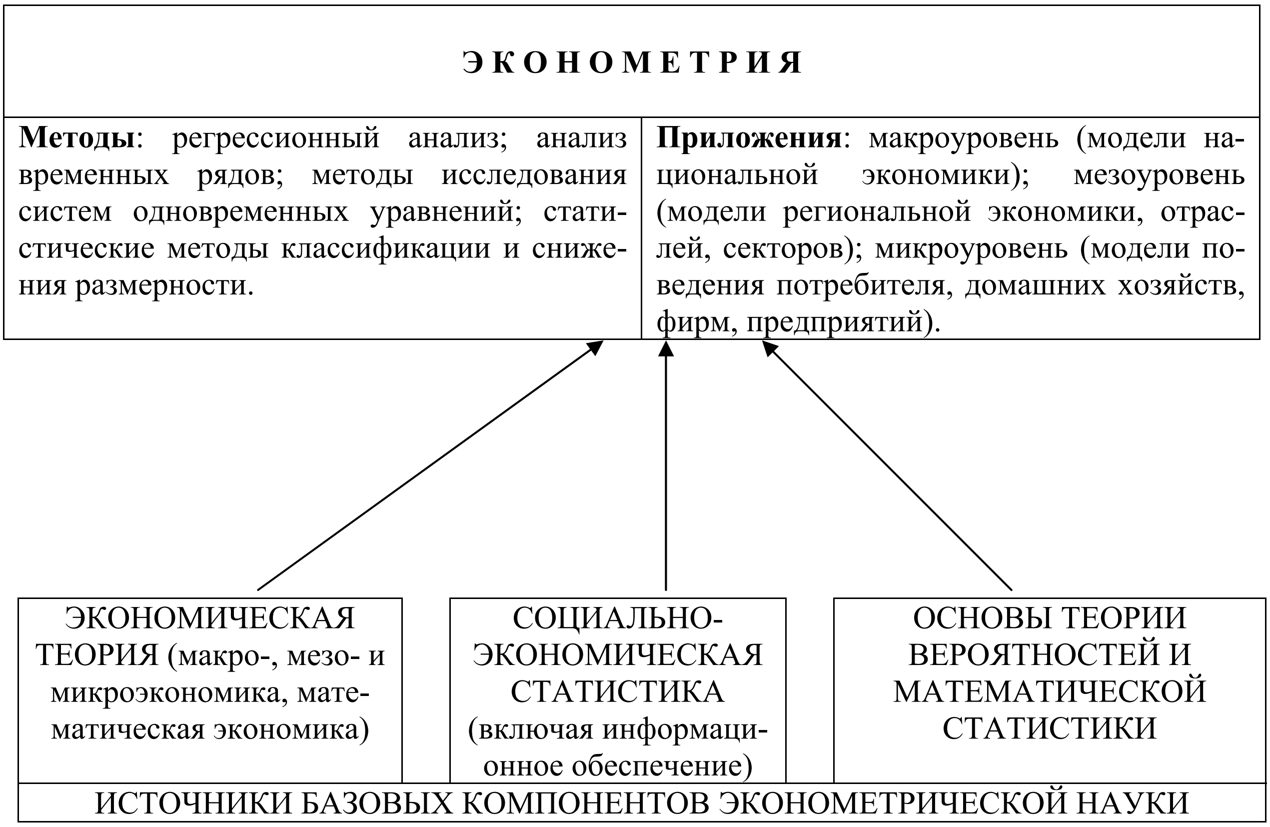

Рис. 1. Эконометрия и ее место в ряду других экономических и математико-статистических дисциплин

Суть эконометрии – синтез экономики, экономической статистики и математики.

Эконометрия — сравнительно молодая область науки, известная под таким названием с 1930 года. Тогда было основано Эконометрическое общество, которое определило себя так: «Международное общество для развития экономической теории и его связи со статистикой и математикой». С 1933 г. выходит журнал «Эконометрия», который издается этим обществом.

Термин «эконометрия» предложил львовский ученый П.Чомпа, опубликовав во Львове в 1910 году книгу «Очерки эконометрии и естественной теории бухгалтерии, которая основана на политической экономии».

За рубежом первые работы по эконометрии, принадлежавшие Муру, вышли из печати в 1914—1917 гг. В 1928 году было опубликовано исследование Ч. Кобба и П. Дугласа о производственной функции. Эта функция вошла в эконометрию как классический пример и до сих пор является важным инструментом экономического анализа.

Термин ***«эконометрия»*** означает ***измерения в экономике.*** Измерение в самом деле есть важная часть эконометрии. Но не все измерения в экономике следует относить к эконометрии. Если раньше некоторые авторы почти все прикладные математические исследования в экономике считали частью эконометрии, то теперь распространена точка зрения, согласно которой ее содержание значительно сужается. Согласно сказанному дадим такое определение.

Эконометрия изучает методы оценивания параметров моделей, которые характеризуют количественные и качественные взаимосвязи между экономическими показателями, а также рассматривает основные направления применения этих моделей в экономических исследованиях.

Тема 1. Предмет, метод и задачи курса «Эконометрия»

1.1. Предмет и метод курса

Определение предмета эконометрии в разных изданиях трактуется по-разному. Сегодня можно условно различать ***пять*** подходов к определению предмета эконометрии, которые характерные для зарубежной эконометрической литературы.

1. *Л. Клейн* [**Ошибка! Источник ссылки не найден.**] определяет эконометрию как науку, которая изучает измерение связей в соответствующем экономическом анализе.

2.Г. Тинтнер [**Ошибка! Источник ссылки не найден.**] отождествляет эконометрию с экономической статистикой.

3. *Г. Хансен* [**Ошибка! Источник ссылки не найден.**] под эконометрией понимает применение математических и статистических методов в экономике.

4. По другим определениям эконометрия есть синтез экономической теории и математики.

5. Некоторые авторы под эконометрией понимают экономическую теорию, математику и статистику, взятые одновременно.

Любое из этих определений предмета эконометрии имеет свое рациональное зерно. Но все они частично содержат в себе только что приведенное определение предмета эконометрии и отвечают поэтому материалу, который изучается в курсе.

Эконометрия не рассматривается как область математики, но математика играет в ней очень важную роль. Поэтому изложение эконометрии дается практически так же, как это принято в математических курсах. Предусматривается введение понятийного аппарата, постановка задачи, а также анализ решения, базирующихся на теоремах и основных определениях. В эконометрии не всегда все утверждения строго доказываются, но алгоритмы задач основываются на методах математической статистики, широко используются матричная алгебра и другие классические разделы высшей математики.

1.2. Место курса среди дисциплин фундаментальной подготовки бакалавров по экономическим специальностям

Эконометрия — одна из основных дисциплин в подготовке бакалавров из экономических специальностей. Она строится на основе математических и экономических знаний. С помощью эконометрических методов могут быть проверены различные экономические гипотезы. В крайнем случае, можно показать невозможность их применения в данных конкретных условиях. Хотя средства эконометрии не дают возможности доказывать теоретические утверждения, но с помощью ее методов можно показать, что-то или другое утверждение не противоречит данным наблюдений. Эконометрические модели могут использоваться для прогнозирования или оценивания влияния принятых решений или правительственных постановлений относительно изменения цен, налогов и т.п., на состояние дел любой фирмы.

Для усвоения курса нужна хорошая математическая подготовка, в частности по ***матричной алгебре, дифференциальному исчислению.*** В одинаковой мере следует владеть ***методами математической статистики.*** Важно также знать ***основные экономические категории и понятия из курсов макро- и микроэкономики.*** Современные эконометрические задачи решаются с использованием ПЭВМ и специализированных пакетов прикладных программ. К ним относятся: SHAZAM, TCP и другие. Достаточно эффективно могут быть использованы статистические пакеты SPSS, STATGRAPHICS, STATISTIKA. Многие задачи могут быть решены в электронной таблице MICRISOFT EXCEL.

1.3. Структура курса

Эконометрия делится на ***две части:***

1) эконометрические методы;
2) эконометрические модели экономических процессов и явлений.

В настоящем пособии изложен материал, который принадлежит к первой части.

Эконометрические методы можно условно разбить на ***четыре группы.*** В ***первую группу*** входят методы оценивания параметров классической эконометрической модели по методу наименьших квадратов. Ко ***второй группе*** принадлежат методы оценивания параметров обобщенной модели, если нарушаются некоторые предпосылки использования метода наименьших квадратов. К ***третьей группе*** входят методы оценивания параметров динамических эконометрических моделей. ***Четвертая группа*** охватывает методы оценивания параметров эконометрических моделей, построенных на основе систем одновременных структурных уравнений.

1.4. Краткая историческая справка

В начале XX столетия в некоторых странах были предприняты попытки составить так называемые «барометры развития». Известнейший из них – «гарвардский барометр», с помощью которого в 20-тые годы старались спрогнозировать закономерности поведения товарного и денежного рынка.

Гарвардская школа считалась в то время центром экономических исследований. Здесь впервые начали системно изучать ряды экономических показателей с учетом взаимосвязи между ними и на основе этих показателей исследовать тенденции и циклы экономических процессов. Кризис 1929—1933 гг. привел к критическому пересмотру методов анализа, которые применялось в экономике. В исследованиях начали учитывать случайные аспекты экономических явлений. Это послужило началом формирования эконометрии как области экономической науки.

Основатели эконометрии — Р. Фриш, Э. Шумпетер, Я. Тинберген. Все они находились под влиянием неоклассической школы и кейнсианства и старались соединить экономическую теорию с математическими и статистическими методами. Сначала ограничивались изучением некоторых моделей спроса и предложения. Только после второй мировой войны были построены комплексные эконометрические модели на макроуровне. В них основное внимание уделялось спросу, финансовому состоянию, налогам, прибыли, ценам и т.п.

На протяжении последних пятидесяти лет развитие эконометрии происходило в таких ***двух направлениях:*** 1) разработка новых методов оценивания параметров моделей с учетом особенностей экономической информации; 2) расширение экономических исследований на основе эконометрических методов.

Классическими эконометрическими проблемами считаются:

1) изучение и учет мультиколлинеарности**;**

2) спецификация ошибок;

3) ковариационный анализ параметров модели;

4) построение моделей с фиктивными переменными;

5) определение лаговых переменных и построение и анализ моделей с распределенным лагом.

1.5. Задачи эконометрического исследования

Роль эконометрического исследования определяется теми задачами, которые может решать эконометрия.

Важнейшей задачей эконометрии является оценивание параметров и проверка значимости эконометрической модели. На первом этапе процесса эконометрического исследования осуществляют спецификацию модели в математической форме. Второй этап — сбор и подготовка экономической информации. На третьем этапе оцениваются параметры модели. Четвертый этап — проверка модели на достоверность. Очень важными на этом этапе есть оценки дисперсии остатков модели. Эти оценки играют решающую роль при выяснении качества эконометрических моделей. Они необходимы для определения надежности вычисленных параметров и для применения разработанных моделей в прогнозировании.

1.6. Краткие выводы

1. Эконометрия является сравнительно молодой областью экономической науки и известна под таким названием с 1930 года. Она изучает методы оценивания параметров моделей, которые характеризуют количественные взаимосвязи между экономическими величинами, а также рассматривает основные направления применение эконометрических моделей в экономических исследованиях.

2. Эконометрия – одна из фундаментальных дисциплин в системе подготовки бакалавров по экономическим специальностям. Для усвоения курса необходима хорошая математическая подготовка и, в особенности достаточно полное знание математической статистики, матричной алгебры и дифференциального исчисления.

3. Эконометрия делится на **две части:**

а) эконометрические методы;

б) эконометрические модели экономических процессов и явлений.

4. Эконометрические методы можно разбить на четыре группы. К первой принадлежат методы оценивания параметров классической эконометрической модели на основе метода наименьших квадратов. Вторая группа объединяет методы оценивания параметров обобщенной модели, если не выполняются некоторые предпосылки использования метода наименьших квадратов. Третью группу составляют методы оценивания параметров динамических эконометрических моделей. Четвертая группа включает методы оценивания параметров эконометрических моделей, построенных на основе системы одновременных структурных уравнений.

5. Развитие эконометрии происходит в двух направлениях:

1) разработка новых методов оценивания параметров моделей с учетом особенностей входной экономической информации; 2) расширение экономических исследований на основе эконометрических методов.

6. Основная задача эконометрии состоит в оценивании параметров и проверке значимости эконометрической модели. Решение этой задачи связано с выполнением таких шагов эконометрического исследования: спецификация модели в математической форме; сбор и подготовка экономической информации; оценивание параметров модели и проверка ее на достоверность.

7. Структура эконометрических исследований содержит в себе эконометрические методы и применение их при построении и анализе эконометрических моделей с учетом особенностей исходной информации.

1.7. Вопросы к теме 1

1. Дайте определения предмета курса эконометрии.
2. Охарактеризуйте альтернативные подходы к определению предмета эконометрии.
3. Какую роль играет этот курс в подготовке бакалавров по экономическим специальностям?
4. Охарактеризуйте структуру курса.
5. Приведите основные этапы развития эконометрии как экономической науки.
6. Назвать задачи эконометрического исследования.
7. Дать характеристику структуры эконометрических исследований.

1.8. Основные термины и понятия

Эконометрия
Верификация
Эконометрические методы
Динамическая эконометрическая модель
Эконометрическая модель
Система одновременных структурных уравнений
Оценивание параметров
Значимость модели

Тема 2. **Основы эконометрического моделирования**

2.1. Особенности эконометрических моделей

Современные методы управления экономическими системами и процессами базируются на широком использовании математических методов и ЭВМ. Применять математику для решения определенных экономических задач начали очень давно, сотни лет тому назад. Но на протяжении последних 50-60 лет экономическая наука достигла определенных рубежей в своем развитии. В ней появились задачи, которые не удается решить с помощью традиционных экономических методов. Математика заняла в экономической науке одно из основных мест. Сформировалось направление теоретико-практических исследований — экономико-математическое моделирование. Математическое моделирование в экономике – это выражение процесса математизации научного экономического знания. Математика, проникая в сущность экономической науки, приносит с собою точность и универсальность решений, строгость и совершенство научных концепций. С развитием математики, электронной вычислительной техники, экономические науки все шире используют математические модели.

Математическая модель объекта (процесса, явления) содержит в себе три группы элементов: 1) характеристику объекта, который нужно определить (неизвестные величины); 2) характеристики внешних (относительно моделированного объекта) условий, которые изменяются; 3) совокупность внутренних параметров объекта.

Множества условий и параметров могут рассматриваться как экзогенные величины (определяемые вне рамок модели), а величины, которые неизвестны — как эндогенные (то есть такие, которые определяются по помощи модели).

Математическую модель можно толковать как особый преобразователь внешних условий объекта $\boldsymbol{X}$ (входа) на характеристики объекта $\boldsymbol{Y}$ (выхода), которые должны быть найдены.

В зависимости от способа выражения соотношений между внешними условиями, внутренними параметрами и характеристиками, которые должны быть найдены, математические модели делятся на две группы: структурные и функциональные.

Объект эконометрического исследования - это экономическая модель, представленная в виде ряда соотношений типа тождеств, уравнений экономического поведения, институциональных, технологических зависимостей.

Процесс построения и исследования экономических моделей условно можно представить следующими этапами:

1) предмодельный анализ экономического объекта;

2) формулирование цели построения модели;

3) конструирование совокупности переменных, необходимых для описания объекта;

4) определение вида структурных уравнений (спецификация модели);

5) формулирование вероятностных гипотез относительно случайных возмущений структурных уравнений в виде предположений о ковариационной матрице случайных возмущений;

6) построение временных рядов на основе наблюдений переменных, участвующих в модели за прошлые периоды;

7) анализ временных рядов с целью выявления корреляционной зависимости между рядами и автокорреляции;

8) корректировка зависимостей (на основе предыдущего анализа), постулированных на шаге 4;

9) выбор метода оценивания структурных коэффициентов модели с учетом гипотез о вероятностных свойствах случайных возмущений и характера временных рядов;

10) оценка структурных коэффициентов выбранным методом;

11) проверка качества полученных оценок, а также гипотез, легших в основу спецификации;

12) проверка всей модели в целом путем экстраполяции назад (на базисный период) и вперед (на период, следующий за базисным и специально не включенный в базу, с тем чтобы иметь возможность проверки). Для проверки можно использовать так называемые коэффициенты соответствия;

13) построение неформальных сценариев развития событий на прогнозируемый период;

14) формализация сценариев в терминах инструментальных и заданных переменных, а также управляемых параметров;

15) получение многовариантного прогноза путем выполнения имитационных расчетов по заданным значениям внесистемных переменных и управляемых параметров;

16) неформальная интерпретация полученных результатов;

17) анализ рассогласования и корректировка модели.

2.2. *Пример модели предложения и спроса на конкурентном рынке*

С целью выполнения предмодельного анализа процессов предложения и спроса на конкурентном рынке рассмотрим рынок из одного продукта. Объект исследования опишем следующими переменными. Пусть q_1 – запрашиваемое и q_2 – предложенное количество некоторого продукта в некоторый день на рынке, P – цена, по которой заключаются сделки.

Необходимые для построения эконометрических моделей связи между переменными определим двумя функциями: $q_1 = f(P)$ – функция спроса; $q_2 = g(P)$ – функция предложения.

Рыночное поведение выразим условием равновесия

$$q_1 = q_2. \tag{2.1}$$

По экономическому смыслу спрос, предложение и цена сделки не могут быть отрицательными:

$$q_1,\ q_2,\ P > 0. \tag{2.2}$$

Проводя линеаризацию функций спроса и предложения, приходим к паутинообразной модели конкурентного рынка следующего вида:

$$q_1 = a + b \times P, \tag{2.3}$$
$$q_2 = c + d \times P, \tag{2.4}$$

где параметры *a, b, c, d* – необходимо оценить ($a, c, d > 0, b < 0$).

Таким образом, получили эконометрическую модель равновесного рынка, заданную соотношениями (2.1)-(2.4), которая может быть исследована в соответствии с этапами, указанными в предыдущем пункте.

2.3. Вопросы к теме 2

1. Дать понятие экономической модели.
2. Что такое эконометрическая модель?
3. Назвать этапы построения и исследования экономической модели.
4. Какими соотношениями задается эконометрическая модель равновесия рынка одного товара?

Тема 3. Простейшая линейная эконометрическая модель

3.1. Предположения и постановка задачи

Придерживаемся позиции, состоящей в том, что эконометрические методы разработаны в связи с оцениванием параметров экономических моделей. Необходимо сразу заметить, что достаточно серьезная экономическая модель содержит несколько уравнений, а в каждом уравнении имеется несколько переменных и параметров.

Для относительно полного понимания метода наименьших квадратов рассмотрим элементарный случай, когда модель представлена одним уравнением с двумя переменными: X – фактор, который описывает факторный признак; Y – фактор, описывающий результирующий признак. Поставим задачу нахождения "адекватной" связи между переменными X и Y.

Первое (исходное) предположение состоит в том, что между факторами Х и У постулируется связь вида $X \rightarrow Y$ (фактор X влияет на фактор Y):

$$Y = f(X). \tag{3.1}$$

Соотношение (3.1) означает, что выполнена *идентификация* переменной X, как воздействующей на переменную Y.

Второе предположение состоит в осуществлении *спецификации формы связи* между Y и X. На этом неформальном этапе привлекаются содержательные соображения, которые должны прояснить конкретный вид функциональной зависимости или подсказать условия, которым должны удовлетворять параметры модели.

Простейшими примерами функциональных зависимостей между переменными X и Y являются:

линейная,

$$Y = \alpha + \beta \times X, \tag{3.2}$$

где α и β – неизвестные параметры, геометрический смысл которых состоит в следующем. Параметр α определяет точку пересечения прямой $Y = \alpha$ с осью ординат, а второй – β – тангенс угла наклона прямой (3.2) к оси абсцисс (рис. 3.1);

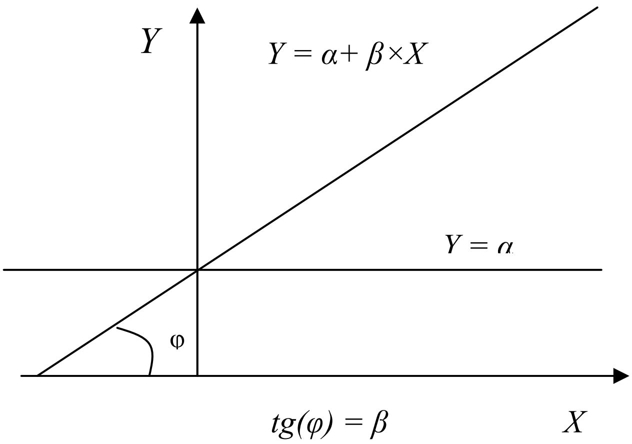

Рис. 3.1. Графики уравнений регрессии

экспоненциальная

$$Y = \alpha \times EXP(\beta \times X); \tag{3.3}$$

степенная

$$Y = \alpha \times X^{\beta}; \tag{3.4}$$

показательная

$$Y = \alpha \times \beta^{X} \tag{3.5}$$

гиперболическая

$$Y = \alpha + \beta / X \tag{3.6}$$

полиномиальная

$$Y = \sum_{i=0}^{n} \beta_i \times X^i \tag{3.7}$$

где $n > 1$, и др. Соотношения (3.3)-(3.6) соответствующими преобразованиями могут быть сведены к линейной зависимости типа (3.2). Для связи типа (3.7) это, вообще говоря, невозможно.

Большинство традиционных экономических теорий содержит отражение связей между экономическими категориями в виде таблиц, графиков или функциональных соотношений.

Выполнение реальных экономических измерений показателей X и Y всегда приводит к ситуации, что их значения не могут быть расположены на некоторой линии типа (3.2)-(3.7). Поэтому первого и второго предположений оказывается недостаточно для реализации *цели исследования*, состоящей в нахождении адекватной формы зависимости между факторами Y и Y.

Один из возможных подходов состоит во введении предположения о наличии *фактора стохастичности* и введении его в модель. Обсудим возможные предпосылки его появления. Предположим, что мы выполнили измерения значений признаков X и Y у N экономических объектов и в результате получили N пар данных (X_i, Y_i), $i = 1, 2, \ldots, N$.

Наши экономические объекты обладают целым рядом других признаков, например, размером, составом и т.д. Считаем, что объекты уже разделены на группы по выбранным признакам тем или иным способом. Анализируем некоторую группу объектов (пусть она содержит первые n единиц). В условиях гипотезы линейности связи между переменными X и Y для данной группы экономических объектов можем предположить, что группа в целом характеризуется связью вида

$$Y' = \alpha + \beta \times X',$$

где каждый представитель группы имеет значение признака X, равное X'.
Естественно, что значения признака Y для них разные. Поэтому

$$Y' = \alpha + \beta \times X' + U_i,\ i = 1, \ldots, n,$$

где U_i – отклонения конкретного значения признака Y_i от центрального, группового.

Может быть предложено минимум *три* пути рационального объяснения включения в уравнение (3.2) стохастической добавки U:

$$Y = \alpha + \beta \times X + U. \qquad (3.8)$$

Первое соображение состоит в том, что в действительности на результирующий фактор Y может оказывать существенное влияние кроме X целый ряд других факторов Z, U, V, W и т.д. Не все они могут быть отслежены или учтены количественно. Считаем, что вместо зависимости вида

$$Y = f(Z, U, V, W, \ldots),$$

где учтены менее важные по сравнению с X факторы $Z, U, V, W \ldots$, имеет место уравнение

$$Y = f(X, U).$$

В нем одна переменная U предназначена для отражения суммарного эффекта факторов Z, U, V, W во влиянии на Y. При этом эффект может быть как положительным, так и отрицательным. Если окажется, что малые значения встречаются чаще чем большие, то об U можно говорить как о случайной переменной с нулевым математическим ожиданием и конечной дисперсией σ_u^2. Ввиду того, что число явно неучтенных факторов значительно, а их можно считать независимыми, привлечение центральной предельной теоремы позволяет трактовать случайную величину U нормально распределенной.

Вторая трактовка наличия случайного фактора в уравнении (3.8) состоит в том, что экономические объекты функционируют с участием людей, влияние которых на результаты деятельности детерминировано предсказать нельзя. Точно так же априори нет никаких оснований утверждать, что вид функциональной зависимости выбран правильно.

Третья трактовка присутствия фактора случайности U в уравнении (3.8) связана с ошибками, возникающими в результате выполнения наблюдений и измерений.

Более того, ошибки, связанные с выбором вида уравнения, могут накладываться, причем не обязательно аддитивно, в виде добавочного слагаемого.

Таким образом, детальный анализ сути спецификации формы связи между переменными X и Y (предположение 2) привел к необходимости выдвижения ряда гипотез о свойствах распределения вероятностей для случайного возмущения U. Гипотезы относятся к математическому ожиданию, дисперсии и ковариациям и состоят в простейшем случае в следующем:

а)

$$Y_i = \alpha + \beta \times X_i + U_i,\ i = 1,\ \ldots,\ n, \tag{3.9}$$

б)

$$M(U_i) = 0; \tag{3.10}$$

$$M(U_i U_j) = \begin{cases} \sigma_n^2, \text{если } i = j,\ i,j = 1,\ \ldots,\ n; \\ 0, \text{если } i \neq j,\ i,j = 1, 2, \ldots, n. \end{cases} \tag{3.11}$$

В соотношениях (3.9) и (3.11) параметры α, β, σ_u^2 считаются неизвестными, а σ_u^2 – постоянной. На основе наблюдений над X и Y необходимо не только статистически оценить указанные параметры (этим занимается статистика), но и проверить по отношению к ним некоторые гипотезы, а именно:

- действительно ли, что $\alpha = 0$?

- с какой достоверностью можно утверждать, что $\beta = \beta_0$, где β_0 – заданная величина?

- является ли адекватной линейная зависимость между переменными X и Y?

- на сколько оправданно утверждение о постоянстве дисперсии σ_u^2?

В теории статистики существуют стандартные методы, позволяющие ответить на поставленные вопросы. Эконометрическая теория развивает новые методы получения статистических выводов относительно экономических моделей и их составляющих в тех случаях, когда стандартные методы неприменимы.

3.2. Метод наименьших квадратов

Получим оценки параметров α и β уравнения линейной регрессии (3.2) методом наименьших квадратов и исследуем свойства этих оценок.

В зависимости от исходных предположений относительно признаков X и Y, а также их характеристик задачи получения статистических выводов могут трактоваться по-разному.

Пусть вначале выполняются первое и второе предположения, а также условия (3.9)-(3.11). Уравнение (3.2) запишем для искомых оценок в виде

$$\widehat{Y} = \widehat{\alpha} + \widehat{\beta} \times X\,, \qquad (3.12)$$

где $\widehat{\alpha}, \widehat{\beta}$ – оценки неизвестных параметров α и β, $\widehat{Y}$ – ордината прямой, соответствующая данному значению X. Совокупность выборочных значений признаков X и Y зададим векторами

$$\vec{X} = [\,X_1, X_2, ..., X_N\,]^T, \vec{Y} = [\,\vec{Y}_1, \vec{Y}_2, ..., \vec{Y}_N\,]^T,$$

где T – знак транспонирования.

Введем обозначения, $\vec{1} = [1,1,...,1]^T, \overline{X} = \frac{1}{n} \times (\vec{1}^T, \vec{X}) = \frac{1}{n}\sum_{i=1}^{n} X_i, \overline{Y} = \frac{1}{n} \times (\vec{1}^T, \vec{Y}) = \frac{1}{n}\sum_{i=1}^{n} Y_i.$

Естественно положить $\hat{\vec{Y}} = \hat{\alpha} \times \vec{1} + \hat{\beta} \times \vec{X}$. Пусть также $\vec{e} = \vec{Y} - \hat{\vec{Y}}$.

Введем, наконец, величину

$$(\vec{e}^{\,T}, \vec{e}\,) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \qquad (3.13)$$

характеризующую сумму квадратов отклонений исходных измерений признака Y от точек прямой (3.12), рассчитанных для значений X_i.

Принцип наименьших квадратов состоит в выборе таких $\hat{\alpha}$ и $\hat{\beta}$, для которых величина (3.13) будет минимальной.

Для нахождения $\hat{\alpha}$ и $\hat{\beta}$ используют необходимые условия экстремума в виде

$$\frac{\partial}{\partial \hat{\vec{\beta}}} (\vec{e}^{\,T}, \vec{e}\,) = \vec{0}, \qquad (3.14)$$

где $\hat{\vec{\beta}} = [\,\hat{\alpha}, \hat{\beta}\,]^T, \vec{0} = [\,0,0\,]^T$.

Соотношение (3.14) представляет собой систему уравнений, которая в окончательном виде записывается как

$$\begin{cases} \hat{\alpha} + \hat{\beta} \times \overline{X} = \overline{Y}, \\ \hat{\alpha} \times \overline{X} + \hat{\beta} \times \overline{X^2} = \overline{X \times Y} \end{cases} \qquad (3.15)$$

и называется *системой нормальных уравнений*. Здесь $\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \overline{X \times Y} = \frac{1}{n}\sum_{i=1}^{n} X_i \times Y_i$.

Из первого уравнения следует, что оценки $\hat{\alpha}$ и $\hat{\beta}$ метода наименьших квадратов удовлетворяют уравнению прямой, проходящей через точку с координатами ($\overline{X},\overline{Y}$).

Решение системы нормальных уравнений (3.15) имеет вид:

$$\hat{\beta} = \frac{\overline{X \times Y} - \overline{X} \times \overline{Y}}{\overline{X^2} - (\overline{X})^2}, \qquad (3.16)$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta} \times \overline{X}.$$

Оценки $\hat{\alpha}$ и $\hat{\beta}$ обладают следующими свойствами.

Свойство 1. Оценки $\hat{\alpha}$ и $\hat{\beta}$ являются линейными комбинациями наблюдений Y_i.

Определение. Оценка $\hat{\gamma}$ параметра γ называется *несмещенной*, если выполнено условие $M(\hat{\gamma}) = \gamma$.

Свойство 2. Величины $\hat{\alpha}$ и $\hat{\beta}$ являются несмещенными оценками параметров α и β.

Можно показать, что дисперсии оценок $\hat{\alpha}$ и $\hat{\beta}$ имеют вид:

$$D(\hat{\alpha}) = \frac{\sum X_i^2}{n \times \sum x_i^2} \times \sigma_u^2, D(\overline{\beta}) = \frac{\sigma_u^2}{\sum x_i^2}, \qquad (3.17)$$

а их ковариация –

$$cov(\hat{\alpha}, \hat{\beta}) = -\frac{\overline{X}}{\sum x_i^2} \times \sigma_u^2, \qquad (3.18)$$

где $x_i = X_i - \overline{X}$.

Определение. Оценка $\hat{\gamma}$ параметра γ называется *наилучшей*, если в классе всех линейных несмещенных операторов оценивания вида $\sum C_i \times Y_i$, где $C_i = \omega_i + d_i$, величина $\omega_i = \frac{x_i}{\sum x_i^2}$, а d_i – произвольная константа, она обладает наименьшей дисперсией.

Свойство 3. Оценки $\hat{\alpha}$ и $\hat{\beta}$ являются наилучшими.

Можно показать, что несмещенная оценка $\hat{\sigma}_u^2$ неизвестной дисперсии σ_u^2 может быть взята в виде

$$\hat{\sigma}_u^2 = \frac{1}{n-2} \sum e_i^2. \qquad (3.19)$$

3.3. Пример

Требуется построить эконометрическую модель зависимости затрат на единицу продукции от уровня ее фондоемкости. Исходные данные в условных единицах представлены во втором и третьем столбцах таблицы 3.1.

Таблица 3.1

Исходные данные и промежуточные расчеты для примера

№ п/п	Y_i	X_i	X_i^2	X_iY_i	$X_i-\bar{X}$	$Y_i-\bar{Y}$	$(X_i-\bar{X})^2$	$(X_i-\bar{X})(Y_i-\bar{Y})$
1	50	90	8100	4500	-10,8	-4,2	116,64	45,36
2	40	75	5625	3000	-25,8	-14,2	665,64	366,36
3	65	120	14400	7800	19,2	10,8	368,64	207,36
4	55	100	10000	5500	-0,8	0,8	0,64	-0,64
5	45	80	6400	3600	-20,8	-9,2	432,64	191,36
6	42	78	6084	3276	-22,8	-12,2	519,84	278,16
7	56	110	12100	6160	9,2	1,8	84,64	16,56
8	60	115	13225	6900	14,2	5,8	201,64	82,36
9	64	115	13225	7360	14,2	9,8	201,64	139,16
10	65	125	15625	8125	24,2	10,8	585,64	261,36
Сумма	542	1008	104784	56221			3177,6	1587,40
Среднее	54,2	100,8	10478,4	5622,1			317,76	158,74

Решение

Пусть зависимость между затратами на единицу продукции и уровнем фондоемкости характеризуется уравнением прямой $Y = \alpha + \beta \times X + u$, где Y – затраты на единицу продукции; X – уровень фондоемкости; u – остатки.

Расчетные значения затрат на единицу продукции (величины $\hat{Y}$) можно найти, воспользовавшись моделью

$$\hat{Y} = \hat{\alpha} + \hat{\beta} X .$$

Продолжение таблицы 3.1

$\hat{Y}_i$	U_i	U_i^2	$(Y_i-\bar{Y})^2$
48,80	1,20	1,43	17,64
41,31	-1,31	1,72	201,64
63,79	1,21	1,46	116,64
53,80	1,20	1,44	0,64
43,81	1,19	1,42	84,64
42,81	-0,81	0,66	148,84
58,80	-2,80	7,82	3,24
61,29	-1,29	1,67	33,64
61,29	2,71	7,32	96,04
66,29	-1,29	1,66	116,64
Сумма		26,60	819,60
Среднее		2,66	81,96

Оценки параметров α и β находим, решая систему нормальных уравнений (3.15):

$$\begin{cases} \hat{\alpha} + 100{,}8\hat{\beta} = 54{,}2 \\ 100{,}8\hat{\alpha} + 10478{,}4\hat{\beta} = 5622{,}1 \end{cases}.$$

Ее решение имеет вид $\hat{\alpha} = 0{,}5; \hat{\beta} = 3{,}8$. Эконометрическая модель выглядит следующим образом

$$Y = 3{,}8 + 0{,}5X + u .$$

Полученный результат допускает следующее толкование. Так как свободный член модели (3,8) не равен нулю, то уровень затрат на единицу продукции не строго пропорционален уровню фондоемкости. Величина 0,5 показывает, что при увеличении фондоемкости продукции на одну условную единицу затраты на единицу продукции возрастают на 0,5.

3.4. Вопросы к теме 3

1. Какие предположения относительно связи вида $X \to Y$ необходимы?
2. Раскрыть сущность метода наименьших квадратов.
3. Назвать примеры функциональных зависимостей между переменными X и Y.
4. Указать цель исследования зависимости между факторами Y и Y.
5. Назвать причины введения в уравнение регрессии фактора случайности.
6. Что такое система нормальных уравнений?
7. Какими свойствами обладают оценки $\hat{\alpha}$ и $\hat{\beta}$?
8. Как интерпретируются параметры уравнения линейной регрессии?

3.5. Задачи для самостоятельного решения

Задана динамика курса валюты за семь месяцев (данные условные). Требуется построить линейную регрессионную модель изменения курса валюты в зависимости от времени и дать интерпретацию полученному результату. Исходные данные представлены в табл. 3.2.

Таблица 3.2

Варианты заданий

Вариант	1							Вариант	14						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,1	3,3	3,4	3,6	3,8	3,9	4	Курс	3,1	3,3	3,4	3,6	3,8	3,9	4,1
Вариант	2							Вариант	15						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,11	3,31	3,41	3,61	3,81	3,91	4,01	Курс	3,09	3,29	3,39	3,59	3,79	3,89	4,09
Вариант	3							Вариант	16						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,12	3,32	3,42	3,62	3,82	3,92	4,02	Курс	3,11	3,31	3,41	3,62	3,82	3,92	4,12
Вариант	4							Вариант	17						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,13	3,33	3,43	3,63	3,83	3,93	4,03	Курс	3,13	3,33	3,43	3,65	3,85	3,95	4,15
Вариант	5							Вариант	18						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,14	3,34	3,44	3,64	3,84	3,94	4,04	Курс	3,15	3,35	3,45	3,68	3,88	3,98	4,18
Вариант	6							Вариант	19						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,15	3,35	3,45	3,65	3,85	3,95	4,05	Курс	3,17	3,37	3,47	3,71	3,91	4,01	4,21
Вариант	7							Вариант	20						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,16	3,36	3,46	3,66	3,86	3,96	4,06	Курс	3,19	3,39	3,49	3,74	3,94	4,04	4,24
Вариант	8							Вариант	21						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,17	3,37	3,47	3,67	3,87	3,97	4,07	Курс	3,21	3,41	3,51	3,77	3,97	4,07	4,27
Вариант	9							Вариант	22						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,18	3,38	3,48	3,68	3,88	3,98	4,08	Курс	3,23	3,43	3,53	3,8	4	4,1	4,3
Вариант	10							Вариант	23						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,19	3,39	3,49	3,69	3,89	3,99	4,09	Курс	3,25	3,45	3,55	3,83	4,03	4,13	4,33
Вариант	11							Вариант	24						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,2	3,4	3,5	3,7	3,9	4	4,1	Курс	3,27	3,47	3,57	3,86	4,06	4,16	4,36
Вариант	12							Вариант	25						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,21	3,41	3,51	3,71	3,91	4,01	4,11	Курс	3,29	3,49	3,59	3,89	4,09	4,19	4,39
Вариант	13							Вариант	26						
День	1	2	3	4	5	6	7	День	1	2	3	4	5	6	7
Курс	3,22	3,42	3,52	3,72	3,92	4,02	4,12	Курс	3,31	3,51	3,61	3,92	4,12	4,22	4,42

3.6. Основные термины и понятия

Идентификация

Спецификация формы связи

Уравнение линейной регрессии

Система нормальных уравнений

Фактор стохастичности Несмещенность

Тема 4. Исследование простейшей эконометрической модели

4.1. Проверка гипотез о значимости параметров эконометрической модели

4.1.1. Предварительные замечания

Чтобы корректно выдвинуть и затем проверить гипотезы типа $\alpha = \alpha_0, \beta = \beta_0$, где α_0, β_0 – заданные величины, необходимо определиться с тем, каким является распределение случайной величины U или совместное распределение совокупности случайных величин $U_1, U_2, ..., U_n$ наряду с условиями (3.9)- (3.11). Дополним эти условия требованием, что $U_1, U_2, ..., U_n$ являются случайными величинами, имеющими совместный нормальный закон распределения.

Как показано в [7] (стр. 33-34), в этом случае оценки $\tilde{\alpha}, \tilde{\beta}$ параметров α и β регрессионного уравнения (3.8) являются нормально распределенными случайными величинами, характеристики которых (средние, дисперсии и ковариации) совпадают с аналогичными выражениями для $\hat{\alpha}$ и $\hat{\beta}$. Это означает и то, что определено двумерное распределение для векторной случайной величины ($\tilde{\alpha}, \tilde{\beta}$). Единственная трудность перед проверкой гипотез относительно и связана с наличием в выражениях для $\hat{\alpha}$ и $\hat{\beta}$ неизвестной дисперсии σ_u^2 возмущающей составляющей U.

Дальнейшее изложение опирается на следующий *результат*: статистика $(n-2)\tilde{\sigma}_u^2 / \sigma_u^2$ имеет χ^2-распределение с (n-2) степенями свободы, и эта статистика не зависит от $\tilde{\alpha}$ и $\tilde{\beta}$. В утверждении использовано обозначение $\tilde{\sigma}_u^2$ для выражения $\sum (Y_i - \tilde{\alpha} - \tilde{\beta} X_i)^2 / n$.

Будет также использоваться *закон распределения Стьюдента* (его еще называют *t-распределением*). Он задается следующим образом. Пусть случайная величина Z имеет стандартное нормальное распределение $N(0,1)$, а случайная величина θ^2 не зависит от Z и распределена по закону χ^2 с r степенями свободы, тогда случайная величина

$$\tilde{t} = Z\sqrt{r} / \theta$$

имеет закон распределения Стьюдента с r степенями свободы. Он задается формулой

$$f(t) = C \times \left(1 + \frac{t^2}{r}\right)^{-\frac{r+1}{2}},$$

где C - константа.

Закон распределения Стьюдента симметричен относительно нуля - своего среднего. Когда $r \to \infty$, случайная величина $\tilde{t}$ стремится к своему предельному значению, которое будет иметь стандартный нормальный закон распределения, т.е. $\tilde{t} \approx N(0,1)$.

Определение. Под *статистической гипотезой* будем понимать логическое высказывание о значениях исследуемых параметров.

Например, запись вида

$$H_0 : \beta = \beta_0 \tag{4.1}$$

обозначает выдвижение гипотезы о том, что некоторый параметр β принимает гипотетическое значение β_0. Выражение (4.1) называют *нулевой гипотезой*. Альтернативное высказывание представляет выражение

$$H_1 : \beta \neq \beta_0 . \tag{4.2}$$

Выражение (4.2) называют *двусторонней альтернативной гипотезой*.

4.1.2. Этапы проверки гипотез

В соответствии с общей схемой проверки статистических гипотез нулевая гипотеза при двусторонней альтернативной гипотезе проверяется по следующим шагам.

Шаг 1. Формулировка нулевой и альтернативной гипотез.

Шаг 2. Выбор статистики, по значениям которой судят о справедливости нулевой гипотезы. Под *статистикой* понимается некоторая функция от случайных величин с известным законом распределения.

Шаг 3. Формулировка правила проверки гипотезы и определение необходимого объема выборки. При этом должны быть заданы уровень доверия и величина, связанная с мощностью критерия. Можно поступать иначе, задав уровень доверия и объем выборки и минимизировать мощность критерия.

Шаг 4. В зависимости от проверяемой гипотезы и ее альтернатив выбираем одно- или двустороннюю проверку. Находим область принятия нулевой гипотезы.

Шаг 5. Сравниваем рассчитанную величину статистики с ее теоретическим значением и принимаем решение о принятии или отклонении гипотезы.

Шаг 6. Интерпретируем результаты проверки гипотез.

4.1.3. Пример

Основываясь на исходных данных результатах расчетов примера п. 3.3, выполнить проверку гипотез о значимости параметров α и β уравнения (3.8).

Решение

Исследуем значимость параметра β.

Шаг 1. Нулевую и альтернативную гипотезы примем в виде

$$H_0 : \beta = 0;$$

$$H_1 : \beta \neq 0.$$

Смысл гипотез состоит в том, что фактически предполагается проверить, существенно или нет влияние уровня фондоемкости продукции на затраты на единицу продукции.

Шаг 2. В качестве статистики можно взять функцию вида $t = \frac{\hat{\beta}}{\hat{\sigma}_u}\sqrt{\sum x_i^2}$, которая в соответствии с результатами п. 4.1.1 имеет t-распределение с (n-2) степенями свободы.

Шаг 3. Правило проверки нулевой гипотезы состоит в сопоставлении эмпирического значения t-критерия с теоретическим. Объем выборки равен 10. Уровень доверия предполагаем равным 95 %. Теоретическое значение t-критерия получаем из статистических таблиц или с помощью функции EXCEL СТЬЮДРАСПОБР(0,025; 8). Оно равно 2,751531.

Шаг 4. Выбираем двустороннюю проверку ($\beta < 0, \beta > 0$). В этом случае область принятия нулевой гипотезы имеет вид:

$$[\,\hat{\beta} - t_{0.025;8} \times \hat{\sigma}_u \,/\, \sqrt{\sum x_i^2}\,;\, \hat{\beta} + t_{0.025;8} \times \hat{\sigma}_u \,/\, \sqrt{\sum x_i^2}\,]\,.$$

По результатам расчетов, представленным в таблице 3.1, имеем:

$$\hat{\beta} = 0{,}5;$$

$$t_{0.025;8} = 2{,}751531;$$

$$\hat{\sigma}_u = \sqrt{\hat{\sigma}_u^2 \times 10/8} = \sqrt{2{,}66 \times 10/8} = 1{,}8234;$$

$$\sqrt{\sum x_i^2} = \sqrt{3177{,}6} = 56{,}3702.$$

Отсюда интервал принятия нулевой гипотезы выглядит следующим образом

$$\underline{[0{,}41; 0{,}59]}\,.$$

Шаг 5. Проверяем принадлежность величины $\beta = 0$ нашему интервалу и убеждаемся в том, что это значение не попадает в область принятия нулевой гипотезы. Следовательно, с вероятностью, не меньшей 95 %, нулевую гипотезу отклоняем и переходим к проверке альтернативной гипотезы.

Шаг 6. Таким образом, параметр β значимо отличается от нуля. Последнее означает, что фактор фондоемкости существенно влияет на затраты, приходящиеся на единицу продукции.

Исследуем теперь значимость параметра α.

Шаг 1. Нулевую и альтернативную гипотезы представим в виде

$$H_0 : \alpha = 0;$$

$$H_1 : \alpha \neq 0.$$

Смысл гипотез состоит в том, что необходимо проверить, не существенность влияния постоянно действующих факторов на затраты на единицу продукции.

Шаг 2. В качестве статистики можно взять функцию вида $t = \dfrac{\hat{\alpha}}{\hat{\sigma}_u} \sqrt{n\sum x_i^2} \Big/ \sqrt{\sum X_i^2}$, которая в соответствии с результатами п. 4.1.1 имеет t-распределение с (n-2) степенями свободы.

Шаг 3. Правило проверки нулевой гипотезы состоит в сопоставлении эмпирического значения t-критерия с теоретическим. Объем выборки равен 10. Уровень доверия предполагаем равным 95 %. Теоретическое значение t-критерия получаем из статистических таблиц или с помощью функции EXCEL СТЬЮДРАСПОБР(0,025; 8). Оно равно 2,751531.

Шаг 4. Выбираем двустороннюю проверку ($\alpha < 0, \alpha > 0$). В этом случае область принятия нулевой гипотезы имеет вид:

$$[\,\hat{\alpha} - t_{0.025;8} \times \hat{\sigma}_u \sqrt{\sum X_i^2} \Big/ \sqrt{n\sum x_i^2}\,; \hat{\alpha} + t_{0.025;8} \times \hat{\sigma}_u \sqrt{\sum X_i^2} \Big/ \sqrt{n\sum x_i^2}\,] .$$

По результатам расчетов, представленным в таблице 3.1, имеем:

$$\hat{\alpha} = 3{,}84;$$

$$t_{0.025;8} = 2{,}751531;$$

$$\hat{\sigma}_u = \sqrt{\hat{\sigma}_u^2 \times 10/8} = \sqrt{2{,}66 \times 10/8} = 1{,}8234;$$

$$\sqrt{n\sum x_i^2} = \sqrt{10 \times 3177{,}6} = 178{,}2582.$$

$$\sqrt{\sum X_i^2} = 323{,}7036.$$

Отсюда интервал принятия нулевой гипотезы выглядит следующим образом

$$\underline{[-5{,}27;12{,}96]}\,.$$

Шаг 5. Проверяем принадлежность величины $\alpha = 0$ нашему интервалу и убеждаемся в том, что это значение попадает в область принятия нулевой гипотезы. Следовательно, с вероятностью, не меньшей 95 %, нулевую гипотезу принимаем.

Шаг 6. Таким образом, параметр α не значимо отличается от нуля. Последнее означает, что постоянные факторы не существенно влияют на затраты, приходящиеся на единицу продукции. Значит, оказывается возможным рассматривать другую регрессионную модель вида $Y = \beta \times X + U$.

4.2. Оценка адекватности модели

4.2.1. Немного теории

В настоящем пособии основное внимание уделяется линейным регрессионным моделям. Поэтому *оценку адекватности модели* понимаем как проверку наличия линейной связи между факторами X и Y.

Проверка адекватности модели может быть проведена с помощью коэффициента детерминации и статистики, построенной на его основе [5, 7]. Оценка коэффициента детерминации имеет вид

$$R^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2},$$

где $y_i = Y_i - \overline{Y}$.

Можно показать, что случайная величина (будем обозначать ее через F) $R^2/((1-R^2)/(n-2))$ удовлетворяет закону распределения Фишера с (1; n-2) степенями свободы. Она может быть задействована в качестве статистики при проверке гипотезы об адекватности линейной модели.

Нулевая гипотеза об отсутствии линейной связи между факторами X и Y имеет вид

$$H_0 : \beta = 0.$$

Проверку этой гипотезы выполняют исходя из неравенства

$$F > F_{\varepsilon;1;n-2}, \tag{4.3}$$

выполнение которого на 100×ε % уровне значимости требует отклонения нулевой гипотезы. Величина $F_{\varepsilon,1,n-2}$ берется из статистических таблиц или рассчитывается с помощью функции FРАСПОБР(0,05;1;n-2) электронной таблицы EXCEL.

4.2.2. Пример

Требуется проверить адекватность модели $Y = 3{,}8 + 0{,}5X + u$, построенной в примере п. 3.3.

<u>Решение</u>

Сначала рассчитаем R^2 . Величины $\sum e_i^2$, $\sum y_i^2$ находятся в предпоследнем столбце и предпоследней строке, в последнем столбце и предпоследней строке табл. 3.1 соответственно. Они равны 26,60 и 819,60. Тогда $R^2 = 1 - \frac{26{,}60}{891{,}60} = 0{,}965746$.

Эмпирическое значение F-критерия равно

$$F = \frac{R^2}{((1-R^2)/(n-2))} = \frac{0{,}965746}{((1-0{,}965746)/8)} = 238{,}502.$$

Теоретическое значение F-критерия получаем с помощью функции FРАСПОБР(0,05;1;n-2) электронной таблицы EXCEL:

$$F_{0,05;1;8} = \text{FРАСПОБР}(0{,}05;1;8) = 5{,}317645.$$

Подставляем найденные значения в неравенство (4.3) и убеждаемся, что оно выполнено. Следовательно, с вероятностью, не меньшей 95 %, можно утверждать, что модель $Y = 3{,}8 + 0{,}5X + u$ адекватно описывает влияние уровня фондоемкости на затраты на единицу продукции.

4.3. Краткие выводы

1. Нулевая и альтернативная гипотезы, применяемые для оценки значимости параметров простейшего уравнения парной линейной регрессии, имеют вид:

$$H_0 : \alpha = \alpha_0;$$
$$H_1 : \alpha \neq \alpha_0$$

для параметра α и

$$H_0 : \beta = \beta_0;$$
$$H_1 : \beta \neq \beta_0.$$

для параметра β.

2. Гипотезы о значимости параметра β и адекватности модели парной линейной регрессии можно представить в форме:

$$H_0 : \beta = 0;$$
$$H_1 : \beta \neq 0.$$

3. Для проверки гипотез используют t-критерий Стьюдента и F-критерий Фишера.

4. В общем случае процедура проверки гипотез состоит из шести шагов.

4.4. Вопросы к теме

1. Какие законы распределения случайных величин используются в процедурах проверки гипотез?
2. Что такое статистическая гипотеза?
3. В чем заключается различие между нулевой и альтернативной гипотезами?

4. Назвать этапы проверки гипотез.

4.5. Основные термины и понятия

Статистика
Закон распределения Стьюдента
Закон распределения Фишера
Статистическая гипотеза
Нулевая гипотеза
Альтернативная гипотеза

4.6. Задачи для самостоятельного решения

Используя результаты решения своего варианта задания из п.3 (табл. 3.2), проверить значимость параметров уравнения регрессии и адекватность модели при 95 % уровне доверия.

Тема 5. Общая линейная модель

5.1. Введение. Предположения

Рассмотрение простейшей двухфакторной модели, методов исследования, свойств охватывает лишь элементарные экономические ситуации. Более реальными являются эконометрические модели, включающие несколько факторов.

Будем исходить из того, что с целью исследования линейной связи между результирующим фактором Y и объясняющими k-1 факторами X_2, ..., X_k было выполнено статистическое наблюдение. В результате наблюдения зарегистрирована выборка объема n, которую представим таблицей вида (табл. 5.1)

Предположим, что существует линейная связь между Y и факторами X_2, ..., X_k вида:

$$Y = \beta_1 + \beta_2 X_2 + ... + \beta_k X_k + U, \tag{5.1}$$

Таблица 5.1

Наблюдения факторов

№ п/п	Y	X_2	X_3	…	X_k
1	Y_1	X_{12}	X_{13}	…	X_{1k}
2	Y_2	X_{22}	X_{23}	…	X_{2k}

...	…	...	…	…	…
n	Y_n	X_{n2}	X_{n3}	…	X_{nk}

где слагаемое U отражает влияние на других факторов, ошибки измерений, ошибки выбора типа модели. Это предположение назовем *гипотезой линейности*. Тогда для наблюденных величин (табл. 5.1) можно записать

$$Y_i = \beta_1 + \beta_2 X_{i2} + ... + \beta_k X_{ik} + U_i, i = 1,2,...,n. \tag{5.2}$$

В системе уравнений (5.2) коэффициенты $\beta_1,...,\beta_k$ и параметры распределения для случайных величин Ui неизвестны и должны быть оценены. В этом состоит одна из задач, которые необходимо решать.

Систему (5.2) запишем в матричной форме:

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{U}, \tag{5.3}$$

где

$$\vec{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & X_{12} & ... & X_{1k} \\ 1 & X_{22} & ... & X_{2k} \\ 1 & ... & ... & ... \\ 1 & X_{n2} & ... & X_{nk} \end{bmatrix}, \vec{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ ... \\ \beta_k \end{bmatrix}, \vec{U} = \begin{bmatrix} U_1 \\ U_2 \\ ... \\ U_n \end{bmatrix}. \tag{5.4}$$

В обозначениях (5.4) матрица наблюдений $\mathbf{X}$ содержит столбец, состоящий из единиц. Этот столбец относится к свободному члену β_1. Матрицу $\mathbf{X}$ называют *матрицей регрессоров.*

Для получения оценок вектора примем ряд предположений относительно способа получения наблюдений. В зависимости от сделанных предположений разрабатываются или подбираются методы оценивания.

Основные и наиболее простые гипотезы наряду с гипотезой линейности (5.1) следующие:

$$M(\vec{U}) = \vec{0}; \tag{5.5}$$

$$M(\vec{U}\vec{U}^T) = \sigma^2 I_n; \tag{5.6}$$

$\mathbf{X}$ - матрица фиксированных чисел; (5.7)

$$\rho(\mathbf{X}) = k < n; \tag{5.8}$$

$$\vec{U} \sim N(\vec{0}, \sigma^2 I_n). \tag{5.9}$$

Гипотеза (5.5) означает, что $M(U_i) = 0$ для всех i , т.е. что случайные переменные Ui имеют нулевое математическое ожидание.

Гипотеза (5.6) очень важна в том смысле, что она отражает совокупность свойств ковариационной матрицы In - единичная матрица размерности n. В этой гипотезе выражены свойства, выполнение которых мы требуем. *Первое*: - это свойство некоррелированности случайных величин

$M(U_iU_j)=0, i\neq j, i,j=1,2,...,n$; *второе*: $M(U_i^2)=\sigma^2$ для всех i - свойство постоянства дисперсии случайных величин Ui , называемое *гомоскедастичностью*.

Гипотеза (5.7) является обобщением на случай нескольких факторов аналогичного предположения о постоянстве фактора *X* в случае линейной двухфакторной модели.

Тем самым оказывается, что у вектора единственным источником возмущений является случайный вектор $\vec{U}$. В этом случае говорят, что свойства искомых оценок и критериев *обусловлены* матрицей наблюдений **X**.

Гипотеза (5.8) характеризует, во-первых, отсутствие строгой линейной зависимости между объясняющими переменными $X_1, X_2, ..., X_k (X_k=1)$ и, во-вторых, что число наблюдений *n* больше числа объясняющих переменных *k*. В соотношении (5.8) обозначение $\rho(X)$ выражает ранг матрицы **X**.

Принадлежность вектора возмущений $\vec{U}$ множеству нормально распределенных векторных случайных величин, выраженная в (5.9), фактически объединяет предположения (5.5) и (5.6).

5.2. Нахождение оценки $\hat{\vec{\beta}}$ вектора $\vec{\beta}$ методом наименьших квадратов

Обозначим через $\hat{\vec{\beta}}=\left[\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k\right]^T$ - вектор-столбец, оценивающий вектор $\vec{\beta}$. Можем записать:

$$\vec{Y}=\mathbf{X}\hat{\vec{\beta}}+\vec{e}, \qquad (5.10)$$

где через $\vec{e}$ обозначен вектор-столбец остатков $\vec{Y}-\mathbf{X}\hat{\vec{\beta}}$.

Критерий - сумма квадратов компонент вектора остатков - имеет вид

$$\vec{e}^T\vec{e}=\sum e_i^2=\vec{Y}^T\vec{Y}-2\hat{\vec{\beta}}^T\mathbf{X}^T\vec{Y}+\hat{\vec{\beta}}^T\mathbf{X}^T\mathbf{X}\hat{\vec{\beta}}. \qquad (5.11)$$

Для нахождения значения $\hat{\vec{\beta}}$, минимизирующего эту сумму квадратов отклонений, продифференцируем (5.11) по $\hat{\vec{\beta}}$. Приравнивая полученное выражение нулевому вектору, получаем систему нормальных уравнений в векторно-матричной форме (сравни с (3.14)):

$$\mathbf{X}^T\mathbf{X}\hat{\vec{\beta}}=\mathbf{X}^T\vec{Y}. \qquad (5.12)$$

На основе гипотезы (5.8) получаем основной результат:

$$\hat{\vec{\beta}}=(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{Y}. \qquad (5.13)$$

Оценка $\hat{\vec{\beta}}$ вектора параметров $\vec{\beta}$ найдена.

5.3. Свойства оценок $\hat{\vec{\beta}}$ вектора параметров $\vec{\beta}$

Подставляя (5.3) в (5.13), можно получить следующее важное представление для оценки $\hat{\vec{\beta}}$:

$$\hat{\vec{\beta}} = \vec{\beta} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\vec{U} \,. \tag{5.14}$$

Отсюда сразу следует, что

$$M(\hat{\vec{\beta}}) = \vec{\beta} \,. \tag{5.15}$$

Результат (5.15) означает, что вектор оценок $\hat{\vec{\beta}}$ является несмещенным.

Можно показать, что ковариационная матрица вектора оценок $\hat{\vec{\beta}}$ имеет вид:

$$D(\hat{\vec{\beta}}) = M[(\hat{\vec{\beta}} - \vec{\beta})(\hat{\vec{\beta}} - \vec{\beta})^T] = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \,. \tag{5.16}$$

Равенство (5.16) означает, что дисперсия компоненты $\hat{\beta}_i$ вектора $\hat{\vec{\beta}}$ может быть оценена путем перемножения *i*-го элемента главной диагонали матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$ на дисперсию случайного возмущения σ^2. Аналогично ковариация пары оценок $\hat{\beta}_i$ и $\hat{\beta}_j$ определяется умножением (*i,j*)-го элемента матрицы $(\mathbf{X}^T\mathbf{X})^{-1}$ на σ^2.

Подытоживая рассмотрение свойств оценок $\hat{\vec{\beta}}$, отметим, что в силу предположения (5.9) элементы вектора $\hat{\vec{\beta}}$ удовлетворяют многомерному нормальному распределению, т.е.

$$\hat{\vec{\beta}} \sim \mathrm{N}[\vec{\beta}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}]. \tag{5.17}$$

5.4. Гипотезы о значимости оценок типа $\hat{\vec{\beta}}$

Если дисперсия σ^2 возмущений U известна, то факты, представленные соотношениями (5.13), (5.14), (5.16) могут быть непосредственно использованы для проверки значимости компонент вектора $\hat{\vec{\beta}}$ и построения доверительных интервалов. В случае незнания σ^2 можно поступить следующим образом. Так как $\vec{e}$ и $\hat{\vec{\beta}}$ есть линейные комбинации нормально распределенных случайных величин, то они тоже распределены нормально. Можно показать, что $M[\vec{e}(\hat{\vec{\beta}} - \vec{\beta})^T] = \mathbf{0}_{n\times k}$, где $\mathbf{0}_{n\times k}$ – нулевая матрица из n строк и k столбцов. Последнее означает, что они распределены независимо друг от друга.

Этот результат позволяет использовать t-распределение для проверки гипотез относительно каждого из регрессионных коэффициентов β_i. Величина $\vec{e}^T\vec{e}/\sigma^2$ имеет независимое от $\hat{\beta}_j$ распределение χ^2 с $n-k$ степенями свободы. Отсюда по определению t-распределения величина

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sum e_i^2/(n-k)} \times \sqrt{a_{jj}}} \tag{5.18}$$

удовлетворяет t-распределению с $n-k$ степенями свободы.

Гипотеза о значимости β_j проверяется следующим образом. В (5.18) подставляем интересующее нас гипотетическое значение β_j и рассчитываем t. Значение t сравниваем с критическим $t_{\varepsilon;n-k}$, соответствующим $n-k$ степеням свободы и $100\times(1-\varepsilon)$ %-му уровню доверия. Если окажется, что выполнено неравенство

$$|t| < t_{\varepsilon;n-k},$$

то гипотеза о значимости β_j отбрасывается.

Примером проверки *отсутствия линейной зависимости* Y от X является проверка гипотезы H_0 : $\beta_j = 0$.

Соотношение (5.18) дает $100\times(1-\varepsilon)$ %-ный доверительный интервал для β_j вида

$$\hat{\beta}_j \pm t_{\varepsilon/2;n-k} \times \sqrt{\sum e_i^2/(n-k)} \times \sqrt{a_{jj}}$$

Рассмотрим подход к *совместной проверке гипотез относительно нескольких или всех* β_j.

Выдвинем нулевую гипотезу

$$H_0 : \vec{\beta}^* = \vec{0} \tag{5.19}$$

против альтернативной H_1, состоящей в том, что не все β_j равны нулю. Вектор $\vec{\beta}^* = [\beta_2, \beta_3, ..., \beta_k]^T$. Нулевая гипотеза предполагает, что отсутствует влияние <u>всех</u> $k-1$ факторов $X_2, X_3, ..., X_k$ на Y.

Можно показать, что для проверки нулевой гипотезы (5.19) применим F-критерий вида

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}, \tag{5.20}$$

где $R^2 = R^2_{Y/X_2,...,X_k}$ – коэффициент детерминации (квадрат коэффициента множественной корреляции).

Гипотеза H_0 (5.19) при $100\times(1-\varepsilon)$ %-ном уровне значимости отклоняется, если выполнено неравенство

$$F > F_{\varepsilon;k-1;n-k}, \tag{5.21}$$

где F рассчитывается по формуле (5.20), а $F_{\varepsilon;k-1;n-k}$ определяется из таблицы теоретических значений F-критерия при $100\times(1-\varepsilon)$ %-ном уровне значимости и (*k-1*), (*n-k*) степенях свободы. Это значение можно определить с помощью функции EXCEL FРАСПОБР(ε, *k*-1, *n-k*) Отклонение гипотезы H_0 содержательно означает, что между фактором *Y* и *всеми* факторами $X_2,...,X_k$ имеется линейная статистическая связь. О силе связи можно судить по величине коэффициента детерминации R^2.

Используя результаты настоящего раздела, можно ставить и решать задачи добавления или исключения одного *и/или* нескольких факторов X_i. Подробнее об этом см. в [6, стр. 144-147].

5.5. Прогноз

Дальнейшее использование соотношений типа (5.1) связано с получением информации об ожидаемом значении фактора *Y*, которое должно соответствовать некоторому сочетанию значений факторов $X_2,...,X_k$, не наблюдавшемуся в выборке. Если отождествить номера наблюдений 1,2,…, n с номерами временных периодов, то конкретизация *задачи прогнозирования* может выглядеть так: по ожидаемым в периоде *n*+1 значениям вектора

$$\vec{C}^T = [1, X_{2.n+1},...,X_{k,n+1}]$$

требуется построить прогноз ожидаемого значения Y_{n+1}, т.е. $M(Y_{n+1}|\vec{C})$. С учетом соотношения (5.2) получим

$$M(Y_{n+1}) = M(\vec{C}^T\vec{\beta} + U_{n+1}) = \vec{C}^T\vec{\beta}.$$

Можно построить либо точечный, либо интервальный прогноз.

Наилучший линейный несмещенный прогноз для $\vec{C}^T\vec{\beta}$ имеет вид $\vec{C}^T\hat{\vec{\beta}}$. Поэтому искомый *точечный прогноз* есть

$$\hat{Y}_{n+1} = \vec{C}^T\hat{\vec{\beta}} = \hat{\beta}_1 + \hat{\beta}_2 X_{2,n+1} + ...\hat{\beta}_k X_{k,n+1}. \tag{5.22}$$

Соответствующий $100\times(1-\varepsilon)$ %-ный доверительный интервал для точечного прогноза имеет вид

$$\vec{C}^T\hat{\vec{\beta}} - t_{\varepsilon/2;n-k}S[\vec{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\vec{C}]^{\frac{1}{2}} \le \vec{C}^T\vec{\beta} \le \vec{C}^T\hat{\vec{\beta}} + t_{\varepsilon/2;n-k}S[\vec{C}^T(\mathbf{X}^T\mathbf{X})^{-1}\vec{C}]^{\frac{1}{2}}. \tag{5.23}$$

В неравенствах (5.25) $S = [\sum e_i^2/(n-k)]^{\frac{1}{2}}$.

Таким образом, ожидаемое значение $M(Y_{n+1}|C)$ величины Y_{n+1} с вероятностью (1-ε) накрывается интервалом (5.23).

Задача интервального прогнозирования состоит в построении доверительного интервала не для среднего $M(Y_{n+1}|C)$, а для изолированного значения Y_{n+1}, которое должно ассоциироваться с вектором $\vec{C}^T = [1, X_{2.n+1},...,X_{k,n+1}]$.

Ее решение сводится к выдвижению и проверке гипотезы H_0 о соответствии значения Y_{n+1} вектору $\vec{C}^T = [1, X_{2.n+1}, ..., X_{k,n+1}]$. При этом используется t-статистика вида

$$t = \frac{\hat{Y}_{n+1} - Y_{n+1}}{S[1 + \vec{C}^T (\mathbf{X}^T \mathbf{X})^{-1} \vec{C}]^{\frac{1}{2}}}, \tag{5.24}$$

удовлетворяющая t – распределению с $n-k$ степенями свободы. Если найденное по (5.24) значение t не превышает критического $t_{\varepsilon;n-k}$, то делаем вывод, что гипотеза H_0 имеет место, т.е. предполагаемое значение Y_{n+1} соответствует значениям факторов $X_{2,n+1}, ..., X_{k,n+1}$.

5.6. Пример

Известны данные о динамике макропоказателей экономики Украины за 1999-2000 гг. (табл. 5.1)[1]. Требуется осуществить прогноз денежной массы, уровня инфляции, уровня безработицы и помесячного ВВП на ноябрь 2000 года.

Решение. Начнем с ВВП. Имеется, по крайней мере, два пути решения задачи прогнозирования ВВП по приведенным данным.

Первый, простейший, состоит в построении прогнозного значения помесячного ВВП на основе данных только последнего столбца табл. 5.1. В качестве прогнозной модели можно взять стандартное уравнение парной линейной регрессии, в котором фактором, определяющим влияние на ВВП, будет время:

$$Y = \alpha + \beta \times t + \varepsilon. \tag{5.25}$$

Второй, более сложный, состоит в привлечении дополнительной информации о поведении экономики Украины, которую отражают факторы: денежная масса, уровень инфляции, уровень безработицы. В этом случае для построения прогноза может быть использована модель множественной регрессии вида:

$$Y = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \beta_3 \times X_3 + \beta_4 \times t + \varepsilon, \tag{5.26}$$

где:

X_1 – денежная масса (агрегат М3), млн. грн.;

X_2 – уровень инфляции (индекс потребительских цен), % к уровню предыдущего месяца;

X_3 – уровень безработицы, %.

Таблица 5.1

Макропоказатели экономики Украины за 1999-2000 гг.

[1] Данные Украинского финансового сервера от 03.12.2000

Год	Месяц	ВПП, млн. грн.	Денежная масса (агрегат М3), млн. грн.	Уровень инфляции (индекс потребительских цен), % к уровню предыдущего месяца	Уровень безработицы, %	ВПП помесячный, млн. грн.
1999	Январь	8017,00	15185,00	101,50	3,81	8017,00
1999	Февраль	15977,00	15366,00	101,00	3,39	7960,00
1999	Март	25157,00	15937,00	101,00	4,04	9180,00
1999	Апрель	34298,00	16680,00	102,30	4,06	9141,00
1999	Май	44341,00	17496,00	102,40	4,03	10043,00
1999	Июнь	55267,00	18579,00	100,10	3,98	10926,00
1999	Июль	66349,00	18816,00	99,00	4,02	11082,00
1999	Август	78329,00	19694,00	101,00	4,08	11980,00
1999	Сентябрь	92324,00	20468,00	101,40	4,12	13995,00
1999	Октябрь	103653,00	20899,00	101,10	4,14	11329,00
1999	Ноябрь	116504,00	21042,00	102,90	4,20	12851,00
1999	Декабрь	127126,00	22079,00	104,10	4,30	10622,00
2000	Январь	9930,00	22052,00	104,60	4,30	9930,00
2000	Февраль	20821,00	22975,00	103,30	4,45	10891,00
2000	Март	32731,00	23632,00	102,00	4,50	11910,00
2000	Апрель	45166,00	25086,00	101,70	4,50	12435,00
2000	Май	57871,00	26031,00	102,10	4,39	12705,00
2000	Июнь	71337,00	27098,00	103,70	4,29	13466,00
2000	Июль	87441,00	28127,00	99,90	4,24	16104,00
2000	Август	104005,00	29273,00	100,00	4,24	16564,00
2000	Сентябрь	121974,00	14822,00	102,60	4,20	17969,00
2000	Октябрь	137384,00	28750,00	101,40	4,20	15410,00

Кроме того, могут быть построены точечный и интервальный прогнозы. Рассмотрим построение точечного прогноза.

I. В соответствии с теоретическими положениями, изложенными в п. 5.5 искомый точечный прогноз, отвечающий вектору $\vec{C}$, получается из (5.22). Применяя, подход, рассмотренный в примере п. 3.3, получим уравнение парной линейной регрессии (называемое *трендом*) в виде:

$$\hat{Y} = 7937,43 + 355,28 \times t .$$

Действуя, как в примере п. 4.1.3, проверяем значимость параметров α и β, а также адекватность модели. Оказывается, что параметры и модель адекватны. Следовательно, она может быть применена для получения точечного прогноза ВВП за ноябрь 2000 года. Подставляя, значение ноября месяца 2000 (равное 23) в уравнение модели, получаем, что $\hat{Y} = 16108,94$. Другими словами, в ноябре 2000 года в Украине следует ожидать величину месячного ВВП в размере 16108,94 млн. грн. Качество точечного прогноза можно охарактеризовать величиной доверительного интервала, определяемого неравенствами (5.23). В нашем примере:

$$\mathbf{X}=\begin{bmatrix}1 & 1\\1 & 2\\1 & 3\\1 & 4\\1 & 5\\1 & 6\\1 & 7\\1 & 8\\1 & 9\\1 & 10\\1 & 11\\1 & 12\\1 & 13\\1 & 14\\1 & 15\\1 & 16\\1 & 17\\1 & 18\\1 & 19\\1 & 20\\1 & 21\\1 & 22\end{bmatrix}$$

$$\mathbf{X^TX}=\begin{bmatrix}\mathbf{22} & \mathbf{253}\\ \mathbf{253} & \mathbf{3795}\end{bmatrix},\ \mathbf{(X^TX)^{-1}}=\begin{bmatrix}\mathbf{0{,}1948} & \mathbf{-0{,}0130}\\ \mathbf{-0{,}0130} & \mathbf{0{,}0011}\end{bmatrix}$$

$$\vec{\mathbf{C}}=\begin{bmatrix}\mathbf{1}\\ \mathbf{23}\end{bmatrix},\ \vec{\mathbf{C}}^{\mathbf{T}}=\begin{bmatrix}\mathbf{1} & \mathbf{23}\end{bmatrix}$$

$$(\vec{\mathbf{C}}^{\mathbf{T}}(\mathbf{X^TX})^{-1}\vec{\mathbf{C}})^{\frac{1}{2}}=$$

$$\left(\begin{bmatrix}\mathbf{1} & \mathbf{23}\end{bmatrix}\begin{bmatrix}\mathbf{0{,}1948} & \mathbf{-0{,}0130}\\ \mathbf{-0{,}0130} & \mathbf{0{,}0011}\end{bmatrix}\begin{bmatrix}\mathbf{1}\\ \mathbf{23}\end{bmatrix}\right)^{\frac{1}{2}}=$$

$$\left(\begin{bmatrix}\mathbf{-0{,}1039} & \mathbf{0{,}0130}\end{bmatrix}\begin{bmatrix}\mathbf{1}\\ \mathbf{23}\end{bmatrix}\right)^{\frac{1}{2}}=(\mathbf{0{,}195})^{\frac{1}{2}}=\mathbf{0{,}441.}$$

$$t_{0,025;20}=\text{СТЬЮДРАСПОБР}(0,025;20)=2,42,$$

$$S=\left[\sum e_i^2/(n-k)\right]^{\frac{1}{2}}=\left[42695344/20\right]^{\frac{1}{2}}=1461,08.$$

$$t_{0,025;20}S\left[\vec{C}^T(X^TX)^{-1}\vec{C}\right]^{\frac{1}{2}}=2,42\times1461,08\times0,441=1562,6.$$

Замечание. Если выполнить фактическое перемножение в последнем выражении, получится 5095,11. Результат в формуле более точен, так как получен с девятью значащими цифрами после десятичной точки.

Таким образом, доверительный интервал для прогнозируемого значения ВВП на ноябрь 2000 года имеет вид:

$$14546,3\le(\text{Прогнозируемое значение ВВП})_{\text{ноябрь 2000}}\le17671,5\,.$$

Окончательный *вывод* заключается в том, что не менее чем в 95 случаях из 100 (т.е. с вероятностью, превышающей 95 %) *истинное прогнозное значение ВВП* накрывается интервалом $\left[14546,3;17671,5\right]$.

II. Не каждого министра финансов удовлетворит такой прогноз. Плюс–минус полтора миллиарда гривен – это что-то значит. Поэтому рассмотрим общую ситуацию, когда наряду с фактором времени для решения задачи точечного прогнозирования привлекаются остальные факторы, имеющиеся в нашем распоряжении: денежная масса (агрегат М3), уровень инфляции (индекс потребительских цен), уровень безработицы.

Привлечем модель линейной множественной регрессии (5.26). Для экономии усилий шире используем возможности электронной таблицы EXCEL и, в частности функцию ЛИНЕЙН(). Краткая справка по применению функции ЛИНЕЙН() приведена в приложении.

Сначала сформируем вектор $\vec{C}$ применительно к расширенной матрице экзогенных переменных **X** модели (5.26). Расширенная матрица имеет вид

X =

1	15185	101,5	3,81	1
1	15366	101,0	3,39	2
1	15937	101,0	4,04	3
1	16680	102,3	4,06	4
1	17496	102,4	4,03	5
1	18579	100,1	3,98	6
1	18816	99,0	4,02	7
1	19694	101,0	4,08	8
1	20468	101,4	4,12	9
1	20899	101,1	4,14	10
1	21042	102,9	4,20	11
1	22079	104,1	4,30	12
1	22052	104,6	4,30	13
1	22975	103,3	4,45	14
1	23632	102,0	4,50	15
1	25086	101,7	4,50	16
1	26031	102,1	4,39	17
1	27098	103,7	4,29	18
1	28127	99,9	4,24	19
1	29273	100,0	4,24	20
1	14822	102,6	4,22	21
1	28750	101,4	4,20	22

Ко второму, третьему и четвертому столбцам матрицы применим функцию EXCEL ПРЕДСКАЗ(). В результате искомый вектор примет вид

$$\vec{C} = \begin{bmatrix} 1 & 27736,23 & 102,21 & 4,47 & 23 \end{bmatrix}^T .$$

Теперь для получения оценок параметров уравнения множественной линейной регрессии (5.25) воспользуемся функцией EXCEL ЛИНЕЙН(). В результате получим итоговую таблицу (см. табл. 5.2).

Отсюда результирующее уравнение линейной множественной регрессии принимает вид

$$Y = 75885,24 - 0,20664 \times X_1 - 618,346 \times X_2 - 566,788 \times X_3 + 508,189 \times t .$$

Таблица 5.2

Расчетные оценки функции ЛИНЕЙН()

508,189	-566,788	-618,346	-0,20664	75885,24

68,529	1606,14	199,2031	0,092151	18824,44
0,858	1134,491			
25,754	17			
132588202	21880205			

Для проверки гипотез о значимости оценок регрессионных параметров определяем теоретическое значение t-критерия: $t_{0,025;17} = СТЬЮДРАСПОБР(0,025;17) = 2,458$. Делением значений первой строки табл. 5.2 на вторую получаем эмпирические значения t-статистики:

7,41 -0,353 -3,104 -2,242 4,031

Сравниваем их с теоретическим и получаем, что неравенство

$$|t| < t_{0,025;17}$$

выполнено для оценок β_1, β_3. Это означает не значимость влияния денежного фактора и фактора безработицы на ВВП с вероятностью, не меньшей 95 %. Остальные факторы оказываются значимыми. Проверка адекватности модели в целом показывает, что модель можно считать адекватной (неравенство

$$25,754 = F_{эмпир} > F_{0,05;5;17} = \text{FРАСПОБР}(0,05;5;17) = 2,81$$

выполнено).

Выполним подстановку значений вектора $\vec{C}$ в уравнение и получим прогнозную оценку

$$\hat{Y} = 16108,94,$$

которая практически совпадает с прогнозной оценкой в случае учета зависимости ВВП *только* от фактора времени *t*.

Полученный результат можно проинтерпретировать как факт неполноты набора использованных факторов и данных. Вывод состоит в том, что для построения удовлетворительного точечного прогноза ВВП по имеющимся факторным признакам, характеризующим состояние макроэкономики Украины, оказывается достаточным использовать фактор времени.

Тема 6. Некоторые аспекты многомерной регрессии

В разделе рассматриваются некоторые проблемы, часто возникающие при практическом использовании многомерных регрессионных моделей.

На практике нередко приходится сталкиваться с ситуацией, когда регрессия является "плохой", т.е. t-статистики большинства оценок малы, что свидетельствует о незначимости соответствующих независимых переменных (регрессоров). В то же время F-статистика может быть достаточно большой, что говорит о значимости регрессии в целом. Характерным является пример из предыдущего раздела. Одна из возможных причин такого явления носит название *мультиколлинеарности* и возникает при наличии высокой корреляции между регрессорами.

Регрессионные модели являются достаточно гибким инструментом, позволяющим, в частности, оценивать влияние качественных признаков (пол, профессия, наличие детей и т.п.) на изучаемую переменную. Это достигается введением в число регрессоров так называемых *фиктивных переменных,* принимающих, как правило, значения 1 или 0 в зависимости от наличия или отсутствия соответствующего признака в очередном наблюдении. С формальной точки зрения фиктивные переменные ничем не отличаются от других регрессоров. Наиболее сложный и интересный вопрос, возникающий при их использовании, — это правильная интерпретация получаемых оценок.

В этом разделе мы также рассмотрим задачу нахождения частной корреляции между переменными и так называемую проблему спецификации модели.

6.1. Мультиколлинеарность

Одним из условий классической регрессионной модели является предположение о линейной независимости объясняющих переменных, что означает линейную независимость столбцов матрицы регрессоров $\mathbf{X}$ или (эквивалентно), что матрица $(\mathbf{X}^T\mathbf{X})^{-1}$ имеет полный ранг k. При нарушении этого условия, т.е. когда один из столбцов матрицы $\mathbf{X}$ есть линейная комбинация остальных столбцов, говорят, что имеет место *полная коллинеарность.* В этой ситуации нельзя построить МНК-оценку вектора параметров $\vec{\beta}$, что формально следует из вырожденности матрицы $\boldsymbol{X^TX}$ и невозможности решить систему нормальных уравнений. Нетрудно также понять и содержательный смысл этого явления. Дело в том что, одни и те же наблюдения могут быть объяснены различными наборами коэффициентов β_i. Эта ситуация тесно связана с проблемой *идентифицируемости* системы, о чем более подробно будет говориться позднее. Если есть полная коллинеарность, то можно выделить в матрице $\boldsymbol{X}$ максимальную линейно независимую систему столбцов и, удалив остальные столбцы, провести новую регрессию.

На практике полная коллинеарность встречается исключительно редко. Гораздо чаще приходится сталкиваться с ситуацией, когда матрица $\boldsymbol{X}$ имеет полный ранг, но между регрессорами имеется высокая степень корреляции, т.е. когда матрица $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, говоря нестрого, близка к вырожденной. Тогда говорят о наличии мультиколлинеарности. В этом случае МНК-оценка формально существует, но обладает "плохими" свойствами.

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

Выделим некоторые наиболее характерные признаки мультиколлинеарности.

1. Небольшое изменение исходных данных (например, добавление новых наблюдений) приводит к существенному изменению оценок коэффициентов модели.

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

3. Оценки коэффициентов имеют неправильные с точки зрения теории знаки или неоправданно большие значения.

Что же делать, если по всем признакам имеется мультиколлинеарность? Однозначного ответа на этот вопрос нет.

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

6.2. Фиктивные переменные

Как правило, независимые переменные в регрессионных моделях имеют "непрерывные" области изменения (национальный доход, уровень безработицы, размер зарплаты и т.п.). 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.

Качественные различия можно формализовать с помощью любой переменной, принимающей два значения, а не обязательно значения 0 или 1. В эконометрической практике почти всегда используют лишь фиктивные переменные типа "0-1", поскольку в этом случае интерпретация выглядит наиболее просто.

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 целесообразнее использовать несколько бинарных переменных. Типичным примером подобной ситуации является исследование сезонных колебаний. Пусть, например, y_t — объем потребления некоторого продукта в месяц t, и есть все основания считать, что потребление зависит от времени года. Для выявления влияния сезонности можно ввести три бинарные переменные d_1, d_2, d_3:

$d_{i1} = 1$, если месяц t является зимним, $d_{i1} = 0$ в остальных случаях;

$d_{i2}. = 1$, если месяц / является весенним, $d_{i2} = 0$ в остальных случаях;

$d_{i3} = 1$, если месяц / является летним, $d_{i3} = 0$ в остальных случаях, и

оценивать уравнение

$$Y_t = \beta_0 + \beta_1 d_{t1} + \beta_2 d_{t2} + \beta_3 d_{t3} + \varepsilon_t, t = 1,...,n. \qquad (6.1)$$

Отметим, что мы не вводим четвертую бинарную переменную d_4, относящуюся к осени, иначе тогда для любого месяца t выполнялось бы тождество $d_{i1} + d_{i2} + d_{i3} + d_{i4} = 1$, что означало бы линейную зависимость регрессоров в (6.1) и, как следствие, невозможность получения МНК-оценок.

Фиктивные переменные, несмотря на свою внешнюю простоту, являются весьма гибким инструментом при исследовании влияния качественных признаков.

В заключение этого раздела 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. Следует только быть внимательным, чтобы при включении нескольких бинарных переменных не нарушить линейную независимость регрессоров (см. выше пример с сезонными колебаниями).

Выводы:

1) для исследования влияния качественных признаков в модель можно вводить бинарные (фиктивные) переменные, которые, как правило, принимают значение 1, если данный качественный признак присутствует в наблюдении, и значение 0 при его отсутствии;

2) способ включения фиктивных переменных зависит от априорной информации относительно влияния соответствующих качественных признаков на зависимую переменную и от гипотез, которые проверяются с помощью модели;

3) от способа включения фиктивной переменной зависит и интерпретация оценки коэффициента при ней.

6.3. Частная корреляция

В том случае, когда имеется одна независимая и одна зависимая переменные, естественной мерой зависимости (в рамках линейного подхода) является (выборочный) коэффициент корреляции между ними. Использование многомерной регрессии позволяет обобщить это понятие на случай, когда имеется несколько независимых переменных. Корректировка здесь необходима по следующим соображениям. Высокое значение коэффициента корреляции между исследуемой зависимой и какой-либо независимой переменной может означать высокую степень зависимости, но может быть обусловлено и другой причиной. А именно, существует 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, исключив (линейное) влияние других факторов. Это можно сделать с помощью коэффициента частной корреляции.

Для простоты предположим, что имеется обычная двумерная регрессионная модель

$$\vec{Y} = \vec{I}\beta_0 + \vec{X}_1\beta_1 + \vec{X}_2\beta_2 + \vec{\varepsilon},$$

где, как обычно, $\vec{Y}$ — $n \times 1$ вектор наблюдений зависимой переменной, $\vec{X}_1$, $\vec{X}_2$ — $n \times 1$ векторы независимых переменных, $\beta_0, \beta_1, \beta_2$ — (скалярные) параметры, $\vec{\varepsilon}$ — $n \times 1$ вектор ошибок. Наша *цель* — определить корреляцию между $\vec{Y}$ и, например, первым регрессором $\vec{X}_1$ после исключения влияния $\vec{X}_2$.

Соответствующая процедура устроена следующим образом.

1. Осуществим регрессию $\vec{Y}$ на $\vec{X}_2$ и константу и получим прогнозные значения $\hat{\vec{Y}} = \hat{\alpha}_1\vec{I} + \hat{\alpha}_2\vec{X}_2$.

2. Осуществим регрессию $\vec{X}_1$ на $\vec{X}_2$ и константу и получим прогнозные значения $\hat{\vec{X}}_1 = \hat{\gamma}_1\vec{I} + \hat{\gamma}_2\vec{X}_2$.

3. Удалим влияние X_2, взяв остатки $\vec{e}_Y = \vec{Y} - \hat{\vec{Y}}$ и $\vec{e}_{X_1} = \vec{X}_1 - \hat{\vec{X}}_1$.

4. Определим (выборочный) *коэффициент частной корреляции* между Y и X_1 при исключении влияния X_2 как (выборочный) коэффициент корреляции между $\vec{e}_Y$ и $\vec{e}_{X_1}$:

$$r(Y, X_1 \mid X_2) = r(\vec{e}_Y, \vec{e}_{X_1}) \qquad (6.2).$$

Из свойств метода наименьших квадратов следует, что $\vec{e}_Y$ и $\vec{e}_{X_1}$ 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:

$$r(Y, X_1 \mid X_2) = \frac{r(Y, X_1) - r(Y, X_2)r(X_1, X_2)}{\sqrt{1 - r^2(X_1, X_2)}\sqrt{1 - r^2(Y, X_2)}}. \qquad (6.3)$$

Значения $r(Y, X_1 \mid X_2)$ лежат в интервале [-1,1], как у обычного коэффициента корреляции. Равенство коэффициента $r(Y, X_1 \mid X_2)$ нулю означает, говоря нестрого, отсутствие *прямого* (линейного) *влияния* переменной X_1 на Y.

Существует тесная связь между коэффициентом частной корреляции $r(Y, X_1 \mid X_2)$ и коэффициентом детерминации R^2, а именно

$$r^2(Y, X_1 \mid X_2) = \frac{R^2 - r^2(Y, X_2)}{1 - r^2(Y, X_2)}.$$

Описанная выше процедура очевидным образом обобщается на случай, когда исключается влияние не одной, а нескольких переменных: достаточно переменную X_2 заменить на набор переменных $\{X_i\}$, сохраняя определение (6.2). Формула (6.3), естественно, усложнится.

Процедура пошагового отбора переменных

Коэффициент частной корреляции часто используется при решении проблемы *спецификации модели* (см. далее п. 6.4). Остановимся на этом аспекте более подробно.

Иногда исследователь заранее знает характер зависимости исследуемых величин, опираясь, например, на экономическую теорию, предыдущие результаты, априорные знания и т.п., и задача состоит лишь в оценивании неизвестных параметров. (По существу, во всех наших предыдущих рассуждениях мы неявно предполагали, что имеется именно такая ситуация.) Классический пример — оценивание параметров производственной функции Кобба-Дугласа $Y = AK^{\alpha}L^{\beta}$, где Y — совокупный выпуск, K — капиталовложения и L — трудозатраты. Логарифмируя это равенство, получаем линейное относительно $\ln A, \alpha, \beta$ уравнение, из которого, например, с помощью метода наименьших квадратов можно получить оценки этих параметров, проверять те или иные гипотезы и т.д.

Однако на практике довольно часто приходится сталкиваться с ситуацией, когда имеется большое число наблюдений различных параметров (независимых переменных), но нет априорной модели изучаемого явления. Возникает естественная проблема, какие переменные включить в регрессионную схему. Теоретические вопросы, связанные с этой проблемой, будут изложены далее, в п. 4.4.

В компьютерные пакеты включены различные эвристические процедуры *пошагового отбора регрессоров.* Основными пошаговыми процедурами являются *процедура последовательного присоединения, процедура присоединения-удаления и процедура последовательного удаления.* Опишем кратко одну из таких процедур, использующую понятие коэффициента частной корреляции.

Процедура присоединения-удаления

На *первом шаге* из исходного набора объясняющих переменных выбирается (включается в число регрессоров) переменная, имеющая наибольший по модулю коэффициент корреляции с зависимой переменной Y.

Второй шаг состоит из двух подшагов. На первом из них, который выполняется, если число регрессоров уже больше двух, делается попытка исключить один из регрессоров. Ищется тот регрессор X_s, удаление которого приводит к наименьшему уменьшению коэффициента детерминации. Затем сравнивается значение F-статистики для проверки гипотезы $\mathbf{H_0}$ о незначимости этого регрессора с некоторым заранее заданным пороговым значением $F_{\varepsilon;k;n-k}$ (см. (4.3)). Если $F < F_{\varepsilon;k;n-k}$, то X_s удаляется из списка регрессоров. Заметим, что гипотеза $\mathbf{H_0}$ о равенстве коэффициента при X_s нулю эквивалентна гипотезе о равенстве коэффициентов детерминации до и после удаления регрессора, а также гипотезе о том, что коэффициент частной корреляции X_s и Y равен 0. Второй подшаг состоит в попытке включения нового регрессора из исходного набора предсказывающих переменных. Ищем переменную X_p *с* наибольшим по модулю частным коэффициентом корреляции (исключается влияние ранее включенных в уравнение регрессоров) и сравниваем значение F-статистики для проверки гипотезы $\mathbf{H_0}$ о незначимости этого регрессора с некоторым заранее заданным пороговым значением $F_{\varepsilon;k;n-k}$. Если $F > F_{\varepsilon;k;n-k}$ то X_p включается в список регрессоров. Второй шаг повторяется до тех пор, пока происходит изменение списка регрессоров. Конечно, ни одна из пошаговых процедур не гарантирует получение оптимального по какому-либо критерию набора регрессоров.

6.4. Обобщенный метод наименьших квадратов

Одно из предположений классической регрессионной модели состоит в том, что случайные ошибки некоррелированы между собой и имеют постоянную дисперсию. В тех случаях, когда наблюдаемые объекты достаточно однородны, не сильно отличаются друг от друга, такое допущение оправдано. Однако во многих ситуациях такое предположение нереалистично. Например, если исследуется зависимость расходов на питание в семье от ее общего дохода, то естественно ожидать, что разброс выданных будет выше для семей с более высоким доходом. Это означает, что дисперсии зависимых величин (а, следовательно, и случайных ошибок) не постоянны. Это явление в эконометрике называется *гетероскедастичностью* (в отличие от *гомоскедастичности* — равенства дисперсий). Кроме того, при анализе временных рядов в довольно редких случаях можно считать, что наблюдения некоррелированы во времени. Как правило, значение исследуемой величины в текущий момент времени статистически зависит от ее значений в прошлом, что означает наличие корреляции между ошибками. Поэтому естественно изучать модели регрессии без предположения, что $\operatorname{cov}(\vec{U}\vec{U}^T) = V(\vec{U}) = \sigma^2\mathbf{I}$ (см. раздел 5.1 формула (5.6)).

В данном разделе мы будем рассматривать так называемую обобщенную регрессионную модель

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{U}\,, \tag{6.4}$$

где $\vec{Y}$ — $n\times 1$ вектор зависимых переменных, $\mathbf{X}$ — $n\times k$ матрица независимых переменных, $\vec{\beta}$ — $k\times 1$ вектор неизвестных параметров, $\vec{U}$— $n\times 1$ вектор случайных ошибок, причем:

1) матрица **X** неслучайна и имеет полный ранг k;

2) $M(\vec{U})=\vec{0}$;

3) $V(\vec{U})=\mathbf{\Omega}$ и матрица $\mathbf{\Omega}$ положительно определена.

Иными словами, обобщенная модель отличается от классической только условием 3).

К системе (6.4) можно применить обычный метод наименьших квадратов. Но оценка матрицы ковариаций вектора $\vec{\beta}$, получаемая при использовании обычного метода наименьших квадратов, является смещенной.

Для получения эффективной оценки надо воспользоваться так называемым *обобщенным методом наименьших квадратов* (ОМНК).

Сущность обобщенного метода наименьших квадратов. Ответ на вопрос об эффективной линейной несмещенной оценке вектора $\vec{\beta}$ для модели (6.4) дает следующая теорема.

Теорема Айткена. В классе линейных несмещенных оценок вектора $\vec{\beta}$ для обобщенной регрессионной модели оценка

$$\hat{\vec{\beta}}^* = (\mathbf{X}^{\mathrm{T}}\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{\Omega}^{-1}\vec{Y} \tag{6.5}$$

имеет наименьшую матрицу ковариаций.

Подчеркнем, что для применения ОМНК необходимо знать матрицу $\mathbf{\Omega}$, что на практике бывает крайне редко. Поэтому вполне естественным кажется такой способ: оценить (каким-нибудь образом) матрицу $\mathbf{\Omega}$, а затем использовать эту оценку в формуле (6.5) вместо $\mathbf{\Omega}$,. Изложенный подход составляет суть так называемого *доступного обобщенного метода наименьших квадратов.*

6.5. Гетероскедастичность и корреляция по времени

Этот раздел посвящен изучению *двух* важных классов обобщенных регрессионных моделей. *Первый* составляют модели с гетероскедастичностью. Термин «*гетероскедастичность*» применяется в ситуации, когда матрица ковариаций вектора ошибок является диагональной, но элементы главной диагонали, вообще говоря, различны. Иными словами, ошибки в разных наблюдениях некоррелированы, но их дисперсии — разные. Модели *второго* класса, как правило, используются при анализе данных, имеющих характер временных рядов. В этих случаях часто приходится принимать во внимание то обстоятельство, что наблюдения в разные моменты времени статистически зависимы (типичный пример — ежедневный обменный курс доллара по отношению к гривне). Следовательно, ошибки, относящиеся к разным наблюдениям (разным моментам времени), могут быть коррелированы, и ковариационная матрица вектора ошибок не является диагональной. Формально проблему оценивания неизвестных параметров решает обобщенный метод наименьших квадратов, рассмотренный в предыдущей главе. Однако, как там отмечалось, его применение требует знания матрицы ковариаций $\mathbf{\Omega}$, вектора ошибок, что бывает крайне редко. Поэтому, помимо теоретических вопросов, в данной главе будут затронуты некоторые аспекты практического использования ОМНК.

6.1. Гетероскедастичность

В этом разделе мы рассмотрим частный случай обобщенной регрессионной модели, а именно, модель с *гетероскедастичностью.* Это означает, что ошибки некоррелированы, но имеют непостоянные дисперсии. (Классическая модель с постоянными дисперсиями ошибок называется *гомоскедастичной.)* Как уже отмечалось, гетероскедастичность довольно часто возникает, если анализируемые объекты, говоря нестрого, неоднородны. Например, если исследуется зависимость прибыли предприятия от каких-либо факторов, скажем, от размера основных фондов, то естественно ожидать, что для больших предприятий колебание прибыли будет выше, чем для малых.

Метод взвешенных наименьших квадратов

Итак, пусть

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{U}, \tag{6.6}$$

и предположим, что ковариационная матрица $\mathbf{\Omega}$ вектора ошибок $\vec{U}$ диагональная, $V(U_t) = \sigma_t^2$, $t = 1, \ldots, n$. Иногда удобно использовать представление $\sigma_t^2 = \sigma^2 \omega_t$, где числа ω_t, нормированы таким образом, что $\sum \omega_t = n$. Тогда при $\omega_t = 1$, $t = 1, \ldots, n$, модель сводится к классической.

Обобщенный метод наименьших квадратов в данном случае выглядит очень просто — вспомогательная система получается делением каждого уравнения в (6.6) на соответствующее σ_t, (здесь нам удобнее выписать каждое уравнение):

$$\frac{Y_t}{\sigma_t} = \sum_{j=1}^{k} \beta_{tj} \frac{X_{tj}}{\sigma_t} + w_t, t = 1, \ldots, n, \tag{6.7}$$

где $w_t = U_t / \sigma_t$ причем $V(w_t) = 1, \mathrm{cov}(w_t, w_s) = 0$, при $t \neq s$. Применяя к (6.7) стандартный метод наименьших квадратов, ОМНК-оценку получаем минимизацией по $\vec{b} = [b_1, ..., b_k]^T$ суммы

$$f(\vec{b}) = \sum_{t=1}^{n} \left[\frac{1}{\sigma_t} \left(Y_t - \sum_{j=1}^{k} b_j X_{tj} \right) \right]^2 .$$

Нетрудно понять содержательный смысл этого преобразования. Используя обычный метод наименьших квадратов, мы минимизируем сумму квадратов отклонений $\varphi(\vec{b}) = \sum_{t=1}^{n} \left(Y_t - \sum_{j=1}^{k} b_j X_{tj} \right)^2$ в которую, говоря нестрого, разные слагаемые дают разный статистический вклад из-за различных дисперсий, что в конечном итоге и приводит к неэффективности МНК-оценки. "Взвешивая" каждое наблюдение с помощью коэффициента $1/\sigma_t$, мы устраняем такую неоднородность. Поэтому часто обобщенный метод наименьших квадратов для системы с гетероскедастичностью называют методом *взвешенных наименьших квадратов.* Можно непосредственно проверить, что применение метода взвешенных наименьших квадратов приводит к уменьшению дисперсий оценок по сравнению а обычным методом наименьших квадратов.

Коррекция на гетероскедастичность

Если числа σ_t, неизвестны (что, как правило, и бывает на практике), необходимо использовать доступный обобщенный метод наименьших квадратов, который требует оценивания дисперсий σ_t^2. Так как число этих параметров равно *n,* то без дополнительных ограничений на структуру матрицы $\mathbf{\Omega}$ нет надежды получить приемлемые оценки дисперсий. Ниже мы рассмотрим один класс моделей с гетероскедастичностью, где такие ограничения накладываются и благодаря этому удается построить удовлетворительные оценки матрицы $\mathbf{\Omega}$, а следовательно, используя доступный обобщенный метод наименьших квадратов, и оценку $\hat{\hat{\beta}}$.

Ошибка пропорциональна независимой переменной. В некоторых ситуациях априорно можно считать, что ошибка прямо пропорциональна одной из независимых переменных, например, X_k: $\sigma_t^2 = \sigma^2 X_{tk}^2$. Тогда, разделив *t*-е. уравнение на X_{tk}, *t=1, ..., n,* и вводя новые независимые переменные $X_{tj}^* = X_{tj} / X_{tk}, t = 1, ..., k$ и новую зависимую переменную $Y_t^* = Y_t / X_{tk}$, *t=1,...,n, j=1,...,k,* получим классическую регрессионную модель. МНК-оценки коэффициентов этой модели дают непосредственно оценки исходной модели. Следует только помнить, что если первый регрессор в **X** есть набор единиц, то оценки свободного члена и коэффициента при $X_{tk}^* = 1/X_{tk}$ в новой модели являются оценками соответственно коэффициента при X_{tk} и свободного члена в исходной модели.

Возникает естественный вопрос, при каких обстоятельствах можно пользоваться описанным выше методом. Существуют процедуры, позволяющие выявлять гетероскедастичность того или иного рода (тесты на гетероскедастичность) [4-6]. Здесь мы ограничимся лишь практическими рекомендациями. Если есть предположение о зависимости ошибок от одной из независимых переменных, то целесообразно расположить наблюдения в порядке возрастания значений этой переменной, а затем провести обычную регрессию и получить остатки. Если размах их колебаний тоже возрастает (это хорошо заметно при обычном визуальном исследовании), то это говорит в пользу исходного предположения. Тогда надо сделать описанное выше преобразование; вновь провести регрессию и исследовать остатки. Если теперь их колебание имеет неупорядоченный характер, то это может служить показателем того, что коррекция на гетероскедастичность прошла успешно. Естественно, следует сравнивать и другие параметры регрессии (значимость оценок, сумму квадратов отклонений и т.п.) и только тогда принимать окончательное решение, какая из моделей более приемлема.

6.2. Корреляция по времени

Авторегрессионный процесс первого порядка

При анализе временных рядов часто приходится учитывать статистическую зависимость наблюдений в разные моменты времени. Иными словами, для многих временных рядов предположение о некоррелированности ошибок не выполняется. В этом разделе мы рассмотрим наиболее простую модель, в которой ошибки образуют так называемый авторегрессионый процесс первого порядка (точное определение будет дано ниже). Как было показано ранее (п. 6.3), применение обычного метода наименьших квадратов к этой системе дает несмещенные и состоятельные оценки параметров, однако можно показать, что оценка дисперсии оказывается смещенной вниз, что может отрицательно сказаться при проверке гипотез о значимости коэффициентов. Образно говоря, МНК рисует более оптимистичную картину регрессии, чем есть на самом деле.

Как и раньше, рассмотрим модель

$$\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\varepsilon}\,, \qquad (6.8)$$

где *t-я* компонента вектора $\vec{Y}$ представляет значение зависимой переменной в момент времени *t*, $t = 1,...,n$, а $\vec{\varepsilon}$ – вектор случайных ошибок. Для удобства запишем подробнее уравнение для наблюдения в момент времени *t:*

$$Y_t = \beta_1 + \beta_2 X_{t2} + ... + \beta_k X_{tk} + \varepsilon_t \tag{6.9}$$

Один из наиболее простых способов учета коррелированности ошибок (в разные моменты времени) состоит в предположении, что случайная последовательность $\{\varepsilon_t, t = 1,...,n\}$ образует *авторегрессионый процесс первого порядка.* Это означает, что ошибки удовлетворяют рекуррентному соотношению

$$\varepsilon_t = \rho\varepsilon_{t-1} + \nu_t \tag{6.10}$$

где $\{\nu_t, t = 1,...,n\}$ — последовательность независимых нормально распределенных случайных величин с нулевым средним и постоянной дисперсией σ_ν^2, а ρ — некоторый параметр, называемый *коэффициентом авторегрессии* ($|\rho| < 1$). Строго говоря, для полного описания модели надо определить ε_0. Будем считать, что ε_0 — нормальная случайная величина с нулевым средним и дисперсией $\sigma_\varepsilon^2 = \sigma_\nu^2(1-\rho)^2$, не зависящая от $\{\nu_t,\ t = 1,...,n\}$. Из дальнейшего станет ясно, почему у ε_0 именно такие параметры. Взяв математическое ожидание от обеих частей (6.10), получим $M(\varepsilon_t) = \rho M(\varepsilon_{t-1})$, откуда следует, что $M(\varepsilon_t) = 0, t = 1,...,n$. Поскольку ε_{t-1} выражается через $\nu_1,...,\nu_{t-1}$ (см. (6.9)), то ε_{t-1}, и ν_t, независимы. Поэтому

$$M(\varepsilon_t^2) = M(\rho\varepsilon_{t-1} + \nu_t)^2 = \rho^2 M(\varepsilon_{t-1}^2) + M(\nu_t^2) = \rho^2 M(\varepsilon_{t-1}^2) + \sigma_\nu^2.$$

Легко проверяется, что если $M(\varepsilon_0^2) = \sigma_\nu^2/(1-\rho^2)$, то

$$\sigma_\varepsilon^2 = M(\varepsilon_t^2) = V(\varepsilon_t) = \sigma_\nu^2/(1-\rho^2).$$

Умножая (6.10) на ε_{t-1}, и вновь пользуясь независимостью ε_{t-1}, и ν_t, получим

$$M(\varepsilon_t, \varepsilon_{t-1}) = Cov(\varepsilon_t, \varepsilon_{t-1}) = \rho V(\varepsilon_{t-1}) = \rho\sigma_\varepsilon^2 \tag{6.11}$$

Аналогично $Cov(\varepsilon_t, \varepsilon_{t-2}) = \rho^2\sigma_\varepsilon^2$, и вообще

$$Cov(\varepsilon_t, \varepsilon_{t-m}) = \rho^m\sigma_\varepsilon^2. \tag{6.12}$$

Таким образом, последовательность $\{e_t\}$ образует стационарный случайный процесс. Именно этим обстоятельством диктовался выбор параметров начальной величины ε_0. На самом деле, с течением времени зависимость ε_t, от ε_0 быстро уменьшается, поэтому в большинстве книг по эконометрике проблему начальных условий для $\{\varepsilon_t\}$ просто не рассматривают, неявно подразумевая, что процесс (6.10) при любом начальном значении быстро сходится к стационарному. Отметим также, что условие $|\rho|<1$ является необходимым для стационарности.

Из (6.11) следует, что

$$\rho = Cov(\varepsilon_t, \varepsilon_{t-1})/\sigma_\varepsilon^2 = Cov(\varepsilon_t, \varepsilon_{t-1})/(V(\varepsilon_t)^{1/2} V(\varepsilon_{t-1})^{1/2}),$$

т.е. ρ есть в точности коэффициент корреляции между двумя соседними ошибками. Пользуясь (6.12), можно выписать ковариационную матрицу случайного вектора $\vec{\varepsilon}$:

$$\mathbf{\Omega} = \frac{\sigma_\nu^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \ldots & \rho^{n-1} \\ \rho & 1 & \rho & \ldots & \rho^{n-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{n-3} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \rho^{n-1} & \rho^{n-2} & \rho^{n-3} & \ldots & 1 \end{bmatrix}.$$

Оценивание в модели с авторегрессией

Проблему оценивания системы (6.8) рассмотрим отдельно для случая, когда коэффициент ρ известен, и отдельно — когда неизвестен.

1. *Значение ρ известно.* В этом случае для оценивания системы (6.5) можно применить обобщенный метод наименьших квадратов. В данном случае нетрудно найти матрицу **P**, для которой $\mathbf{P}^{\mathbf{T}}\mathbf{P} = \mathbf{\Omega}^{-1}$. Здесь весьма просто догадаться, какое линейное преобразование исходной системы (6.8) надо провести, чтобы получить классическую модель. Выпишем (6.9) для момента времени $t-1$ $(t > 0)$

$$Y_{t-1} = \beta_1 + \beta_2 X_{t-1,2} + \ldots + \beta_k X_{t-1,k} + \varepsilon_{t-1},$$

умножим обе части на ρ и вычтем почленно из (6.9). Тогда с учетом (6.10) получим

$$Y_t - \rho Y_{t-1} = \beta_1(1-\rho) + \beta_2(X_{t,2} - \rho X_{t-1,2}) + \ldots + \beta_k(X_{t,k} - \rho X_{t-1,k}) + \nu_t. \quad (6.13)$$

При $t = 1$ достаточно обе части уравнения (6.9) умножить на $\sqrt{1-\rho^2}$:

$$\sqrt{1-\rho^2}Y_1 = \sqrt{1-\rho^2}\beta_1 + \sqrt{1-\rho^2}X_{1,2}\beta_2 + \ldots + \sqrt{1-\rho^2}X_{1,k}\beta_k + \sqrt{1-\rho^2}\varepsilon_1 \quad (6.14)$$

В системе (6.13), (6.14) ошибки удовлетворяют условиям уже обычной регрессионной модели. Действительно, в (6.13) случайные величины $\{\nu_t, t=2,\ldots,n\}$ независимы и имеют постоянную дисперсию σ_ν^2, а в (6.14) ошибка $\sqrt{1-\rho^2}\varepsilon_1$ не зависит от $\{\nu_t,\ t=2,...,n\}$ и, согласно (6.11), также имеет дисперсию σ_ν^2.

На практике часто опускают преобразование (6.14), игнорируя тем самым первое наблюдение. С одной стороны, благодаря этому, преобразование исходной модели (6.8) становится единообразным. В частности, для получения оценки параметра β_1 достаточно оценку свободного члена в (6.13) разделить на $1-\rho$. *С* другой стороны, отбрасывание первого наблюдения может привести к потере важной информации, особенно в выборках небольшого размера.

2. *Значение ρ неизвестно***.** Ситуации, коща параметр авторегрессии *p* известен, встречаются крайне редко. Поэтому возникает необходимость в процедурах оценивания при неизвестном ρ. Как правило, они имеют итеративный характер. Опишем одну из наиболее употребительных. Мы не будем устанавливать сходимость этих процедур. Практика их применения показала, что они достаточно эффективны.

Процедура Кохрейна-Оркатта (Cochrane-Orcutt). Начальным шагом этой процедуры является применение обычного метода наименьших квадратов к исходной системе (6.8) и получение соответствующих остатков $\vec{e}=[e_1,...,e_n]^T$. Далее,

1) в качестве приближенного значения ρ берется его МНК-оценка r в регрессии $e_t=\rho e_{t-1}+\nu_t$;

2) проводится преобразование (6.13) (или (6.13), (6.14)) при $\rho=r$ и находятся МНК-оценки $\hat{\hat{\vec{\beta}}}$ вектора параметров $\vec{\beta}$*;*

3) строится новый вектор остатков $\vec{e}=\vec{Y}-\mathbf{X}\hat{\hat{\vec{\beta}}}$*;*

4) процедура повторяется, начиная с п. 1.

Процесс обычно заканчивается, когда очередное приближение ρ мало отличается от предыдущего. Иногда просто фиксируется количество итераций. Процедура Кохрейна-Оркатта реализована в большинстве эконометрических компьютерных программ. При ее использовании может случиться, что значение параметра ρ будет найдено неточно. Это связано с тем, что при его оценивании может быть фактически найден локальный, а не глобальный минимум квадратов отклонений в регрессии п. 1.

Список рекомендуемой литературы

1. Gujarati D. Basic Econometrics. New York. 1988.
2. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрии. М., 1998.
3. *Грубер Й.* Эконометрия. Т1. — К.: “Астарта”, 1996 г.
4. *Джонстон Дж.* Эконометрические методы. — М. Статистика, 1980.
5. *Доугерти К.* Введение в эконометрику. — М.: ИНФРА-М, 1997.
6. *Дюк В.* Обработка данных на ПК в примерах. — СПб: Питер, 1997.
7. *Задорожный Г.В., Иващенко П.А.* Эконометрика. Ч1. — Х.: ХИБМ, 1996.
8. *Задорожный Г.В., Иващенко П.А.* Эконометрика. Ч2. — Х.: ХИБМ, 1996.
9. *Задорожный Г.В., Иващенко П.А., Тютюнникова С.В.* Экономическая безопасность и теневая экономика. – Х. – 1999.
10. *Иващенко П.А.* Адаптация в экономике. -Х.: Вища шк. Изд-во при Харьк. ун-те, 1986.
11. *Іващенко П.О., Семеняк І.В., Іванов В.В.* Багатовимірні статистичні методи. — Х.: Вид-во "Основа" при Харк. ун-ті. 1992.
12. *Катышев П.К., Пересецкий А.А.* Сборник задач к начальному курсу эконометрики. М., 1999. – 72 с.
13. *Кулинич О.І.* Економетрія. -Хм.: "Поділля", 1997.
14. *Лук'яненко І. Г., Краснікова Л. І.* Економетрика: Підручник. — К.: Тов. «Знання» КОО, 1998.
15. *Лук'яненко І. Г., Краснікова Л. І.* Економетрика: Практикум з використанням комп'ютера.—К.: Тов. «Знания» КОО, 1998.
16. *Магнус Я.Р., Катышев П.К., Пересецкий А.А.* Эконометрика. Начальный курс. - М., 1997.
17. *Наконечний С. І.* Економетрія: Навчальний посібник. К., 1997. –352 с.
18. *Пирогов Г., Федоровский Ю.* Проблемы структурного оценивания в эконометрии.—М., 1979.
19. *Толбатов Ю.А.* Економетрика. К., 1997.
20. *Чупров А.А.* Основные проблемы теории корреляции. — М.: Госстатиздат, 1960.

ПРИЛОЖЕНИЕ

П1. Краткая справка по функции EXCEL ЛИНЕЙН()

Рассчитывает статистику для ряда с применением метода наименьших квадратов, чтобы вычислить прямую линию, которая наилучшим образом аппроксимирует имеющиеся данные. Функция возвращает массив, который описывает полученную прямую. Поскольку возвращается массив значений, функция должна задаваться в виде формулы массива. Для получения дополнительных сведений о формулах массива нажмите кнопку.

Уравнение для прямой линии имеет следующий вид:

$$y = mx + b$$

или

$$y = m_1x_1 + m_2x_2 + ... + b$$

(в случае нескольких диапазонов значений *x*), где зависимое значение *y* является функцией независимого значения *x*. Значения *m* — это коэффициенты, соответствующие каждой независимой переменной *x*, а *b* — это постоянная. Заметим, что y, x и m могут быть векторами. Функция ЛИНЕЙН возвращает массив $\{m_n;m_{n-1};...;m_1;b\}$. ЛИНЕЙН может также возвращать дополнительную регрессионную статистику.

Синтаксис

ЛИНЕЙН(*известные_значения_y; известные_значения_x; конст; статистика*)

Известные_значения_*y* — это множество значений *y*, которые уже известны в соотношении $y = mx + b$.

Если массив *известные_значения_y* имеет один столбец, то каждый столбец массива *известные_значения_x* интерпретируется как отдельная переменная.

Если массив *известные_значения_y* имеет одну строку, то каждая строка массива *известные_значения_x* интерпретируется как отдельная переменная.

Известные_значения_x — это необязательное множество значений *x*, которые уже известны в соотношении $y = mx + b$.

Массив *известные_значения_x* может содержать одно или несколько множеств переменных. Если используется только одна переменная, то *известные_значения_y* и *известные_значения_x* могут иметь любую форму, при условии, что они имеют одинаковую размерность. Если используется более одной переменной, то *известные_значения_y* должны быть вектором (то есть диапазоном высотой в одну строку или шириной в один столбец).

Если *известные_значения_x* опущены, то предполагается, что это массив {1;2;3;...} такого же размера, как и *известные_значения_y*.

Конст — это логическое значение, которое указывает, требуется ли, чтобы константа *b* была равна 0.

Если аргумент *конст* имеет значение ИСТИНА или опущено, то *b* вычисляется обычным образом.

Если аргумент *конст* имеет значение ЛОЖЬ, то *b* полагается равным 0 и значения *m* подбираются так, чтобы выполнялось соотношение $y = mx$.

Статистика — это логическое значение, которое указывает, требуется ли вернуть дополнительную статистику по регрессии.

Если аргумент *статистика* имеет значение ИСТИНА, то функция ЛИНЕЙН возвращает дополнительную регрессионную статистику, так что возвращаемый массив будет иметь вид:

$\{m_n; m_{n-1};...; m_1; b: se_n; se_{n-1};...; se_1; se_b: R^2; se_y: F ;df: ss_{reg};ss_{resid}\}$.

Если аргумент статистика имеет значение ЛОЖЬ или опущена, то функция ЛИНЕЙН возвращает только коэффициенты *m* и постоянную *b*.

Дополнительная регрессионая статистика:

Величина	Описание
$se_n; se_{n-1};...; se_1$	Стандартные значения ошибок для коэффициентов $m_1,m_2,...,m_n$.
se_b	Стандартное значение ошибки для постоянной *b* (seb = #Н/Д, если *конст* имеет значение ЛОЖЬ).
R^2	Коэффициент детерминации. Сравниваются фактические значения *y* и значения, получаемые из уравнения прямой; по ре-

зультатам сравнения вычисляется коэффициент детерминации, нормированный от 0 до 1. Если он равен 1, то имеет место полная корреляция с моделью, т.е. нет различия между фактическим и оценочным значениями *y*. В противоположном случае, если коэффициент детерминации равен 0, то уравнение регрессии неудачно для предсказания значений у. Для получения информации о том, как вычисляется r2, см. "Замечания" в конце данного раздела.

se_y — Стандартная ошибка для оценки *y*.

F — *F*-статистика, или *F*-наблюдаемое значение. *F*-статистика используется для определения того, является ли наблюдаемая взаимосвязь между зависимой и независимой переменными случайной или нет.

df — Степени свободы. Степени свободы полезны для нахождения *F*-критических значений в статистической таблице. Для определения уровня надежности модели нужно сравнить значения в таблице с *F*-статистикой, возвращаемой функцией ЛИНЕЙН.

ss_{reg} — Регрессионая сумма квадратов.

ss_{resid} — Остаточная сумма квадратов.

На приведенном ниже рисунке показано, в каком порядке возвращается дополнительная регрессионная статистика.

	A	B	C	D	E	F
1	m_n	m_{n-1}	…	m_2	m_1	b
2	se_n	se_{n-1}	…	se_2	se_1	se_b
3	R^2	se_v				
4	F	df				
5	$ss_{рег.}$	$ss_{ост.}$				

ОГЛАВЛЕНИЕ