Арженовский С.В., Федосова О.Н.

ЭКОНОМЕТРИКА

Учебное пособие

Ростов-на-Дону
2002

УДК 330.43(075.8)
А80

Арженовский С.В., Федосова О.Н. Эконометрика:Учебное пособие/Рост. гос. экон. унив. – Ростов н/Д., – 2002. - 102 с. - **ISBN 5-7972-0495-9.**

В учебном пособии кратко изложено основное содержание лекционного курса эконометрики. Особое внимание уделено иллюстрации основных теоретических положений примерами из практики эконометрического моделирования.

Для студентов, обучающихся по специальностям экономического направления.

Рецензенты:

Л.И.Ниворожкина, д.э.н., профессор, зав. кафедрой СМиП РГЭУ "РИНХ"

Т.В.Алексейчик, к.э.н., доцент кафедры ФиПМ РГЭУ "РИНХ"

Утверждено в качестве учебного пособия редакционно-издательским советом РГЭУ "РИНХ"

ISBN 5-7972-0495-9

Оглавление

Введение

В последнее время специалисты, обладающие знаниями и навыками проведения прикладного экономического анализа с использованием доступных математических и программных средств, пользуются спросом на рынке труда. Одной из центральных дисциплин в подготовке таких специалистов является дисциплина "Эконометрика".

Эконометрика является областью знаний, которая охватывает вопросы применения статистических методов к теоретическим моделям, описывающим реальные экономические процессы.

Очевидно, что с помощью моделей можно получить много информации об экономических процессах, объяснить те или иные явления или процессы, но никогда не удастся получить всю информацию и однозначно определить истинный механизм экономического процесса или явления.

И даже в тех случаях, когда достаточно адекватная исходным данным эконометрическая модель построена и вопрос только в использовании ее для объяснения экономической ситуации или принятия решения, следует весьма осторожно подходить к выводам и рекомендациям, следующим из модельных оценок.

Эконометрический анализ, как правило, проводят с помощью ПЭВМ. В последние несколько лет сформировался обширный набор из пакетов прикладных программ, позволяющих автоматизировать процессы такого анализа. К наиболее распространенным относятся пакеты SAS, SPSS, Stata, Eviews и др. Имеются простейшие опции для проведения эконометрического анализа в Excel.

В настоящем пособии даются основные понятия, модели и методы эконометрики, рассматриваются примеры.

Содержание пособия полностью соответствует требованиям государственного стандарта высшего профессионального образования за исключением темы "Системы одновременных уравнений".

Для работы с предлагаемым изданием необходимы базовые знания некоторых разделов следующих учебных дисциплин: высшая математика, теория вероятностей, математическая статистика, общая теория статистики.

Эффективным является использование данной книги в сочетании с самостоятельным разбором примеров с использованием доступного статистического программного обеспечения.

Авторы благодарят рецензентов за советы при подготовке учебного пособия.

© Arzhenovsky S.V. , Fedosova O.N.. 2002. All rights reserved. Copies may be made for personal use. Version of 14.03.02.

1. Предмет и задачи дисциплины "Эконометрика"

1.1. Определение эконометрики

Сложность экономических процессов и необходимость их количественного измерения не позволяют современному экономисту ограничиваться в своей работе применением инструментов отдельных экономических дисциплин. Так, например, невозможно сделать прогноз о том, будет ли пользоваться спросом новый продукт (сорт кофе), если рассматривать этот процесс только с точки зрения экономической теории, то есть закона спроса и предложения. На практике для осуществления прогноза экономисту необходимо применить целый комплекс экономических наук, синтез которых и является сутью научной дисциплины - эконометрики.

Основной **целью** эконометрики является модельное описание конкретных количественных взаимосвязей, обусловленных общими качественными закономерностями, изученными в экономической теории.

Эконометрика – относительно молодая научная дисциплина, сформировавшаяся во второй половине XX века и развивающаяся на стыке экономической теории, статистики и математики (см. рис. 1.1).

Рис. 1.1. Эконометрика и ее место в ряду других экономических и статистических дисциплин

Впервые термин эконометрика был введен норвежским ученым Рагнаром Фришем в 1926 году и в буквальном переводе означает «измерение в экономике». Однако на сегодняшний день эта трактовка чересчур широка. Более четко определение эконометрики предложено известным российским ученым, профессором С.А. Айвазяном.

Эконометрика - это самостоятельная научная дисциплина, объединяющая совокупность теоретических результатов, приемов, методов и моделей, предназначенных для того, чтобы на базе
- экономической теории,
- экономической статистики,
- математико-статистического инструментария

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

Таким образом, суть эконометрики состоит в **синтезе** экономической теории, экономической статистики и математико-статистического инструментария.

1.2. Взаимосвязь эконометрики с экономической теорией, статистикой и экономико-математическими методами

Эконометрика не только выявляет объективно существующие экономические законы и связи между экономическими показателями, качественно определенными в экономической теории, но и формирует подходы к их формализации и количественному выражению. Так, к примеру, экономическая теория гласит, что повышение цены на товар, при прочих равных условиях, приводит к падению спроса на него. Однако экономическая теория не может дать ответ на вопрос о величине снижения спроса на конкретный товар в конкретных условиях. Решить эту задачу можно только с помощью эконометрики, которая, таким образом, вносит эмпирическое содержание в экономическую теорию.

В рамках экономического анализа, как правило, выдвигаются какие-либо гипотезы, строятся теории, объясняющие явление или процесс. Узкое место заключается в подтверждении теоретических гипотез фактическими данными. Поэтому в количественном экономическом анализе главную роль играет формирование гипотезы и ее проверка. Интуитивные утверждения должны приобрести форму предположений, которые могут быть либо приняты, либо отвергнуты после сопоставления с наблюдаемыми фактами.

Вопросами применения статистических методов к теоретическим моделям, описывающим реальные хозяйственные процессы, и занимается эконометрика.

Экономическая статистика как элемент информационного обеспечения эконометрики предполагает решение таких задач, как выбор необходимых статистических показателей и обоснование способа их измерения, определение плана статистического обследования и т.д.

Под математико-статистическим инструментарием в эконометрике подразумеваются отдельные расширенные разделы математической статистики, связанные с регрессионным анализом (классическая модель регрессии и классический метод наименьших квадратов, обобщенная модель регрессии и обобщенный метод наименьших квадратов), построением и анализом моделей временных рядов и систем одновременных уравнений.

Вместе с тем, необходимо различать эконометрику и математическую экономику. Именно приземление экономической теории на базу конкретной экономической статистики и извлечение из этого приземления с помощью подходящего математического аппарата вполне определенных количественных взаимосвязей являются ключевыми моментами в понимании сущности эконометрики, разграничении её с математической экономикой, описательной экономической статистикой и математической статистикой.

Так, математическая экономика – это математически сформулированная экономическая теория, которая изучает взаимосвязи между экономическими переменными на абстрактном (неколичественном) уровне. Она становится эконометрикой, когда символически представленные в этих взаимосвязях коэффициенты заменяются конкретными численными оценками, полученными на базе соответствующих экономических данных.

1.3. Области применения эконометрических моделей

Области применения эконометрических моделей напрямую связаны с целями эконометрического моделирования, основными из которых являются:

1) прогноз экономических и социально-экономических показателей, характеризующих состояние и развитие анализируемой системы;

2) имитация различных возможных сценариев социально-экономического развития анализируемой системы.

В качестве анализируемой экономической системы могут выступать страна в целом (макроэкономические системы), регионы, отрасли и корпорации (мезосистемы), а также предприятия, фирмы и домохозяйства (микроэкономические системы).

Кроме того, исследователь должен сформулировать профиль эконометрического моделирования, которое может быть сконцентрировано на проблемах финансового рынка, инвестиционных и социальных проблемах, или же на

целом комплексе проблем одновременно. Понятно, что, чем конкретнее сформулирован профиль исследования, тем более эффективны его результаты.

Например, исследователь изучает проблемы доходов домохозяйств страны. Целесообразнее было бы разделить эту большую задачу на исследование доходов городских и сельских домохозяйств, так как механизм их формирования существенно различен. Эконометрические модели, построенные отдельно для городских и сельских домохозяйств, будут гораздо более адекватны действительности, чем общая модель.

1.4. Методологические вопросы построения эконометрических моделей

В любой эконометрической модели, в зависимости от конечных прикладных целей ее использования все участвующие в ней переменные подразделяются на:

- экзогенные переменные, задаваемые как бы извне, автономно, в определенной степени управляемые (планируемые);
- эндогенные переменные, значения которых формируются в процессе и внутри функционирования анализируемой социально-экономической системы под воздействием экзогенных переменных и во взаимодействии друг с другом, являются предметом объяснения в эконометрической модели;
- предопределенные переменные выступают в роли факторов-аргументов или объясняющих переменных;
- лаговые эндогенные переменные входят в уравнения анализируемой эконометрической системы, но измерены в прошлые моменты, а следовательно, являются уже известными, заданными.

Эконометрическая модель служит для объяснения поведения эндогенных переменных в зависимости от значений экзогенных и лаговых эндогенных переменных.

Весь процесс эконометрического моделирования можно разбить на шесть основных этапов.

1-й этап (постановочный) – определение конечных целей моделирования, набора участвующих в модели факторов и показателей, их роли;

2-й этап (априорный) – предмодельный анализ экономической сущности изучаемого явления, формирование и формализация априорной информации и исходных допущений, в частности относящейся к природе и генезису исходных статистических данных и случайных остаточных составляющих в виде ряда гипотез;

3-й этап (параметризация) – собственно моделирование, т.е. выбор общего вида модели, в том числе состава и формы входящих в неё связей между переменными;

4-й этап (информационный) – сбор необходимой статистической информации, т.е. регистрация значений участвующих в модели факторов и показателей;

5-й этап (идентификация модели) – статистический анализ модели и в первую очередь статистическое оценивание неизвестных параметров модели. Непосредственно связан с проблемой *идентифицируемости* модели, то есть ответа на вопрос «Возможно ли в принципе однозначно восстановить значения неизвестных параметров модели по имеющимся исходным данным в соответствии с решением, принятым на этапе параметризации?». После положительного ответа на этот вопрос необходимо решить проблему идентификации модели, то есть предложить и реализовать математически корректную процедуру оценивания неизвестных параметров модели по имеющимся исходным данным;

6-й этап (верификация модели) – сопоставление реальных и модельных данных, проверка адекватности модели, оценка точности модельных данных. В ходе верификации модели решаются вопросы о том:

- насколько удачно удалось решить проблемы спецификации, идентифицируемости и идентификации, т.е. можно ли рассчитывать на то, что использование полученной модели в целях прогноза даст результаты, адекватные действительности;
- какова точность (абсолютная, относительная) прогнозных и имитационных расчетов основанных на построенной модели;

Получение ответов на эти вопросы с помощью тех или иных математико-статистических методов и составляет содержание верификации модели.

Проблема *спецификации* модели решается на 1, 2, 3 этапах моделирования и включает в себя:

- определение конечных целей моделирования (прогноз, имитация сценариев развития анализируемой системы, управление);
- определение списка экзогенных и эндогенных переменных;
- определение состава анализируемой системы уравнений и тождеств и соответственно списка предопределенных переменных;
- формулировка исходных предпосылок и априорных ограничений относительно стохастической природы остатков (рассмотрение проблемы гомоскедастичности).

Этапы 4, 5 и 6 сопровождаются процедурой *калибровки* модели, которая заключается в переборе большого числа вариантов, обусловленных наличием

«нормативных» ограничений, определенных содержательным смыслом анализируемых связей и определенной нечеткостью (неполнотой) статистической информации. Калибровка модели - трудоемкая процедура, что связано с многократными «вычислительными прогонами» модели.

Наиболее распространенными в эконометрическом моделировании являются следующие образующие четыре группы методы:

- классическая линейная модель множественной регрессии (КЛММР) и классический метод наименьших квадратов (МНК);
- обобщенная КЛММР и обобщенный МНК;
- методы статистического анализа временных рядов;
- методы анализа систем одновременных эконометрических уравнений.

Применение этих методов делает возможным построение следующих типов эконометрических моделей:

1. Регрессионные модели с одним уравнением.

В таких моделях зависимая (объясняемая) переменная y представляется в виде функции

$$y = f(x,\beta) = f(x_1,...,x_k,\beta_1,...,\beta_k),$$

где $x_1, x_2, ..., x_k$ - независимые (объясняющие) переменные,

$\beta_1, ..., \beta_k$ - параметры.

В зависимости от вида функции $f(x,\beta)$ модели делятся на линейные и нелинейные.

Например, можно исследовать уровень дохода семьи как функцию от ряда ее экономических и социально-демографических характеристик (наличие и количество работников в семье, наличие и количество детей и прочих иждивенцев, уровень образования и квалификации главы семьи и т.д.).

2. Модели временных рядов.

К этому классу относятся модели:

- *тренда:* $y(t) = T(t) + \xi_t$,

где t – время,

$T(t)$ - временной тренд заданного параметрического вида (например, линейный $T(t) = a + bt$),

ξ - случайная (стохастическая) компонента;

- *сезонности:* $y(t) = S(t) + \xi_t$,

где $S(t)$ - периодическая (сезонная) компонента,

ξ_t - случайная (стохастическая) компонента.

- *тренда и сезонности*: $y(t) = T(t) + S(t) + \xi_t$ (аддитивная) или

$y(t) = T(t)S(t) + \xi_t$ (мультипликативная)

где $T(t)$ - временной тренд заданного параметрического вида,

$S(t)$ - периодическая (сезонная) компонента,

ξ_t - случайная (стохастическая) компонента.

Кроме того, существуют модели временных рядов, в которых присутствует циклическая компонента, формирующая изменения анализируемого признака, обусловленные действием долговременных циклов экономической, демографической или астрофизической природы (волны Кондратьева, циклы солнечной активности и т.д.).

Модели временных рядов могут применяться для изучения и прогнозирования объема продаж туристических путевок, спроса на железнодорожные и авиабилеты, при краткосрочном прогнозировании процентных ставок и т.д.

3. Системы одновременных уравнений.

Эти модели описываются системами уравнений. Системы могут состоять из тождеств и регрессионных уравнений, каждое из которых, кроме объясняющих переменных, может включать в себя объясняемые переменные из других уравнений системы. Системы одновременных уравнений требуют сложного математического аппарата и могут быть использованы для моделей национальной экономики.

Ярким примером системы одновременных уравнений служит модель спроса и предложения. Пусть Q_t^D - спрос на товар в момент времени t, Q_t^S - предложение товара в момент времени t, P_t - цена на товар в момент времени t, Y_t – доход в момент t.

Составим систему уравнений "спрос – предложение":

$$Q_t^S = \alpha_1 + \alpha_2 P_t + \alpha_3 P_{t-1} + \xi_t \quad \text{(предложение)},$$

$$Q_t^D = \beta_1 + \beta_2 P_1 + \beta_3 Y_t + u_t \quad \text{(спрос)},$$

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

Цена товара P_t и спрос на товар $Q_t = Q_t^D = Q_t^S$ определяются из уравнений модели, то есть являются эндогенными переменными. Объясняющими переменными в данной модели являются доход Y_t и значение цены товара в предыдущий момент времени P_{t-1}.

Для эконометрического моделирования используются данные следующих трех типов.

1. Предположим, что мы располагаем результатами регистрации значений переменных $\left(x^1, x^2, ..., x^p\right)$ на n статистически обследованных объектах. Так что если i – номер обследованного объекта, то имеющиеся исходные статистические данные состоят из n строк вида $\left(x_i^1, x_i^2, ..., x_i^p\right)$, $i = \overline{1,n}$, где x_i^j - значение j

переменной, зарегистрированное на i обследованном объекте. То есть данные могут быть представлены в виде матрицы $n \times p$:

$$X = \begin{pmatrix} x_1^1 & x_1^2 & ... & x_1^p \\ x_2^1 & x_2^2 & ... & x_2^p \\ ... & ... & ... & ... \\ x_n^1 & x_n^2 & ... & x_n^p \end{pmatrix}.$$

Такой тип данных называется пространственной выборкой или данными поперечного среза (cross-section data). Такие данные не имеют временного параметра, и порядок их следования не существенен. Пример: финансовые показатели работы предприятий за истекший год.

2. Предположим, что данные регистрируются на одном и том же объекте, но в разные периоды времени. Тогда аналогом i будет номер периода времени, к которому привязаны соответствующие данные, а n будет общим числом периодов времени. Такие данные называются временно́й выборкой, или временными рядами данных (time series data), или данными продольного среза. Для таких данных существенен порядок следования значений переменных. Пример: финансовые показатели предприятия за последние несколько лет.

3. Наконец, предположим, что отслеживается каждый из n объектов в течение T периодов времени. То есть имеем последовательность матриц вида X, отнесенных к моментам времени $1,2,\ldots,T$:

$$X(t) = \begin{pmatrix} x_1^1(t) & x_1^2(t) & ... & x_1^p(t) \\ x_2^1(t) & x_2^2(t) & ... & x_2^p(t) \\ ... & ... & ... & ... \\ x_n^1(t) & x_n^2(t) & ... & x_n^p(t) \end{pmatrix}.$$

Такие данные называются панельными, или пространственно-временной выборкой (panel data). Данные сочетают в себе свойства как временных рядов, так и данных поперечного сечения. Как правило, значение T мало. Пример: показатели социально-экономического состояния домохозяйств за три года.

2. Парная регрессия

2.1. Основные цели и задачи прикладного корреляционно-регрессионного анализа

Рассмотрим некоторый экономический объект (процесс, явление, систему) и выделим только две переменные, характеризующие объект. Обозначим

переменные буквами Y и X. Будем предполагать, что независимая (объясняющая) переменная X оказывает воздействие на значения переменной Y, которая, таким образом, является зависимой переменной, т.е. имеет место зависимость:

$$Y=f(X). \qquad (2.1)$$

Зависимость (2.1) можно рассматривать с целью установления самого факта наличия или отсутствия значимой связи между Y и X, можно преследовать цель прогнозирования неизвестных значений Y по известным значениям X, наконец возможно выявление причинно-следственных связей между X и Y.

При изучении взаимосвязи между переменными Y и X следует, прежде всего, установить тип зависимости (природу анализируемых переменных Y и X). Возможны следующие ситуации:

- Y и X являются неслучайными переменными, т.е. значения Y строго зависят только от соответствующих значений X и полностью ими определяются. В этом случае говорят о функциональной зависимости, когда Y является некоторой функцией от переменной X и верна модель (2.1). Пример: $y=\sqrt{x}$.
- Y является случайной переменной, а X – неслучайной. В этом случае считают, что между переменными имеет место регрессионная зависимость. То есть верна модель $Y=f(X)+u$, где u – величина случайной ошибки.
- Y и X зависят от множества неконтролируемых факторов, так что являются случайными по своей сущности. В этом случае к проблемам построения конкретного вида зависимости между указанными переменными присоединяется проблема исследования тесноты связи между этими переменными. Речь в этом случае идет о корреляционно-регрессионной зависимости между Y и X.

Будем предполагать наличие второй из указанных ситуаций. Регрессионный анализ является инструментом решения следующих основных задач:

1. Для любых значений объясняющей переменной X построить наилучшие по некоторому критерию оценки для неизвестной функции $f(X)$.

2. По заданным значениям объясняющей переменной X построить наилучший по некоторому критерию прогноз для неизвестного значения результирующей переменной $Y(X)$.

3. Пусть известно, что искомая функция зависит от параметра θ: $f(X, \theta)$. Требуется построить наилучшую в определенном смысле оценку для неизвестного значения этого параметра.

4. Оценить удельный вес влияния переменной X на результирующий показатель Y.

В следующих разделах параграфа рассмотрим процедуру решения этих задач.

2.2. Постановка задачи регрессии

Поставим задачу регрессии Y на X.

Пусть мы располагаем n парами выборочных наблюдений над двумя переменными X и Y:

$$X_1, \quad X_2, \quad \ldots \quad X_n;$$
$$Y_1, \quad Y_2, \quad \ldots \quad Y_n.$$

Функция $f(X)$ называется функцией регрессии Y по X, если она описывает изменение условного среднего значения результирующей переменной Y в зависимости от изменения значений объясняющей переменной X: $f(X)=\mathrm{E}(Y|X)$.

Таким образом, имеет место уравнение регрессионной связи между Y и X:

$$Y_i=f(X_i)+u_i,\ i=1,\ldots,n. \qquad (2.2)$$

Присутствие в модели (2.2) случайной "остаточной" компоненты u, также называемой случайным членом, обусловлено следующими причинами:

1. Ошибки спецификации. Среди них выделяют невключение важных объясняющих переменных, агрегирование (объединение) переменных, неправильную функциональную спецификацию модели.

2. Ошибки измерения. Связаны со сложностью сбора исходных данных и использованием в модели аппроксимирующих переменных для учета факторов, непосредственное измерение которых невозможно.

3. Ошибки, связанные со случайностью человеческих реакций. Обусловлены тем, что поведение и непосредственное участие человека в ходе сбора и подготовки данных может быть достаточно непредсказуемым и вносит, таким образом, свой вклад в случайный член.

Мы хотим на основе выборочных наблюдений с учетом дополнительных требований, налагаемых на u, статистически оценить функцию $f(X)$, проверить оптимальность полученной оценки и использовать уравнение для построения прогноза.

Допущения модели. Относительно u необходимо принять ряд гипотез, известных как условия Гаусса-Маркова:

1. $\mathrm{E}u_i=0,\ i=1,\ldots,n.$

Это требование состоит в том, что математическое ожидание случайного члена в любом наблюдении должно быть равно нулю. Иногда случайный член будет положительным, иногда отрицательным, но он не должен иметь систематического смещения ни в одном из двух возможных направлений. Свойство непосредственно вытекает из смысла функции регрессии. Возьмем в (2.2) матожидание от обеих частей при фиксированном значении X, получим: $\mathrm{E}(Y|X)=\mathrm{E}(f(X))+\mathrm{E}(u)$, по свойству матожидания $\Rightarrow$ $\mathrm{E}(Y|X)=f(X)+\mathrm{E}(u)$, а поскольку с

учетом определения функции регрессии должно быть $f(X)=E(Y|X)$, то необходимо $E(u)=0$.

2. $\mathbf{E}(u_i u_j)=\begin{cases}\sigma_u^2, \text{ при } i=j,\\ 0, \text{ при } i\neq j.\end{cases}$

Первая строчка означает требование постоянства дисперсии регрессионных остатков (независимость от того, при каких значениях объясняющей переменной производятся наблюдения i), которое называют гомоскедастичностью остатков. Вторая строчка предполагает отсутствие систематической связи между значениями случайного члена в любых двух наблюдениях, которые должны быть абсолютно независимы друг от друга.

3. $X_1, \ldots, X_n$ – неслучайные величины.

Таким образом, задача регрессии имеет вид:

$$Y_i = f(X_i)+u_i,\ i=1,\ldots,n.$$

а. $Eu_i=0,\ i=1,\ldots,n.$ (2.3)

б. $\mathbf{E}(u_i u_j)=\begin{cases}\sigma_u^2, \text{ при } i=j,\\ 0, \text{ при } i\neq j.\end{cases}$ (2.4)

в. $X_1, \ldots, X_n$ – неслучайные величины. (2.5)

При выборе вида функции f в (2.2) обычно руководствуются следующими рекомендациями:

- используется априорная информация о содержательной экономической сущности анализируемой зависимости – аналитический способ,
- предварительный анализ зависимости с помощью визуализации – графический способ,
- использование различных статистических приемов обработки исходных данных и экспериментальных расчетов.

2.3. Парная регрессия и метод наименьших квадратов

Будем предполагать в рамках модели (2.2) линейную зависимость между двумя переменными Y и X. Т.е. имеем модель парной регрессии в виде:

$Y_i = \alpha+\beta X_i+u_i,\ i=1,\ldots,n.$

а. $Eu_i=0,\ i=1,\ldots,n.$

б. $\mathbf{E}(u_i u_j)=\begin{cases}\sigma_u^2, \text{ при } i=j,\\ 0, \text{ при } i\neq j.\end{cases}$

в. $X_1, \ldots, X_n$ – неслучайные величины.

Предположим, что имеется выборка значений Y и X.

Обозначим арифметические средние (выборочные математические ожидания) для переменных X и Y:

$$\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i,\ \overline{Y}=\frac{1}{n}\sum_{i=1}^{n}Y_i.$$

Запишем уравнение оцениваемой линии в виде:

$$\hat{Y}=\hat{\alpha}+\hat{\beta}X, \quad (2.6)$$

где $\hat{\alpha}$ и $\hat{\beta}$ - оценки неизвестных параметров α и β, а $\hat{Y}$ - ордината этой линии.

Пусть (X_i, Y_i) одна из пар наблюдений. Тогда отклонение этой точки (см. рис. 2.1) от оцениваемой линии будет равно $e_i=Y_i-\hat{Y}_i$.

Принцип метода наименьших квадратов (МНК) заключается в выборе таких оценок $\hat{\alpha}$ и $\hat{\beta}$, для которых сумма квадратов отклонений для всех точек становится минимальной.

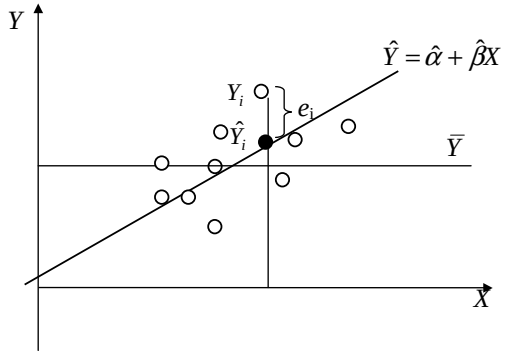

Рис. 2.1. Иллюстрация принципа МНК

Необходимым условием для этого служит обращение в нуль частных производных функционала:

$$\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\sum_{i=1}^{n}(Y_i-\hat{\alpha}-\hat{\beta}X_i)^2$$

по каждому из параметров. Имеем:

$$\frac{\partial}{\partial\hat{\alpha}}\left(\sum_{i=1}^{n}e_i^2\right)=-2\sum_i(Y_i-\hat{\alpha}-\hat{\beta}X_i)=0;$$

$$\frac{\partial}{\partial\hat{\beta}}\left(\sum_{i=1}^{n}e_i^2\right)=-2\sum_i X_i(Y_i-\hat{\alpha}-\hat{\beta}X_i)=0.$$

Упростив последние равенства, получим стандартную форму нормальных уравнений, решение которых дает искомые оценки параметров:

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\alpha} + \hat{\beta}\sum_{i=1}^{n} X_i; \\ \sum_{i=1}^{n} X_i Y_i = \hat{\alpha}\sum_{i=1}^{n} X_i + \hat{\beta}\sum_{i=1}^{n} X_i^2. \end{cases} \qquad (2.7)$$

Из (2.7) получаем:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}, \ \hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X}, \qquad (2.8)$$

где $x_i = X_i - \overline{X}$, $y_i = Y_i - \overline{Y}$.

Пример. Для иллюстрации вычислений при отыскании зависимости с помощью метода наименьших квадратов рассмотрим пример (табл. 2.1).

Таблица 2.1

Индивидуальное потребление и личные доходы (США, 1954-1965 гг.)

Год	Индивидуальное потребление, млрд. долл.	Личные доходы, млрд. долл.
1954	236	257
1955	254	275
1956	267	293
1957	281	309
1958	290	319
1959	311	337
1960	325	350
1961	335	364
1962	355	385
1963	375	405
1964	401	437
1965	431	469

Заметим, что исходные данные должны быть выражены величинами примерно одного порядка. Вычисления удобно организовать, как показано в таблице 2.2. Сначала рассчитываются $\overline{X}, \overline{Y}$, затем x_i, y_i. Результаты заносятся в столбцы 3 и 4. Далее определяются x_i^2, $x_i y_i$ и заносятся в 5 и 6 столбцы таблицы 2.2. По формулам (2.8) получим искомые значения параметров $\hat{\beta}$=43145/46510=0,9276; $\hat{\alpha}$=321,75-0,9276·350=-2,91.

Оцененное уравнение регрессии запишется в виде $\hat{Y}$=-2,91+0,9276X.

Следующая важная проблема состоит в том, чтобы определить, насколько "хороши" полученные оценки и уравнение регрессии. Этот вопрос рассматривается по следующим стадиям исследования: квалифицирование (выяснение условий применимости результатов), определение качества оценок, проверка выполнения допущений метода наименьших квадратов.

Относительно квалифицирования уравнения $\hat{Y}$=-2,91+0,9276X. Оно выражает, конечно, достаточно сильное утверждение. Применять это уравнение для прогнозирования следует очень осторожно. Дело в том, что, даже отвлекаясь от многих факторов, влияющих на потребление, и от систематического изменения дохода по мере варьирования потребления, мы не располагаем достаточно представительной выборкой.

Таблица 2.2

Рабочая таблица расчетов (по данным табл. 2.1)

Год	X	Y	x	y	x^2	xy	$\hat{Y}$	e_i
1954	257	236	-93	-85,75	8649	7974,75	235,48	0,52
1955	275	254	-75	-67,75	5625	5081,25	252,18	1,82
1956	293	267	-57	-54,75	3249	3120,75	268,88	-1,88
1957	309	281	-41	-40,75	1681	1670,75	283,72	-2,72
1958	319	290	-31	-31,75	961	984,25	292,99	-2,99
1959	337	311	-13	-10,75	169	139,75	309,69	1,31
1960	350	325	0	3,25	0	0	321,75	3,25
1961	364	335	14	13,25	196	185,5	334,74	0,26
1962	385	355	35	33,25	1225	1163,75	354,22	0,78
1963	405	375	55	53,25	3025	2928,75	372,77	2,23
1964	437	401	87	79,25	7569	6894,75	402,45	-1,45
1965	469	431	119	109,25	14161	13000,75	432,13	-1,13
Σ	$\overline{X}$ =350,00	$\overline{Y}$ =321,75	0	0,00	46510	43145	$\overline{Y}$ =321,75	0,00

Полученное уравнение $\hat{Y}$=-2,91+0,9276X можно использовать для расчета точечного прогноза, в том числе и на ретроспективу. Подставляя последовательно значения X из второго столбца табл. 2.2 в уравнение $\hat{Y}$=-2,91+0,9276X, получим предпоследний столбец табл. 2.2 для прогнозных значений $\hat{Y}$. Ошибка прогноза вычисляется по формуле $e_i = Y_i - \hat{Y}_i$ и дана в последнем столбце рабочей таблицы.

Заметим, что ошибка прогноза e_i фактически является оценкой значений u_i. График ошибки e_i представлен на рис. 2.2. Следует отметить факт равенства нулю суммы Σe_i=0, что согласуется с первым ограничением модели парной регрессии - Eu_i=0, i=1,…,n. ∇

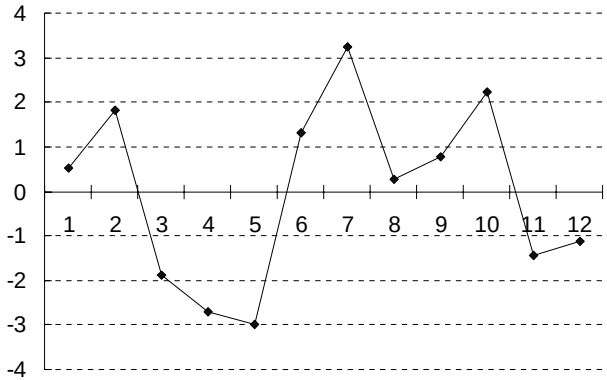

Рис. 2.2. График ошибки прогноза

В модели (2.2) функция f может быть и нелинейной. Причем выделяют два класса нелинейных регрессий:

- регрессии, нелинейные относительно включенной объясняющей переменной, но линейные по параметрам, например полиномы разных степеней - $Y_i = a_0 + a_1 X_i + a_2 X_i^2 + u_i$, $i=1,\ldots,n$ или гипербола - $Y_i = a_0 + a_1/X_i + u_i$, $i=1,\ldots,n$;
- регрессии нелинейные по оцениваемым параметрам, например степенная функция - $Y_i = a_0 X_i^{a_1} u_i$, $i=1,\ldots,n$, или показательная функция - $Y_i = a_0 a_1^{X_i} u_i$, $i=1,\ldots,n$.

В первом случае МНК применяется так же, как и в линейной регрессии, поскольку после замены, например, в квадратичной параболе $Y_i = a_0 + a_1 X_i + a_2 X_i^2 + u_i$ переменной X_i^2 на X_{1i}: $X_i^2 = X_{1i}$, получаем линейное уравнение регрессии $Y_i = a_0 + a_1 X_i + a_2 X_{1i} + u_i$, $i=1,\ldots,n$.

Во втором случае в зависимости от вида функции возможно применение линеаризующих преобразований, приводящих функцию к виду линейной. Например, для степенной функции $Y_i = a_0 X_i^{a_1} u_i$ после логарифмирования получаем $\ln Y_i = \ln a_0 + a_1 \ln X_i + \ln u_i$ линейную функцию в логарифмах и применяем МНК.

Однако для, например, модели $Y_i = a_0 + a_2 X_i^{a_1} + u_i$ линеаризующее преобразование отсутствует, и приходится применять другие способы оценивания (например, нелинейный МНК).

2.4. Коэффициент корреляции, коэффициент детерминации, корреляционное отношение

Для трактовки линейной связи между двумя переменными акцентируют внимание на коэффициенте корреляции.

Пусть имеется выборка наблюдений (X_i, Y_i), $i=1,\ldots,n$, которая представлена на диаграмме рассеяния, именуемой также полем корреляции (рис. 2.3).

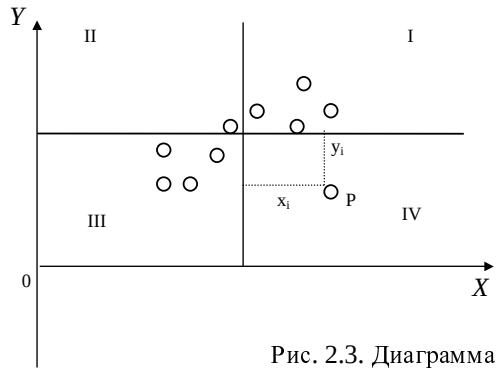

Рис. 2.3. Диаграмма рассеяния

Разобьем диаграмму на четыре квадранта так, что для любой точки $P(X_i, Y_i)$ будут определены отклонения $x_i = X_i - \overline{X}$, $y_i = Y_i - \overline{Y}$.

Ясно, что для всех точек I квадранта $x_i y_i > 0$; для всех точек II квадранта $x_i y_i < 0$; для всех точек III квадранта $x_i y_i > 0$; для всех точек IV квадранта $x_i y_i < 0$. Следовательно, величина $\sum x_i y_i$ может служить мерой зависимости между переменными X и Y. Если большая часть точек лежит в первом и третьем квадрантах, то $\sum x_i y_i > 0$ и зависимость положительная, если большая часть точек лежит во втором и четвертом квадрантах, то $\sum x_i y_i < 0$ и зависимость отрицательная. Наконец, если точки рассеиваются по всем четырем квадрантам $\sum x_i y_i$ близка к нулю и между X и Y связи нет.

Указанная мера зависимости изменяется при выборе единиц измерения переменных X и Y. Выразив $\sum x_i y_i$ в единицах среднеквадратических отклонений, получим после усреднения выборочный коэффициент корреляции:

$$r = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i^2}} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right]\left[n\sum_{i=1}^{n} Y_i^2 - \left(\sum_{i=1}^{n} Y_i\right)^2\right]}} \qquad (2.9)$$

Из последнего выражения можно после преобразований получить следующую формулу для квадрата коэффициента корреляции:

$$R^2 = \frac{\hat{\beta}^2 \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} y_i^2} \text{ или}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} y_i^2} \qquad (2.10)$$

Квадрат коэффициента корреляции называется коэффициентом детерминации. Согласно (2.10) значение коэффициента детерминации не может быть больше единицы, причем это максимальное значение будет достигнуто при $\sum_i e_i^2 = 0$, т.е. когда все точки диаграммы рассеяния лежат в точности на прямой. Следовательно, значения коэффициента корреляции лежат в числовом промежутке от -1 до +1.

Кроме того, из (2.10) следует, что коэффициент детерминации равен доле дисперсии Y (знаменатель формулы), объясненной линейной зависимостью от X (числитель формулы). Это обстоятельство позволяет использовать R^2 как обобщенную меру "качества" статистического подбора модели (2.6). Чем лучше регрессия соответствует наблюдениям, тем меньше $\sum_i e_i^2$ и тем ближе R^2 к 1, и наоборот, чем "хуже" подгонка линии регрессии к данным, тем ближе значение R^2 к 0.

Поскольку коэффициент корреляции симметричен относительно X и Y, то есть $r_{XY}=r_{YX}$, то можно говорить о корреляции как о мере взаимозависимости переменных. Однако из того, что значения этого коэффициента близки по модулю к единице, нельзя сделать ни один из следующих выводов: Y является причиной X; X является причиной Y; X и Y совместно зависят от какой-то третьей переменной. Величина r ничего не говорит о причинно-следственных связях. Эти вопросы должны решаться, исходя из содержательного анализа задачи. Следует избегать и так называемых ложных корреляций, т.е. нельзя пытаться связать явления, между которыми отсутствуют реальные причинно-следственные связи. Например, корреляция между успехами местной футбольной команды и индексом Доу-Джонса. Классическим является пример ложной корреляции, приведенный в начале XX века известным российским статистиком А.А. Чупровым: если в качестве независимой переменной взять число пожарных команд в городе, а в качестве зависимой переменной – сумму убытков от пожаров за год, то между ними есть прямая корреляционная зависимость, т.е. чем больше пожарных команд, тем больше сумма убытков. На самом деле здесь нет причинно-следственной связи, а есть лишь следствия общей причины – величины города.

Проверка гипотезы о значимости выборочного коэффициента корреляции эквивалентна проверке гипотезы о $\beta=0$ (см. ниже) и, следовательно, равносильна проверке основной гипотезы об отсутствии линейной связи между Y и X. Вычисляя значение t-статистики

$$t = r\sqrt{n-2}/\sqrt{1-r^2},$$

вывод о значимости r делается при $|t| > t_\varepsilon$, где t_ε - соответствующее табличное значение t-распределения с $(n-2)$ степенями свободы и уровнем значимости ε.

Пример. Вычислим коэффициент корреляции и проверим его значимость для нашего примера табл. 2.1.

По (2.9) $r=43145/(46510 \cdot 40068{,}25)^{0,5}=0{,}9994$. $R^2=0{,}998$. Значение t-статистики $t=0{,}9994 \cdot [10/(1-0{,}998)]^{0,5}=70{,}67$. Поскольку $t_{0,05;10}=2{,}228$, то $t>t_{0,05;10}$ и коэффициент корреляции значим. Следовательно, можно считать, что линейная связь между переменными Y и X в примере существует. ∇

Если между переменными имеет место нелинейная зависимость, то коэффициент корреляции теряет смысл как характеристика степени тесноты связи. В этом случае используется наряду с расчетом коэффициента детерминации расчет корреляционного отношения.

Предположим, что выборочные данные могут быть сгруппированы по оси объясняющей переменной X. Обозначим s – число интервалов группирования, n_j $(j=1,\ldots,s)$ – число выборочных точек, попавших в j-й интервал группирования, $\bar{Y}_j = \frac{1}{n_j}\sum_{k=1}^{n_j} Y_{ji}$ - среднее значение ординат точек, попавших в j-й интервал группирования, $\bar{\bar{Y}} = \frac{1}{n}\sum_{j=1}^{s} n_j \bar{Y}_j$ - общее среднее по выборке. С учетом формул для оценок выборочных дисперсий среднего значения Y внутри интервалов группирования $\sigma_{\bar{Y}}^2 = \frac{1}{n}\sum_{j=1}^{s} n_j \left(\bar{Y}_j - \bar{\bar{Y}}\right)^2$ и суммарной дисперсии результатов наблюдения $\sigma_Y^2 = \frac{1}{n}\sum_{j=1}^{s}\sum_{k=1}^{n_j} \left(Y_{ji} - \bar{\bar{Y}}\right)^2$ получим:

$$\hat{\rho}_{YX}^2 = \frac{\sigma_{\bar{Y}}^2}{\sigma_Y^2}. \qquad (2.11)$$

Величину $\hat{\rho}_{YX}$ в (2.11) называют корреляционным отношением зависимой переменной Y по независимой переменной X. Его вычисление не предполагает каких-либо допущений о виде функции регрессии.

Величина $\hat{\rho}_{YX}$ по определению неотрицательная и не превышает единицы, причем $\hat{\rho}_{YX}$=1 свидетельствует о наличии функциональной связи между переменными Y и X. Если указанные переменные не коррелированны друг с другом, то $\hat{\rho}_{YX}$=0.

Можно показать, что $\hat{\rho}_{YX}$ не может быть меньше величины коэффициента корреляции r (формула (2.9)) и в случае линейной связи эти величины совпадают.

Это позволяет использовать величину разности $\hat{\rho}^2_{YX} - R^2$ в качестве меры отклонения регрессионной зависимости от линейного вида.

2.5. Оценка статистической значимости регрессии

Перейдем к вопросу о том, как отличить "хорошие" оценки МНК от "плохих". Конечно, предполагается, что существуют критерии качества рассчитанной линии регрессии.

Перечислим способы, которые помогают решить вопрос о достоинствах рассчитанной линии регрессии:

- построение доверительных интервалов и оценка статистической значимости коэффициентов регрессии по t-критерию Стьюдента;
- дисперсионный анализ и F – критерий Фишера;
- проверка существенности выборочного коэффициента корреляции (детерминации).

Перейдем к подробному изложению свойств оценок МНК и способов проверки их значимости.

Несложно показать, что оценки $\hat{\alpha}$ и $\hat{\beta}$ полученные МНК по (2.8) с учетом ограничений (2.3)-(2.5) являются линейными несмещенными оценками и обладают наименьшими дисперсиями (являются эффективными) в классе линейных оценок (теорема Гаусса-Маркова).

Для вычисления интервальных оценок α, β предполагаем нормальное распределение случайной величины u. Для получения интервальных оценок α, β оценим дисперсию случайного члена σ^2_u по отклонениям e_i. В качестве оценки дисперсии ошибки σ^2_u возьмем величину:

$$\sigma_u^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}. \qquad (2.12)$$

Вычислим величину

$$V(\hat{\alpha}) = \frac{\sigma_u^2 \sum X_i^2}{n\sum\left(X_i - \bar{X}\right)^2},$$

и $\sqrt{V(\hat{\alpha})}$ - стандартную ошибку коэффициента регрессии α.

Статистика

$$t = \frac{\hat{\alpha} - \alpha}{\sqrt{V(\hat{\alpha})}},$$

имеет t-распределение Стьюдента. Так как $\hat{\alpha}$ несмещенная оценка, то для заданного 100(1–ε)% уровня значимости доверительный интервал для α суть:

$$\hat{\alpha} \pm t_{\varepsilon,n-2} \frac{\sigma_u \sqrt{\sum X_i^2}}{\sqrt{n\sum (X_i - \bar{X})^2}} \text{ или } \hat{\alpha} \pm t_{\varepsilon,n-2} \sqrt{\frac{\sum e_i^2 \sum X_i^2}{(n-2)n\sum (X_i - \bar{X})^2}}, \qquad (2.13)$$

где $t_{\varepsilon,n-2}$ – табличное значение t распределения для (n-2) степеней свободы и уровня значимости ε.

Вычислим величину

$$V(\hat{\beta}) = \frac{\sigma_u^2}{\sum\left(X_i - \bar{X}\right)^2},$$

и $\sqrt{V(\hat{\beta})}$ - стандартную ошибку[1] коэффициента регрессии β.

Статистика

$$t = \frac{\hat{\beta} - \beta}{\sqrt{V(\hat{\beta})}},$$

имеет t-распределение Стьюдента. Так как $\hat{\beta}$ несмещенная оценка, то для заданного 100(1–ε)% уровня значимости доверительный интервал для β суть:

$$\hat{\beta} \pm t_{\varepsilon,n-2} \frac{\sigma_u}{\sqrt{\sum (X_i - \bar{X})^2}} \text{ или } \hat{\beta} \pm t_{\varepsilon,n-2} \sqrt{\frac{\sum e_i^2}{(n-2)\sum (X_i - \bar{X})^2}}, \qquad (2.14)$$

где $t_{\varepsilon,n-2}$ – табличное значение t распределения для (n-2) степеней свободы и уровня значимости ε.

Проверим гипотезу о равенстве нулю коэффициента α, т.е.

H_0: α=0.

[1] Стандартная ошибка дает только общую оценку степени точности коэффициента регрессии. Ясно, что, чем больше будет величина дисперсии случайного члена (и соответственно ее оценка – выборочная дисперсия остатков), тем существеннее величина стандартной ошибки, и с большей *вероятностью* можно говорить о том, что полученная оценка неточна.

С учетом статистики $t=\dfrac{\hat{\alpha}-\alpha}{\sqrt{V(\hat{\alpha})}}$ для α=0, имея в виду формулу для $V(\hat{\alpha})$, получим:

$$t=\frac{\hat{\alpha}\sqrt{n\sum\left(X_i-\bar{X}\right)^2}}{\sigma_u\sqrt{\sum X_i^2}}. \qquad (2.15)$$

Если вычисленное по (2.15) значение t будет больше t_ε для заданного критического уровня значимости ε, то гипотеза H_0 о равенстве нулю коэффициента α отклоняется, если же $t<t_\varepsilon$, то H_0 принимается.

Аналогично для проверки гипотезы о равенстве нулю коэффициента β, т.е.

$$H_0: \beta=0$$

рассчитаем статистику:

$$t=\frac{\hat{\beta}\sqrt{\sum\left(X_i-\bar{X}\right)^2}}{\sigma_u}. \qquad (2.16)$$

Если вычисленное по (2.16) значение t будет больше t_ε для заданного критического уровня значимости ε, то гипотеза H_0 о равенстве нулю коэффициента β отклоняется, если же $t<t_\varepsilon$, то H_0 принимается.

Заметим, что формула (2.12) может быть упрощена и записана в виде:

$$\sigma_u^2=\frac{\sum Y^2-\hat{\alpha}\sum Y-\hat{\beta}\sum XY}{n-2}. \qquad (2.17)$$

Пример. Приведем расчеты для нашего примера в табл. 2.1. По формуле (2.17) рассчитаем дисперсию ошибки:

σ_u^2=(1282345–(–2,91) · 3861–0,9276 ·1394495)/10=4,6948 или σ_u=2,1667.

Найдем доверительный интервал для α по первой из формул (2.13):

$\alpha=-2{,}91\pm t_{0{,}05;10}\sqrt{1516510}\cdot 2{,}1667/\sqrt{12\cdot 46510}$.

По таблице t-распределения находим

$t_{0,05;10}$=2,228 и α=-2,91±2,228 ·2668,219/747,0743.

Откуда α=-2,91±7,798 или -10,7≤α≤4,9.

С вероятностью 0,95 истинные значения α находятся в интервале 10,7≤α≤4,9.

Аналогично найдем доверительный интервал для β по первой из формул (2.14): $\beta=0{,}9276\pm t_{0{,}05;10}\cdot 2{,}1667/\sqrt{46510}$ =0,9276±0,022 и 0,91≤β≤0,95.

Кроме того по экономическому смыслу переменных примера следует ожидать, что 0≤β≤1. Поскольку доверительный интервал не включает 0 и 1, то результаты регрессии соответствуют гипотезе 0≤β≤1.

Проверим гипотезу о равенстве нулю коэффициента β, т.е. H_0: β=0.

Рассчитаем t-статистику по формуле (2.16):

t=0,9276 · $\sqrt{46510}$/2,1667=92,328.

Табличное значение $t_{0,01;10}$=3,169, так как $t>t_{0,01;10}$, то гипотеза о том, что β=0 отклоняется. Можно говорить о том, что коэффициент β значимо отличен от нуля. ∇

Разложим общую вариацию значений Y около их выборочного среднего $\bar{Y}$ на составляющие (см. рис. 2.1):

$$\sum(Y-\bar{Y})^2=\sum(\hat{Y}-\bar{Y})^2+\sum(Y-\hat{Y})^2. \qquad (2.18)$$

Сумма квадратов отклонений от среднего в выборке равна сумме квадратов отклонений значений $\hat{Y}$, полученных по уравнению регрессии, от выборочного среднего $\bar{Y}$ плюс сумма квадратов отклонений Y от линии регрессии $\hat{Y}$.

Первую связывают с линейным воздействием изменений переменной X и называют "объясненной".

Вторая составляющая является остатком и называется "необъясненной" долей вариации переменной Y.

Отметим, что долю дисперсии, объясняемую регрессией, в общей дисперсии результативной переменной Y характеризует коэффициент детерминации, определяемый по формуле (2.10), которая может быть преобразована с учетом (2.18) к виду:

$$R^2=\frac{\sum(\hat{Y}-\bar{Y})^2}{\sum(Y-\bar{Y})^2}.$$

Предположим, что мы хотим проверить гипотезу об отсутствии линейной функциональной связи между X и Y, т.е. H_0: β=0.

Иначе говоря, мы хотим оценить значимость уравнения регрессии (2.6) в целом. Для проверки гипотезы сведем необходимые вычисления в таблицу (табл. 2.3).

Соотношение

$$F=\frac{Q_1}{Q_2/(n-2)}=\hat{\beta}^2\sum x_i^2\Big/\sigma_u^2 \qquad (2.19)$$

удовлетворяет F - распределению Фишера с (1, n-2) степенями свободы. Критические значения этой статистики F_ε для уровня значимости ε затабулированы.

Если $F>F_\varepsilon$, то гипотеза об отсутствии связи между переменными Y и X отклоняется, в противном случае гипотеза H_0 принимается и уравнение регрессии не значимо.

Таблица 2.3

Таблица дисперсионного анализа

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Среднее квадратов отклонений
X	$Q_1 = \hat{\beta}^2 \sum x_i^2$	1	$\hat{\beta}^2 \sum x_i^2$
Остаток	$Q_2 = (n-2)\sigma_u^2$	n-2	σ_u^2
Общая вариация	$\sum (Y - \overline{Y})^2 = Q_1 + Q_2$	n-1	-

Пример. Для примера табл. 2.1, с учетом предыдущих вычислений, будем иметь таблицу анализа дисперсии - табл. 2.4.

Применяя формулу (2.19), получим $F = \frac{\hat{\beta}\sum x_i^2}{\sigma_u^2} = \frac{0{,}9276^2 \cdot 46510}{4{,}6948} = 8514{,}7$.

Табличное значение $F_{0,01}(1, 10)=10{,}04$, так что имеющиеся данные позволяют отвергнуть гипотезу об отсутствии связи между личными доходами и индивидуальным потреблением. ∇

Таблица 2.4

Таблица анализа дисперсии (пример в табл. 2.1)

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Среднее квадратов отклонений
X	$0{,}9276^2$*46510	1	40019,1
Остаток	10*4,6948	10	4,7
Общая вариация	40066,0	11	-

2.6. Интерпретация уравнения регрессии

Проанализируем, какую информацию дает нам оцененное уравнение регрессии (2.6), т.е. поставим вопрос об интерпретации (содержательном объяснении) коэффициентов уравнения.

Во-первых, можно сказать, что увеличение X на одну единицу (в единицах измерения переменной X) приведет к увеличению/уменьшению (в зависимости от знака коэффициента $\hat{\beta}$) значения Y на $\hat{\beta}$ единиц (в единицах измерения переменной Y).

Во-вторых, необходимо проверить, в каких единицах измерены переменные X и Y и можно ли заменить слово "единица" фактическим количеством (рубли, тонны и т.п.).

В-третьих, константа $\hat{\alpha}$ дает прогнозируемое значение Y, если положить X=0. Это может иметь или не иметь экономического смысла в зависимости от конкретной ситуации.

Часто рассчитывают средний коэффициент эластичности $\overline{Э} = f'(X)\frac{\overline{X}}{\overline{Y}}$, который показывает, на сколько процентов в среднем по совокупности изменится результат Y от своей средней величины при изменении фактора X на 1% от своего среднего значения.

Пример. Продолжая рассмотрение примера п. 2.1, проинтерпретируем уравнение регрессии между индивидуальным потреблением и личными доходами в США: $\hat{Y}$=-2,91+0,9276X.

Поскольку обе переменные измерены в $, то интерпретация облегчается.

Смысл коэффициента $\hat{\beta}$: при увеличении личных доходов граждан США на 1$ расходы на индивидуальное потребление возрастут на 0,9$. Другими словами, из каждого дополнительного доллара дохода 90 центов будут израсходованы на потребление.

Константа в данном случае не имеет никакого смысла применительно к совокупности, поскольку мы не можем сказать, что при нулевых доходах потребление граждан США составит -2,91 млрд. долларов.

Рассчитаем средний коэффициент эластичности:

$$\overline{Э} = f'(X)\frac{\overline{X}}{\overline{Y}} = 0{,}9276 \cdot 350/351{,}75 = 0{,}923.$$

Т.е. при изменении личных доходов на 1% от своего среднего значения в среднем по совокупности индивидуальное потребление изменится на 0,923% от своей средней величины. ∇

При интерпретации уравнения регрессии важно помнить о следующих фактах:

- величины $\hat{\alpha}$ и $\hat{\beta}$ являются только оценками α и β, а следовательно, и вся интерпретация представляет собой тоже оценку;
- уравнение регрессии отражает общую тенденцию для выборки, а каждое отдельное наблюдение при этом подвержено воздействию случайностей;
- верность интерпретации зависит от правильности спецификации уравнения, то есть включения/исключения соответствующих объясняющих переменных и выбора вида функции регрессии.

3. Классическая линейная модель множественной регрессии

Рассмотрим обобщение линейной регрессионной модели для случая более двух переменных.

Всякий раз, когда изучаемый процесс или явление является результатом совместного действия нескольких факторов, у исследователя возникает потреб-

ность в оценке влияния каждого фактора в отдельности. Один из стандартных методов[2], позволяющий успешно решить эту задачу, суть множественная регрессия.

3.1. Предположения модели

Пусть мы располагаем выборочными наблюдениями над k переменными Y_i и X_{ji}, j=1,..., k, i=1,2,…,n, где n – количество наблюдений:

1	2	…	i	…	n
Y_1,	Y_2,	…	Y_i,	…	Y_n
X_{11},	X_{12},	…	X_{1i},	…	X_{1n}
…	…	…	…	…	…
X_{k1},	X_{k2},	…	X_{ki},	…	X_{kn}

Предположим, что существует линейное соотношение между результирующей переменной Y и k объясняющими переменными X_1, X_3, ..., X_k. Тогда с учетом случайной ошибки u_i запишем уравнение:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_i, \quad i = 1,2,...,n \tag{3.1}$$

В (3.1) неизвестны коэффициенты β_j, j=0,2,…,k и параметры распределения u_i. Задача состоит в оценивании этих неизвестных величин. Модель (3.1) называется классической линейной моделью множественной регрессии (КЛММР). Заметим, что часто имеют в виду, что переменная X_0 при β_0 равна единице для всех наблюдений i=1,2,…,n.

Относительно переменных модели в уравнении (3.1) примем следующие основные гипотезы:

$$E(u_i)=0; \tag{3.2}$$

$$E(u_i u_j)=\begin{cases}\sigma^2 & \text{при} \quad i=j,\\ 0 & \text{при} \quad i\neq j,\end{cases} \tag{3.3}$$

X_1, X_3, ..., X_k – неслучайные переменные; (3.4)

Не должно существовать строгой линейной зависимости между переменными X_1, X_3, ..., X_k. (3.5)

Первая гипотеза (3.2) означает, что переменные u_i имеют нулевую среднюю.

Суть гипотезы (3.3) в том, что все случайные ошибки u_i имеют постоянную дисперсию, то есть выполняется условие гомоскедастичности дисперсии (см. подробнее раздел 4).

[2] Другой возможный путь решения - это известная схема управляемого эксперимента – см., например: Джонсон Н., Лион Ф. Статистика и планирование эксперимента в технике и науке. В 2-х т. М.: Мир, 1980.

Согласно (3.4) в повторяющихся выборочных наблюдениях источником возмущений Y являются случайные колебания u_i, а значит, свойства оценок и критериев обусловлены объясняющими переменными X_1, X_3, ..., X_k.

Последняя гипотеза (3.5) означает, в частности, что не существует линейной зависимости между объясняющими переменными, включая переменную X_0, которая всегда равна 1.

Понятно, что условия (3.2)-(3.4) соответствуют своим аналогам для случая двух переменных в п.2.2.

3.2. Оценивание коэффициентов КЛММР методом наименьших квадратов

Применяя к (3.1) с учетом (3.2)-(3.5) МНК, получаем из необходимых условий минимизации функционала:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_k X_{ki})^2,$$

т.е. обращения в нуль частных производных по каждому из параметров:

$$\frac{\partial}{\partial \hat{\beta}_0}\left(\sum_{i=1}^{n} e_i^2\right) = -2\sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_k X_{ki}) = 0;$$

$$\frac{\partial}{\partial \hat{\beta}_1}\left(\sum_{i=1}^{n} e_i^2\right) = -2\sum_i X_{1i}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_k X_{ki}) = 0;$$

...

$$\frac{\partial}{\partial \hat{\beta}_k}\left(\sum_{i=1}^{n} e_i^2\right) = -2\sum_i X_{ki}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{1i} - ... - \hat{\beta}_k X_{ki}) = 0.$$

Упростив последние равенства, получим стандартную форму нормальных уравнений, решение которых дает искомые оценки параметров:

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}; \\ \sum_{i=1}^{n} Y_i X_{1i} = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{1i} X_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{1i} X_{ki}; \\ \quad ... \\ \sum_{i=1}^{n} Y_i X_{ki} = \hat{\beta}_0 \sum_{i=1}^{n} X_{ki} + \hat{\beta}_1 \sum_{i=1}^{n} X_{ki} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{ki} X_{2i} + ... + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2. \end{cases} \tag{3.6}$$

Сложность решения системы линейных уравнений (3.6) с (k+1) неизвестными увеличивается быстрее, чем растет k. В зависимости от количества уравнений система может быть решена методом исключения Гаусса или методом Крамера или другим численным методом решения системы линейных алгебраических уравнений.

Поскольку для большинства практических задач изучаются несколько альтернативных спецификаций модели (3.1), то широкое применение ЭВМ, а также специальных статистических пакетов позволяет значительно упростить процедуру оценивания.

В результате решения системы[3] (3.6) получим оценки коэффициентов $\hat{\beta}_j$, j=0,2,…,k.

Возможна и другая запись уравнения (3.1) в так называемом стандартизованном масштабе:

$$t_Y = b_1 t_{X_1} + b_2 t_{X_2} + ... + b_k t_{X_k} + u, \qquad (3.7)$$

где $t_Y, t_{X_1}, ..., t_{X_k}$ - стандартизованные переменные:

$$t_Y = \frac{Y - \bar{Y}}{\sigma_Y},\ t_{X_j} = \frac{X_j - \bar{X}_j}{\sigma_{X_j}},\ j=1,2,\ldots,k,$$

для которых среднее значение равно нулю:

$$\bar{t}_Y = \bar{t}_{X_j} = 0,\ j=1,2,\ldots,k,$$

а среднее квадратическое отклонение равно единице:

$$\sigma_{t_Y} = \sigma_{t_{X_j}} = 1,\ j=1,2,\ldots,k,$$

b_j, j=1,2,…,k – стандартизованные коэффициенты регрессии.

Нетрудно установить зависимость между коэффициентами "чистой" регрессии β_j и стандартизованными коэффициентами регрессии b_j, j=1,2,…,k, а именно:

$$b_j = \beta_j \frac{\sigma_{X_j}}{\sigma_Y},\ j=1,2,\ldots,k, \qquad (3.8)$$

причем $\beta_0 = \bar{Y} - \beta_1 \bar{X}_1 - ... - \beta_k \bar{X}_k$.

Соотношение (3.8) позволяет переходить от уравнения вида (3.7) к уравнению вида (3.1).

Стандартизованные коэффициенты регрессии показывают, на сколько "сигм" изменится в среднем результат (Y), если соответствующий фактор X_j изменится на одну "сигму" при неизменном среднем уровне других факторов.

[3] С использованием матричной алгебры можно получить аналитическую формулу для оценок коэффициентов, см., например: Магнус Я.Р, Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: Дело, 2000. С. 60-63.

В силу того, что все переменные центрированы и нормированы, коэффициенты b_j, j=1,2,…,k, сравнимы между собой (в этом их отличие от β_j). Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат, что позволяет произвести отсев факторов – исключить из модели факторы с наименьшими значениями b_j.

Нетрудно показать, что оценки МНК $\hat{\beta}_j$, j=0,2,…,k являются наиболее эффективными (в смысле наименьшей дисперсии) оценками в классе линейных несмещенных оценок (теорема Гаусса-Маркова).

Как было уже указано раньше, достоинством метода множественной регрессии является возможность выделения влияния каждого из факторов X_j в условиях, когда воздействие многих переменных на результат эксперимента не удается контролировать. Степень раздельного влияния каждого из факторов характеризуется оценками $\hat{\beta}_j$, j=1,2,…,k.

Пример 1. Исследуется зависимость между стоимостью грузовой автомобильной перевозки Y (тыс. руб), весом груза X_1 (тонн) и расстоянием X_2 (тыс.км) по 20 транспортным компаниям. Исходные данные приведены в таблице 3.1.

Таблица 3.1

Y	51	16	74	7,5	33,0	26,0	11,5	52	15,8	8,0	26	6,0	5,8	13,8	6,20	7,9	5,4	56,0	25,5	7,1
X_1	35	16	18	2,0	14,0	33,0	20	25	13	2,0	21	11,0	3	3,5	2,80	17,0	3,4	24,0	9,0	4,5
X_2	2	1,1	2,55	1,7	2,4	1,55	0,6	2,3	1,4	2,1	1,3	0,35	1,65	2,9	0,75	0,6	0,9	2,5	2,2	0,95

В данном примере мы располагаем пространственной выборкой объема n=20, число объясняющих переменных k=2.

Модель специфицируем в виде линейной функции:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u. \qquad (3.9)$$

Следовательно, система нормальных уравнений для модели (3.9) будет иметь вид

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}; \\ \sum_{i=1}^{n} Y_i X_{1i} = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \hat{\beta}_2 \sum_{i=1}^{n} X_{1i} X_{2i}; \\ \sum_{i=1}^{n} Y_i X_{2i} = \hat{\beta}_0 \sum_{i=1}^{n} X_{2i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{2i} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i}^2. \end{cases} \qquad (3.10)$$

Рассчитаем по данным табл. 3.1 необходимые для составления указанной системы суммы:

ΣY=454,5; ΣX_1=277,2; ΣX_2=31,8;

ΣY^2=18206,89; ΣX_1^2=5860,9; ΣX_2^2=61,45;

$\overline{Y}$=22,73; $\overline{X}_1$=13,86; $\overline{X}_2$=1,59;

$\Sigma X_1 Y$=8912,57; $\Sigma X_2 Y$=908,56; $\Sigma X_1 X_2$=459,24;

Получим систему нормальных уравнений (3.10) в виде:

$$\begin{cases} 454{,}5 = 20\hat{\beta}_0 + 277{,}2\hat{\beta}_1 + 31{,}8\hat{\beta}_2; \\ 8912{,}57 = 277{,}2\hat{\beta}_0 + 5860{,}9\hat{\beta}_1 + 459{,}24\hat{\beta}_2; \\ 908{,}56 = 31{,}8\hat{\beta}_0 + 459{,}24\hat{\beta}_1 + 61{,}45\hat{\beta}_2. \end{cases}$$

Решая последнюю систему линейных алгебраических уравнений, например методом Крамера, получим:

$$\hat{\beta}_0 = -17{,}31;\ \hat{\beta}_1 = 1{,}16;\ \hat{\beta}_2 = 15{,}10.$$

Уравнение регрессии имеет вид:

$$Y = -17{,}31 + 1{,}16 \cdot X_1 + 15{,}10 \cdot X_2.$$

Или, с учетом (3.8) и расчетов:

$$\sigma_Y = \sqrt{\left(\sum Y^2 - \frac{1}{n}\left(\sum Y\right)^2\right)\Big/ n} = \sqrt{(18206{,}89 - (454{,}5)^2/20)/20} = 19{,}85,$$

$$\sigma_{X_1} = \sqrt{\left(\sum X_1^2 - \frac{1}{n}\left(\sum X_1\right)^2\right)\Big/ n} = \sqrt{(5860{,}9 - (277{,}2)^2/20)/20} = 10{,}05,$$

$$\sigma_{X_2} = \sqrt{\left(\sum X_2^2 - \frac{1}{n}\left(\sum X_2\right)^2\right)\Big/ n} = \sqrt{(61{,}45 - (31{,}8)^2/20)/20} = 0{,}74.$$

$$b_1 = \beta_1 \frac{\sigma_{X_1}}{\sigma_Y} = 1{,}16\frac{10{,}05}{19{,}85} = 0{,}77,\ b_2 = \beta_2 \frac{\sigma_{X_2}}{\sigma_Y} = 15{,}10\frac{0{,}74}{19{,}85} = 0{,}56$$

уравнение регрессии в стандартизованном масштабе:

$$t_Y = 0{,}77 t_{X_1} + 0{,}56 t_{X_2}.$$

То есть с ростом веса груза на одну сигму при неизменном расстоянии стоимость грузовых автомобильных перевозок увеличивается в среднем на 0,77 сигмы. Поскольку 0,77>0,56, то влияние веса груза на стоимость грузовых автомобильных перевозок больше, чем фактора расстояния.

Рассчитаем коэффициенты эластичности

$$\overline{Э}_{YX_1} = f'(\overline{X}_1)\frac{\overline{X}_1}{\overline{Y}} = \beta_1 \frac{\overline{X}_1}{\beta_0 + \beta_1\overline{X}_1 + \beta_2\overline{X}_2} = 1{,}16 \cdot 13{,}86/(-17{,}31 + 1{,}16 \cdot 13{,}86 + 15{,}10 \cdot 1{,}59) = 0{,}71,$$

$$\overline{Э}_{YX_1} = f'(\overline{X}_2)\frac{\overline{X}_2}{\overline{Y}} = \beta_2 \frac{\overline{X}_2}{\beta_0 + \beta_1\overline{X}_1 + \beta_2\overline{X}_2} = 1{,}05.$$

С увеличением среднего веса груза на 1% от его среднего уровня средняя стоимость перевозок возрастет на 0,71% от своего среднего уровня, при увеличении среднего расстояния перевозок на 1% средняя стоимость доставки груза увеличится на 1,05%. Различия в силе влияния факторов на результат полученные при сравнении уравнения регрессии в стандартизованном масштабе и коэффициентов эластичности объясняются тем, что коэффициент эластичности рассчитывается исходя из соотношения средних, а стандартизованные коэффициенты регрессии из соотношения средних квадратических отклонений.

Поскольку обычно статистики используют показатель грузооборота, вычисляемый как сумма произведений массы перевезенных грузов на расстояние перевозки, то построим регрессию стоимости 1 км грузовых автомобильных перевозок Y на грузооборот Q ($Q=X_1X_2$):

$$P = 5{,}88 + 0{,}48 \cdot Q - 0{,}003 \cdot Q^2,$$

причем регрессор $Q^2 = Q*Q$ включен исходя из соображений известного экономического закона убывающей предельной полезности, согласно которому в данном случае стоимость перевозки на 1 км должна уменьшаться с ростом грузооборота, т.е. коэффициент при Q^2 должен иметь (и в построенном уравнении имеет) отрицательный знак. ∇

Как уже говорилось в разделе 2.3, регрессионные модели не ограничиваются классом линейных функций. Линеаризация нелинейных функций в уравнении регрессии имеет особенности, рассмотренные в примере.

Пример 2. Исследуется зависимость между выпуском Q (млн. $) и затратами труда L (чел.) и капитала K (млн. $) в металлургической промышленности по 27 американским компаниям. Исходные данные приведены в таблице 3.2.

Таблица 3.2

Q	L	K	Q	L	K
657,29	162,31	279,99	1917,55	536,73	2109,34
935,93	214,43	542,50	9849,17	1564,83	13989,55
1110,65	186,44	721,51	1088,27	214,62	884,24
1200,89	245,83	1167,68	8095,63	1083,10	9119,70
1052,68	211,40	811,77	3175,39	521,74	5686,99
3406,02	690,61	4558,02	1653,38	304,85	1701,06
2427,89	452,79	3069,91	5159,31	835,69	5206,36
4257,46	714,20	5585,01	3378,40	284,00	3288,72
1625,19	320,54	1618,75	592,85	150,77	357,32
1272,05	253,17	1562,08	1601,98	259,91	2031,93
1004,45	236,44	662,04	2065,85	497,60	2492,98
598,87	140,73	875,37	2293,87	275,20	1711,74
853,10	145,04	1696,98	745,67	137,00	768,59
1165,63	240,27	1078,79			

Мы располагаем пространственной выборкой объема n=27, число объясняющих переменных k=2.

Модель зависимости между выпуском и затратами труда и капитала, как правило, специфицируется в виде производственной функции, чаще всего Кобба-Дугласа:

$$Q = AL^{\beta_1}K^{\beta_2}\varepsilon. \qquad (3.11)$$

Поскольку модель (3.11) является нелинейной, преобразуем ее к виду линейной по параметрам. Для этого возьмем логарифм от обеих частей в уравнении (3.11):

$$\ln Q = \ln A + \beta_1 \ln L + \beta_2 \ln K + \ln\varepsilon.$$

Переобозначим для удобства Y=ln$Q$, β_0=ln$A$, X_1=ln$L$, X_2=ln$K$, u=ln ε, тогда имеем линейную модель вида:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + u. \qquad (3.12)$$

Исходные данные к модели вида (3.11) получаются логарифмированием чисел, представленных в таблице 3.2. Соответственно получим табл. 3.3.

После процедуры лианеризации система нормальных уравнений для модели (3.11) будет иметь такой же вид, как и система (3.10)

Рассчитаем по данным табл. 3.3 необходимые для составления указанной системы суммы:

ΣY=200,98;	ΣX_1=155,62;	ΣX_2=201,04;
ΣY^2=1511,07;	ΣX_1^2=908,13;	ΣX_2^2=1521,31;
$\overline{Y}$ =7,44;	$\overline{X}_1$=5,76;	$\overline{X}_2$ =7,45;
$\Sigma X_1 Y$=1170,67;	$\Sigma X_2 Y$=1514,54;	$\Sigma X_1 X_2$=1173,51;

Таблица 3.3

Y	X_1	X_2	Y	X_1	X_2
6,49	5,09	5,63	7,56	6,29	7,65
6,84	5,37	6,30	9,20	7,36	9,55
7,01	5,23	6,58	6,99	5,37	6,78
7,09	5,50	7,06	9,00	6,99	9,12
6,96	5,35	6,70	8,06	6,26	8,65
8,13	6,54	8,42	7,41	5,72	7,44
7,79	6,12	8,03	8,55	6,73	8,56
8,36	6,57	8,63	8,13	5,65	8,10
7,39	5,77	7,39	6,38	5,02	5,88
7,15	5,53	7,35	7,38	5,56	7,62
6,91	5,47	6,50	7,63	6,21	7,82
6,40	4,95	6,77	7,74	5,62	7,45
6,75	4,98	7,44	6,61	4,92	6,64
7,06	5,48	6,98			

Получим систему нормальных уравнений после подстановки соответствующих значений в (3.10) в виде:

$$\begin{cases} 200{,}98 = 27\hat{\beta}_0 + 155{,}62\hat{\beta}_1 + 201{,}04\hat{\beta}_2; \\ 1170{,}67 = 155{,}62\hat{\beta}_0 + 908{,}13\hat{\beta}_1 + 1173{,}51\hat{\beta}_2; \\ 1514{,}54 = 201{,}04\hat{\beta}_0 + 1173{,}51\hat{\beta}_1 + 1521{,}31\hat{\beta}_2. \end{cases}$$

Решая последнюю систему методом Крамера, получим:

$$\hat{\beta}_0$=1,11, $\hat{\beta}_1$=0,56, $\hat{\beta}_2$=0,41.

Уравнение регрессии имеет вид:

$$Y$=1,11+0,56 $\cdot X_1$+0,41 $\cdot X_2$.

Или, с учетом (3.8) и расчетов: σ_Y=0,75, σ_{X_1}=0,65, σ_{X_2}=0,96, $b_1 = \beta_1 \frac{\sigma_{X_1}}{\sigma_Y}$=0,56$\frac{0,65}{0,75}$=0,48, $b_2 = \beta_2 \frac{\sigma_{X_2}}{\sigma_Y}$=0,41$\frac{0,96}{0,75}$=0,52 уравнение регрессии в стандартизованном масштабе:

$$t_Y = 0{,}48t_{X_1} + 0{,}52t_{X_2}.$$

Нетрудно восстановить (учитывая, что A=$e^{1,11}$=3,03) исходную модель (3.9)

$$Q = 3{,}03L^{0,56}K^{0,41}.$$

Эластичность выпуска продукции Q по труду L равна 0,56, а эластичность выпуска продукции Q по капиталу K равна 0,41. Следовательно увеличение затрат труда на 1% приведет к росту выпуска продукции на 0,56%, а увеличение затрат капитала на 1% приведет к росту выпуска продукции на 0,41%.

Очевидно, что обе величины $\hat{\beta}_1$ и $\hat{\beta}_2$ должны находиться между нулем и единицей. Они должны быть положительными, так как увеличение затрат факторов должно вызывать рост выпуска. В то же время, вероятно, они будут меньше единицы, т.к. мы предполагаем, что уменьшение эффекта от масштаба производства приводит к более медленному росту выпуска продукции, чем затрат производственных факторов, если другие факторы остаются постоянными.

Продолжая интерпретацию результатов регрессии $Q = 3{,}03L^{0,56}K^{0,41}$, отметим, что ($\hat{\beta}_1 + \hat{\beta}_2$)<1, т.е. имеет место убывающий эффект от масштаба производства (выпуск увеличивается в меньшей пропорции, чем L и K). ∇

3.3 Парная и частная корреляция в КЛММР

В случаях, когда имеется одна независимая и одна зависимая переменные, естественной мерой зависимости (в рамках линейного подхода) является выборочный (парный) коэффициент корреляции между ними.

Использование множественной регрессии позволяет обобщить это понятие на случай, когда имеется несколько независимых переменных. В этом случае необходима корректировка, так как высокое значение коэффициента корреляции между зависимой и какой-либо независимой переменной может означать высокую степень линейной зависимости, но может означать и то, что третья переменная, оказывает значительное влияние на две первых и, что именно она служит основной причиной их высокой корреляции. Поэтому необходимо найти "чистую" корреляцию между двумя переменными, исключив влияние других факторов путем расчета коэффициента частной корреляции.

Коэффициенты частной корреляции для уравнения регрессии с двумя независимыми переменными рассчитываются как:

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)\cdot(1-r_{x_1x_2}^2)}}, \qquad (3.13)$$

$$r_{yx_2(x_1)} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)\cdot(1-r_{x_1x_2}^2)}}, \qquad (3.14)$$

$$r_{x_1x_2(y)} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)\cdot(1-r_{yx_2}^2)}}, \qquad (3.15)$$

где $r_{yx_1(x_2)}$ - коэффициент частной корреляции между y и x_1 при исключенном влиянии x_2;

$r_{yx_2(x_1)}$ - коэффициент частной корреляции между y и x_2 при исключенном влиянии x_1;

$r_{x_1x_2(y)}$ - коэффициент частной корреляции между x_1 и x_2, исключающий влияние y.

Заметим, что парные линейные коэффициенты корреляции, стоящие в правых частях формул (3.13)-(3.15), могут быть рассчитаны с помощью формулы (2.9).

Коэффициенты частной корреляции более высоких порядков можно определить через коэффициенты частной корреляции более низких порядков по следующей рекуррентной формуле:

$$r_{yx_i(x_1x_2\ldots x_k)} = \frac{r_{yx_i(x_1x_2\ldots x_{k-1})} - r_{yx_k(x_1x_2\ldots x_{k-1})}\, r_{x_ix_k(x_1x_2\ldots x_{k-1})}}{\sqrt{\left(1-r_{yx_k(x_1x_2\ldots x_{k-1})}^2\right)\cdot\left(1-r_{x_ix_k(x_1x_2\ldots x_{k-1})}^2\right)}} \qquad (3.16)$$

Коэффициенты частной корреляции широко используются на стадии формирования модели, при отборе факторов.

Так, например, при построении многофакторной модели применяется метод исключения переменных, в ходе которого строится уравнение регрессии с полным набором переменных, затем рассчитывается матрица частных коэффициентов корреляции. Далее проверяется статистическая значимость каждого из коэффициентов согласно t-критерию Стьюдента. Независимая переменная, имеющая наименьшую и несущественную корреляцию с зависимой переменной, исключается. Затем строится новое уравнение регрессии, и процедура продолжается до тех пор, пока не окажется, что все частные коэффициенты корреляции статистически значимы, то есть существенно отличаются от нуля.

Проверка статистической значимости частного коэффициента корреляции суть проверка гипотезы о том, что он равен нулю

$$H_0:\ r_{yx_i(x_1x_2\ldots x_k)} = 0.$$

Рассчитывается статистика:

$$t = \frac{r_{yx_i(x_1x_2\ldots x_k)}}{\sqrt{1-(r_{yx_i(x_1x_2\ldots x_k)})^2}} \cdot \sqrt{n-(k+1)} \qquad (3.17)$$

Вывод о значимости частного коэффициента корреляции делается при $|t| > t_\varepsilon$, где t_ε соответствующее табличное значение t-распределения с $(n-(k+1))$ степенями свободы.

Пример (продолжение примера 1). Рассчитаем парные линейные коэффициенты корреляции, применяя формулу (2.9) и одновременно проверяя их статистическую значимость.

$$r_{YX_1} = \frac{n\sum_{i=1}^{n} X_1Y - \sum_{i=1}^{n} X_1 \sum_{i=1}^{n} Y}{\sqrt{\left[n\sum_{i=1}^{n} X_1^2 - \left(\sum_{i=1}^{n} X_1\right)^2\right]\left[n\sum_{i=1}^{n} Y^2 - \left(\sum_{i=1}^{n} Y\right)^2\right]}} =$$

$$= \frac{20\cdot 8912{,}57 - 277{,}2\cdot 454{,}5}{\sqrt{(20\cdot 5860{,}9 - 76839{,}84)\cdot(20\cdot 18206{,}89 - 206570{,}3)}} = 0{,}6553,$$

$$t = 0{,}6553\cdot\sqrt{20-2}\,/\sqrt{1-(0{,}6553)^2} = 3{,}68,$$

$$r_{YX_2} = \frac{n\sum_{i=1}^{n} X_2Y - \sum_{i=1}^{n} X_2 \sum_{i=1}^{n} Y}{\sqrt{\left[n\sum_{i=1}^{n} X_2^2 - \left(\sum_{i=1}^{n} X_2\right)^2\right]\left[n\sum_{i=1}^{n} Y^2 - \left(\sum_{i=1}^{n} Y\right)^2\right]}} =$$

$$= \frac{20\cdot 908{,}56 - 31{,}8\cdot 454{,}5}{\sqrt{(20\cdot 61{,}5 - 1011{,}24)\cdot(20\cdot 18206{,}89 - 206570{,}3)}} = 0{,}6346,$$

$$t = 0{,}6346 \cdot \sqrt{20-2} / \sqrt{1-(0{,}6346)^2} = 3{,}60,$$

$$r_{X_1X_2} = \frac{n\sum_{i=1}^{n} X_1X_2 - \sum_{i=1}^{n} X_1 \sum_{i=1}^{n} X_2}{\sqrt{\left[n\sum_{i=1}^{n} X_1^2 - \left(\sum_{i=1}^{n} X_1\right)^2\right]\left[n\sum_{i=1}^{n} X_2^2 - \left(\sum_{i=1}^{n} X_2\right)^2\right]}} =$$

$$= \frac{20 \cdot 8912{,}57 - 277{,}2 \cdot 31{,}8}{\sqrt{(20 \cdot 5860{,}9 - 76839{,}84) \cdot (20 \cdot 61{,}5 - 1011{,}24)}} = 0{,}1247,$$

$$t = 0{,}1247 \cdot \sqrt{20-2} / \sqrt{1-(0{,}1247)^2} = 2{,}80.$$

Составим матрицу парных линейных коэффициентов корреляции (в скобках значение *t*-статистик):

	y	x_1	x_2
y	1,0	0,6553 (3,68)	0,6346 (3,60)
x_1	0,6553 (3,68)	1,0	0,1247(2,80)
x_2	0,6346(3,60)	0,1247(2,80)	1,0

Коэффициент корреляции между y и x_1, свидетельствует о прямой статистически значимой связи между стоимостью перевозки и весом перевозимого груза. Коэффициент корреляции между y и x_2 также свидетельствует о прямой и статистически значимой связи между стоимостью перевозки и расстоянием перевозки. Величина статистически значимого коэффициента корреляции между x_1 и x_2 означает практическое отсутствие взаимосвязи между расстоянием перевозки и весом груза, что не противоречит первоначальным предположениям о том, что расстояние перевозки не может быть обусловлено весом груза и наоборот.

Рассчитаем коэффициенты частной корреляции согласно формулам (3.13)-(3.15) и проверим их значимость согласно (3.17):

$$r_{yx_1(x_2)} = \frac{0{,}6553 - 0{,}6346 \cdot 0{,}1247}{\sqrt{(1-(0{,}6346)^2) \cdot (1-(0{,}1247)^2)}} = 0{,}7513;$$

$$t = \frac{0{,}7513}{\sqrt{1-(0{,}7513)^2}} \cdot \sqrt{20-(2+1)} = 4{,}69,$$

$$r_{yx_2(x_1)} = \frac{0{,}6346 - 0{,}6553 \cdot 0{,}1247}{\sqrt{(1-(0{,}6553)^2) \cdot (1-(0{,}1247)^2)}} = 0{,}7377;$$

$$t = \frac{0{,}7377}{\sqrt{1-(0{,}7377)^2}} \cdot \sqrt{20-(2+1)} = 4{,}51,$$

$$r_{x_1x_2(y)} = \frac{0{,}1247 - 0{,}6553 \cdot 0{,}6346}{\sqrt{(1-(0{,}6553)^2) \cdot (1-(0{,}6346)^2)}} = -0{,}4987;$$

$$t = \frac{-0{,}4987}{\sqrt{1-(-0{,}4987)^2}} \cdot \sqrt{20-(2+1)} = -2{,}37.$$

Составим матрицу частных коэффициентов корреляции (в скобках значение *t*-статистик):

	y	x_1	x_2
y	1,0	0,7513 (4,69)	0,7377 (4,51)
x_1	0,7513 (4,69)	1,0	-0,4987(-2,37)
x_2	0,7377(4,51)	-0,4987(-2,37)	1,0

Как уже говорилось ранее, частные коэффициенты корреляции показывают "чистую" корреляцию пары переменных, исключающую влияние прочих переменных, включенных в уравнение. Таким образом, наиболее сильной является взаимосвязь между стоимостью перевозки и весом груза. Однако заметим, что частные коэффициенты корреляции между y и x_1, y и x_2 свидетельствуют о более сильных взаимосвязях независимых переменных с зависимой, чем это показывают значения парных коэффициентов корреляции. Это произошло потому, что парный коэффициент корреляции завысил тесноту связи между x_1 и x_2, занизив при этом тесноту связи между y и x_1, y и x_2. Отметим также, что все частные коэффициенты корреляции статистически значимы. ∇

3.4 Множественный коэффициент корреляции и множественный коэффициент детерминации

Множественный коэффициент корреляции используется в качестве меры степени тесноты статистической связи между результирующим показателем (зависимой переменной) y и набором объясняющих (независимых) переменных $x_1, x_2, ..., x_k$ или, иначе говоря, оценивает тесноту совместного влияния факторов на результат.

Множественный коэффициент корреляции может быть вычислен по ряду формул[4], в том числе:

- с использованием матрицы парных коэффициентов корреляции

$$R_{yx_1x_2...x_k} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}, \qquad (3.18)$$

[4] Подробнее смотри Эконометрика: Учебник/ Под. ред. Елисеевой И.И. М.:Финансы и статистика, 2001. С.112-120.

где Δr - определитель матрицы парных коэффициентов корреляции $y, x_1, x_2, ..., x_k$,

Δr_{11} - определитель матрицы межфакторной корреляции $x_1, x_2, ..., x_k$;

♦ стандартизованных коэффициентов регрессии b_{x_i} и парных коэффициентов корреляции r_{yx_i}

$$R_{yx_1x_2...x_k} = \sqrt{\sum b_{x_i} \cdot r_{yx_i}}. \qquad (3.19)$$

Для модели, в которой присутствуют две независимые переменные, формула (3.18) упрощается

$$R_{yx_1x_2} = \sqrt{\frac{r_{yx_1}^2 + r_{yx_2}^2 - 2 \cdot r_{yx_1} \cdot r_{yx_2} \cdot r_{x_1x_2}}{1 - r_{x_1x_2}^2}}. \qquad (3.20)$$

Квадрат множественного коэффициента корреляции равен коэффициенту детерминации R^2. Как и в случае парной регрессии, R^2 свидетельствует о качестве регрессионной модели и отражает долю общей вариации результирующего признака y, объясненную изменением функции регрессии $f(x)$ (см. 2.4). Кроме того, коэффициент детерминации может быть найден по формуле

$$R^2 = 1 - \frac{\sigma_{ост}^2}{\sigma_y^2}. \qquad (3.21)$$

Однако использование R^2 в случае множественной регрессии является не вполне корректным, так как коэффициент детерминации возрастает при добавлении регрессоров в модель. Это происходит потому, что остаточная дисперсия уменьшается при введении дополнительных переменных. И если число факторов приблизится к числу наблюдений, то остаточная дисперсия будет равна нулю, и коэффициент множественной корреляции, а значит и коэффициент детерминации, приблизятся к единице, хотя в действительности связь между факторами и результатом и объясняющая способность уравнения регрессии могут быть значительно ниже.

Для того чтобы получить адекватную оценку того, насколько хорошо вариация результирующего признака объясняется вариацией нескольких факторных признаков, применяют скорректированный коэффициент детерминации

$$R_{скорр}^2 = 1 - (1 - R^2) \cdot \frac{n-1}{n-k-1} \qquad (3.22)$$

Скорректированный коэффициент детерминации всегда меньше R^2. Кроме того, в отличие от R^2, который всегда положителен, $R_{скорр}^2$ может принимать и отрицательное значение.

Пример (продолжение примера 1). Рассчитаем множественный коэффициент корреляции, согласно формуле (3.20):

$$R_{yx_1x_2} = \sqrt{\frac{(0{,}6553)^2 + (0{,}6346)^2 - 2 \cdot 0{,}6553 \cdot 0{,}6346 \cdot 0{,}1247}{1 - (0{,}1247)^2}} = 0{,}8601.$$

Величина множественного коэффициента корреляции, равного 0,8601, свидетельствует о сильной взаимосвязи стоимости перевозки с весом груза и расстоянием, на которое он перевозится.

Коэффициент детерминации равен: R^2=0,7399.

Скорректированный коэффициент детерминации рассчитываем по формуле (3.22):

$$R_{скорр}^2 = 1 - (1 - 0{,}7399) \cdot \frac{20-1}{20-2-1} = 0{,}7092.$$

Заметим, что величина скорректированного коэффициента детерминации отличается от величины коэффициента детерминации.

Таким образом, 70,9% вариации зависимой переменной (стоимости перевозки) объясняется вариацией независимых переменных (весом груза и расстоянием перевозки). Остальные 29,1% вариации зависимой переменной объясняются факторами, неучтенными в модели.

Величина скорректированного коэффициента детерминации достаточно велика, следовательно, мы смогли учесть в модели наиболее существенные факторы, определяющие стоимость перевозки. ∇

3.5. Оценка качества модели множественной регрессии

Проверка качества модели множественной регрессии может быть осуществлена с помощью дисперсионного анализа.

Как уже было отмечено (см. 2.5), сумма квадратов отклонений от среднего в выборке равна сумме квадратов отклонений значений $\hat{Y}$, полученных по уравнению регрессии, от выборочного среднего $\bar{Y}$ плюс сумма квадратов отклонений Y от линии регрессии $\hat{Y}$.

С учетом (3.21) получим таблицу дисперсионного анализа (табл. 3.4), аналог таблицы 2.3.

Проверка качества модели множественной регрессии в целом может быть осуществлена с помощью F-критерия Фишера. Для проверки гипотезы о том, что линейная связь между $x_1, x_2, ...x_k$ и y отсутствует:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_к = 0,$$

воспользуемся соотношением

$$F = \frac{R^2}{k} : \frac{1-R^2}{n-(k+1)} \qquad (3.23)$$

которое удовлетворяет F - распределению Фишера с (k, n-(k+1)) степенями свободы. Критические значения этой статистики F_ε для уровня значимости ε затабулированы.

Таблица 3.4

Таблица дисперсионного анализа

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Дисперсия на одну степень свободы
$x_1, x_2, ...x_k$	$Q_1 = \sum\left(\hat{Y} - \bar{Y}\right)^2 = n\sigma_y^2 R^2$	k	$D_1 = \frac{Q_1}{k}$
Остаток	$Q_2 = \sum\left(Y - \hat{Y}\right)^2 = n\sigma_y^2(1-R^2)$	n-k-1	$D_2 = \frac{Q_2}{n-(k+1)}$
Общая вариация	$\sum(Y-\bar{Y})^2 = Q_1 + Q_2 = n\sigma_y^2$	n-1	

Если $F>F_\varepsilon$, то гипотеза об отсутствии связи между переменными $x_1, x_2, ...x_k$ и y отклоняется, в противном случае гипотеза H_0 принимается и уравнение регрессии не значимо.

Пример (продолжение примера 1). Заполним таблицу дисперсионного анализа:

Таблица дисперсионного анализа

Источник вариации	Сумма квадратов отклонений	Число степеней свободы	Дисперсия
$x_1, x_2, ...x_k$	5828,84	2	2914,42
Остаток	2049,54	17	120,56
Общая вариация	7878,38	19	

Получаем $F = \frac{0,74}{2} : \frac{1-0,74}{20-(2+1)} = 24,17$, $F_\varepsilon = F_{(2,17)} = 3,59$.

В нашем примере $F>F_\varepsilon$, следовательно, нулевая гипотеза отклоняется, и уравнение множественной регрессии значимо. ∇

Помимо проверки значимости уравнения в целом, можно проверить статистическую значимость каждого из коэффициентов регрессии в отдельности.

Фактически это означает проверку одной из гипотез:

1) $\begin{matrix} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{matrix}$; …; k) $\begin{matrix} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{matrix}$.

Статистическая значимость каждого из коэффициентов регрессии определяется при помощи t-критерия Стьюдента. Решение о том, что верна нулевая гипотеза, принимается в случае, когда $|t| < t_\varepsilon$, иначе принимается альтернативная гипотеза.

Значение t-статистики Стьюдента в случае множественной регрессии определяется по формуле:

$$t_{\beta_i} = \frac{\hat{\beta}_i}{\mu_{\hat{\beta}_i}}, \qquad (3.24)$$

где $\mu_{\hat{\beta}_i}$ - стандартная ошибка коэффициента регрессии $\hat{\beta}_i$, которая определяется по формуле

$$\mu_{\hat{\beta}_i} = \frac{\sigma_y \cdot \sqrt{1-R^2_{yx_1...x_k}}}{\sigma_{x_i} \cdot \sqrt{1-R^2_{x_i x_1...x_k}}} \cdot \frac{1}{\sqrt{n-k-1}}, \qquad (3.25)$$

здесь σ_y - стандартное отклонение y;

σ_{x_i} - стандартное отклонение x_i;

$R^2_{x_i x_1...x_k}$ - коэффициент детерминации для зависимости фактора x_i от других факторов уравнения множественной регрессии.

Пример (продолжение примера 1). Проверим значимость коэффициентов регрессии. В случае, когда в уравнение регрессии включены две независимые переменные, формула (3.24) упрощается

$$t_{\hat{\beta}_1} = \frac{\sqrt{R^2_{yx_1...x_k} - r^2_{yx_2}}}{\sqrt{1-R^2_{x_i x_1...x_k}}} \cdot \frac{1}{\sqrt{n-k-1}}, \; t_{\hat{\beta}_2} = \frac{\sqrt{R^2_{yx_1...x_k} - r^2_{yx_1}}}{\sqrt{1-R^2_{x_i x_1...x_k}}} \cdot \frac{1}{\sqrt{n-k-1}}.$$

Таким образом:

$$t_{\hat{\beta}_1} = \frac{\sqrt{0,7399-(0,6346)^2}}{\sqrt{1-0,7399}} \cdot \frac{1}{\sqrt{20-2-1}} = 4,69,$$

$$t_{\hat{\beta}_2} = \frac{\sqrt{0,7399-(0,6553)^2}}{\sqrt{1-0,7399}} \cdot \frac{1}{\sqrt{20-2-1}} = 4,50,$$

$t_{\alpha;n-(k+1)} = t_{0,05;17} = 2,11.$

Так как в обоих случаях $|t| > t_\varepsilon$, то коэффициенты регрессии значимы, следовательно, и вес груза, и расстояние грузовой перевозки оказывают существенное, статистически значимое влияние на стоимость перевозки. ∇

3.6 Мультиколлинеарность и методы ее устранения

Одним из важнейших этапов построения регрессии является отбор факторов X_{ji}, j=1,..., k, i=1,2,…,n, включаемых в регрессию (3.1). Наибольшее распространение получили следующие методы построения уравнения множественной регрессии: метод исключения, метод включения, шаговый регрессионный анализ. Перечисленные методы дают близкие результаты: отсев факторов из полного их набора (метод исключения), дополнительное введение фактора (метод включения), исключение ранее введенного фактора (шаговый метод).

Наиболее широко используются для решения вопроса об отборе факторов частные коэффициенты корреляции, оценивающие в чистом виде тесноту связи между фактором и результатом.

При включении факторов следует придерживаться правила, согласно которому число включаемых в модель объясняющих переменных должно быть в 5-6 раз меньше объема совокупности, по которой строится регрессия. Иначе число степеней свободы остаточной вариации будет мало, и параметры уравнения регрессии окажутся статистически незначимы.

Иногда при отборе переменных-факторов нарушается предположение (3.5). В этом случае говорят, что объясняющие переменные X_{ji}, j=1,..., k, i=1,2,…,n модели характеризуются свойством полной (строгой) мультиколлинеарности. В этом случае система (3.6) не может быть разрешена относительно неизвестных оценок коэффициентов. Строгая мультиколлинеарность встречается редко, так как ее несложно избежать на предварительной стадии отбора объясняющих переменных.

Реальная (частичная) мультиколлинеарность возникает в случаях достаточно сильных линейных статистических связей между переменными X_{ji}, j=1,..., k, i=1,2,…,n. Точных количественных критериев для проверки наличия мультиколлинеарности не существует, но имеются некоторые практические рекомендации по выявлению мультиколлинеарности.

1. Если среди парных коэффициентов корреляции между объясняющими переменными имеются значения 0,75-0,80 и выше, это свидетельствует о присутствии мультиколлинеарности.

Пример. В примере 2 между переменными K и L коэффициент корреляции равен 0,96, а между ln$K$ и ln$L$ чуть меньше 0,89. ∇

2. О присутствии явления мультиколлинеарности сигнализируют некоторые внешние признаки построенной модели, являющиеся его следствиями:

- некоторые из оценок $\hat{\beta}_j$, j=1,2,…,k имеют неправильные с точки зрения экономической теории знаки или неоправданно большие по абсолютной величине значения,

- небольшое изменение исходной выборки (добавление или изъятие малой порции данных) приводит к существенному изменению оценок коэффициентов модели вплоть до изменения их знаков,

- большинство оценок коэффициентов регрессии оказываются статистически незначимо отличающимися от нуля, в то время как в действительности многие из них имеют отличные от нуля значения, а модель в целом является значимой при проверке с помощью F-критерия.

Методы устранения мультиколлинеарности.

1. Проще всего удалить из модели один или несколько факторов.

2. Другой путь состоит в преобразовании факторов, при котором уменьшается корреляция между ними. Например, при построении регрессий на основе временных рядов помогает переход от первоначальных данных к первым разностям $\Delta=Y_t-Y_{t-1}$. В примере 2 переход от переменных K и L к их логарифмам уменьшил коэффициент корреляции с 0,96 до 0,89.

3. Использование в уравнении регрессии взаимодействия факторов, например, в виде их произведения.

4. Использование так называемой ридж-регрессии (гребневой регрессии). В этом случае к диагональным элементам системы (3.6) добавляется "гребень" τ (небольшое число, как правило, от 0,1 до 0,4):

$$\begin{cases} \sum_{i=1}^{n} Y_i = n\hat{\beta}_0 + \tau + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}; \\ \sum_{i=1}^{n} Y_i X_{1i} = \hat{\beta}_0 \sum_{i=1}^{n} X_{1i} + \hat{\beta}_1 \sum_{i=1}^{n} X_{1i}^2 + \tau + \hat{\beta}_2 \sum_{i=1}^{n} X_{1i} X_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} X_{1i} X_{ki}; \\ \quad \ldots \\ \sum_{i=1}^{n} Y_i X_{ki} = \hat{\beta}_0 \sum_{i=1}^{n} X_{ki} + \hat{\beta}_1 \sum_{i=1}^{n} X_{ki} X_{1i} + \hat{\beta}_2 \sum_{i=1}^{n} X_{ki} X_{2i} + \ldots + \hat{\beta}_k \sum_{i=1}^{n} X_{ki}^2 + \tau. \end{cases}$$

Это делает получаемые оценки смещенными, но уменьшает средние квадраты ошибок коэффициентов.

5. Использование метода главных компонент[5].

[5] См., например: [1], с. 658-661.

6. Отбор наиболее существенных объясняющих переменных на основе методов исключения, включения, шаговой регрессии, которые используют для принятия решения F-критерий.

4. Спецификация переменных в уравнениях регрессии

4.1. Спецификация уравнения регрессии и ошибки спецификации

При построении эконометрической модели исследователь специфицирует составляющие ее соотношения, выбирает переменные, входящие в эти соотношения, а также определяет вид математической функции, представляющей каждое соотношение. Остановимся на вопросе выбора переменных, которые должны быть включены в модель. До сих пор мы неявно считали, что имеем правильную спецификацию модели.

На практике никогда не получается правильная спецификация модели, возникают так называемые ошибки спецификации. Экономическая теория, положения которой используются при выборе регрессоров, не может быть совершенной. Поэтому исследователь может включить в эконометрическую модель переменные, которых там не должно быть, и может не включить другие переменные, которые должны там присутствовать.

Т.е. изучим две ситуации.

Случай 1. Исключены существенные переменные.

Процесс, порождающий данные:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \gamma_1 Z_{1i} + ... + \gamma_l Z_{li} + u_i,\ i=1,\ldots,n. \quad (4.1а)$$

Модель:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_i, \quad i = 1,2,...,n \quad (4.1б)$$

Случай 2. Включены несущественные переменные.

Процесс, порождающий данные:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + u_i, \quad i = 1,2,...,n \quad (4.2а)$$

Модель:

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \gamma_1 Z_{1i} + ... + \gamma_l Z_{li} + u_i,\ i=1,\ldots,n \quad (4.2б)$$

Часто регрессию (4.1а) называют длинной, а регрессию (4.1б) – короткой.