# Эконометрика

## Содержание

## Введение

Современные социально-экономическибе процессы и явления зависят от большого количества факторов, их определяющих. В связи с этим квалифицированному специалисту необходимо не только иметь четкие представления об основных направлениях развития экономики, но и уметь учитывать сложное взаимосвязанное многообразие факторов, оказывающих существенное влияние на изучаемый процесс. Такие исследования невозможно проводить без знания основ теории вероятностей, математической статистики, многомерных статистических методов и эконометрики, т.е. дисциплин, позволяющих исследователю разобраться в огромном количестве стохастической информации и среди множества различных вероятностных моделей выбрать единственную, наилучшим образом отражающую изучаемый процесс или явление.

Что же такое эконометрика? Можно ли сказать, что эконометрика – это наука об экономических измерениях, как подсказывает само ее название? Не пытаясь более подробно развивать эту проблему, можно привести высказывания признанных авторитетов в экономике и эконометрике.

«Эконометрика позволяет проводить количественный анализ реальных экономических явлений, основываясь на современном развитии теории и наблюдениях, связанных с методами получения выводов» (*Самуэлъсон*).

«Основная задача эконометрики – наполнить эмпирическим содержанием априорные экономические рассуждения» (*Клейн*).

«Цель эконометрики – эмпирический вывод экономических законов. Эконометрика дополняет теорию, используя реальные данные для проверки и уточнения постулируемых отношений» (*Маленво*).

Курс «Эконометрика» занимает важное место в учебных планах экономических вузов. Он входит в цикл естественнонаучных дисциплин в соответствии с государственными образовательными стандартами экономических специальностей.

Настоящее учебное пособие ориентировано на студентов экономических специальностей университетов. Предполагается, что студенты, изучающие эконометрику, уже прослушали базовые курсы по высшей математике, теории вероятностей и математической статистики, микро- и макроэкономике.

В учебном пособии излагаются основы эконометрики. Рассмотрены классическая (парная и множественная) и обобщенная модели линейной регрессии, классический и обобщенный метод наименьших квадратов, анализ временных рядов и систем одновременных уравнений. Большое внимание уделяется различным аспектам многомерной регрессии: мультиколлинеарности, фиктивным переменным.

Для большей наглядности при изложении материала приводится решение сквозной задачи по всем темам пособия.

# 1. Метод наименьших квадратов. Модель парной регрессии

## 1.1. Основы эконометрики

### *1.1.1. Понятие эконометрики. Связь эконометрики с другими областями знаний*

Слово «эконометрика» представляет собой комбинацию двух слов: «экономика» и «метрика» (от греч. «метрон» — правило определения расстояния между двумя точками в пространстве, «метрия» — измерение). Сам термин подчеркивает специфику, содержание эконометрики как науки.

*Эконометрика — наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов.*

Эконометрика представляет собой комбинацию трех областей знания:

- экономической теории;
- экономической статистики;
- математики (математической статистики).

Предметы эконометрики и статистики очень близки. Эконометрика имеет дело с массовыми экономическими явлениями (массовыми, значит повторяющимися в пространстве и во времени). Статистика имеет дело с массовыми явлениями любой природы, в том числе и в экономике.

Специфика эконометрики в том, что она ставит своей задачей при помощи статистики выразить те закономерности, которые экономическая теория и математическая экономика определяют, в общем, схематически. То есть экономическая теория и математическая экономика формулируют гипотезы, которые, в сущности, являются качественными. Эконометрика имеет дело с конкретными экономическими данными и занимается количественным описанием конкретных взаимосвязей, т. е. заменяет коэффициенты, представленные в общем виде в этих взаимосвязях, конкретными численными значениями.

Например, микроэкономическая теория утверждает, что снижение цены товара приводит к увеличению спроса на данный товар (при неизменности всех прочих факторов), т. е. устанавливается связь между спросом на товар и ценой на него. Однако микроэкономическая теория не дает количественных оценок данной связи, т. е. не позволяет ответить на вопрос: на сколько изменится спрос на данный товар в результате изменения его цены на определенную величину. Расчет количественных оценок и есть задача эконометрики.

В эконометрике часто используются математические уравнения и модели, модифицируемые, с тем, чтобы обеспечить возможность проведения эмпирических расчетов.

Большинство эконометрических методов и приемов заимствовано из математической статистики. Однако методы математической статистики универсальны и не учитывают специфики экономических данных. Специфика

экономических данных заключается в том, они не являются результатом контролируемого эксперимента. В экономике невозможно проводить многократные эксперименты (из-за изменения внешних условий). Этот факт рождает ряд специфических проблем, решение которых не входит в математическую статистику.

Кроме того, экономические данные часто содержат ошибки измерения. В эконометрике разрабатываются специальные методы анализа, позволяющие, если не устранить, то, по крайней мере, снизить влияние этих ошибок на полученные результаты.

**Таким** образом, эконометрика связывает между собой экономическую теорию и экономическую статистику и с **помощью** математико-статистических методов придает конкретное количественное выражение общим закономерностям, устанавливаемым экономической теорией.

### *1.1.2. Эконометрическая модель – главный инструмент эконометрических исследований. Задачи, решаемые на ее основе. Этапы эконометрического моделирования*

Главным инструментом эконометрики служит *эконометрическая модель.*

Можно выделить три класса эконометрических моде лей:

1. *Модель временных данных* (в которых результативный признак является функцией переменной времени или переменных, относящихся к другим моментам времени).

К моделям временных данных, представляющих собой зависимость результативного признака от времени, относятся модели:

• тренда (зависимости результативного признака от трендовой компоненты);

• сезонности (зависимости результативного признака от сезонной компоненты);

• тренда и сезонности.

К моделям временных данных, представляющих собой зависимость результативного признака от переменных, да тированных другими моментами времени, относятся модели:

• объясняющие поведение результативного признака в зависимости от предыдущих значений факторных переменных (модели с распределенным лагом);

• объясняющие поведение результативного признака в зависимости от предыдущих значений результативных переменных (модели авторегрессии);

• объясняющие поведение результативного признака в зависимости от будущих значений факторных или результативных переменных (модели ожиданий).

Модели временных данных подразделяют также на модели, построенные по стационарным и нестационарным временным рядам. Стационарные

временные ряды – ряды, имеющие постоянное среднее значение и колеблющиеся вокруг него с постоянной дисперсией. В таких рядах распределение показателя – уровня ряда не зависит от времени, т. е. стационарный временной ряд не содержит трендовой или сезонной компонент. В нестационарных временных рядах распределение уровня ряда зависит от переменной времени.

2. *Регрессионная модель с одним уравнением.*

В таких моделях результативный признак (зависимая переменная) представляется в виде функции факторных признаков (независимых переменных).

Ниже перечислены примеры регрессионных моделей с одним уравнением.

Функция цены: $P = f(Q, Pk)$, где цена определенного товара – $P$ зависит от объема его поставки – $Q$ и от цен конкурирующих товаров $Pk$.

• Функция спроса: $Q_d = f(P, Pk, I)$, где величина спроса на определенный товар – $Q_d$ зависит от цены данного товара – $P$, от цен товаров-конкурентов – $Pk$, а также от реальных доходов потребителей – $I$.

• Производственная функция: $Q = f(L, K)$, представляющая собой зависимость объема производства определенного товара $Q$ от производственных факторов, на пример от затрат капитала $K$ и затрат труда $L$.

3. *Системы одновременных уравнений.*

Эти модели описываются системами взаимосвязанных регрессионных уравнений. Система «объясняет», а также прогнозирует столько результативных признаков, сколько поведенческих уравнений входит в систему.

Уравнения системы могут быть либо тождествами, либо поведенческими уравнениями.

Для тождеств характерно, что их вид и значения пара метров известны.

В поведенческих уравнениях значения параметров требуется оценить. Кроме того, поведенческие уравнения в качестве независимых переменных могут включать не только факторные, но и результативные признаки из других уравнений системы.

Примером системы одновременных уравнений является модель спроса и предложения, включающая 3 *уравнения:*

1 – уравнение предложения: $Q_t^s = a_0 + a_1 \cdot P_t + a_2 \cdot P_{t-1}$;

2 – уравнение спроса: $Q_t^d = b_Q + b_1 \cdot P_t + b_2 \cdot I_t$;

3 – тождество равновесия: $Q_t^s = Q_t^d$ ,

где $Q_t^s$ – предложение товара в момент времени $t$;

$Q_t^d$ – спрос на товар в момент времени $t$;

$P_t$ – цена товара в момент времени $t$;

$P_{t-1}$ – цена товара в предыдущий момент времени $(t - 1)$;

$I_t$ – доход потребителей в момент времени $t$.

Данная модель «объясняет» две результативные переменные: 1) $Q_t$ – объем спроса, равный объему предложения в момент времени *t;* 2) $P_t$ – цену товара в момент времени *t.*

С помощью эконометрической модели могут быть решены самые разнообразные задачи. Их можно классифицировать по трем признакам:

1) по конечным прикладным целям;
2) по уровню иерархии;
3) по профилю анализируемой экономической системы.

По конечным прикладным целям выделяют две основные задачи:

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

• имитация возможных сценариев социально-экономического развития системы для выявления того, как планируемые изменения тех или иных поддающихся управлению параметров скажутся на выходных характеристиках.

По уровню иерархии выделяют задачи, решаемые на:

• макроуровне (страна в целом);
• мезоуровне (уровне регионов, отраслей, корпораций);
• микроуровне (на уровне семьи, предприятия, фирмы).

По профилю анализируемой экономической системы выделяют задачи, направленные на решение проблем:

• рынка;
• инвестиционной, финансовой или социальной политики;
• ценообразования;
• распределительных отношений;
• спроса и потребления;
• на определенный комплекс проблем.

Однако, чем шире комплекс проблем, тем меньше шансов провести эконометрическое исследование достаточно эффективно.

*Основные этапы эконометрического моделирования:*

1) определение конечных целей модели, набора участвующих факторных и результативных признаков;

2) качественный (теоретический) анализ сущности изучаемого явления. Формирование и формализация априорной информации, относящейся к природе исходных статистических данных и случайных составляющих;

3) выбор общего вида модели, состава и формы входящих в нее связей;

4) сбор необходимой информации, анализ ее качества;

5) оценка параметров модели;

6) оценка качества модели (т. е. оценка ее достоверности и надежности). Если качество модели не устраивает исследователя, то следует переход ко второму этапу;

7) интерпретация полученных результатов.

### *1.1.3. Типы данных и виды переменных в эконометрических исследованиях*

При моделировании экономических процессов используют два типа данных:

- пространственные данные (cross-sectional data);
- временные данные (time-series data).

*Пространственными* данными является набор сведений по разным объектам, взятым за один и тот же период или момент времени. Например, набор сведений по разным фирмам (объем производства, численность работников, размер основных производственных фондов и пр.). Другим приме ром могут служить данные об объеме, ценах потребления некоторого товара по потребителям.

*Временными* данными является набор сведений, характеризующий один и тот же объект, но за разные периоды или моменты времени. Примером временных данных могут быть ежеквартальные данные о средней заработной плате, индексе потребительских цен, числе занятых за последние годы или, например, ежедневный курс доллара США или евро на ММВБ. Отличительной особенностью временных данных является то, что они естественным образом упорядочены по времени.

Набор сведений представляет собой множество признаков, характеризующих объект исследования. Признаки являются взаимосвязанными, причем в этой взаимосвязи они могут выступать в одной из двух ролей:

- в роли результативного признака (аналог зависимой переменной $y$ в математике);
- в роли факторного признака, значения которого определяют значение признака-результата (аналог независимой переменной $x$ в математике).

В эконометрической модели результативный признак называют объясняемой переменной, а факторный признак называют объясняющей переменной.

Переменные, участвующие в эконометрической модели любого типа, подразделяются на:

- экзогенные (независимые) – значения которых задаются извне, автономно, в определенной степени они являются управляемыми (планируемыми) ($x$);
- эндогенные (зависимые) – значения которых определяются внутри модели, или взаимозависимые *(y);*
- лаговые – экзогенные или эндогенные переменные эконометрической модели, датированные предыдущими моментами времени и находящиеся в уравнении с текущими переменными. Например: $y_t$ – текущая эндогенная переменная, $y_{t-1}$ – лаговая эндогенная переменная, $y_{t-2}$ – тоже лаговая эндогенная переменная;

• предопределенные переменные (объясняющие переменные). К ним относятся лаговые и текущие экзогенные переменные ($x_t$, $x_{t-1}$,), а также лаговые эндогенные переменные ($y_{t-1}$).

Любая эконометрическая модель предназначена для объяснения значений текущих эндогенных переменных (од ной или нескольких) в зависимости от значений предопределенных переменных.

### *1.1.4. Выборка и генеральная совокупность*

Фундаментальными понятиями статистического анализа являются понятия вероятности и случайной величины (переменной). Случайной переменной мы называем переменную, которая под воздействием случайных факторов может с определенными вероятностями принимать те или иные значения из некоторого множества чисел. Это переменная, которой (даже при фиксированных обстоятельствах) мы не можем приписать определенное значение, но можем приписать несколько значений, которые она принимает с определенными вероятностями. Под вероятностью некоторого события (например, события, состоящего в том, что случайная переменная приняла определенное значение) обычно понимается доля числа исходов, благоприятствующих данному событию, в общем числе возможных равновероятных исходов. Категория "равновероятные исходы" не определяется, а принимается интуитивно. Например, при "бросании монеты" выпадение орла и решки считается равновероятным (вероятность каждого равна 1/2), а случайная величина числа "орлов" при одном "бросании монеты" может быть равна 0 или 1 с вероятностями 1/2.

Совокупность значений {хк} случайной величины х вероятностей {Рк}, с которыми она их принимает, называют законом распределения случайной величины. Функция Р{х}, как и любая функциональная зависимость, может быть представлена в форме таблицы, формулы или графика. Например, закон распределения числа очков при бросании игрального кубика может быть представлен в виде таблицы:

| X | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| p | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 |

Очевидно, что сумма всех этих вероятностей должна равняться единице, поскольку считаем, что с вероятностью "единица" переменная принимает хоть какое-нибудь из этих значений. Обычная (неслучайная, или детерминированная) переменная является предельным случаем случайной переменной, принимая единственное (при фиксированных обстоятельствах) значение с вероятностью "единица".

Различают дискретные и непрерывные случайные величины. Случайная величина дискретна, если результаты наблюдений представляют собой конечный или счетный набор возможных чисел. Случайная величина непрерывна, если ее значения могут лежать в некотором континууме возможных значений. (Это предполагает, что их нельзя пересчитать, ставя в соответствие им натуральные числа 1,2,....). Значения непрерывной случайной величины могут лежать на отрезке, интервале, луче и т. д.

В основе математической статистики лежат понятия генеральной совокупности и выборки (выборочной совокупности).

Под *генеральной совокупностью* мы подразумеваем все возможные наблюдения интересующего нас показателя, все исходы случайного испытания или всю совокупность реализаций случайной величины x. Пример генеральной совокупности – данные о доходах всех жителей какой-либо страны, о результатах голосования населения по какому-либо вопросу и т.д. Однако в большинстве случаев мы имеем дело только с частью возможных наблюдений, взятых из генеральной совокупности, и называем это множество (точнее подмножество) значений выборкой. Таким образом, **выборка** – это множество наблюдений, составляющих лишь часть генеральной совокупности. Выборка объема $n$ – это результат наблюдения случайной величины в вероятностном эксперименте, который повторяется $n$ раз в одних и тех же условиях (которые могут контролироваться), а, следовательно, и при неизменном распределении случайной величины $x$. Процесс, который приводит к получению выборочных данных, называют выборочным исследованием.

Мы обычно говорим о генеральной совокупности, когда используем определенные теоретические модели, но на практике в нашем распоряжении имеются лишь выборочные данные, и поэтому мы можем строить оценки теоретических характеристик, основываясь лишь на данных выборочных наблюдений. Мы обсудим соотношение между теоретическими характеристиками и их выборочными оценками позднее. Подчеркнем лишь, что целью математической статистики является получение выводов о параметрах, виде распределения и других свойствах случайных величин (генеральной совокупности) по конечной совокупности наблюдений – выборке.

Выборку называют репрезентативной (представительной), если она достаточно полно представляет изучаемые признаки и параметры генеральной совокупности. Для репрезентативности выборки важно обеспечить случайность отбора, с тем, чтобы все объекты генеральной совокупности имели равные вероятности попасть в выборку. Для обеспечения репрезентативности выборки применяют следующие способы отбора: простой отбор (последовательно отбирается первый случайно попавшийся объект), типический отбор (объекты отбираются пропорционально представительству различных типов объектов в генеральной совокупности), случайный отбор – например, с помощью таблицы случайных чисел и т.п.

Итак, **выборка** – некоторое количество наблюдений, отобранных из генеральной совокупности, а **наблюдение** – наблюдаемое значение случайной величины или набора случайных величин.

В эконометрике всегда известна только *выборка* из некоторого количества *наблюдений* случайной величины, и по данным *выборки* можно рассчитать только выборочные, а не теоретические характеристики этой случайной величины.

### *1.1.5. Выборочные и теоретические величины. Оценки как случайные величины. Оценки x и $S^2$*

*Математическое ожидание дискретной случайной величины* – это взвешенное среднее всех ее возможных значений, причем в качестве весового коэффициента берется вероятность соответствующего исхода. Вы можете рассчитать его, перемножив все возможные значения случайной величины на их вероятности и просуммировав полученные произведения. Математически, если случайная величина обозначена как *x*, то ее математическое ожидание обозначается как *M(x)*.

Предположим, что *x* может принимать *n* конкретных значений ($x_1$, $x_2$, ... , $x_n$) и что вероятность получения $x_j$ равна $p_i$ Тогда

$$M(x) = x_1 p_1 + x_2 p_2 + ... + x_n p_n = \sum_{i=1}^{n} x_i p_i \qquad (1.1)$$

Рассмотрим простой пример случайной переменной – число очков, выпадающее при бросании лишь одной игральной кости.

В данном случае возможны шесть исходов: $n_1 = 1$, $x_1 = 2$, $x_2 = 3$, $x_3 = 4$, $x_4 = 5$, $x_5 = 6$. Каждый исход имеет вероятность 1/6, поэтому здесь

$$M(x) = \sum_{i=1}^{6} x_i p_i = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3{,}5. \qquad (1.2)$$

В данном случае математическим ожиданием случайной переменной является число, которое само по себе не может быть получено при бросании кости.

Математическое ожидание случайной величины часто называют ее *средним по генеральной совокупности.* Для случайной величины *x* это значение часто обозначается как $\mu$.

Важной функцией переменной *x* является ее теоретическая дисперсия, которая характеризует меру разброса для вероятного распределения. Она определяется как *математическое ожидание* квадрата разности между величиной *x* и ее средним, т.е. величины $(x - \mu)^2$, где $\mu$ – математическое ожидание *x*. Дисперсия обычно обозначается как $\sigma_x^2$, и если ясно, о какой переменной идет речь, то нижний индекс может быть опущен.

Часто вместо рассмотрения случайной величины как единого целого можно и удобно разбить ее на постоянную и чисто случайную составляющие, где постоянная составляющая всегда есть ее математическое ожидание. Если *x* – случайная переменная и $\mu$ – ее математическое ожидание, то декомпозиция случайной величины записывается следующим образом:

$$x = \mu + u \qquad (1.3)$$

где *u – чисто случайная составляющая* (в регрессионном анализе она обычно представлена *случайным членом).*

Случайная составляющая и определяется как разность между *x* и $\mu$*:*

$$u = x - \mu \qquad (1.4)$$

Из определения следует, что математическое ожидание величины *u* равно нулю. Из уравнения (1.4) имеем:

$$M(u) = M(x-\mu) = M(x) - M(\mu) = M(x) - \mu = \mu - \mu = 0. \qquad (1.5)$$

Поскольку весь разброс значений *x* обусловлен *u,* неудивительно, что теоретическая дисперсия *x* равна теоретической дисперсии *u.* Последнее нетрудно доказать. По определению,

$$\sigma_x^2 = M\left\{(x-\mu)^2\right\} = M\left\{u^2\right\} \qquad (1.6)$$

Таким образом, $\sigma^2$ может быть эквивалентно определена как дисперсия *x* или *u.*

Обобщая, можно утверждать, что если *x* – случайная переменная, определенная по формуле (1.3), где $\mu$ – заданное число, *u* – случайный член с *M(u)=0* и дисперсией *D(u)*, то математическое ожидание величины *x* равно $\mu$, а дисперсия – $\sigma^2$.

До сих пор мы предполагали, что имеется точная информация о рассматриваемой случайной переменной, в частности – об ее распределении вероятностей (в случае дискретной переменной) или о функции плотности распределения (в случае непрерывной переменной). С помощью этой информации можно рассчитать теоретическое математическое ожидание, дисперсию и любые другие характеристики, в которых мы можем быть заинтересованы.

Однако на практике, за исключением искусственно простых случайных величин (таких, как число выпавших очков при бросании игральной кости), мы не знаем точного вероятностного распределения или плотности распределения вероятностей. Это означает, что неизвестны также и теоретическое математическое ожидание, и дисперсия. Мы, тем не менее, можем нуждаться в оценках этих или других теоретических характеристик генеральной совокупности.

Процедура оценивания всегда одинакова. Берется выборка из *n* наблюдений, и с помощью подходящей формулы рассчитывается оценка

нужной характеристики. Нужно следить за терминами, делая важное различие между способом или формулой оценивания и рассчитанным по ней для данной выборки числом, являющимся значением оценки.

**Оценка, способ оценивания (estimator)** – общее правило, формула для получения приближенного численного значения какого-либо параметра по данным выборки, а **значение оценки (estimation)** – число, полученное в результате применения оценки к конкретной выборке; является случайной величиной, значение которой зависит от выборки.

В табл. 1.1 приведены формулы оценивания для двух важнейших характеристик генеральной совокупности. *Выборочное среднее x* обычно дает оценку для математического ожидания, а формула $s^2$ в табл. 1.1 – оценку дисперсии генеральной совокупности.

Таблица 1.1

| Характеристики генеральной совокупности | Формулы оценивания |
|---|---|
| Среднее, $\mu$ | $\bar{x} = \frac{1}{n}\sum x_i$ |
| Дисперсия, $s^2$ | $s^2 = \frac{1}{n-1}\sum\left(x_i - \bar{x}\right)^2$ |

Отметим, что это *обычные* формулы оценки математического ожидания и дисперсии генеральной совокупности, однако не единственные. Конечно, не все формулы оценки, которые можно представить, одинаково хороши. Причина, по которой в действительности используется *x*, в том, что эта оценка в наилучшей степени соответствует двум очень важным критериям – несмещенности и эффективности. Эти критерии будут рассмотрены ниже.

Получаемая *оценка* представляет частный случай случайной переменной. Причина здесь в том, что сочетание значений *x* в выборке случайно, поскольку *x* – случайная переменная *u*, следовательно, случайной величиной является и функция набора ее значений. Возьмем, например, $\bar{x}$ – оценку математического ожидания:

$$\bar{x} = \frac{1}{n}\left(x_1 + x_2 + \ldots + x_n\right) \qquad (1.7)$$

Мы только что показали, что величина *x* в *i*-м наблюдении может быть разложена на две составляющие: постоянную часть $\mu$ и чисто случайную составляющую $u_i$

$$x_i = \mu + u_i \qquad (1.8)$$

Следовательно,

$$\bar{x} = \mu + \bar{u} \qquad (1.9)$$

где $\bar{u}$ – выборочное среднее величин $u_i$

Отсюда можно видеть, что $\bar{x}$, подобно *x,* имеет как фиксированную, так и чисто случайную составляющие. Ее фиксированная составляющая $\mu$, то есть математическое ожидание *x*, а ее случайная составляющая $\bar{u}$, то есть среднее значение чисто случайной составляющей в выборке.

Функции плотности вероятности для *x* и $\bar{x}$ показаны на одинаковых графиках (рис. 1.1). Как показано на рисунке, величина *x* считается нормально распределенной. Можно видеть, что распределения, как x, так и $\bar{x}$, симметричны относительно $\mu$ – теоретического среднего. Разница между ними в том, что распределение $\bar{x}$ уже и выше. Величина $\bar{x}$, вероятно, должна быть ближе к $\mu$, чем значение единичного наблюдения *x*, поскольку ее случайная составляющая $\bar{u}$ есть среднее от чисто случайных составляющих $u_1, u_2, \ldots, u_n$ в выборке, которые, по-видимому, "гасят" друг друга при расчете среднего. Далее, теоретическая дисперсия величины $\bar{u}$ составляет лишь часть теоретической дисперсии *u.*

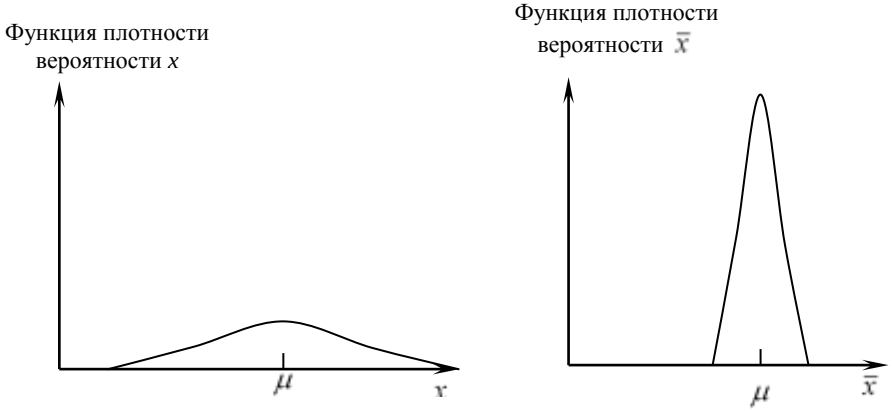


Рис. 1.1. Сравнение функций плотности вероятности одиночного наблюдения и выборочного среднего

Величина $s^2$ – оценка теоретической дисперсии *x* – также является случайной переменной. Вычитая (1.9) из (1.8), имеем:

$$x_i - \bar{x} = u_i - \bar{u} \qquad (1.10)$$

Следовательно,

$$s^2 = \frac{1}{n-1}\sum\left\{\left(x_i - \bar{x}\right)^2\right\} = \frac{1}{n-1}\sum\left\{\left(u_i - \bar{u}\right)^2\right\} \qquad (1.11)$$

Таким образом, $s^2$ зависит от (и только от) чисто случайной составляющей наблюдений $x$ в выборке. Поскольку эти составляющие меняются от выборки к выборке, также от выборки к выборке меняется и величина оценки $s^2$.

### *1.1.6. Выборочная ковариация и выборочная дисперсия*

*Выборочная ковариация* является мерой взаимосвязи между двумя переменными. Данное понятие будет проиллюстрировано на простом примере.

Со времен нефтяного кризиса 1973 г. реальная цена на бензин, т. е. цена бензина, отнесенная к уровню общей инфляции, значительно возросла, и это оказало заметное воздействие на потребительский спрос.

Таблица 1.2

Потребительские расходы на бензин и его реальная цена в США

| Год | Расходы (млрд. долл., цены 1972 г.) | Индекс реальных цен (1972=100) |
|---|---|---|
| 1 | 26,2 | 103,5 |
| 1 | 24,8 | 127,0 |
| 1 | 25,6 | 126,0 |
| 1 | 26,8 | 124,8 |
| 1 | 27,7 | 124,7 |
| 1 | 28,3 | 121,6 |
| 1 | 27,4 | 149,7 |
| 1 | 25,1 | 188,8 |
| 1 | 25,2 | 193,6 |
| 1 | 25,6 | 173,9 |

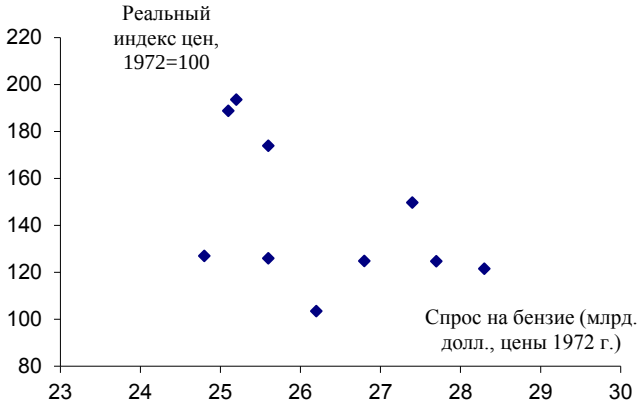


Рис. 1.2. Спрос на бензин в США, 1973 - 1982 гг.

На рис. 1.2 эти данные показаны в виде диаграммы рассеяния. Можно видеть некоторую отрицательную связь между потребительским спросом на бензин и его реальной ценой.

Показатель выборочной ковариации позволяет выразить данную связь единым числом. Для его вычисления мы сначала находим средние (для рассматриваемого выборочного периода) значения цены и спроса на бензин. Обозначив цену через *p* спрос – через *y*, мы, таким образом, определяем $\bar{p}$ и $\bar{y}$ , которые для этой выборки оказываются равными соответственно 143,36 и 26,27. Затем для каждого года вычисляем отклонение величин *p* и *y* от средних и перемножаем их. Для первого года $(p-\bar{p})$ равно (103,5 – 143,36), или – 39,86, и $(y-\bar{y})$ равно (26,2 – 26,27), или –0,07, а произведение $(p-\bar{p})(y-\bar{y})$ составит 2,79. Проделаем это для всех годов выборки и возьмем среднюю величину, она и будет выборочной ковариацией (как видно, не очень сложно вычисляемой).

При наличии *n* наблюдений двух переменных *(x* и *y)* выборочная ковариация между *x* и *y* задается формулой:

$$Cov(x,y)=\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\{(x_1-\bar{x})(y_1-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y})\}$$

Для различения ковариационной генеральной и выборочной совокупности используем обозначение $Cov(x,y)$ с прописной буквы применительно к выборочной ковариации и $cov(x,y)$ или $\sigma_{xy}$ для ковариации между *x* и *y* в генеральной совокупности.

Аналогичные обозначения используются для дисперсии $Var(x)$ – применительно к выборочной дисперсии и $var(x)$ – к дисперсии для генеральной совокупности.

В примере с бензином детали проведенных вычислений для всей выборки приведены в табл. 1.3. Здесь в столбцах 2 и 3 представлены исходные данные для $\bar{p}$ и $\bar{y}$. В результирующих строках вычисляются *p* и *y*. В столбцах 4 и 5 рассчитываются $(p-\bar{p})$ и $(y-\bar{y})$ для каждого года, а в столбце $\sigma$ эти две величины перемножаются. В нижней клетке последнего столбца определяется средняя величина (–16,24), она и является значением выборочной ковариации.

Таблица 1.3

| Наблюдение | $p$ | $y$ | $(p-\bar{p})$ | $(y-\bar{y})$ | $(p-\bar{p})(y-\bar{y})$ |
|---|---|---|---|---|---|
| 1973 | 103,5 | 26,2 | -39,86 | -0,07 | 2,79 |
| 1974 | 127 | 24,8 | -16,36 | -1,47 | 24,05 |
| 1975 | 126 | 25,6 | -17,36 | -0,67 | 11,63 |
| 1976 | 124,8 | 26,8 | -18,56 | 0,53 | -9,84 |
| 1977 | 124,7 | 27,7 | -18,66 | 1,43 | -26,68 |
| 1978 | 121,6 | 28,3 | -21,76 | 2,03 | -44,17 |
| 1979 | 149,7 | 27,4 | 6,34 | 1,13 | 7,16 |
| 1980 | 188,8 | 25,1 | 45,44 | -1,17 | -53,16 |
| 1981 | 193,6 | 25,2 | 50,24 | -1,07 | -53,76 |
| 1982 | 173,9 | 25,6 | 30,54 | -0,67 | -20,46 |
| Сумма | 1433,6 | 262,7 | | | -162,44 |
| Среднее | 143,36 | 26,27 | | | -16,24 |

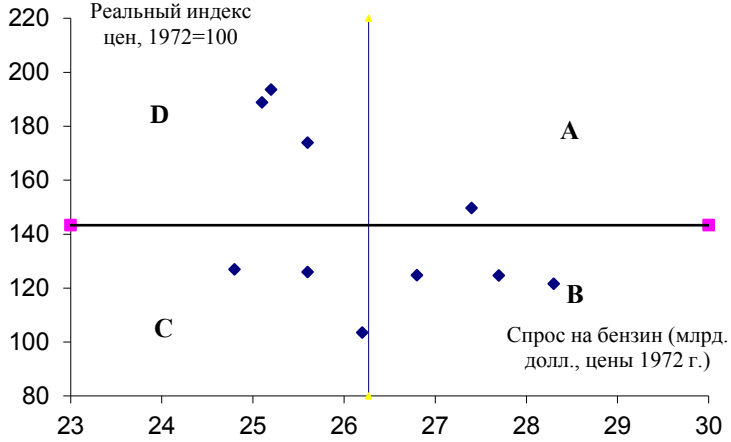


Рис. 1.3.

Ковариация в данном случае отрицательна. Так это и должно быть. Отрицательная связь, как это имеет место в данном примере, выражается отрицательной ковариацией, а положительная связь – положительной ковариацией.

Имеет смысл рассмотреть причину этого. Рисунок 1.3 точно такой же, как и рис. 1.2, но здесь диаграмма рассеяния наблюдений делится на четыре части вертикальной и горизонтальной линиями, проведенными через $\overline{p}$ и $\overline{y}$ соответственно. Пересечение этих линий; образует точку $(\overline{p},\overline{y})$, которая показывает среднюю цену и средний спрос за период времени, соответствующий нашей выборке. Используя аналогию из физики, можно сказать, что эта точка является центром тяжести совокупности точек, представляющих наблюдения.

Для любого наблюдения, лежащего в квадранте **A**, значения реальной цены и спроса выше соответствующих средних значений. Для данных наблюдений как $(p-\overline{p})$, так и $(y-\overline{y})$ являются положительными, а поэтому должно быть положительным и $(p-\overline{p})(y-\overline{y})$. Наблюдение, таким образом, дает положительный вклад в ковариацию. Так, например, наблюдение за 1979г. лежит в этом квадранте и $(p-\overline{p})$=6,34, и $(y-\overline{y})$=1,13, а их произведение равно 7,16. Далее рассмотрим квадрант **B**. Здесь наблюдения имеют реальную цену ниже средней и спрос выше среднего. Поэтому $(p-\overline{p})$ отрицательно, $(y-\overline{y})$ положительно, произведение $(p-\overline{p})(y-\overline{y})$ отрицательно, и наблюдение вносит отрицательный вклад в ковариацию. Например, наблюдение за 1978 г. Имеет $(p-\overline{p})$=–21,76, $(y-\overline{y})$=2,03 и $(p-\overline{p})(y-\overline{y})$, таким образом, равно –44,17.

В квадранте **C** как реальная цена, так и спрос ниже своих средних значений. Таким образом, $(p-\overline{p})$ и $(y-\overline{y})$ оба являются отрицательным, а $(p-\overline{p})(y-\overline{y})$ положительно. (В качестве примера см. наблюдение за 1974 г.)

Наконец, в квадранте **D** реальная цена выше средней, а спрос ниже среднего. Таким образом, $(p-\overline{p})$ положительно, $(y-\overline{y})$ отрицательно, поэтому $(p-\overline{p})(y-\overline{y})$ отрицательно, и в ковариацию, соответственно, вносится отрицательный вклад. (В качестве примера см. наблюдение за 1981 г.)

Поскольку выборочная ковариация является средней величиной произведения $(p-\overline{p})(y-\overline{y})$ для 20 наблюдений, она будет положительной, если положительные вклады будут доминировать над отрицательными, и отрицательной, если будут доминировать отрицательные вклады. Положительные вклады исходят из квадрантов **A** и **C**, и ковариация будет, скорее всего, положительной, если основной разброс пойдет по наклонной вверх. Точно так же отрицательные вклады исходят из квадрантов **B** и **D**. Поэтому если основное рассеяние идет по наклонной вниз, как в данном примере, то ковариация будет, скорее всего, отрицательной.

*Правило 1*

Если $y = v + w$, то $Cov(x,y) = Cov(x,v) + Cov(x,w)$.

*Правило2*

Если $y = ar$, где $a$ – константа, то $Cov(x,y) = aCov(x,r)$.

*Правило 3*

Если $y = a$, где $a$ – константа, то $Cov(x,y) = 0$.

Если $x$ и $y$ – случайные величины, то *теоретическая ковариация* $\sigma_{xy}$ определяется как математическое ожидание произведения отклонений этих величин от их средних значений:

$$cov(x,y) = \sigma_{xy} = M\{(x-\mu_x)(y-\mu_y)\} \qquad (1.12)$$

где $\mu_x$ и $\mu_y$ –теоретические средние значения $x$ и $y$ соответственно.

Как вы и ожидаете, если теоретическая ковариация неизвестна, то для ее оценки может быть использована выборочная ковариация, вычисленная по ряду наблюдений. К сожалению, оценка будет иметь отрицательное смещение, так как

$$M\{Cov(x,y)\} = \frac{n-1}{n} cov(x,y) \qquad (1.13)$$

Причина заключается в том, что выборочные отклонения измеряются по отношению к выборочным средним значениям величин $x$ и $y$ и имеют тенденцию к занижению отклонений от истинных средних значений. Очевидно, мы можем рассчитать несмещенную оценку путем умножения выборочной оценки на $n/(n-1)$.

Если $x$ и $y$ независимы, то их теоретическая ковариация равна нулю, поскольку $M\{(x-\mu_x)(y-\mu_y)\} = M(x-\mu_x)\cdot M(y-\mu_y) = 0\cdot 0$ благодаря свойству и факту, что *M(x)* и *M(y)* равняются соответственно $\mu_x$ и $\mu_y$.

До сих пор термин "дисперсия" использовался в смысле теоретической дисперсии (т.е. относящейся ко всей генеральной совокупности). Для целей, которые прояснятся при обсуждений регрессионного анализа, целесообразно ввести понятие *выборочной дисперсии,* Для выборки из $n$ наблюдений ($x_1$, $x_2$, ... , $x_n$) выборочная дисперсия определяется как среднеквадратичное отклонение в выборке:

$$Var(x) = \frac{1}{n}\sum (x_i - \bar{x})^2 \qquad (1.14)$$

Сделаем следующие замечания:

1. Определенная таким образом выборочная дисперсия представляет собой смещенную оценку теоретической дисперсии, что $s^2$, определенная как

$$\frac{1}{n-1}\sum(x_i-\bar{x})^2$$

является несмещенной оценкой $\sigma^2$. Отсюда следует, что ожидаемое значение величины $Var(x)$ равно $\frac{n-1}{n}\sigma^2$ и что, следовательно, она; имеет отрицательное смещение. Отметим, что если размер выборки *n* становится большим, то $\frac{n-1}{n}$ стремится к единице и, таким образом, математическое ожидание величины $Var(x)$ стремится к $\sigma^2$. Можно легко показать, что ее предел по вероятности (*plim*) равен $\sigma^2$ и, следовательно, она является примером состоятельной оценки, которая смещена для небольших выборок.

2. Так как величина $s^2$ является несмещенной, то в некоторых работах ее часто определяют как выборочную дисперсию и либо избегают ссылок на $Var(x)$, либо дают ей какое-то другое название.

Почему выборочная дисперсия в среднем занижает значение теоретической дисперсии? Причина заключается в том, что она вычисляется как среднеквадратичное отклонение от выборочного среднего, а не от истинного значения. Так как выборочное среднее автоматически находится в центре выборки, то отклонения от него в среднем меньше отклонений от теоретического среднего значения.

Подводя итог вышеизложенному, повторим, что выборочная ковариация

$$Cov(x,y)=\frac{1}{n}\sum(x_i-\bar{x})(y_i-\bar{y})=\frac{1}{n}\{(x_1-\bar{x})(y_1-\bar{y})+...+(x_n-\bar{x})(y_n-\bar{y})\}$$

и выборочная дисперсия

$$Var(x)=\frac{1}{n}\sum(x_i-\bar{x})^2=\frac{1}{n}\{(x_1-\bar{x})^2+(x_2-\bar{x})^2+...+(x_n-\bar{x})^2\}$$

смещенные оценки теоретической ковариации $cov(x,y)$ и дисперсии $s^2(x)$, соответственно. Они имеют отрицательное смещение – то есть в большинстве случаев дают значения оценок, меньшие теоретических величин $cov(x,y)$ и $s^2(x)$, соответственно. Несмещенной оценкой теоретической ковариации является оценка $\frac{n}{n-1}Cov(x,y)$, несмещенной оценкой дисперсии – оценка $s^2=\frac{n}{n-1}Var(x)$.

Существует несколько простых и очень полезных правил для расчета дисперсии, являющихся аналогами правил для ковариации. Эти правила в равной степени можно использовать как для выборочной, так и для теоретической дисперсии.

*Правило дисперсии 1*

Если $y = v + w$, то $Var(y) = Var(v) + Var(w) + 2Cov(v,w)$.

*Правило дисперсии 2*

Если $y = az$, где $a$ – константа, то $Var(y) = a^2 Var(r)$.

*Правило дисперсии 3*

Если $y = a$, где $a$ – константа, то $Var(y) = 0$.

*Правило дисперсии 4*

Если $y = v + a$, где $a$ – константа, то $Var(y) = Var(v)$.

Во-первых, заметим, что дисперсия переменной *x* может рассматриваться как ковариация между двумя величинами *x*:

$$Var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = Cov(x,x) \qquad (1.15)$$

Учитывая это равенство, мы можем воспользоваться правилами расчета выборочной ковариации, чтобы вывести правила расчета дисперсии. Кроме того,

$$Var(x) = \left[\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}\right] - \bar{x}^2 \qquad (1.16)$$

Ковариация весьма удобна с математической точки зрения, что является особенно хорошим измерителем взаимосвязи между величинами. Более точной мерой зависимости является тесно связанный с ней *коэффициент корреляции.*

Подобно дисперсии и ковариации, коэффициент корреляции имеет две формы – теоретическую и выборочную. *Теоретический коэффициент корреляции* традиционно обозначается греческой буквой $\rho$. Для переменных *x* и *y* этот коэффициент определяется следующим образом:

$$\rho_{x,y} = \frac{cov(x,y)}{\sqrt{var(x)var(y)}} = \frac{\sigma_{x,y}}{\sqrt{\sigma_x^2 \sigma_y^2}}. \qquad (1.17)$$

Если *x* и *y* независимы, то $\rho$ равно нулю, так как равна нулю теоретическая ковариация. Если между переменными существует положительная зависимость, то $\sigma_{x,y}$, а следовательно, и $\rho_{x,y}$, будут положительными. Если существует строгая положительная линейная зависимость, то $\rho_{x,y}$ примет максимальное значение, равное 1. Аналогичным

образом при отрицательной зависимости $\rho_{x,y}$ будет отрицательным с минимальным значением –1.

*Выборочный коэффициент корреляции* $r$ определяется путем замены теоретических дисперсий и ковариации в выражении (1.17) на их несмещенные оценки. Мы показали, что такие оценки могут быть получены умножением выборочных дисперсий и ковариации на $n/(n-1)$. Следовательно,

$$r_{x,y} = \frac{\frac{n}{n-1}Cov(x,y)}{\sqrt{\frac{n}{n-1}Var(x)\frac{n}{n-1}Var(y)}}. \qquad (1.18)$$

Множители $n/(n-1)$ сокращаются, поэтому можно определить выборочную корреляцию как

$$r_{x,y} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}. \qquad (1.19)$$

Подобно величине $\rho$, $r$ имеет максимальное значение, равное единице, которое получается при строгой линейной положительной зависимости между выборочными значениями $x$ и $y$ (когда на диаграмме рассеяния все точки находятся точно на восходящей прямой линии). Аналогичным образом $r$ принимает минимальное значение –1, когда существует линейная отрицательная зависимость (точки лежат точно на нисходящей прямой линии). Величина $r = 0$ показывает, что зависимость между наблюдениями $x$ и $y$ в выборке отсутствует. Разумеется, тот факт, что $r = 0$, необязательно означает, что $\rho = 0$ и наоборот.

Выборочная корреляция $r_{x,y} = \frac{Cov(x,y)}{\sqrt{Var(x)Var(y)}}$ – *несмещенная оценка* теоретической корреляции $\rho_{x,y} = \frac{cov(x,y)}{\sqrt{\sigma^2(x)\sigma^2(y)}}$.

## 1.2. Обобщенная линейная регрессионная модель

### *1.2.1. Двумерная (однофакторная) регрессионная модель*

Сформулируем регрессионную проблему для случая одного факторного признака.

Пусть имеется набор значений двух переменных: $y_j$ (объясняемая переменная или результат) и $x_i$ (объясняющая переменная или фактор). Между этими переменными имеется объективная связь:

$$y = f(x) \qquad (1.20)$$

Данное уравнение будем называть «истинным» уравнением регрессии. Необходимо по данным наблюдений ($y_i$, $x_i$, i= 1; n) подобрать функцию: $\hat{y} = f(x)$ «наилучшим» образом описывающую «истинную» зависимость (1.20). Подобрать функцию – значит определить вид функциональной зависимости и значения параметров.

Для определения вида функциональной зависимости можно использовать:

1) теоретические соображения и опыт предыдущих аналогичных исследований;

2) графический способ – на основе корреляционного поля или эмпирической линии регрессии. Корреляционное поле – точечный график в системе координат (*x, y*). Каждая точка соответствует единице наблюдения. Положение каждой точки на графике определяется величиной двух признаков – факторного – *x* и результативного – *y*. Эмпирическая регрессия – регрессия, полученная по эмпирическим (наблюдаемым) данным. Используются результаты аналитической, либо комбинационной группировки. Графически она представляет собой ломаную линию, составленную из точек, абсциссами которых являются средние значения факторного признака, а ординатами – средние значения признака-результата. Число точек равно числу групп в группировке;

3) можно также перебрать несколько функций (построить для каждой из них уравнение регрессии) и выбрать лучшую из них по показателям качества уравнения регрессии.

Наиболее часто используется линейная форма зависимости. Внимание к линейной форме объясняется четкой экономической интерпретацией ее параметров, ограниченной вариацией переменных и тем, что в большинстве случаев нелинейные формы связи для выполнения расчетов преобразуют в линейную форму.

Модель линейной двумерной (однофакторной или парной) регрессии имеет вид:

$$y_i = b_0 + b_1 \cdot x_i + u_i \qquad (1.21)$$

Величина переменной $y_j$ состоит из двух составляющих:

1) неслучайной составляющей $b_0 + b_1 \cdot x_i$;

2) случайной составляющей $u_i$.

На рис. 1.4 показано, как комбинация этих двух составляющих определяет величину $y_i$ для случая парной линейной модели регрессии.

Причины существования случайной составляющей $u_i$:

1) отсутствие в модели «важных» факторов, оказывающих существенное влияние на результат. Парная регрессия почти всегда является большим упрощением. В действительности существуют другие факторы, которые не учтены в формуле (2). Это могут быть факторы, которые мы не можем изме-

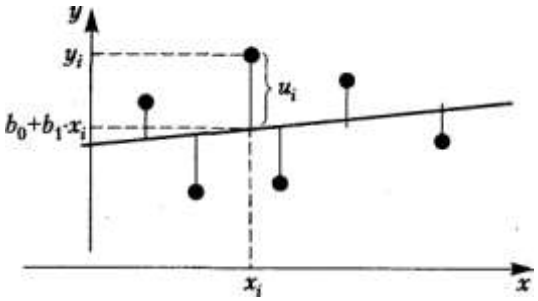


Рис. 1.4. Истинная зависимость между *y* и *x*

рить (например, психологические). Возможно, это факторы, которые мы можем измерить, но которые оказывают очень слабое влияние на результат, и поэтому мы их не учитываем в модели. Кроме того, это могут быть «важные» факторы, которые мы такими не считаем из-за отсутствия опыта. Все это приводит к тому, что наблюдаемые значения лежат вне прямой ($b_0 + b_1 \cdot x_i$);

2) агрегирование переменных. Мы можем пытаться построить зависимость путем агрегирования (объединения) индивидуальных соотношений. Например, функцию суммарного потребления как агрегирование функций потребления по отдельным потребителям. Так как параметры индивидуальных соотношений различны, то агрегированная зависимость будет приближенной;

3) неправильная функциональная спецификация мо дели;

4) ошибки измерения переменных.

Знак коэффициента регрессии $b_1$ в модели указывает направление связи (если $b_1 > 0$, связь прямая, если $b_1 < 0$, то связь обратная). Величина $b_1$ показывает, на какую величину в среднем изменится результат *y*, если фактор *x* увеличить на одну единицу своего измерения.

Формально значение параметра $b_0$ в модели – среднее значение *y* при *x* = 0. Если фактор не имеет и не может иметь нулевого значения, то вышеуказанная трактовка параметра не имеет смысла.

В матричной форме двумерная регрессионная модель имеет вид:

$$Y = X \cdot b + u$$

где Y – случайный вектор-столбец размерности (n×1) наблюдаемых значений результативного признака;

X = ($x_0$, $x_1$,) — матрица размерности (n×2) наблюдаемых значений факторных признаков. Дополнительный фактор $x_0$ связан с наличием в уравнении регрессии свободного члена ($b_o$). Значение фактора ($x_0$) для свободного члена принято считать равным единице;

*b* – вектор-столбец размерности (n×1) неизвестных, подлежащих оценке параметров модели (коэффициентов регрессии);

u – случайный вектор-столбец размерности (n×1) ошибок наблюдений.

***Рассмотрим пример***. Пусть имеются данные о заработной плате и возрасте по 20 рабочим. Требуется построить регрессионную модель заработной платы рабочего. Тогда $y_i$ — заработная плата *i-го* рабочего ($); $x_i$ – возраст *i-го* рабочего (лет), i=l; 20. Исходные данные приведены в табл. 1.4.

Таблица 1.4

| $i$ | $y_i$ | $x_i$ | $i$ | $y_i$ | $x_i$ |
|---|---|---|---|---|---|
| 1 | 300 | 29 | 11 | 400 | 47 |
| 2 | 400 | 40 | 12 | 250 | 28 |
| 3 | 300 | 36 | 13 | 350 | 30 |
| 4 | 320 | 32 | 14 | 200 | 25 |
| 5 | 200 | 23 | 15 | 400 | 48 |
| 6 | 350 | 45 | 16 | 220 | 30 |
| 7 | 350 | 38 | 17 | 320 | 40 |
| 8 | 400 | 40 | 18 | 390 | 40 |
| 9 | 380 | 50 | 19 | 360 | 38 |
| 10 | 400 | 47 | 20 | 260 | 29 |

Для нашего примера параметры линейной парной модели регрессии (1.21) интерпретируются следующим образом. Параметр $b_1$ показывает, на сколько долларов в среднем изменится заработная плата рабочего при увеличении возраста на 1 год. Параметр $b_0$ не интерпретируется, т. к. возраст рабочего не может быть равен 0 лет.

В матричной форме регрессионная модель имеет вид: $Y = X \cdot b + u$

$$Y = \begin{pmatrix} 300 \\ 400 \\ 300 \\ 320 \\ 200 \\ 350 \\ 350 \\ 400 \\ 380 \\ 400 \\ 400 \\ 250 \\ 350 \\ 200 \\ 400 \\ 220 \\ 320 \\ 390 \\ 360 \\ 260 \end{pmatrix} \quad X = \begin{pmatrix} 129 \\ 140 \\ 136 \\ 132 \\ 123 \\ 145 \\ 138 \\ 140 \\ 150 \\ 147 \\ 147 \\ 128 \\ 130 \\ 125 \\ 148 \\ 130 \\ 140 \\ 140 \\ 138 \\ 129 \end{pmatrix} \quad b = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix} \quad u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \\ u_8 \\ u_9 \\ u_{10} \\ u_{11} \\ u_{12} \\ u_{13} \\ u_{14} \\ u_{15} \\ u_{16} \\ u_{17} \\ u_{18} \\ u_{19} \\ u_{20} \end{pmatrix}$$

### *1.2.2. Нормальная линейная регрессионная модель с одной переменной*

Нормальная (классическая или традиционная) линейная регрессионная модель с одной переменной (Classical Normal Regression model) — это модель вида: $y_i = b_0 + b_1 \cdot x_i + u_i$ *(i = 1; n),* для которой выполняются следующие условия (предпосылки):

1) $x_i$ – детерминированная (неслучайная, нестохастическая) величина;

2) $M(u_i) = 0$, (i = 1; n) – математическое ожидание случайной составляющей; равно 0 в любом наблюдении;

3) $\sigma_{ui}^2 = M\left(\left(u_i - \overline{u_i}\right)^2\right) = M\left(u_i^2\right) = \sigma_u^2 = \text{const}$ *(i = 1; n)* – теоретическая дисперсия случайной составляющей; постоянна для всех наблюдений;

4) $\text{Cov}(u_i, u_j) = M\left(\left(u_i - \overline{u_i}\right) \cdot \left(u_j - \overline{u_j}\right)\right) = M(u_i, u_j) = 0$

*(i # j)* – отсутствие систематической связи между значениями случайной составляющей в любых двух наблюдениях (ковариация случайных составляющих в любых двух разных наблюдениях равна нулю);

5)часто добавляется условие: $u_i \approx N\left(0, \sigma_u^2\right)$, т. е. $u_i$ – нормально распределенная случайная величина.

Первая предпосылка – сильное предположение. Иногда достаточно сделать предположение о независимости распределения случайной составляющей $u_i$ от $x_i$.

Третья предпосылка – о независимости дисперсии случайной составляющей от наблюдения называется гомоскедастичностью (homoscedasticity в переводе означает одинаковый разброс). Случай, когда условие гомоскедастичности не выполняется, называется гетероскедастичностью (heteroscedasticity – неодинаковый разброс).

Четвертая предпосылка указывает на некоррелированность случайных составляющих для разных наблюдений. Это условие нарушается в случае, когда данные являются временными рядами. В случае, когда это условие не выполняется, говорят об автокорреляции случайных составляющих.

Запишем теперь матричную форму нормальной линейной модели парной регрессии.

$$Y = X \cdot b + u$$

Предпосылки модели:

1) $x$ — детерминированная (нестохастическая) переменная;

2) $M(u) = 0^n = \begin{pmatrix} 0 \\ 0 \\ \dots \\ 0 \end{pmatrix}$

3, 4) третья и четвертая предпосылка в матричной форме могут быть сформулированы через ковариационную матрицу случайных составляющих

Элементами ковариационной матрицы являются показатели ковариации. В общем случае ковариация – $cov(x,y)$ представляет собой показатель тесноты связи между признаками *x* и *y,* вычисляемый как среднее из произведений отклонений признаков от их средних значений:

$$Cov(x,y) = \overline{(x-\bar{x})(y-\bar{y})} = \overline{x \cdot y} - \bar{x} \cdot \bar{y}$$

Свойства ковариации:

а) Ковариация переменной и константы равна 0: *Cov(x, A) = 0 (A =* const);

б) Ковариация переменной с самой собой равна дисперсии переменной: $\mathrm{cov}(\mathrm{x},\mathrm{x}) = \sigma_{\mathrm{x}}^2$.

Для нормальной линейной модели регрессии ковариационная матрица должна иметь вид

$$C_u = \begin{pmatrix} \sigma_u^2 0...0 \\ 0\sigma_u^2 ...0 \\ \cdots\cdots\cdots \\ 00...\sigma_u^2 \end{pmatrix} = \sigma_u^2 \cdot \begin{pmatrix} 10\ldots0 \\ 01\ldots0 \\ \cdots\cdots\cdots \\ 00\ldots1 \end{pmatrix} = \sigma_u^2 \cdot I$$

где $\sigma_{\mathrm{u}}^2$ – дисперсия случайной составляющей; I – единичная матрица размером *n* × *n.*

5) $\mathrm{u} \approx \mathrm{N}(0,\sigma_{\mathrm{u}}^2)$.

### *1.2.3. Традиционный метод наименьших квадратов – МНК (OLS)*

После определения вида функциональной зависимости – *y = f(x)* оценивают параметры модели. Для определения «наилучших» параметров модели можно использовать следующие критерии:

1) сумму квадратов отклонений наблюдаемых значений зависимой переменной *y* от значений $\hat{y}$, рассчитанных по функции $\mathrm{f}(\mathrm{x})$: $S = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ – метод наименьших квадратов (МНК);

2) сумму модулей отклонений наблюдаемых значений зависимой переменной от ее расчетных значений: $S = \sum_{i=1}^{n} |y_i - \hat{y}_i|$;

3) $S = \sum_{i=1}^{n} g(y_i - \hat{y}_i)$, где g – «мера», с которой отклонение для i-го наблюдения входит в функционал.

Оптимальными будут значения параметров, минимизирующие функционал S.

Для оценки параметров $b_j (j = 0; 1)$ модели линейной пар ной регрессии: $y_i = b_0 + b_1 \cdot x_i + u_i$ (i = 1; *n)* наиболее часто используется традиционный (обычный) метод наименьших квадратов, согласно которому в качестве оценок параметров принимают величины $\tilde{b}_j$ *(j* = 0; 1), минимизирующие сумму квадратов отклонении наблюдаемых значений результативного признака — $y_i$ от расчетных (теоретических) значений – $\hat{y}_1 = \tilde{b}_0 + \tilde{b}_1 \cdot x_1$:

$$S = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \sum_{i=1}^{n} \left(y_i - \left(\tilde{b}_0 + \tilde{b}_1 \cdot x_1\right)\right)^2 \Rightarrow \min_{\tilde{b}_0, \tilde{b}_1}$$

Значения $y_i$ и $x_i$ (*i* = 1; *n)* нам известны, это данные наблюдений. В функции S они представляют собой константы. Переменными в данной функции являются оценки параметров $\tilde{b}_j$ *(j* = 0; 1). Чтобы найти минимум функции двух переменных, необходимо вычислить частные производные данной функции по каждому из параметров и приравнять их к нулю, т.е.

$$\frac{\partial S}{\partial \tilde{b}_0} = 0, \quad \frac{\partial S}{\partial \tilde{b}_1} = 0.$$

В результате получим систему из двух нормальных линейных уравнений:

$$\begin{cases} \sum_{i=1}^{n} y_i = \tilde{b}_0 \cdot n + \tilde{b}_1 \sum_{i=1}^{n} x_i; \\ \sum_{i=1}^{n} y_i x_i = \tilde{b}_0 \sum_{i=1}^{n} x_i + \tilde{b}_1 \sum_{i=1}^{n} x_i^2. \end{cases}$$

Решая данную систему, найдем искомые оценки пара метров:

$$\tilde{b}_1 = \frac{n\sum x_i y_i - \left(\sum x_i\right)\left(\sum y_i\right)}{n\sum x_i^2 - \left(\sum x_i\right)^2} =$$

$$= \frac{\sum x_i y_i - n\bar{x} \cdot \bar{y}}{\sum x_i^2 - n\left(\bar{x}\right)^2} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x^2} = \frac{Cov(x,y)}{\sigma_x^2}$$

$$\tilde{b}_0 = \bar{y} - \tilde{b}_1 \cdot x_1$$

где $\sigma_x^2$ – дисперсия факторного признака;

$\bar{y}$ – среднее значение результативного признака;

$\bar{x}$ – среднее значение факторного признака;

$\overline{x \cdot y}$ – среднее значение произведения фактора на результат.

Правильность расчета параметров уравнения регрессии может быть проверена сравнением сумм $\sum y = \sum \hat{y}$ (при этом, возможно некоторое расхождение из-за округления расчетов).

Результаты многих исследований подтверждают, что число наблюдений должно в 6-7 раз превышать число рассчитываемых параметров при переменной x. Это означает, что искать линейную регрессию, имея менее 7 наблюдений, вообще не имеет смысла.

***Рассмотрим пример***: по данным о заработной плате и возрасте 10 рабочих (см. табл. 1.5) оценить параметры линейной парной регрессии методом наименьших квадратов.

Расчет оценки коэффициента регрессии $b_1$ сведем в табл. 1.5.

Таблица 1.5

| № наблюдения | X – возраст рабочего, лет | Y – заработная плата за месяц, $ | $x \cdot y$ | $\left(x - \bar{x}\right)^2$ |
|---|---|---|---|---|
| 1 | 29 | 300 | 8700 | 44,22 |
| 2 | 40 | 400 | 16000 | 18,92 |
| 3 | 36 | 300 | 10800 | 0,12 |
| 4 | 32 | 320 | 10240 | 13,32 |
| 5 | 23 | 200 | 4600 | 160,02 |
| 6 | 45 | 350 | 15750 | 87,42 |
| 7 | 38 | 350 | 13300 | 5,52 |
| 8 | 40 | 400 | 16000 | 18,92 |
| 9 | 50 | 380 | 19000 | 205,92 |
| 10 | 47 | 400 | 18800 | 128,82 |
| 11 | 28 | 250 | 7000 | 58,52 |
| 12 | 30 | 350 | 10500 | 31,92 |
| 13 | 25 | 200 | 5000 | 113,42 |
| 14 | 48 | 400 | 19200 | 152,52 |
| 15 | 30 | 220 | 6600 | 31,92 |
| 16 | 40 | 320 | 12800 | 18,92 |
| 17 | 40 | 390 | 15600 | 18,92 |
| 18 | 38 | 360 | 13680 | 5,52 |
| 19 | 29 | 260 | 7540 | 44,22 |
| 20 | 25 | 250 | 6250 | 113,42 |
| $\sum$ | 713 | 6400 | 237360 | 1272,55 |
| Среднее значение | 35,65 | 320 | 11868 | 63,63 |

$$\tilde{b}_1 = \frac{11868 - 35{,}65 \cdot 320}{63{,}63} = 7{,}23$$

$$\tilde{b}_0 = 320 - 7{,}23 \cdot 35{,}65 = 62{,}27$$

Тогда линейная парная регрессии, описывающая зависимость заработной платы от возраста рабочего: $\hat{y} = 62{,}27 + 7{,}23 \cdot x$

То есть с увеличением возраста рабочего на 1 год работная плата в среднем повышается на 7,23 руб.

В матричной форме критерий метода наименьших квадратов записывается так:

$$S = \left(Y - X\tilde{b}\right)^T \left(Y - X\tilde{b}\right) = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 = \min_{b_0, b_1}$$

Дифференцируем S по вектору *b* и приравняем производные 0, чтобы найти МНК-оценки *b*. В результате получим систему из двух нормальных линейных уравнении:

$$\frac{\partial S}{\partial \tilde{b}} = -2X^T Y + 2X^T X\tilde{b} = 0.$$

Учитывая обратимость матрицы $X^T X$, находим МНК-оценку вектора *b*:

$\tilde{b} = \left(X^T X\right)^{-1} X^T Y$, где $\tilde{b} = \begin{pmatrix} \tilde{b}_0 \\ \tilde{b}_1 \end{pmatrix}$.

По правилу умножения матриц:

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 29 & 40 & 36 & 32 & 23 & 45 & 38 & 40 & 50 & 47 & 28 & 30 & 25 & 48 & 30 & 40 & 40 & 38 & 29 & 25 \end{pmatrix} \begin{pmatrix} 129 \\ 140 \\ 136 \\ 132 \\ 123 \\ 145 \\ 138 \\ 140 \\ 150 \\ 147 \\ 147 \\ 128 \\ 130 \\ 125 \\ 148 \\ 130 \\ 140 \\ 140 \\ 138 \\ 129 \end{pmatrix} = = \begin{pmatrix} 20 & 713 \\ 713 & 26691 \end{pmatrix}.$$

В матрице $X^TX$ число 20, лежащее на пересечении 1-й строки и 1-го столбца получено как сумма произведений элементов 1-й строки матрицы $X^T$ и 1-го столбца матрицы X. Число 713, лежащее на пересечении 1-й строки и 2-го столбца, получено как сумма произведений элементов 1-й строки матрицы $X^T$ и 2-го столбца матрицы X и т. д

$$X^TY = $$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 29 & 40 & 36 & 32 & 23 & 45 & 38 & 40 & 50 & 47 & 28 & 30 & 25 & 48 & 30 & 40 & 40 & 38 & 29 & 25 \end{pmatrix} \begin{pmatrix} 300 \\ 400 \\ 300 \\ 320 \\ 200 \\ 350 \\ 350 \\ 400 \\ 380 \\ 400 \\ 250 \\ 350 \\ 200 \\ 400 \\ 220 \\ 320 \\ 390 \\ 360 \\ 260 \\ 250 \end{pmatrix} = = \begin{pmatrix} 6400 \\ 327360 \end{pmatrix}.$$

Найдем обратную матрицу:

$$\left(X^T X\right)^{-1} = \frac{1}{20 \cdot 26691 - (713)^2} \begin{pmatrix} 26691 - 713 \\ -71320 \end{pmatrix} = = \begin{pmatrix} 1{,}048721 - 0{,}02801 \\ -0{,}028010{,}00079 \end{pmatrix}.$$

Тогда вектор оценок параметров регрессии равен:

$$\tilde{b} = \begin{pmatrix} 1{,}048721 - 0{,}02801 \\ -0{,}028010{,}00079 \end{pmatrix} \cdot \begin{pmatrix} 6400 \\ 237360 \end{pmatrix} = \begin{pmatrix} 62{,}27 \\ 7{,}23 \end{pmatrix}$$ а оценка уравнения регрессии будет иметь вид: $\hat{y} = 62{,}27 + 7{,}23 \cdot x$

## 1.3. Оценка регрессии

### *1.3.1. Оценка дисперсии случайной составляющей – $\sigma_u^2$. Статистические свойства МНК-оценок (состоятельность, несмещенность, эффективность). Ковариационная матрица МНК-оценок параметров регрессии*

*Оценка дисперсии случайной составляющей в случае ной линейной регрессии.*

Несмещенной оценкой дисперсии случайной составляющей является:

$$s_u^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}\left(y_i - \left(\tilde{b}_0 + \tilde{b}_1 \cdot x_i\right)\right)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(e_i)^2$$

где $e_i$ – остаток, равный разности между фактическим и рассчитанным по уравнению регрессии значениями y:

Оценкой ковариационной матрицы случайных составляющих $C_u$ будет матрица: $\hat{C}_u = s_u^2 \cdot I$.

При повторении выборок того же самого объема n из той же самой генеральной совокупности и при тех же самых значениях объясняющих переменных *x* наблюдаемые значения зависимой переменной *y* будут случайным образом варьироваться (за счет случайного характера случайной составляющей u). Следовательно, будут варьироваться и зависеть от $y_1,.., y_n$ значения оценок параметров регрессии (j = 0; 1] и оценка дисперсии случайной составляющей.

Покажем на примере $\tilde{b}_1$ что значение МНК-оценки параметра регрессии зависит от случайной составляющей *u*.

МНК-оценка коэффициента регрессии, как было показано в п. 1.2.3, рассчитывается как отношение ковариации *x* и *y* к дисперсии *x*:

$$\tilde{b}_1 = \frac{Cov(x,y)}{\sigma_x^2}$$

Так как y зависит от случайной составляющей u ($y=b_0+b_1x+u$), то ковариация $Cov(x,y)$ может быть расписана следующим образом:

$$Cov(x,y) = Cov(x,[b_0 + b_1 \cdot x + u]) =$$
$$= Cov(x,b_0) + Cov(x,b_1 \cdot x) + Cov(x,u).$$

Причем из свойств ковариации (см п. 1.2.2) следует, что:

$$Cov(x,b_0) = 0; \quad Cov(x,b_1 \cdot x) = b_1 \cdot Cov(x,x);$$
$$Cov(x,x) = \sigma_x^2.$$

Тогда $Cov(x,y) = b_1\sigma_x^2 + Cov(x,u).$

В результате МНК-оценка может быть разложена на случайную и неслучайную составляющие:

$$\tilde{b}_1 = \frac{b_1\sigma_x^2 + Cov(x,u)}{\sigma_x^2} = b_1 + \frac{Cov(x,u)}{\sigma_x^2},$$

т.е. МНК-оценка $\tilde{b}_1$ может быть представлена как сумма слагаемых: 1) постоянной величины, равной истинному значению коэффициента $b_1$;

2) случайной составляющей *Cov(x, u)*, которая обусловливает отклонения оценки коэффициента регрессии от истинного значения.

Аналогично можно показать, что МНК-оценка $\tilde{b}_0$, а так же оценка дисперсии $s_u^2$ имеют постоянную составляющую, равную истинному значению, и случайную составляющую зависящую от *u*.

Следует заметить, что на практике мы не можем разложить оценки параметров регрессии и $s_u^2$ на составляющие, так как истинные значения $b_0$,$b_1$ и $\sigma_u^2$ нам не известны, кроме того, мы не знаем фактических значений u в выборке.

Однако приведенное выше разложение оценок позволяет получить некоторую теоретическую информацию об их свойствах.

*Свойства МНК-оценок.* Критериями лучшего способа оценивания является требование состоятельности, несмещенности и эффективности оценок, найденных данным способом.

Способ оценивания дает **состоятельные** оценки, если при бесконечно большом объеме выборки значение статистической оценки стремится к искомому значению параметра (характеристики) генеральной совокупности.

Способ оценивания дает **несмещенные** оценки, если математическое ожидание оценки при данном способе оценивания тождественно искомому параметру (характернее генеральной совокупности (при любом объеме выборки).

Оценка, полученная при данном способе оценивания называется **эффективной**, если ее дисперсия минимальна (при заданном объеме выборки n).

МНК-оценки параметров и дисперсии случайной составляющей являются «наилучшими» (состоятельными, несмещенными и эффективными) оценками.

Докажем, что $\tilde{b}_1$ является несмещенной оценкой $b_1$ если выполняется 1-я предпосылка нормальной линейной модели регрессии. Если мы примем сильную форму 1-й предпосылки нормальной линейной регрессионной модели, т.е. предположим, что *x* –неслучайная величина, то мы можем считать $\sigma_x^2$ известной константой, а математическое ожидание $M\left(Cov\left(x,u\right)\right)$ равным нулю. Тогда:

$$M\left(\tilde{b}_1\right)=M\left(b_1+\frac{Cov(x,u)}{\sigma_x^2}\right)=b_1+M\left(\frac{Cov(x,u)}{\sigma_x^2}\right)=$$

$$=b_1+\frac{1}{\sigma_x^2}M\left(Cov(x,u)\right)=b_1+\frac{1}{\sigma_x^2}\cdot 0=b_1.$$

То есть $M\left(\tilde{b}_1\right)=b_1$, следовательно, $\tilde{b}_1$ является несмещенной оценкой $b_1$.

Аналогично доказывается несмещенность оценки $\tilde{b}_0$.

Эффективность МНК-оценок доказывается с помощью теоремы Гаусса-Маркова, которая гласит:

*Метод наименьших квадратов дает оценки, имеющие наименьшую дисперсию в классе всех линейных несмещенных оценок, если выполняются предпосылки нормальной линейной регрессионной модели* (см. п. 1.2.2).

Ковариационная матрица МНК-оценок параметров регрессии – матрица ковариаций оценок параметров. Для случая парной регрессии это матрица размером 2×2:

$$C_{\tilde{b}}=\begin{pmatrix} Cov\left(\tilde{b}_0,\tilde{b}_0\right) & Cov\left(\tilde{b}_0,\tilde{b}_1\right) \\ Cov\left(\tilde{b}_1,\tilde{b}_0\right) & Cov\left(\tilde{b}_1,\tilde{b}_1\right) \end{pmatrix}=\begin{pmatrix} \sigma_{\tilde{b}_0}^2 & 0 \\ 0 & \sigma_{\tilde{b}1}^2 \end{pmatrix}.$$

Данная матрица в соответствии с методом наименьших квадратов рассчитывается следующим образом:

$$C_{\tilde{b}}=\sigma_u^2\left(X^T\cdot X\right)^{-1}.$$

На главной диагонали данной матрицы, находятся дисперсии МНК-оценок параметров. Для случая парной линейной регрессии формулы расчета дисперсий МНК-оценок параметров следующие:

$$\sigma_{\tilde{b}_0}^2=\frac{\sigma_u^2}{n}\left(1+\frac{\overline{x^2}}{\sigma_x^2}\right),\ \sigma_{\tilde{b}_1}^2=\frac{\sigma_u^2}{n\sigma_x^2},$$

где $\sigma_u^2$ – дисперсия случайной составляющей;

$\sigma_x^2$ – дисперсия факторного признака *x*.

Так как $\sigma_u^2$ нам известна, то при расчете ковариационной матрицы пользуются оценкой дисперсии случайных составляющих – $s_u^2$. Тогда получим оцененную ковариационную матрицу: $C_{\tilde{b}}=s^2\left(X^T\cdot X\right)^{-1}$.

Формулы расчета оценок дисперсий $\sigma^2_{\tilde{b}_0}$ и $\sigma^2_{\tilde{b}_1}$ (полученные через $s_u^2$) в случае парной линейной регрессии будут следующими:

$$\mu^2_{\tilde{b}_0} = \frac{\sum_{i=1}^{n}\left(y-\hat{y}_i\right)^2 \cdot \sum_{i=1}^{n} x_i^2}{n\cdot(n-2)\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2},\ \mu^2_{\tilde{b}_1} = \frac{\sum_{i=1}^{n}\left(y-\hat{y}_i\right)^2}{(n-2)\cdot\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}$$

Корень из оценки дисперсии $\tilde{b}_0$ и $\tilde{b}_1$.

### *1.3.2. Показатели качества регрессии*

Качество модели регрессии связывают с адекватностью модели наблюдаемым (эмпирическим) данным. Проверка адекватности (или соответствия) модели регрессии наблюдаемым данным проводится на основе анализа остатков — $e_i$. Остаток представляет собой отклонение фактического значения зависимой переменной от значения данной переменной, полученное расчетным путем: $e_i = y_i - \hat{y}_i (i=1;n)$. Если $e_i = 0 (i=1;n)$, то для всех наблюдений фактические значения зависимой переменной совпадают с расчетными (теоретическими) значениями: $y_i = \hat{y}_i\ (i=1;n))$. Графически это означает, что теоретическая линия регрессии (линия, построенная по функции $\hat{y}_i = \tilde{b}_0 + \tilde{b}_1 \cdot x_1$ проходит через все точки корреляционного поля, что возможно только при строго функциональной связи. Следовательно, результативный признак *y* полностью обусловлен влиянием фактора *x*.

На практике, как правило, имеет место некоторое рассеивание точек корреляционного поля относительно теоретической линии регрессии, т. е. отклонения эмпирических данных от теоретических $e_i \neq 0$. Величина этих отклонений и лежит в основе расчета показателей качества (адекватности) уравнения.

При анализе качества модели регрессии используется теорема о разложении дисперсии, согласно которой общая дисперсия результативного признака может быть разложена на две составляющие – объясненную и необъясненную уравнением регрессии дисперсии:

$$\sigma^2 = \delta^{*2} + \varepsilon^{*2},$$

где $\delta^{*2} = \frac{\sum_{i=1}^{n}\left(\hat{y}_i-\bar{y}\right)^2}{n}$ — объясненная уравнением регрессии дисперсия результативного признака;

$\varepsilon^{*2} = \frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{n} = \frac{\sum_{i=1}^{n}\varepsilon_i{}^2}{n}$ необъясненная уравнением регрессии (остаточная) дисперсия результативного признака.

На основе теоремы о разложении дисперсии рассчитываются показатели качества модели регрессии.

1. Теоретический коэффициент (индекс для нелинейных форм связей) детерминации: $R^2_{y(x)} = \frac{\delta^{*2}}{\sigma^2}$. Он представляет собой отношение объясненной (уравнением) дисперсии признака-результата к общей дисперсии результативного признака. Коэффициент детерминации характеризует долю вариации (дисперсии) результативного признака, объясняемую регрессией в общей вариации (дисперсии) *y*. Соответственно величина $1 - R^2_{y(x)}$ характеризует долю вариации (дисперсии) *y,* необъясненную уравнением регрессии, а значит, вызванную влиянием прочих неучтенных в модели факторов.

! При парной линейной регрессии коэффициент детерминации равен квадрату парного линейного коэффициента корреляции $\left(r^2_{yx}\right)$: $R^2_{y(x)} = r^2_{yx}$.

Коэффициент линейной парной корреляции – показатель тесноты линейной связи между признаками *y* и *x*:

$$r_{x,y} = \frac{Cov(x,u)}{\sigma_x^2 \cdot \sigma_y^2} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y}$$

где $\sigma_x$ – среднее квадратическое отклонение фактора;

$\sigma_y$ – среднее квадратическое отклонение результата.

Формула его расчета очень похожа на формулу расчета коэффициента регрессии методом наименьших квадратов – $\tilde{b}$. Поэтому коэффициент линейной парной корреляции может быть рассчитан следующим образом:

$$r_{x,y} = \tilde{b} \frac{\sigma_x}{\sigma_y}$$

Область допустимых значений линейного парного коэффициента корреляции от -1 до +1. Если коэффициент корреляции по модулю близок к единице, то связь между признаками может быть охарактеризована как тесная линейная. Если коэффициент корреляции по модулю близок к нулю, то имеет место слабая линейная зависимость.

2. Корень из этого коэффициента (индекса) детерминации $R_{y(x)} = \sqrt{\frac{\delta^{*2}}{\sigma^2}}$ есть коэффициент (индекс) множественной корреляции, или теоретическое корреляционное отношение. Если все точки корреляционного поля лежат на теоретической линии регрессии, то $R_{y(x)} = 1$; следовательно связь между *y* и *x*

— функциональная, и уравнение регрессии очень хорошо описывает фактические данные. Если $R_{y(x)} = 0$, то уравнение плохо описывает данные, а значит, связь между признаками отсутствует.

**!**В случае парной линейной регрессии $R_{y(x)} = \left| r_{yx}^2 \right|$

3. Средняя квадратическая ошибка уравнения регрессии представляет собой среднее квадратическое отклонение наблюдаемых значений результативного признака от теоретических значений, рассчитанных по модели, т.е.: $s_e = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-h}}$, где $h$ равно числу параметров в модели регрессии. Величину средней квадратической ошибки можно сравнить со средним квадратическим отклонением результативного признака $\sigma_y$. Если $s_e$ окажется меньше $\sigma_y$, использование модели регрессии является целесообразным.

4. Средняя ошибка аппроксимации $A = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$. Чем меньше рассеяние эмпирических точек вокруг теоретической линии регрессии, тем меньше средняя ошибка аппроксимации. Ошибка аппроксимации меньше 7% свидетельствует о хорошем качестве модели.

При обработке информации на компьютере выбор вида зависимости (вида уравнения регрессии) обычно осуществляется методом сравнения величины показателя адекватности, рассчитанного при разных видах зависимости. Если показатели адекватности оказываются примерно одинаковыми для нескольких функций, то предпочтение отдается более простым видам функций, ибо они в большей степени поддаются интерпретации и требуют меньшего объема наблюдений.

Для данных табл. 1.6 была построена линейная парная модель регрессии: $\hat{y} = 62{,}27 + 7{,}23 \cdot x$, описывающая зависимость заработной платы рабочего от его возраста.

Рассчитаем показатели качества модели регрессии для этого примера.

Для расчета теоретического коэффициента детерминации определим значение линейного парного коэффициента корреляции $(r_{yx})$ через МНК-оценку коэффициента регрессии: $r_{x,y} = \tilde{b}_1 \frac{\sigma_x}{\sigma_y} = 7{,}23 \cdot \frac{7{,}98}{67{,}6} = 0{,}853$

(где $\sigma_x = 7{,}98$ среднее квадратическое отклонение возраста рабочего; $\sigma_y = 67{,}6$ – среднее квадратическое отклонение заработной платы рабочего).

Тогда теоретический коэффициент детерминации будет равен: $R^2_{yx} = 0{,}853^2 \approx 0{,}728$. Следовательно 72,8% вариации заработной платы рабочего объясняется уравнением линейной регрессии, а значит и возрастом рабочего. А100 – 72,8 = 27,2 % вариации заработной платы обусловлено влиянием не учтенных в модели факторов.

Коэффициент множественной корреляции равен: $R_{y(x)} = r_{yx} = 0{,}853$. Близость к единице данного показателя свидетельствует о хорошей аппроксимации модели фактических данных.

Для расчета средней квадратической ошибки уравнения регрессии нужно рассчитать теоретические значения результативного признака $(\hat{y}_i)$, остатки $(e_i)$ и их квадраты. Результаты расчета сведены в табл.1.6.

Таблица 1.6

| Наблюдение — $i$ | $y_i$> | $\hat{y}_i$ | $e_i = y_i - \hat{y}_i$ | $e_i^2$ |
|---|---|---|---|---|
| 1 | 300 | 271,9233 | 28,0767 | 788,3009 |
| 2 | 400 | 351,4487 | 48,55133 | 2357,232 |
| 3 | 300 | 322,5304 | -22,5304 | 507,6168 |
| 4 | 320 | 293,612 | 26,38796 | 696,3245 |
| 5 | 200 | 228,5458 | -28,5458 | 814,8646 |
| 6 | 350 | 387,5966 | -37,5966 | 1413,501 |
| 7 | 350 | 336,9895 | 13,01049 | 169,2729 |
| 8 | 400 | 351,4487 | 48,55133 | 2357,232 |
| 9 | 380 | 423,7445 | -43,7445 | 1913,577 |
| 10 | 400 | 402,0557 | -2,05571 | 4,22596 |
| И | 250 | 264,6937 | -14,6937 | 215,9056 |
| 12 | 350 | 279,1529 | 70,84712 | 5019,314 |
| 13 | 200 | 243,005 | -43,005 | 1849,429 |
| 14 | 400 | 409,2853 | -9,28529 | 86,21667 |
| 15 | 220 | 279,1529 | -59,1529 | 3499,063 |
| 16 | 320 | 351,4487 ' | -31,4487 | 989,0186 |
| 17 | 390 | 351,4487 | 38,55133 | 1486,205 |
| 18 | 360 | 336,9895 | 23,01049 | 529,4827 |
| 19 | 260 | 271,9233 | -11,9233 | 142,1652 |
| 20 | 250 | 243,005 | 6,99501 | 48,93017 |
| Итого | 6400 | 6400 | 0 | 24887,88 |

Тогда $s_e = \sqrt{\frac{24887{,}88}{20-2}} = 37{,}185$ (в нашем примере $n$ =20, $h$=2).

Среднюю квадратическую ошибку можно найти другим способом – через теоретический коэффициент детерминации, не прибегая к расчетам теоретических значений признака-результата и остатков:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = n \cdot \left(1 - R_{yx}^2\right) \cdot \sigma_y^2$$

$$s_e = \sqrt{\frac{n \cdot \left(1 - R_{yx}^2\right) \cdot \sigma_y^2}{n - h}} = \sqrt{\frac{20 \cdot 0{,}272 \cdot 4570}{20 - 2}} = 37{,}16$$

Величина $s_e = 37{,}18$ оказывается меньше $\sigma_y = 67{,}6$, следовательно, модель регрессии целесообразно использовать.

Рассчитаем среднюю ошибку аппроксимации. Для нашего примера $A = 0{,}1002$ (10 %), что свидетельствует о незначительной погрешности модели.

### *1.3.3. Проверка гипотез о значимости параметров регрессии, коэффициента корреляции и уравнения регрессии в целом*

С помощью метода наименьших квадратов мы получили лишь оценки параметров уравнения регрессии. Чтобы проверить, значимы ли эти параметры (т. е. значимо ли они отличаются от нуля в «истинном» уравнении регрессе $y = b_0 + b_1 \cdot x + u$), используют статистические методы проверки гипотез. С помощью статистических методов проверки гипотез можно также проверить значимость коэффициента парной линейной корреляции (т. е. значимо ли он отличается от нуля в генеральной совокупности).

В качестве основной гипотезы ($H_0$) выдвигают гипотезу о незначимом отличии от нуля «истинного» параметра регрессии или коэффициента корреляции. Альтернативой гипотезой ($H_1$) при этом является гипотеза обратная, т.е. о неравенстве нулю «истинного» параметра или коэффициента корреляции. Мы заинтересованы в том, чтобы основная гипотеза была отвергнута. Для проверки этой гипотезы; используется *t*-статистика критерия проверки гипотезы, имеющая распределение Стьюдента.

Найденное по данным наблюдений значение *t*-статистики (его еще называют наблюдаемым или фактическим) сравнивается с критическим значением *t* -статистики, определяемым по таблицам распределения Стьюдента (которые обычно приводятся в конце учебников и практикумов по статистике или эконометрике). Критическое значение определяется в зависимости от уровня значимости ( $\alpha$ ) и числа степеней свободы, которое равно ($n - h$), $n$ — число наблюдений, $h$ — число оцениваемых параметров в уравнении регрессии. В случае линейной парной регрессии $h = 2$, а число степеней свободы равно (n – 2). Критическое значение может быть также вычислено на компьютере с помощью встроенной функции СТЬЮДРАСПОБР пакета Excel.

Если фактическое значение *t*-статистики, взятое по модулю, больше критического, то основную гипотезу отвергают и считают, что с вероятностью $(1-\alpha)$ «истинный» параметр регрессии (либо коэффициент корреляции) значимо отличается от нуля.

Если фактическое значение *t*-статистики (по модулю) меньше критического, то нет оснований отвергать основную гипотезу, т. е. «истинный» параметр регрессии (либо коэффициент корреляции) незначимо отличается от нуля при уровне значимости $\alpha$.

Для проверки гипотезы: $b_1=0$ статистика критерия проверки имеет вид:

$$t_{(b_1=0)} = \frac{\tilde{b}_1}{\mu_{b_1}}$$

где $\tilde{b}_1$ — оценка коэффициента регрессии $b_1$ полученная по наблюдаемым данным;

$\mu_{b_1}$ — стандартная ошибка оценки коэффициента регрессии $\tilde{b}_1$.

Для линейного парного уравнения регрессий стандартная ошибка коэффициента вычисляется по формуле:

$$\mu_{b_1} = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{(n-2)\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}}.$$

Числитель в этой формуле может быть легко рассчитан через коэффициент детерминации и общую дисперсию признака-результата:

$$\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2 = n \cdot \left(1 - R_{yx}^2\right) \cdot \sigma_y^2.$$

Для проверки гипотезы: $b_0=0$ статистика критерия проверки гипотезы имеет вид:

$$t_{(b_0=0)} = \frac{\tilde{b}_0}{\mu_{b_0}},$$

где $\tilde{b}_0$ — оценка параметра регрессии $b_0$, полученная по наблюдаемым данным;

$\mu_{b_0}$ — стандартная ошибка оценки параметра $\tilde{b}_0$.

Для линейного парного уравнения регрессии:

$$\mu_{b_0} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \sum_{i=1}^{n} x_i^{\,2}}{n \cdot (n-2) \sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Для проверки гипотезы о незначимом отличии от нуля «истинного» коэффициента линейной парной корреляции используют статистику критерия:

$$t_{(r=0)} = \frac{r_{yx}}{\mu_r}$$

$r_{yx}$ – оценка коэффициента корреляции, полученная по наблюдаемым данным (выборочный коэффициент корреляции);

$\mu_r$ – стандартная ошибка выборочного коэффициента корреляции $r_{yx}$.

Для линейного парного уравнения регрессии:

$$\mu_r = \sqrt{\frac{\left(1 - r_{yx}^2\right)}{(n-2)}}$$

! • В парной линейной регрессии между наблюдаемыми значениями статистик критериев существует взаимосвязь: $t_{(b_1=0)} = t_{(r=0)}$.

Рассмотренная формула статистики критерия проверки гипотезы о незначимом отличии от нуля коэффициента корреляции рекомендуется к применению, если:

1) число наблюдений ($n$) большое;
2) величина $|r_{yx}|$ не близка к единице.

Если же величина выборочного коэффициента корреляции по модулю близка к 1, то распределение его оценок отличается от распределения Стьюдента. В данном случае используют подход, предложенный Р. Фишером, а именно, для оценки значимости линейного парного коэффициента корреляции $r$ вводится вспомогательная величина $z$, связанная с данным коэффициентом следующим отношением:

$$z = 0{,}5 \cdot ln\left(\frac{1 + r_{yx}}{1 - r_{yx}}\right).$$

Величину $z$ можно не рассчитывать, а воспользоваться готовыми таблицами $z$-преобразования, в которых приведены значения $z$ для соответствующих значений $r_{yx}$.

При изменении $r_{yx}$ от –1 до +1 величина $z$ изменяется от $-\infty$ до $+\infty$, что соответствует нормальному распределению. Математический анализ

доказывает, что распределение величины *z* мало отличается от нормального даже при близких к единице значениях коэффициента корреляции.

Тогда гипотеза о том, что «истинный» коэффициент корреляции незначимо отличается от нуля, сводится к гипотезе о незначимом отличии от нуля величины *z*. Для проверки данной гипотезы используют статистику критерия: $t_{(z=0)} = \frac{z}{\mu_z}$. Стандартная ошибка $\mu_z$ определяется по формуле:

$$\mu_z = \frac{1}{\sqrt{n-3}},$$

где *n* — число наблюдений.

Критическое значение *t*-статистики — $t_{кр}$ находят по таблицам стандартного нормального распределения по доверительной вероятности ( $1-\alpha$ ). Основную гипотезу отвергают, если $|t_{z=0}| > t_{кр}$.

Оценка значимости уравнения регрессии производится для того, чтобы узнать, пригодно уравнение регрессии для практического использования (например, для прогноза) или нет. При этом выдвигают основную гипотезу о незначимости уравнения в целом, которая формально сводится к гипотезе о равенстве нулю параметров регрессии, или, что тоже самое, о равенстве нулю коэффициента детерминации $R^2=0$. Альтернативная ей гипотеза о значимости уравнения – гипотеза о неравенстве нулю параметров регрессии или о неравенстве нулю коэффициента детерминации:

Для ее проверки используют F-статистику критерия проверки гипотезы: $F = \frac{R^2}{1-R^2} \cdot \frac{n-h}{h-1}$, где *n* –число наблюдений; *h* – число оцениваемых параметров. Данная статистика имеет распределение Фишера-Снедоккора.

По таблицам распределения Фишера-Снедоккора находят критическое значение F-критерия в зависимости от уровня значимости $\alpha$ (обычно его берут равным 0,05) и двух чисел степеней свободы *k1=h – 1 и k2 = n – h.*

Сравнивают значение F-критерия, рассчитанное по данным выборки — $F_{набл}$ с критическим значением $F_{кр(\alpha;k1;k2)}$. Если $F_{набл} < F_{кр(\alpha;k1;k2)}$, то гипотезу о незначимости уравнения регрессии не отвергают. Если $F_{набл} > F_{кр(\alpha;k1;k2)}$, то выдвинутую гипотезу отвергают и принимают альтернативную гипотезу о статистической значимости уравнения регрессии.

! В случае линейной парной регрессии существует взаимосвязь между статистиками критериев проверки гипотез: $t_{(b_1=0)} = t_{(r=0)} = \sqrt{F}$ .

### *1.3.4. Прогноз ожидаемого значения результативного признака по линейному парному уравнению регрессии*

Пусть требуется оценить прогнозное значение признака-результата для заданного значения признака-фактора ($x^p$). Прогнозируемое значение признака-результата с доверительной вероятностью, равной $(1-\alpha)$, принадлежит интервалу прогноза:

$$\left(y^p - t \cdot \mu_p ; y^p + t \cdot \mu_p\right)$$

где $y^p$ – точечный прогноз;

$t$ – коэффициент доверия, определяемый по таблицам распределения Стьюдента в зависимости от уровня значимости $\alpha$ и числа степеней свободы ($n$–2);

$\mu_p$ – средняя ошибка прогноза.

Точечный прогноз рассчитывается по линейному уравнению регрессии как $y^p = a + b \cdot x^p$. Средняя ошибка прогноза определяется по формуле:

$$\mu_p = \sqrt{\frac{\sum\left(e_i^2\right)}{n-2}\left(1+\frac{1}{n}+\frac{\left(x^p-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)} = \sqrt{s_e^2 \cdot \left(1+\frac{1}{n}+\frac{\left(x^p-\bar{x}\right)^2}{\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}\right)}$$

где $s_e^2$ – средняя квадратическая ошибка регрессии.

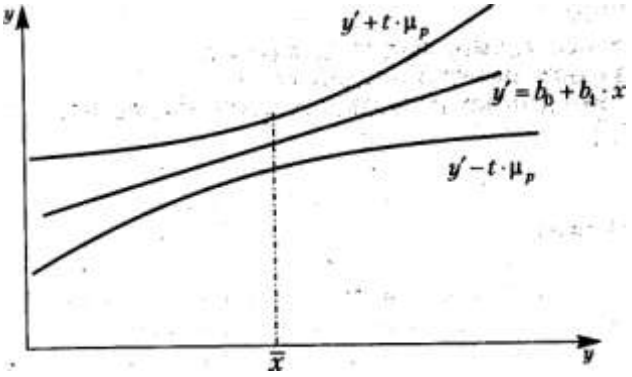


Рис. 1.5. Точечная и интервальная оценки прогноза

По мере удаления $x^p$ от среднего значения ($\bar{x}$) ширина доверительного интервала будет увеличиваться (рис. 1.5).

## 1.4. Нелинейная регрессия

### *1.4.1. Виды нелинейной регрессии. Оценка параметров нелинейной регрессии*

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций, например равносторонней гиперболы: $y_i = a + \frac{b}{x_i} + u_i$; параболы второй степени: $y_i = a + b \cdot x_i + c \cdot x_i^2 + u_i \ \ (i = 1; n)$ и др.

Различают два класса нелинейных регрессий:

- относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам;
- по оцениваемым параметрам.

Нелинейные регрессии по включаемым в нее объясняющим переменным, но линейные по оцениваемым параметрам

Данный класс нелинейных регрессий включает уравнения, в которых *y* линейно связан с параметрами. Примером могут служить следующие функции.

Полиномы разных степеней – $y_i = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2 + b_3 \cdot x_i^3 + ... + b_k \cdot x_i^k + u_i$ – (полином *k*-й степени).

Равносторонняя гипербола – $y_i = a + \frac{b}{x_i} + u_i$.

*Оценка параметров регрессий нелинейных по объясняющим переменным.* При этом используется подход, именуемый «замена переменных». Суть его состоит в замене «нелинейных» объясняющих переменных новыми «линейными» переменными и сведение нелинейной регрессии к линейной. К новой «преобразованной» регрессии может быть применен обычный метод наименьших квадратов (МНК).

Рассмотрим применение данного подхода к параболе второй степени: $y_i = a + b \cdot x_i + c \cdot x_i^2 + u_i$. Заменяя переменную $x^2$ на *z*, получим двухфакторное уравнение линейной регрессии: $y_i = a + b \cdot x_i + c \cdot z_i + u_i$, для оценки параметров которого используется обычный МНК.

Соответственно, для полинома *k*-го порядка $y_i = b_0 + b_1 \cdot x_i + b_2 \cdot x_i^2 + b_3 \cdot x_i^3 + ... + b_k \cdot x_i^k + u_i$,

при замене: $z_1 = x, z_2 = x^2, z_3 = x^3, ..., z_k = x^k$ получим

$y_i = b_0 + b_1 \cdot z_{1i} + b_2 \cdot z_{2i} + b_3 \cdot z_{3i} + ... + b_k \cdot z_{ki} + u_i$

Следовательно, полином любого порядка сводится к линейной регрессии с ее методами оценивания параметров и проверки гипотез.

Среди нелинейной полиноминальной регрессии чаще всего используется парабола второй степени; в отдельных случаях — полином третьего порядка. Ограничение в использовании полиномов более высоких степеней связаны с требованием однородности исследуемой совокупности: чем выше порядок полинома, тем больше изгибов имеет кривая и, соответственно, менее однородна совокупность по результативному признаку.

Парабола второй степени (рис. 1.6) целесообразна к применению, если для определенного интервала значений фактора меняется характер связи с результатом: прямая связь меняется на обратную или обратная на прямую. В этом случае определяется значение фактора, при котором достигается максимальное (или минимальное) значение результативного признака:

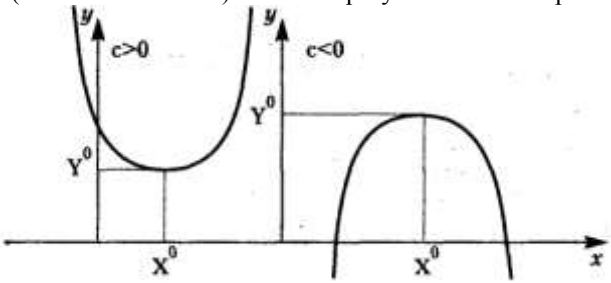


Рис. 1.6. Параболическая зависимость

Если же исходные данные не обнаруживают изменения направленности связи, то параметры параболы второй степени становятся трудно интерпретируемыми, а форма связи часто заменяется другими нелинейными моделями.

Среди класса нелинейных функций, параметры которых без особых затруднений оцениваются МНК, следует назвать хорошо известную в эконометрике равностороннюю гиперболу: $y_i = a + \frac{b}{x_i} + u_i$. Она может быть использована, например, для характеристики связи удельных расходов сырья, материалов и топлива с объемом выпускаемой продукции.

Для оценки параметров равносторонней гиперболы используется тот же подход «замены переменных»: заменив $1/x$ на $z$, получим линейное уравнение регрессии: $y_i = a + b \cdot z_i + u_i$, для которого может быть применен обычный МНК.

При $b > 0$ имеем обратную зависимость, которая при $x \to \infty$ характеризуется нижней асимптотой, т. е. минимальным предельным значением $y$, оценкой которого служит параметр $a$ (рис. 1.7, а).

При $b < 0$ имеем медленно повышающуюся функцию с верхней асимптотой при $x \to \infty$, т. е. максимальным предельным уровнем $y$, оценкой которого служит параметр $a$(рис. 1.7, б).

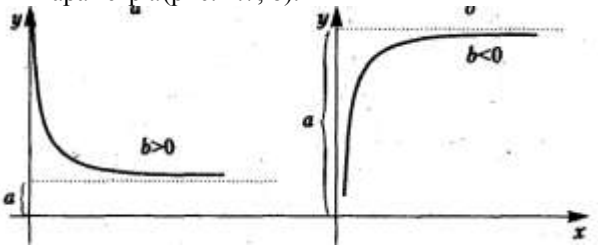


Рис. 1.7. Зависимость в виде равносторонней гиперболы

*Регрессии нелинейные по оцениваемым параметрам.* К данному классу регрессий относятся уравнения, в которых $y$ нелинейно связан с параметрами. Примером таких нелинейных регрессий являются функции:

- степенная – $y_i = a \cdot x_i^b \cdot u_i$;
- показательная – $y_i = a \cdot b^{x_i} \cdot u_i$;
- экспоненциальная – $y_i = e^{a+b \cdot x_i} \cdot u_i$.

Данный класс нелинейных моделей подразделяется на два типа:

1) нелинейные модели внутренне линейные;
2) нелинейные модели внутренне нелинейные.

Если *нелинейная модель внутренне линейна*, то она с помощью соответствующих преобразований может быть приведена к линейному виду (например, логарифмированием и заменой переменных). Если же *нелинейная модель внутренне нелинейна*, то она не может быть сведена к линейной функции.

Примером нелинейной по параметрам регрессии внутренне линейной является степенная функция, которая широко используется в эконометрических исследованиях при изучении спроса от цен:

$$y_i = a \cdot x_i^b \cdot u_i$$

где$y$ – спрашиваемое количество;

$x$ – цена;

$u$ – случайная составляющая.

Данная модель нелинейна относительно оцениваемых параметров, т. к. включает параметры $a$ и $b$ неаддитивно. Однако ее можно считать внутренне линейной, ибо логарифмирование данного уравнения по основанию $e$ приводит

его к линейному виду: $\ln y_i = \ln a + b \cdot \ln x_i + \ln u_i$. Заменив переменные и параметры, получим линейную регрессию, оценки параметров которой *a* и *b* могут быть найдены МНК.

В рассматриваемой выше степенной функции предполагалось, что случайная составляющая *u* мультипликативно связана с объясняющей переменной *x*. Если же модель представить в виде $y_i = a \cdot x_i^b \cdot u_i$, то она становится внутренне нелинейной, т. к. ее невозможно преобразовать к линейному виду.

Если модель внутренне нелинейна по параметрам, то для оценки параметров используются итеративные процедуры, успешность которых зависит от вида уравнений и особенностей применяемого итеративного подхода.

*Применение МНК для оценки параметров нелинейных моделей внутренне линейных.* В моделях, нелинейных по оцениваемым параметрам, но приводимых к линейному виду, МНК применяется к преобразованным уравнениям. В таких моделях преобразованию подвергается *результативный признак y*, в отличие от нелинейных моделей 1-го типа, где результативный признак *y* остается неизменным, а преобразуется факторный признак.

Если в линейной модели и моделях, нелинейных по переменным, при оценке параметров исходят из критерия $\sum (y - \hat{y})^2 \to min$, то в моделях, нелинейных по оцениваемым параметрам, требование МНК применяется не к исходным данным результативного признака, а к их преобразованным величинам, т. е. $\ln y$, $1/y$. Это значит, что оценка параметров основывается на минимизации суммы квадратов отклонений логарифмов:

$$\sum (\ln y - (\ln \hat{y}))^2 \to min.$$

Соответственно, если в линейных моделях и моделях, нелинейных по переменным, $\sum (y - \hat{y})^2 = 0$, то в моделях, нелинейных по оцениваемым параметрам, $\sum (\ln y - (\ln \hat{y}))^2 = 0$, а $\sum (y - \hat{y})^2 \neq 0$. Вследствие этого оценка параметров с помощью МНК для нелинейных моделей, внутренне линейных, оказывается несколько смещенной.

При исследовании взаимосвязей среди функций, использующих $\ln y$, в эконометрике преобладают степенные зависимости – это кривые спроса и предложения, кривые Энгеля, производственные функции, кривые освоения для характеристики связи между трудоемкостью продукции и масштабами производства в период освоения выпуска нового вида изделий, а также зависимость валового национального дохода от уровня занятости.

Для оценки параметров степенной функции $y_i = a \cdot x_i^b \cdot u_i$ применяется МНК к линеаризованному уравнению $\ln y_i = \ln a + b \cdot \ln x_i + \ln u_i$, т. е. решается система нормальных уравнений:

$$\begin{cases} \sum_{i=1}^{n} \ln y_i = n \cdot \ln a + b \sum_{i=1}^{n} \ln x_i; \\ \sum_{i=1}^{n} \ln y_i \cdot \ln x_i = \ln a \cdot \sum_{i=1}^{n} \ln x_i + b \sum_{i=1}^{n} (\ln x_i)^2. \end{cases}$$

Параметр *b* определяется непосредственно из системы, а параметр *a* – косвенным путем после потенцирования величины $\ln a$.

Поскольку параметр *a* экономически не интерпретируется, то нередко зависимость записывается в виде логарифмически линейной, т. е. как: $\ln y_i = A + b \cdot \ln x_i + \ln u_i \ (A = \ln a)$.

В виде степенной функции изучается эластичность не только спроса, но и предложения. При этом обычно в функциях спроса параметр $b < 0$, а в функциях предложения – $b > 0$.

В отдельных случаях может использоваться так называемая обратная функция: $y_i = 1/(a + b \cdot x_i + u_i)$, являющаяся разновидностью гиперболы. Но если в равносторонней гиперболе $y_i = a + b/x_i + u_i$, преобразованию подвергается объясняющая переменная $1/x = z$ и $y_i = a + b \cdot z_i + u_i$, то для получения линейной формы зависимости в обратной модели преобразовывается *y*: $1/y = z$. Тогда модель обратной зависимости принимает вид: $z_i = a + b \cdot x_i + u_i$.

Обратная модель является внутренне линейной по параметрам. Требование МНК при этом выполняется для обратных значений результативного признака – $1/y = z$, а именно: $\sum (z - \hat{z})^2 \to min$.

Поскольку уравнение обратной функции линейно относительно величин $1/y$, то, если обратные значения $1/y$ имеют экономический смысл, коэффициент регрессии интерпретируется так же, как в линейном уравнении регрессии. Если, например, под *y* подразумеваются затраты на рубль продукции, а под *x* – производительность труда (выработка продукции на одного работника), то обратная величина характеризует затратоотдачу, и параметр *b* имеет экономическое содержание – средний прирост продукции в стоимостном измерении на 1 руб. затрат с ростом производительности труда на единицу своего измерения.

### *1.4.2. Корреляция для нелинейной регрессии*

Уравнение нелинейной регрессии, так же как и в линейной зависимости, дополняется показателями корреляции:

$$R^2 = \frac{\delta^{*2}}{\sigma^2} = \left(1 - \frac{\varepsilon^{*2}}{\sigma^2}\right) = 1 - \frac{n\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2};\ R^2 = \sqrt{\frac{\delta^{*2}}{\sigma^2}},$$

где $\sigma^2$ – общая дисперсия результативного признака *y,*

$\delta^{*2}$ – объясненная уравнением регрессии $y = f(x)$ дисперсия *y;*

$\varepsilon^{*2}$ – остаточная (необъясненная уравнением) дисперсия признака *y.*

Величину $R^2$ (равную отношению объясненной уравнением регрессии дисперсии результата *y* к общей дисперсии *y)* для нелинейных связей называют **индексом детерминации,** а корень из данной величины R называют **индексом корреляции.**

Величина индекса корреляции R находится в границах от 0 до 1. Чем ближе она к единице, тем теснее связь рассматриваемых признаков, тем более надежно уравнение регрессии.

Если после преобразования уравнение регрессии (нелинейное по объясняющим переменным) принимает форму линейного *парного* уравнения регрессии, то для оценки тесноты связи может быть использован линейный коэффициент корреляции $R_{yx} = r_{yz}$, где *z* – преобразованная величина признака-фактора, например $z = 1/x$ или $z = \ln x$.

Иначе обстоит дело, когда преобразования уравнения в линейную форму связаны с результативным признаком (нелинейность по параметрам). В этом случае линейный коэффициент корреляции по преобразованным значениям признаков дает лишь приближенную оценку тесноты связи и численно не совпадает с индексом корреляции.

Вследствие близости результатов и простоты расчета с использованием компьютерных программ для характеристики тесноты связи по нелинейным функциям широко используется линейный коэффициент корреляции ( $r_{\ln y,\ln x}$ или $r_{\ln y,x}$ *).* Несмотря на близость значений $R_{y(x)}$ и $r_{\ln y,\ln x}$ или $R_{y(x)}$ и $r_{\ln y,x}$ следует помнить, что $R_{y(x)}$ для функции $\hat{y} = f(x)$ не равен $R_{x(y)}$ для регрессии $\hat{x} = \varphi(y)$ (в отличие от уравнений регрессии линейных или нелинейных по объясняющим переменным, для которых $r_{y,x} = r_{x,y}$ и $R_{y(x)} = R_{x(y)}$.

Поскольку в расчете индекса корреляции используется соотношение объясненной (факторной) и общей суммы квадратов отклонений, то $R^2$ имеет тот же смысл, что и коэффициент детерминации. Оценка существенности индекса корреляции производится так же, как и оценка надежности коэффициента корреляции.

Индекс детерминации $R^2$ используется для проверки существенности уравнения нелинейной регрессии в целом по F-критерию Фишера:

$$F = \frac{R^2 \cdot (n-h)}{(1-R^2)\cdot(h-1)}$$

где $R^2$ – индекс детерминации;

$n$ – число наблюдений;.

$h$ – число параметров в уравнении.

Величина $(h - 1)$ характеризует число степеней свободы для объясненной (факторной) суммы квадратов, а $(n - h)$ – число степеней свободы для остаточной суммы квадратов.

Индекс детерминации $R^2$ можно сравнивать с коэффициентом детерминации $r^2$ для обоснования возможности применения линейной функции. Чем больше кривизна линии регрессии, тем величина коэффициента детерминации $r^2$ меньше индекса детерминации $R^2$. Близость этих показателей означает, что нет необходимости усложнять форму уравнения регрессии и можно использовать линейную функцию. Практически, если величина $(R^2 - r^2)$ не превышает 0,1, то предположение о линейной форме связи считается оправданным. В противном случае проводится оценка существенности различий $R^2$, вычисленных по одним и тем же исходным данным, через $t$-критерий Стьюдента:

$$t_{((R-r)=0)} = \frac{R^2 - r^2}{m_{R-r}}$$

где $m_{R-r}$ – ошибка разности между $R^2$ и $r^2$, определяемая по формуле:

$$m_{R-r} = \sqrt{\frac{\left(R^2 - r^2\right) - \left(R^2 - r^2\right)^2 \cdot \left(2 - \left(R^2 + r^2\right)\right)}{n}}$$

Если $t_{набл} > t_{кр}$, , различия между рассматриваемыми показателями корреляции существенны и замена нелинейной регрессии уравнением линейной функции невозможна. Практически, если величина $t<2$, то различия между $R$ и $r$ несущественны, и, следовательно, возможно применение линейной регрессии, даже если есть предположения о некоторой нелинейности рассматриваемых соотношений признаков фактора и результата.

### *1.4.3. Коэффициент эластичности как характеристика силы связи фактора с результатом.*

Коэффициент эластичности представляет собой показатель силы связи фактора *x*г с результатом *y,* показывающий, на сколько процентов изменится значение *y* при изменении значения фактора на 1 %. Коэффициент эластичности *(Э)* рассчитывается как относительное изменение *y* на единицу относительного изменения *x*:

$$Э = \frac{dy}{dx} : \frac{y}{x} = \frac{dy}{dx} \cdot \frac{x}{y}.$$

Различают обобщающие (средние) и точечные коэффициенты эластичности.

Обобщающий коэффициент эластичности рассчитывается для среднего значения $\bar{x}$: $Э_{(\bar{x})} = \frac{dy}{dx} \cdot \frac{\bar{x}}{y(\bar{x})}$ и показывает, на сколько процентов изменится *y* относительно своего среднего уровня при росте *x* на 1 % относительно своего среднего уровня.

Точечный коэффициент эластичности рассчитывается для конкретного значения $x = x_0$: $Э_{(x_0)} = \frac{dy}{dx} \cdot \frac{x_0}{y(x_0)}$ и показывает, на сколько процентов изменится *y* относительно уровня $y(x_0)$ при увеличении *x* на 1% от уровня $x_0$.

В зависимости от вида зависимости между *x* и *y* формулы расчета коэффициентов эластичности будут меняться. Основные формулы приведены в табл. 1.7.

Таблица 1.7

| Вид функции $y = f(x)$ | Точечный коэффициент эластичности | Средний коэффициент эластичности |
|---|---|---|
| Линейная $y = b_0 + b_1 \cdot x$ | $Э_{(x_0)} = \frac{b_1 \cdot x_0}{b_0 + b_1 \cdot x_0}$ | $Э_{(\bar{x})} = b_1 \frac{\bar{x}}{y(\bar{x})}$ |
| Парабола $y = a + b \cdot x + c \cdot x^2$ | $Э_{(x_0)} = \frac{(b + 2c \cdot x_0) \cdot x_0}{a + b \cdot x_0 + c \cdot x_0^{\,2}}$ | $Э_{(\bar{x})} = \frac{(b + 2c \cdot \bar{x})\bar{x}}{y(\bar{x})}$ |
| Равносторонняя гипербола $y = a + b / x$ | $Э_{(x_0)} = \frac{-b}{a \cdot x_0 + b}$ | $Э_{(\bar{x})} = \frac{-b}{a \cdot \bar{x} + b}$ |
| Степенная $y = a \cdot x^b$ | $Э_{(x_0)} = b$ | $Э_{(\bar{x})} = b$ |
| Показательная $y = a \cdot b^x$ | $Э_{(x_0)} = x_0 \cdot \ln b$ | $Э_{(\bar{x})} = \bar{x} \cdot \ln b$ |

Только для степенных функций $y = a \cdot x^b$ коэффициент эластичности представляет собой постоянную независящую от $x$ величину (равную в данном случае параметру *b).* Именно поэтому степенные функции широко используются в эконометрических исследованиях. Параметр $b$ в таких функциях имеет четкую экономическую интерпретацию – он показывает процентное изменение результата при увеличении фактора на 1 %. Так, если зависимость спроса $y$ от цен $p$ характеризуется уравнением вида: $y = 200\,p^{-1,5}$, то, следовательно, с увеличением цен на 1 % спрос снижается в среднем на 1,5 %.

Несмотря на широкое использование в эконометрике коэффициентов эластичности, возможны случаи, когда их расчет экономического смысла не имеет. Это происходит тогда, когда для рассматриваемых признаков бессмысленно определение изменения значений в процентах. Например, бессмысленно определять, на сколько процентов изменится заработная плата с ростом возраста рабочего на 1 %-. В такой ситуации степенная функция, даже если она оказывается наилучшей по формальным соображениям (исходя из наибольшего значения $R^2$), не может быть экономически интерпретирована.

**Тесты по разделу**

**1. Тест Фишера является**

a) односторонним
b) многосторонним
c) многокритериальным
d) двусторонним

**2. При стремлении размера выборки к бесконечности стандартное отклонение математического ожидания стремится к**

a) 0
b) ½
c) 2
d) 1

**3. Если выборка достаточно полно отражает изучаемые параметры генеральной совокупности, то ее называют**

a) полной
b) хорошей
c) параметрической
d) репрезентативной

**4. Необходимость применения специальных статистических методов для обработки экономической информации вызвана _____ данных**

a) регулярной периодичностью
b) большой размерностью
c) стохастической природой
d) взаимозависимостью

**5. Эконометрика – часть экономической науки, занимающаяся разработкой и применением _____ методов анализа экономических процессов**

a) математических
b) качественных
c) структурных
d) экспертных

**6. Общая, объясненная и необъясненная суммы квадратов отклонений находятся в следующих соотношениях**

a) RSS=TSS/ESS
b) TSS=RSS+ESS
c) ESS=TSS/RSS
d) TSS=RSS–ESS

**7. Эластичность *y* по *x* рассчитывается _____ величины относительного изменения *y* на величину относительного изменения *x*:**

a) делением
b) умножением
c) увеличением
d) уменьшением

**8. Оценка *a* для параметра уравнения парной регрессии при использовании МНК вычисляется по формуле *a* =**

a) $\bar{y}+\bar{x}$
b) $b\bar{x}$
c) $\bar{y}-b\bar{x}$
d) $\bar{y}+b\bar{x}$

**9. Случайный член *v* в уравнении $y=\alpha x^{\beta}v$ изменяет выражение $\alpha x^{\beta}$**

a) в случайной пропорции
b) в несколько раз
c) на случайную величину
d) на фиксированную величину

**10. Если все наблюдения лежат на линии регрессии, то коэффициент детерминации $R^2$ для модели парной регрессии равен**

a) единице
b) нулю
c) 2/3
d) 1/2

**11. Совокупность значений случайной величины и вероятностей, с которыми она их принимает, называют _____ случайной величины**

a) математическим ожиданием
b) законом распределения
c) дисперсией
d) ковариацией

**12. Оценка стандартного отклонения случайной величины, полученная по данным выборки, называется стандартной _____ случайной величины**

a) ошибкой
b) записью
c) оценкой
d) поправкой

**13. Логарифмическое преобразование позволяет осуществить переход от нелинейной модели $y=\alpha x^{\beta}u$ к модели**

a) $ln\,y = \alpha + \beta x + u$

b) $y = ln\alpha + \beta\, ln\, x + ln u$

c) $ln\,y = ln\alpha + \beta x + u$

d) $ln\,y = ln\alpha + \beta\, ln\, x + ln u$

**14. Если случайный член в исходном нелинейном уравнении регрессии является аддитивным, то логарифмирование _____ линеаризации**

a) не приводит к
b) приводит к
c) может привести к
d) препятствует

**15. Метод наименьших квадратов – метод нахождения оценок параметров регрессии, основанный на минимизации _____ квадратов остатков всех наблюдений**

a) разности
b) суммы
c) произведения
d) среднего арифметического

**16. Первое условие Гаусса-Маркова заключается в том, что _____ для любого *i***

a) $M(u_i) = 1$

b) $\sigma^2(u_i) = 0$

c) $\sigma^2(u_i) = 1$

d) $M(u_i) = 0$

**17. Из перечисленного: 1) увеличение количества наблюдений в выборке *n*; 2) уменьшение количества наблюдений в выборке *n*; 3) уменьшение диапазона наблюдений *Vаг(x);* 4) увеличение диапазона наблюдений *Vаг(x);* 5) уменьшение $\sigma^2(u)$; 6) увеличение $\sigma^2(u)$ – для улучшения точности оценок по МНК можно использовать**

a) 1,4,5
b) 1,3,6
c) 2,4,5
d) 2,3,6

**18. Регрессором в уравнении парной линейной регрессии называется**

a) первый параметр
b) объясняющая переменная
c) зависимая переменная
d) случайный член

**19. Доля объясненной дисперсии зависимой переменной во всей выборочной дисперсии *y* выражается коэффициентом**

a) регрессии
b) корреляции
c) детерминации
d) вариации

**20. Граничное значение области принятия гипотезы *p*%-ной вероятностью совершить ошибку I рода определяется _____ при *p*-процентном уровне значимости**

a) стандартной ошибкой коэффициента
b) гипотетическим значением коэффициента
c) критическим значением теста
d) стандартным отклонением коэффициента

**21. Остаток в наблюдении равен _____, если $y_i$ - истинное значение переменной *y* в *i*-ом наблюдении, $x_i$ - результат *i*-го наблюдения, *a* и *b* - оценки параметров модели линейной регрессии**

a) $\sum y_i - \sum (a + bx_i)$
b) $y_i / (a + bx_i)$
c) $y_i - \sum (a + bx_i)$
d) $y_i - (a + bx_i)$

**22. Чем ближе коэффициент детерминации $R^2$ к 1, тем ближе выборка $\{(x_i, y_i)\}$ к**

a) центру тяжести
b) форме круга
c) линии регрессии
d) истинной прямой

**23. Отличие одностороннего теста от двустороннего заключается в том, что он имеет только**

a) одно распределение
b) одну оценку
c) один параметр
d) одно критическое значение

**24. При введении случайного компонента в экономическую модель связи ее остальных переменных приобретают _____характер**

a) логарифмический
b) детерминированный
c) линейный
d) стохастический

**25. Наука, изучающая методы обработки результатов наблюдений массовых случайных явлений**

a) математическая статистика
b) математический анализ
c) теория вероятностей
d) эконометрика

## Вопросы для повторения раздела

**1.** Что такое генеральная совокупность и выборка?

**2.** Дайте определения и приведите как определяются основные числовые характеристики по результатам выборки: выборочное среднее, дисперсия, среднее квадратическое отклонение?

**3.** Как связаны между собой случайные величины, имеющие стандартизированное нормальное распределение, распределения Стьюдента, $\chi^2$ и Фишера?

**4.** Справедливо или ложно утверждение, что при увеличении числа степеней свободы распределения Стьюдента, $\chi^2$ и Фишера стремятся к стандартизированному нормальному распределению?

**5.** Перечислите свойства ковариации.

**6.** Приведите свойства коэффициента корреляции.

**7.** Что такое функция регрессии?

**8.** Назовите основные причины наличия в регрессионной модели случайного отклонения.

**9.** Назовите основные этапы регрессионного анализа.

**10.** Что понимается под спецификацией модели, и как она осуществляется?

**11.** Дайте определения несмещенности, эффективности и состоятельности оценок.

**12.** Какие выводы можно сделать об оценках коэффициентов регрессии и случайного отклонения, полученных по МНК?

**13.** Что такое статистическая гипотеза и какова цель ее проверки?

**14.** Что такое нулевая и альтернативная гипотеза? Назовите принципы их построения. Приведите общую схему проверки гипотез.

# 2. Модель множественной регрессии

## 2.1. Множественная линейная регрессия

### *2.1.1. Нормальная линейная модель множественной регрессии*

Естественным обобщением линейной регрессии с двумя переменными является многомерная регрессионная модель (multiple regression model) или модель множественной регрессии:

$$y_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + \ldots + b_j x_{ji} + \ldots + b_m \cdot x_{mi} + u_i$$

где $y_i$ – значение признака-результата (зависимой переменной) для $i$-го наблюдения;

$x_{ji}$ – значение $j$-го фактора (независимей или объясняющей переменной) $(j = 1;m)$ для $i$-го наблюдения;

$u_i$ – случайная составляющая результативного признака для $i$-го наблюдения;

$b_o$ – свободный член, который формально показывает среднее значение $y$ при $x_1 = x_2 = \ldots = x_m = 0$;

$b_i$ – коэффициент «чистой» регрессии при $j$-$m$ факторе ($j=1,m$). Он характеризует среднее изменение признака-результата $y$ с изменением соответствующего фактора $x_j$. на единицу, при условии, что прочие факторы модели не изменяются и фиксированы на средних уровнях.

Обычно для многомерной регрессионной модели делаются следующие предпосылки.

1. $\left(x_{1i}, x_{2i}, \ldots, x_{ji}, \ldots, x_{mi}\right)$ – детерминированные (нестохастические) переменные.

2. $M\left(u_i\right) = 0$, $(i = 1, n)$ – математическое ожидание случайной составляющей равно 0 в любом наблюдении.

3. $M\left(u_i^{\,2}\right) = \sigma_{ui}^2 = \sigma_u^2 = const$, $(i = 1, n)$ – теоретическая дисперсия случайной составляющей; постоянна для всех наблюдений.

4. $M\left(u_i \cdot u_j\right) = Cov\left(u_i, u_j\right) = 0, \left(i \neq j\right)$ – отсутствие систематической связи между значениями случайной составляющей в любых двух наблюдениях.

5. Часто добавляется условие: $u_i \approx N\left(0, \sigma^2\right)$, т. е. $u_i$ – нормально распределенная случайная величина.

Модель линейной множественной регрессии, для которой выполняются данные предпосылки, называется нормальной линейной регрессионной (Classical Normal Regression model).

В матричной форме нормальная (классическая) регрессионная, модель имеет вид:

$$Y = X \cdot b + u\,,$$

где $Y$ – случайный вектор-столбец размерности ($n\times1$) наблюдаемых значений результативного признака;

$X$ – матрица размерности ($n\times(m+1)$) наблюдаемых значений факторных признаков. Добавление 1 к общему числу факторов $m$ учитывает свободный член $b_0$ в уравнении регрессии. Значения фактора $x_0$ для свободного члена принято считать равным единице;

$b$ – вектор-столбец размерности (($m+1)\times1$) неизвестных, подлежащих оценке параметров модели (коэффициентов регрессии);

$u$ – случайный вектор-столбец размерности (n×1) ошибок наблюдений.

$$Y=\begin{pmatrix} y_1 \\ \cdots \\ y_i \\ \cdots \\ y_n \end{pmatrix} \quad X=\begin{pmatrix} 1x_{11}\ldots x_{m1} \\ \ldots\ldots\ldots\ldots \\ 1x_{1i}\ldots x_{mi} \\ \ldots\ldots\ldots\ldots \\ 1x_{1n}\ldots x_{mn} \end{pmatrix} \quad b=\begin{pmatrix} b_0 \\ b_1 \\ \cdots \\ b_j \\ \cdots \\ b_m \end{pmatrix} \quad u=\begin{pmatrix} u_1 \\ \cdots \\ u_i \\ \cdots \\ u_n \end{pmatrix}$$

Предпосылки данной модели:

1) $\left(x_1,x_2,\ldots,x_j,\ldots,x_m\right)$ – детерминированные (нестохастические) переменные, т. е. ранг матрицы $X$ равен $m+1<n$;

2) $M\left(u\right)=0^n$;

3,4) ковариационная матрица должна иметь вид:

$$C_u=\begin{pmatrix} \sigma_u^2 0\ldots0 \\ 0\sigma_u^2\ldots0 \\ \cdots\cdots\cdots \\ 00..\sigma_u^2 \end{pmatrix}=\sigma_u^2\cdot I\,,$$

где $\sigma_u^2$ – дисперсия случайной составляющей;

$I$ – единичная матрица размером $n\times m$;

5) $u_i\approx N\left(0,\sigma^2\right)$.

Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям.

1. Они должны быть количественно измеримы. Если не обходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность (например, в модели урожайности качество почвы задается в виде баллов).

2. Каждый фактор должен быть достаточно тесно связан с результатом (т. е. коэффициент парной линейной корреляции между фактором и результатом должен существенно отличаться от нуля).

3. Факторы не должны сильно коррелировать друг с другом, тем более находиться в строгой функциональной связи (т. е. они не должны быть интеркоррелированны).

### *2.1.2. Традиционный метод наименьших квадратов для многомерной регрессии (OLS)*

Основная задача регрессионного анализа заключается в нахождении по выборке объемом $n$ оценки неизвестных коэффициентов регрессии $(b_0, b_1, \ldots b_m)$ модели или вектора $b$.

Оценка параметров многомерной модели, как и в случае парной регрессии, осуществляется обычно традиционным методом наименьших квадратов (МНК). Согласно данному методу, в качестве оценки вектора $b$ принимают вектор $\tilde{b}$, который минимизирует сумму квадратов отклонений наблюдаемых значений $y_i$ от рассчитанных по модели $\hat{y}_i$.

В матричной форме функционал $S$ будет записан так:

$$S = \left(Y - X\tilde{b}\right)^T \left(Y - X\tilde{b}\right) = \sum_{i=1}^{m} \left(y_i - \hat{y}_i\right) \Rightarrow \underset{\tilde{b}0,\tilde{b}1,\ldots\tilde{b}k}{min}$$

МНК-оценки в матричной форме находят по формулам:

$$\tilde{b} = \left(X^T X\right)^{-1} X^T Y, \text{ где } \tilde{b} = \begin{pmatrix} \tilde{b}_0 \\ \tilde{b}_1 \\ \vdots \\ \tilde{b}_j \\ \vdots \\ \tilde{b}_m \end{pmatrix}.$$

Оценим с помощью МНК параметры линейной двухфакторной модели: $y_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + u_i$, *i=1; n.* Для этого минимизируем функционал:

$$S = \sum_{i=1}^{m} \left(y_i - \hat{y}_i\right) = \sum_{i=1}^{m} \left( y_i - \left(\tilde{b}_0 + \tilde{b}_1 \cdot x_{1i} + \tilde{b}_2 \cdot x_{2i}\right)^2 \right) \Rightarrow \underset{\tilde{b}0,\tilde{b}1,\tilde{b}2}{min}$$

Функционал S является функцией трех переменных $\tilde{b}_0, \tilde{b}_1, \tilde{b}_2$. Чтобы найти экстремум функции нескольких переменных, нужно взять ее частные производные по этим переменным и приравнять их нулю:

$$\frac{\partial S}{\partial \tilde{b}_0} = 0, \ \frac{\partial S}{\partial \tilde{b}_1} = 0, \ \frac{\partial S}{\partial \tilde{b}_2} = 0.$$

Получим следующую систему нормальных линейных уравнений:

$$\begin{cases} \sum_{i=1}^{n} y_i = \tilde{b}_0 \cdot n + \tilde{b}_1 \sum_{i=1}^{n} x_{1i} + \tilde{b}_2 \sum_{i=1}^{n} x_{2i}; \\ \sum_{i=1}^{n} y_i \cdot x_{1i} = \tilde{b}_0 \cdot x_{1i} + \tilde{b}_1 \sum_{i=1}^{n} x_{1i}^2 + \tilde{b}_2 \sum_{i=1}^{n} x_{2i} \cdot x_{1i}; \\ \sum_{i=1}^{n} y_i \cdot x_{2i} = \tilde{b}_0 \cdot x_{2i} + \tilde{b}_1 \sum_{i=1}^{n} x_{1i} \cdot x_{2i} + \tilde{b}_2 \sum_{i=1}^{n} x_{2i}^2. \end{cases}$$

Параметры этой системы могут быть найдены, например, методом К. Гаусса, либо методом итераций.

Рассмотрим **пример**. Для данных табл. 2.1 найдем МНК-оценки параметров линейного двухфакторного уравнения регрессии: $y_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + u_i$.

Расчет необходимых сумм для системы нормальных линейных уравнений сведем в табл. 2.1.

Таблица 2.1

| $i$ | $y$ – заработная плата, \$ | $x_1$ – возраст, лёт | $x_2$ – выработка, шт./смену | $yx_1$ | $yx_2$ | $x_1^2$ | $x_2^2$ | $x_1x_2$ |
|---|---|---|---|---|---|---|---|---|
| 1 | 300 | 29 | 17 | 8700 | 5100 | 841 | 289 | 493 |
| 2 | 400 | 40 | 25 | 16000 | 10000 | 1600 | 625 | 1000 |
| 3 | 300 | 36 | 15 | 10800 | 4500 | 1296 | 225 | 540 |
| 4 | 320 | 32 | 17 | 10240 | 5440 | 1024 | 289 | 544 |
| 5 | 200 | 23 | 15 | 4600 | 3000 | 529 | 225 | 345 |
| 6 | 350 | 45 | 18 | 15750 | 6300 | 2025 | 324 | 810 |
| 7 | 350 | 38 | 17 | 13300 | 5950 | 1444 | 289 | 646 |
| 8 | 400 | 40 | 25 | 16000 | 10000 | 1600 | 625 | 1000 |
| 9 | 380 | 50 | 19 | 19000 | 7220 | 2500 | 361 | 950 |
| 10 | 400 | 47 | 23 | 18800 | 9200 | 2209 | 529 | 1081 |
| 11 | 250 | 28 | 15 | 7000 | 3750 | 784 | 225 | 420 |
| 12 | 350 | 30 | 18 | 10500 | 6300 | 900 | 324 | 540 |
| 13 | 200 | 25 | 16 | 5000 | 3200 | 625 | 256 | 400 |
| 14 | 400 | 48 | 23 | 19200 | 9200 | 2304 | 529 | 1104 |
| 15 | 220 | 30 | 18 | 6600 | 3960 | 900 | 324 | 540 |
| 16 | 320 | 40 | 18 | 12800 | 5760 | 1600 | 324 | 720 |
| 17 | 390 | 40 | 25 | 15600 | 9750 | 1600 | 625 | 1000 |
| 18 | 360 | 38 | 23 | 13680 | 8280 | 1444 | 529 | 874 |
| 19 | 260 | 29 | 18 | 7540 | 4680 | 841 | 324 | 522 |
| 20 | 250 | 25 | 17 | 6250 | 4250 | 625 | 289 | 425 |
| $\Sigma$ | **6400** | **713** | **382** | **237360** | **125840** | **26691** | **7530** | **13954** |

Тогда система нормальных линейных уравнений будет иметь вид:

$$\begin{cases} 6400 = \tilde{b}_0 \cdot 20 + \tilde{b}_1 \cdot 713 + \tilde{b}_2 \cdot 382; \\ 237360 = \tilde{b}_0 \cdot 713 + \tilde{b}_1 \cdot 26691 + \tilde{b}_2 \cdot 13954; \\ 125840 = \tilde{b}_0 \cdot 382 + b_1 \cdot 13954 + b_2 \cdot 7530. \end{cases}$$

Решив систему, найдем значения $\tilde{b}_0, \tilde{b}_1, \tilde{b}_2$:

$\tilde{b}_0 = -16{,}04$; $\tilde{b}_1 = 5{,}1$; $\tilde{b}_2 = 8{,}08$.

Найдем МНК-оценки для нашего примера матричным способом.

Воспользовавшись правилами умножения матриц будем иметь:

$X^T X =$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 29 & 40 & 36 & 32 & 23 & 45 & 38 & 40 & 50 & 47 & 28 & 30 & 25 & 48 & 30 & 40 & 40 & 38 & 29 & 25 \\ 17 & 25 & 15 & 17 & 15 & 18 & 17 & 25 & 19 & 23 & 15 & 18 & 16 & 23 & 18 & 18 & 25 & 23 & 18 & 17 \end{pmatrix} \begin{pmatrix} 12917 \\ 14025 \\ 13615 \\ 13217 \\ 12315 \\ 14518 \\ 13817 \\ 14025 \\ 15019 \\ 14723 \\ 12815 \\ 13018 \\ 12516 \\ 14823 \\ 13018 \\ 14018 \\ 14025 \\ 13823 \\ 12918 \\ 12517 \end{pmatrix} =$$

$$= \begin{pmatrix} 20 & 713 & 382 \\ 713 & 26691 & 13954 \\ 382 & 13954 & 7530 \end{pmatrix}.$$

$X^T X =$

$$= \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 29 & 40 & 36 & 32 & 23 & 45 & 38 & 40 & 50 & 47 & 28 & 30 & 25 & 48 & 30 & 40 & 40 & 38 & 29 & 25 \\ 17 & 25 & 15 & 17 & 15 & 18 & 17 & 25 & 19 & 23 & 15 & 18 & 16 & 23 & 18 & 18 & 25 & 23 & 18 & 17 \end{pmatrix} \begin{pmatrix} 300 \\ 400 \\ 300 \\ 320 \\ 200 \\ 350 \\ 350 \\ 400 \\ 380 \\ 400 \\ 250 \\ 350 \\ 200 \\ 400 \\ 220 \\ 320 \\ 390 \\ 360 \\ 260 \\ 250 \end{pmatrix} = \begin{pmatrix} 6400 \\ 237360 \\ 125840 \end{pmatrix}$$

Найдем обратную матрицу.

Матрицей, обратной к матрице $A$, называется матрица $A^{-1}$ такая, что $AA^{-1}=I$ ($I$ – единичная матрица).

Обозначим $a_{ij}$ элементы матрицы $A^{-1}$. Тогда $a_{ij}=\dfrac{(-1)^{i+j}\left|M_{ij}\right|}{\left|A\right|}$, где $M_{ij}$ – матрица, получающаяся из $A$ вычеркиванием $i$-й строки и $j$-го столбца. Для нашего примера:

$$\left|A\right|=\begin{vmatrix}20 & 713 & 382\\ 713 & 26691 & 13954\\ 382 & 13954 & 7530\end{vmatrix}=3696554;$$

$$\left|M_{11}\right|=\begin{vmatrix}26691 & 13954\\ 13954 & 7530\end{vmatrix}=6269114;$$

$a_{11}=(-1)^2\cdot 6269114/3696554=1{,}6959$ и т.д.

В результате получим:

$$\left(X^T X\right)^{-1}=\begin{pmatrix}20 & 713 & 382\\ 713 & 26691 & 13954\\ 382 & 13954 & 7530\end{pmatrix}^{-1}=$$

$$=\begin{pmatrix}1{,}695935 & -0{,}0104 & -0{,}06675\\ -0{,}0104 & 0{,}001265 & -0{,}00182\\ -0{,}06675 & -0{,}00182 & 0{,}006885\end{pmatrix}$$

Тогда вектор оценок коэффициентов регрессии равен:

$$\tilde{b}=\begin{pmatrix}1{,}695935 & -0{,}0104 & -0{,}06675\\ -0{,}0104 & 0{,}001265 & -0{,}00182\\ -0{,}06675 & -0{,}00182 & 0{,}006885\end{pmatrix}\cdot\begin{pmatrix}6400\\ 237360\\ 125840\end{pmatrix}=\begin{pmatrix}-16{,}039\\ 5{,}099019\\ 8{,}076387\end{pmatrix}$$

То есть $\tilde{b}_0=-16{,}039$; $\tilde{b}_1=5{,}099019$; $\tilde{b}_2=8{,}076387$ (оценки такие же, что и найденные 1-м способом).

Кроме того, для линейной множественной регрессии существует другой способ оценки параметров – через $\beta$-коэффициенты (параметры уравнения регрессии в стандартных масштабах).

При построении *уравнения регрессии в стандартном масштабе* все значения исследуемых признаков переводятся в стандарты (стандартизованные значения) по формулам:

$$t_{x_{ji}}=\frac{x_{ji}-\bar{x}_j}{\sigma_{x_j}},\ j=1;m,$$

где $x_{ji}$ – значение переменной $x_j$ в $i$-м наблюдении.

$$t_{y_i}=\frac{y_i-\bar{y}}{\sigma_y}.$$

Таким образом, начало отсчета каждой стандартизованной переменной совмещается с ее средним значением, а в качестве единицы изменения принимается ее среднее квадратическое отклонение ($\sigma$). Если связь между переменными в естественном масштабе линейная, то изменение начала отсчета и единицы измерения этого свойства не нарушат, так что и стандартизованные переменные будут связаны линейным соотношением:

$$t_y' = \sum_{j=1}^{m} \beta_j \cdot t_{x_j} .$$

$\beta$-коэффициенты могут быть оценены с помощью обычного МНК.

При этом система нормальных уравнений будет иметь вид:

$$\begin{cases} r_{x1y} = \beta_1 + r_{x1x2}\beta_2 + \ldots + r_{x1xm}\beta_m; \\ r_{x2y} = r_{x2x1}\beta_1 + \beta_2 + \ldots + r_{x2xm}\beta_m; \\ \ldots \\ r_{xmy} = r_{xmx1}\beta_1 + r_{xmx2}\beta_2 + \ldots + \beta_m. \end{cases}$$

(так как $r_{x_j x_k} = r_{t_j t_k} = \sum_{i=1}^{n} t_{x_j i} \cdot t_{x_k i}$ ).

Найденные из данной системы $\beta$-коэффициенты позволяют определить значения коэффициентов регрессии в естественном масштабе по формулам:

$$\tilde{b}_j = \beta_j \cdot \frac{\sigma_y}{\sigma_x}, \; j=1;m; \; a = y - \sum_{j=1}^{m} b_j \cdot \bar{x}_j .$$

Найдем $\beta$-коэффициенты для нашего примера. Система нормальных линейных уравнений будет иметь вид (воспользуемся данными корреляционной матрицы, рассчитанной в предыдущем вопросе):

$$\begin{cases} 0{,}853056 = \beta_1 + 0{,}615448 \cdot \beta_2; \\ 0{,}778766 = \beta_2 + 0{,}615448 \cdot \beta_1. \end{cases}$$

Тогда $\beta_1 = 0{,}60166$, $\beta_2 = 0{,}408476$.

Отсюда $\tilde{b}_1 = 0{,}60166 \cdot \frac{67{,}6018}{7{,}9767} = 5{,}1$. $\left(\sigma_y = 67{,}6018; \sigma_{x1} = 7{,}9767\right)$;

$\tilde{b}_2 = 0{,}408476 \cdot \frac{67{,}6018}{3{,}4191} = 8{,}1$ $\left(\sigma_y = 67{,}6018; \sigma_{x2} = 3{,}4191\right)$;

$\tilde{b}_0 = 320 - 5{,}1 \cdot 35{,}65 - 8{,}1 \cdot 19{,}1 = -16$ $\left(\bar{y} = 320; \bar{x}_1 = 35{,}65; \bar{x}_2 = 19{,}1\right)$

Оцененное уравнение регрессии для нашего примера будет иметь вид: $\hat{y}_{x1,x2} = -16{,}04 + 5{,}1 \cdot x_1 + 8{,}08 \cdot x_2$ .

Дадим интерпретацию параметров данного уравнения.

Параметр $\tilde{b}_1$ =–5,099019 показывает, что заработная плата рабочего в среднем увеличивается на 5$ при увеличении возраста рабочего на 1 год при условии, что выработка рабочего не меняется и фиксирована на среднем

уровне; параметр $\tilde{b}_2$=8,076387 показывает, что заработная плата рабочего в среднем увеличивается на 8$ при увеличении выработки рабочего за смену на 1 штуку при условии, что возраст рабочего не изменился и фиксирован на среднем уровне.

Параметр $\tilde{b}_0$ мы не интерпретируем, т. к. в выборке отсутствуют значения признаков $x_1$ и $x_2$, близкие к нулю.

### *2.1.3. Показатели тесноты связи фактора с результатом. Коэффициенты частной эластичности и стандартизированные коэффициенты регрессии ($\beta$ – коэффициенты)*

Если факторные признаки различны по своей сущности и/или имеют различные единицы измерения, то коэффициенты регрессии $b_j$ уравнения: $y_i = b_0 + b_1 \cdot x_{1i} + b_2 \cdot x_{2i} + u_i$ являются несопоставимыми. Поэтому уравнение регрессии дополняют соизмеримыми показателями тесноты связи фактора с результатом, позволяющими ранжировать факторы по силе влияния на результат. К таким показателям тесноты связи относят: частные коэффициенты эластичности, $\beta$ -коэффициенты и другие.

Частные коэффициенты эластичности $Э_j$ рассчитываются по формуле:

$$Э_j = \frac{\partial Y}{\partial X_j} \cdot \frac{\overline{X_j}}{\overline{Y_{x1,\ldots,xm}}} \quad (j = 1; m),$$

где $\overline{X_j}$ – среднее значение фактора $x_j$;

$\overline{Y}$ – среднее значение результата $y$.

Частный коэффициент эластичности показывает, насколько процентов в среднем изменяется признак-результат $y$ с увеличением признака-фактора $x_j$ на 1 % от своего среднего уровня при фиксированном положении других факторов модели. В случае линейной зависимости $Э_j$ рассчитываются по формуле:

$$Э_j = b_j \cdot \frac{\overline{X_j}}{\overline{Y_{x1,\ldots,xm}}},$$

где $b_j$ – коэффициент регрессии при $j$-м факторе.

Стандартизированные частные коэффициенты регрессии $\beta$ -коэффициенты ($\beta_j$) показывают, на какую часть своего среднего квадратического отклонения $\sigma_y$ изменится признак-результат $y$ с увеличением

соответствующего фактора $x_j$ на величину своего среднего квадратического отклонения ($\sigma_{xj}$) при неизменном влиянии прочих факторов модели.

Частные коэффициенты эластичности и стандартизованные частные коэффициенты регрессии можно использовать для ранжирования факторов по силе влияния на результат. Чем больше величина $Э_j$ или $\beta_j$, тем сильнее влияет фактор $x_i$ на результат $y$.

**Пример.** Рассмотрим ранжирование факторов на примере. Исходные данные были приведены в табл. 2.1. Воспользуемся результатами оценивания регрессии заработной платы рабочих $y$ по возрасту $x_1$ и выработке $x_2$:

$$\hat{y}_{x1,x2} = -16{,}04 + 5{,}1 \cdot x_1 + 8{,}08 \cdot x_2 \qquad (2.1)$$

(см. п. 2.1.2).

Частный коэффициент эластичности для фактора «возраст» будет равен: $Э_1 = 5{,}1 \cdot \frac{35{,}65}{320} = 0{,}568 \quad (\bar{y} = 320; \bar{x} = 19{,}1)$.

Частный коэффициент эластичности для фактора «выработка» равен: $Э_1 = 8{,}08 \cdot \frac{19{,}1}{320} = 0{,}482 \quad (\bar{y} = 320; \bar{x} = 19{,}1)$. Так как $Э_1 > Э_2$, то фактор «возраст» сильнее влияет на заработную плату рабочего (т. е. вызывает более существенное изменение заработной платы), чем фактор «выработка».

$\beta$-коэффициенты для данного примера были рассчитаны в п. 2.1.2 и составили, соответственно, для фактора«возраст», ($x_1$): $\beta_1$ = 0,60166, для фактора «выработка» ($x_2$): $\beta_2$ = 0,408476. Так как $\beta_1$, > $\beta_2$, то фактор «возраст» сильнее влияет на заработную плату рабочего (т. е. вызывает более существенное изменение заработной платы), чем фактор«выработка».

По коэффициентам эластичности и $\beta$-коэффициентам могут быть сделаны противоположные выводы. Причины этого: а) вариация одного фактора очень велика; б) разнонаправленное воздействие факторов на результат.

Коэффициент $\beta_j$ может также интерпретироваться как показатель прямого (непосредственного) влияния $j$-го фактора $x_i$ на результат $y$. Во множественной регрессии $j$ -й фактор оказывает не только прямое, но и косвенное (опосредованное) влияние на результат (т. е. влияет на результат через другие факторы модели). Косвенное влияние измеряется величиной: $\sum_{i=1,\ldots j-1,j+1,\ldots m} \beta_i \cdot r_{xj,xi}$, где $m$ – число факторов в модели. Полное влияние $j$-го фактора на результат, равное сумме прямого и косвенного влияний, измеряет коэффициент линейной парной корреляции данного фактора и результата – $r_{xj,y}$

.

Так для нашего примера непосредственное влияние фактора «возраст» на результат «заработную плату» в уравнении регрессии (2.1) измеряется $\beta_1$ и составляет 0,60166; косвенное (опосредованное) влияние данного фактора на результат определяется как: $r_{x1x2} \cdot \beta_2 = 0{,}615448 \cdot 0{,}408476 = 0{,}251$.

Непосредственное влияние фактора «выработка» на результат «заработная плата» в уравнении регрессии (3) измеряется $\beta_2$ и составляет 0,408476. Косвенное влияние данного фактора на результат определяется как: $r_{x1x2} \cdot \beta_1 = 0{,}615448 \cdot 0{,}60166 = 0{,}370$.

### *2.1.4. Частная корреляция*

Показатели парной корреляции – $r_{yx}$ характеризуют тесноту связи результата и фактора, не принимая во внимание возможного влияния на результат других факторных признаков. Поэтому во множественном регрессионном анализе возникает проблема определения тесноты связи между двумя признаками в чистом виде, т. е. при устранении воздействия других факторов. Нам под силу исключить влияние только *учтенных в модели* факторов.

Показателем «чистого» влияния фактора на результат при устранении влияния прочих факторов, включенных в модель регрессии, является *частный коэффициент корреляции* или частный индекс корреляции (в зависимости от формы связи).

Пусть исследовалась зависимость $y = f(x_1, u_1)$, для которой остаточная (необъясненная уравнением регрессии) дисперсия равна:

$$\varepsilon^2_{y(x1)} = \frac{\sum\left(y_i - \hat{y}_{(x1)i}\right)^2}{n}.$$

Включив в уравнение регрессии дополнительный фактор $x_2$, т. е. найдя зависимость $y = f(x_1, x_2, u_{12})$, мы получим остаточную дисперсию результата:

$\varepsilon^2_{y(x1x2)} = \dfrac{\sum\left(y_i - \hat{y}_{(x1,x2)i}\right)^2}{n}$, которая будет не больше $\varepsilon^2_{y(x1)}$.

Сокращение остаточной дисперсии за счет дополнительного включения в уравнение регрессии фактора $x_2$ составит: $\varepsilon^2_{y(x1)} - \varepsilon^2_{y(x1,x2)}$. Чем выше доля этого сокращения в исходной дисперсии, т. е. чем выше соотношение $\dfrac{\left(\varepsilon^2_{y(x1)} - \varepsilon^2_{y(x1,x2)}\right)}{\varepsilon^2_{y(x1)}}$, тем теснее связь между $y$ и $x_2$ при постоянном действии $x_1$.

Корень квадратный из этой величины и есть коэффициент частной корреляции результата со вторым фактором при постоянном действии первого фактора:

$$r_{yx2/x1} = \sqrt{\frac{\varepsilon^2_{y(x1)} - \varepsilon^2_{y(x1,x2)}}{\varepsilon^2_{y(x1)}}}.$$

Аналогично можно определить коэффициент частной корреляции результата с первым фактором при постоянном действии второго фактора:

$$r_{yx1/x2} = \sqrt{\frac{\varepsilon^2_{y(x2)} - \varepsilon^2_{y(x1,x2)}}{\varepsilon^2_{y(x2)}}}.$$

В общем виде тесноту связи $y$ и $x_1$ в модели, содержащей $m$ факторов, можно оценить по следующей формуле:

$$r_{yx1/x2\ldots xm} = \sqrt{\frac{\varepsilon^2_{y(x2\ldots xm)} - \varepsilon^2_{y(x1,x2\ldots xm)}}{\varepsilon^2_{y(x2,x3\ldots xm)}}}$$

(в обозначении коэффициента частной корреляции после знака «/» перечисляются те факторы модели, влияние которых устраняется).

Рассмотренные частные коэффициенты корреляции являются коэффициентами частной корреляции первого порядка. Порядок частного коэффициента корреляции определяется числом факторов, влияние которых исключается. Для коэффициента парной корреляции $r_{yx}$ порядок равен 0.

Для расчета частных коэффициентов корреляции могут быть использованы парные коэффициенты корреляции. Для двухфакторной модели регрессии можно вычислить следующие коэффициенты частной корреляции первого порядка:

$$r_{yx1/x2} = \frac{r_{x1y} - r_{x2y} \cdot r_{x1x2}}{\sqrt{\left(1 - r^2_{x1x2}\right)\left(1 - r^2_{x2y}\right)}}$$

(фактор $x_2$ фиксирован).

Можно рассчитать также коэффициент частной корреляции, измеряющий тесноту связи между $x_1$ и $x_2$ при фиксации признака-результата $y$:

$$r_{x1x2/y} = \frac{r_{x2x1} - r_{x1y} \cdot r_{x2y}}{\sqrt{\left(1 - r^2_{yx2}\right)\left(1 - r^2_{x1y}\right)}}.$$

Для трехфакторной модели регрессии можно рассчитать частные коэффициенты корреляции 2-го порядка:

$$r_{yx1/x2x3} = \frac{r_{x1y/x2} - r_{yx3/x2} \cdot r_{x1x3/x2}}{\sqrt{\left(1 - r^2_{x1x3/x2}\right)\left(1 - r^2_{yx3/x2}\right)}};$$

$$r_{yx2/x1x3} = \frac{r_{x2y/x1} - r_{yx3/x1} \cdot r_{x2x3/x1}}{\sqrt{\left(1 - r^2_{x2x3/x1}\right)\left(1 - r^2_{yx3/x1}\right)}};$$

$$r_{yxi/x1x2\ldots xi-1,xi+1\ldots xm} =$$

$$= \frac{r_{yxi/x1x2\ldots xi-1,xi+1\ldots xm-1} - r_{yxm/x1x2\ldots xi-1,xi+1\ldots xm-1} \cdot r_{xi,xm/x1\ldots xi-1,xi+1\ldots xm-1}}{\sqrt{\left(1-r^2_{xi,xm/x1\ldots xi-1,xi+1\ldots xm-1}\right)\left(1-r^2_{yxm/x1x2\ldots xi-1,xi+1\ldots xm-1}\right)}}$$

Если рассматривается регрессия с числом факторов $m$, то возможны частные коэффициенты корреляции не только первого, но и второго,..., $(m-1)$-го порядка;

На практике наибольший интерес представляют частные коэффициенты корреляции самого высокого порядка.

Данные формулы для расчета частных коэффициентов корреляции $j$-го порядка через коэффициенты частной корреляции $(j-1)$-го порядка называются рекуррентными.

Частные коэффициенты корреляции могут быть рассчитаны и другим способом. Если выразить остаточную дисперсию через коэффициент множественной детерминации, то в двухфакторной модели:

$$r_{yx2/x1} = \sqrt{1-\frac{1-R^2_{y(x1x2)}}{1-r^2_{yx1}}} = \sqrt{1-\frac{R^2_{y(x1x2)}-r^2_{yx1}}{1-r^2_{yx1}}}\ .$$

Для общего случая (модель содержит $m$ факторов) частные коэффициенты корреляции можно определить таким образом:

$$r_{yxi/x1\ldots xi-1,xi+1\ldots xm} = \sqrt{1-\frac{1-R^2_{y(x1\ldots xm)}}{1-R^2_{y(x1\ldots xj-1,xj+1\ldots xm)}}}\ .$$

где $R^2_{y(x1\ldots xm)}$ — коэффициент множественной детерминации $y$ с комплексом факторов: $x_1 \ldots x_{\text{m}}$.

$R^2_{y(x1\ldots xj-1,xj+1\ldots xm)}$ — коэффициент множественной детерминации $y$ с комплексом факторов: $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_{\text{m}}$.

Частные коэффициенты корреляции, рассчитанные 1-м способом по рекуррентным формулам, изменяются от –1 до +1, а рассчитанные 2-м способом, через множественные коэффициенты детерминации, — от 0 до 1. Чем ближе к единице модуль частного коэффициента корреляции, тем теснее связь фактора с результатом при устранении влияния прочих факторов, включенных в модель регрессии.

Частные коэффициенты корреляции используются не только для ранжирования факторов модели по степени влияния на результат, но и также для отсева факторов. При малых значениях $r_{yxm/x1,x2,\ldots xm-1}$ нет смысла вводить в уравнение $m$-й фактор, т. к. качество уравнения регрессии при его введении возрастет незначительно (т. е. коэффициент множественной детерминации увеличится незначительно).

Значимость частных коэффициентов корреляции, так же как и парных коэффициентов корреляции, проверяется с помощью *t*-критерия Стьюдента. Наблюдаемое значение находится по формуле:

$$t = \frac{r}{\sqrt{1-r^2}} \cdot \sqrt{n-k-2}\ ,$$

где r – оценка частного коэффициента корреляции;

k – порядок частного коэффициента корреляции.

Квадрат частного коэффициента корреляции — частный коэффициент детерминации. Коэффициенты частной детерминации не могут быть сравнимы, т. к. представляют собой доли от разных величин.

### *2.1.5. Коэффициенты множественной детерминации и корреляции Скорректированный коэффициент множественной детерминации*

*Коэффициенты множественной детерминации и корреляции* характеризуют совместное влияние всех факторов на результат.

По аналогии с парной регрессией можно определить долю дисперсии результата, объясненной вариацией включенных в модель факторов ( $\delta^{*2}$ ), в его общей дисперсии ( $\sigma_y^2$ ). Ее количественная характеристика — теоретический множественный коэффициент детерминации ( $R^2_{y(x1\ldots xm)}$ ).

Для линейного уравнения регрессии данный показатель может быть рассчитан через ( $\beta$ -коэффициенты:

$$R^2_{y(x1\ldots xm)} = \sum_{j=1}^{m} \beta_j \cdot r_{xyj}\ .$$

$R_{y(x1\ldots xm)} = \sqrt{R^2_{y(x1\ldots xm)}}$ – коэффициент множественной корреляции. Он принимает значения от 0 до 1 (в отличие от парного коэффициента корреляции, который может принимать отрицательные значения). Поэтому R не может быть использован для интерпретации направления связи. Чем плотнее фактические значения $y_i$ располагаются относительно теоретической линии регрессии, тем меньше остаточная дисперсия и, следовательно, больше величина $R_{y(x1\ldots xm)}$. Таким образом, при значении *R,* близком к 1, уравнение регрессии лучше описывает фактические данные, и факторы сильнее влияют на результат. При значении *R,* близком к 0, уравнение регрессии плохо описывает фактические данные, и факторы оказывают слабое воздействие на результат.

! Важное свойство коэффициента детерминации состоит в том, что это неубывающая функция от числа факторов, т.е. включение в модель любого дополнительного фактора $x_{m+1}$ не приведет к снижению

коэффициента детерминации: $R^2_{y(x1\ldots xm)} \le R^2_{y(x1\ldots xm,xm+1)}$.

Рассмотрим подробнее формулу расчета коэффициента детерминации через необъясненную дисперсию:

$$R^2_{y(x1\ldots xm,xm+1)} = 1 - \frac{\sum\left(y - \hat{y}_{y(x1\ldots xm,xm+1)}\right)^2}{\sum\left(y - \bar{y}\right)^2}.$$

Знаменатель в данной формуле от числа факторов не зависит. А числитель снижается с введением в модель дополнительного фактора. Поэтому при сравнении двух моделей иногда не совсем ясно, за счет чего возрос показатель $R^2$: за счет реального влияния дополнительного фактора на результат, либо просто ввиду увеличения числа факторов.

Для того чтобы значения $R^2$ были сравнимы по разным моделям, необходимо учесть число независимых переменных в модели. Это можно сделать, если определить коэффициент детерминации не через суммы квадратов, а через дисперсии на 1 степень свободы. В результате получим скорректированный коэффициент детерминации — $R^2_{скор}$:

$$R^2_{скор} = 1 - \frac{\dfrac{\sum\left(y_i - \hat{y}_i\right)^2}{(n-h)}}{\dfrac{\sum\left(y_i - \bar{y}\right)^2}{(n-1)}} = 1 - \left(1 - R^2\right)\frac{(n-1)}{(n-h)},$$

где $h$ – общее число параметров в уравнении регрессии;

$n$ – число наблюдений.

Если $n$ велико, то $R^2$ и $R^2_{скор}$, будут незначительно отличаться.

Рассмотрим **пример.** Пусть по данным о 20 рабочих оценена регрессия заработной платы рабочего по возрасту ($x_1$) и выработке ($x_2$):

$\hat{y}_{x1,x2} = -16{,}04 + 5{,}1 \cdot x_1 + 8{,}08 \cdot x_2$.

Оценим качество данного уравнения регрессии, т.е. рассчитаем коэффициент множественной детерминации:

$R^2_{y(x1,x2)} = \beta_1 \cdot r_{yx1} + \beta_2 \cdot r_{yx2} =$

$=0{,}60166 \cdot 0{,}85305 + 0{,}408476 \cdot 0{,}778766 = 0{,}831356$

(расчет $\beta_1$ и $\beta_2$ см. в п. 2.1.2, расчет $r_{yx1}$ и $r_{yx2}$ – в п. 2.2.1).

Это значит, что 83,14 % вариации заработной платы рабочего определяется уравнением регрессии, а, следовательно, и факторами: «возраст» и «выработка».

Значение коэффициента множественной корреляции $R_{y(x1,x2)} = 0{,}911787$ близко к единице, что свидетельствует об очень тесной зависимости между факторами и результатом.

Сравним результаты оценки двухфакторной регрессии с однофакторной регрессией заработной платы рабочего по возрасту рабочего *x*. Оцененное уравнение однофакторной регрессии: $y_{x1} = 62{,}27 + 7{,}23 \cdot x_1$ (см. п. 1.2.1), коэффициент детерминации $R^2_{y(x1)} = 0{,}728$. Коэффициент детерминации в однофакторной регрессии меньше, чем в двухфакторной.

Чтобы определить, какое уравнение регрессии лучше, рассчитаем скорректированные коэффициенты детерминации:

- для однофакторной регрессии:
$$R^2_{скору(x1)} = 1 - (1 - 0{,}728)\frac{(20-1)}{(20-2)} = 0{,}7126;$$
- для двухфакторной регрессии:
$$R^2_{скору(x1,x2)} = 1 - (1 - 0{,}831)\frac{(20-1)}{(20-3)} = 0{,}8115$$

Так как скорректированный коэффициент детерминации для двухфакторной модели больше, чем для однофакторной, делаем вывод, что двухфакторная модель регрессии предпочтительнее. Улучшение качества уравнения регрессии при введении дополнительного фактора *($x_2$* — выработка за смену) существенно.

### *2.1.6. Оценка значимости уравнения множественной регрессии. Оценка значимости фактора, дополнительно включенного в модель регрессии. Общий и частный F- критерий*

Оценка значимости уравнения множественной регрессии осуществляется путем проверки основной гипотезы H0: $R^2_{y(x1\ldots xm)} = 0$ или $\tilde{b}_1 = \tilde{b}_2 = \ldots \tilde{b}_m = 0$ (гипотеза о статистической незначимости уравнения регрессии).

Альтернативной гипотезой – H1 (которая принимается, если основная окажется не верна) является гипотеза о статистической значимости уравнения регрессии: $R^2_{y(x1\ldots xm)} \neq 0$ или $\tilde{b}_1 \neq \tilde{b}_2 \neq \ldots \tilde{b}_m \neq 0$.

Для проверки основной гипотезы используют общий F-критерий Фишера.

При этом вычисляют фактическое (наблюдаемое) значение F-статистики критерия, например, через коэффициент детерминации ($R^2_{y(x1\ldots xm)}$) с учетом изменения числа степеней свободы:

$$F = \frac{R^2_{y(x1\ldots xm)}}{1 - R^2_{y(x1\ldots xm)}} \cdot \frac{n-h}{h-1},$$

где $n$ — число наблюдений;

$h$ — число оцениваемых параметров (в случае двухфакторной линейной регрессии $h = 3$).

По таблицам распределения Фишера-Снедоккора находят критическое значение F-статистики — $F_{кр}$. Для определения $F_{кр}$ задаются уровнем значимости $\alpha$ (обычно его берут равным 0,05) и двумя числами степеней свободы $k1=h-1$ и $k2=n-h$.

Сравнивают фактическое значение F-статистики критерия, вычисленное по данным наблюдений — ($F_{набл}$) с критическим – $F_{кр(\alpha;k1;k2)}$. Если $F_{набл} < F_{кр(\alpha;k1;k2)}$, то основную гипотезу о незначимости уравнения регрессии не отвергают. Если $F_{набл} > F_{кр(\alpha;k1;k2)}$, то основную гипотезу отвергают и принимают альтернативную гипотезу о статистической значимости уравнения регрессии.

*Оценка значимости дополнительного включения фактора (частный F-критерий).* Необходимость такой оценки связана с тем, что не каждый фактор, вошедший в модель, может существенно увеличить долю объясненной вариации результативного признака. Это может быть связано с последовательностью вводимых факторов (т. к. существует корреляция между самими факторами).

Мерой оценки значимости улучшения качества модели, после включения в нее фактора $x_j$, служит частный *F-критерий* – $F_{xj}$:

$$F_{xj} = \frac{R^2_{y(x1\ldots xj\ldots xm)} - R^2_{y(x1\ldots xj-1,xj+1\ldots xm)}}{1 - R^2_{y(x1\ldots xj\ldots xm)}} \cdot \frac{n-h}{1},$$

где $h$ – число оцениваемых параметров.

В числителе – прирост доли вариации $y$ за счет дополнительно включенного в модель фактора $x_j$.

Если наблюдаемое значение $F_{xj}$ больше $F_{кр(\alpha;k1=1;k2=m-p)}$, то дополнительное введение фактора $x_j$ в модель статистически оправдано.

Допустим, что оценивается значимость фактора $x_1$ как дополнительно включенного в модель $y = f(x2)$. Тогда частный F-критерий будет вычисляться по формуле:

$$F_{x1} = \frac{R^2_{y(x1x2)} - R^2_{y(x2)}}{1 - R^2_{y(x1x2)}} \cdot \frac{n-3}{1} = \frac{R^2_{y(x1x2)} - r^2_{yx2}}{1 - R^2_{y(x1x2)}} \cdot \frac{n-3}{1}.$$

Частный F-критерий оценивает значимость коэффициентов «чистой» регрессии ($b_j$). Существует взаимосвязь между частным F-критерием — $F_{xj}$ и

*t*-критерием, используемым для оценки значимости коэффициента регрессии при *j*-м факторе: $t_{(bj=0)} = \sqrt{F_{xj}}$ .

Рассмотрим методику расчета общего и частного F-критериев на примере. Пусть по данным о 20 рабочих исследуется влияние на заработную плату рабочего за месяц *y* факторов: возраста $x_1$, и выработки за смену $x_2$. Данные приведены в табл. 2.1.

Оценим с помощью общего F-критерия значимость уравнения регрессии, построенного по имеющимся данным (см. п.2.1.2):

$$\hat{y}_{x1,x2} = -16{,}04 + 5{,}1 \cdot x_1 + 8{,}08 \cdot x_2 . \qquad (2.2)$$

Теоретический коэффициент детерминации для данного уравнения равен $R^2_{y(x1,x2)} = 0{,}831$ (расчет смотри в п. 2.1.5).

Тогда: $F_{набл} = \dfrac{R^2_{y(x1x2)}}{1 - R^2_{y(x1x2)}} \cdot \dfrac{n-h}{n-1} = \dfrac{0{,}831}{1-0{,}831} \cdot \dfrac{20-3}{3-1} = 41{,}9$ .

$F_{кр(\alpha=0{,}05;k1=3-1=2;k2=20-3=17)} = 3{,}59$ .

$F_{набл}$=41,9>$F_{кр}$, следовательно, уравнение регрессии (2.2) статистически значимо и может быть использовано на практике.

Оценим с помощью частного F-критерия:

1) целесообразность включения в модель регрессии фактора $x_2$ после введения $x_1$ ($F_{x2}$);

2) целесообразность включения в модель регрессии фактора $x_1$, после введения $x_2$ ($F_{x1}$);

3) значимость коэффициентов регрессии.

Для проверки гипотезы о целесообразности включения фактора $x_2$ в модель регрессии после введения $x_1$, определим наблюдаемое значение частного F-критерия:

$$F_{x2} = \frac{R^2_{y(x1x2)} - R^2_{y(x1)}}{1 - R^2_{y(x1x2)}} \cdot \frac{n-3}{1} = \frac{0{,}831 - 0{,}728}{1 - 0{,}831} \cdot \frac{20-3}{1} = 10{,}45 \text{ где}$$

$R^2_{y(x1)} = r^2_{yx1} = 0{,}853^2 = 0{,}728$ (коэффициент парной корреляции $r_{yx1}$ рассчитан в п. 2.2.1).

$F_{кр(\alpha=0{,}05;k1=1;k2=20-3=17)} = 4{,}45$ .

Сравним наблюдаемое значение частного F-критерия с критическим: $F_{x2} > F_{кр}$, следовательно, фактор $x_2$ (выработка рабочего) целесообразно включать в модель после введения фактора $x_1$ (возраст рабочего).

Для проверки гипотезы о целесообразности включения фактора $x_1$, в модель регрессии после введения $x_2$ определим наблюдаемое значение частного F-критерия:

$$F_{x1} = \frac{R^2_{y(x1x2)} - R^2_{y(x2)}}{1 - R^2_{y(x1x2)}} \cdot \frac{n-3}{1} = \frac{0{,}831 - 0{,}6065}{1 - 0{,}831} \cdot \frac{20-3}{1} = 22{,}67$$ где

$R^2_{y(x2)} = r^2_{yx2} = 0{,}7788^2 = 0{,}6065$ (коэффициент парной корреляции $r_{yx2}$ рассчитан в п. 2.2.1).

Сравним наблюдаемое значение частного F-критерия с критическим: $F_{x1} > F_{кр(\alpha=0{,}05;k1=1;k2=20-3=17)} = 4{,}45$, следовательно, фактор $x_1$ – возраст рабочего целесообразно включать в модель после введения фактора $x_2$ – выработка за смену.

Для проверки гипотезы о значимости (значимом отличии от нуля) коэффициента $b_1$, при факторе $x_1$ (возраст) определим наблюдаемое значение *t*-статистики: $t_{(b1=0)} = \sqrt{F_{x1}} = 4{,}76$. Сравним его с критическим значением $t_{кр(0{,}1;k=20-3=17)} = 2{,}11$. Так как наблюдаемое значение больше критического, то гипотеза о незначимости коэффициента регрессии отвергается, следовательно, коэффициент регрессии $b_1$ значимо отличается от нуля.

Для проверки гипотезы о значимости (значимом отличии от нуля) коэффициента $b_2$ при факторе $x_2$ (выработка рабочего) определим наблюдаемое значение *t*-статистики: $t_{(b2=0)} = \sqrt{F_{x2}} = 3{,}23$. Сравним его с критическим значением $t_{кр(0{,}1;k=20-3=17)} = 2{,}11$. Так как наблюдаемое значение больше критического, то гипотеза о незначимости коэффициента регрессии отвергается, следовательно, коэффициент регрессии $b_2$ значимо отличается от нуля.

## 2.2. Различные аспекты множественной регрессии

### *2.2.1. Проблема мультиколлинеарности*

*Мультиколлинеарность* – это нестрогая линейная зависимость между факторными признаками (что противоречит 1-й предпосылке нормальной линейной множественной регрессионной модели о независимости факторных признаков $(x_1, x_2, ..., x_j, ..., x_m)$, которая может привести к следующим нежелательным последствиям.

1. Оценки параметров становятся ненадежными. Они обнаруживают большие стандартные ошибки, малую значимость. В то же время модель в целом является значимой, т. е. значение множественного коэффициента корреляции завышено.

2. Небольшое изменение исходных данных приводит к существенному изменению оценок параметров модели.

3. Оценки параметров модели имеют неправильные с точки зрения теории знаки или неоправданно большие значения, что делает модель непригодной для анализа и прогнозирования.

4. Становится невозможным определить изолированное влияние факторов на результативный показатель.

Нестрогая линейная зависимость между факторными признаками совсем необязательно дает неудовлетворительные оценки. Если все другие условия благоприятствуют, т. е. если число наблюдений значительно, выборочные дисперсии факторных признаков велики, а дисперсия случайной составляющей мала, то в итоге можно получить вполне хорошие оценки. Рассмотрение данной проблемы начинается только тогда, когда это серьезно влияет на результаты оценки регрессии.

Данная проблема является обычной для регрессий временных рядов. Если независимые переменные имеют ярко выраженный временной тренд, то они будут тесно коррелированны, и это может привести к мультиколлинеарности.

На практике о наличии мультиколлинеарности судят по матрице парных линейных коэффициентов корреляции (корреляционной матрице):

$$\begin{pmatrix} 1 r_{01} \dots r_{0j} \dots r_{0m} \\ r_{10} 1 \dots r_{1j} \dots r_{1m} \\ \dots\dots\dots\dots\dots \\ r_{j0} r_{j1} \dots 1 \dots r_{jm} \\ \dots\dots\dots\dots\dots \\ r_{m0} r_{m1} \dots r_{mj} \dots 1 \end{pmatrix}$$

где $r_{jk}$ – коэффициент парной линейной корреляции между $j$-м и $k$-м факторными признаками ($j, k=1; m$)

$r_{0j}$ – коэффициент парной линейной корреляции между результативным признаком и $j$-м фактором ($j=1; m$).

Коэффициент корреляции, измеряющий связь признака с самим собой, равен единице, т. к. в этом случае имеет место максимально тесная связь. Поэтому на главной диагонали в корреляционной матрице стоят единицы. Корреляционная матрица является симметричной относительно главной диагонали, т. к. $r_{jk} = r_{kj}$.

Если имеет место мультиколлинеарность, то в модель следует включать не все факторы, а только те, которые в меньшей степени ответственны за мультиколлинеарность (при условии, что качество модели снижается несущественно).

В наибольшей степени ответственным за мультиколлинеарность будет тот признак, который теснее связан с другими факторами модели (имеет более высокие по модулю значения коэффициентов парной линейной корреляции).

Еще один способ определения факторов, ответственных за мультиколлинеарность, основан на вычислении коэффициентов множественной детерминации $\left(R^2_{xj(x1,..xj-1,xj+1,...xm)}\right)$, показывающих зависимость фактора $x_j$ от других факторов модели $\left(x_1,x_2,...,x_{j-1},x_{j+1},...,x_m\right)$. Чем ближе значение коэффициента множественной детерминации к единице, тем больше ответственность за мультиколлинеарность фактора, выступающего в роли зависимой переменной. Сравнивая между собой коэффициенты множественной детерминации для различных факторов, можно проранжировать переменные по степени ответственности за мультиколлинеарность.

**Пример**. Допустим, имеются данные о заработной плате $y$ ($), возрасте $x_1$ (лет), стаже работы по специальности $x_2$ (лет), выработке — $x_3$ (шт./смену) по 10 рабочим (табл. 2.2). Требуется построить регрессионную модель заработной платы.

Таблица 2.2

| № наблюдения | y – заработная плата, $ | $x_1$ – возраст, лёт | $x_2$ – стаж работы по специальности, лет | $x_3$ – выработка, шт./смену |
|---|---|---|---|---|
| 1 | 300 | 29 | 6 | 17 |
| 2 | 400 | 40 | 19 | 25 |
| 3 | 300 | 36 | 10 | 15 |
| 4 | 320 | 32 | 10 | 17 |
|  | 200 | 23 | 3 | 15 |
| 6 | 350 | 45 | 20 | 18 |
| 7 | 350 | 38 | 17 | 17 |
| 8 | 400 | 40 | 23 | 25 |
| 9 | 380 | 50 | 31 | 19 |
| 10 | 400 | 47 | 25 | 23 |
| 11 | 250 | 28 | 7 | 15 |
| 12 | 350 | 30 | 7 | 18 |
| 13 | 200 | 25 | ,6 | 16 |
| 14 | 400 | 48 | 20 | 23 |
| 15 | 220 | 30 | 5 | 18 |
| 16 | 320 | 40 | 15 | 18 |
| 17 | 390 | 40 | 20 | 25 |
| 18 | 360 | 38 | 20 | 23 |
| 19 | 260 | 29 | 10 | 18 |
| 20 | 250 | 25 | 5 | 17 |

Проверим наличие мультиколлинеарности между факторами для данного примера. Для этого построим корреляционную матрицу (табл. 2.3).

Таблица 2.3

| | $y$ | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| $y$ | **1** | **0,853056** | **0,849877** | **0,778766** |
| $x_1$ | **0,853056** | 1 | 0,935263 | 0,615448 |
| $x_2$ | **0,849877** | 0,935263 | 1 | 0,69661 |
| $x_3$ | **0,778766** | 0,615448 | 0,69661 | 1 |

Из корреляционной матрицы видно, что между признаками $x_1$(возраст) и $x_2$(стаж работы по специальности) имеет место довольно сильная линейная зависимость, т. к. $r_{1,2}$= 0,935 > 0,8.

Продемонстрируем отрицательное влияние мультиколлинеарности. Для этого построим регрессионную модель заработной платы, включив в нее все исходные показатели:

$$\hat{y}_{x1,x2,x3} = -10{,}9 + 4{,}92 \cdot x_1 + 0{,}22 \cdot x_2 + 7{,}98 \cdot x_3$$

(1,98) (0,08) (2,79)

В скобках указаны расчетные значения, *t*-критерия для проверки гипотезы о значимости коэффициента регрессии. Критическое значение $t_{кр}$= 1,746 при уровне значимости $\alpha = 0{,}1$ и числе степеней свободы $(n - h)$ = 20 – 4 = 16. Из уравнения следует, что статистически значимыми являются коэффициенты регрессии только при $x_1$ и $x_3$, т. к. $|t_1|$ = 1,98 > $t_{кр}$= 1,746, $|t_3|$ = 2,79 > $t_{кр}$= 1,746. Таким образом, полученное уравнение регрессии неприемлемо.

Из модели следует исключить фактор $x_2$, т. к. он теснее связан с третьим фактором (выработкой), чем фактор $x_1$,: $r_{x2,x3г}$ = 0,697 >$r_{x1,x3}$= 0,615.

Построим теперь уравнение регрессии, исключив фактор $x_2$. Оно будет иметь вид:

$$\hat{y}_{x1,x3} = -16{,}04 + 5{,}1 \cdot x_1 + 8{,}08 \cdot x_3$$

(4,76) (3,23)

В этом уравнении все коэффициенты регрессии значимы. Причем значения *t*-статистик больше, чем в первом уравнении. Новое уравнение значимо в целом при $\alpha = 0{,}05$, т. к. $F_{набл} = 41{,}9 > F_{кр(\alpha=0{,}05;k1=3;k2=17)} = 3{,}2$

### *2.2.2. Фиктивные переменные множественной регрессии*

До сих пор мы рассматривали в качестве факторов количественные признаки (признаки, принимающие числовые значения).

Вместе с тем, может оказаться необходимым включить в модель качественный (атрибутивный) фактор (факторы). Примером качественных признаков может служить пол, образование, климатические условия.

Чтобы ввести такие признаки в модель, они должны быть преобразованы в количественные, т. е. им должны быть присвоены цифровые метки. Сконструированные на основе качественных факторов числовые переменные называют фиктивными переменными.

**Рассмотрим применение фиктивных переменных на примере**. Пусть по данным о 20 рабочих цеха оценивается регрессия заработной платы рабочего за месяц $y$ (\$) от количественного фактора $x_1$ – возраст рабочего (лет) и качественного фактора $x_2$ – пол. Мы предполагаем, что у мужчин зарплата выше, чем у женщин. Введем в модель: $y_i = b_0 + b_1 \cdot x_{1i} + u_i$ фиктивную переменную $z$, которая принимает 2 значения: 1 — если пол рабочего мужской; 0 — если пол женский.

Построим модель:

$$y_i = b_0 + b_1 \cdot x_{1i} + c \cdot z_i + u_i \qquad (2.3)$$

Исходные данные приведены в табл. 2.4.

Таблица 2.4

| № наблюдения | $y$ – заработная плата рабочего за месяц, \$ | $X$ – возраст рабочего, лет | Пол, м/ж |
|---|---|---|---|
| 1 | 300 | 29 | ж |
| 2 | 400 | 40 | м |
| 3 | 300 | 36 | ж |
| 4 | 320 | 32 | ж |
| 5 | 200 | 23 | м |
| 6 | 350 | 45 | м |
| 7 | 350 | 38 | ж |
| 8 | 400 | 40 | м |
| 9 | 380 | 50 | м |
| 10 | 400 | 47 | м |
| 11 | 250 | 28 | ж |
| 12 | 350 | 30 | м |
| 13 | 200 | 25 | м |
| 14 | 400 | 48 | м |
| 15 | 220 | 30 | ж |
| 16 | 320 | 40 | м |
| 17 | 390 | 40 | м |
| 18 | 360 | 38 | м |
| 19 | 260 | 29 | ж |
| 20 | 250 | 25 | м |

Для оценки параметров модели (2.3) используем обычный МНК. Построим систему нормальных линейных уравнений:

$$\begin{cases} \sum y_i = b_0 \cdot n + b_1 \cdot \sum x_{1i} + c \cdot \sum z_i; \\ \sum y_i \cdot x_{1i} = b_0 \cdot \sum x_{1i} + b_1 \cdot \sum x_{1i}^2 + c \cdot \sum z_i \cdot x_{1i}; \\ \sum y_i \cdot z_i = b_0 \cdot \sum z_i + b_1 \cdot \sum x_{1i} \cdot z_i + c \cdot \sum z_i^2. \end{cases}$$

В результате решения системы получим оценки: $\tilde{b}_0 = 63{,}52$; $\tilde{b}_1 = 7\,4$; $\tilde{c} = 10{,}32$.

$$\hat{y}_i = 63{,}52 + 7 \cdot x_{1i} + 10{,}32 \cdot z_i$$

(1,63) (6,14) (0,541);

$$R^2 = 0{,}732;\ R_{скор}{}^2 = 0{,}701;\ F = 23{,}25.$$

В скобках указаны значения *t*-критерия.

Результаты анализа регрессии свидетельствуют, что коэффициент при фиктивной переменной незначимо отличается от нуля, т. к.

$$t_{набл} = 0{,}541 < t_{кр(0б05жл=20-3=17)} = 2{,}1.$$

Это можно объяснить малым размером выборки (20 наблюдений). Возможно, если мы рассмотрим регрессию на реальных данных, то результаты будут иными.

Интерпретация параметра $c = 10{,}32$ при фиктивной переменной: у мужчин-рабочих зарплата в среднем выше, чему женщин-рабочих при одном и том же возрасте мужчины и женщины на \$10,32.

Сравним полученные результаты с результатами оценивания однофакторной модели:

$$y_{x1} = 62{,}27 + 7{,}23 \cdot x_1$$

(4,29) (4,104)

$$R^2 = 0{,}728;\ R_{скор}{}^2 = 0{,}713;\ F = 48{,}1.$$

Из модели, включающей фиктивную переменную, можно вывести *частные уравнения регрессии* для различных частей полной совокупности. Всю совокупность наблюдений можно разделить на 2 части: одна из них представляет те наблюдения, у которых $z = 1$; другая — те наблюдения, у которых $z = 0$.

В случае нашего примера совокупность рабочих можно разбить на 2 части (по полу) и построить для них частные уравнения регрессии:

а) $\hat{y}_i = 73{,}84 + 7 \cdot x_{1i}$ при $z$=1 (рабочий – мужчина);

б) $\hat{y}_i = 63{,}52 + 7 \cdot x_{1i}$ при $z$=0 (рабочий – женщина).

Сопоставляя эти частные уравнения регрессии, видим, что модели, описывающие заработную плату рабочего для мужчин и женщин, различаются значениями свободного члена. В случае а) (рабочий – мужчина) свободный член больше, чем в случае б) (рабочий – женщина). Если изобразить эти уравнения графически в системе координат *($x_1$; y),* то данные уравнения будут представлять собой параллельные линии, сдвинутые относительно друг друга по оси ординат. График частного уравнения регрессии для; мужчин будет располагаться выше, чем график частного уравнения регрессии для женщин.

В рассмотренном примере качественный признак принимает только 2 значения. Если же число градаций (значений) качественного фактора больше 2, в модель вводится несколько фиктивных переменных. Их число должно быть на 1 меньше числа градаций качественного фактора. Например, введем в

модель регрессии заработной платы рабочего *(y)* от возраста ($x_1$), качественный фактор – образование, принимающий 3 градации (значения): «до 8 классов»; «среднее»; «специальное». Для придания этому фактору численных значений введем 2 фиктивные переменные $z_1$ и $z_2$. Их возможные значения приведены в табл. 2.5.

Таблица 2.5

| Образование | $z_1$ | $z_2$ |
|---|---|---|
| до 8 классов | 0 | 0 |
| среднее | 1 | 0 |
| специальное | 0 | 1 |

Модель регрессии будет иметь вид:

$y_i = b_0 + b_1 \cdot x_{1i} + c_1 \cdot z_{1i} + c_2 \cdot z_{2i} + u_i$ '.

В результате оценивания с помощью МНК получим уравнение:

$\hat{y}_i = \tilde{b}_0 + \tilde{b}_1 \cdot x_{1i} + \tilde{c}_1 \cdot z_{1i} + \tilde{c}_2 \cdot z_{2i}$ .

Частные уравнения регрессии, соответствующие различным значениям качественного признака «образование»:

- «до 8 классов»: $\hat{y}_i = \tilde{b}_0 + \tilde{b}_1 \cdot x_{1i}$ ;
- «среднее»: $\hat{y}_i = \tilde{b}_0 + \tilde{c}_1 + \tilde{b}_1 \cdot x_{1i}$ ;
- «специальное»: $\hat{y}_i = \tilde{b}_0 + \tilde{c}_2 + \tilde{b}_1 \cdot x_{1i}$ .

Значение качественного фактора, для которого все фиктивные переменные равны нулю ($z_1 = z_2 = 0$), называют базовым значением.

В нашем примере базовым значением фактора «образование» является образование «до 8 классов».

Параметр при фиктивной переменной $z_1$ – $c_1$ означает, что при одном и том же возрасте рабочие со средним образованием получают заработную плату на $c_1$ долларов выше по сравнению с рабочими, имеющими образование «до 8 классов».

Параметр при фиктивной переменной $z_2$ – $c_2$ означает, что при одном и том же возрасте рабочие со специальным образованием получают заработную плату на $c_2$ долларов выше по сравнению с рабочими, имеющими образование «до 8 классов».

Графически частные уравнения регрессии представляют собой прямые линии, сдвинутые по оси ординат.

### *2.2.3. Тест Чоу*

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

Обозначим объединенную (общую) регрессию ***P***, а отдельные регрессии подвыборок как ***A*** и ***B***. Пусть суммы квадратов остатков для регрессий подвыборок равны соответственно: $E_A$ и $E_B$. Пусть $E_A^p$ и $E_B^p$ – суммы квадратов остатков в объединенной регрессии для наблюдений, относящихся к двум рассматриваемым подвыборкам.

Отдельные регрессии для подвыборок должны соответствовать наблюдениям, по меньшей мере так же хорошо, и даже лучше, чем объединенная регрессия. Поэтому должны выполняться следующие соотношения: $E_A < E_A^p$ и $E_B < E_B^p$ или $(E_A + E_B) \le E_p$ где $E_p = E_A^p + E_B^p$ – общая сумма остатков в объединенной регрессии.

Поясним графически суть данного подхода. Предположим, что имеются данные временного ряда по двум переменным, и что в период выборки произошло структурное изменение, разделяющее наблюдения на подвыборки ***A*** и ***B***. На рис. 2.1,*б* подвыборки обеспечивают вполне адекватное соответствие данным (им соответствуют низкие значения $E_A$ и $E_B$). Для случая объединенной регрессии (рис. 2.1,*а)* остатки в обеих подвыборках в целом оказываются значительно больше.

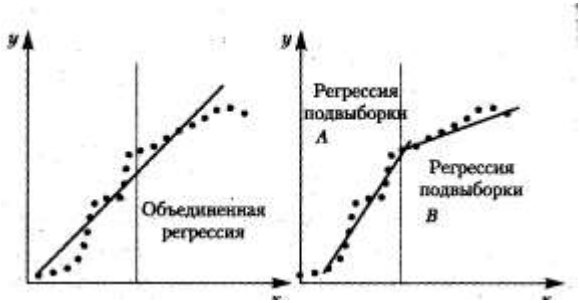


Рис. 2.1. Применение теста Чоу

*а* – объединенная регрессия; б – отдельные регрессии подвыборок

Равенство между $E_p$ и $(E_A + E_B)$ будет иметь место только при совпадении коэффициентов регрессии для объединенной регрессии и регрессий подвыборок. В общем случае при разделении выборки будет наблюдаться улучшение качества уравнения. Улучшение качества измеряется величиной: $E_p - E_A + E_B$. Однако при разделении выборки растет число степеней свободы: дополнительно приходится оценивать $(m + 1)$ параметров, где $m$ – число объясняющих переменных (факторов). Поэтому число степеней свободы возрастает на $(m + 1)$. После разделения выборки остается необъясненная сумма квадратов остатков $(E_A + E_B)$ с $(n - 2m - 2)$ степенями свободы.

Для того чтобы определить, является ли значимым улучшение качества уравнения после разделения выборки, используют *F*-статистику:

$$F = \frac{\dfrac{Улучшение\_качества\_уравнения}{Использованные\_степени\_свободы}}{\dfrac{Необъясненная\_дисперсия}{Число\_остающихся\_степеней\_свободы}} =$$

$$= \frac{\dfrac{\left(E_p - E_A - E_B\right)}{\left(m+1\right)}}{\dfrac{\left(E_A + E_B\right)}{\left(n-2m-2\right)}}.$$

Если вычисленное по данным выборки наблюдаемое значение *F*-статистики больше критического значения: $F_{кр(\alpha;m+1;n-2m-2)}$, то улучшение качества регрессии после разделения выборки существенно, т. е. не следует оценивать объединенную регрессию.

Рассмотрим применение теста Чоу **на примере**. Воспользуемся данными табл. 2.4. Пусть мы решили, что следует построить 2 отдельных уравнения регрессии для рабочих-мужчин и рабочих-женщин. Тогда оценивание объединенной регрессии и регрессий для подвыборок дает результаты, приведенные в табл. 2.6.

Таблица 2.6

| Выборка | Оцененное уравнение | $R^2$ | Сумма квадратов остатков |
|---|---|---|---|
| Объединенная выборка, | $\hat{y}_{x1} = 62{,}27 + 7{,}23 \cdot x_1$<br>$T_{набл}$ (4,29) (4,104) | 0,728 | 24888 |
| Мужчины | $y_{x1} = 55 + 7{,}39 \cdot x_1$<br>$T_{набл}$ (1,39) (6,88) | 0,735 | 24027 |
| Женщины | $y_{x1} = 59{,}43 + 7{,}3 \cdot x_1$<br>$T_{набл}$ (1,43) (6,48) | 0,712 | 24831 |

Соответствующая *F*-статистика будет равна:

$$F_{набл} = \frac{\dfrac{\left(24888 - 24027 - 24831\right)}{\left(1+1\right)}}{\dfrac{\left(24027 + 24831\right)}{\left(20-2-2\right)}} = -3{,}92.$$

Полученное значение меньше нуля. Так как $F_{кр}$ всегда положительно, то, следовательно, $F_{набл}$ будет меньше $F_{кр}$ и разбивать выборку на части не следует. Улучшение качества регрессии после разделения выборки на части не существенно.

Тест Чоу может применяться для выявления стабильности тенденции временного ряда. Допустим, что ряд динамики имеет нестабильную тенденцию. Это значит, что начиная с некоторого момента *t**, происходит изменение характера динамики изучаемого показателя, что приводит к изменению параметров тренда, описывающего эту динамику. Момент времени *t** сопровождается значительными изменениями ряда факторов, оказывающих сильное воздействие на изучаемый показатель (например, начало крупных экономических реформ, нефтяные кризисы, изменение экономического курса и прочее).

При этом весь ряд динамики представляет собой выборку, которую можно разделить на подвыборки:

1) до момента *t*;*
2) после момента *t*.*

Выдвигается основная гипотеза Н0: о структурной стабильности тенденции. В соответствии с тестом Чоу рассчитывается $F_{набл}$, определяется $F_{кр}$. Если $F_{набл} < F_{кр}$, то гипотезу Н0 не отвергаем, и наоборот.

### *2.2.4. Нелинейная множественная регрессия. Производственная функция*

Для линейного регрессионного анализа требуется линейность только по параметрам, поскольку нелинейность по объясняющим переменным может быть устранена с помощью метода замены переменных. Например, зависимость: $y = b_0 + b_1 \cdot x_1^{\,2} + b_2 \cdot x_2^{\,0,5} +$ может быть переписана в форме, которая будет линейной по объясняющим переменным $z_1 = x_1^2$, $z_2 = x_2^{0,5}$ следующим образом: $y = b_0 + b_1 \cdot z_1 + b_2 \cdot z_2 + u$. Если случайная составляющая *u* в начальном уравнении регрессии удовлетворяет предпосылкам нормальной модели, то она будет им удовлетворять и в преобразованном уравнении регрессии.

Нелинейность по параметрам является более серьезной проблемой. Однако если имеет место показательная, либо степенная зависимость и случайная составляющая входит в модель мультипликативно, то модель может быть линеаризована посредством логарифмирования обеих ее частей. Например, функция спроса: $y = a \cdot x^b \cdot p^c \cdot u$, где *y* – объем потребления товара; *x* – доход; *p* – цена; *u* – случайная .составляющая, может быть преобразована в форму, которая является линейной по параметрам:

$$Lny = Lna + b \cdot Lnx + c \cdot Lnp + Lnu$$

При этом коэффициент при Ln*x* будет непосредственной оценкой *b* – эластичности спроса по доходу, коэффициент при Ln*p* будет оценкой *c* – эластичности спроса по цене.

*Производственная функция.* Производственная функция! представляет собой математическую модель, характеризующую зависимость объема выпускаемой продукции от факторов производства. При этом модель может быть построена как для отдельной фирмы и отрасли, так и для всей национальной экономики. Рассмотрим производственную функцию, включающую два фактора производства: затраты капитала *K* и трудовые затраты *L*, определяющие объем выпуска *Q*, Тогда можно записать: *Q=f(K,L).*

Определенного уровня выпуска можно достичь с помощью различного сочетания капитальных и трудовых затрат. Кривые, описываемые условиями *f(K,L)=const*, называют изоквантами. Предполагается, что по мере роста значений одного из факторов предельная норма замещения данного фактора производства уменьшается. Поэтому при сохранении постоянного объема производства экономия (сокращение) одного вида затрат, связанная с увеличением затрат другого фактора, постепенно уменьшается.

Рассмотрим подробно производственную функцию Кобба-Дугласа, которая для двух факторов производства имеет вид: $Q = A \cdot K^{\alpha} \cdot L^{\beta}$ , где *A,* $\alpha$, $\beta$ – параметры модели.

Величина *A* зависит от единиц измерения *Q, K, L*, а также от эффективности производственного процесса. При фиксированных значениях *K* и *L* функции, характеризующейся большей величиной параметра *A,* соответствует большее значение *Q*, следовательно, и производственный процесс, описываемый такой функцией, более эффективен.

Параметры $\alpha$ и $\beta$ называют коэффициентами эластичности. Они показывают, на сколько процентов в среднем изменится *Q*, если $\alpha$ и $\beta$ увеличить соответственно на 1 %. Можно предположить, что обе величины $\alpha$ и $\beta$ находятся между нулем и единицей. Они должны быть положительными, так как увеличение затрат производственных факторов должно вызвать рост объема выпуска. Скорее, всего, они будут меньше единицы, так как разумно предположить, что рост объема выпуска происходит медленнее, чем рост производственных затрат, если другие факторы остаются постоянными.

*Эффект от масштаба производства.* Рассмотрим поведение производственной функции при изменении масштабов производства. Предположим для этого, что затраты каждого фактора производства увеличились в *C* раз. Тогда новое значение будет определяться следующим образом:

$$\hat{Q} = A \cdot (C \cdot K)^{\alpha} \cdot (C \cdot L)^{\beta} = C^{\alpha+\beta} \cdot Q .$$

Если ($\alpha$ + $\beta$) > 1, то говорят, что функция имеет возрастающий эффект от масштабов производства. Это значит, что если *K* и *L* увеличиваются в некоторой пропорции (С раз), то *Q* растет в большей пропорции ($C^{\alpha+\beta}$ раз, что больше *C*).

Если $(\alpha + \beta) = 1$, то говорят, что функция имеет постоянный эффект от масштабов производства. Это значит, что $Q$ увеличивается в той же пропорции, что и $K$ и $L$.

Если $(\alpha + \beta) < 1$, то говорят, что функция имеет убывающий эффект от масштабов производства. Это значит, что $Q$ увеличивается в меньшей пропорции, чем $K$ и (т. к. $C^{\alpha+\beta} < C$).

### *2.2.5. Гетероскедастичность случайной составляющей*

При оценке параметров уравнения регрессии чаще всего применяется традиционный метод наименьших квадратов. При этом должны выполняться определенные предпосылки относительно случайной составляющей $u_i$ и объясняющих переменных $x_i$ (предпосылки нормальной линейной модели). Напомним, что $u_i$, имеет смысл отклонения в линейной модели регрессии: $u_i = y_i - (b_0 + b_1 \cdot x_i)$.

Третья предпосылка гласит: $\sigma_{ui}^2 = \sigma_u^2 = const$, $i=1;n$, что означает постоянство дисперсий случайных составляющих для каждого наблюдения $i$.

Поясним данную предпосылку на **примере.** Случайная составляющая $u_i$ в каждом наблюдении может иметь только одно значение. Что же означает дисперсия $u_i$? Имеется в виду *возможное* поведение $u_i$ до того, как проведено наблюдение. То есть нет основания *apriori* ожидать появления особенно больших отклонений в любом наблюдении $i=1;n$. Иными словами вероятность того, что величина $u_i$ примет какое-то данное значение, будет одинакова для всех $i$. Это условие известно как условие гомоскедастичности, что означает одинаковый разброс.

Вместе с тем, для некоторых выборок можно предположить, что теоретическое распределение случайной составляющей $u_i$ является различным для разных наблюдений в выборке, а следовательно, различными будут и дисперсии случайных составляющих. Если дисперсии случайных составляющих неодинаковы в разных наблюдениях: $\sigma_{ui}^2 \neq \sigma_u^2 \neq const$, $i, j = 1;n$ $(i \neq j)$, говорят, что имеет место гетероскедастичность (т. е. неодинаковый разброс случайных составляющих). Например, если исследуется зависимость расходов на питание в семье от ее общего дохода, то можно ожидать, что разброс данных будет выше для семей с более высоким доходом. Это означает, что дисперсии зависимых величин – расходов на питание, (а следовательно, и случайных ошибок) не постоянны для отдельных значений объясняющей переменной – дохода.

Гетероскедастичность может иметь место и при использовании в качестве данных наблюдений временных рядов ($x_t$, $y_t$). Если значения $x_t$ и $y_t$ увеличиваются со временем, то, возможно, и дисперсия случайной составляющей также будет расти со временем.

*Наличие гетероскедастичности можно наглядно видеть из поля корреляции* (рис. 2.2).

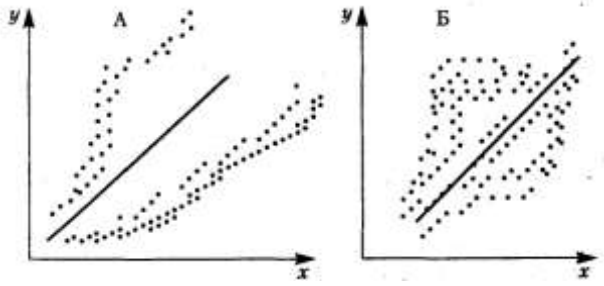


Рис. 2.2. Корреляционное поле. Случаи гетероскедастичности

На рис. 2.2,*а* дисперсия случайных составляющих растет по мере увеличения *x*. На рис. 2.2,*б* дисперсия случайных составляющих достигает максимальной величины при средних значениях *x* и уменьшается при минимальных и максимальных значениях *x*.

Кроме того, наличие гетероскедастичности можно проследить из графика зависимости остатков $e_i$ от расчетного значения признака-результата $\hat{y}_i$. Гетероскедастичность, соответствующая полю корреляции *а* на рис. 2.2, приведена на рис. 2.3,*а*, гетероскедастичность, соответствующая полю корреляции *б* на рис. 2.2, приведена на рис. 2.3,*б*.

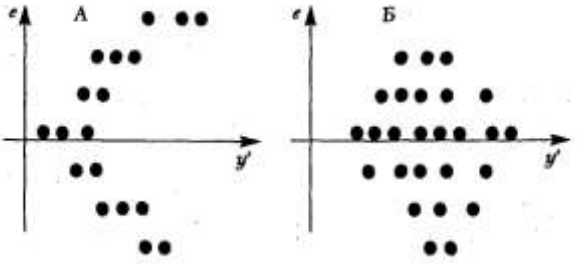


Рис. 2.3. Графики зависимости остатков от теоретических значений результата. Случаи гетероскедастичности

Последствия гетероскедастичности:

- оценки параметров уравнения регрессии становятся неэффективными;
- оценки стандартных ошибок параметров регрессии будут неверными. (Например, оценки стандартных ошибок могут оказаться заниженными. Тогда значения *t*-критерия окажутся завышенными. Мы решим, что параметр регрессии значим, а на самом деле это будет не так. То есть могут быть получены неверные выводы о надежности уравнения регрессии.)

*Обнаружение гетероскедастичности.* Наиболее популярным является тест Голдфелда-Квандта.

Данный тест используется для проверки следующего типа гетероскедастичности: когда среднее квадратическое отклонение случайной составляющей $\sigma_{ui}$ *пропорционально значению признака-фактора* $x_i$ в $i$-м наблюдении. При этом делается предположение, что случайная составляющая $u_i$ распределена нормально.

Алгоритм-тест Голдфелда-Квандта приведен ниже.

Все наблюдения $i = 1; n$ упорядочиваются по значению $x_i$.

Оценивается регрессия: $\hat{y}_{1i} = b_{01} + b_{11} \cdot x_i$ ( $i = 1; n'$) для первых $n'$ наблюдений.

Оценивается регрессия: $\hat{y}_{2i} = b_{02} + b_{12} \cdot x_i$ ( $i = n - (n' + 1); n$) для последних $n'$ наблюдений. ($n' < n/2$).

Рассчитывают суммы квадратов отклонений фактических значений признака-результата от его расчетных значений для обеих регрессий:

$$S1 = \sum_{i=1}^{n'} (y_i - \hat{y}_{1i})^2 \quad и \quad S2 = \sum_{i=n-(n'+1)}^{n} (y_i - \hat{y}_{2i})^2 .$$

Находят отношение сумм квадратов отклонений: *S1/S2* (или *S2/S1*). В числителе должна быть наибольшая из сумм квадратов отклонений. Данное отношение имеет *F*-распределение со степенями свободы: $k_1 = n' - h$ и $k_2 = n' - h$, где $h$ – число оцениваемых параметров в уравнении регрессии.

Если $F_{набл} = \frac{S1}{S2} > F_{кр(\alpha;k1;k2)}$, то гетероскедастичность имеет место.

Если в модели более одного фактора, то наблюдения должны упорядочиваться по тому фактору, который, как предполагается, теснее связан с $\sigma_{ui}^2$, и $n'$ должно быть больше, чем $h$.

*Устранение гетероскедастичности.* Для этого нужно найти способ придать наибольший вес наблюдению $i$, у которого среднее квадратическое отклонение случайной составляющей $\sigma_{ui}$ максимально (такие наблюдения обладают самым низким качеством); и малый вес наблюдению, у которого среднее квадратическое отклонение случайной составляющей $\sigma_{ui}$ минимально (такие наблюдения обладают самым высоким качеством). Тогда мы получим более точные (эффективные) оценки параметров уравнения регрессии: $y_i = b_0 + b_1 \cdot x_i + u_i$.

Разделим правую и левую части уравнения на $\sigma_{ui}$. Получим:

$$\frac{y_i}{\sigma_{ui}} = \frac{b_0}{\sigma_{ui}} + b_1 \frac{x_1}{\sigma_{ui}} + \frac{u_i}{\sigma_{ui}}.$$

Введем новые переменные:

$$Y_i = \frac{y_i}{\sigma_{ui}}; X_i = \frac{x_1}{\sigma_{ui}}; v_i = \frac{1}{\sigma_{ui}}; U_i = \frac{u_i}{\sigma_{ui}}.$$

Тогда уравнение регрессии примет вид:

$$Y_i = b_0 \cdot v_i + b_1 \cdot X_i + U_i.$$

Преобразованное уравнение относится к двухфакторному уравнению регрессии (1-й фактор – *X*, 2-й фактор — *v*). Данное уравнение представляет собой так называемую взвешенную регрессию (с весами $1/\sigma_{ui}$). При этом наблюдениям высокого качества с меньшими $\sigma_{ui}$ придаются большие веса $1/\sigma_{ui}$ и наоборот. Случайная составляющая в *i*-м наблюдении $-\frac{u_i}{\sigma_{ui}}$ имеет постоянную дисперсию:

$$M\left(\frac{u_i}{\sigma_{ui}}\right)^2 = \frac{1}{\sigma_{ui}} M\left(u_i^2\right) = \frac{1}{\sigma_{ui}} \sigma_{ui} = 1,$$

т. е. модель будет гомоскедастичной.

Данный способ устранения гетероскедастичности применим, если известны фактические значения $\sigma_{ui}$, что не встречается на практике.

Однако, если мы сможем подобрать некоторую величину, пропорциональную $\sigma_{ui}$ в каждом наблюдении *i = 1; n*, и разделим на нее обе части уравнения, то гетероскедастичность будет устранена. Например, может оказаться целесообразным предположить, что $\sigma_{ui}$ приблизительно пропорциональна $x_i$, как в критерии Голдфелда-Квандта $\left(x_i = \lambda \cdot \sigma_{ui}\right)$.

Тогда: $\frac{y_i}{x_i} = \frac{b_0}{x_i} + b_1 \frac{x_1}{x_i} + \frac{u_i}{x_i} = b_0 \cdot \frac{1}{x_i} + b_1 + \frac{u_i}{x_i}$.

Если «повезет», новая случайная составляющая $\frac{u_i}{x_i}$ будет иметь постоянную дисперсию. Оценим регрессию новой зависимой переменной $Y_i = \frac{y_i}{x_i}$ на новую независимую переменную $X_i = \frac{1}{x_i}$. Тогда коэффициент при этой переменной – эффективная оценка параметра $b_0$, а постоянный член – эффективная оценка параметра $b_1$ исходного уравнения регрессии:

$y_i = b_0 + b_1 \cdot x_i + u_i$. Дисперсия случайной составляющей в этом уравнении может быть записана как

$$M\left[\frac{u_i}{x_i}\right]^2 = M\left[\frac{u_i}{\lambda \cdot \sigma_{ui}}\right]^2 = \frac{1}{\lambda^2} \cdot \frac{\sigma_{ui}^2}{\sigma_{ui}^2} = \frac{1}{\lambda^2}.$$

То есть она будет постоянна для всех наблюдений. Следовательно, гетероскедастичность в преобразованном уравнении регрессии отсутствует.

### *2.2.6. Автокорреляция случайных составляющих. Обнаружение автокорреляции случайных составляющих. Критерии Дарбина-Уотсона*

Автокорреляция – корреляционная зависимость между текущими уровнями некоторой переменной и уровнями этой же переменной, сдвинутыми на несколько периодов времени назад.

Автокорреляция случайной составляющей $u$ – корреляционная зависимость текущих $u_i$ и предыдущих $u_{i-L}$ значений случайной составляющей. Величина $L$ называется запаздыванием, сдвигом во времени или лагом. Лаг определяет порядок автокорреляции.

Автокорреляция случайной составляющей нарушает 4-ю предпосылку нормальной линейной модели регрессии: $Cov(u_i, u_j) = 0, \forall i, j = 1; n, i \neq j$ (условие независимости случайных составляющих в различных наблюдениях, $i$,$j$ – номера наблюдений).

Наличие случайной составляющей $u_i$ в уравнении регрессии может быть обусловлено: невключением в уравнение регрессии объясняющих переменных (*неучтенными факторами*); агрегированием переменных; неправильной функциональной спецификацией модели; ошибками измерения.

Обычно автокорреляция встречается при использовании данных *временных рядов*. Допустим, что случайная составляющая обусловлена только невключением в модель объясняющих переменных. Тогда, если значение $u_i$ в $i$-м наблюдении должно быть независимым от его значения в предыдущем ($i$ – $L$)-м наблюдении $u_{i-L}$, то и значение любой факторной переменной, «скрытой» в $u$, должно быть некоррелированным с ее значением в предыдущем наблюдении.

Рассмотрим **пример** автокорреляции случайных составляющих (взятый из учебника Доугерти [6]). Проанализируем модель зависимости спроса на мороженое $y$ от дохода $x$ (учтенный фактор). На $y$ оказывают влияние не только доход $x$, но и другие факторы, которые не учтены в модели. Допустим, что один из таких факторов – время года. Летом спрос на мороженое выше, чем зимой. Данный фактор находит свое отражение в случайной составляющей.

Автокорреляция может быть как положительной, так и отрицательной.

Положительная автокорреляция означает постоянное в одном направлении действие неучтенных факторов на результат. Например, спрос на мороженое

всегда выше линии тренда летом (т. е. для летних наблюдений $u > 0$) и ниже зимой (т. е. для зимы $u < 0$) (рис. 2.4).

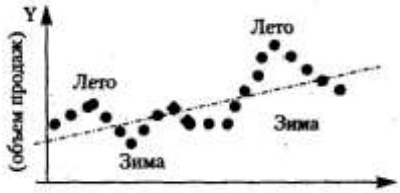


Рис. 2.4. Пример положительной автокорреляции

Отрицательная автокорреляция означает разнонаправленное действие неучтенных в модели факторов на результат, что приводит к отрицательной корреляции между последовательными значениями случайной составляющей. То есть за положительными значениями случайной составляющей и в одном наблюдении следуют отрицательные значения и в следующем, и наоборот. Заметим, что отрицательная автокорреляция в экономике встречается относительно редко.

Последствия автокорреляции случайной составляющей:

• коэффициенты регрессии становятся неэффективными, хоть и несмещенными и состоятельными;

• стандартные ошибки коэффициентов регрессии становятся заниженными, а значения *t*-критерия завышенными.

*Обнаружение автокорреляции случайной составляющей.* Оценкой случайной составляющей является остаток – разность между фактическим и рассчитанным по оцененному уравнению регрессии значениями признака-результата. Так как автокорреляция случайных составляющих имеет место, в основном, когда исходные данные являются временными рядами, обозначим номер наблюдения как *t (t = 1; n).* Тогда для *t*-го наблюдения остаток будет равен: $e_t = y_t - \hat{y}_t = y_t - \left(\tilde{b}_0 + \tilde{b}_1 x_{1t} + \tilde{b}_2 x_{2t} + \ldots + \tilde{b}_m x_{mt}\right)$, где $\tilde{b}_0, \tilde{b}_1, \tilde{b}_2, \ldots \tilde{b}_m$ – МНК- оценки коэффициентов истинного уравнения регрессии – $a, b_1, b_2, \ldots b_m$.

Рассмотрим способы обнаружения автокорреляции остатков (а следовательно, и случайных составляющих).

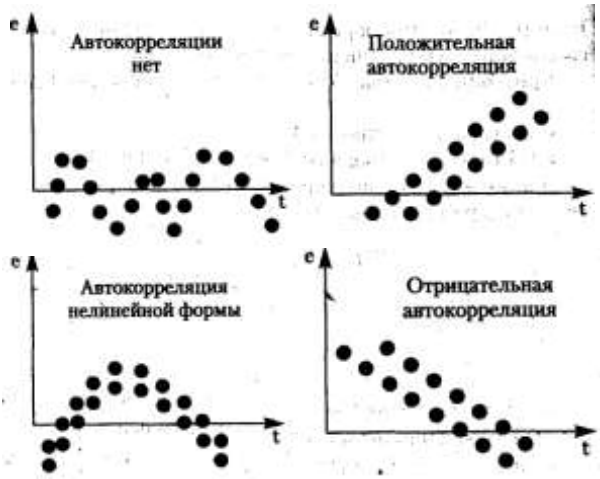


Рис. 2.5. Обнаружение автокорреляции остатков

**1-й способ** – визуальный (графический). С помощью МНК оценивается регрессия $y = f(x_1, x_2, \ldots, x_m)$. Рассчитываются остатки $e_i$. Строится график зависимости остатков $e_t$ от номера наблюдения – *t (t = 1; n)*. (рис. 2.5).

**2-й способ** – основан на применении критерия Дарбина-Уотсона.

Данный метод применяют для обнаружения автокорреляции, подчиняющейся авторегрессионному процессу 1-го порядка: $u_t = \rho \cdot u_{t-1} + e_t$ *(t = 1; n)*.

Предполагается, что величина $e_t$ в каждом *t*-м наблюдении не зависит от его значений во всех других наблюдениях. Если $\rho$ положительна, то автокорреляция положительна, если $\rho$ отрицательна, то автокорреляция отрицательна. Если $\rho = 0$, то автокорреляции нет (т. е. четвертая предпосылка нормальной линейной модели выполняется).

*Критерий Дарбина-Уотсона сводится к проверке гипотезы:*

- H0 (основная гипотеза): $\rho = 0$
- H1 (альтернативная гипотеза): $\rho > 0$ или $\rho < 0$.

Для проверки основной гипотезы используется статистика критерия Дарбина-Уотсона – DW:

$$DW = \frac{\sum_{t=2}^{N}(e_t - e_{t-1})^2}{\sum_{t=1}^{N} e_t^2}, \text{ где } e_t = y_t - \hat{y}_t .$$

На больших выборках $DW \approx 2(1 - r_{et,et-1})$, где $r_{et,et-1}$ выборочный коэффициент автокорреляции 1-го порядка.

Если $r_{et,et-1} = +1$, то DW = 0. Если $r_{et,et-1} = -1$, то DW = 4. Если $r_{et,et-1} = 0$, то DW = 2.

Данная статистика имеет распределение Дарбина-Уотсона. Существуют специальные статистические таблицы для определения нижней и верхней критических границ DW-статистики – $d_L$ и $d_u$. Они определяются в зависимости от *n* и числа степеней свободы $(h - 1)$, где *h* – число оцениваемых параметров.

Если $DW_{набл} < d_L$, то принимается гипотеза H1: $\rho > 0$ (положительная автокорреляция).

Если $d_u < DW_{набл} < 2$, то принимается гипотеза H0: $\rho = 0$ (автокорреляции нет).

Если $2 < DW_{набл} < 4 - d_u$, то принимается гипотеза H0: $\rho = 0$ (автокорреляции нет).

Если $DW_{набл} > 4 - d_u$ ,то принимается гипотеза H1: $\rho < 0$ (отрицательная автокорреляция).

Если $4 - d_u < DW_{набл} < 4 - d_L$, $d_L < DW_{набл} < d_u$, то имеет место случай неопределенности.

Графически тест Дарбина-Уотсона представлен на рис. 2.6 (штриховкой отмечена область неопределенности):

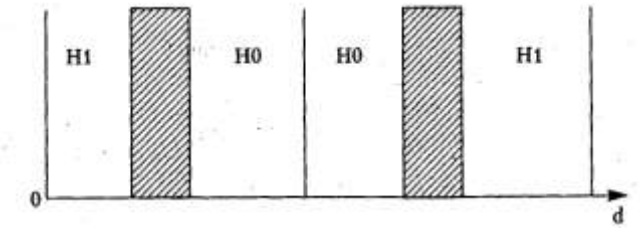


Рис. 2.6. Тест Дарбина-Уотсона на автокорреляцию

Ограничения применения DW-статистики:

• данный критерий неприменим к моделям авторегрессии: $y_t = f(y_{t-1})$;

• данный критерий проверяет только гипотезу о наличии автокорреляции *первого* порядка;

• достоверные результаты получаются только на больших выборках.

Вместе с тем встречаются случаи, когда целесообразно предположить, что случайные составляющие связаны авторегрессионным процессом более высокого порядка L:

$$u_t = \rho_1 \cdot u_{t-1} + \rho_2 \cdot u_{t-2} + \ldots + \rho_L \cdot u_{t-L} + e_t (t = 1; n).$$

Например, автокорреляция 4-го порядка может иметь место при использовании в качестве данных ежеквартальных наблюдений (когда сезонные колебания переходят из года в год).

Возможны случаи, когда случайные составляющие связываются не авторегрессионным процессом, а процессом скользящих средних:

$$u_t = \gamma_1 \cdot e_t + \gamma_2 \cdot e_{t-1} + \ldots + \gamma_L \cdot e_{t-L} (t = 1; n).$$

Для обнаружения автокорреляции высокого порядка используют различные методы. Рассмотрим один из них – метод Лагранжа.

Суть данного подхода заключается в следующем.

1. Решают вопрос о величине L – максимальной величине запаздывания во времени.

2. С помощью МНК оценивается исходная регрессия: $y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \ldots + b_m x_{mt} + u_t$ и рассчитываются остатки $e_t$.

3. Оценивается регрессия:

$$e_t = C_0 + C_1 \cdot e_{t-1} + C_2 \cdot e_{t-2} + \ldots + C_s \cdot e_{t-L} + d_1 \cdot x_{1t} + \; + d_2 \cdot x_{2t} + \ldots + d_m \cdot x_{mt} + \varepsilon_t$$

где $e_{t-p}$ – остаток в наблюдении *(t-p) (p = 1; L) (t = 1; n)*

$x_j$ *(j = 1; m)* – объясняющие переменные первоначальной регрессии;

$\varepsilon_t$ – случайная составляющая для данной зависимости.

Регрессия оценивается по данным для периодов от ($L$ + 1) до $n$, так как величины $e_{t-L}$ не определены для первых L периодов.

4. Выдвигается основная гипотеза H0: автокорреляция отсутствует. Для ее проверки вычисляется статистика: $\chi^2 = (n \cdot R^2)$, которая имеет распределение $\chi^2$ («хи-квадрат») с $L$ степенями свободы. Определяется критическое значение $\chi^2_{кр}$, по соответствующим статистическим таблицам. Фактическое значение $\chi^2$ сравнивается с критическим $\chi^2_{кр}$. Если $\chi^2 > \chi^2_{кр}$, то гипотезу об отсутствии автокорреляции отвергают. Если $\chi^2 < \chi^2_{кр}$, то гипотезу об отсутствии автокорреляции не отвергают.

Данный тест рассчитан на работу с большими выборками, поэтому нужно проявлять осторожность при толковании результатов, полученных на малых выборках.

### *2.2.7. Устранение автокорреляции случайных составляющих*

Что можно сделать в отношении автокорреляции?

Во-первых, определить фактор, ответственный за автокорреляцию, и включить его в уравнение регрессии. Но это практически невозможно.

Во-вторых, в случае автокорреляции, подчиненной авторегрессионному процессу 1-го порядка (т. е. $u_t = \rho \cdot u_{t-1} + e_t, \forall t = 1; n$), мы могли бы полностью устранить автокорреляцию, если бы знали величину $\rho$.

Предположим, что истинная модель задается выражением:

$$y_t = b_0 + b_1 x_t + u_t \qquad (2.4)$$

*(t = 1; n).*

Тогда наблюдение $(t - 1)$ формируется как:

$$y_{t-1} = b_0 + b_1 x_{t-1} + u_{t-1}.$$

Умножим полученное выражение на $\rho$ и вычтем его из обеих частей уравнения (2.4). В результате получим:

$$y_t - \rho \cdot y_{t-1} = (1-\rho)b_0 + b_1(x_t - \rho x_{t-1}) + u_t - \rho \cdot u_{t-1}. \qquad (2.5)$$

Обозначим $y_t^* = y_t - \rho \cdot y_{t-1}; x_t^* = x_t - \rho x_{t-1}, q_t = 1 - \rho, t = 2; n.$

Тогда формулу (2.5) можно переписать как:

$y_t^* = q_t \cdot b_0 + b_1 x_t^* + u_t - \rho u_{t-1} = q_t \cdot b_0 + b_1 x_t^* + \varepsilon_t$

$\left(u_t = \rho \cdot u_{t-1} + \varepsilon_t \Rightarrow \varepsilon_t = u_t - \rho u_{t-1}\right).$

Значения $\varepsilon_t$ (случайной составляющей в модели зависимости $u_t$ от $u_{t-1}$) для различных $t$ не зависят друг от друга, поэтому проблемы автокорреляции остатков в полученном уравнении для преобразованных переменных нет. Параметры данного уравнения оцениваются обычным МНК. Получаем оцененную регрессию: $\hat{y}_t^* = \tilde{b}_0^* + \tilde{b}_1^* x_t^*$.

Затем определяются оценки параметров исходного уравнения регрессии (2.4): $\tilde{b}_0 = \dfrac{\tilde{b}_0^*}{(1-\rho)}, \tilde{b}_1 = \tilde{b}_1^*$. Окончательно оцененное уравнение регрессии будет иметь вид: $\hat{y}_t = \tilde{b}_0 + \tilde{b}_1 x_t$.

Однако, если в выборке нет данных, предшествующих 1-му наблюдению, мы не можем вычислить значения $y_1^*$, $x_1^*$. Уменьшение числа степеней свободы на единицу приведет к потере эффективности, которая может в небольших выборках перевесить повышение эффективности от устранения автокорреляции. Эту проблему решают с помощью поправки Прайса-Уинстена: $y_1^* = \sqrt{1-\rho^2} \cdot y_1,\ x_1^* = \sqrt{1-\rho^2} \cdot x_1,\ q_1 = \sqrt{1-\rho^2}$.

На практике величина $\rho$, конечно, неизвестна, его оценка получается одновременно с оценками параметров $b_0$ и $b_1$. В качестве оценки $\rho$ обычно используют выборочный коэффициент автокорреляции остатков 1-го порядка $r_{et,et-1}$: $\rho = r_{et,et-1}$.

$$r_{et,et-1} = \frac{\sum_{t=2}^{N} e_t e_{t-1}}{\sum_{t=2}^{N} e_t^{\,2}},\ e_t = y_t - \left(\tilde{b}_0 + \tilde{b}_1 x_t\right),$$

где $\tilde{b}_0, \tilde{b}_1$ – МНК-оценки параметров $b_0$, $b_1$.

Другой способ оценки $\rho$ используется в методе Кохрейна-Оркатта. Алгоритм метода Кохрейна-Оркатта носит итеративный характер:

1) обычным МНК оценивается регрессия (2.4);
2) вычисляются остатки $e_t$ ($t = 1; n$);
3) оценивается регрессионная зависимость $e_t$ от $e_{t-1}$: $\hat{e}_t = \rho^* \cdot e_{t-1}$, коэффициент при $e_{t-1}$, равный $\rho^*$, представляет собой оценку $r_{et,et-1}$;
4) с этой оценкой ($\rho^*$) уравнение (2.4) преобразуется к виду:

$$y_t^* = q_t^* \cdot b_0 + b_1 \cdot x_t^* + u_t^* \qquad (2.6)$$

где $y_t^* = y_t - \rho \cdot y_{t-1}$; $x_t^* = x_t - r^* \cdot x_{t-1}$; $q_t^* = 1 - \rho^*$; $u_t^* = u_t - r^* \cdot u_{t-1}$;

5)обычным МНК оценивается уравнение (2.6): $\hat{y}_t = \tilde{b}_0 + \tilde{b}_1 x_t^*$. Затем вычисляют оценки параметров исходного уравнения (2.4), как $\tilde{b}_0 = \frac{\tilde{b}_0^*}{\left(1-\rho^*\right)}, b_1 = \tilde{b}_1$ ;

6)повторно вычисляются остатки $e_t$ и процесс возвращается к этапу 3.

Процесс обычно заканчивается, когда очередное приближение $\rho^*$ мало отличается от предыдущего. Иногда просто фиксируется количество итераций.

Другой способ оценки $\rho$ – метод Хилдрета-Лу:

Для каждого значения $\rho$ из определенного диапазона с заданным шагом внутри его оценивается преобразованное уравнение регрессии:

$$y_t^* = q_t^* \cdot b_0 + b_1 \cdot x_t^* + u_t^* \qquad (2.7)$$

где $y_t^* = y_t - r^* \cdot y_{t-1}$; $x_t^* = x_t - r^* \cdot x_{t-1}$; $q_t^* = 1 - r^*$; $u_t^* = u_t - r^* \cdot u_{t-1}$;

Например, из диапазона от $\rho \in [-1;+1]$ с шагом 0,01.

Значение $\rho^*$, которое дает минимальную сумму квадратов отклонений для преобразованного уравнения (2.7), принимается в качестве оценки $\rho$. Оценки

коэффициентов исходного уравнения регрессии (2.4) определяются как $b_0 = \frac{\tilde{b}_0}{\left(1-\rho^*\right)}, b_1 = \tilde{b}_1$ .

## 2.3. Некоторые обобщения множественной регрессии

### *2.3.1. Обобщенный метод наименьших квадратов –ОМНК (GLS)*

Обобщенный метод наименьших квадратов (ОМНК) применяется в том случае, если нарушена третья, либо четвертая предпосылки нормальной линейной регрессионной модели, т. е. если случайные составляющие не имеют постоянной дисперсии или коррелированны между собой. Это характерно для неоднородной совокупности, состоящей из сильно отличающихся друг от друга единиц.

В этом случае имеет место обобщенная линейная модель множественной регрессии, которая в матричной форме выглядит так:

$$Y = X \cdot b + U .$$

$X$ – неслучайная матрица, имеющая полный ранг, т, е. $x_1, x_2,..., x_m$ – неслучайные переменные.

$M(U) = 0$.

$C_U = \sigma_u^2 C_0$ , где $C_0$ – некоторая заранее известная матрица.

Основное отличие обобщенной регрессионной модели от нормальной состоит в виде матрицы ковариации случайных составляющих – $C_U$. В нормальной модели предполагается, что матрица $C_0$ равна единичной матрице. В обобщенной модели допускается, что ковариации (следовательно, дисперсии и корреляции) остатков могут быть произвольными, следовательно, и матрица $C_0$ может содержать произвольные значения. В этом – суть обобщения нормальной модели.

Применение традиционного (обычного) МНК к обобщённой регрессионной модели дает состоятельные и несмещенные оценки. Однако данные оценки становятся неэффективными. Поэтому для оценивания параметров обобщенной модели пользуются обобщенным МНК. При этом вектор оценок параметров модели определяется по формуле:

$$\tilde{b}_{OMHK} = \left(X^T \cdot C_0^{-1} \cdot X\right)^{-1} X^T \cdot C_0^{-1} \cdot Y \qquad (2.8)$$

Следует отметить, что для обобщенной регрессионной модели, в отличие от нормальной, коэффициент детерминации $R^2$ не может служить удовлетворительной мерой качества модели. Он даже не обязан лежать в интервале от 0 до 1, а добавление или удаление независимой переменной не обязательно приводит к его увеличению или уменьшению.

Проверка гипотез о значимости коэффициентов регрессии проводится так же, как и в случае нормальной линейной модели.

Ковариационная матрица МНК-оценок параметров регрессии (которая используется для определения дисперсий и стандартных ошибок оценок параметров) будет рассчитываться как $C_{\tilde{b}} = \sigma_u^2 \left( X^T \cdot C_0^{-1} \cdot X \right)^{-1}$.

В отличие от нормальной модели регрессии $\sigma_u^2$, уже нельзя интерпретировать как величину дисперсии случайной составляющей. Однако для получения ковариационной матрицей МНК-оценок ее нужно оценить. Для этого используют следующую формулу:

$$s^2 = \frac{1}{n-h} \left( Y - X \cdot \tilde{b}_{OMHK} \right)^T \cdot C_0^{-1} \left( Y - X \cdot \tilde{b}_{OMHK} \right).$$

Для применения ОМНК необходимо знать матрицу $C_U$, что на практике бывает редко. Рекомендуется оценить каким-либо образом матрицу $C_U$, а затем использовать эту оценку в формуле (2.8) вместо $C_0$. Для этого вводят априорные ограничения на структуру матрицы $C_U$. Данный подход составляет суть так называемого доступного обобщенного метода наименьших квадратов (FGLS) и используется при оценивании параметров в модели с гетероскедастичностью, а также в случае автокорреляции случайных составляющих.

Рассмотрим ковариационную матрицу случайных составляющих в случае их *гетероскедастичности* и взаимной независимости; Если предположить, что $\sigma_{ui}$ приблизительно пропорциональна $x_i$ т.е. $x_i = \lambda \cdot \sigma_{ui}$, то ковариационная матрица может быть представлена в виде:

$$C_u = \begin{pmatrix} \frac{x_i^2}{\lambda^2} 0 \ldots 0 \\ 0 \frac{x_i^2}{\lambda^2} \ldots 0 \\ \ldots\ldots\ldots\ldots \\ 00 \ldots \frac{x_i^2}{\lambda^2} \end{pmatrix} = \frac{1}{\lambda^2} \begin{pmatrix} x_i^2 0 \ldots 0 \\ 0 x_i^2 \ldots 0 \\ \ldots\ldots\ldots\ldots \\ 00 \ldots x_i^2 \end{pmatrix}.$$

Рассмотрим ковариационную матрицу случайных составляющих в случае их *автокорреляции.* При сохранении свойства гомоскедастичности и в предположении, что случайные составляющие связаны автокорреляционной зависимостью 1-го порядка: $u_i = \rho \cdot u_{i-1} + e_i$, ковариационная матрица может быть представлена в виде:

$$C_u = \frac{\sigma_u^2}{1-\rho^2} \begin{pmatrix} 1 \rho \rho^2 \ldots \rho^{n-2} \rho^{n-1} \\ \rho 1 \rho \ldots \rho^{n-3} \rho^{n-2} \\ \ldots\ldots\ldots\ldots\ldots \\ \rho^{n-1} \rho^{n-2} \rho^{n-3} \ldots \rho 1 \end{pmatrix},$$

где $\sigma_u^2$ – дисперсия случайной составляющей, оценкой которой служит величина:

$$s_u^2 = \frac{1}{n-h}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-h}\sum_{i=1}^{n}(e_i)^2 ,$$

где $h$ – число оцениваемых параметров;
$n$ – объем выборки.

### *2.3.2. Стохастические объясняющие переменные. Обнаружение корреляции объясняющих переменных и случайной составляющей*

В нормальной линейной регрессионной модели предполагается (1-я предпосылка), что объясняющие переменные $x_j$ *(j=1;m)* являются детерминированными (нестохастическими). Это означает, что если бы пришлось повторить регрессионный анализ с другой выборкой, то значения объясняющих переменных остались бы неизменными. При этом значения зависимой переменной *y* изменились бы, т.к. изменились бы значения случайной составляющей в новой выборке.

На практике предположение о нестохастичности объясняющих переменных оказывается нереалистичным. Обычно обнаруживается, что объясняющие переменные сами были определены из других экономических зависимостей. Кроме того, при измерении объясняющих переменных могут возникать случайные ошибки.

Можно выделить три типа моделей со стохастическими объясняющими переменными:

1) *x* и *u независимо распределены*, т.е. объясняющие переменные распределены независимо от случайной составляющей;

2) *x* и *u одномоментно не коррелированны*, т. е. объясняющие переменные зависят от случайной составляющей, но их значения в каждый момент времени не коррелированы. Например, когда в качестве одной из объясняющих переменных используется лаговая зависимая переменная (модели с лаговыми переменными);

3) *x* и *u одномоментно коррелированны*, т. е. значения объясняющих переменных и случайной составляющей коррелируют в каждый момент времени. Например, когда данные подвержены воздействию ошибок измерения. Или когда осуществляется оценка параметров уравнения, входящего в состав системы одновременных уравнений (см. п. 2.4).

Рассмотрим применение традиционного (обычного) МНК для таких моделей.

*При этом будем предполагать, что ковариации, дисперсии и средние значения объясняющей переменной в генеральной совокупности стремятся к конечным пределам.*

**Замечание**: в моделях, где используются данные временных рядов, это предположение не является целесообразным. Когда модель включает переменные с трендом, имеет смысл считать, что дисперсия объясняющей переменной неограниченно увеличивается по мере расширения периода выборки.

**Случай 1:** распределение *x* имеет конечное математическое ожидание и конечную дисперсию.

1) *x* и *u независимо распределены*.

В этом случае обычный МНК сохраняет все свои важные свойства (несмещенность, состоятельность и эффективность);

2) *x* и *u одномоментно некоррелированы*. Например, когда в качестве одной из объясняющих переменных используется лаговая зависимая переменная: $y_t = a + b_0 \cdot x_t + c_1 \cdot y_{t-1} + u_t$. В данном случае $y_{t-1}$ находится непосредственно под воздействием $u_{t-1}$ и косвенно – под влиянием всех предшествующих значений случайной составляющей. Следовательно, одна из объясняющих переменных в этой модели не имеет независимого распределения от случайной составляющей, и МНК не дает несмещенных оценок. Тем не менее МНК-оценки остаются состоятельными, если $y_t$ и $u_{t-1}$ некоррелированы. Таким образом, МНК сохраняет желаемые свойства в больших выборках, хотя в малых выборках это необязательно так;

3) *x* и *u одномоментно коррелированы*. В этом случае МНК-оценка даже в больших выборках будет смещенной и несостоятельной.

**Случай 2:** дисперсия *x* неограниченно возрастает.

Если $Cov(x,u)$ имеет конечный предел, даже при корреляции *x* и *u* в каждый момент времени (3 вид моделей), МНК обеспечивает состоятельность получаемых оценок. Если $cov(x,u)$ не имеет конечного предела, то мало что можно сказать о свойствах МНК-оценок.

Обнаружить нестохастический характер объясняющих переменных (выполнение предпосылки о независимости распределения случайной составляющей $u_i$ от $x_i$) можно из графика зависимости остатков $e_i = y_i - \hat{y}_i$ от *теоретических* значений результативного признака $\hat{y}_i$ (представляющих собой функцию от $x_i$).

Если на графике получена горизонтальная полоса, то остатки $e_i$ представляют собой случайные величины и МНК оправдан. Теоретические значения $\hat{y}_i$ хорошо аппроксимируют фактические значений *y* (рис. 2.7,*а*).

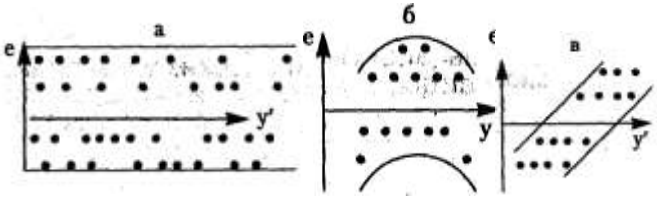


Рис. 2.7. Обнаружение корреляции *x* и *u*

Если $e_i$ зависит от $\hat{y}_i$, то применение МНК не оправдано. При этом возможны следующие случаи (рис. 2.7): *б* – криволинейной зависимости, *в* – прямолинейной зависимости.

### 2.4. Системы эконометрических уравнений. Их виды. Структурная и приведенная форма модели

Не всегда получается описать адекватно сложное социально-экономическое явление с помощью только одного соотношения (уравнения). Кроме того, некоторые переменные могут оказывать взаимные воздействия и трудно однозначно определить, какая из них является зависимой, а какая независимой переменной. Поэтому при построении эконометрической модели прибегают к системам уравнений.

Системы эконометрических уравнений включают множество эндогенных (зависимых) переменных $y_{kt}$ $(k = 1; K)$ и множество предопределенных переменных, к которым относятся лаговые и текущие экзогенные переменные – $x_{jt}$, $x_{jt}$ $(j=1;m)$, а также лаговые эндогенные переменные – $y_{kt}$ $(k = 1; K)$ $(t=1;n$ – номер наблюдения). Все эконометрические модели предназначены для объяснения текущих значений *эндогенных* переменных по значениям *предопределенных* переменных.

Система уравнений в эконометрических исследованиях может быть построена по-разному. Выделяют следующие три вида эконометрических систем.

*Система независимых уравнений*, когда каждая зависимая переменная *y* рассматривается как функция только от предопределенных переменных *x*:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + u_1; \\ y_2 = a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + u_2; \\ \ldots \\ y_k = a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{km}x_m + u_k. \end{cases}$$

*Система рекурсивных уравнений*, когда в каждом последующем уравнении системы зависимая переменная представляет функцию от всех зависимых и предопределенных переменных предшествующих уравнений:

$$\begin{cases} y_1 = a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + u_1; \\ y_2 = b_{21}y_1 + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + u_2; \\ y_3 = b_{31}y_1 + b_{32}y_2 + a_{31}x_1 + a_{32}x_2 + \ldots + a_{3m}x_m + u_3; \\ \ldots \\ y_k = b_{k1}y_1 + b_{k2}y_2 + \ldots + b_{kk-1}y_k + a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{km}x_m + u_k. \end{cases}$$

В рассмотренных двух видах систем каждое уравнение может рассматриваться самостоятельно, и параметры таких уравнений можно определить с помощью традиционного метода наименьших квадратов (МНК).

*Система взаимозависимых (совместных, одновременных) уравнений*, когда зависимые переменные в одних уравнениях входят в левую часть (т. е, выступают в роли признаков-результатов), а в других уравнениях – в правую часть системы (т. е. выступают в роли признаков-факторов):

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 \ldots + b_{1k}y_k + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + u_1; \\ y_2 = b_{21}y_1 + b_{23}y_3 \ldots + b_{2k}y_k + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + u_2; \\ \ldots \\ y_k = b_{k1}y_1 + b_{k2}y_2 + \ldots + b_{kk-1}y_k + a_{k1}x_1 + a_{k2}x_2 + \ldots + a_{km}x_m + u_k. \end{cases}$$

Название «система одновременных уравнений» подчеркивает тот факт, что в системе одни и те же переменные одновременно рассматриваются как зависимые в одних уравнениях и как независимые в других.

В отличие от предыдущих систем каждое уравнение системы одновременных уравнений не может рассматриваться самостоятельно, и для нахождения его параметров традиционный МНК неприменим, так как нарушаются предпосылки, лежащие в основе МНК:

1) причинно-следственная зависимость между переменными в уравнении. В первом уравнении $y_1$ есть функция от $y_2$, во втором уравнении $y_2$ есть функция от $y_1$;
2) факторы в такой системе мультиколлинеарны. Как следует из 2-го уравнения системы, $y_2$ зависит от $x_1$. Но в других уравнениях системы признаки $y_2$ и $x_1$ фигурируют как факторные (объясняющие переменные);
3) случайные составляющие оказываются коррелированными с объясняющими переменными.

Таким образом, нарушается 1-я предпосылка нормальной регрессионной модели о нестохастичности объясняющих переменных. В результате оценки параметров получаются смещенными и несостоятельными

*Структурная и приведенная формы системы одновременных уравнений.* Структурная форма модели описывает реальное экономическое явление или процесс. Чаще всего реальные явления или процессы настолько сложны, что для их описания системы независимых или рекурсивных уравнений не подходят, поэтому прибегают к системам одновременных уравнений. Параметры структурной формы называются структурными параметрами, или коэффициентами.

*Некоторые из уравнений структурной формы могут быть представлены в виде тождеств, т. е. уравнений заданной формы с известными параметрами.*

От структурной формы легко перейти к так называемой приведенной форме модели. Приведенная форма модели – система независимых уравнений, в которой все текущие эндогенные переменные модели выражены через предопределенные переменные модели:

$$\begin{cases} y_1 = A_{11}x_1 + \ldots + A_{1m}x_m + U_1; \\ y_2 = A_{21}x_1 + \ldots + A_{2m}x_m + U_2; \\ y_k = A_{k1}x_1 + \ldots + A_{km}x_m + U_k. \end{cases}$$

Поэтому параметры каждого из уравнений системы в приведенной форме можно определить независимо традиционным МНК. Параметры приведенной формы модели называются приведенными параметрами или коэффициентами. Приведенная фирма строится для того, чтобы по МНК-оценкам ее параметров определить оценки структурных коэффициентов.

## 2.5. Проблема идентификации модели

### *2.5.1. Необходимое и достаточное условие идентификации*

Зная оценки приведенных коэффициентов, можно определить параметры структурной формы модели. Но не всегда, а только если модель является идентифицируемой.

Модель считается точно идентифицированной, если все ее уравнения точно идентифицированы.

Модель считается неидентифицированной, если среди уравнений модели есть хотя бы одно неидентифицированное.

Модель считается сверхидентифицированной, если среди уравнений модели есть хотя бы одно сверхидентифицированное.

Уравнение называется точно идентифицированным, если оценки структурных параметров можно однозначно (единственным способом) найти по коэффициентам приведённой модели.

Уравнение сверхидентифицировано, если для некоторых структурных параметров можно получить более одного численного значений.

Уравнение называется неидентифицированным, если оценки его структурных параметров невозможно найти по коэффициентам приведенной модели.

***Правила идентификации*** **(применяются только структурной форме модели)**

Введем следующие обозначения:

$M$ – число предопределенных переменных в модели;

$m$ – число предопределенных переменных в данном уравнении;

$K$ – число эндогенных переменных в модели;

$k$ – число эндогенных переменных в данном уравнении;

$A$ – матрица коэффициентов при переменных, не входящих в данное уравнение.

*Необходимое* (но недостаточное) условие идентификации уравнения модели:

*Для того чтобы уравнение модели было идентифицируемо, необходимо, чтобы число предопределенных переменных, не входящих в уравнение, было не меньше «числа эндогенных переменных, входящих в уравнение минус 1», т.е.* $M - m > k - 1$.

Если $M - m = k - 1$, уравнение точно идентифицировано.

Если $M - m > k - 1$, уравнение сверхидентифицировано.

*Достаточное* условие идентификации уравнения модели:

Для того чтобы уравнение было идентифицируемым, достаточно, чтобы ранг матрицы А был равен (К – 1).

Ранг матрицы – размер наибольшей ее квадратной подматрицы, определитель которой не равен нулю.

Сформулируем необходимое и достаточное условия идентификации уравнения модели:

1. Если $M - m > k - 1$ и ранг матрицы $A$ равен $K - 1$, то уравнение сверхидентифицировано.

2. Если $M - m = k - 1$ и ранг матрицы $A$ равен $K - 1$, то уравнение точно идентифицировано.

3. Если $M - m >= k - 1$ и ранг матрицы $A$ меньше $K - 1$, то уравнение неидентифицировано.

4. Если $M - m < k - 1$, то уравнение неидентифицировано. В этом случае ранг матрицы $A$ будет меньше $K - 1$.

Рассмотрим **пример.**

Пусть имеется система:

$$\begin{cases} Y_1 = b_{12}Y_2 + b_{13}Y_3 + a_{11}X_1 + a_{12}X_2 + u_1; \\ Y_2 = b_{21}Y_1 + a_{21}X_1 + u_2; \\ Y_3 = b_{32}Y_2 + a_{31}X_1 + a_{33}X_3 + u_3. \end{cases}$$

*Требуется* составить приведенную форму модели, проверить каждое уравнение структурной модели на идентификацию.

Решение:

В данной системе:

$Y_1, Y_2, Y_3$ – эндогенные переменные ($K = 3$);

$X_1, X_2, X_3$ – предопределенные переменные ($M = 3$).

Составим приведенную форму модели:

$$\begin{cases} Y_1 = A_{11}X_1 + A_{12}X_2 + A_{13}X_3 + U_1; \\ Y_2 = A_{21}X_1 + A_{22}X_2 + A_{23}X_3 + U_2; \\ Y_3 = A_{31}X_1 + A_{32}X_2 + A_{33}X_3 + U_3. \end{cases}$$

Проверим, как выполняется необходимое условие идентификации для каждого уравнения.

Для 1-го уравнения имеем: $k_1 = 3$; $m_1 = 2$;

$M - m_1 = 1 < k_1 - 1 = 2$, следовательно, 1-е уравнение неидентифицировано.

Для 2-го уравнения имеем: $k_2 = 2$; $m_2 = 1$;

$M - m_2 = 1 > k_2 - 1 = 1$, следовательно, 2-е уравнение сверхидентифицировано.

Для 3-го уравнения имеем: $k_3 = 2$; $m_3 = 2$;

$M - m_3 = 1 = k_3 - 1 = 1$, следовательно, 3-е уравнение точно идентифицировано.

Рассмотрим как выполняется достаточное условие идентификации для каждого уравнения системы. Для того чтобы оно выполнялось, необходимо, чтобы определитель матрицы $A$ (матрицы коэффициентов при переменных, не входящих в это уравнение) был равен $K - 1 = 2$.

Составим матрицу $A$ для 1-го уравнения системы. В 1-м уравнении отсутствует лишь одна переменная системы $X_3$. Поэтому матрица $A$ будет иметь вид:

$X_3$

$\begin{pmatrix} 0 \\ a_{33} \end{pmatrix}$ – во 2-м уравнении;
– в 3-м уравнении.

Ранг данной матрицы равен 1, что меньше $K - 1 = 2$, следовательно, 1-е уравнение модели неидентифицировано.

Составим матрицу $A$ для 2-го уравнения системы. Во 2-м уравнении отсутствуют переменные $Y_3$, $X_2$, $X_3$:

$Y_3$, $X_2$, $X_3$

$\begin{pmatrix} b_{13} a_{13} 0 \\ 1 a_{32} a_{33} \end{pmatrix}$

Ранг данной матрицы равен 2, что равно $K - 1 = 2$, следовательно, 2-е уравнение модели точно идентифицировано.

Составим матрицу $A$ для 3-го уравнения системы. В 3-м уравнении отсутствуют переменные $Y_1$, $X_2$:

$\begin{pmatrix} 1 a_{12} \\ b_{21} 0 \end{pmatrix}$ – в 1-м уравнении;
– во 2-м уравнении.

Ранг данной матрицы равен 2, что равно $K - 1 = 2$, следовательно, 3-е уравнение модели идентифицировано.

Сделаем выводы: 1-е уравнение системы неидентифицировано (т. к. не выполняются достаточное и необходимое условия идентификации). 2-е

уравнение системы сверхидентифицировано. Следовательно, система в целом является неидентифицируемой.

Рассмотрим еще один **пример.** Дана структурная модель спроса и предложения:

1 – уравнение предложения: $Q_t^s = a_0 + a_1 \cdot P_t + a_2 \cdot P_{t-1}$;

2 – уравнение спроса: $Q_t^d = b_0 + b_1 \cdot P_t + b_2 \cdot I_t$;

3 – тождество равновесия: $Q_t^s = Q_t^d$.

Проверим модель на идентифицикацию и составим приведенную форму модели.

Учитывая тождество, система примет вид:

$$\begin{cases} Q_t = a_0 + a_1 \cdot P_t + a_2 \cdot P_{t-1} + u_{1t}; \\ Q_t = b_0 + b_1 \cdot P_t + b_2 \cdot I_t + u_{2t}. \end{cases}$$

В данной модели:

- эндогенными переменными являются взаимозависимые переменные: $P_t$ – цена; $Q_t$ –количество товара;
- предопределенными переменными являются: $I_t$ – до ход (экзогенная переменная); $P_{t-1}$ – цена в предыдущий период времени (лаговая эндогенная переменная);
- случайными переменными являются: $u_{1t}$, $u_{2t}$;
- структурными параметрами модели являются: $a_0$, $a_1$, $a_2$, $b_0$, $b_1$, $b_2$.

Таким образом, число эндогенных переменных модели $K$ = 2; число предопределенных переменных модели $M$ = 2.

Проверим выполнение необходимого условия идентификаций.

Для 1-го уравнения модели (функции спроса) $k$ = 2; $m$ = 1.

$M–m$=2–1=1=$k$–1=2–1=1, следовательно, уравнение точно идентифицировано.

Для 2-го уравнения (функция предложения) $k$ = 2; $m$ = 1.

$M–m$=2–1=1=$k$–1=2–1=1, следовательно, уравнение точно идентифицировано.

Проверим выполнение достаточного условия идентификации. Так как $K$ – 1 = 1, то достаточно, чтобы хотя бы один из коэффициентов матрицы $A$ был не равен нулю.

В 1-м уравнении отсутствует только переменная $I_t$ Матрица $A = (b_2)$. Определитель данной матрицы не равен нулю, следовательно ранг = 1 = $K$ – 1 и уравнение идентифицируемо.

Во 2-м уравнении отсутствует только переменная $P_{t-1}$. Матрица $A = (a_2)$. Определитель данной матрицы не равен нулю, следовательно ранг = 1 = $K$ – 1 и уравнение идентифицируемо.

Сделаем выводы: 1-е и 2-е уравнения системы точно идентифицированы. Следовательно, система в целом является точно идентифицируемой.

Составим приведенную форму. Приведенная форма содержит 2 уравнения (число уравнений равно числу текущих эндогенных переменных модели). Каждое уравнение представляет зависимость эндогенной переменной от предопределенных переменных модели (дохода и цены в предыдущий период).

Для количества товара – $Q_t = A_1 + A_2 \cdot I_t + A_3 \cdot P_{t-1} + U_1$,

Для цены – $P_t = B_1 + B_2 \cdot I_t + B_3 \cdot P_{t-1} + U_2$.

### *2.5.2. Оценка точно идентифицированного уравнения. Косвенный метод наименьших квадратов (КМНК – ILS)*

Оценка точно идентифицированного уравнения осуществляется с помощью *косвенного метода наименьших квадратов* (КМНК).

Алгоритм КМНК включает 3 шага:

1. составление приведенной формы модели и выражение каждого коэффициента приведенной формы через структурные параметры;
2. применение обычного МНК к каждому уравнению приведенной формы и получение численных оценок приведенных параметров;
3. определение оценок параметров структурной формы по оценкам приведенных коэффициентов, используя соотношения, найденные на 1-м шаге.

*Примечание:* при небольшом числе переменных можно не определять приведенные коэффициенты через структурные параметры, что приводит к необходимости решения нелинейной системы уравнений, а воспользоваться более простым приемом – получить из имеющихся приведенных уравнений структурные уравнения.

Рассмотрим **пример.** Пусть дана структурная форма модели спроса и предложения:

$Q_t = a_0 + a_1 \cdot P_t + a_2 \cdot P_{t-1} + u_{1t}$;

$Q_t = b_0 + b_1 \cdot P_t + b_2 \cdot I_t + u_{2t}$.

В данной модели эндогенными переменными (т.е. определяемыми внутри модели) являются взаимозависимые переменные: $P_t$ – цена; $Q_t$ –количество товара.

• *предопределенными переменными (которые определяют значения эндогенных переменных) являются:*

$I_t$ – доход (экзогенная переменная);

$P_{t-1}$ – цена в предыдущий период времени (лаговая эндогенная переменная);

- случайными переменными являются: $u_{1t}$, $u_{2t}$;
- структурными параметрами модели являются: $a_0$, $a_1$, $a_2$, $b_0$, $b_1$, $b_2$.

Имеются данные за 6 периодов времени по переменным (табл. 2.7).

Таблица 2.7

| $t$ | $Q_t$ | $P_t$ | $I_t$ | $P_{t-1}$ |
|---|---|---|---|---|
| 1 | 105 | 15 | 14 | 12 |
| 2 | 130 | 12 | 16 | 15 |
| 3 | 100 | 14 | 12 | 12 |
| 4 | 120 | 15 | 15 | 14 |
| 5 | 125 | 14 | 16 | 15 |
| 6 | 120 | 15 | 17 | 14 |
| Итого | 700 | 85 | 90 | 82 |

Найдем оценки структурных параметров модели.

Ранее было показано, что данная модель точно идентифицирована.

Поэтому для оценки структурных параметров модели можно применить косвенный МНК:

*1 шаг.* Составим приведенную форму:

$$Q_t = A_1 + A_2 \cdot I_t + A_3 \cdot P_{t-1} + U_1;$$

$$P_t = B_1 + B_2 \cdot I_t + B_3 \cdot P_{t-1} + U_2.$$

*2 шаг.* С помощью обычного МНК найдем оценки приведенных коэффициентов. В соответствии с методикой МНК, система нормальных уравнений для расчета параметров 1-го приведенного уравнения ($A_1$, $A_2$, $A_3$) примет вид:

$$\begin{cases} \sum Q_t = n \cdot A_1 + A_2 \cdot \sum I_t + A_3 \cdot \sum P_{t-1}; \\ \sum Q_t \cdot I_t = A_1 \cdot \sum I_t + A_2 \cdot \sum I_t^{\,2} + A_3 \cdot \sum I_t \cdot P_{t-1}; \\ \sum Q_t \cdot P_{t-1} = A_1 \cdot \sum P_{t-1} + A_2 \cdot \sum P_t \cdot I_t + A_3 \cdot \sum P_{t-1}^2. \end{cases}$$

По данным табл. 11 имеем:

$$\begin{cases} 700 = 6A_1 + 90A_2 + 82A_3; \\ 10590 = 90A_1 + 1366A_2 + 1240A_3; \\ 9645 = 82A_1 + 1240A_2 + 1130A_3. \end{cases}$$

Решив систему находим, что $A_1$ = 1,55; $A_2$ = 1,15; $A_3$ = 7,16.

Аналогично определяем параметры 2-го приведенного уравнения ($B_1$, $B_2$, $B_3$).

Окончательно оцененная приведенная форма модели имеет вид:

$$Q_t = 1{,}55 + 1{,}15 \cdot I_t + 7{,}16 \cdot P_{t-1} \quad (1 - e\_уравнение);$$

$$P_t = 19{,}34 + 0{,}55 \cdot I_t - 0{,}99 \cdot P_{t-1} \quad (2 - e\_уравнение).$$

*3 шаг.* По оцененной приведенной форме получим 1-е уравнение структурной формы модели – зависимость $Q_t$ от $P_t$ и $P_{t-1}$. Для этого выразим из 2-го уравнения приведенной формы $I_t$:

$I_t$ = –19,34/0,55+(0,99/0,55) $P_{t-1}$+*(1/0,55)*$P_t$ =
= –34,9+1,78 $P_{t-1}$+1,8 $P_t$ .

Затем подставим полученное выражение в 1-е уравнение приведенной формы:

$$Q_t = 1{,}55 + 1{,}15 \cdot (-34{,}9 + 1{,}78\,P_{t-1} + 1{,}8\,P_t) + 7{,}16 \cdot P_{t-1} =$$

=–38,54+2,07 $P_t$+9,21$P_{t-1}$.

Следовательно, $a_0$ = 38,54; $a_1$ = –2,07; $a_2$ =–9,21.

Аналогично можно определить параметры 2-го Структурного уравнения, представляющего собой зависимость $Q_t$ от $P_t$ и $I_t$. Для этого выразим из 2-го уравнения приведенной формы $P_{t-1}$:

$P_{t-1}$ =19,34/0,99 + (0,55/0,99) $I_t$– (1/0,99) $P_t$ =
=19,6 + 0,56 $I_t$ –1,01 $P_t$.

Затем подставим полученное выражение в 1-е уравнение приведенной формы:

$$Q_t = 1{,}55 + 1{,}15 \cdot I_t + 7{,}16 \cdot (19{,}6 + 0{,}56\,I_t - 1{,}01\,P_t) =$$

=141,95–7,26 $P_t$+5,17$I_t$.

Следовательно, $b_0$ =141,95; $b_1$ = –7,26; $b_2$ = 5,17.

В результате структурная форма модели имеет вид:

$Q_t$ =–38,54+2,07 $P_t$+9,21$P_{t-1}$ $R^2$ = 0,923;

$Q_t$ =141,95–7,26 $P_t$+5,17$I_t$ $R^2$ = 0,77.

Для сравнения обычный МНК дает следующие результаты:

$Q_t$ =5,2–0,166 $P_t$+8,33$P_{t-1}$ $R^2$ = 0,962;

$Q_t$ =90,38–3,85 $P_t$+5,38$I_t$ $R^2$ = 0,887.

Это смещенные оценки структурных параметров модели.

### *2.5.3. Оценка сверхидентифицированного уравнения. Двухшаговый метод наименьших квадратов (ДМНК – 2 SLS)*

Оценка сверхидентифицируемого уравнения осуществляется при помощи *двухшагового метода наименьших квадратов* (ДМНК – 2 SLS).

Алгоритм двухшагового МНК включает следующие шаги:

1) составление приведенной формы модели;

2) применение обычного МНК к каждому уравнению приведенной формы и получение численных оценок приведенных параметров;

3) определение расчетных значений эндогенных переменных, которые фигурируют в качестве факторов в структурной форме модели;

4) определение структурных параметров каждого уравнения в отдельности обычным МНК, используя в качестве факторов входящие в это уравнение предопределенные переменные и расчетные значения эндогенных переменных, полученные на шаге 1.

Параметры сверхидентифицированной функции предложения нельзя определить косвенным МНК. Обычный МНК также нельзя применять, так как

в этом случае были бы нарушены предпосылки нормальной линейной модели регрессии. Нарушение этих предпосылок связано с наличием в уравнении в качестве фактора эндогенной переменной $y_t$.

Предположим, что мы нашли переменную $\hat{y}_t$, которая имеет 2 свойства:

- тесно коррелирует с переменной $y_t$;
- не коррелирует со случайной составляющей.

Такие переменные в эконометрике называются *инструментальными* переменными. Они отвечают предпосылкам нормальной линейной регрессионной модели.

Если мы теперь заменим в уравнении регрессии переменную $y_t$ (выступающую в роли фактора) инструментальной переменной $\hat{y}_t$, то к такому преобразованному уравнению регрессии можно применить обычный МНК.

Рассмотрим **пример.** Вернемся к модели спроса и предложения (см. п.2.5.2) и введем новую предопределенную переменную $R_t$ – благосостояние потребителей в уравнение спроса:

$Q_t = a_0 + a_1 \cdot P_t + a_2 \cdot P_{t-1} + u_{1t}$ – функция предложения;

$Q_t = b_0 + b_1 \cdot P_t + b_2 \cdot I_t + a_3 \cdot R_t + u_{2t}$ – функция спроса.

Исходные данные приведены в табл. 2.8.

Таблица 2.8

| $t$ | $Q_t$ | $P_t$ | $I_t$ | $P_{t-1}$ | $R_t$ | $\hat{P}_t$ |
|---|---|---|---|---|---|---|
| 1 | 105 | 15 | 14 | 12 | 25 | 14,85 |
| 2 | 130 | 12 | 16 | 15 | 25 | 12,23 |
| 3 | 100 | 14 | 12 | 12 | 25 | 14,05 |
| 4 | 120 | 15 | 15 | 14 | 30 | 15,20 |
| 5 | 125 | 14 | 16 | 15 | 28 | 13,57 |
| 6 | 120 | 15 | 17 | 14 | 28 | 15,10 |
| Итого | 700 | 85 | 90 | 82 | 161 | 85 |

Найдем оценки параметров структурной формы модели.

В этой модели функция спроса точно идентифицирована, а функция предложения сверхидентифицирована; при этом выполняется достаточное условие идентификации.

Параметры сверхидентифицированной функции предложения нельзя определить косвенным МНК. Обычный МНК также нельзя применять, т. к. в этом случае были бы нарушены предпосылки применения этого метода. Нарушение этих предпосылок связано с наличием в уравнении в качестве фактора эндогенной переменной $P_t$.

Используем двухшаговый МНК для оценок параметров уравнения предложения.

*1 шаг.* Выпишем приведенную форму модели:

$Q_t = A_1 + A_2 \cdot I_t + A_3 \cdot P_{t-1} + A_4 \cdot R_t + U_1$;

$P_t = B_1 + B_2 \cdot I_t + B_3 \cdot P_{t-1} + B_4 \cdot R_t + U_2$.

Из второго уравнения модели можно найти расчетные значения $\hat{P}_t$:

$\hat{P}_t = \tilde{B}_1 + \tilde{B}_2 \cdot I_t + \tilde{B}_3 \cdot P_{t-1} + \tilde{B}_4 \cdot R_t$.

Второе уравнение приведенной формы можно переписать в виде:

$P_t = \hat{P}_t + U_2$.

Таким образом, переменная $P_t$ состоит из двух элементов:

• $\hat{P}_t$, которая есть линейная комбинация преопределенных переменных ($I_t$, $R_{t-1}$ и $R_t$);

• $U_2$, которая в соответствии с предпосылкой МНК не коррелирует с переменной $\hat{P}_t$.

Сверхидентифицированную функцию предложения можно переписать в виде:

$Q_t = a_0 + a_1 \cdot \left(\hat{P}_t + U_2\right) + a_2 \cdot P_{t-1} + u_{1t}$ или

$Q_t = a_0 + a_1 \cdot \hat{P}_t + a_2 \cdot P_{t-1} + \hat{u}_{1t}$

где $\hat{u}_{1t} = u_{1t} + a_1 \cdot U_2$.

Полученное уравнение отличается от исходной функции предложения только тем, что переменная $P_t$ заменена на ее оценку $\hat{P}_t$, и ошибкой $\hat{u}_1$.

Переменная $\hat{P}_t$ имеет следующие свойства:

1) она тесно коррелирует с $P_t$;

2) она не коррелирует с ошибкой $\hat{u}_1$.

Таким образом, она является инструментальной переменной.

*2 шаг*. Определим обычным МНК параметры приведенной формы модели, используя данные таблицы 2.8:

$\hat{Q}_t = 7{,}8 + 1{,}28 \cdot I_t + 7{,}29 \cdot P_{t-1} - 0{,}37 \cdot R_t$;

$\hat{P}_t = 11{,}79 + 0{,}4 \cdot I_t - 1{,}14 \cdot P_{t-1} + 0{,}445 \cdot R_t$.

4 *шаг*. Подставим во 2-е уравнение приведенной формы фактические значения $I_t$, $R_{t-1}$, $R_t$ из таблицы, найдем расчетные значения $\hat{P}_t$ и добавим их в табл. 2.8.

Например $\hat{P}_t$ = 11,79 + 0,4 · 14 - 1,14 · 12 + 0,445 · 25 = 14,85.

5 *шаг.* Применим обычный МНК к уравнению:

$Q_t = a_0 + a_1 \cdot \hat{P}_t + a_2 \cdot P_{t-1} + \hat{u}_{1t}$

В результате получим:

$\hat{Q}_t = 0{,}019 + 0{,}1 \cdot \hat{P}_t + 8{,}43 \cdot P_{t-1} \quad R^2 = 0{,}962;$

Таким образом, структурные параметры уравнения предложения найдены.

Параметры точно идентифицированной функции спроса можно определить двумя способами:

Двухшаговым МНК:

$Q_t = b_0 + b_1 \cdot \hat{P}_t + b_2 \cdot I_t + a_3 \cdot R_t + u_{2t}$

$Q_t = 83{,}37 - 6{,}41 \cdot \hat{P}_t + 3{,}83 \cdot I_t + 2{,}48 \cdot R_t \quad R^2 = 0{,}976.$

Косвенным МНК (т.к. функция спроса точно идентифицирована).

Для этого выразим из 2-го уравнения приведенной формы переменную $P_{t-1}$:

$P_{t-1} = 10{,}37 - 0{,}88 \cdot P_t + 0{,}35 \cdot I_t + 0{,}39 \cdot R_t.$

Подставим найденное выражение вместо $P_{t-1}$ в 1-е уравнение приведенной формы и получим:

$Q_t = 83{,}37 - 6{,}41 \cdot P_t + 3{,}83 \cdot I_t + 2{,}48 \cdot R_t.$

Это 2-е уравнение структурной формы, параметры которого найдены КМНК.

Выпишем *структурную форму* модели:

$Q_t = 0{,}019 + 0{,}1 \cdot \hat{P}_t + 8{,}43 \cdot P_{t-1} \quad R^2 = 0{,}962.$

$Q_t = 83{,}37 - 6{,}41 \cdot \hat{P}_t + 3{,}83 \cdot I_t + 2{,}48 \cdot R_t \quad R^2 = 0{,}976.$

Для сравнения обычный МНК дает следующие результаты:

$Q_t = 5{,}2 - 0{,}166 \cdot \hat{P}_t + 8{,}328 \cdot P_{t-1} \quad R^2 = 0{,}962.$

$Q_t = 82{,}16 - 6{,}26 \cdot \hat{P}_t + 3{,}87 \cdot I_t + 2{,}43 \cdot R_t \quad R^2 = 0{,}983.$

! Хотя функция предложения по сравнению с предыдущей моделью, в которую не входила переменная $R_t$ не изменилась, оценки ее структурных параметров изменились, т.к. изменилась вся модель: в нее была включена дополнительная предопределенная переменная. А обычный МНК не привел бы к изменению оценок структурных параметров функции.

Выделим 3 главные особенности двухшагового МНК.

1. Двухшаговый МНК может применяться для оценки не только сверхидентифицированных, но и точно идентифицированных уравнений. В этом случае оценки, полученные двухшаговым и косвенным МНК, совпадут.
2. Если значения коэффициентов детерминации по у нениям приведенной формы велико и превышает 0,8 ($R^2 > 0{,}8$), то оценки структурных параметров, полученные двухшаговым и обычным МНК, будут близки. Причина в том, что при высоком $R^2$ расчетные значения

инструментальных переменных не будут сильно отличаться от фактического значения соответствующих эндогенных переменных.

3. Однако если коэффициент детерминации $R^2$ для приведенного уравнения низкий, то расчетные значения эндогенной переменной будут плохой аппроксимацией ее фактических значений и применение двухшагового МНК может оказаться неэффективным.

**Тесты по разделу**

**1. Из перечисленных условий: 1) большое число наблюдений, 2) незначительное число наблюдений, 3) маленькие выборочные дисперсии объясняющих переменных, 4) большие выборочные дисперсии объясняющих переменных, 5) большая дисперсия случайного члена, 6) малая дисперсия случайного члена - к условиям, благоприятствующим получению надежных оценок регрессии, относятся**

a) 2,3,5
b) 1,3,5
c) 2,4,6
d) 1,4,6

**2. При проведении теста Голдфелда Квандта предполагается, что стандартное отклонение остаточного члена регрессии растет с _____ переменной**

a) падением зависимой
b) ростом зависимой
c) ростом объясняющей
d) падением объясняющей

**3. Коэффициент ранговой корреляции имеет дисперсию**

a) $n/(n+1)$
b) $n/(n-1)$
c) $1/(n-1)$
d) $n-1$

**4. Совокупность фиктивных переменных – некоторое количество фиктивных переменных, предназначенное для описания**

a) эталонной категории
b) регрессионной модели
c) набора категорий
d) одной категории

**5. Модель множественной регрессии без свободного коэффициента имеет вид: *y* =**

a) $\alpha + x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + u$

b) $\alpha + x_1 + x_2 + \ldots + x_m + u$

c) $\alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m$

d) $\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + u$

**6. Фиктивная переменная для коэффициента наклона предназначена для установление влияния категории на**

a) коэффициент при нефиктивной переменной
b) коэффициент при фиктивной переменной
c) случайный член регрессии
d) свободный член регрессии

**7. Ловушка duymmy trap приводит к**

a) полной коллинеарности
b) потере эффективности оценок
c) смещению оценок регрессии
d) мультиколлинеарности

**8. Функция спроса $y = \alpha x^{\beta} p^{\gamma} v$ может быть линеаризована посредством**

a) потенцирования
b) возведения в степень
c) дифференцирования
d) логарифмирования

**9. Пересмотр оценок в методе Кокрана-Оркатта выполняется до тех пор, пока не будет _____ оценок**

a) получена требуемая точность
b) получено необходимое значение
c) получено необходимое количество
d) выполнено заданное число итераций

**10. Значения t-статистики для фиктивных переменных незначимо отличаются от**

a) –1
b) 0
c) ½
d) 1

**11. Строгая линейная зависимость между переменными – ситуация, когда _____ двух переменных равна 1 или –1**

a) теоретическая корреляция
b) сумма
c) разность
d) выборочная корреляция

**12. При положительной автокорреляции *DW***

a) =0

b) <2
c) >2
d) >1

**13. Функция Кобба-Дугласа является**

a) функцией спроса
b) функцией предложения
c) производственной функцией
d) целевой функцией потребления

**14. Явление, когда строгая линейная зависимость между переменными приводит к невозможности применения МНК, называется**

a) полной коллинеарностью
b) неопределенностью
c) мультиколлинеарностью
d) детерминированностью

**15. Если вычисленное значение статистики Спирмена превысит некое критическое значение, то принимается решение о**

a) наличии гетероскедастичности
b) отсутствий гетероскедастичности
c) отсутствии мультиколлинеарности
d) наличии мультиколлинеарности

**16. Стандартные отклонения коэффициентов регрессии обратно пропорциональны величине _____, где *n* – число наблюдений**

a) $n^2$
b) $n^3$
c) $n$
d) $\sqrt{n}$

**17. Если независимые переменные имеют ярко выраженный временной тренд, то они оказываются**

a) имеющими большое влияние
b) тесно коррелированными
c) малозначимыми
d) слабо коррелированными

**18. Процесс выбора необходимых для регрессии переменных и отбрасывание лишних переменных называется**

a) прогнозированием
b) унификацией переменных
c) моделированием
d) спецификацией переменных

**19. Явление, когда нестрогая линейная зависимость между объясняющими переменными в модели множественной регрессии приводит к получению ненадежных оценок регрессии, называют**

a) коррелированностью
b) мультиколлинеарностью
c) детерминированностью
d) смещенностью

**20. Гетероскедастичность заключается в том, что дисперсия случайного члена регрессии _____ наблюдений**

a) зависит от числа
b) зависит от номера
c) зависит от времени проведения
d) одинакова для всех

**21. Из перечисленного: 1) число объясняющих переменных, 2) количество наблюдений в выборке, 3) конкретные значений переменных. Критическое значение статистики Дарбина-Уотсона зависят от**

a) 1,2,3
b) 1,2
c) 3
d) 1,3

**22. Как правило в эталонной категории**

a) все фиктивные переменные равны 1
b) все фиктивные переменные равны 0
c) только одна из фиктивных переменных равна 1
d) только одна из фиктивных переменных равна 0

**23. Автокорреляция – нарушение условия Гаусса-Маркова**

a) второго
b) первого
c) четвертого
d) третьего

**24. Положительная автокорреляция – ситуация, когда случайный член регрессии в следующем наблюдении ожидается**

a) того же знака, что и в первом наблюдении
b) того же знака, что и в настоящем наблюдении
c) противоположного знака по сравнению с настоящим наблюдением
d) противоположного знака по сравнению с настоящим наблюдением

**25. Коэффициенты при сезонных фиктивных переменных показывают _____ при смене сезона**

a) трендовые изменения

b) численную величину изменения, происходящего
c) изменения числа потребителей
d) направление изменения, происходящего

**Вопросы для повторения раздела**

**1.** Как определяется модель множественной линейной регрессии?
**2.** В чем суть МНК для построения множественного линейного уравнения регрессии?
**3.** Как определяется статистическая значимость коэффициентов регрессии?
**4.** Что такое автокорреляция остатков и каковы ее виды?
**5.** В чем суть статистики Дарбина-Уотсона и как она связана с коэффициентом корреляции между соседними отклонениями?
**6.** Как анализируется статистическая значимость статистики Дарбина-Уотсона?
**7.** Каковы признаки качественной регрессионной модели?
**8.** В чем суть гетероскедастичности?
**9.** Почему при наличии гетероскедастичности МНК позволяет получить более эффективные оценки, чем обычный МНК?
**10.** Что такое автокорреляция? Назовите основные причины автокорреляции.
**11.** Какие последствия автокорреляции? Перечислите основные методы обнаружения автокорреляции.
**12.** Объясните значения терминов «коллинеарность» и «мультиколлинеарность».
**13.** Каковы основные последствия мультиколлинеарности? Перечислите основные методы устранения мультиколлинеарности.
**14.** Каковы основные причины использования фиктивных переменных в регрессионных моделях?

## 3. Временные ряды и динамические процессы

### 3.1. Автокорреляция уровней временного ряда и выявление его структуры

Модели, построенные по временным данным, представляют модели временных рядов.

Временной ряд $X_t$ *(t=1; n)* – ряд значений какого-либо показателя, характеризующих один и тот же объект за несколько последовательных моментов или периодов времени. Уровень временного ряда $X_t$ складывается из следующих основных компонентов:

• трендовой компоненты, характеризующей основную тенденцию уровней ряда (*T*);

• циклической, или периодической, компоненты, характеризующей циклические или периодические колебания изучаемого явления. Различают конъюнктурную компоненту (*K*), связанную с большими экономическими циклами, и сезонную компоненту (*S*), связанную с внутригодовыми колебаниями уровней ряда;

• случайной компоненты, которая является результатом воздействия множества случайных факторов (E).

Тогда уровень ряда можно представить как функцию от этих компонент: $X = f(T,K,S,E)$.

В зависимости от вида связи между этими компонентами может быть построена либо аддитивная модель: $X = T + K + S + E$, либо мультипликативная модель: $X = T \cdot K \cdot S \cdot E$ ряда динамики.

Для выявления структуры ряда (т. е. состава компонент) строят автокорреляционную функцию. Дело в том, что наличии во временном ряде трендовой и циклической компонент значения последующего уровня ряда зависят от предыдущих.

Автокорреляция уровней ряда – корреляционная между последовательными уровнями одного и того же ряда динамики (сдвинутыми на определенный промежуток времени *L* – лаг). То есть связь между рядом: $X_1$, $X_2$, .... $X_{n-L}$ и рядом $X_{1+L}$, $X_{2+L}$, .... $X_n$, где *L* – положительное целое число. Автокорреляция может быть измерена коэффициентов автокорреляции:

$$r_{t,t-L} = \frac{\overline{X_t \cdot X_{t-L}} - \overline{X_t} \cdot \overline{X_{t-L}}}{\sigma_t \cdot \sigma_{t-L}},$$

где $\overline{X_t \cdot X_{t-L}} = \frac{\sum_{i=1+L}^{n} X_i \cdot X_{i-L}}{n-L}$;

$\overline{X_t} = \frac{\sum_{i=1+L}^{n} X_i}{n-L}$ – средний уровень ряда

($X_{1+L}$, $X_{2+L}$, .... $X_n$);

$\overline{X_{t-L}} = \frac{\sum_{i=1+L}^{n} X_{i-L}}{n-L}$ – средний уровень ряда

($X_1$, $X_2$, .... $X_{n-L}$).

$\sigma_t \cdot \sigma_{t-L}$ – средние квадратические отклонения, для рядов ($X_{1+L}$, $X_{2+L}$, .... $X_n$) и

($X_1$, $X_2$, .... $X_{n-L}$) соответственно.

Лаг (сдвиг во времени) определяет порядок коэффициента автокорреляции. Если $L$ = 1, то имеем коэффициент автокорреляции 1-го порядка $r_{t,t-1}$,. если $L$ = 2, то коэффициент автокорреляции 2-го порядка $r_{t,t-2}$ и т.д. Следует учитывать, что с увеличением лага на единицу число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается на 1. Поэтому обычно рекомендуют максимальный порядок коэффициента автокорреляции, равный $n/4$.

Рассчитав несколько коэффициентов автокорреляции, можно определить лаг (I), при котором автокорреляция ($r_{t,t-L}$) наиболее высокая, выявив тем самым *структуру временного ряда.* Если наиболее высоким оказывается значение $r_{t,t-1}$, то исследуемый ряд додержит только тенденцию. Если наиболее высоким оказался $r_{t,t-L}$, то ряд содержит (помимо тенденции) колебания периодом $L$. Если ни один из $r_{t,t-l}$ ($l=1;L$) не является значимым, можно сделать одно из двух предположений:

- либо ряд не содержит тенденции и циклических колебаний, а его уровень определяется только случайной компонентой;
- либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ.

Последовательность коэффициентов автокорреляции 1, 2 и т.д. порядков называют автокорреляционной функцией временного ряда. График зависимости значений коэффициентов автокорреляции от величины лага (порядка коэффициента автокорреляции) называют коррелограммой.

Рассмотрим **пример:** Пусть имеются данные предприятия об объемах выпуска некоторого товара по кварталам за 3 года в тыс. шт. (табл. 3.1)

Таблица 3.1

| Год | 1 | | | | г | | | | 3 | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Квартал | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 | 1 | 2 | 3 | 4 |
| Объем выпуска($X_t$) | 410 | 400 | 715 | 600 | 585 | 560 | 975 | 800 | 765 | 720 | 1235 | 1100 |

Определим структуру данного временного ряда. Для этого рассчитаем коэффициенты автокорреляции 1, 2, 3, 4, 5 порядков.

Чтобы найти коэффициент корреляции 1-го порядка, нужно найти корреляцию между рядами (расчет производится не по 12, а по 11 парам наблюдений):

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $X_t$ | 400 | 715 | 600 | 585 | 560 | 975 | 800 | 765 | 720 | 1235 | 1100 |
| $X_{t-1}$ | 410 | 400 | 715 | 600 | 585 | 560 | 975 | 800 | 765 | 720 | 1235 |

Тогда коэффициент автокорреляции 1-го порядка будет равен $r_{t,t-1} = 0{,}538$.

Коэффициент корреляции 2-го порядка между рядами:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $X_t$ | 715 | 600 | 585 | 560 | 975 | 800 | 765 | 720 | 1235 | 1100 |
| $X_{t-2}$ | 410 | 400 | 715 | 600 | 585 | 560 | 975 | 800 | 765 | 720 |

будет равен $r_{t,t-2} = 0{,}286$ (расчет в данном случае производится не по 12, а по 10 парам наблюдений).

Аналогично рассчитываются коэффициенты автокорреляции 3-го, 4-го и 5-го порядков. Результаты расчета представим в виде таблицы коррелограммы (табл. 3.2).

Таблица 3.2

| Лаг (порядок) | $r_{t,t-l}$ | Коррелограмма |
|---|---|---|
| 1 | 0,538 | **** |
| 2 | 0,286 | * |
| 3 | 0,432 | *** |
| 4 | 0,992 | ***** |
| 5 | 0,373 | ** |

*Вывод:* в данном ряду динамики имеется тенденция и периодические колебания с периодом, равным 4.

### 3.2. Моделирование тенденции временного ряда (построение тренда)

Для выявления основной тенденции (тренда) в уровнях ряда, т. е. выравнивания ряда динамики, используются различные методы:

- методы механического выравнивания (без использования количественной модели);
- метод аналитического выравнивания (с использованием количественной модели).

Методы механического выравнивания (скользящих средних, экспоненциального сглаживания и др.) рассматривались подробно в статистике. В эконометрике основное внимание уделяется методу аналитического выравнивания. Данный метод заключается в построении уравнения регрессии, характеризующего зависимость уровней ряда от временной переменной: $\hat{X} = f(t)$.

При выборе вида функции тренда можно воспользоваться методом конечных разностей (обязательным условием применения данного подхода является равенство интервалов между уровнями ряда).|

Конечными разностями первого порядка являются разности между последовательными уровнями ряда:

$\Delta_t^1 = X_t - X_{t-1}$ ($t = 2; n$).

*Конечными разностями второго порядка являются разности между последовательными конечными разностями 1-гопорядка:*

$\Delta_t^2 = \Delta_t^1 - \Delta_{t-1}^1 = X_t - X_{t-1} - X_{t-1} - X_{t-2} = X_t - 2X_{t-1} - X_{t-2}$ (t = 3; n).

Конечными разностями $j$-го порядка являются разности между последовательными конечными разностями ($j$–1)-го порядка:

$\Delta_t^j = \Delta_t^{j-1} - \Delta_{t-1}^{j-1}$ ($t = j+1; n$).

Если общая тенденция выражается линейным уравнением $\hat{X} = a + b \cdot t$, тогда конечные разности первого порядка постоянны: $\Delta_2^1 = \Delta_3^1 = \ldots \Delta_n^1$, а разности второго порядка равны нулю.

Если общая тенденция выражается параболой второго порядка: $\hat{X} = a + bt + Ct^2$, то получим постоянными конечные разности второго порядка: $\Delta_3^2 = \Delta_4^2 = \ldots \Delta_n^2$, нулевыми – разности третьего порядка.

Порядок конечных разностей $j$, остающихся примерно равными друг другу, принимается за степень выравнивающего многочлена:

Если примерно постоянными оказываются темпы роста, то для выравнивания применяется показательная функция: $\hat{X} = \sum_{i=1}^{j} b_i \cdot t^i$.

При выборе формы уравнения следует исходить из объема имеющейся информации. Чем больше параметров содержит уравнение, тем больше должно быть наблюдений при одной и той же степени надежности оценивания.

Выбор формы кривой может осуществляться и на основе принятого критерия качества уравнения регрессии, в качестве которого может служить сумма квадратов отклонений фактических значений уровня ряда от значений уровней, рассчитанных по уравнению тренда. Из совокупности кривых выбирается та, которой соответствует минимальное значение критерия. Другим

статистическим критерием является коэффициент множественной детерминации $R^2$.

Интерпретация параметров линейного тренда:

$\hat{X} = a + b \cdot t$ ,

где $a$ – уровень ряда за период времени $t = 0$;

$b$ – средний абсолютный прирост уровня ряда за единичный промежуток времени.

Интерпретация параметров тренда, имеющего вид показательной функции:

$\hat{X} = a \cdot b^t$ ,

где $a$ – уровень ряда за период (в момент) времени $t = 0$;

$b$ – средний коэффициент роста за единичный промежуток времени.

*Расчет параметров уравнения тренда*. Расчет параметров при аналитическом выравнивании чаще всего производится с помощью метода наименьших квадратов (МНК). При этом поиск параметров для линейного уравнения тренда можно упростить, если отсчет времени производить так, чтобы сумма показателей времени изучаемого ряда динамики была равна нулю. То есть вводится новая условная переменная времени $t^y$, такая, что сумма значений этой переменной по всем элементам динамического ряда равна нулю: $\sum t^y = 0$.

При нечетном числе уровней ряда динамики для получения $\sum t^y = 0$ уровень, находящийся в середине ряда, принимается за условное начало отсчета времени (периоду или моменту времени, соответствующему данному уровню, присваивается нулевое значение). Даты времени, расположенные левее этого уровня, обозначаются натуральными числами со знаком минус (-1,-2, -3...), а даты времени, расположенные правее этого уровня – натуральными числами со знаком плюс (1, 2, 3 ...).

Если число уровней ряда четное, периоды времени левой половины ряда (до середины) нумеруются -1, -3, -5 и т. д. А периоды правой половины +1, +3, +5 и. т. д. При этом $\sum t^y = 0$.

При оценке параметров линейного тренда: $X_t = a + b \cdot t_t^y + u_t$ методом наименьших квадратов система нормальных уравнений преобразуется к виду:

$$\begin{cases} \sum_{t=1}^{n} X_t = a \cdot n; \\ \sum_{t=1}^{n} X_t \cdot t_t^y = b \sum_{t=1}^{n} \left(t_t^y\right)^2. \end{cases}$$

Тогда параметры линейного уравнения тренда рассчитываются по формулам:

$$b = \frac{\sum_{t=1}^{n} X_t \cdot t_t^y}{\sum_{t=1}^{n} \left(t_t^y\right)^2}, \quad a = \frac{\sum_{t=1}^{n} X_t{}^S}{n}.$$

Рассмотрим **пример.** Пусть имеются поквартальные данные за 3 года об объемах выпуска продукции некоторым предприятием в тыс. шт. Данные приведены в табл. 3.3 (строки 1,2,4).

Таблица 3.3

| Год | Квартал | $t$ – номер наблюдения | Объем выпуска($X_t$) | $t_t^y$ | $\hat{X}_t$ |
|---|---|---|---|---|---|
| 1999 | 1 | 1 | 477 | -11 | 416,1 |
| | 2 | 2 | 402 | -9 | 474,8 |
| | 3 | 3 | 552 | -7 | 533,4 |
| | 4 | 4 | 695 | -5 | 592,1 |
| 2000 | 1 | 5 | 652 | -3 | 650,8 |
| | 2 | 6 | 562 | -1 | 709,4 |
| | 3 | 7 | 812 | 1 | 768,1 |
| | 4 | 8 | 895 | 3 | 826,7 |
| 2001 | 1 | 9 | 832 | 5 | 885,4 |
| | 2 | 10 | 722 | 7 | 944,1 |
| | 3 | 11 | 1072 | 9 | 1002,7 |
| | 4 | 12 | 1195 | 11 | 1061,4 |

Требуется построить линейное уравнение тренда. В нашем примере четное число уровней ряда: $n = 12$. Следовательно, условная переменная времени $t^y$ для 6-го элемента ряда будет равна -1, а для 7-го +1. Внесем значения новой условной переменной времени в табл. 3.3 (графа 5).

Рассчитаем параметры линейного уравнения тренда приведенным выше формулам:

$$b = \frac{477 \cdot (-11) + 402 \cdot (-9) + \ldots + 1195 * 11}{(-11)^2 + (-9)^2 + \ldots + (11)^2} = \frac{16778}{572} = 29{,}3;$$

$$a = \frac{477 + 402 + \ldots + 1195}{12} = 739.$$

Дадим интерпретацию параметров тренда. С каждым кварталом объем выпуска товара в среднем увеличивается на 29,3 усл. ед. А средний за период с 1993 по 1995г. объем выпуска составил 739 усл. ед.

Рассчитаем значения трендовой компоненты по формуле $\hat{X}_t = 739 + 29{,}3 \cdot t_t^y$. Внесем значения $\hat{X}_t$ в табл. 3.3(графа 6).

### 3.3. Моделирование сезонных и циклических колебаний

Существуют несколько подходов при моделировании сезонных или циклических колебаний;

- расчет значений сезонной компоненты и построение аддитивной или мультипликативной модели временно го ряда;
- применение сезонных фиктивных переменных;
- использование рядов Фурье и др.

Рассмотрим 1-й из перечисленных подходов более подробно, т.к. он является наиболее простым. Причем будем моделировать только сезонные колебания, учитывая, что моделирование прочих циклических колебаний осуществляется аналогично.

#### *3.3.1. Расчет сезонной компоненты и построение модели временного ряда*

Аддитивную модель: $X=T+S+E$ применяют в случае, когда амплитуда сезонных колебаний со временем не меняется. В противном случае используют мультипликативную модель: $X=T \cdot S \cdot E$.

Введем обозначения.

Пусть имеется временной ряд – $X_{ij}$,

где $i$ – номер сезона (периода времени внутри года, например месяца, квартала); $i=1;L$ ($L$ – число сезонов в году);

$j$ – номер года, $j=1;m$ ($m$ – всего лет).

Тогда количество исходных уровней ряда равно $L \cdot m=n$.

Построение модели начинается с расчета сезонной компоненты. Только потом рассчитывают трендовую компоненту.

В качестве сезонной компоненты для аддитивной модели применяют абсолютное отклонение — $Sa_i$, для мультипликативной модели — индекс сезонности — $Is_i$. Сезонные компоненты должны отвечать определенным требованиям:

- в случае аддитивной модели сумма всех сезонных компонент должна быть равна нулю;
- в случае мультипликативной модели произведение всех сезонных компонент должно быть равно единице.

Перед расчетом сезонных компонент ряд динамики выравнивают. Чаще всего используют механическое выравнивание (например, метод скользящей средней). В результате получают выровненный ряд: $X_{ij}^{в}$, который не содержит сезонной компоненты.

Абсолютное отклонение в 1-м сезоне определяется как среднее арифметическое из отклонений фактического и выровненного уровней ряда:

$$Sa_i = \frac{1}{m}\sum_{j=1}^{m}\left(X_{ij} - X_{ij}^{B}\right)$$

Индекс сезонности в $i$-м сезоне определяется как среднее арифметическое из отношений фактического уровня ряда к выровненному:

$$Ia_i = \frac{1}{m}\sum_{j=1}^{m}\frac{X_{ij}}{X_{ij}^{B}}.$$

При построении трендовой компоненты модели временного ряда используют аналитическое выравнивание (см. п. 3.2). Данный метод выравнивания применяют не к фактическому ряду динамики, а к ряду, в котором исключена сезонная составляющая. Это означает, что исходные уровни ряда корректируются на величину сезонной компоненты. В случае аддитивной модели из исходных уровней вычитают $Sa_i$ . В случае мультипликативной модели исходные уровни ряда делят на $Is_i$ .

Рассмотрим на примере построение аддитивной модели временного ряда. Пусть имеются поквартальные данные за 3 года об объемах выпуска продукции некоторым предприятием (в тыс. шт.). Данные приведены в табл. 3.4 (строки 1, 2,3).

Таблица 3.4

| Год | Квартал – $i$ | Объем выпуска ($X_{tij}$) | $X_{ij}^{c}$ | $X_{ij}^{в}$ | $X_{ij} - X_{ij}^{в}$ | $X_{ij}^{S} =$ $= X_{ij} -$ $-Sa_i$ | T | $\hat{T} + Sa_j$ |
|---|---|---|---|---|---|---|---|---|
| 1999 | 1 | 410 | | | | 477,15 | 416,1 | 348,9 |
| (1) | 2 | 400 | | | | 401,60 | 474,8 | 473,2 |
| | 3 | 715 | 531,25 | 553,13 | 161,88 | 551,60 | 533Д | 696,8 |
| | 4 | 600 | 575,00 | 595,00 | 5,00 | 694,66 | 592,1 | 497,40 |
| 2000 | 1 | 585 | 615,00 | 647,50 | -62,50 | 652,15 | 650,8 | 583,6 |
| (2) | 2 | 560 | 680,00 | 705,00 | -145,00 | 561,60 | 709,4 | 707,8 |
| | 3 | 975 | 730,00 | 752,50 | 222,50 | 811,60 | 7684 | 931,5 |
| | 4 | 800 | 775,00 | 795,00 | 5,0 | 894,66 | 826,7 | 732,1 |
| 2001 | 1 | 765 | 815,00 | 847,50 | -82,50 | 832,15 | 885,4 | 818,3 |
| (3) | 2 | 720 | 880,00 | 917,50 | -197,50 | 721,60 | 944,1 | 942,5 |
| | 3 | 1235 | 955,00 | | | 1071,6 | 1002,7 | 1166,1 |
| | 4 | 1100 | | | | 1194,6 | 1061,4 | 966,7 |

В нашем примере $L = 4$; $m = 3$; $n = 12$.

Для расчета сезонной компоненты проведем выравнивание уровней ряда методом скользящей средней. Период усреднения примем равным 4. Рассчитанная по 4-м уровням средняя $X_{ij}^{C}$ будет относиться к середине интервала усреднения (см. табл. 3.4, строка 4). Чтобы полученные средние привести в соответствие с фактическими моментами времени, найдем средние значения из двух последовательных скользящих средних - центрированные скользящие средние – $X_{ij}^{B}$ (табл. 3.4, строка 5).

Для расчета абсолютных отклонений $Sa_i$ ($i$ = 1; $L$) найдем разности между исходными – $X_{ij}$ и выровненными – $X_{ij}^{B}$ уровнями ряда (табл. 3.4, строка 6). Для дальнейшего расчета $Sa_i$ построим отдельную таблицу. Строки данной таблицы соответствуют сезонным компонентам, столбцы – годам. В теле таблицы находятся значения: $X_{ij} - X_{ij}^{B}$. По этим данным рассчитываются средние арифметические из абсолютных отклонений по каждой строке – ($S_i^S$).

Если сумма всех средних оценок равна нулю $\sum_{i=1}^{L} S_i^S = 0$, то данные величины и будут окончательными значениями сезонных компонент ($Sa_i = S_i^C$). Если их сумма не равна нулю, то рассчитываются скорректированные значения сезонных компонент вычитанием из средней оценки величины, равной отношению суммы средних оценок сезонных компонент к их общему числу $Sa_i = S_i^c - \frac{\sum_{i=1}^{L} S_i^c}{L}$. Для нашего примера расчет значений $Sa_i$, представлен в табл. 3.5.

Таблица 3.5

| Номер компо-ненты | Год 1 | Год 2 | Год 3 | Средняя оценка сезонной компоненты $S_i^C$ | Скорректиро-ванная сезонная компонента $Sa_i$ |
|---|---|---|---|---|---|
| 1 | – | –66,67 | –70,00 | –68,33 | –67,15 |
| 2 | –1,67 | –5,00 | –1,67 | –2,78 | –1,60 |
| 3 | 123,33 | 180,00 | 183,33 | 162,22 | 163,40 |
| 4 | –78,33 | –113,33 | – | –95,83 | –94,66 |
| Итого | | | | –4,72 | 0 |

Для определения трендовой компоненты устраним сезонные колебания из уровней исходного ряда: $X_{ij}^{S} = X_{ij} - Sa_i$ . Результаты расчета $X_{ij}^{S}$ для нашего примера представлены в табл. 3.4, строка 7. Далее строим уравнение регрессии для уровней $X_{ij}^{S}$ – уравнение тренда: $\hat{T} = a + b \cdot t^{y}$ (где $t^{y}$ – условная переменная времени). Расчет параметров см. в п. 3.2. Окончательно имеем: $\hat{T} = 739 + 29{,}3 \cdot t^{y}$ . Рассчитанные по уравнению тренда уровни ряда $\hat{T}$ занесем в табл. 3.4 строка 8.

Теперь смоделируем уровни ряда в соответствии с аддитивной моделью, т. е. прибавим к $\hat{T}$ вычисленное ранее значение абсолютного отклонения – $Sa_i$. Результаты занесем в последнюю строку табл. 3.4.

Результаты моделирования представлены на рис. 3.1.

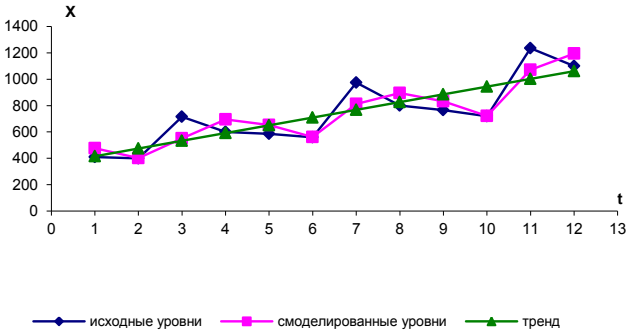


Рис. 3.1. Исходный ряд динамики и ряд, построенный по аддитивной модели

### *3.3.2. Использование сезонных фиктивных компонент при моделировании сезонных колебаний*

При этом подходе строится модель регрессии, включающая наряду с фактором времени фиктивные переменные. Количество фиктивных переменных в модели должно быть меньше на единицу числа сезонов внутри года, т. е. равно $L$–1 в наших обозначениях. Например, при моделировании поквартальных данных модель должна содержать три фиктивные компоненты наряду с фактором времени.

Каждому сезону соответствует определенное сочетание значений фиктивных переменных. Тот из сезонов, для которого значения всех фиктивных переменных равны нулю, принимается за эталон сравнения. Для остальных сезонов одна из фиктивных переменных должна принимать значение, равное единице. Например, если имеются поквартальные данные, то

значения фиктивных переменных ($Z_2$, $Z_3$, $Z_4$) для каждого квартала будут следующими:

| Квартал | $Z_2$ | $Z_3$ | $Z_4$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 |
| 4 | 0 | 0 | 1 |

Общий вид модели будет следующим:

$x_t = a + b \cdot t + C_2 \cdot Z_2 + C_3 \cdot Z_3 + C_4 \cdot Z_4 + u_t$.

Уравнение тренда для каждого квартала будут иметь вид:

Для 1-го квартала: $x_t = a + b \cdot t + u_t$.

Для 2-го квартала: $x_t = a + b \cdot t + c_2 + u_t$.

Для 3-го квартала: $x_t = a + b \cdot t + c_3 + u_t$.

Для 4-го квартала: $x_t = a + b \cdot t + c_4 + u_t$.

Данные уравнения отличаются только величиной свободного члена уравнения регрессии. Параметр $b$ в этой модели характеризует среднее абсолютное изменение уровней ряда под воздействием тенденции. Данная модель является аналогом аддитивной модели временного ряда.

Усреднив значения свободных членов в частных уравнениях регрессии, найдем их среднюю оценку – $\bar{a}$. Разность между средней оценкой и значением постоянного члена в уравнении регрессии для некоторого сезона дает оценку *сезонного отклонения*. Например, в случае поквартальных данных оценка сезонного отклонения может быть рассчитана следующим образом.

Для 1-го квартала: $a - \bar{a}$.

Для 2-го квартала: $a + C_2 - \bar{a}$.

Для 3-го квартала: $a + C_3 - \bar{a}$.

Для 4-го квартала: $a + C_4 - \bar{a}$,

где $\bar{a} = \dfrac{a + (a + c_2) + (a + c_3) + (a + c_4)}{4}$.

Заметим, что сумма сезонных отклонений в аддитивной модели должна равняться нулю.

### 3.4. Специфика изучения взаимосвязей по временным рядам. Исключение сезонных колебаний. Исключение тенденции

Предположим, изучается зависимость между рядами $X_t$ и $Y_t$.

При моделировании взаимосвязи двух или более временных рядов могут возникнуть следующие проблемы.

1. Искажение показателей тесноты и силы связи:

- если ряды содержат циклические или сезонные колебания одинаковой периодичности, то это приведет к завышению истинных показателей тесноты связи изучаемых временных рядов;
- если только один из рядов содержит циклические или сезонные колебания или периодичность колебаний различна, то это приведет к занижению истинных показателей тесноты связи изучаемых временных рядов.

2. Проблема «ложной корреляции»:

- если ряды имеют тренды одинаковой направленности, то между уровнями этих рядов всегда будет наблюдаться положительная корреляция, независимо от того, существует причинная связь между этими рядами или нет;
- если ряды имеют тренды разной направленности, то корреляция рядов окажется отрицательной.

Для того чтобы избавиться от данных проблем, необходимо устранить в уровнях ряда трендовую и сезонную (циклическую) компоненты.

Устранить сезонную компоненту в случае аддитивной модели можно, оценив абсолютные разности $Sa_i$ ($i$ = 1; $L$, где $L$ – число сезонов) и вычтя их из исходных уровней ряда. В случае мультипликативной модели необходимо оценить индексы сезонности — $Is_i$ ($i$ = 1; $L$, где $L$ – число сезонов) и разделить исходные уровни ряда на них. Как рассчитать сезонные компоненты, см. в п. 3.3.

Чтобы избавиться от «ложной корреляции» необходимо исключить тенденцию из уровней ряда. Предположим, что по двум временным рядам $X_t$ и $Y_t$ строится уравнение парной регрессии: $Y_t = a + b \cdot X_t + u_t$. Обнаружить «ложную корреляцию» можно, проанализировав остатки в данном уравнении регрессии. Если имеет место автокорреляция остатков, то, следовательно, имеет место и «ложная корреляция», и наоборот.

Для устранения тенденции обычно применяют следующие методы.

### *3.4.1. Метод отклонений от тренда*

При этом вычисляют значения $u_{xt}$, $u_{yt}$, представляющие собой отклонения уровней $X_t$ и $Y_t$ от их значений, рассчитанных по уравнениям трендов: $\hat{x}_t = f(t)$ и $\hat{y}_t = f(t)$: $u_{xt} = x_t - \hat{x}_t$, $u_{yt} = y_t - \hat{y}_t$. Затем измеряют корреляцию между этими отклонениями ($u_x$ и $u_y$), например, с помощью коэффициента корреляции:

$$r_{ux,uy} = \frac{\sum_{i=1}^{N}\left(u_{xt} \cdot u_{yt}\right)}{\sqrt{\sum_{i=1}^{N} u_{xt}^2 \cdot \sum_{i=1}^{N} u_{yt}^2}}.$$

Если предполагаемая функция регрессии линейная, то можно построить уравнение регрессии, измеряющее зависимость отклонения $u_x$ от отклонения $u_y = a + b \cdot u_x$. Параметры данного уравнения могут быть оценены с помощью МНК по формулам:

$$b = \frac{\sum_{i=1}^{N}\left(u_{xt} \cdot u_{yt}\right)}{\sqrt{\sum_{i=1}^{N} u_{xt}^2}}; \quad a = \overline{u_y} - b \cdot \overline{u_x} = 0.$$

То есть уравнение имеет вид: $u_y = b \cdot u_x$.

Содержательная интерпретация параметра этой модели затруднительна. Так, параметр $b$ показывает, насколько в среднем за период отклонилось значение $Y$ от тренда при отклонении $X$ от своего тренда на 1 единицу измерения.

Однако данное уравнение регрессии можно использовать; для прогнозирования. Для этого необходимо определить трендовое значение факторного признака $\hat{X}$ и оценить величину предполагаемого отклонения фактического значения $X$ от трендового. Далее определяют по уравнению тренда $Y$ значение $\hat{Y}$. По уравнению регрессии отклонений от трендов находят величину $u_{yt} = Y_t - \hat{Y}_t$. Затем находят точечный прогноз фактического значения $Y_{t+1}$ по формуле: $Y_{t+1} = \hat{Y}_t + u_{yt}$.

### *3.4.2. Метод последовательных разностей*

При этом вычисляют разности между текущим и предыдущим уровнями, т.е. величины абсолютных цепных приростов: $\Delta y_t = Y_t - Y_{t-1}$; $\Delta x_t = X_t - X_{t-1}$.

Тогда показатель тесноты связи – коэффициент линейной корреляции абсолютных приростов будет выглядеть так:

$$r_{\Delta x,\Delta y} = \frac{\sum_{i=1}^{N}\left(\Delta x_t \cdot \Delta y_t\right)}{\sqrt{\sum_{i=1}^{N} \Delta x_t^2 \cdot \sum_{i=1}^{N} \Delta y_t^2}}.$$

Уравнение регрессии по абсолютным приростам будет иметь вид:

$\Delta y_t = a + b \cdot \Delta x_t$.

В отличие от уравнения регрессии по отклонениям параметрам данного уравнения (по абсолютным разностям) легко дать интерпретацию. Параметр *b* показывает прирост *Y* в среднем при изменении прироста *X* на 1 единицу измерения. Параметр *a* характеризует прирост *Y* при нулевом приросте *X*.

Недостатком данного метода является сокращение числа пар наблюдений, т. е. потеря информации.

Разности первого порядка исключают автокорреляцию только в тех рядах динамики, в которых основной тенденцией является прямая линия.

Для рядов, с основной тенденцией близкой к экспоненте, следует рекомендовать исследовать корреляцию цепных коэффициентов (темпов) роста.

Для рядов, с основной тенденцией, близкой к параболе 2-го порядка, следует рекомендовать исследовать корреляцию конечных разностей второго порядка: $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1}$; $\Delta^2 x_t = \Delta x_t - \Delta x_{t-1}$.

Если ряды динамики имеют разные типы тенденций, вполне допустимо коррелировать соответствующие разные цепные показатели: например, абсолютные изменения в одном ряду с темпами изменений в другом.

### *3.4.3. Включение в модель регрессии фактора времени*

При этом в уравнение регрессии включают переменную времени в качестве дополнительного фактора: $Y = f(X, t)$. Например, $Y_t = a + b \cdot X_t + t + u_t$. Преимущество данного метода в том, что модель регрессии, построенная таким образом, позволяет учесть всю информацию. Кроме того, интерпретация параметров не вызывает затруднений.

## 3.5. Динамические эконометрические модели (ДЭМ). Общая характеристика. Модели авторегрессии. Интерпретация параметров

К динамическим эконометрическим моделям (ДЭМ) относят модели, которые в данный момент времени учитывают значения входящих в нее переменных, относящихся к текущему и к предыдущему моментам времени. Например, $y_t = f(x_t, x_{t-1})$; $y_t = f(x_t, y_{t-1})$ – ДЭМ, а $y_t = f(x_t)$ – не ДЭМ.

Выделяют 2 типа динамических эконометрических моделей ДЭМ:

1. Модели, в которых лаговые значения переменных (переменных, относящихся к предыдущим моментам времени) непосредственно включены в модель. Это модели авторегрессии и модели с распределенным лагом.

   Модели авторегрессии – это ДЭМ, в которых в качестве факторных переменных содержатся лаговые значения результативной переменной. Например, $y_t = a + b_0 \cdot x_t + c_1 \cdot y_{t-1} + u_t$.

Модели с распределенным лагом – это ДЭМ, в которых содержатся не только текущие, но и лаговые значения факторных переменных. Например, $y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + \ldots + b_L \cdot x_{t-L} + u_t$.

2. Модели, в которых включены переменные, характеризующие ожидаемый или желаемый уровень признака-результата или одного из факторов в момент времени t. Этот уровень считается неизвестным и определяется с учетом информации, которой располагают в предыдущий момент времени t – 1.

Ожидаемые значения показателей определяют различными способами. В зависимости от способа различают модели:

- неполной корректировки;
- адаптивных ожиданий;
- рациональных ожиданий.

Оценка параметров этих моделей сводится к оценке параметров моделей авторегрессии.

*Модели авторегрессии.* Модели авторегрессии – это ДЭМ, в которых в качестве факторных переменных содержатся лаговые значения результативной переменной. Например, $y_t = a + b_0 \cdot x_t + c_1 \cdot y_{t-1} + u_t$.

Коэффициент регрессии $b_0$ в данной модели характеризует краткосрочное изменение $y$ под влиянием изменения $x$ на единицу своего измерения.

Коэффициент $c_1$ характеризует изменение $y$ в момент $t$ под воздействием своего изменения в предшествующий момент времени $(t-1)$.

Произведение коэффициентов $(b_0 \cdot c_1)$ – называют промежуточным мультипликатором. Данный показатель характеризует общее абсолютное изменение результата $y$ в момент $(t+1)$.

Показатель $b = b_0 + b_0 \cdot c_1 + b_0 \cdot c_1^2 + b_0 \cdot c_1^3 + \ldots$ – называют долгосрочным мультипликатором. Он характеризует общее абсолютное изменение $y$ в долгосрочном периоде.

Практически во все модели авторегрессии вводят условие стабильности, состоящее в том, что $|c_1| < 1$. Тогда при наличии бесконечного лага:

$$b = b_0\left(1 + c_1 + c_1^2 + c_1^3 + \ldots\right) = \frac{b_0}{1 - c_1}.$$

Применение МНК к моделям авторегрессии неприемлемо, т. к. нарушается 1-я предпосылка нормальной линейной модели регрессии, а именно, одна из объясняющих переменных ($y_{t-1}$) частично зависит от случайной составляющей ($u_t$). Это приводит к получению смещенной оценки параметра при переменной $y_{t-1}$.

## 3.6. Регрессионные модели с распределенными лагами

### *3.6.1. Модели с распределенным лагом. Интерпретация параметров. Средний и медианный лаги. Изучение структуры лагов*

Модели с распределенным лагом — это ДЭМ, в кс содержатся не только текущие, но и лаговые значения факторных переменных. Например, $y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + \ldots + b_L \cdot x_{t-L} + u_t$.

Данная модель позволяет определить влияние изменения независимой переменной $x$ на результат $y$. Если в момент времени $t$ происходит изменение независимой переменной $x$, то это изменение будет влиять на значение зависимой переменной $y$ в течение $L$ следующих моментов времени. Коэффициент регрессии $b_0$ называют краткосрочным мультипликатором, он характеризует среднее абсолютное изменение $y_t$ при изменении $x_t$ на единицу своего измерения в некоторый фиксированный момент $t$ без учета воздействия лаговых значений фактора $x$.

Коэффициент регрессии $b_1$ характеризует среднее абсолютное изменение $y_t$ вследствие изменения фактора $x$ на единицу своего измерения в момент времени $(t - 1)$.

Сумму коэффициентов $b_0+b_1$ называют промежуточным мультипликатором. Она характеризует совокупное воздействие фактора $x$ на результат $y$ в момент $(t + 1)$, т.е. изменение $x$ на единицу в момент $t$ влечет изменение $y$ на $b_0$ единиц в момент $t$ и изменение $y$ на $b_1$ в момент $(t + 1)$.

Сумма коэффициентов $b_0 + b_1 + b_2$ – тоже промежуточный мультипликатор, характеризующий совокупное воздействие фактора $x$ на результат $y$ в момент$(t + 2)$ и т.д.

Сумма $b = b_0 + b_1 + b_2 + \ldots + b_L$ — это долгосрочный мультипликатор, характеризующий общее изменение результата $y$ в момент $(t + L)$ под влиянием изменения $x$ на единицу своего измерения в момент $t$.

Введем новые показатели.

1. Весовые коэффициенты: $\beta_j = \frac{b_j}{b}$, $j = 0; L$. Если все коэффициенты регрессии $b_j$ одного знака, то $\sum_{j=0}^{L} \beta_j = 1$.

2. Средний лаг – $\overline{L} = \sum_{j=0}^{L} j \cdot \beta_j$ – это средний период, в течение которого будет происходить изменение результата под воздействием изменения фактора $x$ в момент $t$. Если значение среднего лага небольшое, то это говорит о довольно быстром реагировании $y$ на изменение $x$. Если значение среднего лага

большое, это говорит о медленном воздействии фактора на результат.

3. Медианный лаг $L_{Me}$ – это величина лага, для которого $\sum_{j=0}^{L_{Me}} \beta_j = 0{,}5$.

   То есть это период времени, в течение которого с текущего момента $t$ будет реализована половина общего воздействия фактора на результат.

*Применение МНК к моделям с распределенным лагом в большинстве случаев неприемлемо, т. к.:*

- текущие и лаговые значения независимой переменной, как правило, тесно связаны друг с другом. Тем самым, нарушается 1-я предпосылка нормальной линейной модели регрессии, возникает проблема мультиколлинеарности;
- при большой величине лага (L) снижается число наблюдений, по которому строится модель и увеличивается число факторных признаков (xt, xt-1, xt-2,...). Это ведет к потере числа степеней свободы в модели;
- в таких моделях часто возникает проблема автокорреляции остатков.

Следствием этого является нестабильность оценок параметров модели, т. е. при изменении спецификации модели коэффициенты существенно меняются, снижается их точность и эффективность. Чистое влияние фактора на результат в таких условиях выявить невозможно.

Поэтому на практике оценку параметров таких ДЭМ проводят с помощью специальных методов (метод Алмон, метод Койка).

Если для модели с распределенным лагом $y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + \ldots + b_L \cdot x_{t-L} + u_t$ построить график зависимости коэффициентов при факторных переменных $b_j$ от величины лага $j$, то можно получить графическое изображение структуры лага. Структура лага может быть различной (см. рис. 3.2).

Основная трудность в выявлении структуры лага состоит в том, как получить значения параметров $b_j$ . Обычно МНК редко бывает полезным. Поэтому в большинстве случаев предположения о структуре лага основаны на общих положениях экономической теории, на исследованиях взаимосвязи показателей, либо на результатах проведенных ранее эмпирических исследований или иной априорной информации.

Лаги Алмон (метод Алмон) используют для описания модели с распределенным лагом, имеющей полиномиальную структуру Лага и конечную величину лага ($L$). Примером лагов, образующих полином 2-й степени, является рис. 3.2,*в*. Частным случаем полиноминальной структуры лага является линейная модель (рис. 3.2,*а*).

Метод Койка обычно применяется, если величина лага — $L$ бесконечно большая, а также в предположении геометрической структуры лага (рис. 3.2,*б*).

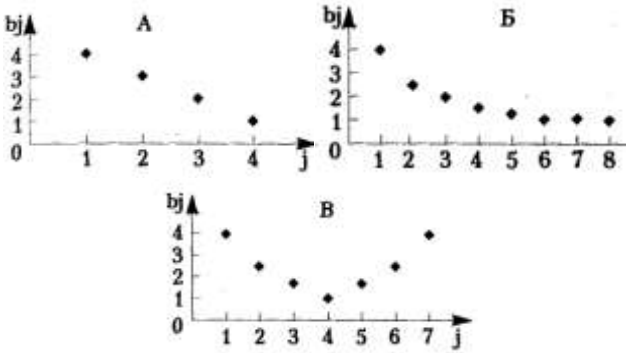


Рис. 3.2. Основные формы структуры лага:
*а* – линейная; *б* – геометрическая; *в* — полиномиальная

### *3.6.2. Оценивание параметров модели с распределенным лагом. Метод Алмон*

Лаги Алмон (метод Алмон) используют для описания модели с распределенным лагом:

$$y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + \ldots + b_L \cdot x_{t-L} + \hat{u}_t \qquad (3.1)$$

имеющей полиноминальную структуру лага и конечную величину лага $L$.

Ниже мы опишем суть метода Алмон.

1. Формализуют зависимость коэффициентов $b_j$ от величины лага $j$. Модель зависимости представляет собой полином:

- либо 1-й степени: $b_j = C_0 + C_1 \cdot j$;
- либо 2-й степени: $b_j = C_0 + C_1 \cdot j + C_2 \cdot j^2$;
- либо 3-й степени: $b_j = C_0 + C_1 \cdot j + C_2 \cdot j^2 + C_3 \cdot j^3$;
- • либо $K$-й степени (общий случай): $b_j = C_0 + C_1 \cdot j + C_2 \cdot j^2 + \ldots + C_K \cdot j^{K3}$.

2. Тогда каждый коэффициент модели (3.1) – $b_i$ ($j$= 0; $L$) можно выразить следующим образом:

$b_0 = C_0$ ;

$b_1 = C_0 + C_1 + \ldots + C_K$ ;

$b_2 = C_0 + 2C_1 + 4C_2 + \ldots + 2^K C_K$ ;

$b_3 = C_0 + 3C_1 + 9C_2 + \ldots + 3^K \cdot C_K$ и т.д.

$b_L = C_0 + LC_1 + L^2 \cdot C_2 + \ldots + L^K \cdot C_K$ .

Подставим найденные соотношения для $b_i$ в модель (3.1) и получим:

$y_t = a + b_0 \cdot x_t + (C_0 + C_1 + \ldots + C_K) \cdot x_{t-1} +$

$$+\left(C_0+2C_1+4C_2+\ldots+2^K C_K\right)\cdot x_{t-2}+\ldots+$$
$$+\left(C_0+LC_1+L^2\cdot C_2+\ldots+L^K\cdot C_K\right)\cdot x_{t-L}+\hat{u}_t$$

3. Перегруппируем слагаемые:

$$y_t = a+C_0\cdot\left(x_t+x_{t-1}+x_{t-2}+\ldots+x_{t-L}\right)\cdot+$$
$$+C_1\cdot\left(x_{t-1}+2x_{t-2}+3x_{t-3}+\ldots+L\cdot x_{t-L}\right)+$$
$$+C_2\cdot\left(x_{t-1}+4x_{t-2}+9x_{t-3}+\ldots+L^2\cdot x_{t-L}\right)+$$
$$+C_K\cdot\left(x_{t-1}+2^K x_{t-2}+3^K x_{t-3}+\ldots+L^K\cdot x_{t-L}\right)+u_t.$$

Обозначим слагаемые в скобках при коэффициентах $C_i$ *(i=0; K)* как новые переменные:

$$z_0 = x_t+x_{t-1}+x_{t-2}+\ldots+x_{t-L}=\sum_{j=0}^{L}x_{t-j};$$

$$z_1 = x_{t-1}+2x_{t-2}+3x_{t-3}+\ldots+L\cdot x_{t-L}=\sum_{j=0}^{L}j\cdot x_{t-j};$$

$$z_2 = x_{t-1}+4x_{t-2}+9x_{t-3}+\ldots+L^2\cdot x_{t-L}=\sum_{j=0}^{L}j^2\cdot x_{t-j};$$

…

$$z_K = x_{t-1}+2^K x_{t-2}+3^K x_{t-3}+\ldots+L^K\cdot x_{t-L}=\sum_{j=0}^{L}j^K\cdot x_{t-j}$$

Тогда модель примет вид:

$$y_t = a+C_0\cdot z_0+C_1\cdot z_1+\ldots+C_K z_K+u_t. \qquad (3.2)$$

4. Определим параметры новой модели (3.2) с помощью обычного МНК. Затем от параметров $C_i$ *(i* = 0; *K)* перейдем к параметрам $b_j$ (*j* = 0; *L*), используя соотношения, полученные на 1-м шаге.

Применение метода Алмон сопряжено с рядом проблем.

Величина максимального лага *L* должна быть известна заранее. Лучший способ определения величины *L* – использование показателей тесноты связи (например, линейных парных коэффициентов корреляции) между результатом *y* и лаговым значением фактора *x*: $r_{y,xt-1}$; $r_{y,xt-2}$. Если показатель тесноты связи значим (значимо отличается от нуля), то его следует включать в модель с распределенным лагом. Порядок максимального значимого показателя тесноты связи принимается в качестве *L*.

Необходимо каким-то образом определить степень полинома *K* (обычно ограничиваются *K* = 2, 3). Выбранная степень должна быть на единицу меньше числа экстремумов в структуре лага.

Переменные $z_i$ $(i = 0; K)$, которые определяются как линейная комбинация исходных факторов $x$, будут коррелировать между собой в случаях, когда наблюдается высокая связь между самими факторными признаками. Поэтому проблема мультиколлинеарности остается актуальной и в модели (3.2). Однако мультиколлинеарность новых переменных $z_i$ сказывается на оценках параметров $b_j$ $(j = 0; L)$ в меньшей степени, чем в случае применения обычного МНК к модели (3.1).

*Преимущества метода Алмон.* Он достаточно универсален и может быть применен для моделирования процессов, которые характеризуются разнообразными структурами лагов.

При относительно небольшом числе переменных в модели (3.2) $(K = 2,3)$, не приводящем к значительной потере числа степеней свободы, с помощью метода Алмон можно построить модели с распределенным лагом (3.1) любой длины (максимальный лаг $L$ может быть достаточно большим).

### *3.6.3. Оценивание параметров моделей с геометрической структурой лага. Метод Койка*

Метод Койка обычно применяют, если в моделях с распределенным лагом $y_t = a + b_0 \cdot x_t + b_1 \cdot x_{t-1} + \ldots + b_L \cdot x_{t-L} + \hat{u}_t$ величина максимального лага $L$ бесконечна. При этом используют допущение о геометрической структуре лага, т.е. воздействие лаговых значений фактора на результат уменьшается с увеличением величины лага в геометрической прогрессии. Если имеется одна объясняющая переменная, то модель имеет вид:

$$y_t = a + b_0 \cdot x_t + b_0 \cdot \lambda \cdot x_{t-1} + b_0 \cdot \lambda^2 \cdot x_{t-2} + b_0 \cdot \lambda^3 \cdot x_{t-3} + u_t \qquad (3.3)$$

где $b_j = b_0 \cdot \lambda^j \left(j = 0; \infty\right), \lambda \in \left[-1; +1\right]$

В данной зависимости всего три параметра: $a$, $b_0$ и $\lambda$. Для их оценивания нельзя применять обычный МНК, так как:

1) возникает проблема мультиколлинеарности;
2) из полученных МНК-оценок не удалось бы вывести значения $b_0$ и $\lambda$. Здесь можно получить одно значение оценки $b_0$ с помощью коэффициента при $x_t$, и совершенно другое, возведя в квадрат коэффициент при $x_{t-1}$ и разделив его на коэффициент при $x_{t-2}$.

Для оценки параметров $a$, $b_0$ и $\lambda$ можно использовать два метода: нелинейный МНК либо преобразование Койка.

*Суть нелинейного МНК.* Задаем для $\lambda$ значения в пределах от 0 до 1 с шагом, например, 0,01 (чем меньше шаг, тем более точным будет результат).

Для каждого значения $\lambda$ рассчитывается переменная:

$$z_t = x_t + \lambda \cdot x_{t-1} + \lambda^2 \cdot x_{t-2} + \lambda^3 \cdot x_{t-3} + \ldots + \lambda^L \cdot x_{t-L}$$

с таким значением $L$, при котором дальнейшие лаговые значения $x$ не оказывают существенного воздействия на $z$.

С помощью обычного МНК оценивается уравнение регрессии:

$$y_t = a + b_0 \cdot z_t + u_t \qquad (3.4)$$

и определяется теоретический коэффициент детерминации $R^2$.

Такие расчеты проделывают для всех $\lambda \in [-1;+1]$.

В качестве окончательных оценок *a, $b_0$* и λ выбирают те, которые обеспечивают наибольшее значение $R^2$ для уравнения (3.4).

*Суть метода Койка (преобразование Койка).* Если выражение (3.3) выполняется для периода *t*, то оно должно выполняться и для периода (*t* – 1):

$$y_{t-1} = a + b_0 \cdot x_{t-1} + b_0 \cdot \lambda \cdot x_{t-2} + b_0 \cdot \lambda^2 \cdot x_{t-3} + b_0 \cdot \lambda^3 \cdot x_{t-4} + \ldots + u_{t-1}$$

Умножив обе части этого уравнения на λ и вычтя их из уравнения (3.3), получим:

$$y_t - \lambda \cdot y_{t-1} = a(1-\lambda) + b_0 \cdot x_t + u_t - \lambda \cdot u_{t-1} \text{ или}$$

$$y_t = a(1-\lambda) + b_0 \cdot x_t + \lambda \cdot y_{t-1} + u_t - \lambda \cdot u_{t-1}.$$

Полученная модель относится к моделям авторегрессии.

Эта форма позволяет анализировать краткосрочные и долгосрочные динамические свойства модели.

В краткосрочном аспекте (в текущем периоде) значение $y_{t-1}$ нужно рассматривать как фиксированное. Воздействие *x* на *y* характеризует коэффициент $b_0$.

В долгосрочном периоде (без учета случайной составляющей) если $x_t$ стремится к некоторому равновесному значению $\bar{x}$, то $y_t$ и $y_{t-1}$, будут также стремиться к равновесному уровню $\bar{y}$, определяемому как:

$$\bar{y} = a(1-\lambda) + b_0 \bar{x} + \lambda \bar{y},$$

из которого следует: $\bar{y} = a + \frac{b_0}{1-\lambda} \cdot \bar{x}$.

Таким образом, долгосрочное воздействие *x* на *y* отражается коэффициентом $\frac{b_0}{1-\lambda}$. Если $\lambda \in [-1;+1]$, то этот коэффициент превысит $b_0$, т.е. долгосрочное воздействие оказывается сильнее краткосрочного.

Модель преобразования Койка привлекательна с практической точки зрения, т.к. оценивание парной регрессии с помощью МНК позволяет получить оценки *a, $b_0$* и λ. Метод Койка требует гораздо меньших усилий при оценивании параметров, чем нелинейный МНК. Однако применение данного метода сопряжено с серьезной эконометрической проблемой – нарушением 1-го условия нормальной линей ной модели регрессии: объясняющая переменная $y_{t-1}$ частично зависит от $u_{t-1}$ и поэтому коррелирует с одной из случайных

составляющих $\left(-\lambda \cdot u_{t-1}\right)$. В итоге оценки, полученные с помощью МНК, оказываются смещенными и несостоятельными.

### 3.7. Оценивание параметров моделей авторегрессии. Метод инструментальных переменных

При построении моделей авторегрессии:

$$y_t = a + b_0 \cdot x_t + c_1 \cdot y_{t-1} + u_t \qquad (3.5)$$

возникает проблема: нарушается 1-я предпосылка нормальной линейной модели регрессии об отсутствии связи между факторным признаком и случайной составляющей. В модели авторегрессии факторный признак $y_{t-1}$ связан со случайной составляющей $u_{t-1}$. Поэтому применение обычного МНК для оценки параметров уравнения регрессии приводит к получению смещенной оценки параметра при переменной $y_{t-1}$.

Для оценивания параметров уравнения регрессии может быть использован метод инструментальных переменных.

Суть метода инструментальных переменных состоит в следующем.

Переменную $y_{t-1}$ из правой части уравнения, для которой нарушается предпосылка МНК, заменяют на новую переменную, удовлетворяющую следующим требованиям:

1) она должна тесно коррелировать с $y_{t-1}$ ;
2) она не должна коррелировать со случайной составляющей $u_t$.

Затем оценивают регрессию с новой инструментальной переменной с помощью обычного МНК.

Рассмотрим один из методов получения инструментальной переменной.

Так как $y_t$ зависит от $x_t$, предположим, что имеет место зависимость $y_{t-1}$ от $x_{t-1}$, т. е. $y_{t-1} = d_0 + d_1 \cdot x_{t-1} + v_t = \hat{y}_{t-1} + v_t$ .

Оценка $\hat{y}_{t-1}$ может быть найдена с помощью обычного МНК.

Новая переменная $\hat{y}_{t-1}$ тесно коррелирует с $y_{t-1}$ и не коррелирует со случайной составляющей $u_t$, т. е. может служить инструментальной переменной для фактора $y_{t-1}$ .

В результате модель авторегрессии примет вид:

$$y_t = a + b_0 \cdot x_t + c_1 \cdot \hat{y}_{t-1} + \hat{u}_t \qquad (3.6)$$

$\left(\hat{u}_t = c_1 \cdot v_t + u_t\right)$.

Оценки параметров данной модели находят обычным МНК. Полученные оценки являются искомыми оценками модели авторегрёссии (3.5).

Отметим, что практическая реализация метода инструментальных переменных осложняется появлением проблемы мультиколлинеарности факторов в модели: функциональная связь между $\hat{y}_{t-1}$ и $x_{t-1}$ ( $\hat{y}_{t-1} = f\left(x_{t-1}\right)$)

приводит к появлению высокой корреляционной связи между $\hat{y}_{t-1}$ и $x_t$ В некоторых случаях эту проблему можно решить включением в модель (3.6) фактора времени *t*.

### 3.8. Модель адаптивных ожиданий

Если в моделях учитывают не фактическое значение переменной, а ее желаемое (ожидаемое) значение, то такие модели относят ко 2-му типу динамических эконометрических моделей – моделям адаптивных ожиданий либо к моделям частичной (неполной) корректировки.

Модель адаптивных ожиданий (МАО) учитывает желаемое (ожидаемое) значение *факторного* признака $x^*_{t+1}$. Пример МАО: ожидаемое в будущем (в период

*(t* + 1)) значение курса доллара $x^*_{t+1}$ влияет на наши инвестиции в текущем периоде $y_t$. Или другой пример: ожидаемое значение заработной платы $x^*_{t+1}$ влияет на уровень безработицы в текущем периоде $y_t$.

В общем виде модель адаптивных ожиданий можно за писать так:

$$y_t = a + b_0 \cdot x^*_{t+1} + u_t \,. \qquad (3.7)$$

Желаемое (ожидаемое) значение переменных определяется по значению реальных (фактических) переменных в предыдущий период времени *(t).*

Механизм формирования ожиданий в модели адаптивных ожиданий следующий:

$$x^*_{t+1} - x^*_t = \lambda \cdot \left(x_t - x^*_t\right) (0 \le \lambda \le 1)$$

$$\text{или } x^*_{t+1} = \lambda \cdot x_t + (1-\lambda) \cdot x^*_t \,. \qquad (3.8)$$

То есть значение переменной, ожидаемое в следующий период $x^*_{t+1}$, формирующееся как среднее арифметическое взвешенное ее реального и ожидаемого значений в текущем периоде. Чем больше величина $\lambda$, тем быстрее ожидаемое значение адаптируется предыдущим реальным значениям. Чем меньше $\lambda$, тем ожидаемое значение в будущем $x^*_{t+1}$ ближе к ожидаемому значению предыдущего периода $x^*_t$ (т. е. тенденции в ожиданиях сохраняются).

Для того чтобы оценить параметры данной модели (3.7), обычный МНК применить невозможно, т. к. модель включает ожидаемые значения факторной переменной, которые нельзя получить эмпирическим путем. Поэтому для оценки параметров исходную модель преобразуют.

Подставим в модель (3.7) вместо $x^*_{t+1}$ соотношение (3.8):

$$y_t = a + b_0 \cdot \left(\lambda \cdot x_t + (1-\lambda) \cdot x^*_t\right) + u_t =$$

$$a + \lambda \cdot b_0 \cdot x_t + (1-\lambda) b_0 x_t^* + u_t . \qquad (3.9)$$

Если модель (3.7) имеет место для периода *t,* то она будет иметь место и для периода *(t* – 1). Таким образом, в период *(t*–- 1) получим:

$$y_{t-1} = a + b_0 \cdot x_t^* + u_{t-1} .$$

Умножим это выражение на (1 – *λ)* и получим:

$$(1-\lambda) \cdot y_{t-1} = (1-\lambda) \cdot a + (1-\lambda) \cdot b_0 \cdot x_t^* + (1-\lambda) \cdot u_{t-1}$$

Вычтем почленно полученное выражение из (3.9):

$y_t - (1-\lambda) \cdot y_{t-1} = a - (1-\lambda) \cdot a + \lambda \cdot b_0 \cdot x_t + u_t - (1-\lambda) \cdot u_{t-1}$ или

$y_t = \lambda \cdot a + \lambda \cdot b_0 \cdot x_t + (1-\lambda) \cdot y_{t-1} + u_t^*$ ,

где $u_t^* = u_t - (1-\lambda) \cdot u_{t-1}$.

Мы получили модель авторегрессии, определив параметры которой, можно легко перейти к параметрам исходной модели (3.7).

Полученная модель включает только фактические значения переменных, поэтому ее параметры можно определять с помощью стандартных статистических методов.

Исходная модель (3.7), характеризующая зависимость результативного признака от ожидаемых значений факторного признака, называется долгосрочной функцией МАО.

Преобразованная модель, которая описывает зависимость результативного признака от фактических значений факторного признака, называется краткосрочной функцией МАО.

### 3.9. Модель частичной (неполной) корректировки

Если в моделях учитывают не фактическое значение переменной, а ее желаемое (ожидаемое) значение, то такие модели относят ко 2-му типу динамических эконометрических моделей – моделям ожиданий, либо к моделям частичной (неполной) корректировки.

Модель частичной корректировки (МЧК) учитывает желаемое (ожидаемое) значение *результативного* признака $y_t^*$. Примером МЧК может служить модель Литнера: фактический объем прибыли $x_t$ оказывает влияние на величину желаемого объема дивидендов $y_t^*$. В общем виде модель частичной корректировки можно записать так:

$$y_t^* = a + b_0 \cdot x_t + u_t . \qquad (3.10)$$

Желаемое (ожидаемое) значение переменных определяется по значению реальных (фактических) переменных в предыдущий период времени *t* - 1.

В таких моделях предполагается, что фактическое при ращение зависимой переменной $y_t - y_{t-1}$ пропорционально разнице между ее желаемым уровнем и фактическим значением в предыдущий период $y_t^* - y_{t-1}$:

$$y_t - y_{t-1} = \lambda\left(y_t^* - y_{t-1}\right) + v_t \left(0 \le \lambda \le 1\right)$$

или $y_t = \lambda \cdot y_t^* + (1-\lambda)\cdot y_{t-1} + v_t$. (3.11)

Из этого следует, что $y_t$ получается как среднее арифметическое взвешенное желаемого уровня $y_t^*$ и фактического значения этой переменной в предыдущем периоде $y_{t-1}$. Чем больше величина $\lambda$, тем быстрее происходит процесс корректировки. Если значение $\lambda$. = 1, то $y_t = y_t^*$ и полная корректировка происходит за 1 период. Если $\lambda$ = 0, то корректировка $y_t$ не происходит совсем.

Подставив (3.10) в (3.11) получим:

$$y_t = a \cdot \lambda + b_0 \cdot \lambda \cdot x_t + (1-\lambda)\cdot y_{t-1} + v_t + \lambda \cdot u_t =$$

$$= a \cdot \lambda + b_0 \cdot \lambda \cdot x_t + (1-\lambda)\cdot y_{t-1} + \varepsilon_t . \left(\varepsilon_t = v_t + \lambda \cdot u_t\right).$$

Параметры преобразованного уравнения регрессии $a$, $\lambda$, $b_0$ могут быть оценены с помощью обычного МНК. Данная модель, как и в методе Койка (см. п. 3.6.3) включает стохастическую объясняющую переменную $y_{t-1}$. Но теперь эта переменная не коррелирует с текущим значением совокупного случайного члена ( $\varepsilon_t$ ) уравнения (поскольку $v_t$, так же как и $u_t$, рассчитывается после того, как определилось значение $y_{t-1}$). При таких условиях обычный МНК позволяет получать асимптотически несмещенные и эффективные оценки (исключение составляют малые выборки).

Соотношение: $y_t = a \cdot \lambda + b_0 \cdot \lambda \cdot x_t + (1-\lambda)\cdot y_{t-1} + \varepsilon_t$ включает только фактические значения переменных, его еще называют краткосрочной функцией МЧК.

Исходное уравнение (3.10) называют долгосрочной функцией МЧК.

## Тесты по разделу

**1. На больших временах _____ факторы описываются монотонной функцией**

a) долговременные
b) сезонные
c) случайные
d) циклические

**2. Если математическое ожидание и дисперсия случайной величины временного ряда *x(t)* не зависят от времени, то такой ряд будет**

a) стационарным в узком смысле
b) квазистационарным
c) стационарным в обоих смыслах
d) стационарным в широком смысле

**3. В методе выделения неслучайной составляющей (МНК) необходимо, чтобы величина _____ была минимальной**

a) $Mx(t)$
b) $\sum_{t=1}^{n}[x(t)-f(t)]^2$
c) $\sum_{t=1}^{n}x(t)^2$
d) $\sum_{t=1}^{n}f(t)^2$

**4. Когда делается предсказание на момент времени *T* + *p*, предполагается, что известна величина**

a) $y(T)$
b) $x(T)$
c) $x(T+p)$
d) $y(T+p)$

**5. В критерии серий, основанном на медиане, общее число серий временного ряда 1, 3, 5, 4, 2 равно**

a) 3
b) 2
c) 5
d) 4

**6. В критерии серий, основанном на медиане, проверяется гипотеза**

a) $Mx(t)=0$
b) $Mx(t)=const$
c) $x(t)=const$
d) $\frac{x(t-1)+x(t+1)}{2}=0$

**7. В критерии серий, основанном на медиане, временному ряду 2, 5, 4, 6, 3 соответствует последовательность**

a) – + + –
b) – + – + –
c) + – + –
d) – + + + –

**8. Пусть имеется матрица исходных статистических данных** $(и.с.д.)=\begin{pmatrix} x_1^1(t)x_1^2(t)\ldots x_1^m(t) \\ x_2^1(t)x_2^2(t)\ldots x_2^m(t) \\ \ldots\ldots\ldots\ldots \\ x_n^1(t)x_n^2(t)\ldots x_n^m(t) \end{pmatrix}$ **Одномерным временным рядом будет ряд значений _____ матрицы и.сд. в последовательные моменты времени**

a) одного из столбцов
b) одной из строк
c) всей
d) одного из элементов

**9. Если элементы набора данных не являются одинаково распределенными, то речь идет о**

a) случайной выборке
b) стационарном временном ряде
c) временном ряде
d) генеральной совокупности

**10. Процесс АР(2) имеет автокорреляционную функцию, которая**

a) обращается в ноль после некоторой точки
b) имеет бесконечную протяженность
c) имеет максимум в точке $\tau = 2$
d) не меняется после $\tau = 2$

**11. Целевая переменная в модели частичного приспособления имеет вид**

a) $y^*(t) = \tilde{\beta}_0 + \tilde{\beta}_1 x(t) + \tilde{\delta}(t)$
b) $y^*(t) = \tilde{\beta}_0 x(t) + \tilde{\beta}_1 y(t) + \tilde{\delta}(t)$
c) $y^*(t) = \tilde{\beta}_0 + \tilde{\beta}_1 y(t) + \tilde{\delta}(t)$
d) $y^*(t) = \tilde{\beta}_0 x(t) + \tilde{\beta}_1 y(t-1) + \tilde{\delta}(t)$

**12. Если в методе последовательных разностей** $\hat{\sigma}^2(2) \succ \hat{\sigma}^2(5)$, **а** $\hat{\sigma}^2(8) \approx \hat{\sigma}^2(5)$, **то неслучайная составляющая аппроксимируется полиномом степени**

a) $4 \le p \le 7$

b) $2 \le p \le 4$

c) $1 \le p \le 7$

d) $p \ge 1$

**13. Если общий линейный процесс описывается классической линейной моделью множественной регрессии, то он имеет вид $\varepsilon(t)$=**

a) $\sum_{k=1}^{\infty} \lambda_k \delta(t-k) \varepsilon(t-k)$

b) $\delta(t) + \sum_{k=1}^{\infty} \beta_k \delta(t-k)$

c) $\delta(t) + \sum_{k=1}^{\infty} \lambda_k \varepsilon(t-k)$

d) $\sum_{k=1}^{\infty} \lambda_k \varepsilon(t-k)$

**14. Если неслучайная составляющая временного ряда x(t) имеет линейный вид $f(t) = \beta_0 + \beta_1 t$, то $M\Delta^2 x(t)$ равно**

a) 0

b) 1

c) $\beta_0$

d) $\beta_1$

**15. Сглаженное значение $\hat{f}(t)$ вычисляется по формуле**

a) $\sum_{k=0}^{m} w_k x(t+k)$

b) $\sum_{k=-m}^{m} w_k x(t+k)$

c) $\sum_{k=0}^{m} w_k x(k)$

d) $\sum_{k=-m}^{m} w_k x(t)$

**16. Для идентификации АР и СС моделей сначала делают оценки**

a) автокорреляционной функции

b) частной автокорреляции

c) автоковариационной функции
d) спектральной плотности

**17. В модели АР(1) частная автокорреляционная функция случайных остатков, разделенных двумя тактами времени, равна**

a) $\alpha^2$
b) $2\alpha$
c) 1
d) 0

**18. Если элементы набора данных не являются статистически независимыми, то речь идет о**

a) стационарном временном ряде
b) генеральной совокупности
c) случайной выборке
d) временном ряде

**19. В лаговой структуре Койка веса $w_k$ равны _____ , где** $0 \prec \lambda \prec 1$

a) $w_k = (1-\lambda)\lambda^k$
b) $w_k = \dfrac{\lambda^k}{1-\lambda}$
c) $w_k = \dfrac{\lambda^k}{1-k^{\lambda}}$
d) $w_k = \lambda^{k-1}$

**20. В критерии восходящих и нисходящих серий, общее число серий временного ряда 5, 7, 6, 4, 3,1 равно**

a) 1
b) 3
c) 2
d) 5

**21. Коэффициент автокорреляции $r[\varepsilon(t),\varepsilon(t\pm\tau)]$ случайных остатков в модели АР(1) равен**

a) $\alpha^r$
b) $\dfrac{\sigma_0^2}{1-\tau\alpha}$
c) $\tau\alpha$
d) $\tau\sigma_0^2$

**22. Критерий восходящих и нисходящих серий позволяет**

a) найти доверительный интервал прогноза
b) определить среднеквадратичное отклонение
c) найти минимальные и максимальные значения
d) выявить неслучайную составляющую

**23. Коэффициент Тейла лежит в пределах**

a) от –1 до 1
b) от 0 до $\infty$
c) от $-\infty$ до $\infty$
d) от 0 до 1

**24. В критерии серий, основанном на медиане, протяженность самой длинной серии временного ряда 5,1, 4, 2 равна**

a) 1
b) 3
c) 4
d) 2

**25. Для весовых коэффициентов в методе скользящего среднего справедлива формула**

a) $\sum_{k=-m}^{m} w_k^{\,2} = \sigma^2$

b) $\sum_{k=-m}^{m} w_k = 2m+1$

c) $\sum_{k=-m}^{m} w_k^{\,2} = 1$

d) $\sum_{k=-m}^{m} w_k = 1$

## Вопросы для повторения раздела

**1.** В чем состоит различие между моделями с распределенными лагами и авторегрессионными моделями?

**2.** Каковы основные причины лагов в эконометрических моделях?

**3.** Перечислите основные способы определения оценок для моделей с распределенными лагами?

**4.** В чем суть преобразования Койка?

**5.** В чем суть модели адаптивных ожиданий? В чем состоит отличие модели адаптивных ожиданий от модели частичной корректировки?

**6.** Опишите суть метода определения оценок на основе использования распределенных лагов Алмон?

**7.** Как определяется автокорреляция остатков в авторегрессионных моделях?

## Тренировочные задания

При выполнении тренировочных заданий в таблицы с данными вместо знаков «xx» необходимо подставить две последние цифры зачетной книжки.

**1.** В выборке представлены данные по цене P некоторого блага и количеству (Q) данного блага, приобретаемому хозяйством ежемесячно в течение года.

| Месяц | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| P | 10,xx | 20,xx | 15,xx | 25,xx | 30,xx | 35,xx |
| Q | 110,xx | 75,xx | 100,xx | 80,xx | 60,xx | 55,xx |

| Месяц | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|
| P | 40,xx | 35,xx | 25,xx | 40,xx | 45,xx | 40,xx |
| Q | 40,xx | 80,xx | 60,xx | 30,xx | 40,xx | 30,xx |

1) Постройте корреляционное поле и по его виду определите формулу зависимости между P и Q.
2) Оцените по МНК параметры уравнения линейной регрессии.
3) Оцените выборочный коэффициент корреляции rpq.
4) Проинтерпретируйте результаты.

**2.** Имеются данные за 10 лет по прибылям X и Y (в %) двух компаний:

| Год | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| X | 19,2xx | 15,8xx | 12,5xx | 10,3xx | 5,7xx | –5,8xx | –3,5xx | 5,2xx | 7,3xx | 6,7xx |
| Y | 20,1xx | 18,0xx | 10,3xx | 12,5xx | 6,0xx | –6,8xx | –2,8xx | 3,0xx | 8,5xx | 8,0xx |

1) Постройте регрессионную модель $Y=b_0+b_1X+e$.
2) Оцените статистическую значимость коэффициентов регрессии.
3) Оцените коэффициент детерминации R2 данного уравнения.
4) Постройте регрессионную модель $Y=bX+u$.
5) Приведите формулы расчета коэффициента b, его стандартной ошибки Sb и стандартной ошибки регрессии S (обратите внимание на число степеней свободы при расчете данной оценки).
6) Значимо или нет различаются коэффициенты b1 и b?
7) Какую из построенных моделей вы предпочтете?
8) Можно ли на основе построенных регрессий утверждать, что прибыль одной из компаний является следствием прибыли другой?

**3.** Для прогноза возможного объема экспорта на основе ВНП предложено использовать линейную регрессионную модель. При этом используются данные за 1995 – 2004 годы.

| Годы | 1995 | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|---|---|---|
| ВНП | 1000 | 1090 | 1150 | 1230 | 1300 | 1360 | 1400 | 1470 | 1500 | 1580 |
| Экспорт | 190 | 220 | 240 | 240 | 260 | 250 | 280 | 290 | 310 | 350 |

1) Сформулируйте соответствующую регрессионную модель, дав интерпретацию ее параметров.
2) Рассчитайте на основе имеющихся данных оценки параметров модели.
3) Вычислите стандартную ошибку регрессии.
4) Рассчитайте стандартные ошибки коэффициентов.
5) Определите 90 и 95%-е доверительные интервалы для теоретических коэффициентов регрессии.
6) Проанализируйте статистическую значимость коэффициентов при уровнях значимости $\alpha$=0,1 и $\alpha$=0,05.
7) Оцените коэффициент корреляции между ВНП и экспортом.
8) Дайте прогнозы по объему экспорта на 2006 и 2009 годы.
9) Определите 95%-е доверительные интервалы для этих прогнозов.
10) Рассчитайте коэффициент детерминации и сравните его с коэффициентом корреляции.
11) Какие предпосылки относительно случайного отклонения модели необходимы для обоснованности выводов по предыдущим пунктам?
12) Сделайте выводы по предыдущим пунктам.

**4.** Предполагается, что объем Q предложения некоторого блага для функционирующей в условиях конкуренции фирмы зависит линейно от цены P данного блага и заработной платы W сотрудников фирмы, производящих данное благо: $Q = \beta_0 + \beta_1 P + \beta_2 W + \varepsilon$.

Статистические данные, собранные за 16 месяцев, занесены в следующую таблицу:

| Q | 20 | 35 | 30 | 45 | 60 | 69 | 75 | 90 | 105 | 110 | 120 | 130 | 130 | 130 | 135 | 140 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P | 10 | 15 | 20 | 25 | 40 | 37 | 43 | 35 | 38 | 55 | 50 | 35 | 40 | 55 | 45 | 65 |
| W | 12 | 10 | 9 | 9 | 8 | 8 | 6 | 4 | 4 | 5 | 3 | 1 | 2 | 3 | 1 | 2 |

1) Оцените по МНК коэффициенты уравнения регрессии.
2) Проверьте гипотезы о том, что при прочих равных условиях рост цены товара увеличивает предложение; рост заработной платы снижает предложение.

3) Определите интервальные оценки коэффициентов при уровне значимости $\alpha$=0,1. Как с их помощью проверить гипотезу о статистической значимости коэффициентов регресс сии?
4) Оцените общее качество уравнения регрессии.
5) Является ли статистически значимым коэффициент детерминации R2?
6) Проверьте гипотезу об отсутствии автокорреляции остатков.
7) Сделайте выводы по построенной модели.

**5.** Анализируя прибыль предприятии Y(млн $) в зависимости от расходов на рекламу X(млн $). По наблюдениям за 9 лет получены следующие данные:

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Y | 5,xx | 7,xx | 13,xx | 15,xx | 20,xx | 25,xx | 22,xx | 20,xx | 17,xx |
| X | 0,8xx | 1,0xx | 1,8xx | 2,5xx | 4,0xx | 5,7xx | 7,5xx | 8,3xx | 8,8xx |

1) Постройте корреляционное поле и и выдвиньте предположение о формуле зависимости между рассматриваемыми показателями.
2) Оцените по МНК коэффициенты линейной регрессии $Y=b_0+b_1X+e$.
3) Оцените качество построенной регрессии.
4) Оцените по МНК коэффициенты квадратичной регрессии $Y=b_0+b_1X+b_2X^2+ e$.
5) Оцените качество построенной регрессии. Какую из моделей вы предпочтете?

**6.** В таблице приведены статистические данные по процентному изменению заработной платы (Y) , росту производительности труда (X1) и уровню инфляции (X2) за 20 лет:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 6,0xx | 8,9xx | 9,0xx | 7,1xx | 3,2xx | 6,5xx | 9,1xx | 14,6xx | 11,9xx | 9,4xx |
| $X_1$ | 2,8xx | 6,3xx | 4,5xx | 3,1xx | 1,5xx | 7,6xx | 6,7xx | 4,2xx | 2,7xx | 3,5xx |
| $X_2$ | 3,0xx | 3,1xx | 3,8xx | 3,8xx | 1,1xx | 2,3xx | 3,6xx | 7,5xx | 8,0xx | 6,3xx |

Продолжение таблицы

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Y | 12,0xx | 12,5xx | 8,5xx | 5,9xx | 6,8xx | 5,6xx | 4,8xx | 6,7xx | 5,5xx | 4,0xx |
| $X_1$ | 5,0xx | 2,3xx | 1,5xx | 6,0xx | 2,9xx | 2,8xx | 2,6xx | 0,9xx | 0,6xx | 0,7xx |
| $X_2$ | 6,1xx | 6,9xx | 7,1xx | 3,1xx | 3,7xx | 3,9xx | 3,9xx | 4,8xx | 4,3xx | 4,8xx |

1) По МНК постройте уравнение регрессии $y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + e_t$.
2) Оцените качество построенного уравнения, включая наличие автокорреляции и гетероскедастичности.
3) По МНК постройте уравнение регрессии $y_t = c_0 + c_1 x_{1t-1} + c_2 x_{2t-1} + v_t$, учитывая, что x10=3,5; x20=4,5.
4) Оцените качество построенного уравнения.
5) Сравните построенные модели. Какая из них предпочтительнее и почему?

## Вопросы к экзамену

1. Эконометрика как наука.
2. Предмет, методы, задачи и основные принципы эконометрики.
3. Эконометрический эксперимент и его результаты.
4. Особенности эконометрического метода.
5. История возникновения эконометрики.
6. Основные моменты эконометрического моделирования.
7. Методологические вопросы построения эконометрических моделей: обзор используемых методов.
8. Основные математические предпосылки эконометрического моделирования.
9. Эконометрическая модель и экспериментальные данные.
10. Пространственная выборка.
11. Временной и динамический ряд.
12. Понятие о функциональной, статистической и корреляционной связях. Основные задачи прикладного корреляционно-регрессионного анализа.
13. Линейная регрессионная модель.
14. Системы одновременных уравнений.
15. Основные этапы и проблемы эконометрического моделирования.
16. Функциональная, статистическая и корреляционная зависимости.
17. Линейная парная регрессии.
18. Коэффициент корреляции.
19. Основные положения регрессионного анализа. Оценка параметров парной регрессионной модели.
20. Метод наименьших квадратов и условия его применения для определения параметров уравнения парной регрессии.
21. Теорема Гаусса-Маркова для случая парной регрессионной модели.
22. Метод максимального правдоподобия.
23. Интервальная оценка функции регрессии и ее параметров. Доверительный интервал для функции регрессии.
24. Интервальная оценка функции регрессии и ее параметров. Доверительный интервал для индивидуальных значений зависимой переменной и для параметров регрессионной модели.

25. Оценка значимости уравнения регрессии. Коэффициент детерминации.
26. Геометрическая интерпретация регрессии и коэффициента детерминации.
27. Оценка статистической значимости показателей корреляции, параметров уравнения регрессии, уравнения регрессии в целом: t-критерий Стьюдента, F-критерий Фишера.
28. Обобщенная линейная модель множественной регрессии. Обобщенный метод наименьших квадратов.
29. Коэффициент ранговой корреляции Спирмена.
30. Множественная линейная регрессия.
31. Метод наименьших квадратов для множественной линейной модели.
32. Теорема Гаусса-Маркова для множественной линейной модели.
33. Проверка общего качества оценивания. Коэффициент детерминации для множественной модели.
34. Оценивание значимости коэффициента детерминации.
35. Спецификация переменных. Отбор объясняющих переменных.
36. Спецификация переменных. Мультиколлинеарность.
37. Мультиколлинеарность. Методы устранения мультиколлинеарности.
38. Явление автокорреляции.
39. выявление автокорреляции.
40. Устранение автокорреляции.
41. Стохастические объясняющие переменные. Стохастические регрессоры.
42. Стохастические объясняющие переменные. Метод инструментальных переменных.
43. Стохастические объясняющие переменные. Гетероскедастичность.
44. Примеры нелинейных моделей и преобразование переменных.
45. Нелинейные модели, неприводимые к линейному виду.
46. Анализ линейной модели множественной регрессии при гетероскедастичности и автокорреляции.
47. Фиктивные переменные для коэффициентов наклона. Тест Чоу.
48. Множественный коэффициент корреляции и множественный коэффициент детерминации. Оценка надежности показателей корреляции.

49. Оценка качества модели множественной регрессии: F-критерий Фишера, t-критерий Стьюдента.
50. Структурная и приведенная формы модели системы линейных уравнений.
51. Оценивание коэффициентов структурной модели.
52. Моделирование динамических процессов. Основные понятия.
53. Модели с распределенным лагом.
54. Динамические процессы.
55. Стохастическая природа экономических данных. Понятие случайной переменной.
56. Точечные оценки. Характеристики генеральной совокупности: математическое ожидание и дисперсия.
57. Оценка как случайная величина. Несмещенность. Эффективность. Состоятельность
58. Выборочное среднее как оценка математического ожидания.
59. Оценка теоретической дисперсии. Ковариация.
60. Основные статистические распределения, используемые в регрессионном анализе.
61. Основные правила проверки гипотез.
62. Зависимость между критериями в парном регрессионном анализе.
63. Мощность критерия. Доверительные интервалы.
64. Алгоритм поиска уравнения регрессии по МНК.
65. Нелинейные регрессии, сводящиеся к линейным. Степенная регрессия. Функция Кобба-Дугласа.
66. Аналитическое выравнивание временных рядов. Оценка параметров уравнения тренда.
67. Автокорреляция в остатках, ее измерение и интерпретация. Критерий Дарбина-Уотсона в оценке качества трендового уравнения регрессии.
68. Анализ временных радов при наличии периодических колебаний: аддитивная и мультипликативная модели.
69. Системы одновременных уравнений. Структурная и приведенная формы модели.
70. Оценивание коэффициентов структурной модели.

## ГЛОССАРИЙ

**Эконометрика** – часть экономической науки, занимающаяся разработкой и применением математических, и прежде всего экономико-статистических, методов анализа экономических процессов, обработки статистической экономической информации.

**Эконометрические методы –** методы исследования экономики, изучающие экономические процессы с количественной стороны.

**Выборка –** некоторое количество наблюдений, отобранных из генеральной совокупности.

**Наблюдение –** наблюдаемое значение случайной величины или набора случайных величин.

**Оценка, способ оценивания (estimator) –** общее правило, формула для получения приближенного численного значения какого- либо параметра по данным выборки.

**Значение оценки (estimator) –** число, полученное в результате применения оценки к конкретной выборке.

**Смещение** – разность между математическим ожиданием оценки и истинным значением оцениваемого параметра.

**Несмещенная оценка –**оценка, имеющая нулевое смещение.

**Эффективная оценка** – несмещенная оценка, имеющая наименьшую дисперсию среди всех несмещенных оценок.

**Эксперимент по методу Монте-Карло** – искусственный, контролируемый эксперимент, проводимый для проверки и сравнения эффективности различных статистических методов.

**Состоятельная оценка** – оценка, у которой смещение и дисперсия стремятся к 0 при увеличении объема выборки**.**

**Модель** – совокупность переменных и связей между ними в форме уравнений, описывающая зависимость между наблюдаемыми переменными.

**Модель парной регрессии** – простейшая линейная модель зависимости между двумя переменными: $y = \alpha + \beta x + \varepsilon$.

**Зависимая переменная регрессии** – переменная величина в модели парной регрессии, которую считают (по экономическим соображениям) зависящей от другой переменной (в модели $y = \alpha + \beta x + \varepsilon$ зависимая переменная - *y)*.

**Объясняющая переменная регрессии (регрессор)** – переменная величина в модели парной регрессии, от которой зависит (по экономическим соображениям) зависимая переменная (в модели $y = \alpha + \beta x + \varepsilon$ объясняющая переменная - *x*).

**Случайный член регрессии** – слагаемое и в модели $y = \alpha + \beta x + \varepsilon$, которое описывает воздействие случайных факторов.

**Уравнение линейной регрессии** – уравнение $y = a + bx$, где *a* и *b*-оценки параметров $\alpha$ и $\beta$, полученные в результате оценивания модели регрессии $y = \alpha + \beta x + \varepsilon$ по данным выборки.

**Остаток в наблюдении** – разность $y_i - (a + bx_i)$ между истинным значением переменной *y* в *i*-ом наблюдении ($y_i$) и значением $a + bx_i$ *i*-ом наблюдении, полученным подстановкой наблюдения $x_i$ в уравнение линейной регрессии.

**Метод наименьших квадратов *(МНК)* (OLS – Ordinary Least Squares)** – метод нахождения оценок параметров регрессии, основанный на минимизации суммы квадратов остатков всех наблюдений.

**Объясненная дисперсия зависимой переменной** – выборочная дисперсия расчетных значений величины *y*: $Var(a + bx)$.

**Необъясненная дисперсия зависимой переменной** – выборочная дисперсия остатков в наблюдениях: $Var(y - (a + bx))$.

**Общая сумма квадратов отклонений (TSS – Total Sum of Squares)** – сумма квадратов отклонений величины *y* от своего выборочного среднего $\bar{y}$.

**Объясненная сумма квадратов отклонений (ESS – Explained Sum of Squares)** – сумма квадратов отклонений величины $a + bx$ от своего выборочного среднего $a + b\bar{x}$.

**Необъясненная (остаточная) сумма квадратов отклонений (RSS – Unexplained Sum of Squares)** – сумма квадратов остатков всех наблюдений.

**Коэффициент детерминации $R^2$** – доля объясненной дисперсии зависимой переменной во всей выборочной дисперсии *y*:

$$R^2 = \frac{Var(a + bx)}{Var(y)} = \frac{ESS}{TSS}.$$

**Стандартное отклонение случайной величины** – корень квадратный из теоретической дисперсии случайной величины; среднее ожидаемое расстояние между наблюдениями этой случайной величины и ее математическим ожиданием.

**Стандартная ошибка случайной величины** – оценка стандартного отклонения случайной величины, полученная по данным выборки..

**Нулевая гипотеза ($H_o$)** – утверждение о том, что неизвестный параметр модели принадлежит заданному множеству *A*.

**Альтернативная гипотеза** – утверждение о том, что неизвестный параметр модели принадлежит другому заданному множеству *B*, $A \cap B = \varnothing$.

**Область принятия гипотезы** – множество значений оценок параметра, при попадании в которое нулевая гипотеза не отвергается.

**Ошибка I рода** – ситуация, когда оценка параметра не попала в область принятия нулевой гипотезы, нулевая гипотеза была отвергнута, хотя та была истинной.

**Ошибка II рода** – ситуация, когда не отвергнута ложная гипотеза.

**Цена ошибки** – численное выражение ущерба от ошибки, величина «штрафа» за ошибку.

**Функция цены** – функция, где аргументом является род ошибки, а значением функции – цена ошибки..

**T-тест (тест Стьюдента)** – проверка гипотезы $H_0 : \beta = \beta_0$ о значении коэффициента $\beta$ с помощью распределения Стьюдента.

**Число степеней свободы** – натуральное число, характеристика таких законов распределения, как распределение Стьюдента, распределение Фишера и некоторых других.

**Критическое значение теста при *p*-процентном уровне значимости** – граничное значение области принятия гипотезы, проверяемой тестом, *p*-процентной вероятностью совершить ошибку I рода.

**Доверительный интервал** – интервал с центром в полученной оценке параметра, который содержит истинное значение параметра с доверительной вероятностью.

**Односторонний тест** – тест на проверку гипотезы, в котором область принятия гипотезы имеет только одно критическое значение.

**Двусторонний тест** – тест на проверку гипотезы, в котором область принятия гипотезы имеет два критических значения - меньшее и большее.

**F-тест (тест Фишера)**' – проверка гипотезы $H_0 : R^2 = 0$ (значимость всей регрессии) с помощью распределения Фишера.

**Нелинейная по переменным модель** – нелинейная модель $y = f(x)$, в которой возможна замена переменной $z = g(x)$, приводящая получившуюся модель $y = F(z)$ – к линейной.

**Нелинейная по параметрам модель** – модель, которую нельзя привести заменами переменных к линейной.

**Логарифмическое преобразование** – переход от нелинейной и по переменным, и по параметрам модели $y = \alpha x^{\beta} \varepsilon$ к логарифмической модели $\ln y = \ln \alpha + \beta \ln x + \ln \varepsilon$.

**Метод Зарембки** – процедура выбора между линейной и логарифмической моделями.

**Тест Бокса-Кокса (решетчатый поиск)** – прямой компьютерный метод выбора наилучших значений параметров нелинейной модели в заданных исследователем пределах с заданным шагом (решеткой).

**Итерационные методы** – компьютерные сходящиеся методы поиска наилучших значений параметров нелинейной модели.

**Модель множественной регрессии** – линейная модель зависимости между переменными: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon$, содержащая более двух переменных

**Модель множественной регрессии без свободного коэффициента** – линейная модель зависимости между переменными: $y = \alpha + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon$, не содержащая коэффициента $\alpha$

**Плоскость регрессии** – $m$-мерная плоскость $y = a + b_1 x_1 + b_2 x_2 + ... + b_m x_m$ в $(m + 1)$-мерном пространстве

**Нестрогая линейная зависимость между переменными** – ситуация, когда теоретическая корреляция двух переменных близка к 1 или –1

**Строгая линейная зависимость между переменными** – ситуация, когда выборочная корреляция двух переменных равна 1 или –1

**Мультиколлинеарность** – явление, когда нестрогая линейная зависимость между объясняющими переменными в модели множественной регрессии приводит к получению ненадежных оценок регрессии

**Полная коллинеарность** – явление, когда строгая линейная зависимость между переменными приводит к невозможности применения МНК

**Лишняя переменная** – объясняющая переменная, включенная 'в модель множественной регрессии, в то время, как по экономическим причинам ее присутствие в модели не нужно

**Отсутствующая переменная** – необходимая по экономическим причинам объясняющая переменная, отсутствующая в модели

**Спецификация переменных** – выбор необходимых для регрессии переменных и отбрасывание лишних переменных

**Замещающая переменная** – объясняющая переменная, используемая в регрессии вместо трудноизмеримой, но важной переменной

**Лаговая переменная** – наблюдение зависимой переменной регрессии в предшествующий момент, используемое как объясняющая переменная

**Фиктивная переменная** – переменная, принимающая в каждом наблюдении только два значения: 1 – «да» или 0 – «нет»

**Категория** – событие, про которое для каждого наблюдения можно определенно сказать - произошло оно в этом наблюдении или нет

**Набор категорий** – конечный набор взаимоисключающих событий, полностью исчерпывающий все возможности

**Совокупность фиктивных переменных** – некоторое количество фиктивных переменных, предназначенное для описания набора категорий

**Эталонная категория** – категория, с которой сравниваются другие категории

**Сезонные фиктивные переменные** – совокупность фиктивных переменных, предназначенная для обозначения различных лет, времен года, месяцев и т.п.

**Ловушка dummy trap** – выбор такой совокупности фиктивных переменных, у которой сумма этих переменных тождественно равна константе

**Фиктивная переменная взаимодействия** – фиктивная переменная, предназначенная для установления влияния на регрессию одновременного наступления сразу нескольких независимых друг от друга событий, каждое из которых описывается своей фиктивной переменной

**Гетероскедастичность** – нарушение второго условия теоремы Гаусса-Маркова, которое заключается в том, что дисперсия случайного члена регрессии зависит от номера наблюдения: $\sigma^2(\varepsilon_i)$ – зависит от $i$

**Ранг наблюдения переменной** – номер наблюдения переменной в упорядоченной по возрастанию последовательности

**Тест ранговой корреляции Спирмена** – тест на гетероскедастичность, устанавливающий, что стандартное отклонение остаточного члена регрессии имеет нестрогую линейную зависимость с объясняющей переменной

**Тест Голдфелда-Квандта** – тест на гетероскедастичность, устанавливающий, что стандартное отклонение остаточного члена регрессии растет, когда растет объясняющая переменная

**Тест Глейзера** – наиболее тонкий тест на гетероскедастичность, улавливающий нелинейную связь между стандартным отклонением остаточного члена регрессии и объясняющей переменной

**Автокорреляция (случайного члена в уравнении регрессии)** – нарушение третьего условия Гаусса-Маркова, которое заключается в том, что случайные члены регрессии в разных наблюдениях являются зависимыми: $cov(\varepsilon_k, \varepsilon_i) \neq 0$, при $k \neq i$

**Положительная автокорреляция (случайного члена)** – ситуация, когда случайный член регрессии в следующем наблюдении ожидается того же знака, что и случайный член в настоящем наблюдении

**Отрицательная автокорреляция (случайного члена)** – ситуация, когда случайный член регрессии в следующем наблюдении ожидается знака, противоположного знаку случайного члена в настоящем наблюдении

**Автокорреляция первого порядка** – ситуация, когда коррелируют случайные члены регрессии в последовательных наблюдениях

**Авторегрессионная схема первого порядка** – частный случай автокорреляции первого порядка, когда зависимость между последовательными случайными членами, описывается формулой $\varepsilon_{k+1} = p\varepsilon_k + u_{k+1}$, где $p$ – константа, $U_{k+1}$ -новый случайный член

Критерий Дарбина-Уотсона – метод обнаружения автокорреляции первого порядка с помощью статистики Дарбина-Уотсона

**Зона неопределенности критерия Дарбина-Уотсона** – промежуток значений статистики Дарбина-Уотсона, при попадании в который критерий не дает определенного ответа о наличии или отсутствии автокорреляции первого порядка

**Поправка Прайса-Уинстена** – метод спасения первого наблюдения в автокорреляционной схеме первого порядка

**Метод Кокрана-Оркатта** – компьютерный итерационный метод устранения автокорреляции первого порядка.

**Панельные данные –** данные нескольких одновременных временных рядов

**Временной ряд (time series) –** наблюдения экономического показателя одного объекта в равноотстоящие моменты времени

**Член временного ряда –** наблюдение экономического показателя одного объекта в некоторый момент времени

**Перекрестные данные (cross-section data) –** выборка из экономических показателей, полученная для большого количества однотипных объектов (семей, фирм, регионов, стран); все наблюдения или одновременные, или считаются независимыми от времени

**Долговременные факторы –** неслучайные факторы, формирующие тенденцию

**Тренд –** тенденция, которую формируют долговременные факторы

**Сезонные факторы –** факторы, обусловленные периодичностью (сезонной, квартальной)

**Циклические (конъюнктурные) факторы –** факторы, обусловленные действием долгосрочных циклов (солнечная активность, демографические «ямы», волны Кондратьева, политические выборы)

**Случайные факторы –** факторы, не поддающиеся учету и регистрации

**Разладочные случайные факторы –** случайные факторы, приводящие к резкому изменению (слому) всей модели

**Эволюционные остаточные случайные факторы –** случайные факторы, влияние которых не приводит к резкому изменению ни характера, ни параметров модели

**Строго стационарный (стационарный в узком смысле) временной ряд –** временной ряд *x(t),* у которого совместное распределение вероятностей *m*

наблюдений $x(t_1), x(t_2), \ldots, x(t_m)$ такое же, как и для $m$ наблюдений $x(t_1+\tau), x(t_2+\tau), \ldots, x(t_m+\tau)$ для любого $m, t_1, t_2, t_m$ и $\tau$.

**Стационарный (стационарный в широком смысле) временной ряд** – Временной ряд $x(t)$ с постоянным математическим ожиданием $M(x(t))$ и дисперсией $D(x(t))$, не зависящими от $t$

**Нестационарный временной ряд** – Временной ряд отличающийся от стационарного на неслучайную составляющую (тренд)

**Автоковариационная функция** – функция $\gamma(\tau) = cov(x(t), x(t+\tau))$ для стационарного ряда (зависит только от τ)

**Автокорреляционная функция** – функция $r(\tau) = cor(x(t), x(t+\tau))$ для стационарного ряда (зависит только от τ)

**Коррелограмма** – график автокорреляционной функции

**Частная (очищенная) автокорреляционная функция** – функция, измеряющая корреляцию $x(t)$ и $x(t+\tau)$ напрямую, без влияния промежуточных между ними наблюдений

**Спектральная плотность временного ряда** – сумма ряда $p(\omega) = 1 + 2\sum_{t=1}^{+\infty} r(\tau)cos(t\omega)$, где $r(\tau)$ – автокорреляционная функция.

**Серия** – Последовательность подряд идущих плюсов или минусов

**Критерий серий** – критерий, основанный на исследовании количества серий и их длин в последовательности

**Метод скользящего среднего** – метод сглаживания временного ряда для уменьшения влияния случайных факторов

**Метод последовательных разностей** – метод поиска степени многочлена, описывающего тренд

**Белый шум** – временной ряд $\delta(t), M\delta(t) = 0, M(\delta(t)\delta(t \pm \tau)) = \begin{cases} \sigma_0^2 при \tau = 0 \\ 0 при \tau \neq 0 \end{cases}$, серия импульсов, генерирующая случайные остатки анализируемого временного ряда.

**Модель авторегрессии 1-го порядка АР(1), марковский процесс (AR(1) models) –** временной ряд, описываемый формулой $\varepsilon(t) = \alpha\varepsilon(t-1) + \delta(t)$,-где $\delta(t)$ – белый шум

**Модель авторегрессии 2-го порядка АР(2), модель Юла (AR(2) – models) –** временной ряд, описываемый формулой $\varepsilon(t) = \alpha_1\varepsilon(t-1) + \alpha_2\varepsilon(t-2) + \delta(t)$, где $\delta(t)$ – белый шум

**Модель скользящего среднего 1-го порядка СС(1) (MA(1) models) –** временной ряд, описываемый формулой $\varepsilon(t)=\delta(t)-\beta\delta(t-1)$, где $\delta(t)$ – белый шум

**Условия стационарности –** условия на параметры модели, при которых временной ряд получается стационарным

**Идентификация модели временного ряда –** оценка параметров модели временного ряда - чисел $\alpha,\alpha_1,\alpha_2,\beta,\beta_1,\beta_2\ldots\sigma_0^2$

**Модель скользящего среднего 2-го порядка СС(2) (MA(2) - models ) –** временной ряд, описываемый формулой $\varepsilon(t)=\delta(t)-\beta_1\delta(t-1)-\beta_2\delta(t-2)$, где $\delta(t)$ – белый шум

**Условия обратимости –** условия на параметры модели, при которых зависимость значения временного ряда от прошлых значений уменьшается с отдалением прошлого

**Модель Бокса-Дженкинса (АРПСС(p,q,k)) (ARIMA- models) –** временной ряд, у которого тренд – алгебраический полином степени *(k-1),* а остаток - АРСС*(p, q)*

**Регрессионная модель с распределенными лагами –** модель зависимости $y(t)=c_0+\beta_0x(t)+\beta_1x(t-1)+\ldots+\beta_px(t-p)+\delta(t)$, где $\delta(t)$ – белый шум

**Лаговая структура Ш. Алмон –** Регрессионная модель с распределенными лагами, в которой параметры получаются по формуле $\beta_k=A_0+A_1k+A_2k^2+\ldots+A_mk^m$, где $A_0$, $A_1$, $A_2,\ldots A_m$ –неизвестные параметры

**Лаговая структура Койка –** Регрессионная модель с распределенными лагами, в которой параметры убывают в геометрической прогрессии: $y(t)=c_0+\beta x(t)+\beta\lambda x(t-1)+\beta\lambda^2x(t-p)+\ldots+\delta(t)$, где $\delta(t)$ – белый шум, $-1<\lambda<1$, но чаше всего $0<\lambda<1$

**Модель частичного приспособления (частичной корректировки) –** модель, в которой предполагается, что уравнение определяет не фактическое *y(t),* а желаемое $\hat{y}(t)$ значение: $\hat{y}(t)=\beta_0+\beta_1x(t)+\delta(t)$, причем фактическое приращение зависимой переменной $(y(t)-y(t-1))$ пропорционально разнице между ее желаемым уровнем и значением в предыдущий период: $y(t)-y(t-1)=\mu(\hat{y}(t)-y(t-1))+\varepsilon(t)$, где $\varepsilon(t)$ – случайный член

**Модель Линтнера –** модель частичного приспособления, описывающая выплату дивидендов

**Модель адаптивных ожиданий –** модель, в которой корректируется ожидаемое значение объясняющей переменной $\hat{x}(t+1)$ (но экспертно формируемое в момент *t*)

**Модель Кейгана –** модель, описывающая гиперинфляцию с помощью модели адаптивных ожиданий

**Предсказание (prediction) (безусловное прогнозирование) –** оценка значения зависимой переменной *y* в момент *T+p,* полученная на основе первых *T* наблюдений, когда значения объясняющих переменных в этот момент известны: *x(T+p)*.

**Прогноз (forecast)** – оценка значения зависимой переменной *y* в момент *T+p,* полученная на основе первых *T* наблюдений, когда значения объясняющих переменных *x(T+p)* в этот момент неизвестны

**Относительная ошибка прогноза (RFE)** – величина $\frac{\Delta\hat{y}(T+p)-\Delta y(T+p)}{\Delta y(T+p)}$, где $\hat{y}(T+p)$ – прогноз; $\Delta\hat{y}(T+p)=\hat{y}(T+p)-y(T)$ – предсказываемый прирост; $\Delta y(T+p)=y(T+p)-y(T)$ – действительный прирост.

**Тест Чоу** – проверка гипотезы о неудачности предсказания с помощью статистики Фишера.

## СПИСОК ЛИТЕРАТУРЫ


1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов. - М.: ЮНИТИ, 1998. - 1022 с.
2. Боровиков В.П., Боровиков И.П. STATISTICA – Статистический анализ и обработка данных в среде Windows. – М.: Информационно-издательский дом «Филин», 1997. - 608 с.
3. Бородич С.А. Эконометрика: Учеб. пособие / С.А. Бородич. – 2-е изд., испр. – Мн.: Новое знание, 2004. – 416с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 1998. - 479 с.
5. Грицан В. К Эконометрика: учебное пособие. — М: Издательско-торговая корпорация «Дашков и К°», 2002.
6. Доугерти К. Введение в эконометрику: Пер с англ. - М.: ИНФРА-М, 1996.-416 с.
7. Ежеманская С.Н. Эконометрика / Серия «Учебники, учебные пособия». – Ростов н/Д: Феникс 2003. – 160с.
8. Ефимова М.Р., Петрова Е.В., Румянцев В.Н. Общая теория статистики. Учебник. - М.: ИНФРА-М, 1996. - 416 с.
9. Замков О.О., Толстопятенко А.В., Черемных Ю.Н. Математические методы в экономике. Учебник. - М.: МГУ им. М.В. Ломоносова, Издательство «ДИС», 1998.-368 с.
10. Катышев П.К., Пересецкий А.А. Сборник задач к начальному курсу эконометрики. - М.: Дело, 1999. - 72 с.
11. Колемаев В.А. и др. Теория вероятностей и математическая статистика: Учебное пособие для экон. спец. вузов/ Под ред. В.А. Колемаева. — М.: Высш. шк., 2000. — 400 с.
12. Колемаев В.А. Эконометрика: Учебник. – М.: ИНФРА – М, 2004.
13. Кремер Н.Ш., Путко Б.А. Эконометрика. - М.: ЮНИТИ-ДАНА, 2002.
14. Лева О.В. Эконометрика: Учеб. пособие. – Белгород: Изд-во БелГТАСМ, 2002, – 80 с.
15. Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. Учебное пособие. 4-е изд. - М.: Дело, 2000. - 400 с.
16. Мардас А.Н. Эконометрика. – СПб: Питер, 2001. –144с.
17. Уотшем Т. Дж., Паррамоу К. Количественные методы в финансах: Учебное пособие для вузов: Пер. с англ. / Под ред. М.Р. Ефимовой. - М.: Финансы, ЮНИТИ, 1999. - 527 с.
18. Эконометрика / Под ред. члена-корреспондента Российской академии наук И. И. Елисеевой. — М.: Финансы и статистика, 2002.