МИНИСТЕРСТВО СЕЛЬСКОГО ХОЗЯЙСТВА
РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГОУ ВПО «СМОЛЕНСКАЯ ГОСУДАРСТВЕННАЯ СЕЛЬСКОХОЗЯЙСТВЕННАЯ АКАДЕМИЯ»

БЕЛОКОПЫТОВ А.В.

ОСНОВЫ ЭКОНОМЕТРИКИ

Учебное пособие

Смоленск 2013

УДК 330.43:330.45
ББК 65.050
Б-114

Р е ц е н з е н т :

Зав. кафедрой менеджмента, рекламы и таможенного дела филиала ГОУ ВПО РГТЭУ в г. Смоленске, к.э.н. Зюскин А.А.

Б 114 Белокопытов А. В.

Основы эконометрики. – Смоленск, 2013. – 156 с.

В учебном пособии рассмотрены особенности эконометрических исследований, применение методов корреляционно-регрессионного анализа, оценка моделей временных рядов и системы одновременных линейных уравнений. Раскрывается ряд специальных вопросов, связанных с качественным анализом экономических переменных и особенностями применения в АПК.

Предназначено для студентов экономических специальностей сельскохозяйственных вузов, изучающих дисциплины «Эконометрика», «Статистика», «Математическое моделирование производственно-экономических процессов и систем», «Планирование и прогнозирование развития АПК», а также аспирантов и научных работников.

ВВЕДЕНИЕ

Эконометрика как дисциплина федерального компонента по циклу общих математических и естественно-научных дисциплин впервые включена в основную образовательную программу подготовки экономистов в соответствии с Государственным образовательным стандартом второго поколения. Несмотря на то, что в настоящее время появилось большое количество учебной литературы по данному курсу, ощущается острая нехватка доступных и понятных учебников и учебных пособий по эконометрике для студентов экономических специальностей вузов. Применение аспектов математики в различных областях знаний (экономика, физика, химия, биология, социология и т.д.) принесло значительные успехи. В настоящее время идет накопление информации в различных областях экономических знаний с использованием эконометрики.

Для экономической теории и практики все большее значение приобретают вопросы прогнозирования развития общественного производства, выбора альтернатив экономического роста в условиях ограниченных ресурсов, определения наиболее эффективных путей достижения экономичных результатов.

Действенным инструментом научного анализа тенденций и прогнозирования перспектив экономического роста является экономико-математическое моделирование, в частности использование методов корреляционно-регрессионного анализа и временных моделей. Особенно эффективно использование эконометрических моделей при изучении динамики и тенденций развития экономики, выявления влияния важнейших факторов на конечные результаты.

Пособие содержит курс лекций по основным разделам эконометрики: парная и множественная регрессия, системы эконометрических уравнений и временные ряды. По всем главам имеются примеры задач с подробным анализом их решения. Типовые упражнения, представленные в отдельном разделе, предназначены для самостоятельной работы студентов и дают возможность глубже и полнее усвоить пройденный материал. Необходимые для решения задач математико-статистические таблицы приведены в приложении пособия. По своему содержанию учебное пособие соответствует требованиям образовательного стандарта Министерства образования РФ.

РАЗДЕЛ I. ПАРНАЯ И МНОЖЕСТВЕННАЯ РЕГРЕССИЯ И КОРРЕЛЯЦИЯ

1.1 Особенности эконометрического метода и взаимосвязей в экономике

Эконометрика – одна из базовых дисциплин экономического образования во всем мире. Однако до недавнего времени она не была признана в СССР и России. Это было связано с тем, что из трех основных составляющих эконометрики – экономической теории, экономической статистики и математики – две первые были представлены в нашей стране неудовлетворительно. Но теперь ситуация изменилась коренным образом.

Язык экономики все больше становится языком математики, а экономику все чаще называют одной из наиболее математизированных наук. Достижения современной экономической науки предъявляют новые требования к высшему профессиональному образованию экономистов. Современное экономическое образование, — утверждает директор ЦЭМИ РАН академик В. Л. Макаров, — держится на трех китах: макроэкономике, микроэкономике и эконометрике. Если в период централизованной плановой экономики упор делался на балансовых и оптимизационных методах исследования, на описании «системы функционирования социалистической экономики», построении оптимизационных моделей отраслей и предприятий, то в период перехода к рыночной экономике возрастает роль эконометрических методов. Без знания этих методов невозможно ни исследование и теоретическое обобщение эмпирических зависимостей экономических переменных, ни построение сколько-нибудь надежного прогноза в банковском деле, финансах или бизнесе.

Термин «эконометрика» был впервые введен бухгалтером П. Цьемпой (Австро-Венгрия, 1910 г.). Цьемпа считал, что если к данным бухгалтерского учета применить методы алгебры и геометрии, то будет получено новое, более глубокое представление о результатах хозяйственной деятельности. Это употребление термина, как и сама концепция, не прижилось, но название «эконометрика» оказалось весьма удачным для определения нового направления в экономической науке, которое выделилось в 1930 гг.

Зарождение эконометрики является следствием междисциплинарного подхода к изучению экономики. Эта наука возникла в результате взаимодействия и объединения в особый «сплав» трех компонент: экономической теории, статистических и математических методов. Впоследствии к ним присоединилось развитие вычислительной техники как условие развития эконометрики.

Единое общепринятое определение эконометрики отсутствует. На наш взгляд, наиболее рациональным является следующее определение: «Эконометрика — это наука, которая дает количественное выражение взаимосвязей экономических явлений и процессов».

Основные результаты экономической теории носят качественный характер, а эконометрика вносит в них эмпирическое содержание. Математическая экономика выражает экономические законы в виде математических соотношений, а эконометрика осуществляет опытную проверку этих законов. Экономическая статистика дает информационное обеспечение исследуемого процесса в виде исходных (обработанных) статистических данных и экономических показателей, а эконометрика, используя традиционные математико-статистические и специально разработанные методы, проводит анализ количественных взаимосвязей между этими показателями.

Устойчивое финансовое положение предприятия зависит от надежных экономически обоснованных управленческих решений по осуществлению хозяйственной деятельности. Обоснование процесса принятия управленческих решений является одной из наиболее важных задач эконометрики. Сам ход принятия управленческих решений должен учитывать их многовариантность, наличие неопределенности, оценку влияния факторов на каждый отдельно взятый вариант, установление параметров оптимальности и т.д. Выбор наилучшего варианта проводится путем применения эконометрических расчетов, которые помогают хозяйственному руководителю выработать правильное решение.

Эконометрический метод складывался в преодолении следующих неприятностей, искажающих результаты применения классических статистических методов:

• асимметричности связей;

• мультиколлинеарности объясняющих переменных;

• закрытости механизма связи между переменными в изолированной регрессии;

• эффекта гетероскедастичности, т. е. отсутствия нормального распределения остатков для регрессионной функции;

• автокорреляции;

• ложной корреляции;

• наличия лагов.

Эконометрическое исследование включает решение следующих проблем:

• качественный анализ связей экономических переменных — выделение зависимых (y_i) и независимых переменных (x_k);

• подбор данных;

• спецификация формы связи между y и x_k,

• оценка параметров модели;

• проверка ряда гипотез о свойствах распределения вероятностей для случайной компоненты (гипотезы о средней, дисперсии и ковариации);

• анализ мультиколлинеарности объясняющих переменных, оценка ее статистической значимости, выявление переменных, ответственных за мультиколлинеарность;

• введение фиктивных переменных;

• выявление автокорреляции, лагов;

• выявление тренда, циклической и случайной компонент;

• проверка остатков на гетероскедастичность;

• анализ структуры связей и построение системы одновременных уравнений;

• моделирование на основе системы временных рядов;

Пример эконометрическое исследования можно увидеть на следующем пример. Допустим, мы продаем автомобиль и решили дать объявление о продаже в газете «Из рук в руки». Возникает вопрос: какую цену указать в объявлении? Очевидно, мы будем руководствоваться информацией о цене, которую выставляют другие продавцы подобных автомобилей. Что значит «подобные автомобили», т.е. автомобили, обладающие близкими значениями таких факторов, как год выпуска, пробег, мощность двигателя и т.д. После анализа колонок объявлений мы назначаем цену. Этот конкретный пример помог нам проследить основные моменты эконометрического моделирования.

Таким образом, в качестве этапов эконометрического исследования можно указать:

• постановку проблемы;

• получение данных, анализ их качества;

• спецификацию модели;

• оценку параметров;

• интерпретацию результатов.

Эконометрическая модель, как правило, основана на теоретическом предположении о круге взаимосвязанных переменных и характере связи между ними. Взаимосвязанные признаки подразделяются на факторные (под их воздействием изменяются другие, зависящие от них, признаки) и результативные.

Связи по степени тесноты могут быть функциональными (при которых определенному значению факторного признака соответствует строго определенное значение результативного признака) и статистическими или корреляционными (когда одному и тому же значению факторного признака может соответствовать несколько значений результативного признака).

Корреляционная зависимость проявляется только в средних величинах и выражает числовое соотношение между ними в виде тенденции к возрастанию или убыванию одной переменной величины при возрастании или убывании другой. Объяснение этому – сложность взаимосвязей между анализируемыми факторами, на взаимодействие которых влияют неучтенные случайные величины. Поэтому связь между признаками проявляется лишь в среднем, в массе случаев.

Корреляционная связь является свободной, неполной и неточной связью. Например, себестоимость величины продукции зависит от уровня производительности труда: чем выше производительность труда, тем ниже себестоимость. Но себестоимость зависит также и от ряда других факторов: стоимости сырья и материалов, топлива, электроэнергии, их расхода на единицу продукции, цеховых и общезаводских расходов и т.д. Поэтому нельзя утверждать что при повышении производительности труда, допустим на 10% себестоимость снизится также на 10%. Может случиться, что несмотря на рост производительности труда, себестоимость не только не снизится, но даже несколько повысится, если на нее окажут более сильное влияние действующие в обратном направлении другие факторы.

По направлению связи бывают прямыми, когда зависимая переменная растет с увеличением факторного признака, и обратными, при которых рост последнего сопровождается уменьшением функции. Относительно своей аналитической формы связи бывают линейными и нелинейными. Существует еще одна достаточно важная характеристика связей с точки зрения взаимодействующих факторов. Если характеризуется связь между двух признаков, то ее принято называть парной, если изучаются более двух переменных – множественная. По силе связи различают слабые и сильные связи. Это формальная характеристика выражается конкретными величинами в соответствии с принятыми критериями силы связи.

Методы оценки тесноты связи подразделяются на корреляционные (параметрические) и непараметрические. Первые основаны на использовании оценок нормального распределения, в случаях, когда изучаемые величины подчиняются закону нормального распределения.

Непараметрические методы не накладывают ограничений на закон распределения изучаемых величин. Их преимущество – в простоте вычислений.

В пособие рассматриваются именно параметрические методы оценки, где для определения тесноты связи используется шкала Чеддока:

Показания тесноты связи	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99
Характер силы связи	слабая	заметная	умеренная	сильная	очень сильная

Контрольные вопросы:

- Дайте определение дисциплине «Эконометрика»
- Какие четыре составляющие определяют эконометрику как науку?
- Какие недостатки преодолеваются при использовании эконометрического метода?
- Что делает шкала Чеддока?
- Назовите и охарактеризуйте различные типы связей между явлениями.
- Какие проблемы решает эконометрическое исследование?
- В чем специфика параметрических методов оценивания?
- Перечислите этапы эконометрического исследования

1.2. Спецификация модели парной регрессии

Простая регрессия представляет собой регрессию между двумя переменными — y и x, т. е. модель вида:

$$y_x = f(x) \qquad (1.1),$$

где y — зависимая переменная (результативный признак);

x - независимая, или объясняющая переменная (признак-фактор).

Основные типы функций, используемые при количественной оценке связей:

$y=a+bx$ – линейная функция;
$y= a+b/x$ - гипербола;
$y=a+bx+cx^2$ – парабола;
$y=a+bx+cx^2+dx^3$ -полином 3 степени;
$y=ax^b$ – степенная функция;
$y=ab^x$ -показательная функция;
$y=a+b\lg x$ - логарифмическая функция;

Для изучения связей между признаками x и y строится **корреляционная таблица**, включающая группировку значений x и y, а также производные величины, необходимые для будущих вычислений и построения уравнения регрессии. Наглядным изображением корреляционной таблицы служит **корреляционное поле**. Оно представляет собой график, где на оси абсцисс откладываются значения x, по оси ординат – y. По расположению точек, их концентрации в определенном направлении можно судить о наличии и форме связи.

Любое эконометрическое исследование начинается со спецификации модели, т. е. с формулировки вида модели, исходя из соответствующей теории связи между переменными. Прежде всего из всего круга факторов, влияющих на результативный признак, необходимо выделить наиболее существенно влияющие факторы. Парная регрессия достаточна, если имеется доминирующий фактор, который и используется в качестве объясняющей переменной. Предположим, что выдвигается гипотеза о том, что урожайность картофеля находится в прямой зависимости от внесения удобрений под эту культуру, т. е. $y=a+bx$. В этом случае необходимо знать, какие остальные факторы предполагаются неизменными, возможно, в дальнейшем их придется учесть в модели и от простой регрессии перейти к множественной.

В уравнении регрессии корреляционная по сути связь признаков представляется в виде функциональной связи, выраженной соответствующей математической функцией. Практически в каждом отдельном случае величина у складывается из двух слагаемых:

$$y=y_x+\varepsilon \quad (1.2),$$

где у - фактическое значение результативного признака;

y_x - теоретическое значение результативного признака, найденное из уравнения регрессии;

ε - случайная величина, характеризующая отклонения реального значения результативного признака от теоретического, найденного по уравнению регрессии.

Случайная величина ε называется также возмущением. Она включает влияние не учтенных в модели факторов, случайных ошибок и особенностей измерения. Ее присутствие в модели порождено тремя источниками: спецификацией модели, выборочным характером исходных данных, особенностями измерения переменных.

К ошибкам спецификации относятся:

- **Невключение объясняющих переменных.** Соотношение между у и х почти наверняка является очень большим упрощением. Недоучет в уравнении регрессии какого-либо существенного фактора, т.е. использование парной регрессии вместо множественной, может привести к ошибке. Часто происходит так, что имеются переменные, которые мы хотели бы включить в регрессионное уравнение, но не можем этого сделать потому, что не знаем, как их измерить, например психологические факторы. Возможно, что существуют также другие факторы, которые мы можем измерить, но которые оказывают такое слабое влияние, что их не стоит учитывать. Кроме того, могут быть факторы, которые являются существенными, но которые мы из-за отсутствия опыта таковыми не считаем. Объединив все эти составляющие, мы получаем то, что обозначено как ε.

- **Агрегирование переменных.** Во многих случаях рассматриваемая зависимость — это попытка объединить вместе некоторое число микроэкономических соотношений. Например, функция суммарного потребления — это попытка общего выражения совокупности решений отдельных индивидов о расходах. Так как отдельные соотношения, вероятно, имеют разные параметры, любая попытка определить соотношение между совокупными расходами и доходом является лишь аппроксимацией. Наблюдаемое расхождение при этом приписывается наличию случайного члена.

- **Неправильное описание структуры модели.** Структура модели может быть описана неправильно или не вполне правильно. Здесь можно привести один из многих возможных примеров. Если зависимость относится к данным о временном ряде, то значение у может зависеть не от фактического значения x, а от значения, которое ожидалось в предыдущем периоде. Если ожидаемое и фактическое значения тесно связаны, то будет казаться, что между у и x существует зависимость, но это будет лишь аппроксимация, и расхождение вновь будет связано с наличием случайного члена.

- **Неправильная функциональная спецификация.** Функциональное соотношение между у и x математически может быть определено неправильно. Например, истинная зависимость может не являться линейной, а быть более сложной.

Наряду с ошибками спецификации могут иметь место **ошибки выборки**, поскольку исследователь чаще всего имеет дело с выборочными данными при установлении закономерной связи между признаками. Ошибки выборки возможны и в силу неоднородности данных в исходной статистической совокупности, что, как правило, бывает при изучении экономических процессов. Если совокупность неоднородна, то уравнение регрессии не имеет практического смысла. Для получения хорошего результата обычно исключают из совокупности единицы с аномальными значениями исследуемых признаков.

Наибольшую опасность в практическом использовании методов регрессионного анализа представляют **ошибки измерения**. Если ошибки спецификации можно уменьшить, изменяя форму модели, а ошибки выборки — увеличивая объем исходных данных, то ошибки измерения практически сводят на нет все усилия по количественной оценке связи между признаками.

Особенно велика роль ошибок измерения при исследовании на макроуровне. Так, в исследованиях спроса и потребления в качестве объясняющей переменной широко используется «доход на душу населения». Вместе с тем статистическое измерение величины дохода сопряжено с рядом трудностей и не лишено возможных ошибок, например в результате невозможности учесть «черную» зарплату.

Предполагая, что ошибки измерения и выборки сведены к минимуму, основное внимание в эконометрических исследованиях уделяется ошибкам спецификации модели, а именно выбору формы связи.

В парной регрессии выбор вида математической функции $y_x = f(x)$ может быть осуществлен тремя методами:

- **графическим;**
- **аналитическим**, т. е. исходя из теории изучаемой взаимосвязи;
- **экспериментальным**.

При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден и основан на поле корреляции. Определенное расположение точек в системе координат дает возможность для моделирования применить ту или иную кривую.

Значительный интерес представляет аналитический метод выбора типа уравнения регрессии. Он основан на изучении материальной природы связи исследуемых признаков.

Например, затраты предприятия могут быть подразделены на два вида:

- условно-переменные, изменяющиеся пропорционально изменению объема продукции (расход материала, оплата труда и др.) – bx;
- условно-постоянные, не изменяющиеся с изменением объема произ-

водства (арендная плата, содержание администрации и др.) - а.

Тогда зависимость всех затрат у на производство зерна х характеризуется линейной функцией:

$$y_x = a + bx \quad (1.3),$$

где х – валовой сбор зерна;

у – все производственные затраты.

Если теперь разделить обе части уравнения на величину собранного объема зерна х, то получим выражение зависимости себестоимости единицы продукции z=y/x от объема произведенного зерна х в виде уравнения равносторонней гиперболы: z = b+a/x.

При обработке информации на компьютере выбор вида уравнения регрессии обычно осуществляется экспериментальным методом, т. е. путем сравнения величины **остаточной дисперсии** $\sigma^2_{ост}$ рассчитанной при разных моделях.

Если уравнение регрессии проходит через все точки корреляционного поля, что возможно только при функциональной связи, когда все точки лежат на линии регрессии, то фактические значения результативного признака совпадают с теоретическими y_x т. е. они полностью обусловлены влиянием фактора х. В этом случае остаточная дисперсия $\sigma^2_{ост} = 0$. В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих не учитываемых в уравнении регрессии факторов. Иными словами, имеют место отклонения фактических данных от теоретических ($y — y_x$). Величина этих отклонений и лежит в основе расчета остаточной дисперсии:

$$\sigma^2{}_{ост} = \frac{\sum (y - y_x)^2}{n} \quad (1.4)$$

При обработке статистических данных на компьютере перебираются разные математические функции в автоматическом режиме, и выбирается та, для которой остаточная дисперсия является наименьшей.

Если остаточная дисперсия оказывается примерно одинаковой для нескольких функций, то на практике предпочтение отдается более простым видам функций, ибо они в большей степени поддаются интерпретации и требуют меньшего объема наблюдений.

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. Если вид функции усложняется, то требуется увеличение объема наблюдений, ибо каждый параметр при x должен рассчитываться хотя бы по 7 наблюдениям.

Контрольные вопросы:

- Что такое спецификация модели регрессии?
- Перечислите ошибки спецификации.
- Что такое случайная величина и теоретическое значение y?
- Какими способами осуществляется выбор формы связи?
- Как рассчитывается остаточная дисперсия?
- Охарактеризуйте аналитический метод выбора вида математической функции.
- Какое минимальное количество наблюдений необходимо для оценки параметров параболы?

Решение типовых задач

Пример 1. В результате моделирования на ПК зависимости спроса y на товар А от его цена x были получены теоретические значения y, найденные по двум уравнениям: линейное y_1 и гиперболы y_2 – таблица 1.3.

Таблица 1.3

№ наблю-дения	y	y_1	y_2	№ наблю-дения	y	y_1	y_2
1	4,3	4,1	4,3	8	11,1	10,7	11,2
2	6,2	6	6,1	9	9,3	9,1	9,4
3	3,2	3,1	3,3	10	13,4	13,8	13,1
4	12	12,2	12,1	11	3,2	3	2,8
5	8,8	8,9	8,7	12	4,4	4,1	4,6
6	4,7	4,8	4,8	13	5,9	5,8	5,7
7	6,9	7,3	7,1	14	8,2	8	7,9

Задание. Определить какая из математических функций больше подходит для данной модели.

Решение. Для ответа на вопрос найдем остаточную дисперсию для линейной и гиперболической функции по формуле:

$$\sigma^2{}_{ост} = \frac{\sum (y - y_x)^2}{n}$$

Для линейной $\sigma^2{}_{ост\ 1}$=0,061, а для гиперболы $\sigma^2{}_{ост\ 2}$=0,038. Так как $\sigma^2{}_{ост1}$ >$\sigma^2{}_{ост\ 2}$, то гиперболическое уравнение регрессии лучше описывает данную зависимость.

1.3. Линейная регрессия и корреляция

Линейная регрессия сводится к нахождению параметров уравнения вида:

$$y_x = a + b \cdot x \quad (1.5)$$

Уравнение вида (1.5) позволяет по заданным фактическим значениям фактора x иметь теоретические значения результативного признака.

Классический же подход к оцениванию параметров линейной регрессии основан на **методе наименьших квадратов** (МНК). МНК позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных (теоретических) y_x минимальна:

$$\sum_i (y_i - y_{x_i})^2 \to \min \text{ или } \sum_i \varepsilon_i^2 \to \min \qquad (1.6)$$

Чтобы найти минимум функции из формулы (1.6), надо вычислить частные производные по каждому из параметров a и b и приравнять их к нулю. Обозначим искомую функцию через f(x), тогда:

$$f(x) = \sum (y_i - \hat{y}_x)^2 = \sum (y - a - b \cdot x)^2; \qquad (1.7)$$

$$\begin{cases} \dfrac{df}{da} = -2\sum y + 2 \cdot n \cdot a + 2 \cdot b \sum x = 0; \\ \dfrac{df}{da} = -2 \cdot \sum y \cdot x + 2 \cdot a \sum x + 2 \cdot b \sum x^2 = 0. \end{cases} \qquad (1.8)$$

Преобразуя формулу (1.8), получим следующую систему нормальных уравнений для оценки параметров a и b:

$$\begin{cases} n \cdot a + b \sum x = \sum y, \\ a \sum x + b \sum x^2 = \sum y \cdot x. \end{cases} \qquad (1.9)$$

Решая систему нормальных уравнений (1.9) либо методом последовательного исключения переменных, либо методом определителей найдем искомые оценки параметров a и b. Можно воспользоваться следующими готовыми формулами:

$$a = \bar{y} - b \cdot \bar{x}, \qquad b = \frac{\operatorname{cov}(x, y)}{\sigma_x^2} = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2}. \qquad (1.10)$$

где $\operatorname{cov}(x, y)$ — ковариация признаков;

σ_x^2 — дисперсия признака x.

Параметры a и b методом определителей находятся так:

$$\Delta = \begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix},\quad \Delta_a = \begin{vmatrix} \sum y & \sum x \\ \sum yx & \sum x^2 \end{vmatrix},\quad \Delta_b = \begin{vmatrix} n & \sum y \\ \sum x & \sum yx \end{vmatrix}, \text{ откуда } \mathrm{a} = \frac{\Delta_\mathrm{a}}{\Delta},\quad \mathrm{b} = \frac{\Delta_\mathrm{b}}{\Delta} \quad (1.11).$$

Линейная регрессия находит широкое применение в эконометрике в виде четкой экономической интерпретации ее параметров. Параметр b называется **коэффициентом регрессии**. Его величина показывает среднее изменение результата с изменением фактора на одну единицу. Формально a — значение y при x = 0. Если признак-фактор x не имеет и не может иметь нулевого значения, то вышеуказанная трактовка свободного члена a не имеет смысла.

Присутствуют три особенности при интерпретации модели регрессии:

- a и b лишь оценки истинных значений параметров α и β;
- уравнение регрессии отражает лишь общую тенденцию для выборки (отдельное наблюдение подвержено воздействию случайностей);
- верность интерпретации зависит от правильности спецификации модели уравнения.

Уравнение регрессии всегда дополняется показателем, характеризующим степень тесноты связи между признаками. При использовании линейной регрессии в качестве такого показателя выступает **линейный коэффициент корреляции** r_{xy}. Можно пользоваться разными формулами для расчета, но результат всегда будет одинаков:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{yx} - \overline{y}\cdot\overline{x}}{\sigma_x\sigma_y} \qquad (1.12)$$

Линейный коэффициент корреляции находится в границах: $-1 \le r_{xy} \le 1$. Положительное значение коэффициента говорит о том, что связь между признаками прямая, отрицательное – обратная. Если коэффициент регрессии b> 0, то $0 \le r_{xy} \le 1$, и, наоборот, при b< 0, $-1 \le r_{xy} \le 0$. После нахождения линейного коэффициента корреляции качественная оценка тесноты связи дается с помощь шкалы Чеддока.

Следует иметь в виду, что величина линейного коэффициента корреляции оценивает тесноту связи рассматриваемых признаков в ее линейной форме. Поэтому близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствие связи между признаками. При иной спецификации модели связь между признаками может оказаться достаточно тесной.

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r^2_{yx}, называемый **коэффициентом детерминации**. Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака:

$$r^2_{yx} = \frac{\sigma^2_{y\,объясн}}{\sigma^2_{y\,общ}} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_{y\,общ}}, \tag{1.13}$$

$$где\, \sigma^2_{y\,объясн} = \frac{\sum (y_x - \bar{y})^2}{n}, \sigma^2_{y\,общ} = \frac{\sum (y - \bar{y})^2}{n}, \sigma^2_{ост} = \frac{\sum (y_x - y)^2}{n}$$

Соответственно величина $1 - r^2$ характеризует долю дисперсии y, вызванную влиянием остальных не учтенных в модели факторов. Можно показать, что коэффициент детерминации равен квадрату линейного коэффициента корреляции.

Коэффициент вариации случайной величины x (V_x.) – мера относительного разброса случайной величины. Показывает, какую долю среднего значения случайной величины составляет ее средний разброс.

$$V_x = \frac{\sigma_x}{\bar{x}} \cdot 100\%$$

Контрольные вопросы:

- Напишите уравнение линейной регрессии и дайте ее интерпретацию.
- С помощью какой системы можно найти параметры a и b в линей-

ной функции?

- Как рассчитывается коэффициент вариации и детерминации?
- Для чего нужен линейный коэффициент корреляции?
- Перечислите особенности интерпретации линейного уравнения регрессии.
- Какой знак у коэффициента b, если r>0 ?

Решение типовых задач

Пример 1. Предположим по группе сельхозпредприятий, специализирующих на производстве молока, рассматривается функция издержек y=a+bx (табл.1.1).

Задание. Построить уравнение регрессии, дать его экономическую интерпретацию. Оценить степень тесноты связи между y и x, рассчитать коэффициент вариации и детерминации.

Решение. Исходные данные и показатели для расчета оценок параметров a и b сгруппируем в таблице 1.1.

Таблица 1.1

№ предприятия	Уровень производства молока, тыс.тонн, x	Затраты на производство, млн. руб., y	yx	x^2	y^2	y_x
1	1	3	3	1	9	3,1
2	2	7	14	4	49	6,8
3	4	15	60	16	225	14,2
4	3	10	30	9	100	10,5
5	5	17	85	25	289	17,8
6	3	10	30	9	100	10,5
7	4	15	60	16	225	14,2
Итого	22	77	282	80	997	77,1

Система нормальных уравнений будет иметь вид:

$$\begin{cases} 7\cdot a + 22\cdot b = 77, \\ 22\cdot a + 80\cdot b = 282. \end{cases}$$

Решая ее, получим: a = -0,58; b = 3,68. Этот же результат можно было получить, используя уже готовые формулы для a и b. Получим следующее уравнение регрессии: y_x = -0,58 + 3,68x.

Подставив в уравнение значения x, найдем теоретические значения y_x. В данном случае величина параметра a не имеет экономического смысла. В рассматриваемом примере имеем: $\bar{x} = 3{,}14; \bar{y} = 11{,}0$. Среднеквадратическое отклонение σ и **коэффициент вариации** V для x и y будут:

$$\sigma_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} = 1{,}25; \quad \sigma_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}} = 4{,}63$$

$$V_x = \frac{\sigma_x}{\bar{x}} \cdot 100\% = 39{,}8; \ V_y = 42{,}1\%$$

Это означает, что колеблемость показателя производства молока достаточно высокая возле среднего уровня.

Найдем линейный коэффициент корреляции:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = 3{,}68\frac{1{,}25}{4{,}63} = 0{,}994.$$

Полученное значение достаточно близко к 1 и по шкале Чеддока означает, что связь между изучаемыми признаками очень сильная. Таким образом, в нашем примере наблюдается очень тесная зависимость между производством молочной продукции и затратами на ее получение.

Коэффициент детерминации равен 0,987. Следовательно, уравнением регрессии объясняется 98,7% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,3% ее дисперсии (т.е. остаточная дисперсия). Чем больше доля объясненной вариации, тем соответственно меньше роль прочих факторов, и, следовательно, линейная модель хорошо аппроксимирует исходные данные, ей можно воспользоваться для прогноза значений результативного признака.

Так, полагая, что объем молочной продукции предприятия может составить 6 тыс.тонн, прогнозное значение для издержек производства окажется 21,5 млн. руб. (для расчета используется полученное уравнение).

Пример 2. Известно уравнение тренда, определяющего изменение среднегодовых надоев молока от одной коровы в кг по предприятию за 1999-2008г.: y_t = 2130 – 46t, где t=0,1 и т.д. По нему определены теоретические значения признака (табл.1.2).

Задание. Дайте интерпретацию параметров уравнения регрессии и вычислите коэффициент детерминации.

Решение. Коэффициент регрессии показывает, что ежегодно происходит абсолютное снижение надоев молока от одной коровы в среднем на 46 кг. Параметр a=2130 характеризует уровень надоев молока за 1999 год. Для вычисления коэффициента детерминации определяем производные величины (табл.1.2).

Таблица 1.2

годы	y	y_x	$y-y_x$	$(y-y_x)^2$	$y-\bar{y}$	$(y-\bar{y})^2$
1991	2124	2130	-6	36	201	40401
1992	2071	2084	-13	169	148	21904
1993	2040	2038	2	4	117	13689
1994	1996	1992	4	16	73	5329
1995	1942	1946	-4	16	19	361
1996	1908	1900	8	64	-15	225
1997	1849	1854	-5	25	-74	5476
1998	1825	1808	17	289	-98	9604
1999	1754	1762	-8	64	-169	28561
2000	2124	2130	5	25	-202	40804
Итого (в среднем)	1923	1923	0	708	0	166354

Таким образом, коэффициент детерминации равен:

$$r^2_{yx} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_{yобщ}} = 1 - \frac{708}{166354} = 0{,}996$$

Мы получили, что 99,6% вариации результативного признака (уровень надоев молока от одной коровы) объясняется полученным уравнением регрессии, т.е. фактором времени. Только на 0,4% оставшейся вариации y приходятся прочие неучтенные в модели факторы.

1.4. Нелинейная регрессия и корреляция

Связь между признаками не всегда можно полностью проявить на основе прямолинейного уравнения связи. Взаимодействие между показателями, характеризующими определенные процессы, нередко имеет более сложный характер, чем просто пропорциональные зависимости. Например, биологическим процессам присущи нелинейные формы связи. Рост и развитие растений, как правило, развиваются нелинейно. Известно, что продуктивность коров в зависимости от числа отелов вначале имеет тенденцию к росту, а затем после достижения максимума начинает закономерно снижаться. Поэтому для характеристики всего многообразия зависимостей в сельском хозяйстве нельзя использовать только один вид уравнения.

Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций. Различают **два класса** нелинейных регрессий:

• регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам:

полиномы разных степеней $y_x=a+b_1x+b_2x^2+\ldots+b_nx^n$,

гипербола $y_x=a+b/x$.

• регрессии, нелинейные по оцениваемым параметрам:

степенная $y=ax^b$, показательная $y_x= ab^x$, экспоненциальная $y_x=e^{ax+b}$.

Нелинейная регрессия по включенным переменным не таит каких-либо сложностей в оценке ее параметров. Она определяется, как и в линейной регрессии, методом наименьших квадратов (МНК), ибо эти функции линейны по параметрам. Например, в параболе $y_x=a+b_1x+b_2x^2$ заменяя переменные $x=x_1$, $x^2=x_2$, получим двухфакторное уравнение линейной регрессии, для оценки параметров которого используется МНК:

$$\begin{cases} n\cdot a+b_1\sum x_1+b_2\cdot\sum x_2=\sum y, \\ a\cdot\sum x_1+b_1\cdot\sum x_1^2+b_2\cdot\sum x_2\cdot x_1=\sum y\cdot x_1, \\ a\cdot\sum x_2+b_1\cdot\sum x_1\cdot x_2+b_2\cdot\sum x_2^2=\sum y\cdot x_2. \end{cases} \quad (1.14.)$$

Следовательно, полином любого порядка сводится к линейной регрессии с ее методами оценивания параметров и проверки гипотез. Как показывает опыт большинства исследователей, среди нелинейной полиномиальной регрессии чаще всего используется парабола второй степени; в отдельных случаях — полином третьего порядка. Ограничения 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.

Для параметров нелинейных уравнений регрессий экономическая интерпретация резко усложняется. Увеличение количества внесенных удобрений приводит, при прочих равных условиях, к росту урожайности, но чрезмерное внесение их без изменения других элементов к дальнейшему повышению урожайности не приводит, а, наоборот, снижает ее. При изучении зависимости заработной платы работников физического труда от возраста — с увеличением возраста повышается заработная плата ввиду одновременного увеличения опыта и повышения квалификации работника. Однако с определенного возраста ввиду старения организма и снижения производительности труда дальнейшее повышение возраста может приводить к снижению заработной платы работника.

Параметр b_2 в параболе $y=a+b_1x_1+b_2x^2$ характеризует степень ускорения или замедления кривизны параболы. При $b_2 > 0$ парабола имеет минимум, а при $b_2 < 0$ — максимум. Параметр b_1 характеризует угол наклона кривой, а параметр а — начало кривой.

Среди класса нелинейных функций, параметры которых без особых затруднений оцениваются МНК, следует назвать хорошо известную в эконометрике **равностороннюю гиперболу** $y_x = a+b/x$. Она может быть использована не только для характеристики связи удельных расходов сырья, материалов, топлива с объемом выпускаемой продукции, но и времени обращения товаров от величины товарооборота, т.е. и на микроуровне, и на макроуровне.

Классическим ее примером является кривая Филлипса, характеризующая нелинейное соотношение между нормой безработицы x и процентом прироста заработной платы y: y_x =0,007+0,18/x. Полученное уравнение регрессии показывает наличие обратной зависимости между процентом прироста заработной платы и уровнем безработицы. Величина параметра а, равная 0,007, означает, что с ростом уровня безработицы темп прироста заработной платы в пределе стремится к нулю. Соответственно можно определить тот уровень безработицы, при котором заработная плата оказывается стабильной и темп ее прироста равен нулю. При b< 0 имеем медленно повышающуюся функцию с верхней асимптотой, т. е. с максимальным предельным уровнем y, оценку которого в уравнении дает параметр а.

Иначе обстоит дело с регрессией, **нелинейной по оцениваемым параметрам**. Данный класс нелинейных моделей подразделяется на два типа: нелинейные модели **внутренне линейные** и нелинейные модели **внутренне нелинейные**. Если нелинейная модель внутренне линейна, то она с помощью соответствующих преобразований может быть приведена к линейному виду. Если же нелинейная модель внутренне нелинейна, то она не может быть сведена к линейной функции.

В ряде случаев для проведения качественного анализа уравнение линейной связи целесообразно заменить логарифмическим, коэффициенты которого дадут относительную оценку связей между изучаемыми показателями. К такому логарифмическому уравнению приводится уравнение степенной функции, являющейся **внутренне линейной:** $y=ax^b$.

Данная модель нелинейна относительно оцениваемых параметров, однако ее можно считать внутренне линейной, ибо **логарифмирование** данного уравнения по основанию e (или по любому другому) приводит его к линейному виду:

lny=lna+b lnx, после соответствующей замены У=A+bX.

Соответственно оценки параметров a и b могут быть найдены также МНК:

$$\begin{cases} n\cdot\ln a + b\sum\ln x = \sum\ln y \\ \ln a\cdot\sum\ln x \ + b\cdot\sum(\ln x)^2 = \sum\ln y\cdot\ln x, \end{cases} \qquad (1.15)$$

В рассматриваемой степенной функции предполагается, что случайная ошибка мультипликативно связана с объясняющей переменной х, т.е. $y_x = ax^b\varepsilon$. Если же модель представить в виде $y_x = ax^b + \varepsilon$, то она становится внутренне нелинейной, ибо ее невозможно превратить в линейный вид. Также внутренне нелинейными будут и модели вида: $y=a+bx^c$, $y=a+1/x^b$.

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

Широкое применение степенной функции связано еще и с тем, что параметр b в ней имеет четкое экономическое истолкование, т. е. он является **коэффициентом эластичности**. Это значит, что величина коэффициента b показывает, на сколько процентов изменится в среднем результат, если фактор изменится на 1 %.

О правомерности подобного истолкования параметра b для степенной функции можно судить, если рассмотреть общую формулу расчета коэффициента эластичности:

$$Э = f^{'}(x)\cdot\frac{x}{y} \quad , (1.16)$$

где f ’ (x) — первая производная, характеризующая соотношение приростов результата и фактора для соответствующей формы связи.

Для степенной функции $f^{`}(x)=abx^{b-1}$, откуда $Э= abx^{b-1}x/(ax^b)=b$.

Коэффициент эластичности, естественно, можно определять и при наличии других форм связи, но только для степенной функции он представляет собой постоянную величину, равную параметру b. В других функциях коэффициент эластичности зависит от значений фактора х.

В силу того, что коэффициент эластичности для линейной функции не является величиной постоянной, а зависит от соответствующего значения х, то часто рассчитываются **средний показатель эластичности**:

$$Э = f'(x) \cdot \frac{\bar{x}}{\bar{y}} \qquad (1.17)$$

где $\bar{x}$ и $\bar{y}$ – средние значения фактора и результата.

В моделях, нелинейных по оцениваемым параметрам, но приводимых к линейному виду МНК применяется к преобразованным уравнениям. Если в линейной модели и моделях, нелинейных по переменным, при оценке параметров исходят из критерия $\Sigma(y\text{-}y_x)^2$ —> rnin, то в моделях, нелинейных по оцениваемым параметрам, требование МНК применяется не к исходным данным результативного признака, а к их преобразованным величинам, т. е. к lny, 1/y и т.д. Вследствие этого оценка параметров для линеаризуемых функций методом МНК оказываются несколько смещенной (обычно заниженной).

Тесноту связи в уравнении нелинейной регрессии определяют **индексом корреляции** (R):

$$R = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}} = \sqrt{\frac{\sigma^2_{факт}}{\sigma^2_y}} \qquad (1.18)$$

Величина данного показателя находится в границах: $0<R<1$, чем ближе к единице, тем теснее связь рассматриваемых признаков, тем более надежно найденное уравнение регрессии.

Если же нелинейное относительно объясняемой переменной уравнение регрессии при линеаризации принимает форму линейного уравнения парной регрессии, то для оценки тесноты связи может быть использован линейный коэффициент корреляции, величина которого в этом случае совпадет с индексом корреляции $R_{yx} = r_{yz}$ где z — преобразованная величина признака-фактора, например z=1/x в уравнении гиперболы.

Иначе обстоит дело, когда преобразования уравнения в линейную форму связаны с зависимой переменной y. В этом случае линейный коэффициент корреляции по преобразованным значениям признаков дает лишь приближенную оценку тесноты связи и численно не совпадает с индексом корреляции. Так, для степенной функции после перехода к логарифмически линейному уравнению ln y = ln a + b lnx может быть найден линейный коэффициент корреляции не для фактических значений переменных x и y, а для их логарифмов. Соответственно квадрат его значения будет характеризовать отношение факторной суммы квадратов отклонений к общей, но не для y, а для его логарифмов:

$$r^2_{\ln y \ln x} = 1 - \frac{\sum(\ln y - \ln y_x)^2}{\sum(\ln y - \ln \bar{y})^2} \qquad (1.19)$$

где $\ln y_x$ – расчетные значения, вычисленные по логарифмическому уравнению; $\ln \bar{y}$ – среднее значение логарифма y.

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.

Между тем при расчете индекса корреляции используются суммы квадратов отклонений признака y, а не их логарифмов, С этой целью определяются теоретические значения результативного признака, т. е. y_x как антилогарифм $\ln y_x$ После чего истинный индекс корреляции можно определить по формуле:

$$R_{yx} = \sqrt{1 - \frac{\sum(y - anti\log(\ln y_x))^2}{\sum(y - \bar{y})^2}} \qquad (1.20)$$

Поскольку в расчете индекса корреляции используется соотношение факторной и общей суммы квадратов отклонений, то R^2 имеет тот же смысл, что и r^2 коэффициент детерминации. Иногда его называют индексом детерминации.

Контрольные вопросы:

- Назовите классы нелинейных регрессий и охарактеризуйте их.
- К какому классу нелинейных регрессии относится показательная функция?
- Как рассчитывается коэффициент эластичности?
- Для какого уравнения регрессии коэффициент эластичности является постоянным числом?
- Что такое средний коэффициент эластичности
- Приведите пример внутренне нелинейной регрессии.
- Что такое индекс корреляции и детерминации?
- Как связаны линейный коэффициент корреляции и индекс корреляции?
- Как оценить степень тесноты связи между признаками в моделях внутренне линейных?

Решение типовых задач

Пример 1. Имеются данные об урожайности культуры и себестоимости 1ц, представленные в таблице 1.3.

Задание.

1.Выбрать уравнение регрессии для изучаемой связи и построить модель.

2. Рассчитать средний коэффициент эластичности.

3. Сделать точечный прогноз уровня себестоимости 1 ц при плановой урожайности зерновых 15 ц/га.

Решение.

Таблица 1.3.

№	себестоимость 1ц, руб, x_0	урожайность ц с 1 га, x_1	расчетные величины			
			$z=1/x_0$	zx_1	x_1^2	z^2
1	22,49	9,7	0,0445	0,431	94,09	0,0020
2	20,71	10,5	0,0483	0,507	110,25	0,0023
3	18,83	11,6	0,0531	0,616	134,56	0,0028
4	17,26	12,1	0,0579	0,701	146,41	0,0034
5	18,10	12,2	0,0552	0,674	148,84	0,0031
6	16,33	13,2	0,0612	0,808	174,24	0,0037
7	15,18	13,4	0,0659	0,883	179,56	0,0043

8	16,30	13,8	0,0613	0,847	190,44	0,0038
9	13,30	16,2	0,0752	1,218	262,44	0,0057
10	12,28	17,1	0,0814	1,393	292,41	0,0066
11	11,53	17,6	0,0867	1,526	309,76	0,0075
12	10,70	18,5	0,0935	1,729	342,25	0,0087
13	8,81	22,1	0,1135	2,509	488,41	0,0129
Итого	201,85	188	0,8978	13,842	2873,66	0,0668

При анализе нелинейных связей между признаками особое значение приобретают выбор и обоснование типа уравнения, которое наиболее полно отразит имеющуюся связь. Себестоимость и урожайность – два показателя, имеющие устойчивые связи, изменяются в большинстве случаев нелинейно. Часть затрат на 1 га посева является относительно устойчивой (постоянные затраты), другая часть затрат, обеспечивает повышение урожайности (удобрения, обработка посевов и др. – переменные затраты).

Изменение сочетаний между компонентами затрат, различных по воздействию на величину урожайности, являются причиной неравномерного изменения себестоимости по сравнению с изменениями урожайности. Таким образом, форма связи изменения себестоимости в зависимости от меняющейся урожайности позволяет предполагать, что подходящим может быть следующее математическое уравнение: $x_0 = \frac{1}{a + bx_1}$

Рассмотрим предложенное выше уравнение. Предварительно проведем преобразования, т.е. $x_0 = \frac{1}{a + bx_1}$, заменим $\frac{1}{x_0} = a + bx_1$.

Для упрощения обозначим $\frac{1}{x_0} = z$. тогда уравнение можно записать как $z=a+bx_1$.

Фактически получено уравнение прямолинейной зависимости, для решения которого составим систему из двух нормальных уравнений:

$$\begin{cases} \sum z = na + b\sum x_1 \\ \sum zx_1 = a\sum x_1 + b\sum x_1^2 \end{cases}$$

В таблице 1.3. определим $z = \frac{1}{x_0}$, $\sum z$, $\sum x_0$, $\sum x_1$, $\sum x_1^2$, а также и $\sum z^2$.

Подставим полученные величины в уравнения:

$$\begin{cases} 0{,}8978=13a+188b \\ 13{,}842=188a+2873{,}66 \end{cases}$$

После решения системы получим z= -0,011+0,0055x_1, но поскольку z=1/x_0, а соответственно x_0=1/z, то в конечном виде уравнение, выражающее связь межу урожайностью и себестоимостью, будет следующим:

$$x_0 = \frac{1}{a+bx_1} = \frac{1}{-0{,}011+0{,}0055x_1}$$

Следовательно, повышение урожайности зерновых на x_1 единицу приведет к снижению себестоимости на $\frac{1}{-0{,}011+0{,}0055x_1}$ руб.

В отличие от прямолинейной зависимости здесь уже нельзя указать, каким будет снижение себестоимости на 1 ц прибавки урожайности, так как темп снижения себестоимости при различном уровне урожайности будет неодинаковым. Например, при урожайности x_1=9,5 ц с га и x_2=10,5 ц с га, вычисленная по уравнению себестоимость составит:

при x_1=9,5 ц с га x_0=1/(-0,011+0,0055*9,5)=27,40 руб.

при x_2=10,5 ц с га x_0=1/(-0,011+0,0055*10,5)=21,16 руб.

То есть снижение себестоимости при росте урожайности на 1 ц (с 9,5 до 10,5 ц с га) составит 6 руб. 24 коп. (27,40-21,16).

Для нахождения среднего коэффициента эластичности определяем средние значения признаков и производную f $'$(x_1) гиперболы: x_0=15,52, x_1=14,46,

$$f''(x_1) = -\frac{1}{(a+bx_1)^2} \cdot b$$

Теперь вычислим средний коэффициент эластичности для нашего уравнения: $Э = f'(x_1) \cdot \frac{\bar{x}_1}{\bar{x}_0} = \frac{-1}{-0{,}011+0{,}0055 \cdot 14{,}46} \cdot 0{,}0055 \cdot \frac{14{,}46}{15{,}52} = -0{,}075$

Это означает, что с увеличением урожайности зерновых на 1% от его среднего уровня себестоимость 1 ц продукции снижается на 0,075% от своего среднего уровня.

Для определения прогнозного значения себестоимости 1 ц зерновых следует предполагаемое значение урожайности подставить в уравнение регрессии: при x_1=15 ц с га x_0=1/(-0,011+0,0055*9,5)=14 руб.

Таким образом, прогнозное значение себестоимости 1 ц продукции при плановой урожайности 15 ц/га составит 14 рублей.

1.5. Оценка надежности результатов парной регрессии и корреляции. Прогнозирование по линейному уравнению.

Статистические критерии — это компактные формулировки правил проверки достоверности выводов анализа или правильности выдвигаемых гипотез. Они позволяют вместо субъективных оценок использовать объективные количественные характеристики. Дело в том, что совершенно очевидные, на первый взгляд, субъективные по своей природе оценки экономических ситуаций и процессов не всегда совпадают с истинным положением дел.

Забывая о возможности влияния случайных причин, многие исследователи даже самые незначительные изменения показателей относят на счет влияния изучаемых факторов. Исключить субъективные ошибочные выводы позволяет аппарат статистических гипотез.

В процессе определения значимости параметров регрессии и корреляции обычно происходит формулировка двух гипотез: **нулевой** H_0 и **альтернативной** H_1.

Нулевая гипотеза (H_0) — гипотеза, проверяемая с помощью тех или иных критериев. Альтернативная гипотеза (H_1) — гипотеза, противопоставляемая нулевой, которая может оказаться верной, если H_0 будет отвергнута.

Выбор того, следует отвергнуть или принять нулевую гипотезу, зависит от **уровня значимости** α, при котором она рассматривается. В большинстве работ по экономике за критический уровень значимости берется 5 или 1%. Исследователи обычно представляют свои результаты при двух уровнях значимости в 5 и 1%, а не ограничиваются только одним уровнем. Причина заключается в том, что обычно делается попытка найти баланс между риском допущения ошибок I и II рода. Ошибка I рода имеет место в том случае, когда вы отвергаете истинную нулевую гипотезу. Ошибка II рода возникает, когда вы не отвергаете ложную гипотезу.

Проблема, как избежать ошибок I и II рода, известна всем. Типичным примером этого является расследование уголовного преступления.

Если за нулевую гипотезу принять вариант, что подсудимый невиновен, то ошибка I рода происходит, когда суд присяжных признает его виновным. Ошибка II рода имеет место в том случае, когда суд присяжных ошибочно оправдывает виновного подсудимого.

Большинство людей выбирают достаточно простую форму обеспечения гарантий и осуществляют проверку на обоих уровнях значимости, представляя результаты каждой такой проверки. На самом деле часто нет необходимости непосредственно ссылаться на оба результата. Если вы отклоняете гипотезу H_0 при однопроцентном уровне, то из этого автоматически следует, что вы отклоните ее и при уровне значимости в 5%, и нет необходимости упоминать об этом. Если же вы не отвергаете гипотезу при уровне значимости в 5%, то из этого автоматически следует, что вы не отвергнете ее и при однопроцентном уровне значимости, и вновь нет смысла об этом говорить. Только в одном случае вы должны представить оба результата: если гипотеза отвергается на 5-процентном, но не на однопроцентном уровне значимости.

В последнем случае можно либо пойти на риск принятия гипотезы, либо взять гипотезу под сомнение. В такой ситуации следует признать целесообразным провести повторный эксперимент для получения данных, на основании которых можно было бы сделать более определенные выводы.

С помощью корреляционно-регрессионного анализа можно получить оценки параметров уравнений зависимостей. Однако они являются лишь оценками. Поэтому возникает вопрос о том, насколько они надежны. В качестве оценки статистической значимости результатов парной регрессии и корреляции могут выступать:

- **показатель детерминации** уравнения регрессии;
- значения **F-критерия Фишера** и **t-критерия Стьюдента**;
- **средняя ошибка аппроксимации**.

Величина r^2 коэффициента детерминации служит одним из критериев оценки качества линейной модели (см.1.3). Чем больше доля объясненной вариации, тем соответственно меньше роль прочих факторов, и, следовательно, линейная модель хорошо аппроксимирует исходные данные, ей можно воспользоваться для прогноза значений результативного признака.

Оценка значимости уравнения регрессии в целом дается с помощью F-критерия Фишера. При этом выдвигается нулевая гипотеза H_0, что коэффициент регрессии равен нулю, т. е. b=0, и, следовательно, фактор x не оказывает влияния на результат.

Непосредственному расчету F-критерия предшествует **анализ дисперсии**. Центральное место в нем занимает разложение общей суммы квадратов отклонений переменной от среднего значения y на две части — «**объясненную**» и «**необъясненную**»:

$$\sum(y-\bar{y})^2 = \sum(y_x-\bar{y})^2 + \sum(y_x-y)^2 \quad (1.21)$$

общая сумма квадратов отклонений | сумма квадратов отклонений, объясненная регрессией | остаточная сумма квадратов отклонений

Очевидно, что если сумма квадратов отклонений, обусловленная регрессией, будет больше остаточной суммы квадратов, то уравнение регрессии статистически значимо и фактор оказывает существенное воздействие на результат. Это равносильно тому, что коэффициент детерминации будет приближаться к единице.

Факторную дисперсию можно рассчитать и без теоретических значений y_x: $\sigma^2_{факт} = \frac{1}{n}\cdot(a\sum y + b\sum x\cdot y) - y^2$ для уравнения $y_x = a + bx$

Любая сумма квадратов отклонений связана с **числом степеней свободы** (df), т. е. с числом свободы независимого варьирования признака. Число степеней свободы связано с числом единиц совокупности n и с числом определяемых по ней констант. При расчете объясненной или факторной суммы квадратов на нее приходится **df=m** степеней свободы, а именно, величина равная количеству определяемых параметров при переменных x, а оставшиеся степени **df=n-1-m** приходятся на остаточную сумму квадратов. Например, для уравнения параболы $y_x=a+bx+cx^2$ для факторной суммы m=2, остаточное число степеней n-3.

Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим F-критерий:

$$F_{факт} = \frac{S^2_{факт}}{S^2_{ост}} = \frac{\sum (y_x - \bar{y})^2}{\sum (y_x - y)^2} \cdot \frac{n-m-1}{m} \quad (1.22.)$$

где $F_{факт}$ - критерий Фишера для проверки нулевой гипотезы H_o.

Если нулевая гипотеза справедлива, то факторная и остаточная дисперсии не отличаются друг от друга. Для опровержения гипотезы необходимо, чтобы факторная дисперсия превышала остаточную в несколько раз. Разработанные таблицы **критических значений $F_{кр}$** (приложение А) при разных уровнях существенности нулевой гипотезы α и различном числе степеней свободы ($к_1$=m и k_2=n-m-1), дают возможности проверки значимости уравнения регрессии.

Если $F_{факт} > F_{кр}$, то H_0 отклоняется и уравнение считается статистически значимым в целом при соответствующем уровне значимости α. Если $F_{факт} < F_{кр}$, то H_0 не может быть отклонена без серьезного риска сделать неправильный вывод о наличии связи. В этом случае с вероятностью p=1-α уравнение регрессии считается статистически незначимым и, скорее всего, связь между признаками случайна.

Величина F-критерия связана с показателем детерминации. Можно показать, что значение F-критерия равно:

$$F = \frac{r^2}{1-r^2} \cdot \frac{n-m-1}{m} \quad \text{или} \quad F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}, \quad (1.23.)$$

где r^2 – линейный коэффициент детерминации;

R^2 – индекс детерминации в нелинейной регрессии.

В линейной регрессии обычно оценивается значимость не только уравнения в целом, но и отдельных его параметров.

При оценке адекватности линейной регрессии с одновременным построением **доверительного интервала** для параметров a и b исходят из того, что эти коэффициенты вычислены стандартным методом МНК. С этой целью по каждому из параметров определяется его **стандартная ошибка**. Исходя из логики построения линейной модели видно, что стандартная ошибка коэффициента регрессии b прямо пропорциональна дисперсии случайной величины (остаточная дисперсия), обратно пропорциональна числу наблюдений и дисперсии x. С учетом того, что для линейной функции m=1, получим формулу стандартной ошибки для b:

$$m_b = \sqrt{\frac{\sum(y_x - y)^2/(n-2)}{n\cdot\sum(x-\bar{x})^2/n}} = \sqrt{\frac{S^2_{ост}}{\sum(x-\bar{x})^2}} \quad (1.24.)$$

где $S^2_{ост}$ - остаточная дисперсия на одну степень свободы.

Стандартная ошибка параметра a определяется по формуле:

$$m_a = \sqrt{\frac{S^2_{ост}\cdot\bar{x}^2}{\sum(x-\bar{x})^2}} \quad (1.25.)$$

Значимость линейного коэффициента корреляции проверяется на основе величины ошибки коэффициента корреляции m_r:

$$m_r = \sqrt{\frac{1-r^2}{n-2}} \quad (1.26.)$$

Величина стандартной ошибки совместно с t-распределением Стьюдента при n-2 степенях свободы применяется для проверки существенности коэффициента регрессии и для расчета его доверительных интервалов. Для оценки существенности параметров и коэффициента корреляции их значения сравниваются со стандартной ошибкой, т. е. определяется фактическое значение t-критерия Стьюдента:

$$t_b = \frac{b}{m_b}, t_a = \frac{a}{m_a}, t_r = \frac{r}{m_r} \quad (1.27.)$$

Фактическое значение критерия сравнивается с **табличным** (критическим) при определенном уровне значимости α и числе степеней свободы n-2 (для нелинейных функций df=n-m-1). Если $t_{факт} > t_{кр}$, то гипотезу о не существенности параметра регрессии и равенстве его нулю можно отклонить, а сам параметр является статистически значимым. Если $t_{факт} < t_{кр}$, то нулевая гипотеза H_0 принимается с вероятностью $1\text{-}\alpha$ и параметр статистически незначим.

Критические значения t-критерия Стьюдента используются для построения доверительных интервалов для параметров регрессии и корреляции. Это позволяет определить с определенной степенью точности, в каких пределах находятся истинные значения этих величин. Так, например, истинное значение коэффициента регрессии β определяется как:

$$b - t_{кр}m_b \le \beta \le b + t_{кр}m_b \quad (1.28.)$$

Аналогичные рассуждения можно провести и для параметра а и коэффициента корреляции r. Отметим также, что, так как значение $t_{кр}$ зависит от выбора уровня значимости α, границы доверительного интервала также зависят от этого выбора. Если принимается 5-процентный уровень значимости, то соответствующим доверительным интервалом считается 95-процентный интервал.

Поскольку коэффициент регрессии в эконометрических исследованиях имеет четкую экономическую интерпретацию, то доверительные границы интервала для коэффициента регрессии не должны содержать противоречивых результатов, например, $-10 < b < 40$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего быть не может, а значит оценка такого параметра ненадежна.

В парной линейной регрессии между t-критерием Стьюдента и F-критерием Фишера существует взаимосвязь:

$$F = t^2_b = t^2_r \quad (1.29.)$$

Таким образом, проверка гипотез о значимости коэффициентов регрессии и корреляции равносильна проверке гипотезы о существенности линейного уравнения регрессии в целом.

В качестве еще одного критерия оценки адекватности построения модели регрессии выступает **средняя ошибка аппроксимации**. Величина отклонений фактических и расчетных значений результативного признака (y-y_x) по каждому наблюдению представляет собой ошибку аппроксимации. Отклонения (y-y_x) несравнимы между собой. Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации как среднюю арифметическую простую:

$$\bar{A} = \frac{1}{n} \cdot \sum \left| \frac{y - y_x}{y} \right| \cdot 100\% \quad (1.30.)$$

На практике предполагают, что если средняя ошибка аппроксимации не превосходит 8-10%, то качество модели оценивается как хорошее.

Прогнозирование по линейному уравнению. В большинстве случаев после построения уравнения регрессии требуется провести прогноз результативного показателя при определенных значениях x. Прогнозное значение y_x получается путем подстановки в уравнение регрессии значения x и вычисления ошибки прогноза (доверительный интервал). Просто точечный прогноз не дает полного представления о прогнозируемом уровне показателя. Поэтому он дополняется расчетом стандартной ошибки m_{Yx} , и соответственно интервальной оценкой истинного прогнозного значения $\bar{y}_{x\ пр}$:

$$y_x - t_{кр} m_{Yx} \le y_{x\ пр} \le y_x + t_{кр} m_{Yx} \quad (1.31)$$

Стандартная ошибка прогноза для линейного уравнения регрессии зависит от остаточной дисперсии, приходящейся на одну степень свободы, дисперсии x и насколько прогнозное значение x отклоняется от среднего значения. **Величина** стандартной ошибки достигает **минимума** при прогнозном значении $x_{пр} = \bar{x}$ и возрастает по мере того, как «удаляется» от среднего значения $x_{пр}$ в любом направлении. Можно ожидать наилучшие результаты прогноза, если признак-фактор x находится в центре области наблюдений и нельзя ожидать хороших результатов прогноза при удалении $x_{пр}$ от $\bar{x}$. Таким образом, ошибка m_{Yx} рассчитывается как:

$$m_{Yx} = S_{ост}\sqrt{1+\frac{1}{n}+\frac{(x_{np}-\bar{x})^2}{\sum(x-\bar{x})^2}},$$

$$S_{ост} = \sqrt{\frac{\sum(y-y_x)^2}{n-2}} \qquad (1.32)$$

Прогнозируемый интервал, рассчитываемый по формуле (1.31.) зависит от t-Стьюдента, т.е. от уровня значимости α. Чем больше вероятность с которой мы делаем наш вывод, тем шире будет доверительный интервал, содержащий истинное значение прогнозируемого показателя y_x.

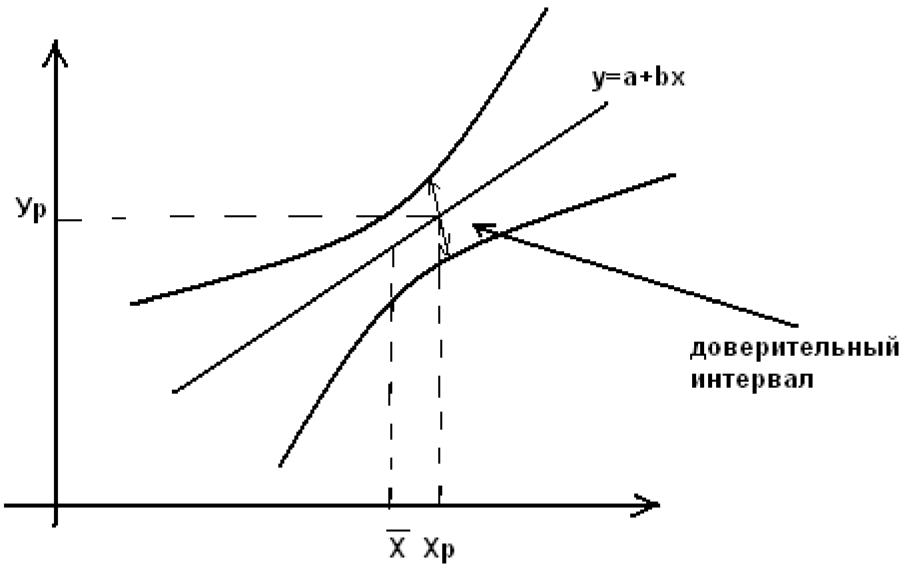


Рис.1. Доверительный интервал прогнозируемого значения линейной регрессии:

При прогнозировании на основе уравнения регрессии следует помнить, что величина прогноза зависит не только от стандартной ошибки индивидуального значения y, но и от точности прогноза значения фактора x. Его величина может задаваться на основе анализа других моделей исходя из конкретной ситуации, а также из анализа динамики данного фактора.

Контрольные вопросы:

- Назовите основные показатели, определяющие качество и надежность уравнения регрессии.
- Для чего используется F-критерий Фишера?
- Чем отличаются гипотезы H_0 и H_1?
- Для чего используется t-критерий Стьюдента?
- Какая взаимосвязь между F-критерием Фишера и t-критерием Стьюдента?
- Если средняя ошибка аппроксимации получилась 15%, что можно сказать о качестве модели?
- Как определяется доверительный интервала для параметра уравнения регрессии?
- Что такое уровень значимости α?

Решение типовых задач

Пример 1. По сельскохозяйственным предприятиям региона известны данные за 2008 г.

Номер хозяйства	Расход кормов на корову за год, ц. корм.ед., x	Удой молока от коровы за год, кг, y
1	39	2660
2	41	2960
3	43,5	2680
4	39,5	3080
5	44,5	3240
6	53	3900
7	33,5	2780
8	44	3160
9	36,5	3040
10	43,5	3240
11	38	3180
12	57,5	3460

Задание.

1. Построить линейное уравнение парной регрессии y от x.
2. Рассчитать линейный коэффициент парной корреляции и среднюю ошибку аппроксимации.
3. Оценить статистическую значимость параметров регрессии и корреляции.

4. Построить доверительный интервал для коэффициента регрессии с вероятностью 95%.

Решение. Для расчета параметров уравнения линейной регрессии строим расчетную таблицу 1.4.

Таблица 1.4.

№	y	x	xy	x^2	y^2	y_x	$\|y-y_x\|$	$\|(y-y_x)/y\|$
1	2660	39,0	103740	1521	7075600	2975	315	0,119
2	2960	41,0	121360	1681	8761600	3049	89	0,030
3	2680	43,5	116580	1892	7182400	3141	461	0,172
4	3080	39,5	121660	1560	9486400	2994	86	0,028
5	3240	44,5	144180	1980	10497600	3178	62	0,019
6	3900	53,0	206700	2809	15210000	3491	409	0,105
7	2780	33,5	93130	1122	7728400	2773	7	0,003
8	3160	44,0	139040	1936	9985600	3159	1	0,000
9	3040	36,5	110960	1332	9241600	2883	157	0,052
10	3240	43,5	140940	1892	10497600	3141	99	0,031
11	3180	38,0	120840	1444	10112400	2939	241	0,076
12	3460	57,5	198950	3306	11971600	3657	197	0,057
Итого (в среднем)	3115	42,8	134840	1873	9812567	1540	x	0,690

$$b = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{134840 - 3115\cdot 42{,}8}{1873 - 42{,}8^2} = 36{,}8$$

$$a = \bar{y} - b * \bar{x} = 3115 - 36{,}8 * 42{,}8 = 1539{,}5$$

Отсюда уравнение регрессии имеет вид: $y_x = 1539{,}5 + 36{,}8x$.

С увеличением расхода кормов на корову на 1 ц. корм.ед. среднегодовой надой на 1 корову в среднем возрастает на 36,8 кг. Тесноту линейной связи оценим с помощью коэффициента корреляции. Для этого сначала найдем среднеквадратические отклонения x и y по формулам:

$$\sigma_x = \sqrt{\bar{x^2} - \bar{x}^2} = \sqrt{41{,}16} = 6{,}42$$

$$\sigma_y = \sqrt{\bar{y^2} - \bar{y}^2} = \sqrt{109342} = 330{,}7$$

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\sigma_y} = \frac{134840 - 3115\cdot 42{,}8}{6{,}42\cdot 330{,}7} = 0{,}721$$

По шкале Чеддока можно сказать, что связь между x и y характеризуется как сильная. Определим качество модели через среднюю ошибку аппроксимации: $\bar{A} = \frac{1}{n} \cdot \sum \left| \frac{y - y_x}{y} \right| \cdot 100\% = \frac{0,690}{12} \cdot 100 = 5,75\%$

Качество модели можно оценить как хорошее, так как $\bar{A}$ не превышает 8-10%. Коэффициент детерминации r^2 равен 0,52. Это означает, что 52% вариации уровня удоев молока от одной коровы объясняется вариацией фактора х – расход кормов на корову.

Оценим значимость уравнения в целом с помощью F-критерия:

$$F_{факт} = \frac{r^2}{1 - r^2} \cdot \frac{n - 1 - 1}{1} = \frac{0,52}{1 - 0,52} \cdot (12 - 2) = 10,8$$

Определим критическое значения критерия по таблице при $к_1$=1, $к_2$=10 и уровне значимости α=0,05. Оно равно 4,96. Так как $F_{факт} > F_{кр}$, то гипотезу H_0 о случайном характере связи отклоняем с вероятностью 95%. Уравнение регрессии статистически значимо.

Оценку статистической значимости параметров регрессии проведем с помощью t-статистики Стьюдента. Выдвигаем гипотезы H_0 о статистически незначимом отличии показателей от нуля: a=b=r= 0. $t_{кр}$ для числа степеней свободы df= n-2=12-2=10 и α= 0,05 составит 2,23.

Определим случайные ошибки m_b, m_a, m_r:

$$m_b = \sqrt{\frac{S^2_{ост}}{\sum (x - \bar{x})^2}} = \sqrt{\frac{62997}{503,5}} = 11,19$$

$$m_a = \sqrt{\frac{S^2_{ост} \cdot \bar{x^2}}{\sum (x - \bar{x})^2}} = \sqrt{\frac{62997 \cdot 1873}{503,5}} = 484$$

$$m_r = \sqrt{\frac{1 - r^2}{n - 2}} = \sqrt{\frac{1 - 0,52}{12 - 2}} = 0,219$$

Тогда t_b=36,8/11,19=3,29, t_a=1539,5/484=3,18, t_r=0,721/0,219=3,29 . Фактические значения t-статистики превосходят табличные значения:

$$t_b = 3,29 > t_{кр} = 2,23,\ t_a = 3,18 > t_{кр} = 2,23,\ t_r = 3,29 > t_{кр} = 2,23\ .$$

Поэтому гипотезы H_0 отклоняется, т.е. a,b и r не случайно отличаются от нуля, а статистически значимы с вероятностью 95%. Зная связь между F и t-критериями можно было найти t-критерий для b и r по-другому:

$$t_b = t_r = \pm\sqrt{F} = \sqrt{10,8} = 3,29$$

Результаты проверки надежности отдельных параметров согласуются с результатами проверки уравнения в целом. Рассчитаем доверительный интервал для коэффициента регрессии b. Для этого определим предельную ошибку $\Delta = t_{кр} * m_b = 2,23 * 11,19 = 25,0$. Соответственно доверительный интервал при 5% уровне значимости будет:

$$36,8 - 25,0 \le \beta \le 36,8 + 25,0$$

$$11,8 \le \beta \le 61,8$$

Анализ верхней и нижней границ доверительного интервала приводит к выводу о том, что с вероятностью p =1-α=0,95 коэффициент регрессии b, находясь в указанных границах, не принимает нулевого значения, т.е. не являются статистически незначимым и существенно отличен от нуля.

Пример 2. Изучается зависимость потребления материалов y от объема производства продукции x. По 20 наблюдениям были получены следующие варианты уравнения регрессии:

1. y =3+2x.

   (6,48)

2. *ln y=2,5+ 0,2 lnx,* $r^2=0,68$.

   (6,19)

3. ln y=1,1+0,8x, $r^2=0,69$.

   (6,2)

4. $y=3+1,5x+0,1 \cdot X^2$, $r^2=0,701$.

   (3,0) (2,65)

В скобках указаны фактические значения t-критерия.

Задание.

1. Определите коэффициент детерминации для 1-го уравнения.
2. Запишите функции, характеризующие зависимость y от x во 2-м и 3-м уравнениях.
3. Выберите наилучший вариант уравнения регрессии.

Решение. Используем связь между F и t-критериями: $F = t^2_b = 6,48^2 = 42$, также известно, что для линейного уравнения F-критерий можно рассчитать:

$$F = \frac{r^2}{1-r^2} \cdot (12-2) = 42$$

Решая это уравнение с одной неизвестной, получаем, что r^2=0,81. Для выражения уравнения в форме зависимости у от х потенцируем обе части:

$$e^{\ln y} = e^{2,5+0,2\ln x},$$
$$y = e^{2,5} \cdot x^{0,2},$$
$$y = 12 \cdot x^{0,2}$$

Аналогично для 3-го уравнения получим:

$$e^{\ln y} = e^{1,1+0,8x},$$
$$y = e^{1,1} \cdot 2,2^{x},$$
$$y = 3 \cdot 2,2^{x}$$

Для определения лучшего уравнения оценим значимость параметров каждого уравнения:

1. Так как t_b=6,48 > $t_{кр}$=2,88 (α=0,01 и k=n-2=20-2=18), то параметр b статистически значим с вероятностью 99%.

2. Так как t_b=6,19 > $t_{кр}$=2,88 (α=0,01 и k=n-2=18), то параметр b статистически значим с вероятностью 99%.

3. Так как t_b=6,2 > $t_{кр}$=2,88 (α=0,01 и k=n-2=18), то параметр b статистически значим с вероятностью 99%.

4. Так как t_{b1}=3,0 > $t_{кр}$=2,90 (α=0,01 и k=n-3=20-3=17), то параметр b_1 статистически значим с вероятностью 99%. Но, так как t_{b2}=2,65 < $t_{кр}$=2,90 (α=0,01 и k=n-3=17), то следует принять с вероятностью 99% гипотезу H_0 о статистической незначимости параметра регрессии b_2 и равенстве его нулю.

На основе выше изложенного можно сделать вывод, что 4-ое уравнение менее надежно и его нельзя использовать в качестве наилучшего варианта уравнения регрессии. Так как первые три уравнения имеют статистические значимые параметры, то в качестве дополнительного критерия качества построенной модели рассмотрим коэффициент детерминации. С учетом того, что r^2_1=0,81 > r^2_2=0,68 и r^2_3=0,69, то наиболее приемлемым и надежным в данной ситуации представляется первое уравнение регрессии.

1.6. Спецификация модели множественной регрессии и корреляции

Любое эконометрическое исследование начинается со **спецификации модели**, т.е. с формулировки вида модели, исходя из соответствующей теории связи между переменными. Иными словами, исследование начинается с теории, устанавливающей связь между явлениями. Прежде всего, из всего круга факторов, влияющих на результативный признак, необходимо выделить наиболее существенно влияющие факторы. От правильно выбранной спецификации модели зависит величина случайных ошибок.

До сих пор мы рассматривали корреляционные связи между двумя признаками: результативным (у) и факторным (х). Например, выпуск продукции зависит не только от размера основного капитала, но и от уровня квалификации рабочих, состояния оборудования, обеспеченности и качества сырья и материалов, организации труда и т.д. В связи с этим возникает необходимость в изучении, измерении связи между результативным признаком и двумя и более факторными.

Последние годы в условиях стагнации производства в сельском хозяйстве положительная тенденция, за редким исключением, отсутствует. Показатели урожайности сельскохозяйственных культур, продуктивности животных, как правило, снижаются. В связи с этим для прогнозирования основных производственно-экономических показателей целесообразно применять другие методы, в частности, корреляционно-регрессионный анализ на основе многофакторных моделей с целью выявления влияния тех или иных факторов на результаты производства.

Такая взаимосвязь может быть представлена в виде уравнения множественной регрессии:

$$y=f(x_1,x_2,\dots,x_p), \qquad (1.33.)$$

где у – зависимая переменная (результативный признак),

$x_1,x_2,\dots,x_p$ – независимые переменные (факторы).

Множественный корреляционно-регрессионный анализ решает три задачи:

- определение формы связи, параметров уравнения;
- оценка степени тесноты связи;
- установление влияния отдельных факторов на общий результат.

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели. Она включает в себя два круга вопросов: отбор факторов и выбор вида уравнения регрессии. Их решение при построении модели множественной регрессии имеет некоторую специфику.

Включение в уравнение множественной регрессии того или иного набора факторов связано прежде всего с представлением исследователя о природе взаимосвязи моделируемого показателя с другими экономическими явлениями. Факторы, включаемые во множественную регрессию, должны отвечать следующим требованиям:

1. Они должны быть количественно измеримы. Если необходимо 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. Факторы не должны быть сильно коррелированны между собой и тем более находиться в точной функциональной связи. Включение в модель признаков с высокой межфакторной корреляцией (например, когда $r_{yx1} < r_{x1x2}$ в модели с двумя факторами) может привести к нежелательным последствиям — система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии. Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результативный показатель и параметры уравнения регрессии оказываются не интерпретируемыми.

Так, в уравнении $y=a+b_1x_1+b_2x_2$ предполагается, что факторы x_1 и x_2 независимы друг от друга, т. е. $r_{x1x2} = 0$. Тогда можно говорить, что параметр b_1, измеряет силу влияния фактора x_1 на результат у при неизменном значении фактора x_2. Если же $r_{x1x2} = 1$, то с изменением фактора x_1 фактор x_2 не может оставаться неизменным. Отсюда b_1 и b_2 нельзя интерпретировать как показатели раздельного влияния x_1 и x_2 на у.

3. Включаемые во множественную регрессию факторы должны объяснять вариацию зависимой переменной. Если строится модель с набором m факторов, то для нее рассчитывается показатель детерминации R^2_m, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии m факторов. Влияние других, не учтенных в модели факторов, оценивается как $1-R^2_m$ с соответствующей остаточной дисперсией. При дополнительном включении в регрессию m+1 фактора коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться, т.е. $R^2_{m+1}>R^2_m$ и $\sigma^2_{ост\ m+1}<\sigma^2_{ост\ m}$. Если же этого не происходит и данные показатели практически мало отличаются друг от друга, то включаемый в анализ фактор не улучшает модель и практически является лишним фактором.

Так, если для регрессии, включающей пять факторов, коэффициент детерминации составил 0,857 и включение шестого фактора дало коэффициент детерминации 0,858, то вряд ли целесообразно дополнительно включать в модель этот фактор. Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и не увеличивает показатель детерминации, но и приводит к статистической незначимости параметров регрессии по t-критерию Стьюдента.

4. Замещающие переменные. Часто бывает, что вы не можете найти данных по переменной, которую хотелось бы включить в уравнение регрессии. Некоторые переменные, относящиеся к социально-экономическому положению или к качеству образования, имеют такое расплывчатое определение, что их в принципе даже невозможно измерить. Другие могут поддаваться измерению, но оно требует столько времени и энергии, что на практике их приходится отбрасывать.

Независимо от причины обычно бывает полезно вместо отсутствующей переменной использовать некоторый ее заменитель (proxy), а не пренебрегать ею совершенно. В качестве показателя общего социально-экономического положения вы можете использовать его заменитель — показатель дохода, если данные о нем имеются. В качестве показателя качества образования можно использовать отношение числа преподавателей и сотрудников к числу студентов или расходы на одного студента. Вместо переменной, опущенной в каком-либо обзоре, вы можете обратиться к другим, уже фактически собранным данным, если в них имеется подходящая замещающая переменная.

Имеются две причины для поиска такой переменной. Во-первых, если вы просто опустите важную переменную, то регрессия может пострадать от смещения оценок и статистическая проверка будет ненадежной. Во-вторых, результаты оценки регрессии с включением замещающей переменной могут дать косвенную информацию о той переменной, которая замещена данной переменной. Так, фактор времени можно в некоторых ситуациях использовать как замещающую переменную для показателя технического прогресса.

Может быть ситуация непреднамеренного использования замещающих переменных. Случается, что вы используете замещающую переменную, не осознавая этого. Вы полагаете, что y зависит от z, а в действительности эта величина зависит от x. Если корреляция между величинами z и x незначительна, то результаты будут плохими, и вы поймете, что тут что-то неладно. Но если корреляция тесная, то результаты окажутся удовлетворительными (коэффициент R^2 будет близок к желаемому уровню и т. п.), и вы можете даже не подозревать, что полученное соотношение неверно.

Как это может отразиться на результатах анализа? Это, во-первых, зависит от того, с какой целью вы строите данную регрессию. Если целью оценивания регрессии является предсказание будущих значений величины y, то использование замещающей переменной не будет иметь большого значения при условии, конечно, что корреляция тесная и не является в то же время счастливой статистической случайностью. Однако если вы намерены использовать объясняющую переменную в качестве инструмента экономической политики для оказания влияния на поведение зависимой переменной, то последствия могут оказаться катастрофическими. Если только не будет функциональной связи между замещающей переменной и истинной объясняющей переменной, манипулирование замещающей переменной не окажет никакого влияния на зависимую переменную. Если мотивом построения регрессии является чисто научное любопытство, то исход будет столь же неудовлетворительным.

Таким образом, хотя теоретически регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости. Отбор факторов производится на основе качественного теоретико-экономического анализа. Однако теоретически анализ часто не позволяет однозначно ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения фактора в модель. Поэтому отбор факторов обычно осуществляется в две стадии: на первой подбираются факторы исходя из сущности проблемы; на второй - на основе **матрицы показателей корреляции** определяют t-статистики для параметров регрессии.

Коэффициенты корреляции между факторами (т. е. корреляции между объясняющими переменными) позволяют исключать из модели дублирующие факторы. Считается, что две переменные явно **коллинеарны**, т. е. находятся между собой в линейной зависимости, если $r_{x1x2} \geq 0,7$.

Поскольку одним из условий построения уравнения множественной регрессии является независимость действия факторов, т.е. $r_{xi\ xj}=0$, то коллинеарность факторов нарушает это условие. Если факторы явно коллинеарны, то они дублируют друг друга и один из них рекомендуется исключить из регрессии. Предпочтение при этом отдается не фактору, более тесно связанному с результатом, а тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами. В этом требовании проявляется специфика множественной регрессии как метода исследования комплексного воздействия факторов в условиях их независимости друг от друга.

По величине парных коэффициентов корреляции обнаруживается лишь явная коллинеарность факторов. Наибольшие трудности в использовании аппарата множественной регрессии возникают при наличии **мультиколлинеарности** факторов, когда более чем два фактора связаны между собой линейной зависимостью, т.е. имеет место совокупное воздействие факторов друг на друга. Включение в модель мультиколлинеарных факторов нежелательно в силу следующих последствий:

• затрудняется интерпретация параметров множественной регрессии как характеристик действия факторов в «чистом» виде, ибо факторы коррелированны, параметры линейной регрессии теряют экономический смысл;

• оценки параметров ненадежны, колеблются с изменением объема наблюдений не только по величине, но и по знаку, что делает модель непригодной для анализа и прогнозирования; обнаруживаются большие стандартные ошибки.

Для оценке мультиколлинеарности факторов может быть использован определитель матрицы парных коэффициентов корреляции между факторами. Если бы факторы не коррелировали между собой, то матрица парных коэффициентов была бы единичной, т.к. все недиагональные элементы были бы равны 0, т.е. Det|R|=1. Если, наоборот, то все коэффициенты корреляции равны 1 и тогда определитель равен 0. Следовательно, чем ближе к 0 определитель матрицы, тем сильнее будет мультиколлинеарность факторов и ненадежнее результаты множественной регрессии.

Например, для 3 коррелируемых факторов определитель матрицы будет выглядеть следующем образом:

$$Det|R| = \begin{vmatrix} r_{x1x1} & r_{x1x2} & r_{x1x3} \\ r_{x2x1} & r_{x2x2} & r_{x2x3} \\ r_{x3x1} & r_{x3x2} & r_{x3x3} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{vmatrix} = 0 \quad (1.34.)$$

Проверка мультиколлинеарности факторов может быть проведена методом испытания гипотезы о независимости переменных H_0: Det|R|=1. Для этого находят фактическое значение критерия χ^2 :

$$\chi^2 =(n-1)-1/6*(2*m+5)*lgDet|R| \qquad (1.35.)$$

После чего сравнивают его с табличным, определенном при n*(n-1)/2 степенях свободы и уровне значимости α. Если фактическое превосходит табличное, то гипотеза H_0 отклоняется.

Существует ряд подходов преодоления сильной межфакторной корреляции:

- **исключение** из модели одного или нескольких факторов - является самым простым;
- **преобразование факторов**, при котором уменьшается корреляция между ними;
- переход к **совмещенным уравнениям** регрессии, т. е. к уравнениям, которые отражают не только влияние факторов, но и их взаимодействие. Например, возможно построение следующего совмещенного уравнения: $y=a+b_1x_1+b_2x_2+b_3x_1x_2$.

Рассматриваемое уравнение включает взаимодействие первого порядка (взаимодействие двух факторов). Возможно включение в модель и взаимодействий более высокого порядка, если будет доказана их статистическая значимость по F-критерию Фишера, например, взаимодействие второго порядка и т. д.

Отбор факторов, включаемых в регрессию, является одним из важнейших этапов практического использования методов регрессии. Подходы к отбору факторов на основе показателей корреляции могут быть разные. Наиболее широкое применение получили следующие методы построения уравнения множественной регрессии:

- **метод исключения;**
- **метод включения;**
- **шаговый регрессионный анализ**.

Каждый из этих методов по-своему решает проблему отбора факторов, давая в целом близкие результаты — отсев факторов из полного его набора (метод исключения), дополнительное введение фактора (метод включения), исключение ранее введенного фактора (шаговый регрессионный анализ).

Вследствие взаимодействия факторов парные коэффициенты корреляции не могут в полной мере решать вопрос о целесообразности включения в модель того или иного фактора. В качестве еще одного критерия используются показатели частной корреляции, оценивающие в чистом виде тесноту связи фактора с результатом.

Выбор формы уравнения регрессии. Как и в парной зависимости, возможны разные виды уравнений множественной регрессии: линейные и нелинейные. Ввиду четкой интерпретации параметров наиболее широко используются линейная и степенная функции.

В линейной множественной регрессии $y=a+b_1*x_1+b_2*x_2+\ldots+b_p*x_p$ параметры при переменных x называются **коэффициентами «чистой» регрессии**. Они характеризуют среднее изменение результата с изменением соответствующего фактора на единицу при неизмененном значении других факторов, закрепленных на среднем уровне.

В производственной функции вида $y=ax_1^{b1}x_2^{b2}\ldots x_m^{bm}$ экономический смысл имеют не только коэффициенты b_i каждого фактора, но и их сумма, т.е. **сумма эластичностей**: $B=b_1+\ldots+b_m$. Эта величина фиксирует обобщенную характеристику эластичности производства.

Для построения уравнения множественной регрессии чаще всего используются следующие функции:

- линейная – $y=a+b_1x_1+b_2x_2+\ldots+b_mx_m$,
- степенная – $y=ax_1^{b1}x_2^{b2}\ldots x_m^{bm}$,
- экспонента – $y=e^{a+b1x1+b2x2+\ldots+bmxm}$,
- гипербола – $y=1/(a+b_1x_1+b_2x_2+\ldots+b_mx_m)$.

Стандартные компьютерные программы обработки регрессионного анализа позволяют перебирать различные функции и выбрать ту из них, для которой остаточная дисперсия и ошибка аппроксимации минимальны, а коэффициент детерминации максимален.

Контрольные вопросы:

- Охарактеризуйте спецификацию модели множественной регрессии.
- Какие основные задачи решает множественный корреляционно-регрессионный анализ?
- Какие требования предъявляют к факторам, которые включают в модель множественной регрессии?
- Опишите виды математических функций часто используемых для построения уравнения множественной регрессии.
- Коэффициенты «чистой» регрессии – это…
- С помощью каких методов проводится отбор факторов в модель?
- Что такое мультиколлинеарность факторов и как ее можно проверить?
- Какие два фактора называются коллинеарными?
- Для чего нужны замещающие переменные?

Решение типовых задач

Пример 1. При изучении зависимости $y= f(x_1,x_2,x_3)$ известна матрица парных коэффициентов корреляции:

	y	x_1	x_2	x_3
y	1			
x_1	0,86	1		
x_2	0,74	0,81	1	
x_3	0,58	0,44	0,21	1

Задание. Определить коллинеарные факторы и устранить их влияние.

Решение. Так как r_{x1x2}=0,81>0,7, то факторы x_1 и x_2 коллинеарны между собой. Значит, они дублируют друг друга и необходимо исключить один из них. В модели целесообразно оставить фактор x_2, а не x_1, так как хотя корреляция x_2 с результатом у слабее, чем корреляция фактора x_1 с у, но вместе с тем межфакторная корреляция r_{x2x3} слабее межфакторной корреляции r_{x1x3}. Следовательно, в данной модели необходимо строить уравнение регрессии у от x_2 и x_3.

Пример 2. По 20 предприятиям исследуется зависимость объема производства продукции (у) от энергетических мощностей (x_1), фондооснащенности (x_2) и трудовой нагрузки (x_3). Были получены следующие варианты уравнений регрессии:

1. $y_x = 23 + 2x_1 + 4x_2 + 3{,}2x_3$

(0,5) (2) (2,4)

2. $y_x = 19 + 1{,}8x_1 + 3{,}1x^2_1 + 4{,}2x_2$

(0,7) (3,4) (1,8)

3. $y_x = 21 + 2{,}4x_1 + 4{,}3x_2$

(0,8) (1,3)

В скобках указаны стандартные ошибки для коэффициентов регрессии.

Задание. Выберите наилучшее уравнение регрессии.

Решение. Для выбора наилучшего уравнения необходимо проверить надежность коэффициентов регрессии в каждом уравнении. Для этого используется t-критерий Стьюдента:

$$t_{факт}=b/m_b,$$

где b - коэффициент регрессии; m_b - стандартная ошибка коэффициента.

Для первого уравнения фактические значения t-критерия равны соответственно 4; 2; 1,33. Находим критическое значение t-критерия из таблицы (см. приложение 3) для степеней свободы v=n-m-1=20-3-1=16 и уровня значимости α=0,05: $t_{кр}$=2,12.

Так как $t_{кр} > t_{факт}$ для второго и третьего коэффициентов регрессии b_2 и b_3, то с вероятностью 95% можно утверждать, что они равны нулю и статистически незначимы. Первый коэффициент регрессии является значимым.

Аналогичные рассуждения для второго уравнения говорят о том, что только второй коэффициент регрессии оказался ненадежным. В третьем уравнении все коэффициенты регрессии оказались статистически значимыми. Следовательно, из всех представленных уравнений наиболее надежным является третье. В данном примере для описания зависимости лучше всего подходит линейное уравнение регрессии с двумя факторами x_1 и x_2. Фактор x_3 следует исключить из модели.

Пример 3. По 20 предприятиям региона изучается зависимость выработки продукции на 1 работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса работников высокой квалификации в общей численности работников x_2 (%) – табл.2.1.

Таблица 2.1.

№ предприятия	y	x_1	x_2	№ предприятия	y	x_1	x_2
1	7,0	3,9	10,0	11	9,0	6,0	21,0
2	7,0	3,9	14,0	12	11,0	6,4	22,0
3	7,0	3,7	15,0	13	9,0	6,8	22,0
4	7,0	4,0	16,0	14	11,0	7,2	25,0
5	7,0	3,8	17,0	15	12,0	8,0	28,0
6	7,0	4,8	19,0	16	12,0	8,2	29,0
7	8,0	5,4	19,0	17	12,0	8,1	30,0
8	8,0	4,4	20,0	18	12,0	8,5	31,0
9	8,0	5,3	20,0	19	14,0	9,6	32,0
10	10,0	6,8	20,0	20	14,0	9,0	36,0

Известна матрица парных коэффициентов корреляции:

	y	x_1	x_2
y	1		
x_1	0,970	1	
x_2	0,941	0,943	1

Задание.

1. Постройте линейное уравнение множественной регрессии.
2. Проанализировать включение в модель каждого из факторов и наличие коллинеарности. Сделайте выводы о возможности улучшения построенной модели.

Решение. Линейное уравнение множественной регрессии представляет собой следующий вид: $y=a+b_1x_1+b_2x_2$

Для нахождения ее параметров решим систему из трех уравнений.

$$\begin{cases} n\cdot a+b_1\sum x_1+b_2\cdot\sum x_2=\sum y, \\ a\cdot\sum x_1+b_1\cdot\sum x_1^2+b_2\cdot\sum x_2\cdot x_1=\sum y\cdot x_1, \\ a\cdot\sum x_2+b_1\cdot\sum x_1\cdot x_2+b_2\cdot\sum x_2^2=\sum y\cdot x_2. \end{cases}$$

По результатам вычислений составим уравнение множественной регрессии линейного вида: $y=1{,}84+0{,}946x_1+0{,}086x_2$.

Полученное уравнение говорит о том, что с ростом ввода новых основных фондов на 1% выработка продукции на 1 работника возрастет в среднем на 0,946 тыс. руб. при том же уровне квалификации персонала. Увеличение доли высококвалифицированных специалистов среди работников на 1% при том же уровне ввода новых основных фондов предполагает дополнительный рост выработки продукции на 0,086 тыс. руб.

Найдем множественный коэффициент корреляции:

$$R_{yx1x2}=\sqrt{\frac{r_{yx1}^2+r_{yx2}^2-2\cdot r_{yx1}\cdot r_{yx2}\cdot r_{x1x2}}{1-r_{x1x2}^2}}=0{,}973$$

Показатель корреляции показывает на наличие очень сильной связи между факторами и результатом.

Значение коэффициентов парной корреляции указывают на сильную связь выработки продукции у как с коэффициентом обновления фондов – x_1, так и с долей работников высокой квалификации – x_2. Но в то же время $r_{x1x2}=0{,}943$ показывает на наличие явной коллинеарности между факторами, причем межфакторная связь настолько тесная, что даже превышает тесноту связи у с x_2. В связи с этим для улучшения данной модели можно исключить один из дублирующих факторов, являющихся менее информативным.

Для анализа существенности каждого фактора в модели используем частные F-критерии Фишера и оценку значимости каждого из параметров уравнения.

Найдем частные F-критерии Фишера по формулам:

$$F_{частx1} = \frac{R^2_{yx1x2} - r^2_{yx2}}{1 - R^2_{yx1x2}} \cdot \frac{n-m-1}{1} = \frac{0,947 - 0,885}{1 - 0,947} \cdot \frac{20-2-1}{1} = 19,9$$

$$F_{частx2} = \frac{R^2_{yx1x2} - r^2_{yx1}}{1 - R^2_{yx1x2}} \cdot \frac{n-m-1}{1} = \frac{0,947 - 0,941}{1 - 0,947} \cdot \frac{20-2-1}{1} = 1,9$$

Частный F-критерий – $F_{частx1}$ показывает статистическую значимость включения фактора x_1 в модель после того, как в нее включен фактор x_2. Табличное значение F-критерия при k_1=1, k_2=17 и уровне значимости α=0,05 равняется 4,45. Так как фактическое значение $F_{частx1}$=19,9>4,45, то фактор x_1 целесообразно включить в модель после фактора x_2, его дополнение в модель является статистически значимым. Прирост факторной дисперсии за счет дополнительного фактора x_1 является существенным.

Если поменять первоначальный порядок включения факторов в модель и рассмотреть включения фактора x_2 после x_1, то результат проверки значимости будет иным. Так как фактическое значение $F_{частx2}$=1,9<4,45, то фактор x_2 нецелесообразно включать в модель после фактора x_1, прирост факторной дисперсии за счет дополнительного признака x_2 оказался незначительным. Это говорит о случайной природе включения фактора x_2 в модель.

Теперь проведем анализ с учетом значимости коэффициентов регрессии. Для этого вычислим t-критерии коэффициентов b_1 и b_2.

$$t_{b1} = \sqrt{F_{частx1}} = 4,5$$
$$t_{b2} = \sqrt{F_{частx2}} = 1,4$$

По таблице критических значений находим $t_{кр}$ при к=17 и α=0,01: $t_{кр}$=2,9.

При сравнении фактических значений критерия Стьюдента с критическим получим, что статистически значимым является коэффициент b_1, а коэффициент регрессии b_2 сформировался под воздействием случайных причин, незначительно отличается от нуля. Это утверждение верно с вероятностью 99%. Поэтому фактор x_2, силу которого оценивает b_2 можно исключить как несущественно влияющий на результат.

Таким образом, обе проверки дали одинаковый результат. Общий вывод состоит в том, что множественная модель с факторами x_1 и x_2 содержит неинформативный фактор x_2. Для улучшения модели необходимо перейти от множественной регрессии к парной, выражающей зависимость у от x_1. При этом являющейся более простой, хорошо детерминированной и пригодной для анализа и прогноза.

1.7. Построение многофакторной регрессионной модели и множественная корреляция

Парная регрессия может дать хороший результат при моделировании, если влиянием других факторов, воздействующих на объект исследования, можно пренебречь. Поведение отдельных экономических переменных контролировать нельзя, т. е. не удается обеспечить равенство всех прочих условий для оценки влияния одного исследуемого фактора. В этом случае следует попытаться выявить влияние других факторов, введя их в модель, т. е. построить уравнение множественной регрессии.

Линейное уравнение **множественной регрессии** имеет вид:

$$y=a+b_1x_1+b_2x_2+\ldots+b_nx_n \text{ , (1.36.)}$$

где a, $b_1,b_2\ldots b_n$ – параметры уравнения;

$x_1,x_2\ldots x_n$ – факторы, объясняющие переменные;

y – результативный признак, зависимая переменная.

Параметры уравнения множественной регрессии оцениваются, как и в парной регрессии, методом наименьших квадратов (МНК):

$$\begin{cases} n\cdot a+b_1\sum x_1+b_2\cdot\sum x_2+\ldots+b_n\cdot\sum x_n=\sum y, \\ a\cdot\sum x_1+b_1\cdot\sum x_1^2+b_2\cdot\sum x_2\cdot x_1+\ldots+b_n\cdot\sum x_n\cdot x_1=\sum y\cdot x_1, \\ \ldots \\ a\cdot\sum x_n+b_1\cdot\sum x_1\cdot x_n+b_2\cdot\sum x_2\cdot x_n+\ldots+b_n\cdot\sum x_n^2=\sum y\cdot x_n. \end{cases} \quad (2.4.)$$

Для ее решения можно применить обычный метод подстановок или использовать методом определителей.

Еще один способ получения параметров многофакторного регрессионного уравнения с помощью **коэффициентов парной корреляции**. Все переменные и соотношения между ними будем брать в стандартизированном масштабе. За начало отсчета принимаем среднее арифметическое, а за единицу измерения – среднее квадратическое отклонение. Тогда значение признака в стандартизированном масштабе запишется:

$$t_{xi} = \frac{x_i - \bar{x}}{\sigma_{xi}}, \quad \mathrm{t}_\mathrm{y} = \frac{y - \bar{y}}{\sigma_y} \quad (1.37.)$$

В стандартизированном масштабе среднее значение признака t_y=0, а среднее σ_y =1. Получим уравнение регрессии в **стандартизированном масштабе**:

$$t_y=\beta_1 t_{x1}+\beta_2 t_{x2}+\ldots+\beta_p t_{xp}, \quad (1.38.)$$

где β_i – стандартизированные коэффициенты.

К такому уравнению также применим метод МНК. Стандартизированные коэффициенты регрессии β_i определяются из следующей системы уравнений:

$$\begin{cases} r_{yx1}=\beta_1+\beta_2 r_{x2x1}+\ldots+\beta_p r_{xpx1}, \\ r_{yx2}=\beta_1 r_{x1x2}+\beta_2+\ldots+\beta_p r_{xpx2}, \\ \ldots \\ r_{yxp}=\beta_1 r_{x1xp}+\beta_2 r_{x2xp}+\ldots+\beta_p. \end{cases} \quad (1.39.)$$

Решение может быть записано в виде:

$$\beta_i = \frac{\Delta_i}{\Delta}, \quad где \quad \mathrm{i}=1..\mathrm{p}, \quad \Delta = \begin{vmatrix} 1 & r_{x2x1} & \ldots & r_{xpx1} \\ r_{x1x2} & 1 & \ldots & r_{xpx2} \\ \ldots & & & \\ r_{x1xp} & r_{x2xp} & \ldots & 1 \end{vmatrix} \quad (1.40.)$$

Определитель Δ_i получается из определителя Δ заменой в нем соответствующего столбца столбцом свободных членов исходной системы, т.е применим также метод определителей.

Существует связь между коэффициентами множественной регрессии b_i и стандартизированными β_I:

$$\beta_i = b_i \cdot \frac{\sigma_{xi}}{\sigma_y} \quad (1.41.)$$

Параметр а определяется как а= $\bar{y}$ $-b_1$ $\bar{x}_1$-b_2 $\bar{x}_2$-…-b_p $\bar{x}_p$. **Стандартизированные коэффициенты** выражают скорость изменения среднего значения функции по каждому из аргументов при постоянном значении прочих аргументов. Они показывают, на сколько сигм изменится в среднем результат, если соответствующий фактор x_i, изменится на одну сигму при неизменном среднем уровне других факторов. В силу того, что все переменные заданы как центрированные и нормированные, стандартизованные коэффициенты регрессии сравнимы между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов регрессии в отличие от коэффициентов «чистой» регрессии b_i, которые несравнимы между собой.

При нелинейной зависимости признаков, приводимой к линейному виду, параметры множественной регрессии также определяются МНК с той лишь разницей, что он используется не к исходной информации, а к преобразованным данным. Поскольку параметры степенной функции представляют собой коэффициенты эластичности, то они сравнимы по разным факторам.

Практическая значимость уравнения множественной регрессии оценивается с помощью **показателя множественной корреляции** и его квадрата - коэффициента детерминации. Независимо от формы связи показатель множественной корреляции может быть найден как индекс множественной корреляции:

$$R_{yx1...xp}=\sqrt{1-\frac{\sigma^2_{ост}}{\sigma^2_y}}=\sqrt{\frac{\sigma^2_{факт}}{\sigma^2_y}} \quad (1.42.)$$

где σ_y - общая дисперсия результативного признака;

$\sigma_{ост}$ - остаточная дисперсия, определяемая исходя из уравнения регрессии;

$\sigma_{факт}$ - факторная дисперсия.

Методика построения индекса множественной корреляции аналогична построению индекса корреляции для парной зависимости. Границы его изменения те же: от 0 до 1. Чем ближе его значение к 1, тем теснее связь результативного признака со всем набором исследуемых факторов. Величина индекса множественной корреляции должна быть больше или равна максимальному парному индексу корреляции:

$$R_{yx1x2\ldots xp} \geq \max(r_{yxi}) \;, i=1..p \quad (1.43.)$$

При линейной зависимости признаков формула индекса корреляции может быть получена через стандартизированные коэффициенты:

$$R_{yx1x2\ldots xp}=\sqrt{\sum \beta_i \cdot r_{yxi}} \quad (1.44.)$$

где r_{yxi} - парные коэффициенты корреляции результата с каждым фактором.

При правильном включении факторов в регрессионный анализ величина индекса множественной корреляции будет существенно отличаться от индекса корреляции парной зависимости. Если же дополнительно включенные в уравнение множественной регрессии факторы третьестепенны, то индекс множественной корреляции может практически совпадать с индексом парной корреляции. Отсюда ясно, что, сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

При линейной зависимости коэффициент множественной корреляции можно определить также через матрицу парных коэффициентов корреляции:

$$R_{yx1x2\ldots xp}=\sqrt{1-\frac{\Delta r}{\Delta r_1}}\;, \quad (1.45.)$$

$$\Delta r = \begin{vmatrix} 1 & r_{yx1} & \ldots & r_{yxp} \\ r_{yx1} & 1 & \ldots & r_{x1xp} \\ \ldots & & & \\ r_{yx1} & r_{x1xp} & \ldots & 1 \end{vmatrix}$$

- определитель матрицы парных коэффициентов корреляции;

$$\Delta r_1 = \begin{vmatrix} 1 & r_{x1x2} & \ldots & r_{x1xp} \\ r_{x2x1} & 1 & \ldots & r_{x2xp} \\ \ldots & & & \\ r_{xpx1} & r_{xpx2} & \ldots & 1 \end{vmatrix}$$

- определитель матрицы межфакторной корреляции.

Как видим, величина множественного коэффициента корреляции зависит не только от корреляции результата с каждым из факторов, но и от **межфакторной корреляции**. Рассмотренная формула позволяет определять совокупный коэффициент корреляции, не обращаясь при этом к уравнению множественной регрессии, а используя лишь парные коэффициенты корреляции.

При трех переменных для двухфакторного уравнения регрессии данная формула совокупного коэффициента корреляции легко приводится к следующему виду:

$$R_{yx1x2}=\sqrt{\frac{r_{yx1}^2+r_{yx2}^2-2\cdot r_{yx1}\cdot r_{yx2}\cdot r_{x1x2}}{1-r_{x1x2}^2}} \quad (1.46.)$$

Величина показателя множественной корреляции определяет степень тесноты связи между признаками в линейной и нелинейной формах. Низкое значение коэффициента (индекса) множественной корреляции означает, что в регрессионную модель не включены существенные факторы — с одной стороны, а с другой стороны — рассматриваемая форма связи не отражает реальные соотношения между переменными, включенными в модель. Требуются дальнейшие исследования по улучшению качества модели и увеличению ее практической значимости.

На основе линейного уравнения множественной регрессии $y=a+b_1*x_1+b_2*x_2+...+b_p*x_p$ могут быть найдены частные уравнения регрессии, то есть уравнения регрессии, которые связывают результативный признак с соответствующими факторами x при закреплении других учитываемых во множественной регрессии факторов на среднем уровне. **Частные уравнения** регрессии имеют следующий вид:

$$\begin{aligned}
&y_{x_1/x_2,x_3,...,x_p}=a+b_1\cdot x_1+b_2\cdot\bar{x}_2+b_3\cdot\bar{x}_3+...+b_p\cdot\bar{x}_p,\\
&y_{x_2/x_1,x_3,...,x_p}=a+b_1\cdot\bar{x}_1+b_2\cdot x_2+b_3\cdot\bar{x}_3+...+b_p\cdot\bar{x}_p,\\
&\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots, \qquad (1.47.)\\
&y_{x_p/x_1,x_2,...,x_{p-1}}=a+b_1\cdot\bar{x}_1+b_2\cdot\bar{x}_2+...+b_{p-1}\cdot\bar{x}_{p-1}+b_p\cdot x_p.
\end{aligned}$$

При подстановке в уравнения (1.47.) средних значений соответствующих факторов они принимают вид парных уравнений линейной регрессии, то есть имеем:

$$\begin{cases} y_{x_1.x_2x_3...x_p} = A_1 + b_1x_1, \\ y_{x_2.x_1x_3...x_p} = A_2 + b_2x_2, \\ ...., \\ y_{x_p.x_1x_2...x_{p-1}} = A_p + b_px_p, \end{cases} \quad (1.48.)$$

В отличие от парной регрессии частные уравнения регрессии характеризуют изолированное влияние фактора на результат, ибо другие факторы закреплены на неизменном уровне. Эффекты влияния других факторов присоединены в них к свободному члену уравнения множественной регрессии. Это позволяет на основании частных уравнений регрессии определять **частные коэффициенты эластичности**:

$$Э_{y_{xi}} = b_i \cdot \frac{x_i}{y_{x_i / x_1x_2 ... x_{i-1}x_{i+1} ... x_p}}, \quad (1.49.)$$

где b_i – коэффициенты регрессии для фактора x_i в линейном уравнении множественной регрессии;

$y_{x_i / x_1x_2 ... x_{i-1}x_{i+1} ... x_p}$ - частное уравнение регрессии.

Частные коэффициенты эластичности в общем виде для любого уравнения множественной регрессии можно рассчитать как:

$Э_{yxi} = f'_{xi}(x_1\ x_2\ \ldots x_p) * x_i / y$, (1.50.)

где $f'_{xi}(x_1\ x_2\ \ldots x_p)$ – частная производная по переменной x_i .

Контрольные вопросы:

- Запишите линейное уравнение множественной регрессии для трех факторов.
- Перечислите все основные способы оценки параметров в линейном уравнении множественной регрессии.
- Что такое стандартизированное уравнение множественной регрессии?

- Как связаны β-коэффициенты с коэффициентами линейное уравнение множественной регрессии в естественной форме?
- Что определяют β-коэффициенты?
- Как оценивается степень тесноты связи в уравнение множественной регрессии?
- Как связаны коэффициента корреляции между переменными с множественным коэффициентом корреляции?
- В каких пределах находится множественным коэффициентом корреляции?

Решение типовых задач

Пример 1. В таблице 1.6. представлены данные за ряд лет по экономическому району Смоленской области о размерах валового дохода, стоимости производственных фондов и численности работников.

Задание.

1. Обоснуйте выбор уравнения регрессии.
2. Рассчитайте параметры регрессионной модели и оцените его значимость через F-критерий Фишера.
3. Определите частные коэффициенты эластичности.

Таблица 1.6.

№	Валовой доход, млн.руб.	Основные производственные фонды, млн.руб.	Численность работников сферы материального производства, тыс. чел.
1	1110	1604	970
2	1289	1725	981
3	1443	1855	1003
4	1652	1994	1028
5	1793	2144	1080
6	1997	2305	1091
7	2091	2408	1116
8	2068	2673	1136
9	2344	2879	1136
10	2602	3089	1167
11	2820	3522	1205
12	2816	3816	1223

Решение. В качестве зависимой переменной следует выбрать валовой доход (у) как результативный показатель хозяйственной деятельности, а ресурсы производства (основные фонды, трудовые ресурсы) выступают в роли влияющих факторов. Такого рода зависимости в практике исследований выражаются уравнением производственной функции Кобба-Дугласа. В нашем случае оно примет вид: $y=a\ x_1^{\ b1}\ x_2^{\ b2}$.

Для нахождения его параметров следует вначале прологарифмировать его: $lgy=lga+b_1lgx_1+b_2lgx_2$. Обозначим $k_0=lgy$, $k_1=lgx_1$, $k_2=lgx_2$, $a_0=lga$. В результате получим $k_0=a_0+b_1k_1+b_2k_2$. По форме получено линейное уравнение для трех переменных. Для определения параметров составим систему нормальных уравнений на основе МНК:

$$\begin{cases} n\cdot a_0+b_1\sum k_1+b_2\cdot\sum k_2=\sum k_0, \\ a\cdot\sum k_1+b_1\cdot\sum k_1^2+b_2\cdot\sum k_2\cdot k_1=\sum k_0\cdot k_1, \\ a\cdot\sum k_2+b_1\cdot\sum k_1\cdot k_2+b_2\cdot\sum k_2^2=\sum k_0\cdot k_2. \end{cases}$$

Логарифмы анализируемых показателей, а также производные величины занесем в отдельную таблицу 1.7.:

Таблица 1.7

№	k_0	k_1	k_2	k_1^2	k_2^2	k_0^2	k_1k_2	k_0k_1	k_0k_2
1	3,045	3,205	2,987	10,273	8,921	9,274	9,573	9,761	9,096
2	3,110	3,237	2,992	10,477	8,950	9,674	9,683	10,067	9,305
3	3,159	3,268	3,001	10,682	9,008	9,981	9,809	10,326	9,482
4	3,218	3,300	3,012	10,888	9,072	10,356	9,939	10,619	9,693
5	3,254	3,331	3,033	11,097	9,202	10,586	10,105	10,838	9,869
6	3,300	3,363	3,038	11,308	9,228	10,892	10,215	11,098	10,026
7	3,320	3,382	3,048	11,436	9,288	11,025	10,306	11,228	10,119
8	3,316	3,427	3,055	11,744	9,335	10,993	10,471	11,362	10,130
9	3,370	3,459	3,055	11,966	9,335	11,357	10,569	11,657	10,296
10	3,415	3,490	3,067	12,179	9,407	11,664	10,704	11,919	10,475
11	3,450	3,547	3,081	12,580	9,492	11,904	10,928	12,237	10,630
12	3,450	3,582	3,087	12,828	9,532	11,900	11,058	12,355	10,650
Сумма	39,408	40,590	36,457	137,458	110,771	129,605	123,360	133,468	119,772

Подставим соответствующие значения в систему уравнений:

$$\begin{cases} 39,408=12a_0+40,59b_1+36,457b_2\ , \\ 133,468=40,59a_0+137,458b_1+123,36b_2\ , \\ 119,772=36,457a_0+123,36b_1+110,771b_2\ . \end{cases}$$

Решая уравнения, получим k_0=-6,732+0,215k_1+3,058k_2 .

Для осуществления перехода к начальному уравнению потенцируем по основанию 10:

$y_{x1x2} = 10^{-6,732}x_1^{0,215}x_2^{3,058}$, откуда $y_{x1x2}=1,85\cdot10^{-7}\cdot x_1^{0,215}x_2^{3,058}$.

Полученные коэффициенты регрессии определяют относительный прирост валового дохода на единицу прироста каждого из факторов. В степенной функции коэффициент эластичности постоянен и равен коэффициенту регрессии, поэтому $Э_{x1}=b_1=0,215$, $Э_{x2}=b_2=3,058$. При формировании валового дохода в большей степени оказывает влияние численность трудовых ресурсов, гораздо в меньше степени – основные производственные фонды. Так, с увеличением численности работников на 1% валовой доход растет в среднем на 3,06% при фиксированном значении основных фондов, а рост обеспеченности производственными фондами на 1% дает прирост валового дохода на 0,22% при неизменной численности трудовых ресурсов.

Если бы только трудовые ресурсы и основные фонды определяли величину валового дохода и темпы роста всех трех переменных (y_1,x_1,x_2) за изучаемый период были неизменны, то сумма b_1 и b_2 была бы равна единице. В рассматриваемом примере это не наблюдается. Действительно, если сопоставить данные последнего года с начальным, то валовой доход за весь период возрастает на 254%, фонды – на 238%, а число работников – на 126%. Поскольку темп роста валового дохода превосходит темпы роста факторов его определяющих, то сумма b_1 и b_2 превышает единицу. Если бы темп роста валового дохода был ниже, чем темп роста факторов x_1 и x_2, то сумма b_1 и b_2 оказалось бы меньше единицы. следовательно уравнение $y=a\ x_1^{\ b1}\ x_2^{\ b2}$ позволяет рассматривать коэффициенты регрессии b_1 и b_2 как показатели, при помощи которых оценивается эффективность факторов, участвующих в создании валового дохода.

Чтобы проверить значимость построенного уравнения регрессии вычислим индекс корреляции по формуле (табл.1.8.):

$$R_{yx} = \sqrt{1 - \frac{\sum (y - anti\log(\lg y_{x1x2}))^2}{\sum (y - \bar{y})^2}}$$

После вычисления производных показателей получим:

$$R = \sqrt{1 - \frac{119194}{3603681}} = 0{,}983$$

Индекс корреляции говорит о наличии сильной корреляционной связи между изучаемыми признаками. Значимость уравнения проверим с помощью F-критерия Фишера:

$$F_{факт} = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = \frac{0{,}966}{1-0{,}966} \cdot (12-1-1) = 284$$

Таблица 1.8

№	$\lg y_{x1x2}$	antilog ($\lg y_{x1x2}$)	$[\lg y_{x1x2}$ - antilog $(\lg y_{x1x2})]^2$	$(y - \bar{y})^2$
1	3,089	1227	13791	795813
2	3,111	1291	2	508488
3	3,147	1403	1615	312574
4	3,186	1536	13414	122558
5	3,259	1814	459	43716
6	3,279	1901	9234	26
7	3,313	2056	1190	7906
8	3,346	2221	23280	4345
9	3,353	2256	7688	116907
10	3,396	2487	13159	359900
11	3,451	2822	3	668988
12	3,478	3004	35357	662460
Сумма	39,408	24019	119194	3603681

Определим критическое значения критерия по таблице при $к_1=1$, $к_2=10$ и уровне значимости $\alpha=0{,}05$. Оно равно 4,96. Так как $F_{факт} > F_{кр}$, то гипотезу H_0 о случайном характере связи отклоняем с вероятностью 95%. Уравнение регрессии статистически значимо.

Пример 2. Имеется информация по 25 наблюдениям.

Признак	Среднее значение	Коэффициент вариации, %	Уравнение регрессии
У	23	20	$y = 20 - 2x_1 - 0{,}5x_2$
x_1	6	40	$y = 9 - 1x_1$
x_2	8	10	$y = 4 - 0{,}6x_2$

Задание.

1. Найти линейное уравнение множественной регрессии в стандартизированной форме. Определить силу влияния каждого фактора.
2. Рассчитать показатель множественной корреляции и детерминации, если известно, что $r_{x1x2} = -0{,}5$.

Решение: Уравнение в стандартизированной форме представляет собой следующий вид: $t_y = \beta_1 t_{x1} + \beta_2 t_{x2}$

Для нахождения его параметров используем уравнение в естественной форме и среднеквадратические отклонения признаков σ: $\beta_i = b_i \cdot \dfrac{\sigma_{xi}}{\sigma_y}$

Значение среднеквадратических отклонений σ найдем с помощью показателя вариации:

$$V_y = \frac{\sigma_y}{\bar{y}} \cdot 100\%, \quad V_{x1} = \frac{\sigma_{x1}}{\bar{x}_1} \cdot 100\% \qquad V_{x2} = \frac{\sigma_{x2y}}{\bar{x}_2} \cdot 100\%$$

Отсюда выразив искомые величины, получим:

$$\sigma_y = \frac{V_y \cdot \bar{y}}{100} = \frac{20 \cdot 23}{100} = 4{,}6$$

$$\sigma_{x1} = \frac{V_{x1} \cdot \bar{x}_1}{100} = \frac{40 \cdot 6}{100} = 2{,}4$$

$$\sigma_{x2} = \frac{V_{x2} \cdot \bar{x}_2}{100} = \frac{8 \cdot 10}{100} = 0{,}8$$

Теперь находим β-коэффициенты: $\beta_1 = -2 * 4{,}6/2{,}4 = -3{,}83$, $\beta_2 = -0{,}5 * 4{,}6/0{,}8 = -2{,}88$. Отсюда $t_y = -3{,}83t_{x1} - 2{,}88t_{x2}$.

По β-коэффициентам проранжируем факторы по силе влияния. Так как $|\beta_1| = 3{,}83 > |\beta_2| = 2{,}88$, то фактор x_1 оказывает большее влияние на у, чем x_2.

Расчет показателя множественной корреляции проведем через показатели парной корреляции:

$$R_{yx1x2} = \sqrt{\frac{r_{yx1}^2 + r_{yx2}^2 - 2 \cdot r_{yx1} \cdot r_{yx2} \cdot r_{x1x2}}{1 - r_{x1x2}^2}}$$

Предварительно определим парные коэффициенты корреляции r_{yx1} и r_{yx2}, используя парные уравнения регрессии:

$$r_{yx1} = b_{x1}\frac{\sigma_{x1}}{\sigma_y} = -1 \cdot \frac{2,4}{4,6} = -0,52$$

$$r_{yx2} = b_{x2}\frac{\sigma_{x2}}{\sigma_y} = -0,6 \cdot \frac{0,8}{4,6} = -0,10$$

Окончательно, получим:

$$R_{yx1x2} = \sqrt{\frac{(-0,52)^2 + (-0,1)^2 + 2 \cdot 0,52 \cdot 0,10 \cdot 0,5}{1 - (-0,5)^2}} = \sqrt{0,443} = 0,67$$

$$R^2_{yx1x2} = 0,67^2 = 0,443$$

Зависимость y от x_1 и x_2 характеризуется заметной связью, в которой 44% вариации y определяется вариацией учтенных в модели факторов. Прочие факторы, не включенные в модель, составляют соответственно 56% от общей вариации y. Это говорит о том, что следует улучшить модель, введя дополнительные факторы или изменив форму связи.

Пример 3. Известно, что по ряду регионов имеется зависимость величины импорта на определенный товар y относительно отечественного его производства x_1, изменения запасов x_2 и потребления на внутреннем рынке x_3, выражающаяся следующем уравнением множественной регрессии:

$y=66,028+0,135 \cdot x_1+0,476 \cdot x_2+0,343 \cdot x_3$.

При этом $\bar{y} = 31,5; \bar{x}_1 = 245,7; \bar{x}_2 = 3,7; \bar{x}_3 = 182,5.$

Задание.

1.Определить средние коэффициенты эластичности. Пояснить их смысл.

2. Вычислить частные уравнения регрессии.

3. Найдите частные коэффициенты эластичности для региона, если известно, что в нем x_1=160,2; x_2=4,0; x_3=190,5.

Решение. Определим средние показатели эластичности по каждому фактору:

$\overline{Э}_{y_{x_1}} = 0{,}135 \cdot \frac{245{,}7}{31{,}5} = 1{,}053\%,$ то есть с ростом величины отечественного производства на 1% размер импорта в среднем по совокупности регионов возрастет на 1,053% при неизменных запасах и потреблении семей.

Для второй переменной коэффициент эластичности составляет:

$\overline{Э}_{y_{x_2}} = 0{,}476 \cdot \frac{3{,}7}{31{,}5} = 0{,}056\%,$ то есть с ростом изменения запасов на 1% при неизменном производстве и внутреннем потреблении величина импорта увеличивается в среднем на 0,056%.

Для третьей переменной коэффициент эластичности составляет:

$\overline{Э}_{y_{x_3}} = 0{,}343 \cdot \frac{182{,}5}{31{,}5} = 1{,}987\%,$ то есть при неизменном объеме производства и величины запасов с увеличением внутреннего потребления на 1% импорт товара возрастает на 1,987%.

Средние показатели эластичности можно сравнивать друг с другом и соответственно ранжировать факторы по силе их воздействия на результат. В рассматриваемом примере наибольшее воздействие на величину импорта оказывает размер внутреннего потребления товара x_3, а наименьшее – изменение запасов x_2. Наряду со средними показателями эластичности в целом по совокупности регионов на основе частных уравнений регрессии могут быть определены частные коэффициенты эластичности для каждого региона. Частные уравнения регрессии в нашем случае составят:

$y_{x1/x2x3} = a + b_1 \cdot x_1 + b_2 \cdot \bar{x}_2 + b_3 \cdot \bar{x}_3,$

то есть $y_{x1/x2x3} = -66{,}028 + 0{,}135 \cdot x_1 + 0{,}476 \cdot 3{,}7 + 0{,}343 \cdot 182{,}5 = -1{,}669 + 0{,}135 \cdot x_1;$

$y_{x2/x1x3} = a + b_1 \cdot \bar{x}_1 + b_2 \cdot x_2 + b_3 \cdot \bar{x}_3,$

то есть $y_{x2/x1x3} = -66{,}028 + 0{,}135 \cdot 245{,}7 + 0{,}476 \cdot x_2 + 0{,}343 \cdot 182{,}5 = 29{,}739 + 0{,}476 \cdot x_2;$

$y_{x3/x1x2} = a + b_1 \cdot \bar{x}_1 + b_2 \cdot \bar{x}_2 + b_3 \cdot x_3,$

то есть $y_{x3/x1x2} = -66{,}028 + 0{,}135 \cdot 245{,}7 + 0{,}476 \cdot 3{,}7 + 0{,}343 \cdot x_3 = -31{,}097 + 0{,}343 \cdot x_3.$

Подставляя в данные уравнения фактические значения по отдельным регионам соответствующих факторов, получим значения моделируемого показателя у при заданном уровне одного фактора и средних значениях других факторов. Эти расчетные значения результативного признака используются для определения частных коэффициентов эластичности по приведенной выше формуле. Так, если в нашем случае x_1=160,2; x_2=4,0; x_3=190,5, то частные коэффициенты эластичности составят:

$$Э_{y_{x1}} = b_1 \cdot \frac{x_1}{y_{x1/x2,x3}} \text{ или } Э_{y_{x1}} = 0{,}135 \cdot \frac{160{,}2}{-1{,}669+0{,}135160{,}2} = 1{,}084\%;$$

$$Э_{y_{x2}} = b_2 \cdot \frac{x_2}{y_{x2/x1,x3}} \text{ или } Э_{y_{x2}} = 0{,}476 \cdot \frac{4{,}0}{29{,}739+0{,}4764{,}0} = 0{,}060\%;$$

$$Э_{y_{x3}} = b_3 \cdot \frac{x_3}{y_{x3/x1,x2}} \text{ или } Э_{y_{x3}} = 0{,}343 \cdot \frac{190{,}5}{-31{,}097+0{,}343190{,}5} = 1{,}908\%.$$

Как видим, частные коэффициенты эластичности для региона несколько отличаются от аналогичных средних показателей по совокупности регионов. Они могут быть использованы при принятии решений относительно развития конкретных регионов.

1.8. Адекватность построения модели множественной регрессии

Оценка надежности и значимости модели множественной регрессии проводится аналогично оценке адекватности построения парной регрессии. Значимость уравнения множественной регрессии в целом, так же как и в парной регрессии, оценивается с помощью **F-критерия Фишера**:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}, \quad (1.51.)$$

где R^2 – множественный коэффициент (индекс) детерминации.

В определенной степени качество построенной модели множественной регрессии оценивает сам **множественный показатель детерминации**. Чем ближе его величина к 1, тем лучше модель аппроксимирует исходные данные, надежнее построенное уравнение.

Независимо от формы связи этот показатель может быть найден как квадрат индекса множественной корреляции.

$$R^2_{yx1...xp} = 1 - \frac{\sigma^2_{ост}}{\sigma^2_{общ\,y}} \quad (1.52.)$$

где $\sigma^2_{ост}$ - остаточная дисперсия для уравнения $y = f(x_1, x_2, ..., x_p)$;

$\sigma^2_{общ\ y}$ - общая дисперсия результативного признака

В рассматриваемых показателях множественной корреляции используется остаточная дисперсия, которая имеет систематическую ошибку в сторону преуменьшения, тем более значительную, чем больше параметров определяется в уравнении регрессии при заданном объеме наблюдений n. Если число параметров при x_j равно m и приближается к объему наблюдений, то остаточная дисперсия будет близка к нулю и коэффициент корреляции приблизится к единице даже при слабой связи факторов с результатом. Для того чтобы не допустить возможного преувеличения тесноты связи, используется скорректированный коэффициент (индекс) множественной корреляции и детерминации. Скорректированный индекс множественной детерминации содержит поправку на число степеней свободы, а именно остаточная сумма квадратов делится на число степеней свободы остаточной вариации (n-m-1), а общая сумма квадратов отклонений – на число степеней свободы в целом по совокупности (n-1). Формула **скорректированного индекса множественной детерминации** имеет вид:

$$\overline{R}^2 = 1 - \frac{\sum (y - y_{x1...xp})^2 : (n-m-1)}{\sum (y - \overline{y})^2 : (n-1)}, \quad (1.53.)$$

где m – число параметров при переменных x; n – число наблюдений;

$y_{x1...xp}$ – теоретические значения y, вычисленные по уравнения регрессии.

Поскольку $\sum (y - y_{x1...xp})^2 / \sum (y - \overline{y})^2 = 1 - R^2$, то величину скорректированного индекса детерминации можно представить в виде:

$$\overline{R}^2 = 1 - (1 - R^2) \cdot \frac{(n-1)}{(n-m-1)}, \quad (1.54.)$$

где R^2 – обычный индекс множественной детерминации.

Чем больше величина m, тем сильнее различия $\overline{R}^2$ и R^2. Для линейной зависимости признаков скорректированный коэффициент множественной корреляции определяется по той же формуле, что и индекс множественной корреляции, т.е. как корень квадратный из $\overline{R}^2$. Отличие состоит лишь в том, что в линейной зависимости под m подразумевается число факторов, включенных в регрессионную модель, а в криволинейной зависимости m – число параметров при x и их преобразованиях, которое, может быть, больше числа факторов как экономических переменных. Так, если $y = f(x_1, x_2)$, то для линейной регрессии m=2, а для регрессии вида: $y = a + b_1 \cdot x_1 + b_{12} \cdot x_1^2 + b_2 \cdot x_2 + b_{22} \cdot x_2^2$ число параметров при x равно 4, т.е. m=4. При заданном объеме число параметров при прочих равных условиях с увеличением числа независимых переменных (параметров) скорректированный коэффициент множественной детерминации убывает. Его величина может стать и отрицательной при слабых связях результата с факторами. В этом случае он должен считаться равным нулю.

Также присутствует оценка значимости отдельных коэффициентов множественной регрессии по **t-критерию Стьюдента**, позволяющая определять включение каждого фактора в модель. В этом случае, как и в парной регрессии, для каждого фактора используются формулы:

$$t_{bi} = \frac{b_i}{m_{bi}}, \quad \mathrm{m_{bi}} = \frac{\sigma_y}{\sigma_{xi}} \cdot \sqrt{\frac{1-R^2_{yx1...xp}}{1-R^2_{xi\ \mathrm{x1...xp}}} \cdot \frac{1}{n-m-1}}, \ (1.55.)$$

где σ_y – среднее квадратическое отклонение для признака y;

σ_{xi} – среднее квадратическое отклонение для признака x_i;

$R^2_{yx1\ldots xp}$ – коэффициент детерминации для уравнения множественной регрессии;

$R^2_{xi\ x1\ldots xp}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии.

Кроме этого способа, во множественной регрессии появляется еще один – F-критерий частный, определяющий целесообразность включения отдельного фактора после всех остальных. Необходимость такой оценки связана с тем, что не каждый фактор, вошедший в модель, может существенно увеличивать долю объясненной вариации результативного признака. Кроме того, при наличии в модели нескольких факторов они могут вводиться в модель в разной последовательности. Ввиду корреляции между факторами значимость одного и того же фактора может быть разной в зависимости от последовательности его введения в модель. **Частный F-критерий** построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительно включенного фактора, с остаточной дисперсией на одну степень свободы по регрессионной модели в целом. Используется следующая формула:

$$F_{част\ xi} = \frac{R^2{}_{yx1...xp} - R^2_{yx1...x_{i-1}x_{i+1}...xp}}{1-R^2_{yx1...xp}} \cdot \frac{n-m-1}{1}, \ (1.56.)$$

где $R^2_{x1\ldots xp}$ – индекс множественной детерминации для модели с полным набором факторов;

$R^2_{x1\ldots xi-1,xi+1\ldots xp}$ – индекс множественной детерминации, но без включения в модель фактора x_i.

Фактическое значение частного F-критерия сравнивается с критическим (табличным) при 5%- или 1%-ом уровне значимости и числе степеней свободы: $k_1=1$ и $k_2=n-m-1$. Если фактическое превышает критическое, то дополнительное включение фактора x_i в модель после факторов $x_1,x_2..x_{i-1},x_{i+1}\ldots x_p$ статистически оправданно и коэффициент чистой регрессии b_i при факторе x_i статистически значим. Если наоборот, то дополнительное включение в модель фактора x_i после факторов $x_1,x_2..x_{i-1},x_{i+1}\ldots x_p$ не увеличивает существенно долю объясненной вариации признака, следовательно, нецелесообразно его включение в модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

С помощью частного F-критерия можно проверить значимость всех коэффициентов регрессии в предположении, что каждый соответствующий фактор x_i вводился в уравнение множественной регрессии последним. Так как частный F-критерий косвенно оценивает значимость коэффициентов чистой регрессии, то существует взаимосвязь между ним и t-критерием для коэффициента регрессии при i-м факторе:

$$t_{bi} = \pm\sqrt{F_{част\ \mathrm{xi}}} \quad (1.57.)$$

Знак критерия зависит от знака коэффициента регрессии, для которого его определяют.

Взаимосвязь показателей частного F-критерия и t-критерия для коэффициентов чистой регрессии может использоваться в процедуре отбора факторов. Отсев факторов при построении уравнения регрессии методом исключения можно осуществлять не только по частным коэффициентам корреляции, исключая на каждом шаге фактор с наименьшим незначимым значением частного коэффициента корреляции, но и по величине t_{bi} и $F_{част\ xi}$.

Контрольные вопросы:

- С помощью каких критериев оценивается надежность в уравнении множественной регрессии?

- Какая связь между частными F-критериями и t_{bi} –критериями параметров уравнения?
- Для чего нужны скорректированные индекс множественной корреляции и индекс детерминации?
- Что показывает множественный коэффициент детерминации?
- Как оценить целесообразность включения в модель фактора?
- Можно ли оценить надежность уравнения множественной регрессии с помощью средней ошибки аппроксимации?
- В чем различия между скорректированным и нескорректированным коэффициентом множественной корреляции?

Решение типовых задач

Пример 1. По 30 сельскохозяйственным предприятиям Смоленской области имеются данные, характеризующие зависимость y (валовая продукция на 1 работника, тыс.руб.) от x_1 (среднедневная заработная плата одного работающего, руб.) и x_2 (удельный вес сельскохозяйственной техники, находящейся за пределами сроков эксплуатации, %): $\bar{y}=86{,}8$, $\bar{x}_1=54{,}9$, $\bar{x}_2=33{,}5$, $\sigma_y=11{,}44$, $\sigma_{x1}=5{,}86$, $\sigma_{x2}=0{,}58$. Коэффициенты корреляции: $r_{yx1}=0{,}841$, $r_{yx2}=-0{,}210$, $r_{x1x2}=-0{,}116$.

Задание.

1. Построить уравнение множественной регрессии в стандартизированной и естественной формах, рассчитать частные коэффициенты эластичности, сравнить, пояснить различия между ними.

2.Определить коэффициент множественной корреляции (скорректированный и нескорректированный).

Решение. Линейное уравнение множественной регрессии в естественной форме y от x_1 и x_2 имеет вид: $y=a+b_1x_1+b_2x_2$ Для расчета его параметров применим метод стандартизации переменных, предварительно построив уравнение в стандартизированной форме $t_y=\beta_1 t_{x1}+\beta_2 t_{x2}$.

Расчет β-коэффициентов из системы нормальных уравнений (2.7.) выполним по формулам:

$$\beta_1 = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{1 - r^2_{x1x2}} = \frac{0{,}841 - 0{,}21 \cdot 0{,}116}{1 - 0{,}116^2} = 0{,}827$$

$$\beta_2 = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{1 - r^2_{x1x2}} = \frac{-0{,}21 + 0{,}841 \cdot 0{,}116}{1 - 0{,}116^2} = -0{,}114$$

Получим уравнение t_y =0,827*t_{x1} - 0,114*t_{x2}. Стандартизированные коэффициенты регрессии можно использовать при отсеве факторов в модели, исключая из уравнения те из них, которые имеют наименьшее значение β-коэффициента. Для построения уравнения в естественной форму используем формулы связи между b_i и β_i :

$$\beta_i = b_i \cdot \frac{\sigma_{xi}}{\sigma_y}, \quad откуда\ b_i = \beta_i \cdot \frac{\sigma_y}{\sigma_{xi}}$$

b_1=0,827*11,44/5,86=1,62 , b_2=-0,114*11,44/0,58=-2,25

Значение а определяем из соотношения: a= $\bar{y}$ –b_1 $\bar{x}_1$-b_2 $\bar{x}_2$ = 86,8-1,62*54,9+2,25*33,5 = -73,5. Теперь можно записать уравнение:

y_{x1x2}= - 73,5 + 1,62x_1 – 2,25x_2

Данное уравнение показывает, что с ростом заработной платы на 1 руб. валовая продукция на 1 работающего в среднем увеличится на 1,62 тыс. руб., а увеличение удельного веса изношенной техники наоборот, снижает объем валовой продукции на 1 работника на 2,25 тыс. руб.

Для характеристики относительной силы влияния x_1 и x_2 на y рассчитаем средние коэффициенты эластичности:

$$Э = b_i \cdot \frac{\bar{x}}{\bar{y}}$$

Следовательно, $\bar{Э}_{x1}$=1,62*54,9/86,8=1,025%, $\bar{Э}_{x2}$=-2,25*33,5/86,8=00,868%.

С увеличением средней заработной платы на 1% от ее среднего уровня валовая продукция на 1 работника возрастает на 1,03% от своего среднего уровня, а рост износа сельскохозяйственной техники на 1% снижает валовую продукцию на 0,87% от своего среднего уровня. Очевидно, что сила влияния средней заработной платы x_1 на производительность труда оказалась большей, чем сила влияния степени износа техники x_2. К аналогичным выводам о силе связи приходим при сравнении модулей значений β_1 и β_2 : $|\beta_1|=0,827 > |\beta_2|=0,114$.

Различия в силе влияния фактора на результат, полученные при сравнении $Э_{xi}$ и β_i , объясняются тем, что коэффициент эластичности исходит из соотношения средних , а β-коэффициент – из соотношения квадратических отклонений.

Расчет линейного коэффициента множественной корреляции выполним с использованием коэффициентов r_{yxi} и β_i :

$$R_{yx1x2}=\sqrt{\sum \beta_i \cdot r_{yxi}}=\sqrt{0,841\cdot 0,827+0,21\cdot 0,114}=\sqrt{0,72}=0,848$$

Связь y от x_1 и x_2 по шкале Чеддока характеризуется как сильная, в которой 72% вариации производительности труда определяется вариацией учтенных в модели факторов: средней заработной платой и степени износа техники. Прочие факторы, не включенные в модель, составляют соответственно 28% от общей вариации y.

Скорректированный коэффициент корреляции рассчитаем по формуле:

$$\overline{R}=\sqrt{1-(1-R^2)\cdot\frac{(n-1)}{(n-m-1)}}=\sqrt{1-(1-0,72)\cdot\frac{30-1}{30-2-1}}=0,836$$

Скорректированный коэффициент корреляции немного уменьшил обычный показатель корреляции, сделав поправку на число степеней свободы.

Пример 2. По 20 хозяйствам Смоленской области изучается зависимость объема валовой продукции на 100 га с/х угодий y (тыс.руб.) от трудообеспеченности на 100 га с/х угодий x_1 (чел.) и от доли квалифицированных кадров в общей численности x_2(%) – табл. 1.8.

Задание.

1. Определить F-критерий и статистическую значимость уравнения множественной регрессии и его показателя тесноты связи.
2. С помощью частных F-критериев Фишера оценить, насколько целесообразно включение в уравнение множественной регрессии фактора x_1 после

фактора x_2 и насколько целесообразно включение x_2 после x_1.

3. Оценить с помощью t-критерия Стьюдента статистическую значимость коэффициентов при переменных x_1 и x_2 множественного уравнения регрессии.

Таблица 1.8

признак	среднее значение	среднее квадратическое отклонение	характеристика тесноты связи	уравнение связи
y	112,76	31,58	R_{yx1x2}=0,773	y_{x1x2}=-30,49+6,14*x_1+4,13*x_2
x_1	5,4	3,34	r_{yx1}=0,746	y_{x1}=74,4+7,1*x1
x_2	50,88	1,74	r_{yx2}=0,507 r_{x1x2}=0,432	y_{x2}=-355,3+9,2*x2

Решение. Общий F-критерий проверяет гипотезу H_0 о статистической значимости уравнения регрессии и показателя тесноты связи:

$$F_{факт}=\frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m}=\frac{0,773^2}{1-0,773^2}\cdot\frac{20-2-1}{2}=12,62$$

F $_{кр}$ (2;17;0,05) = 3,59. $F_{факт} > F_{кр}$, гипотезу H_0 отклоним. С вероятностью 95% можно говорить о статистической значимости уравнения в целом. Показатель детерминации R^2 = 0,60 или 60% вариации результативного признака y объясняется влиянием факторов x_1 и x_2.

Сначала оценим целесообразность включения в уравнение множественной регрессии фактора x_1 после x_2. Для этого используем частный F- критерий Фишера с учетом того, что в уравнении два фактора:

$$F_{част\ x1}=\frac{R^2{}_{yx1x2}-r^2_{yx2}}{1-R^2_{yx1x2}}\cdot\frac{n-m-1}{1}=\frac{0,60-0,507^2}{1-0,60}\cdot\frac{20-2-1}{1}=14,4$$

F $_{кр}$ = (1;17;0,05)= 4,45, Так как $F_{част\ x1}$ > F $_{кр}$, то включение x_1 после x_2 оказалось статистически значимым, этот фактор оказывает существенное влияние на результат. Проверим целесообразность включения x_2 после x_1.

$$F_{част\ x2}=\frac{R^2{}_{yx1x2}-r^2_{yx2}}{1-R^2_{yx1x2}}\cdot\frac{n-m-1}{1}=\frac{0,60-0,746^2}{1-0,60}\cdot\frac{20-2-1}{1}=1,73$$

Для сравнения используем то же табличное (критическое) значение $F_{кр}$ = 4,45. В связи с тем, что F $_{част\ x2}$ < F $_{кр}$, включение x_2 после x_1 оказалось бесполезным.

Оценка с помощью t – критерия Стьюдента значимости коэффициентов b_1 и b_2 связано с сопоставлением их значений с величиной случайных ошибок m_{b1} и m_{b2} .

$$t_{b1} = \sqrt{F}_{\text{част } x1} = 3{,}79$$

$$t_{b2}=1{,}32$$

$$t_{кр}(17;0{,}05)=2{,}1098$$

Так как $t_{b1}>t_{кр}$., то коэффициент регрессии b_1является надежным и статистически значимым с вероятностью 95%, его можно использовать при анализе и прогнозе. Но так как $t_{b2}< t_{кр.}$, то величина b_2 – статистически незначима и с вероятностью 95% можно утверждать, что он незначительно отличается от нуля.

Проверка по t-критерию Стьюдента еще раз показала статистическую значимость одновременного влияния x_1 и x_2 на y и не значимость влияния x_2. В дальнейшем при анализе и прогнозе лучше использовать модель парной регрессии y от x_1 .

Пример 3. По 30 наблюдениям матрица парных коэффициентов корреляции оказалась следующей:

	y	x_1	x_2	x_3
y	1,00			
x_1	0,30	1,00		
x_2	0,60	0,10	1,00	
x_3	0,40	0,15	0,80	1,00

Задание.

1. Постройте уравнение регрессии в стандартизованном виде и сделайте выводы.
2. Определите показатель множественной корреляции (нескорректированный и скорректированный).
3. Оцените целесообразность включения переменной x_1 в модель после введения в нее переменных x_2 и x_3.

Решение. Кроме линейного уравнения множественной регрессии в естественной форме $y=a+b_1x_1+b_2x_2+b_3x_3$ применяется уравнение в стандартизированной форме $t_y=\beta_1 t_{x1}+\beta_2 t_{x2}+\beta_3 t_{x3}$. Расчет β-коэффициентов осуществим из системы нормальных уравнений:

$$\begin{cases} r_{yx1}=\beta_1+\beta_2\cdot r_{x2x1}+\beta_3\cdot r_{x3x1} \\ r_{yx2}=\beta_1\cdot r_{x1x2}+\beta_2+\beta_3\cdot r_{x3x2} \\ r_{yx3}=\beta_1\cdot r_{x1x3}+\beta_2\cdot r_{x2x3}+\beta_3 \end{cases}, \text{ откуда } \begin{cases} 0{,}3=\beta_1+\beta_2\cdot 0{,}10+\beta_3\cdot 0{,}15 \\ 0{,}6=\beta_1\cdot 0{,}10+\beta_2+\beta_3\cdot 0{,}8 \\ 0{,}4=\beta_1\cdot 0{,}15+\beta_2\cdot 0{,}8+\beta_3 \end{cases}$$

Получим β_1=0,262 , β_2=0,792 , β_3= -0,273.

Значит уравнение будет $t_y=0{,}262*t_{x1}+0{,}792*t_{x2}-0{,}273*t_{x3}$. Сравнивая модули β-коэффициентов можно определить силу влияния каждого фактора на результативный признак. Так как $|\beta_2=0{,}792|>|\beta_3=-0{,}273|>|\beta_2=0{,}262|$, то наибольшее влияние на y оказывает фактор x_2, затем идет x_3 и x_1. Причем зависимость между y и фактором x_3 имеет обратную направленность. Парный коэффициент корреляции r_{yx3} неправильно характеризует направление связи из-за сильной мультиколлинеарности между факторами (например, факторы x_2 и x_3 явно коллинеарны, так как $r_{x2x3}=0{,}8>0{,}7$).

Показатель множественной корреляции вычисляется через β-коэффициенты и коэффициенты парной корреляции:

$$R_{yx_1x_2x_3}=\sqrt{\beta_1\cdot r_{yx1}+\beta_2\cdot r_{yx2}+\beta_3\cdot r_{yx3}}=\sqrt{0{,}262\cdot 0{,}3+0{,}792\cdot 0{,}6-0{,}273\cdot 0{,}4}=0{,}744$$

Линейный коэффициент множественной корреляции говорит о сильном характере связи между y и тремя факторами. Определим скорректированный коэффициент множественной корреляции как корень квадратный из скорректированного индекса детерминации $\bar{R}^2$:

$$\bar{R}=\sqrt{\bar{R}^2}=\sqrt{1-(1-R^2)\cdot\frac{(n-1)}{(n-m-1)}}=\sqrt{1-(1-0{,}774^2)\cdot\frac{30-1}{30-3-1}}=0{,}7435$$

Скорректированный показатель $\bar{R}$, содержащий поправку на число степеней свободы, незначительно отличается от обычного R. Эта дает возможность при анализе ограничиться обычным показателем корреляции.

Оценим целесообразность включения переменной x_1 в модель после введения в нее переменных x_2 и x_3, использовав частный F-критерий Фишера.

$$F_{част\ x1} = \frac{R^2_{yx1x2x3} - R^2_{yx2x3}}{1 - R^2_{yx1x2x3}} \cdot \frac{n-m-1}{1},$$

где R^2_{yx2x3} – линейный коэффициент множественной детерминации для модели с факторами x_2 и x_3 без фактора x_1.

Для расчета критерия следует вначале найти R^2_{yx2x3}. Для этого воспользуемся формулой:

$$R_{yx2x3} = \sqrt{\frac{r^2_{yx2} + r^2_{yx3} - 2 \cdot r_{yx2} \cdot r_{yx3} \cdot r_{x2x3}}{1 - r^2_{x2x3}}} = 0{,}615$$

отсюда $F_{част\ x1} = \dfrac{0{,}5535 - 0{,}3782}{1 - 0{,}5535} \cdot \dfrac{30-3-1}{1} = 10{,}2$

$F_{кр}$ = (1;26;0,05)= 4,22, Так как $F_{част\ x1} > F_{кр}$, то включение x_1 после факторов x_2 и x_3 оказалось статистически значимым, этот фактор оказывает существенное влияние на результат.

РАЗДЕЛ II. ВРЕМЕННЫЕ РЯДЫ И СИСТЕМА ОДНОВРЕМЕННЫХ УРАВНЕНИЙ

2.1. Временной ряд и спецификация его исследований

Для построения эконометрической модели используют два типа исходных данных:

• данные, характеризующие совокупность различных объектов в определенный момент (период) времени;

• данные, характеризующие один объект за ряд последовательных моментов (периодов) времени.

Модели, построенные по данным первого типа, называются **пространственными** моделями. Модели, построенные на основе второго типа данных, называются **моделями временных рядов**.

Временной ряд — это совокупность значений какого-либо показателя за несколько последовательных моментов или периодов времени. Каждый уровень временного ряда формируется под воздействием большого числа факторов, которые условно можно подразделить на **три группы (три компоненты)**:

• факторы, формирующие тенденцию ряда;

• факторы, формирующие циклические колебания ряда;

• случайные факторы.

В большинстве случаев фактический уровень временного ряда можно представить как сумму или произведение трендовой, циклической и случайной компонент. Модель, в которой временной ряд представлен как сумма перечисленных компонент, называется аддитивной моделью временного ряда. Модель, в которой временной ряд представлен как произведение перечисленных компонент, называется мультипликативной моделью временного ряда.

Основная задача эконометрического исследования отдельного временного ряда — выявление и придание количественного выражения каждой из перечисленных выше компонент с тем, чтобы использовать полученную информацию для прогнозирования будущих значений ряда или при построении моделей взаимосвязи двух или более временных рядов.

Наибольшую сложность представляет взаимосвязанное изучение двух и более временных рядов, т.е. рассмотрение причинно-следственных зависимостей переменных, представленных в форме временных рядов. Применение в этих целях традиционных методов корреляционно-регрессионного анализа, рассмотренных ранее, может привести к ряду серьезных проблем, возникающих как на этапе построения, так и на этапе анализа эконометрических моделей. В первую очередь эти проблемы связаны со спецификой временных рядов как источника данных в эконометрическом моделировании.

Рассмотрим, например, как сказывается на результатах корреляционно-регрессионного анализа, присутствие во временном ряду трех основных компонент**.**

На этапе предварительного анализа выявляется структура изучаемых временных рядов. Если на этом этапе было выявлено, что временные ряды содержат сезонные или циклические колебания, то перед проведением дальнейшего исследования взаимосвязи необходимо устранить **сезонную** или **циклическую** компоненту из уровней каждого ряда, поскольку ее наличие приведет к завышению истинных показателей силы и тесноты связи изучаемых временных рядов в случае, если оба ряда содержат циклические колебания одинаковой периодичности, либо к занижению этих показателей в случае, если сезонные или циклические колебания содержат только один из рядов или периодичность колебаний в рассматриваемых временных рядах различна. Устранение сезонной компоненты из уровней временных рядов можно проводить в соответствии с методикой построения аддитивной и мультипликативной моделей.

Кроме этого, если рассматриваемые временные ряды имеют **тенденцию**, то коэффициент корреляции между признаками по абсолютной величине будет высоким (положительным в случае совпадения и отрицательным в случае противоположной направленности тенденций рядов x и y). Однако из этого еще нельзя делать вывод о том, что x причина y, или наоборот. Высокий коэффициент корреляции в данном случае есть результат того, что x и y зависят от времени, или содержат тенденцию. При этом одинаковую или противоположную тенденцию могут иметь ряды, совершенно не связанные друг с другом причинно-следственной зависимостью.

Например, коэффициент корреляции между числом больничных коек и количеством учреждений культурно-досугового типа в Смоленской области за период с 1985 по 2005 г. составил 0,91. Это, естественно, не означает, что увеличение количества учреждений культуры и досуга способствует росту числа больничных коек или увеличение числа последних стимулирует спрос на культурно-увеселительные мероприятия.

При наличии во временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Корреляционную зависимость между последовательными уровнями временного ряда называют **автокорреляцией уровней ряда**.

Количественно ее можно измерить с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями этого ряда, сдвинутыми на несколько шагов во времени.

Пусть дан временной ряд y_1 , y_2, ..y_n , тогда коэффициент корреляции между рядами y_t и y_{t-1} (т.е., например, между текущем и предыдущем годом, если заданы годичные данные) можно определить как:

$$r_1 = \frac{\sum_{t=2}^{n}(y_t - \bar{y}_1)\cdot(y_{t-1} - \bar{y}_2)}{\sqrt{\sum_{t=2}^{n}(y_t - \bar{y}_1)^2 \cdot \sum_{t=2}^{n}(y_{t-1} - \bar{y}_2)^2}}, \quad (2.1)$$

где $\bar{y}_1 = \frac{\sum_{t=2}^{n} y_t}{n-1}$, $\bar{y}_2 = \frac{\sum_{t=2}^{n} y_{t-1}}{n-1}$

Величину r_1 называют **коэффициентом автокорреляции** уровней ряда

первого порядка, так как он измеряет зависимость между соседними уровнями ряда t и t-1, т. е. при лаге 1. Аналогично можно определить коэффициенты автокорреляции второго и более высоких порядков.

Отметим два важных свойства коэффициента автокорреляции. **Во-первых**, он строится по аналогии с линейным коэффициентом корреляции и таким образом характеризует тесноту только линейной связи текущего и предыдущего уровней ряда. Поэтому по коэффициенту автокорреляции можно судить о наличии линейной (или близкой к линейной) тенденции. Для некоторых временных рядов, имеющих сильную нелинейную тенденцию (например, параболу второго порядка или экспоненту), коэффициент автокорреляции уровней исходного ряда может приближаться к нулю. **Во-вторых**, по знаку коэффициента автокорреляции нельзя делать вывод о возрастающей или убывающей тенденции в уровнях ряда. Большинство временных рядов экономических данных содержит положительную автокорреляцию уровней, однако при этом могут иметь убывающую тенденцию.

Последовательность коэффициентов автокорреляции уровней первого, второго и т. д. порядков называют **автокорреляционной функцией** временного ряда. График зависимости ее значений от величины лага (порядка коэффициента автокорреляции) называется **коррелограммой**.

Анализ автокорреляционной функции и коррелограммы позволяет определить лаг, при котором автокорреляция наиболее высокая, а следовательно, и лаг, при котором связь между текущим и предыдущими уровнями ряда наиболее тесная, т. е. при помощи анализа автокорреляционной функции и коррелограммы можно выявить **структуру ряда**.

Если наиболее высоким оказался коэффициент автокорреляции первого порядка, исследуемый ряд содержит только тенденцию. Если наиболее высоким оказался коэффициент автокорреляции порядка г, ряд содержит не только тенденцию, но и циклические колебания с периодичностью в r моментов времени. Если ни один из коэффициентов автокорреляции не является значимым, можно сделать одно из двух предположений относительно структуры этого ряда; либо ряд не содержит тенденции и циклических колебаний и имеет структуру, сходную со структурой ряда, подверженного лишь воздействию **случайного** фактора, либо ряд содержит сильную нелинейную тенденцию, для выявления которой нужно провести дополнительный анализ.

Наличие тенденции в каждом из временных рядов означает, что на зависимую переменную y и независимую x оказывает воздействие фактор времени, который непосредственно в модели не учтен. Влияние фактора времени будет выражено в корреляционной зависимости между значениями остатков ε за текущий и предыдущие моменты времени, которая получила название «**автокорреляция в остатках**».

Автокорреляция в остатках есть нарушение одного из условий применения МНК — предпосылки о случайности остатков, полученных по уравнению регрессии. Один из возможных путей решения этой проблемы состоит в применении к оценке параметров модели обобщенного МНК. При построении уравнения множественной регрессии по временным рядам данных, помимо двух вышеназванных проблем, возникает также проблема мультиколлинеарности факторов, входящих в уравнение регрессии, в случае если эти факторы содержат тенденцию.

Контрольные вопросы:

- Какие модели называются пространственными моделями?
- Что такое временной ряд?
- Как рассчитывается коэффициент автокорреляции уровней первого порядка, второго и т.д.?
- Что такое коррелограмма?
- Для чего служит автокорреляционная функция?
- Какие компоненты может содержать временной ряд?
- Что такое лаг?
- Охарактеризуйте явление «автокорреляция в остатках».
- Что представляет собой аддитивная и мультипликативная модели

временного ряда?

Решение типовых задач

Пример 1. На основе данных по продуктивности коров за 24 месяца получены следующие значения коэффициентов автокорреляции уровней первого и т.д. порядков:

$r_1 = 0,67$; $r_2 = 0,43$; $r_3 = 0,75$; $r_4 = 0,57$; $r_5 = 0,38$; $r_6 = 0,88$.

Задание.

1. Охарактеризуйте структуру ряда, определив наличие тенденции и циклических колебаний.

2. Обоснуйте выбор уравнения регрессии для прогноза будущих значений ряда.

Решение:

Так как значение коэффициентов автокорреляции первого порядка и нескольких других достаточно высокие, то ряд содержит тенденцию и циклические колебания. Наиболее ощутимые периодические колебания наблюдаются с циклом равном 6 (максимальным по абсолютной величине является $r_6 = 0,88$).

Наиболее целесообразным будет построение уравнение авторегрессии (в качестве переменных выступают уровни ряда сдвинутые на несколько периодов) $y_t=a+by_{t-6}$, так как $r_6 = 0,88$ свидетельствует о наличии очень тесной связи между уровнями ряда с лагом в 6 месяцев.

Кроме этого, возможно построение и множественного уравнения авторегрессии y_t от y_{t-3} и y_{t-6}, так как r_3 тоже достаточно высок:

$y_t=a+b_1y_{t-3}+b_2y_{t-6}$

Выбор лучшего уравнения можно осуществить на основе показателя детерминации.

2.2. Моделирование тенденции временного ряда. Автокорреляция в остатках.

При описании одномерных динамических рядов изменения уровней может быть охарактеризовано плавной траекторией, характеризующей зависимость уровней временного ряда от времени (тренд).

Непосредственное выделение тренда может осуществляться 3 способами:

- метод укрупнения интервалов;
- аналитическое выравнивание временного ряда;
- метод скользящей средней.

Первый метод основан на **укрупнении периодов** времени, к которым относятся уровни ряда. Например, ряд ежегодного производства можно заменить рядом 5-летнего производства.

Аналитическое выравнивание является самым распространенным способов моделирования тенденции временного ряда, сущность которого заключается в построении временной аналитической функции, характеризующей зависимость уровней ряда от времени, или **тренда**. Уравнение тренда позволяет получить выровненные значения уровней динамического ряда или теоретические значения.

Процедура выравнивания ряда включает в себя следующие этапы:

- выбор одного или нескольких уравнений тренда, форма которых соответствует характеру изменения временного ряда;
- оценка параметров выбранных кривых;
- проверка адекватности построенных моделей прогнозируемому процессу и окончательный выбор уравнения тренда;
- расчет точечного и интервального прогноза.

Существует несколько способов определения типа тенденции. К числу наиболее распространенных способов относятся **качественный анализ** изучаемого процесса, построение и визуальный анализ **графика** зависимости уровней ряда от времени. В этих же целях можно использовать и коэффициенты **автокорреляции уровней** ряда. Если временной ряд имеет линейную тенденцию, то его соседние уровни y_t, и y_{t-1} тесно коррелируют. В этом случае коэффициент автокорреляции первого порядка уровней исходного ряда должен быть высоким. Если временной ряд содержит нелинейную тенденцию, например, в форме экспоненты, то коэффициент автокорреляции первого порядка по логарифмам уровней исходного ряда будет выше, чем соответствующий коэффициент, рассчитанный по уровням ряда.

Если динамика характеризуется более или менее стабильным абсолютным приростом, то математическим выражением такой тенденции будет являться прямая линия. В случае, если абсолютный прирост равномерно увеличивается, в качестве приближенного математического выражения этой тенденции можно принять параболу второго порядка, если же относительно стабильными являются темпы роста, то закономерность будет выражаться показательной кривой.

Для построения трендов чаще всего применяются следующие функции:

- линейный тренд – $y_t=a+bt$
- гипербола – $y_t=a+b/t$
- экспоненциальный тренд – $y_t=e^{a+bt}$
- тренд в форме степенной функции – $y_t=a*t^b$
- парабола второго и более высоких порядков – $y_t=a+b_1t+b_2t^2+\ldots$

Параметры каждого из перечисленных выше трендов можно определить обычным методом МНК, используя в качестве независимой переменной время (t=1,2,..., n), а в качестве зависимой переменной — фактические уровни временного ряда у. Для нелинейных трендов предварительно проводят стандартную процедуру их линеаризации.

Оценка качества аналитического выравнивания проводится с использованием уже рассмотренных критериев оценки адекватности модели в корреляционно-регрессионном анализе (показателя корреляции и детерминации, F- и t-критерии и т.д.). В качестве критерия можно выбирать остаточную дисперсию, коэффициент автокорреляции ряда отклонений, определенного по разнице фактических и выровненных значений по какой-либо аналитической функции данных.

Автокорреляция ряда отклонений (остатков) есть нарушение одного из условий применения МНК - предпосылки о случайности остатков, полученных по уравнению регрессии. При ее наличии надежность построенной модели будет низкой. Проверка гипотезы о наличии или отсутствии автокорреляции остатков проводится с помощью критерия **Дарбина-Уотсона**:

$$d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})}{\sum \varepsilon_t^{\,2}} \quad (2.2)$$

Алгоритм выявления автокорреляции остатков на основе критерия Дарбина — Уотсона следующий. Выдвигается гипотеза H_0 об отсутствии автокорреляции остатков. Альтернативные гипотезы H_1 и H_1^* состоят, соответственно, в наличии положительной или отрицательной автокорреляции в остатках. Далее по специальным таблицам (см. прил. 2) определяются критические значения критерия Дарбина — Уотсона d_L и d_u для заданного числа наблюдений n, числа независимых переменных модели k и уровня значимости α. По этим значениям числовой промежуток [0;4] разбивают на пять отрезков (рис.2.1).

Есть положительная автокорреляция остатков, H_0 отклоняется, с вероятностью	Зона неопределенности	Нет оснований отклонять H_0, автокорреляция отсутствует	Зона неопределенности	Есть от… тельная корреляц… остатков, отклоняет… вероятнос…
0 1-α принимается H_1	d_L d_u	4-d_L	4 - d_u	1-α 4 прин… ся H_1^*

Рис.2.1. Зоны определения автокорреляции по критерию Дарбина-Уотсона

Если фактическое значение критерия Дарбина — Уотсона попадает в зону неопределенности, то на практике предполагают существование автокорреляции остатков и отклоняют гипотезу.

Во многих случаях присутствие автокорреляции в остатках обусловлено неправильным выбором формы связи или недоучетом какого-либо существенного фактора.

Метод скользящей средней основан на укрупнении периодов, образованных последовательным исключением начального ряда и замены его очередным. Пусть дан следующий временной ряд y_1, y_2,… y_n. Тогда, если выравнивание происходит по пяти членам временного ряда, то новый ряд будет:

$$\bar{y}_1 = \frac{y_1 + y_2 + \ldots + y_5}{5},$$

$$\bar{y}_2 = \frac{y_2 + y_3 + \ldots + y_6}{5}, \quad (2.3)$$

$$и\ т.д.$$

Число уровней, по которому рассчитывается среднее значение, называется **интервалом сглаживания**. Может быть четным или нечетным. При нечетном сглаживании полученное среднеарифметическое закрепляется за серединой расчетного интервала, при четном проводится дополнительное центрирование расчетных средних.

Интервал сглаживания обеспечивает взаимное погашение случайных отклонений во временном ряду. Выбор величины интервала должен осуществляться с учетом особенностей динамики рассматриваемого показателя. Если наблюдается определенная цикличность изменения показателей, интервал скольжения должен быть равен продолжительности цикла. При отсутствии цикличности в изменении показателей рекомендуется производить многовариантный расчет при изменяющемся параметре сглаживания. При этом появляется возможность определять среднее прогнозное значение для планового периода в целом.

Этот метод используется часто в тех случаях, когда ряды динамики характеризуются резкими колебаниями показателей по годам. Такие ряды, как правило, имеют слабую связь со временем и не обнаруживают четкой тенденции изменения. В этом случае методы аналитического выравнивания и экспоненциального сглаживания малоэффективны, так как достоверность расчетов резко падает.

Контрольные вопросы:

- В чем заключается метод укрупнения интервалов?
- Охарактеризуйте способ скользящей средней.
- Как моделируется тенденция с помощью метода аналитического выравнивания?
- Привидите примеры видов трендов?
- Что выступает в качестве независимой переменной при построении уравнения тренда?
- Что такое автокорреляция в остатках?
- Для чего необходим критерий Дарбина-Уотсона?
- В каких пределах находится значение критерий Дарбина-Уотсона?
- Что такое интервал сглаживания?
- В чем разница при сглаживании методом скользящей средней по четному и нечетному количеству уровней?

Решение типовых задач

Пример 1. Имеется следующий временной ряд:

t	1	2	3	4	5	6	7	8	9	10
x_t	15	…	…	…	…	…	…	…	…	25

Известно также, что $\Sigma x_t=250$, $\Sigma x_t^2 =8000$, $\sum_{t=2}^{n} x_t \cdot x_{t-1} = 47900$.

Задание.

1. Определить коэффициент автокорреляции уровней этого ряда первого порядка.

2. Установить содержит ли временной ряд тенденцию, возможно ли построение уравнения регрессии зависимости данного ряда от другого временного ряда.

3. Построено уравнение регрессии $y_t = 2+3{,}4x_t +2{,}1t$, характеризующее зависимость временного ряда y_t от временного ряда x_t и фактора времени t.

Известны суммы остаточных величин $\Sigma\varepsilon_t^2$=6900, $\Sigma(\varepsilon_t-\varepsilon_{t-1})^2$=25000. Рассчитать критерий Дарбина-Уотсона и указать, пригодно ли уравнение для прогноза.

Решение:

Для определения коэффициента автокорреляции уровней первого порядка следует рассмотреть зависимость между данным рядом и этим же, но сдвинутым на один период (лаг):

x_1 , x_2 …x_{n-1}

x_2 , x_3 …x_n

Соответственно формула расчета будет:

$$r_1 = \frac{\sum_{t=2}^{10}(x_t - \bar{x}_1)\cdot(x_{t-1} - \bar{x}_2)}{\sqrt{\sum_{t=2}^{10}(x_t - \bar{x}_1)^2 \cdot \sum_{t=2}^{10}(x_{t-1} - \bar{x}_2)^2}}, \text{ где}$$

$$\bar{x}_1 = \frac{\sum_{t=2}^{10} x_t}{9}, \quad \bar{x}_2 = \frac{\sum_{t=2}^{10} x_{t-1}}{9}$$

Вспомнив общие формулы для коэффициента корреляции преобразуем выражение согласно нашим исходным данным:

$$r_1 = \frac{\overline{x_t \cdot x_{t-1}} - \bar{x}_t \cdot \bar{x}_{t-1}}{\sqrt{(\overline{x^2}_t - \bar{x}_t^{\,2})\cdot(\overline{x^2}_{t-1} - \bar{x}_{t-1}^{\,2})}} = \frac{\frac{\sum x_t \cdot x_{t-1}}{9} - \frac{\sum_{t=2}^{10} x_t}{9}\cdot\frac{\sum_{t=2}^{10} x_{t-1}}{9}}{\sqrt{\left(\frac{\sum_{t=2}^{10} x_t^2}{9} - \frac{(\sum_{t=2}^{10} x_t)^2}{9}\right)\cdot\left(\frac{\sum_{t=2}^{10} x_{t-1}^2}{9} - \frac{(\sum_{t=2}^{10} x_{t-1})^2}{9}\right)}} =$$

$$= \frac{\frac{\sum x_t \cdot x_{t-1}}{9} - \frac{\sum x_t - 15}{9}\cdot\frac{\sum x_t - 25}{9}}{\sqrt{\left(\frac{\sum x_t^2 - 225}{9} - \frac{(\sum x_t - 15)^2}{9}\right)\cdot\left(\frac{\sum x_t^2 - 625}{9} - \frac{(\sum x_t - 25)^2}{9}\right)}} = -$$

$$= \frac{\frac{47900}{9} - \frac{235\cdot 225}{9\cdot 9}}{\sqrt{\left(\frac{7775}{9} - \frac{235^2}{9}\right)\cdot\left(\frac{7375}{9} - \frac{225^2}{9}\right)}} = 0{,}928$$

Так как коэффициент автокорреляции достаточно высокий, то ряд содержит линейную тенденцию. При построении уравнения регрессии зависимости данного ряда от любого другого временного ряда оценки параметров могут быть смещенными и ненадежными. Чтобы это не произошло, вначале следует устранить или элиминировать присутствующую тенденцию во временном ряду.

Рассчитаем критерий Дарбина-Уотсона по формуле:

$$d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})}{\sum \varepsilon_t^{\ 2}} = \frac{25000}{6900} = 3{,}63$$

Табличные значения критерия Дарбина-Уотсона для n=10 (число наблюдений), к=2 (число факторов), при уровне значимости α=0,05 будут d_L=0,70 d_u=1,64. Строим искомый интервал:

Есть положительная автокорреляция остатков, H_0 отклоняется, с вероятностью	Зона неопределенности	Нет оснований отклонять H_0, автокорреляция отсутствует	Зона неопределенности	Есть от тельная корреляц остатков, отклоняет вероятнос
0 1-α принимается H_1	0,70 1,64	2,36	3,30	1-α 4 прин ся H_1^*

Фактическое значение 3,63 попадает в 5 зону, что говорит о наличии отрицательной автокорреляции в остатках. Уравнение регрессии не может быть использовано для прогноза, так в нем не устранена автокорреляция в остатках, которая может иметь разные причины: не включение какого-либо существенного фактора, наличие общей тенденции, неточна форма связи т д.

Пример 2. Имеются данные об изменении спроса на хлебобулочные изделия по месяцам:

t	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
спрос на хлебобулочные изделия, млн.руб, y_t	27	30	31	34	35	39	38	34	36	38	43	46	36	41	48	48	53	52

Задание. Используя различные методы моделирования тенденции, осуществить прогнозирование спроса во втором полугодие второго года.

Решение:

Анализ графика временного ряда говорит о наличии неустойчивой тенденции во времени, характеризующей спрос на хлебопродукты.

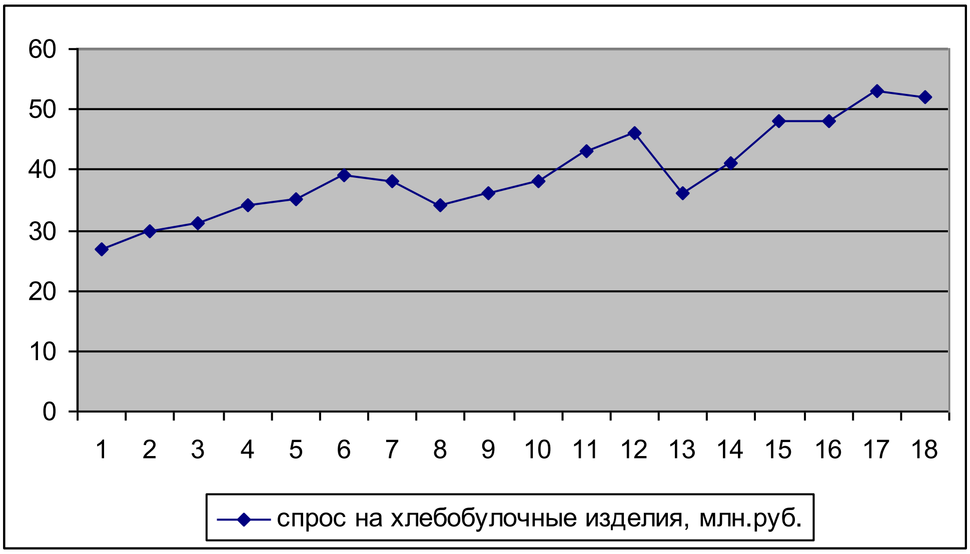


Рис. 2.2. График временного ряда, характеризующего спрос на хлебобулочные изделия

Промоделируем тенденцию методом укрупнения интервалов, заменив месячные данные квартальными:

квартал	1	2	3	4	5	6
спрос, млн.руб.	88	108	108	127	125	153

Теперь имеющаяся линейная тенденция к возрастанию спроса с течением времени имеет ярко выраженный характер.

Аналогичную динамику можно получить, применив метод скользящей средней. В качестве интервала сглаживания возьмем 5 уровня ряда и рассчитаем y_t^c – уровни нового временного ряда.

$y_1^c=(y_1+y_2+y_3+y_4+y_5)/5=(27+30+31+34+35)/5=31{,}4$,

$y_2^c=(y_2+y_3+y_4+y_5+y_6)/5= =(30+31+34+35+39)/5=33{,}8$ и т.д.

t	y_t	y_t^c
1	27	

t	y_t	y_t^c
2	30	
3	31	31,4
4	34	33,8
5	35	35,4
6	39	36,0
7	38	36,4
8	34	37,0
9	36	37,8
10	38	39,4
11	43	39,8
12	46	40,8
13	36	42,8
14	41	43,8
15	48	45,2
16	48	48,4
17	53	
18	52	

Для прогнозирования тенденции на перспективу проведем аналитическое выравнивание исходного временного ряда. График временного ряда говорит о том, что в качестве математической функции удобно выбрать линейный тренд y=a+bt, параметры которого определяются методом МНК из системы:

$$\begin{cases} n \cdot a + b\sum t = \sum y, \\ a\sum t + b\sum t^2 = \sum y \cdot t. \end{cases}$$

подставим

$$\begin{cases} 18 \cdot a + b \cdot 171 = 709 \\ a \cdot 171 + b \cdot 2109 = 7359 \end{cases}$$

Откуда a=27,16, b=1,29, получим уравнение тренда y=27,16+1,29t (t=1,2..18)

Найденные r^2=0,83 показывает хорошее качество построенной модели, объясняя 83% вариации уровней ряда имеющейся тенденцией.

Спрогнозируем будущий уровень спроса на второе полугодие, рассчитав значения для месяцев t=19,20,21,22,23,24. По уравнению регрессии получим соответствующие будущие уровни ряда: 51,6, 52,9, 54,2, 55,5, 56,8, 58,0. Окончательно во втором полугодие второго года планируется спрос на хлебобулочные изделия в размере 329 млн.руб.

Аналогично можно построить прогноз и используя метод скользящей средней. Для этого проведем аналитическое выравнивание нового ряда y_t^c с использованием линейной функции. Находя параметры уравнения обычным МНК, получим y^c =28,67+1,10t (t=3,4…16).

Определим будущий уровень y_{19} из соотношения:

$(y_{15}+y_{16}+y_{17}+y_{18}+y_{19})/5=y_{17}^c$

Так как y_{17}^c=28,67+1,10*17=47,41, то y_{19}=47,41*5-48-48-53-52=36,05. Таким же способом находятся y_{20}=53,51, y_{21}=53,51, y_{22}=58,51 y_{23}=57,51 y_{24}=41,57. Таким образом, с применением метода скользящей средней прогноз уровня спроса на второе полугодие составляет 301 млн.руб., что на 9,3% меньше первого варианта прогноза.

2.3. Исключение тенденции и периодические колебания в рядах динамики

Как было сказано выше, построение одно- и многофакторного уравнения регрессии на основе данных временных рядов иногда сталкивается с проблемой присутствия тенденции в каждом из признаков. Это свойство не дает возможность использования метода МНК, так как приводит к несостоятельности оценок параметров уравнения.

Для устранения данного недостатка применяются разнообразные методы исключения тенденции, сущность которых заключается удалении или фиксации воздействия фактора времени на формирование уровней ряда. Их можно разделить на две группы:

• методы, основанные **на преобразовании** уровней исходного ряда в новые переменные, не содержащие тенденции (метод последовательных разностей, метод отклонений от трендов);

• методы, основанные на изучении взаимосвязи исходных уровней временных рядов **при элиминировании воздействия** фактора времени на зависимую и независимые переменные модели (метод включения в модель фактора времени).

Метод отклонений от тренда. Пусть имеются два временных ряда x и y каждый из которых содержит трендовую компоненту T и случайную компоненту ε. Проведение аналитического выравнивания по каждому из этих рядов позволяет найти параметры соответствующих уравнений трендов и определить расчетные по тренду уровни x_t и y_t соответственно. Эти расчетные значения можно принять за оценку трендовой компоненты T каждого ряда. Поэтому влияние тенденции можно устранить путем вычитания расчетных значений уровней ряда из фактических:

тренды: $y=a_1+b_1 t+\varepsilon$, $x=a_2+b_2 t+\varepsilon$ (2.4)

отклонения: $\Delta y=y - y_t$, $\Delta x=x - x_t$

Дальнейший анализ взаимосвязи рядов проводят с использованием не исходных уровней, а отклонений от тренда Δy и Δx при условии, что последние не содержат тенденции. Наличие линейной тенденции в рядах динамики можно определить по коэффициентам автокорреляции уровней первого порядка

Содержательная интерпретация параметров этой модели затруднительна, однако ее можно использовать для прогнозирования. Для этого необходимо определить трендовое значение факторного признака x_t и с помощью одного из методов оценить величину предполагаемого отклонения фактического значения от трендового. Далее по уравнению тренда для результативного признака определяют трендовое значение y_t,, а по уравнению регрессии по отклонениям от трендов находят величину отклонения y_t-y. Затем находят точечный прогноз фактического значения y по формуле:

$y=y_t+\Delta y$ (2.5)

Метод последовательных разностей. В ряде случаев вместо аналитического выравнивания временного ряда с целью устранения тенденции можно применить более простой метод — метод последовательных разностей.

Если временной ряд содержит ярко выраженную линейную тенденцию, ее можно устранить путем замены исходных уровней ряда цепными абсолютными приростами (первыми разностями).

Пусть ряд содержит тенденцию $y=a+bt+\varepsilon$, тогда цепные абсолютные приросты будут $\Delta_t = y_t - y_{t-1} = a+bt + \varepsilon_t - (a+b(t-1)+\varepsilon_{t-1}) = b + \varepsilon_t - \varepsilon_{t-1}$ (2.6)

Коэффициент b - константа, которая не зависит от времени. При наличии сильной линейной тенденции остатки ε_t достаточно малы и в соответствии с предпосылками МНК носят случайный характер. Поэтому первые разности уровней ряда Δ_t, не зависят от переменной времени, их можно использовать для дальнейшего анализа.

Первые разности могут устранить только **линейную** тенденцию. Если временной ряд имеет параболическую тенденцию, то для ее устранения можно заменить исходные уровни ряда на **вторые разности**: $\Delta_t^2 = \Delta_t - \Delta_{t-1}$.

Тенденция временного ряда в форме экспоненциального или **степенного** тренда хорошо удаляется при применении метода последовательных разностей не к исходным уровням ряда, а к их **логарифмам**.

В отличие от уравнения регрессии по отклонениям от тренда, параметрам данного уравнения легко дать интерпретацию. Например, если зависимость производительности аграрного труда от уровня оплаты работников характеризуется уравнением $\Delta y=3{,}5+1{,}4\Delta x$, то это означает, что при изменении прироста уровня оплаты труда на 1 руб. прирост производительности увеличится в среднем на 1,4 руб.

При всей своей простоте метод последовательных разностей имеет два существенных недостатка. Во-первых, его применение связано с сокращением числа пар наблюдений, по которым строится уравнение регрессии, и, следовательно, с потерей числа степеней свободы. Во-вторых, использование вместо исходных уровней временных рядов их приростов или ускорений приводит к потере информации, содержащейся в исходных данных – отсутствие прогноза.

Включение в модель регрессии фактора времени. В данном методе тенденция фиксируется через включение фактора времени в модель в качестве независимой переменной, т.е. модель вида:

$y=a+b_1x+b_2t$ (2.7)

Преимущество данной модели по сравнению с методами отклонений от трендов и последовательных разностей в том, что она позволяет учесть всю информацию, содержащуюся в исходных данных, поскольку значения y и x есть уровни исходных временных рядов. Кроме того, модель строится по всей совокупности данных за рассматриваемый период в отличие от метода последовательных разностей, который приводит к потере числа наблюдений. Параметры a и b модели с включением фактора времени определяются обычным МНК.

Моделирование сезонных и циклических колебаний. Существует несколько подходов к анализу структуры временных рядов, содержащих сезонные или циклические колебания. Одним из подходов является расчет значений сезонной компоненты методом скользящей средней и построение **аддитивной** или **мультипликативной** модели временного ряда:

У=T+S+E или Y=T*S*E   (2.8)

Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строят аддитивную модель временного ряда, в которой значения сезонной компоненты предполагаются постоянными для различных циклов. Если амплитуда сезонных колебаний возрастает или уменьшается, строят мультипликативную модель временного ряда, которая ставит уровни ряда в зависимость от значений сезонной компоненты.

Построение аддитивной и мультипликативной моделей сводится к расчету значений **трендовой** (Т), **сезонной** (S) и **случайной** (Е) компонент для каждого уровня ряда. При построении аддитивной или мультипликативной моделей целесообразно использовать следующий алгоритм:

1. Выравнивание исходного ряда методом скользящей средней и получение уровней $У^с$.

Например, изменение расхода горючего (тыс.литров) на предприятии по месяцам за 2007-2008 гг. характеризуется следующими данными:

1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
56,6	53,2	52,8	55,1	51,7	51,9	53,9	50,1	51,2	53,2	49,6	50,7
13	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**
52,4	48,8	49,8	51,8	47,9	48,2	50,3	47,0	46,9	48,7	46,1	45,8

Рассчитаем компоненты аддитивной модели, проведя выравнивание исходного ряда методом скользящей средней. Известно, что колебания расходов горюче-смазочных материалов поквартальные (осень, зима, весна, лето), поэтому в качестве интервала сглаживания возьмем 3. Так, $у_1^с$=(56,6+53,2+52,8)/3=54,2, $у_2^с$=(53,2+52,8+55,1)/4=53,7 и т.д. (табл.2.1).

Полученные значения закрепляем за серединой периода.

2. Расчет значений сезонной компоненты S: на первом этапе выделение сезонной составляющей S^* (для аддитивной $S^*=Y-Y^c$, для мультипликативной $S^*=Y/Y^c$), расчет среднего значения сезонной компоненты по одноименным периодам S^{**} и, наконец, окончательный расчет S путем ввода поправочного коэффициента, позволяющего выполнить условие сезонных компонент (для аддитивной - $\Sigma S=0$, для мультипликативной - ΣS=числу периодов в цикле).

Рассчитаем сезонную компоненту, найдя разность для аддитивной модели $S^*=Y-Y^c$ (табл.2.1):

Таблица 2.1

t	y	Y^c	$Y-Y^c$	S^{**}
1	56,6			1,8
2	53,2	54,2	-1,0	-1,3
3	52,8	53,7	-0,9	-0,5
4	55,1	53,2	1,9	
5	51,7	52,9	-1,2	
6	51,9	52,5	-0,6	
7	53,9	52,0	1,9	
8	50,1	51,7	-1,6	
9	51,2	51,5	-0,3	
10	53,2	51,3	1,9	
11	49,6	51,2	-1,6	
12	50,7	50,9	-0,2	
13	52,4	50,6	1,8	
14	48,8	50,3	-1,5	
15	49,8	50,1	-0,3	
16	51,8	49,8	2,0	
17	47,9	49,3	-1,4	
18	48,2	48,8	-0,6	
19	50,3	48,5	1,8	
20	47,0	48,1	-1,1	
21	46,9	47,5	-0,6	
22	48,7	47,2	1,5	
23	46,1	46,9	-0,8	
24	45,8			

После этого находим средние оценки сезонной компоненты S^{**} за каждый квартальный месяц по временам года (учитывая, что у нас рабочий период представляет собой времена года, то одноименными месяцами будут №3,6,9,12,15,18,21 и т.д.):

$S_{3,6,9,12,15,18,21,24}^{**}$=(-0,9-0,6-0,3-0,2-0,3-0,6-0,6)= -0,5 – 1 месяц времени года;

$S_{1,4,7,10,13,16,19,22}^{**}$= 1,8 – 2 месяц времени года;

$S_{2,5,8,11,14,17,20,23}^{**}$= - 1,3 – 3 месяц времени года.

Взаимопогашаемость сезонных воздействий в аддитивной модели выражается в том, что сумма значений сезонной компоненты по месяцам должна быть равна нулю. Проверим это условие: $S_1^{**}+ S_2^{**}+ S_3^{**}$=1,8-1,3-0,5=0. Если бы оно не выполнялось, то необходимо было бы ввести поправочный коэффициент. Таким образом, окончательно значения сезонной компоненты:

S_1= -0,5 – 1 месяц времени года;

S_2= 1,8 – 2 месяц времени года;

S_3= -1,3 – 3 месяц времени года.

3. Устранение сезонной компоненты из исходных уровней ряда и получение выровненных данных (Т+Е) в аддитивной или (Т*Е) в мультипликативной модели (табл.2.2).

4. Аналитическое выравнивание уровней (Т+ Е) или (Т*Е) и расчет значений Т с использованием полученного уравнения тренда.

На основе ряда Т+Е в нашем случае рассчитаем параметры линейного тренда: Т=-0,3326t + 54,717 (t=1,2…24). Подставляя в это уравнение регрессии значения t=1,2 и т.д., получим оценку трендовой компоненты временного ряда Т (табл.2.2).

5. Расчет полученных по модели значений (Т+ S) или (Т*S) – табл.2.2.

6. Расчет абсолютных и/или относительных ошибок. Проводить можно как в относительной форме E_o=Y/(T*S), так и в абсолютной форме E_a= Y - (T*S).

В большинстве случаев рассчитывается по аналогии с корреляционно-регрессионным анализом так называемый «коэффициент детерминации»:

$Д=1-\Sigma E_a^2/\Sigma E^2$, (2.9)

где ΣE^2 – общая дисперсия y_t;

ΣE_a^2 – сумма квадратов отклонений фактических уровней ряда от теоретических.

Д=1-2,9/185,2=0,98

Таблица 2.2

t	Y_t	Y-S	T	T+S	E_o	E_a	E_a^2	$y-y_{ср}$	E^2
1	56,6	54,8	54,4	56,2	1,007	0,401	0,161	6,0	36,4
2	53,2	54,5	54,1	52,8	1,008	0,419	0,176	2,6	6,9
3	52,8	53,3	53,7	53,2	0,992	-0,410	0,168	2,2	5,0
4	55,1	53,3	53,4	55,2	0,998	-0,101	0,010	4,5	20,5
5	51,7	53,0	53,1	51,8	0,998	-0,083	0,007	1,1	1,3
6	51,9	52,4	52,7	52,2	0,994	-0,312	0,097	1,3	1,8
7	53,9	52,1	52,4	54,2	0,994	-0,303	0,092	3,3	11,1
8	50,1	51,4	52,1	50,8	0,987	-0,685	0,470	-0,5	0,2
9	51,2	51,7	51,7	51,2	1,000	-0,014	0,000	0,6	0,4
10	53,2	51,4	51,4	53,2	1,000	-0,005	0,000	2,6	6,9
11	49,6	50,9	51,1	49,8	0,996	-0,188	0,035	-1,0	0,9
12	50,7	51,2	50,7	50,2	1,010	0,484	0,234	0,1	0,0
13	52,4	50,6	50,4	52,2	1,004	0,193	0,037	1,8	3,3
14	48,8	50,1	50,1	48,8	1,000	0,010	0,000	-1,8	3,1
15	49,8	50,3	49,7	49,2	1,012	0,582	0,338	-0,8	0,6
16	51,8	50,0	49,4	51,2	1,012	0,590	0,348	1,2	1,5
17	47,9	49,2	49,1	47,8	1,002	0,108	0,012	-2,7	7,1
18	48,2	48,7	48,7	48,2	1,000	-0,021	0,000	-2,4	5,6
19	50,3	48,5	48,4	50,2	1,002	0,088	0,008	-0,3	0,1
20	47,0	48,3	48,1	46,8	1,004	0,206	0,042	-3,6	12,7
21	46,9	47,4	47,7	47,2	0,993	-0,323	0,104	-3,7	13,5
22	48,7	46,9	47,4	49,2	0,990	-0,514	0,264	-1,9	3,5
23	46,1	47,4	47,1	45,8	1,007	0,304	0,092	-4,5	20,0
24	45,8	46,3	46,7	46,2	0,991	-0,425	0,181	-4,8	22,8
Итого (в среднем)	50,57	50,56	X	50,57	1,0	0,0	2,9	X	185,2

7. Прогнозирование будущих уровней ряда на основе модели.

Прогнозное значение P_t в мультипликативной модели есть произведение соответствующей трендовой T и сезонной S компонент, а в аддитивной их сумма.

Контрольные вопросы:

- Какие методы исключения тенденции вы знаете?
- Раскройте сущность метода последовательных разностей.
- Охарактеризуйте модель с включением фактора времени.

- Охарактеризуйте метод отклонения от трендов.

- Что представляют собой сезонные и циклические колебания в рядах динамики?

- Какие модели временных рядов вы знаете.

- Опишите алгоритм построение аддитивной модели временного ряда.

- Опишите алгоритм построение мультипликативной модели временного ряда.

- Чему равна сумма сезонных компонент в аддитивной модели?

- Чему равна сумма сезонных компонент в мультипликативной модели?

- Как рассчитывается трендовая компонента в аддитивной модели?

- Как рассчитывается трендовая компонента в мультипликативной модели?

Решение типовых задач

Пример 1. Исследуется зависимость у (объем производства сельскохозяйственной продукции, млн.руб.) от х (уровень квалификации работников, %) по данным за 28 лет. Были получены следующие результаты:

1. Уравнения линейных трендов:

$y_t=2,1+1,29t$ $r^2=0,91$ $d=2,31$

$x_t=1,2+5,8t$ $r^2=0,89$ $d=2,08$

1. Уравнения регрессии по уровням временных рядов:

$y=-3,5+0,3x$ $r^2=0,94$ $d=2,21$

3.Уравнения регрессии по первым разностям уровней временных рядов:

$\Delta y_t=1,1+0,02\Delta x_t$ $r^2=0,85$ $d=2,23$

4.Уравнение регрессии по вторым разностям уровней временных рядов:

$\Delta^2 y_t=0,6+0,01\Delta^2 x_t$ $r^2=0,49$ $d=2,71$

5.Уравнение регрессии по уровням временных рядов с включением фактора времени:

$y_t=3,97+0,27x_t+0,53t$, $R^2=0,98$ $d=0,8$

Задание.

1. Сформулируйте свои выводы относительно величины коэффициента автокорреляции остатков первого порядка в каждом из рядов.

2. Выберите наилучшее уравнение регрессии для прогнозирования и дайте его интерпретацию.

3. По известной информации за последние 3 года, дайте точечный прогноз y_t на ближайший год с учетом выбранного уравнения.

год	26	27	28	29
y_t	34	39	40	?
x_t	117	123	122	125

Решение:

Коэффициент автокорреляции остатков первого порядка связан критерием Дарбина-Уотсона следующим соотношением:

$$d=2*(1-r^1),$$

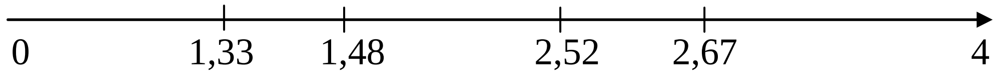
где r^1- коэффициент автокорреляции остатков первого порядка.

Таким образом, что ответить на первый вопрос необходимо сделать проверку на автокорреляцию по критерию Дарбина-Уотсона по каждому уравнению.

1. $n=28$, $k=1$, $\alpha=0,05$ по таблице определили $d_L=1,33$ $d_u=1,48$.

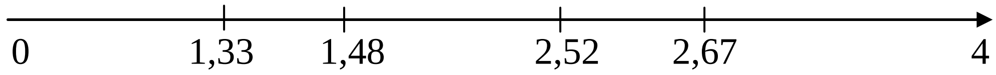
Фактические значения $d=2,31$ и $d=2,08$ попадают в 3 зону, говорящую об отсутствии автокорреляции. Поэтому r^1 в уравнениях трендов приблизительно равен нулю.

2. $n=28$, $k=1$, $\alpha=0,05$ по таблице определили $d_L=1,33$ $d_u=1,48$.



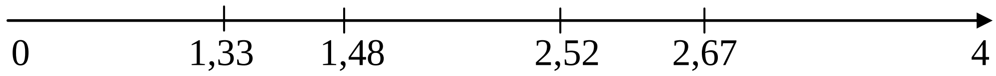
Фактическое значение d=2,21 попадает в 3 зону, говорящую об отсутствии автокорреляции. Поэтому r^1 в уравнении по уровням временных рядов приблизительно равен нулю.

3. n=28, k=1, α=0,05 по таблице определили d_L=1,33 d_u=1,48.



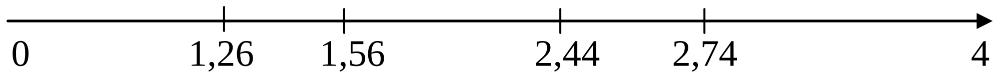
Фактическое значение d=2,23 попадает в 3 зону, говорящую об отсутствии автокорреляции. Поэтому r^1 в уравнении по первым разностям приблизительно равен нулю.

4. n=28, k=1, α=0,05 по таблице определили d_L=1,33 d_u=1,48.



Фактическое значение d=2,71 попадает в 5 зону, говорящую о присутствии отрицательной автокорреляции. Поэтому r^1 в уравнении по вторым разностям приближается к –1.

5. n=28, k=2, α=0,05 по таблице определили d_L=1,26 d_u=1,56.



Фактическое значение d=0,8 попадают в 1 зону, говорящую о присутствии положительной автокорреляции. Поэтому r^1 в уравнении с включением фактора времени приближается к 1.

Осуществляя выбор лучшего уравнения регрессии следует сказать, что 4 и 5 уравнения сразу отпадают в силу наличия в них автокорреляции остатков, делающих оценки параметров смещенными и ненадежными. Из 2 и 3 уравнений по критерию – показатель детерминации лучше второе, так как 0,94>0,85. Вместе с тем анализ первых двух трендов говорит о наличии сильной линейной тенденции в обоих взаимосвязанных рядах динамики x и y. Поэтому для ее устранения следует использовать один из известных методов (метод отклонения от трендов, уравнение по первым разностям, модель с включением фактора времени). В нашем случае наиболее рациональным будет использование для этих целей 3-го уравнения, устраняющего именно линейную тенденцию. То, что наличие тенденции завышает истинный показатель связи между признаками можно увидеть, если сравнить показатели детерминации 2 и 3 уравнений – 0,94>0,85. Таким образом, наиболее надежным и пригодным для прогноза является 3 уравнение регрессии.

Уравнение по первым разностям имеет свою интерпретацию: увеличение прироста уровня квалификации работников на 1% позволяет расти объему производства сельскохозяйственной продукции со средним ускорением, равным 0,02 млн.руб.

Для получения прогноза рассчитаем Δx_t =125-122=3, затем Δy_t=1,1+0,02*3=1,16. Отсюда объем производства сельскохозяйственной продукции в будущем году прогнозируется в размере $y_{29}=y_{28}+\Delta y_t$=40+1,16=41,16 млн.руб.

Пример 2. Известные изменения затрат труда (тыс.чел-час) в сельском хозяйстве региона по кварталам:

квартал	1	2	3	4	5	6	7	8
затраты труда, млн. чел.-час	93,8	100,0	105,6	96,6	96,3	104,3	106,9	96,7
квартал	9	10	11	12	13	14	15	16
затраты труда, млн. чел.-час	98,8	108,8	101,9	97,8	98,5	112,3	102,2	98,7

Задание.

1. Построить мультипликативную модель временного ряда.
2. Оцените качество модели и сделайте прогноз затрат сельскохозяйственного труда на 1 квартал и 1 полугодие ближайшего следующего года.

Решение:

График данного временного ряда (рис.2.3) свидетельствует о наличии сезонных колебаний и общей возрастающей тенденции уровней ряда. Затраты труда в весенне-летний период выше, чем в осенне-зимний период.

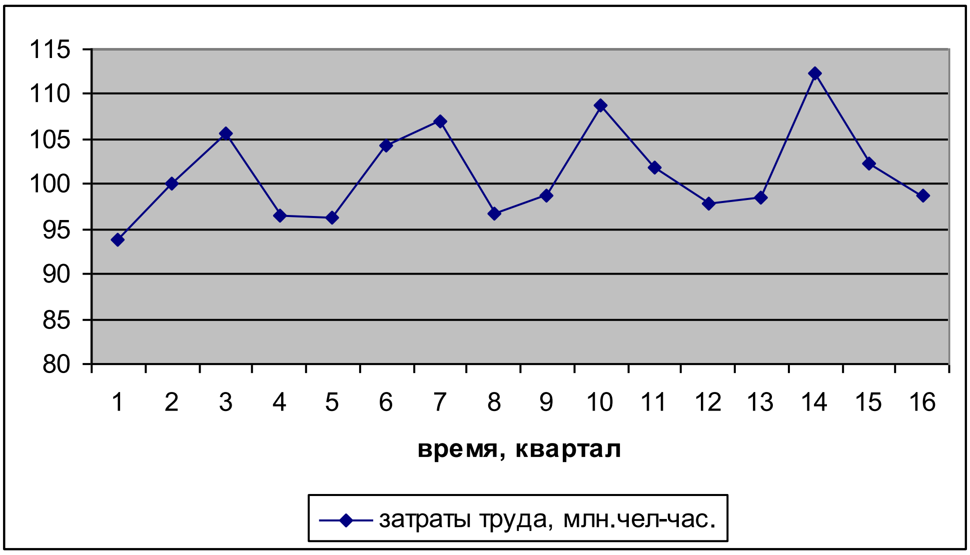


Рис.2.3. Затраты труда в сельском хозяйстве региона

Рассчитаем компоненты мультипликативной модели, проведя вначале выравнивание исходного ряда методом скользящей средней. Интервал сглаживания в связи годичными колебаниями возьмем равным 4. Так, y_1^c=(93,8+100,0+105,6+96,6)/4=99,0, y_2^c=(100,0+105,6+96,6+96,3)/4=100,7 и т.д. (табл.2.3).

Таблица 2.3

№ квартала	y_t	y_t^c	y_t^{cc}	S^*
1	93,8			
2	100			
		99,0		
3	105,6		99,9	1,057
		100,7		
4	96,6		100,7	0,959
		100,7		
5	96,3		100,9	0,955
		101,0		
6	104,3		101,0	1,032
		101,1		
7	106,9		101,4	1,055
		101,7		
8	96,7		102,2	0,946
		102,8		
9	98,8		102,2	0,967
		101,6		
10	108,8		101,7	1,070
		101,8		
11	101,9		101,8	1,001

		101,8		
12	97,8		102,2	0,957
		102,6		
13	98,5		102,7	0,959
		102,7		
14	112,3		102,8	1,092
		102,9		
15	102,2			
16	98,7			

Так как выравнивание проводится по четному числу периодов, то необходимо провести еще центрирование скользящей: $y_1^{cc}=(y_1^c+y_2^c)/2=(99{,}0+100{,}7)/2=99{,}9$ и т.д. Оценка сезонной компоненты проводится путем деления фактических уровней ряда на центрированные значения скользящей средней: $S^*=y_t : y_t^{cc}$. После этого находим средние за каждый квартал оценки сезонной компоненты S^{**}:

$S_1^{**}=(0{,}955+0{,}967+0{,}959)/3=0{,}960$ – 1 квартал;

$S_2^{**}=(1{,}032+1{,}070+1{,}092)/3=1{,}065$ – 2 квартал;

$S_3^{**}=(1{,}057+1{,}055+1{,}001)/3=1{,}038$ – 3 квартал;

$S_4^{**}=(0{,}959+0{,}946+0{,}957)/3=0{,}954$ – 4 квартал.

Взаимопогашаемость сезонных воздействий в мультипликативной модели выражается в том, что сумма значений сезонной компоненты по кварталам должна быть равна числу периодов в цикле (в нашем случае год, значит должно быть равно 4). Проверим это условие: $S_1^{**}+ S_2^{**}+ S_3^{**}+ S_4^{**}=0{,}960+1{,}065+1{,}038+0{,}954=4{,}017$.

Чтобы поправит создавшуюся ситуацию и уменьшить значение до 4, введем поправочный коэффициент к=4,017/4=1,00425.

Теперь окончательно определим скорректированные значения сезонной компоненты S:

$S_1=0{,}960/1{,}00425=0{,}956$ – 1 квартал;

$S_2=1{,}65/1{,}00425=1{,}061$ – 2 квартал;

$S_3=1{,}038/1{,}00425=1{,}033$ – 3 квартал;

$S_4=0{,}954/1{,}00425=0{,}950$ – 4 квартал.

Проверка необходимого условия дает положительный результат – сумма в точности равна числу периодов (4). Теперь устраним сезонную компоненту из исходных уровней ряда, определив величину T*E=Y/S.

t	Y_t	Y/S	T	T*S	E_o	E_a	E_a^2	$y-y_{cp}$	E^2
1	93,8	98,1	98,9	94,5	0,992	-0,741	0,549	-7,4	54,8
2	100	94,3	99,2	105,2	0,951	-5,159	26,610	-1,2	1,4
3	105,6	102,2	99,5	102,8	1,027	2,792	7,797	4,4	19,4
4	96,6	101,7	99,8	94,8	1,019	1,784	3,183	-4,6	21,2
5	96,3	100,7	100,1	95,8	1,006	0,546	0,298	-4,9	24,0
6	104,3	98,4	100,4	106,5	0,979	-2,203	4,855	3,1	9,6
7	106,9	103,5	100,8	104,1	1,027	2,782	7,738	5,7	32,5
8	96,7	101,8	101,1	96,0	1,007	0,679	0,461	-4,5	20,3
9	98,8	103,3	101,4	97,0	1,019	1,833	3,360	-2,4	5,8
10	108,8	102,6	101,7	107,8	1,009	0,952	0,906	7,6	57,8
11	101,9	98,6	102,0	105,4	0,967	-3,529	12,452	0,7	0,5
12	97,8	103,0	102,3	97,2	1,006	0,574	0,330	-3,4	11,6
13	98,5	103,0	102,7	98,2	1,003	0,320	0,102	-2,7	7,3
14	112,3	105,9	103,0	109,2	1,028	3,107	9,653	11,1	123,2
15	102,2	98,9	103,3	106,7	0,957	-4,539	20,606	1,0	1,0
16	98,7	103,9	103,6	98,4	1,003	0,269	0,073	-2,5	6,3
Итого (в среднем)	101,2	101,2	X	X	16,0	-0,532	99,0	X	396,4

На основе ряда T*E рассчитаем параметры линейного тренда:

T=98,54+0,32t (t=1,2…16)

Подставляя в это уравнение регрессии значения t=1,2 и т.д. получим оценку трендовой компоненты временного ряда T. График уравнения тренда представлен на рис.2.4.

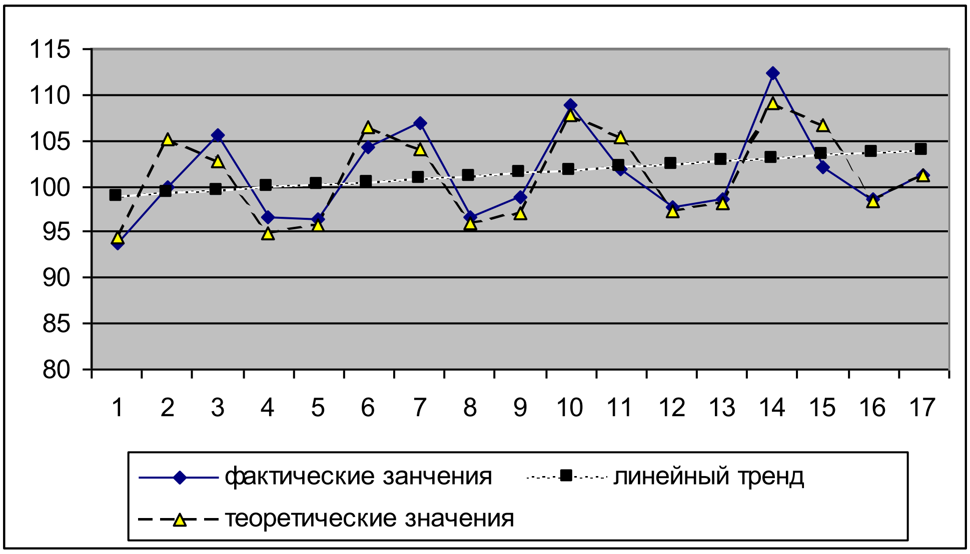


Рис.2.4. Затраты труда в сельском хозяйстве (фактические и выровненные по мультипликативной модели значения уровней ряда)

Найдем теоретические уровни ряда по мультипликативной модели, умножив T на значение сезонной компоненты S для соответствующих кварталов.

Расчет ошибок в мультипликативной модели можно проводить как в относительной форме $E_o=Y/(T*S)$, так и в абсолютной форме $E_a= Y - (T*S)$.

Если временной ряд относительных ошибок E_o не содержит автокорреляции остатков, то его можно использовать вместо исходного ряда для изучения его взаимосвязи с другими временными рядами.

По аналогии с корреляционно-регрессионным анализом абсолютные ошибки применяются для сравнения с другими моделями и определения качества построенной мультипликативной модели. Рассчитаем подобие коэффициента детерминации для данной модели:

$Д=1-\Sigma E_a^2/\Sigma E^2$,

$Д=1-99,0/396,4=0,750$

Таким образом, доля объясненной дисперсии уровней ряда равна 75,0%, что говорит о неплохом качестве построенной мультипликативной модели исходного временного ряда.

Прогнозное значение P_t в мультипликативной модели есть произведение соответствующей трендовой T и сезонной S компонент. Для первого и второго кварталов следующего года это будет:

$P_{17}=T_{17}*S_{17}$, $P_{18}=T_{18}*S_{18}$,

$S_{17}=S_1=0,956$, $S_{18}=S_2=1,061$.

Для определения трендовой компоненты воспользуемся уравнением тренда $T=98,54+0,32t$. Получим:

$T_{17}=98,54+0,32*17=103,9$, $T_{18}=98,54+0,32*18=104,2$.

Отсюда, $P_{17}=103,9*0,956=99,3$, а $P_{18}=104,2*1,061=110,6$ млн.чел-час.

Таким образом, прогноз затрат сельскохозяйственного труда на первый квартал ближайшего года составил 99,3 млн.чел-часов, а в первом полугодии он будет равен:

$P_I=P_{17}+P_{18}=99,3+110,6=209,9$ млн.чел-часов.

4.3. Система эконометрических уравнений и проблема идентификации

Многие экономические взаимосвязи допускают моделирование одним уравнением. В большинстве случаев использование МНК для оценки параметров таких моделей является наиболее подходящей процедурой. Однако, при использовании отдельных уравнений регрессии, например для экономических расчетов, в большинстве случаев предполагается, что аргументы (факторы) можно изменять независимо друг от друга. Однако это предположение является очень грубым: практически изменение одной переменной, как правило, не может происходить при абсолютной неизменности других. Ее изменение повлечет за собой изменения во всей системе взаимосвязанных признаков. Следовательно, отдельно взятое уравнение множественной регрессии не может достаточно точно характеризовать влияния отдельных признаков на вариацию результирующей переменной. Именно поэтому в последние десятилетия в социально-экономических исследованиях важное место заняла проблема описания структуры связей между переменными системой – **система одномерных уравнений**.

Одной из простейших систем одновременных уравнений является модель спроса-предложения в рыночной экономике. Как предложение на отдельный вид продукции P_t , так и спрос S_t в произвольный момент времени t являются линейными функциями от цены c_t , поэтому получается следующая система уравнений:

$$\begin{array}{ll} \text{функция предложения} & \left\{ \begin{array}{l} P_t = a_0+a_1c_t\,, \quad a_1<0 \\ \\ S_t = b_0+b_1c_t\,, \quad b_1<0 \quad (2.10) \end{array} \right. \\ \text{функция спроса} & \end{array}$$

условие равновесия $\quad P_t = S_t\,.$

Наличие случайных отклонений в данных моделях связано с отсутствием некоторых важных объясняющих переменных (дохода, цен на товарызаменители и т.д.). Например, если в функцию спроса добавить доход потребителей y_t, то получим систему:

$$\begin{cases} \text{функция предложения} \quad P_t = a_0+a_1c_t\,, \quad a_1<0 \\ \text{функция спроса} \quad S_t = b_0+b_1c_t + b_2y_t\,, \quad b_1<0 \end{cases} \qquad (2.11)$$

условие равновесия $P_t = S_t$.

Система уравнений в эконометрических исследованиях может быть построена по-разному. Возможна **система независимых уравнений,** когда каждая зависимая переменная (y_i) рассматривается как функция одного и того же набора факторов (x_i):

$$\begin{cases} y_1=f(x_1,x_2\ldots x_n) \\ y_2=f(x_1,x_2\ldots x_n) \\ \ldots \\ y_n=f(x_1,x_2\ldots x_n) \end{cases} \qquad (2.12)$$

Причем число факторов в каждом уравнении может видоизменяться. Отсутствием какого-либо фактора в уравнении системы может быть следствием как нецелесообразности его включением в модель, так и несущественности его воздействия на результативный признак.

Примером такой модели может служить модель экономической эффективности производства, где в качестве используемых переменных выступают показатели, характеризующие эффективность производства, — производительность труда, прибыль и рентабельность производства, а в качестве факторов - специализация хозяйства, энерговооруженность, уровень механизации и т.д. Каждое уравнение системы независимых уравнений может рассматриваться самостоятельно, а для нахождения его параметров используется обычный метод МНК.

Уравнения, составляющие исходную модель, называют **структурными уравнениями** модели. Обычно их подразделяют на **поведенческие** уравнения и уравнения **тождества**. В первых из них описываются взаимодействия между переменными. Во вторых – соотношения, которые должны выполняться во всех случаях.

Наибольшее распространение в эконометрических исследованиях получила система взаимозависимых уравнений**.** В ней одни и те же зависимые переменные в одних уравнениях входят в левую часть, а в других уравнениях — в правую часть системы. Примером может служить модель производительности труда и фондоотдачи:

$$\begin{cases} y_1=a_{10}+a_{11}x_1+a_{12}x_2 \\ y_2=a_{20}+a_{21}y_1+a_{22}x_1+a_{23}x_2 \end{cases} \qquad (2.13)$$

где y_1 – производительность труда;

y_2 – фондоотдача;

x_1 – фондовооруженность труда;

x_2 – уровень комплексной автоматизации и механизации производства.

В таких системах одни и те же переменные (y) одновременно рассматриваются как зависимые в одних уравнениях и как независимые в других, представляя так называемую структурную форму модели. Эта форма не позволяет оценивать параметры уравнения методом МНК, так как каждое уравнение не является самостоятельным.

В структурной форме модели можно выделить две группы переменных: эндогенные и экзогенные.

Эндогенные переменные, обозначающиеся как y, представляют собой зависимые переменные, число которых равно числу уравнений в системе.

Экзогенные переменные, обозначающиеся как x, определяют влияние на эндогенные, но не зависят от них. К таким относятся и предопределенные переменные (лаговые эндогенные переменные, оказывающие определенное влияние на зависимые эндогенные переменные). Так, запасы сельскохозяйственной продукции текущего года (y_t) могут зависеть не только от ряда экономических факторов, но и от уровня запасов в предыдущем году (y_t).

Все переменные в модели могут быть выражены в отклонениях от среднего уровня ($x - \bar{x}$, $y - \bar{y}$), в связи с чем свободные члены в **структурной форме** будут отсутствовать:

$$\begin{cases} y_1=a_{11}y_2+a_{12}x_1+a_{13}x_2 \\ y_2=a_{21}y_1+a_{22}x_1+a_{23}x_2 \end{cases} \qquad (2.14)$$

Непосредственное применение МНК к структурной форме дает несостоятельные и смещенные оценки, поэтому необходимо преобразование данной формы в **приведенную форму** систему одновременных уравнений. Эта система уравнений, где эндогенные переменные выражены только через экзогенные или предопределенные переменные:

$$\begin{cases} y_1=b_{11}x_1+b_{12}x_2+\ldots+b_{1m}x_m \\ y_2=b_{21}x_1+b_{22}x_2+\ldots+b_{2m}x_m \\ y_n=b_{n1}x_1+b_{n2}x_2+\ldots+b_{nm}x_n \end{cases} \qquad (2.15)$$

Данная форма аналогична системы независимых уравнений, а, значит, ее параметры могут быть определены обычным МНК. После этого определяются значение эндогенных переменных через экзогенные. Рассматривая процесс преобразования из одной формы модели в другую можно показать, что коэффициенты приведенной формы модели представляют собой нелинейные функции коэффициентов структурной формы модели. Кроме этого, при переходе возникает еще проблема идентификации.

Исходную систему уравнений называют **идентифицируемой**, если по коэффициентам приведенных уравнений можно однозначно определить значение коэффициентов структурных уравнений, т.е. количество коэффициентов приведенной и структурной форм совпадает.

Модель **неидентифицируема**, если по коэффициентам приведенных уравнений можно получить несколько вариантов значений коэффициентов структурных уравнений, т.е. число приведенных коэффициентов меньше числа структурных коэффициентов.

Модель **сверхидентифицируема**, если по коэффициентам приведенных уравнений невозможно определить значения коэффициентов структурных уравнений, т.е. число приведенных коэффициентов больше числа структурных коэффициентов.

Например, имеется модель со структурной формой:

$$\begin{cases} y_1=a_{11}y_2+a_{12}x_1+a_{13}x_2 \\ y_2=a_{21}y_1+a_{22}x_1+a_{23}x_2. \end{cases}$$

Соответствующее ей приведенная форма будет:

$$\begin{cases} y_1=b_{11}x_1+b_{12}x_2 \\ y_2=b_{21}x_1+b_{22}x_2 \end{cases}$$

Получим, что четырем коэффициентам приведенной формы соответствует шесть коэффициентов структурной, а, значит, модель является неидентифицируемой. В предположении, что a_{12} и a_{23} равны 0 можно получить идентифицируемую модель. Если же еще структурный коэффициент a_{11}=0, то модель превращается в сверхидентифицируемую. Возможно несколько путей уменьшения числа оцениваемых структурных коэффициентов: присутствие слабой взаимосвязи признаков с эндогенной переменной из левой части системы; предполагая, что оказываемое влияние на эндогенную переменную одинаково по абсолютной величине для нескольких признаков.

Структурная модель в общем виде, состоящая из n уравнений с набором n эндогенных и m экзогенных переменных в каждом, максимально содержит n(n-1+m) параметров.

Структурная модель всегда представляет собой систему совместных уравнений, каждое из которых требуется проверять на идентификацию.

Модель считается **идентифицируемой**, если каждое уравнение системы идентифицируемо. Если хотя бы одно из уравнений системы не идентифицируемо, то и вся модель считается **неидентифицируемой**. **Сверхидентифицируемая** модель содержит хотя бы одно сверхидентифицируемое уравнение.

Для быстрого формального определения идентифицируемости структурных уравнений применяются ряд необходимых и достаточных условий.

Пусть система состоит из N уравнений относительно N эндогенных переменных с К экзогенными переменными. В проверяемом на идентифицируемость уравнении количество эндогенных и экзогенных переменных соответственно равно n и к.

Первое необходимое условие.

Чтобы уравнение было идентифицируемо, необходимо, чтобы число экзогенных переменных, отсутствующих в данном уравнении, но присутствующих в системе, было равно числу эндогенных переменных в данном уравнении без одного, т.е. (К-к)+1=n.

Второе необходимое условие.

Уравнение идентифицируемо, если оно исключает N-1 переменную (экзогенную или эндогенную), присутствующую в модели, т.е. (N-n)+(К-к)=N-1.

При замене знаков равенства в обоих необходимых условиях на знак больше уравнение становится сверхидентифицируемым, на знак меньше – неидентифицируемым.

Необходимое и достаточное условие.

В модели, содержащей N уравнений относительно N эндогенных переменных, условие идентифицируемости выполняется тогда и только тогда, когда ранг матрицы коэффициентов, составленной из исключенных из данных уравнений переменных (экзогенную или эндогенную), но входящих в другие уравнения системы, равен N-1. Это в том числе означает, что определитель матрицы не должен быть равен 0.

Контрольные вопросы:

- Что представляет собой система одновременных уравнений?
- Какие переменные называются эндогенными?
- Какие переменные называются экзогенными?
- Что такое структурные уравнения?
- Приведите пример системы одновременных уравнений.
- Какая модель называется индентифицируемой и неидентифицируемой?

- Назовите необходимые и достаточные условия идентифицируемости системы одновременных уравнений.
- Что такое сверхидентифицируемая модель?
- Какая форма называется приведенной формой системы уравнений?
- Какая форма называется структурной формой системы уравнений?

Решение типовых задач

Пример 1. Имеется система одновременных уравнений:

$$\begin{cases} y_1=a_{11}y_2+a_{12}x_2+a_{13}x_3 \\ y_2=a_{21}y_1+a_{22}x_1+a_{23}x_3 \\ y_3=a_{31}y_1+a_{32}y_2+a_{33}x_3 \end{cases}$$

Задание:

1. Проверить выполнение необходимых и достаточного условий идентификации.

2. Что произойдет, если коэффициент $a_{12}=0$ или третье уравнение измениться: $y_3=a_{31}y_1+a_{32}y_2+a_{33}x_1+a_{34}x_3$?

Решение:

Общее количество эндогенных переменных N=3, экзогенных К=3. Проверка первого необходимого условия:

Для первого уравнения к=2, n=2, откуда (К-к)+1=3-2+1=2. Тогда (К-к)+1=n, что означает выполнение необходимого условия.

Для второго уравнения к=2, n=2, откуда (К-к)+1=3-2+1=2. Тогда (К-к)+1=n, что означает выполнение необходимого условия.

Для третьего уравнения к=1, n=3, откуда(К-к)+1=3-1+1=3. Тогда (К-к)+1=n, что означает выполнение необходимого условия.

Таким образом, первое необходимое условие выполнимо в целом во всей системе.

Проверка второго необходимого условия:

Для первого уравнения к=2, n=2, откуда (N-n)+(К-к)=3-2+3-2=2. Тогда (N-n)+(К-к)=N-1, что означает выполнение необходимого условия.

Для второго уравнения к=2, n=2, откуда (N-n)+(К-к)=3-2+3-2=2. Тогда (N-n)+(К-к)=N-1, что означает выполнение необходимого условия.

Для третьего уравнения к=1, n=3, откуда (N-n)+(К-к)=3-3+3-1=2. Тогда (N-n)+(К-к)=N-1, что означает выполнение необходимого условия.

Следовательно, и второе необходимое условие выполнено для системы в целом.

Рассмотрим выполнение достаточного и необходимого условия.

Целесообразность проверки условия идентификации модели через определитель матрицы коэффициентов, объясняется тем, что возможна ситуация, когда для каждого уравнения системы выполнено необходимое равенство, а определитель матрицы структурных коэффициентов равен нулю. В этом случае соблюдается лишь необходимое, но недостаточное условие идентификации.

Составим матрицу коэффициентов при отсутствующих в первом уравнении переменных:

	y_3	x_1
второе уравнение	0	a_{22}
третье уравнение	-1	0

Согласно таблице определитель матрицы (detA) не равен нулю и ранг самой матрицы равен 2, что соответствует наличию идентифицируемого уравнения.

Составим матрицу коэффициентов при отсутствующих во втором уравнении переменных:

	y_3	x_2
первое уравнение	0	a_{12}
третье уравнение	-1	0

Согласно таблице определитель матрицы (detA) не равен нулю и ранг самой матрицы равен 2, что соответствует наличию идентифицируемого уравнения.

Составим матрицу коэффициентов при отсутствующих в третьем уравнении переменных:

	x_1	x_2
первое уравнение	0	a_{12}
второе уравнение	a_{22}	0

Согласно таблице определитель матрицы (detA) не равен нулю и ранг самой матрицы равен 2, что соответствует наличию идентифицируемого уравнения.

Таким образом, на основании достаточного и необходимого условия можно сделать вывод, что данная система одновременных уравнений в целом идентифицируема.

2. Если предположить, что коэффициент a_{12}=0, то первое уравнение станет сверхидентифицируемо и вся система в целом будет сверхидентифицируемой.

Если предположить, что третье уравнение примет вид $y_3=a_{31}y_1+a_{32}y_2+a_{33}x_1+a_{34}x_3$, то (К-к)+1=3-2+1=2. Получим (К-к)+1<n, т.е. уравнение станет неидентифицируемым. И даже при выполнении обоих предположений (первое уравнение – сверхидентифицируемо, второе – идентифицируемо, третье – неидентифицируемо), вся система становится неидентифицируемой.

4.4. Оценивание параметров структурной модели.

Коэффициенты структурной модели могут быть оценены разными способами в зависимости от вида системы одновременных уравнений. Наибольшее распространение в литературе получили следующие методы оценивания коэффициентов структурной модели:

- **косвенный метод наименьших квадратов;**
- **двухшаговый метод наименьших квадратов;**
- **трехшаговый метод наименьших квадратов;**
- **метод максимального правдоподобия с полной информацией;**
- **метод максимального правдоподобия при ограниченной информации.**

Косвенный и двухшаговый методы наименьших квадратов являются традиционными методами оценки коэффициентов структурной модели. Эти методы достаточно легкореализуемы.

Косвенный метод наименьших квадратов (КМНК) применяется для идентифицируемой системы одновременных уравнений, а двухшаговый метод наименьших квадратов (ДМНК) используется для оценки коэффициентов сверхидентифицируемой модели.

Метод максимального правдоподобия рассматривается как наиболее общий метод оценивания, результаты которого при нормальном распределении признаков совпадают с МНК. Однако при большом числе уравнений системы этот метод приводит к достаточно сложным вычислительным процедурам. Поэтому в качестве модификации используется метод максимального правдоподобия при ограниченной информации, разработанный в 1949 г. Т.Андерсоном и Н.Рубиным.

В отличие от метода максимального правдоподобия в данном методе сняты ограничения на параметры, связанные с функционированием системы в целом. Это делает решение более простым, но трудоемкость вычислений остается достаточно высокой. Несмотря на его значительную популярность, к середине 60-х годов он был практически вытеснен двухшаговым методом наименьших квадратов (ДМНК) в связи с гораздо большей простотой последнего. Дальнейшим развитием ДМНК является трехшаговый МНК (ТМНК), предложенный в 1962 г. А. Зельнером и Г. Тейлом. Этот метод оценивания пригоден для всех видов уравнений структурной модели. Однако при некоторых ограничениях на параметры более эффективным оказывается ДМНК.

Одним из традиционных методов является **косвенный метод наименьших квадратов** (КМНК), используемый в случае точно идентифицируемой структурной модели. Он включает в себя несколько этапов:

1. Преобразование структурных уравнений в уравнения приведенной формы.
2. Оцениваются методом МНК параметры уравнений в приведенной форме (b_{ij}).
3. Коэффициенты приведенной формы трансформируются в параметры структурной модели.

Компьютерная программа применения КМНК предполагает, что система уравнений содержит в правой части в каждом уравнении как эндогенные, так и экзогенные переменные. Между тем могут быть системы, в которых в одном из уравнений, например, отсутствуют экзогенные переменные. Для такой модели непосредственное получение структурных коэффициентов невозможно. В этом случае сначала определяется система приведенной формы модели, решаемая обычным МНК, а затем путем алгебраических преобразований переходят к коэффициентам структурной модели.

Если система сверхидентифицируема, то КМНК не используется, ибо он не дает однозначных оценок для параметров структурной модели. В этом случае могут использоваться разные методы оценивания, среди которых наиболее распространенным и простым является **двухшаговый метод наименьших квадратов** (ДМНК).

Основная идея ДМНК — на основе приведенной формы модели получить для сверхидентифицируемого уравнения теоретические значения эндогенных переменных, содержащихся в правой части уравнения. Далее, подставив их вместо фактических значений, можно применить обычный МНК к структурной форме сверхидентифицируемого уравнения. Метод получил название двухшагового МНК, ибо дважды используется МНК: на первом шаге при определении приведенной формы модели и нахождении на ее основе оценок теоретических значений эндогенной переменной $y_n=b_{n1}x_1+ b_{n2}x_2+ \dots +b_{nm}x_m$ и на втором шаге применительно к структурному сверхидентифицируемому уравнению при определении структурных коэффициентов модели по данным теоретических (расчетных) значений эндогенных переменных.

Сверхидентифицируемая структурная модель может быть двух типов:

- все уравнения системы сверхидентифицируемы;
- система содержит наряду со сверхидентифицируемыми точно идентифицируемые уравнения.

Если все уравнения системы сверхидентифицируемые, то для оценки структурных коэффициентов каждого уравнения используется ДМНК. Если в системе есть точно идентифицируемые уравнения, то структурные коэффициенты по ним находятся из системы приведенных уравнений.

Контрольные вопросы:

- Какими способами могут быть оценены коэффициенты структурной модели?
- Опишите косвенный метод наименьших квадратов.
- Опишите двухшаговый метод наименьших квадратов.
- Для какой системы используется косвенный метод наименьших квадратов?
- Для какой системы используется двухшаговый метод наименьших квадратов?
- Кем разработан метод максимального правдоподобия?
- Кем был предложен трехшаговый метод наименьших квадратов?

Решение типовых задач

Пример 1. Имеется следующие данные.

годы	Потребление говядины на душу населения, кг, y_1	Средняя цена говядины, у.е., y_2	Материальные затраты по производству, % к цене, x_1	Среднедневная заработная плата, у.е., x_2
2004	62	5,1	60	1100
2005	64	4,1	57	1100
2006	66	4,4	57	1550
2007	61	5,0	63	1570
2008	66	3,9	50	1750

Задание.

1. Построить модель :

$$\begin{cases} y_1=f(y_2,x_2) \\ y_2=f(y_1,x_1) \end{cases}$$

в общем виде и проверить условие идентифицируемости системы.

2. Рассчитать структурные коэффициенты построенной модели.

Решение:

Общий вид системы будет:

$$\begin{cases} y_1=a_{11}y_2+a_{12}x_2 \\ y_2=a_{21}y_1+a_{22}x_1 \end{cases}$$

Воспользуемся первым необходимым условием для проверки идентифицируемости системы. Для первого уравнения к=1, n=2, откуда (К-к)+1=2-1+1=2. Тогда (К-к)+1=n, что означает выполнение необходимого условия.

Для второго уравнения к=1, n=2, откуда (К-к)+1=2-1+1=2. Тогда (К-к)+1=n, что означает выполнение необходимого условия. Таким образом, система в целом идентифицируема и ее параметры можно рассчитать косвенным методом наименьших квадратов (КМНК). С этой целью запишем приведенную форму данной модели:

$$\begin{cases} y_1=b_{11}x_1+b_{12}x_2 \\ y_2=b_{21}x_1+b_{22}x_2 \end{cases}$$

Оценка коэффициентов приведенной модели проводится обычным МНК из системы нормальных уравнений:

для первого уравнения $\begin{cases} \Sigma y_1x_1=b_{11}\Sigma x_1^2+b_{12}\Sigma x_1x_2 \\ \Sigma y_1x_2=b_{11}\Sigma x_1x_2+b_{12}\Sigma x_2^2 \end{cases}$

для второго уравнения $\begin{cases} \Sigma y_2x_1=b_{21}\Sigma x_1^2+b_{22}\Sigma x_1x_2 \\ \Sigma y_2x_2=b_{21}\Sigma x_1x_2+b_{22}\Sigma x_2^2 \end{cases}$

При решений этих систем, предполагается, что свободные члены отсутствуют, и переменные x и y выражены через отклонения от средних уровней. Корреляционная таблица исходных данных представлена в табл.2.6.

Таблица 2.6

№	y_1	y_2	x_1	x_2
1	-1,8	0,6	2,6	-314
2	0,2	-0,4	-0,4	-314
3	2,2	-0,1	-0,4	136
4	-2,8	0,5	5,6	156
5	2,2	-0,6	-7,4	336

Решив системы методом подстановок, получим следующую приведенную форму:

$$\begin{cases} y_1=-0{,}3903x_1+0{,}0005x_2\,, \\ y_2=0{,}0990x_1+0{,}0001x_2\,. \end{cases}$$

Из приведенной формы модели определим коэффициенты структурной формы:

$$\begin{cases} y_1=-0{,}3903x_1+0{,}0005x_2\,, \\ x_1=(y_2-0{,}0001x_2)/0{,}0990\,, \end{cases}$$

$y_1=-0{,}3903*(y_2-0{,}0001x_2)/0{,}0990+0{,}0005x_2=-3{,}942y_2+0{,}001x_2$,

$$\begin{cases} x_2=(y_1+0{,}3903x_1)/0{,}0005\,, \\ y_2=0{,}0990x_1+0{,}0001x_2\,, \end{cases}$$

$y_2=0{,}0990x_1+0{,}0001*(y_1+0{,}3903x_1)/0{,}0005=0{,}2y_1+0{,}177x_1.$

Таким образом, структурная модель примет вид:

$$\begin{cases} y_1=-3{,}942y_2+0{,}001x_2\,, \\ y_2=0{,}200y_1+0{,}177x_1\,. \end{cases}$$

Эту же систему можно записать, включив в нее свободный член уравнения, т. е. перейти от переменных в виде отклонений от среднего уровня к исходным переменным y и x. Свободные члены уравнений определим по формулам:

$$a_{10}=\bar{y}_1 - a_{11}\,\bar{y}_2 - a_{12}\,\bar{x}_2=63{,}8 - (-3{,}942)*4{,}5 - 0{,}001*1414=80{,}125$$

$$a_{20}=\bar{y}_2 - a_{21}\,\bar{y}_1 - a_{22}\,\bar{x}_1=4{,}5 - 0{,}2*63{,}8 - 0{,}177*57{,}4=-18{,}420$$

Тогда получим:

$$\begin{cases} y_1=80{,}125 - 3{,}942y_2+0{,}001x_2\,, \\ y_2= -18{,}420+0{,}200y_1+0{,}177x_1\,. \end{cases}$$

Если к каждому уравнению структурной формы модели применить традиционный МНК, то результаты будут резко отличаться:

$$\begin{cases} y_1=77{,}32 - 3{,}38y_2+0{,}001x_2\,, \\ y_2= 6{,}20 - 0{,}08y_1+0{,}06x_1\,. \end{cases}$$

Как видим, не совпадают даже знаки коэффициентов при переменных: во втором уравнении структурной формы коэффициенты больше нуля, а в уравнении регрессии один оказался меньше нуля; свободный член изменил знак. Различия между коэффициентами регрессии и структурными коэффициентами модели численно могут быть и менее существенными.

Раздел III . Типовые задания для закрепления.

Задача 1. По семи территориям за 2008 г. известны значения двух признаков, представленных в таблице.

Район	Расходы на покупку продовольственных товаров в общих расходах, %	Среднемесячная заработная плата одного работающего, руб.
1	68,8	4510
2	61,2	5900
3	59,9	5720
4	56,7	6180
5	55,0	5880
6	54,3	4720
7	49,3	5520

Задание:

1.Определите факторный, результативный признаки и постройте уравнение линейной регрессии. Дайте интерпретацию его параметров.

2.Вычислите коэффициент детерминации через общую и факторную дисперсии, что показывает данный коэффициент.

3. Выполните прогноз уровня расходов на покупку продовольственных товаров при прогнозном значении среднемесячной заработной платы, составляющим 115% от среднего уровня.

4. Для характеристики зависимости у от х рассчитайте также параметры следующих функций: степенная, показательная, равносторонняя гипербола.

Задача 2. Зависимость среднемесячной производительности труда от возраста рабочих характеризуется моделью: $y = a + bx + cx^2$. Ее использование привело к результатам, представленным в таблице 3.1.

Таблица 3.1

№ п/п	Производительность труда рабочих, тыс. руб., у		№ п/п	Производительность труда рабочих, тыс. руб., у	
	фактическая	расчетная		фактическая	расчетная
1	12	10	6	11	12
2	8	10	7	12	13
3	13	13	8	9	10
4	15	14	9	11	10
5	16	15	10	9	9

Задание. Оцените качество модели, определив ошибку аппроксимации, индекс корреляции и F-критерий Фишера.

Задача 3. Пусть имеется следующая модель регрессии, характеризующая зависимость y от x: y_x = 8-7x. Известно также, что r_{xy}= -0,5 ; n = 20.

Задание.

1. Постройте доверительный интервал для коэффициента регрессии в этой модели: а) с вероятностью 90%;

б) с вероятностью 99%.

2. Проанализируйте результаты, полученные в п.1, и поясните причины их различий.

Задача 4. По 20 регионам страны изучается зависимость уровня безработицы y (%) от индекса потребительских цен x (% к предыдущему году). Информация о логарифмах исходных показателей представлена в таблице.

Показатель	ln x	ln y
Среднее значение	0,6	1,0
Среднее квадратическое отклонение	0,4	0,2

Известно также, что коэффициент корреляции между логарифмами исходных показателей составил $r_{\ln x \ln y}$ =0,8.

Задание.

1. Постройте уравнение регрессии зависимости уровня безработицы от индекса потребительских цен в степенной форме.

2. Дайте интерпретацию коэффициента эластичности данной модели регрессии.

3. Определите значение коэффициента детерминации и поясните его смысл.

4. Оцените значимость построенного уравнения в целом.

Задача 5. По группе 10 заводов, производящих однородную продукцию, получено уравнение регрессии себестоимости единицы продукции y (тыс. руб.) от уровня технической оснащенности x (тыс.руб.): y =20+700/x .

Доля остаточной дисперсии в общей составив 0,19.

Задание.

1. Определить коэффициент эластичности, предполагая, что стоимость активных производственных фондов составляет 200 тыс. руб.
2. Рассчитать индекс корреляции и F-критерий Фишера. Сделайте выводы.

Задача 6. Изучалась зависимость вида $y = ax^{b}$. Для преобразованных в логарифмах переменных получены следующие данные:

$$\sum xy = 4{,}2087; \sum x^2 = 9{,}2334; \sum x = 4{,}2370; \sum y = 3{,}9310; \sum (y - y_x)^2 = 0{,}0014.$$

Задание.

1. Найдите параметры a и b.
2. Найдите показатель корреляции и оцените его значимость, если $\sigma_y = 0{,}08$.
3. Оцените значимость уравнения, если известно, что n=9.

Задача 7. По совокупности 30 предприятий концерна изучается зависимость прибыли y (тыс.руб.) от выработки продукции на одного работника x_1 (ед.) и индекса цен на продукцию x_2 (%). Данные приведены в табл. 2.2.

Таблица 2.2.

Признак	Среднее значение	Среднее квадратическое отклонение	Парный коэффициент корреляции
y	250	38	$r_{yx_1} = 0{,}68$
x_1	47	12	$r_{yx_2} = 0{,}63$
x_2	112	21	$r_{x_1x_2} = 0{,}42$

Задание:

1. Найти линейное уравнение множественной регрессии в стандартизированной форме и естественной форме.
2. Определить силу влияния каждого фактора.
3. Рассчитать показатель множественной корреляции и детерминации.
4. Рассчитать общий и частные F-критерии Фишера.

Задача 8. По 19 предприятиям оптовой торговли изучается зависимость объема реализация (y) от размера торговой площади (x_1) и товарных запасов (x_2). Были получены следующие варианты уравнения регрессии:

1. $y=25+15x_1$ $r^2=0,9$;
2. $y= 42+27x_2$ $r^2=0,84$;
3. $y= 30+ 10x_1+8x_2$ $R^2=0,92$;
   (2,5) (4,0)
4. $y=21 + 14x_1 + 20x_2 + 0,6x_2^2$ $R^2=0,95$.
   (5,0) (12,0) (0,2)

В скобках указаны значения стандартных ошибок для коэффициентов регрессии.

Задание:

1. Проанализируйте тесноту связи результата с каждым из факторов.
2. Выберите лучшее уравнение регрессии, обоснуйте свое решение.

Задача 9. Для изучения рынка жилья в городе по данным о 46 коттеджах было построено уравнение множественной регрессии:

$y=21,1 - 6,2x_1 + 0,95x_2 + 3,57x_3$; $R^2=0,7$,

(1,8) (0,54) (0,83)

где y- цена объекта, тыс.долларов; x_1- расстояние до центра города, км; x_2- полезная площадь объекта, кв. м; x_3- число этажей в доме, ед.

В скобках указаны значения стандартных ошибок для коэффициентов множественной регрессии.

Задание.

1. Проверьте гипотезу о том, что коэффициент регрессии b_1 в генеральной совокупности равен 0.
2. Проверьте гипотезу о том, что коэффициент регрессии b_2 генеральной совокупности равен 0.
3. Проверьте гипотезу о том, что коэффициент регрессии b_3 в генеральной совокупности равен 0.
4. Проверьте гипотезу о том, что коэффициенты регрессии b_1, b_2 и b_3 в генеральной совокупности одновременно равны 0.
5. Поясните причины расхождения результатов, полученных в пп. 1, 2 и 3, с результатами, полученными в п. 4.

Задача 10. Зависимость спроса на свинину x_1 от цены на нее x_2 и от цены на говядину x_3 представлены уравнением $\lg x_1=0,13-0,21*\lg x_2+2,83*\lg x_3$.

Задание.

1. Представить данное уравнение в естественной форме.
2. Оценить значимость параметров данного уравнения, если известно, что t-критерий для параметра b_2 при x_2 составил 0,827, а для параметра b_3 при x_3 – 1,015.
3. Сделайте выводы о возможности использования уравнения для прогноза.

Задача 11. Анализируется зависимость объема производства продукции предприятиями отрасли черной металлургии от затрат труда и расхода чугуна. Для этого по 20 предприятиям собраны следующие данные: y - объем продукции предприятия в среднем за год (млн. руб.), x_1 - среднегодовая списочная численность рабочих предприятия (чел.), x_2 - средние затраты чугуна за год (млн. т). В таблице 3.2 представлены результаты корреляционного анализа этого массива данных в виде матрицы парных коэффициентов корреляции.

Таблица 3.2

для исходных переменных				для натуральных логарифмов исходных переменных			
	y	x_1	x_2		ln y	x_1	x_2
y	1,00			ln y	1,00		
x_1	0,78	1,00		ln x_1	0,86	1,00	
x_2	0,86	0,96	1,00	ln x_2	0,90	0,69	1,00

Задание.

1. Поясните смысл приведенных выше коэффициентов.

2. Используя эту информацию, опишите ваши предположения относительно:

а) знаков коэффициентов регрессии в уравнениях парной линейной регрессии y по x_1 ($y=a+b\ x_1$) и y по x_2 ($y=a+bx_2$);

б) статистической значимости коэффициентов регрессии при переменных x_1 и x_2 в линейном уравнении множественной регрессии и в уравнении множественной регрессии в форме функции Кобба -Дугласа.

3. Определите значения коэффициентов детерминации в уравнениях

парной линейной регрессии $y=a+bx_1$ и $y=a+bx_2$. Какое из этих уравнений лучше?

4. Найдите уравнение множественной линейной регрессии в стандартизованном масштабе и сделайте выводы.

Задача 12. По данным, полученным от 20 фермерских хозяйств одного из регионов, изучается зависимость объема выпуска продукции растениеводства y (млн руб.) от трех факторов: численности работников L (чел.), количества минеральных удобрений на 1 га посева M (кг) и количества осадков в период вегетации - R (г). Были получены следующие варианты уравнений регрессии и доверительные интервалы для коэффициентов регрессий (табл. 3.3 и 3.4):

l) $y = -5+0,8L+l,2M$, $R^2 =0,75$.

Таблица 3.3

Граница	Доверительные интервалы для коэффициентов регрессии при факторе	
	L	M
Нижняя	0,4	???
Верхняя	???	1,4
Примечание. Доверительные интервалы построены с вероятностью P = 0,95.		

2) $y=2+0,5L+1,7M-2R$, $R^2 =0,77$.

Таблица 3.4

Граница	Доверительные интервалы для коэффициентов регрессии при факторе		
	L	M	R
Нижняя	0,1	???	???
Верхняя	???	2,3	1,5
Примечание. Доверительные интервалы построены с вероятностью P = 0,95.			

Задание.

1. Восстановите пропущенные границы доверительных интервалов в каждом уравнении.

2. Выберите наилучшее уравнение регрессии. Дайте интерпретацию их параметров и доверительных интервалов для коэффициентов регрессии.

3. Каковы ваши предложения относительно значения t-критерия Стьюдента для коэффициента регрессии при факторе R во 2-м уравнении?

Задача 13. Имеется информация по 25 наблюдениям.

Признаки	Среднее значение	Коэффициент вариации, %	Уравнение регрессии
У	35	20	$y = 20 + x_1 - 2{,}0x_2$
x_1	16	30	$y = 9 + 1{,}1x_1$
x_2	8	10	$y = 4 - 4{,}l\ x_2$

Задание.

1. Оцените значимость каждого уравнения регрессии, если известно, что $r_{x1\ x2} = -0{,}35$.

2. Оцените значимость коэффициентов регрессии уравнения с двумя объясняющими переменными.

3. Определите показатели частной корреляции.

4. Найдите частные средние коэффициенты эластичности.

Задача 14. Зависимость объема продаж у (тыс. долл.) от расходов на рекламу х (тыс. долл.) характеризуется по 12 предприятиям концерна следующим образом:

Уравнение регрессии $y=10{,}6+ 0{,}6x$

Среднеквадратичное отклонение х $\sigma_x =4{,}7$

Среднеквадратичное отклонение у $\sigma_y=3{,}4$

Задание.

1. Определите коэффициент корреляции.

2. Оценить значимость уравнения регрессии в целом.

3. Найдите стандартную ошибку оценки коэффициента регрессии.

4. Оцените значимость коэффициента регрессии через t-критерий Стьюдента.

5. Определите доверительный интервал для коэффициента регрессии с вероятностью 0,95 и сделайте экономический вывод.

Задача 15. Покажите, что коэффициент детерминации равен квадрату линейного коэффициента корреляции между х и у.

Задача 16. По 8 наблюдениям построена модель парной регрессии следующих функций:

$y=a+b_1x^3$; $y=a+b/x$; $y=b_1x+b_2x^4$; $y=a+b_1x+b_2/x$.

Задание. Определить какие из построенных моделей точно будут статистически ненадежными.

Задача 17. Экспериментальным способом построены модели, форма связи которых выражаются следующими математическими функциями: $y=a+b_1x+b_2x^3$; $y=a+b/x$; $y=ax^b$. Рассчитанные для них остаточные дисперсии получились соответственно 3,45; 4,12 и 3,11.

Задание. Определить какая из функций лучшего всего описывает требуемую зависимость.

Задача 18. Производственная функция, полученная по данным за 1990 - 1997 гг., характеризуется уравнением

$$\lg P = 0{,}552 + \underset{(0{,}584)}{0{,}276} \cdot \lg Z + \underset{(0{,}065)}{0{,}521} \cdot \lg K, \; R^2 = 0{,}984, \; r^2_{PZ} = 0{,}7826, \; r^2_{PK} = 0{,}9836.$$

где P - индекс промышленного производства; Z - численность рабочих; K - капитал.

В скобках указаны значения стандартных ошибок для коэффициентов регрессии.

Задание.

1. Дайте интерпретацию параметров уравнения регрессии.
2. Оцените значимость параметров регрессии с помощью t-критерия Стьюдента и сделайте соответствующие выводы о целесообразности включения факторов в модель.
3. Оцените значимость уравнения регрессии в целом с помощью F-критерия Фишера.
4. Найдите величины частных значений F-критерия и сделайте соответствующие выводы.
5. Какова роль факторов, не учтенных в модели, в вариации индекса промышленного производства.

Задача 19. По 30 наблюдениям получены следующие данные:

Уравнение регрессии	$y = a + 0{,}176x_1 + 0{,}014x_2 - 7{,}75x_3$
Коэффициент детерминации	0,65
$\bar{y}$	200
$\bar{x}_1$	150
$\bar{x}_2$	20
$\bar{x}_3$	100

Задание. Найдите скорректированный коэффициент корреляции, оцените значимость уравнения регрессии в целом.

Задача 20. Получены функции:

$$y = a + bx^3 + \varepsilon,$$
$$y = a + b\ln x + \varepsilon,$$
$$\ln y = a + b\ln x + \varepsilon,$$
$$y = a + bx^c + \varepsilon,$$
$$y^a = b + cx^2 + \varepsilon,$$
$$y = 1 + a(1 - x^b) + \varepsilon,$$
$$y = a + b\frac{x}{10} + \varepsilon.$$

Задание. Определите, какие из представленных функций линейны по переменным, линейны по параметрам, нелинейны ни по переменным , ни по параметрам.

Задача 21. В таблице 3.5. представлен реальный доход на душу населения y (тыс.долл.), процент рабочей силы, занятой в сельском хозяйстве x_1 и средний уровень образования населения в возрасте после 25 лет x_2 (число лет, проведенных в учебных заведениях) для развитых стран.

Таблица 3.5

№ страны	y	x_1	x_2	№ страны	y	x_1	x_2
1	7	8	9	9	10	6	12
2	9	9	13	10	11	7	14
3	9	7	11	11	11	6	11
4	8	6	11	12	12	4	15
5	8	10	12	13	9	8	15
6	14	4	16	14	10	5	10
7	9	5	11	15	12	8	13
8	8	5	11				

Задание.

1. Постройте множественную регрессию y от x_1 и x_2. Интерпретируйте

полученные результаты.

2. Определите коэффициенты эластичности.

3. Почему, как правило, константа а не играет существенной роли при рассмотрении надежности регрессии?

4. Постройте 95% доверительные интервалы для коэффициентов регрессии.

5. Вычислите коэффициент детерминации (нескорректированный и скорректированный).

6. Проверьте на 5% уровне значимость коэффициентов b_1 и b_2.

Задача 22. По группе предприятий, производящих однородную продукцию, известно, как зависит себестоимость единицы продукции у от факторов, приведенных в таблице.

Признак-фактор	Уравнение парной регрессии	Среднее значение фактора
Объем производства, млн руб.,x_1	$y_x = 0{,}62 + 58{,}74/x$	2,64
Трудоемкость единицы продукции, чел-час, x_2	$y_x = 9{,}30 + 9{,}83x_2$	1,38
Оптовая цена за 1 т. энергоносителя, млн.руб., x_3	$y_x=11{,}75+x_3^{1,63}$	1,503
Доля прибыли, изымаемой государством, %, x_4	$y_x = 14{,}87 \bullet 1{,}016^{x4}$	26,3

Задание.

1. Определить с помощью средних коэффициентов эластичности силу влияния каждого фактора на результат.

2. Ранжировать факторы по силе влияния.

Задача 23. По 20 предприятиям отрасли были получены следующие результаты регрессионного анализа зависимости объема выпуска продукции у (млн руб.) от численности занятых на предприятии x_1 (чел.) и среднегодовой стоимости основных фондов x_2 (млн руб.):

Коэффициент детерминации	0,81
Множественный коэффициент корреляции	???
Уравнение регрессии	$\ln y = ??? + 0{,}48 \ln x_1 + 0{,}62 \ln x_2$

Стандартные ошибки параметров	2	0,06	???
t-критерий для параметров	5	???	5

Задание.

1. Напишите уравнение регрессии, характеризующее зависимость у от x_1 и x_2.
2. Восстановите пропущенные характеристики.
3. С вероятностью 0,95 постройте доверительные интервалы для коэффициентов регрессии.
4. Проанализируйте результаты регрессионного анализа.

Задача 24. По 40 предприятиям одной отрасли исследовалась зависимость производительности труда - у от уровня квалификации рабочих – x_1 и энерговооруженности их труда – x_2. Результаты оказались следующими:

Уравнение регрессии	$y = a + 10\ x_1 + 2\ x_2$		
Стандартные ошибки параметров	0,5	2	7
t-критерий для параметров	3	?	5
Множественный коэффициент корреляции		0,85	

Задание.

1. Определите параметр a и заполните пропущенные значения.

2. Оцените значимость уравнения в целом, используя значение множественного коэффициента корреляции.

3. Какой из факторов оказывает более сильное воздействие на результат?

Задача 25. По данным за 20 месяцев построено уравнение регрессии зависимости удоев молока на одну корову (ц) от расходов на корма (ц.к.ед.) и квалификации животноводов (% работников мастеров 1 и 2 класса):

$У=-18 + 1,5*x_1 +0,02*x_2$

Имеются следующие данные:

№	y	x_1	x_2
1	28	30	80
2	25	28	85
3	34	34	78

…	…	…	…

$\Sigma\varepsilon^2_t$ =10400 $\Sigma(\varepsilon_t - \varepsilon_{t-1})^2$=38500

Задание:

1. По имеющимся данным рассчитать теоретические значения y_t, ε_t, ε_{t-1} , ε^2_t ,$(\varepsilon_t - \varepsilon_{t-1})^2$.
2. Рассчитать критерий Дарбина-Уотсона и оценить наличие автокорреляции в остатках при уровне значимости 0,05.
3. Указать, пригодно ли уравнение для прогноза.

Задача 26. Имеются данные о величине заработной платы и расходах на непродовольственные товары.

годы	1998	1999	2000	2001	2002	2003
Расходы на непродовольственные товары, тыс. руб.	31	35	39	43	50	53
Заработная плата, % к 1998 г.	100	103	106	108	115	118

Задание:

1. Определить ежегодные абсолютные приросты зарплаты и расходов, сделать выводы о тенденции развития каждого ряда.
2. Перечислить основные пути устранения тенденции для построения модели спроса на непродовольственные товары в зависимости от заработной платы.
3. Рассчитать линейную модель спроса, используя первые разности уровней исходных динамических рядов. Дать ей экономическую интерпретацию.
4. Построить модель спроса с включением фактора времени, интерпретировать ее параметры.

Задача 27. На основе данных по приросту мяса КРС за 24 месяца получены следующие значения коэффициентов автокорреляции уровней первого и т.д. порядков:

r_1 = 0,71; r_2 = 0,83; r_3 = 0,91; r_4 = 0,51; r_5 = 0,68; r_6 = 0,38.

Задание:

3. Охарактеризуйте структуру ряда, определив наличие тенденции и циклических колебаний.

4. Обоснуйте выбор уравнения регрессии (авторегрессии) для прогноза будущих значений ряда.

Задача 28. Пусть имеется следующий временной ряд:

t	1	2	3	4	5	6	7	8
x_t	20	...	...	...				10

Известно также, что $\Sigma x_t = 150$, $\Sigma x^2_t = 8100$, $\Sigma x_t x_{t-1} = 7350$ (t=2..n).

Задание:

1. Определите коэффициент автокорреляции уровней этого ряда первого порядка.

2. Установите, включает ли исследуемый временной ряд тенденцию.

Задача 29. Имеются поквартальные данные изменения объемов сельскохозяйственного производства (в сопост. ценах % к предыдущему году).

Номер квартала	Объем производства, % к предыдущему периоду	Номер квартала	Объем производства, % к предыдущему периоду
1	100,0	11	98,8
2	93,9	12	101,9
3	96,5	13	113,1
4	101,8	14	98,4
5	107,8	15	97,3
6	96,3	16	102,1
7	95,7	17	97,6
8	98,2	18	83,7
9	104,0	19	84,3
10	99,0	20	88,4

Задание:

1. Постройте аддитивную модель временного ряда.

2. Оцените качество модели через показатели средней абсолютной ошибки и «коэффициента детерминации».

3. Спрогнозируйте изменение объемов производства в I полугодии шестого года.

Задача 30. Администрация банка изучает динамику депозитов физических лиц за ряд лет (млн долл. в сопоставимых ценах). Исходные данные представлены ниже:

Время, лет	1	2	3	4	5	6	7
Депозиты физических лиц, x	2	6	7	3	10	12	13

Известно также следующие $\Sigma x^2=511$

Задание:

1. Постройте уравнение линейного тренда и дайте интерпретацию его параметров.

2. Определите коэффициент детерминации для линейного тренда.

3. Администрация банка предполагает, что среднегодовой абсолютный прирост депозитов физических лиц составляет не менее 2,5 млн. долл. Подтверждается ли это предположение результатами, которые вы получили?

Задача 31. Для прогнозирования объема продаж колбасных изделий (млн.руб.) на основе поквартальных данных за 2004-2008 гг. была построена аддитивная модель временного ряда объема продаж. Уравнение, моделирующее динамику трендовой компоненты этой модели, имеет вид: T= 100 + 2 • t (при построении тренда для моделирования переменной времени использовались натуральные числа, начиная с 1). Показатели за 2007 г., полученные в ходе построения аддитивной модели, представлены в таблице.

Время года	Фактический объем продаж в 2007 г.	Компонента, полученная по аддитивной модели		
		трендовая	сезонная	случайная
Зима	100			+4
Весна			10	+5
Лето	150		25	
Осень				

Задание: Определите недостающие в таблице данные, учитывая, что объем продаж перерабатывающего цеха колбасных изделий за 2007 г. в целом составил 490 млн. руб.

Задача 32. В целях прогнозирования объема экспорта страны на будущие периоды были собраны данные за 30 лет по следующим показателям: y_t -

объем экспорта (млрд долл., в сопоставимых ценах); x_t -индекс физического объема промышленного производства (в % к предыдущему году). Ниже представлены результаты предварительной обработки исходных данных.

1 .Уравнения линейных трендов:

а) для ряда Y_t

$Y_t = 3{,}1 + 1{,}35\, t + \varepsilon, \quad R^2 = 0{,}91 \quad d = 2{,}31;$

б) для ряда X_t

$X_t = -8{,}4 + 4{,}8t + \varepsilon, \quad R^г = 0{,}89 \quad d = 2{,}08.$

2. Уравнение регрессии по уровням временных рядов:

$Y_t = -10{,}5 + 0{,}5X_t + \varepsilon \qquad R^2 = 0{,}95\ d = 2{,}21$

3. Уравнение регрессии по первым разностям уровней временных рядов:

$\Delta Y_t = 1{,}4 + 0{,}03\, \Delta X_t + \varepsilon \quad R^2 = 0{,}86 \quad d = 2{,}25.$

4. Уравнение регрессии по вторым разностям уровней временных рядов:

$\Delta^2 Y_t = 0{,}7 + 0{,}012\, \Delta^2 X_t + \varepsilon \quad R^2 = 0{,}47 \quad d = 2{,}69.$

5. Уравнение регрессии по уровням временных рядов с включением фактора времени:

$Y_t = 4{,}23 + 0{,}24\, x_t + 0{,}78\, t + \varepsilon, \quad R^2 = 0{,}97 \quad d = 0{,}9.$

Задание:

1. Сформулируйте свои предположения относительно величины коэффициента автокорреляции первого порядка в каждом из рядов. Ответ обоснуйте.

2. Выберите наилучшее уравнение регрессии, которое можно использовать для прогнозирования объема экспорта, и дайте интерпретацию его параметров.

3. Пусть известна информация за последние три года:

Год (номер периода t)	28	29	30	31
$У_t$	38	742	43	???
X_t	120	126	121	124

Используя выбранное вами в п. 2 уравнение, дайте точечный прогноз ожидаемого значения y_t на ближайший год (период 31).

Задача 33. На основе поквартальных данных об уровне безработицы в летнем курортном городе (% от экономически активного населения) за последние 5 лет была построена мультипликативная модель временного ряда. Скорректированные значения сезонной компоненты за каждый квартал приводятся ниже:

1 квартал – 1,4   3 квартал – 0,7

2 квартал – 0,8   4 квартал - ?

Уравнение тренда выглядит следующим образом: $T = 9{,}2 - 0{,}3t$; (при расчете параметров тренда для нумерации кварталов использовались натуральные числа t = 1: 20).

Задание:

1. Определите значения сезонной компоненты за IV квартал.

2. На основе построенной модели дайте точечные прогнозы уровня безработицы на I и II квартал следующего года.

Задача 34. Изучается зависимость падежа телят (y_t) от расхода кормов на 1 голову (x_t) по кварталам:

Показатель	1 кв.	2 кв.	3 кв.	4 кв.	5 кв.	6 кв.
Расход кормов на 1 голову, % к кварталу 1	100	104	112	117	121	126
Падеж телят, голов	89	83	80	77	75	72

Известно также, что $\Sigma x_t = 680$, $\Sigma y_t = 476$, $\Sigma x_t y_t = 53648$, $\Sigma x^2_t = 77566$.

Задание:

1. Постройте модель зависимости падежа молодняка скота от расхода кормов с включением фактора времени.

2. Дайте интерпретацию параметров полученной вами модели.

Задача 35. Управление сельского хозяйства изучает динамику численности тракторов и комбайнов, имеющихся в хозяйствах района за ряд лет (тыс. единиц).

Время, лет	1	2	3	4	5	6	7

Численность тракторов и комбайнов, тыс. ед.	2	6	7	3	10	12	13

Известно также следующие $\Sigma x^2=511$

Задание:

1. Постройте уравнение линейного тренда и дайте интерпретацию его параметров.

2. Определите коэффициент детерминации для линейного тренда.

3. Работники управления предполагают, что среднегодовой абсолютный прирост численности техники составляет не менее 2,5 тыс. единиц. Подтверждается ли это предположение результатами, которые вы получили?

Задача 36. Имеются данные об урожайности зерновых в хозяйстве:

годы	1	2	3	4	5	6	7	8
урожайность, ц/га	10,2	10,8	11,8	13,0	14,8	17,2	20,1	23,2

Задание:

1. Обоснуйте выбор типа уравнения тренда.

2. Рассчитайте параметры уравнения тренда.

3. Дайте прогноз урожайности зерновых на следующий год.

Задача 37. Имеются данные об объеме экспорта зерна из региона (тонн) за 2003-2008 гг.

Номер Квартала	Экспорт, тонн	Номер квартала	Экспорт, тонн
1	4087	13	6975
2	4737	14	6891
3	5768	15	7527
4	6005	16	7971
5	5639	17	5875
6	6745	18	6140
7	6311	19	6248
8	7107	20	6041
9	5741	21	4626
10	7087	22	6501

Задание:

1. Постройте график временного ряда.

2. Постройте аддитивную и мультипликативную модели временного ряда.

3. Оцените качество каждой модели через показатели средней абсолютной ошибки и «коэффициент детерминации». Выберите лучшую модель.

Задача 38. На основе помесячных данных о потреблении электроэнергии в регионе (млн кВт*ч) за последние 3 года была построена аддитивная модель временного ряда. Скорректированные значения сезонной компоненты за соответствующие месяцы приводятся ниже:

Январь	+25	Май	-32	Сентябрь	+3
Февраль	+10	Июнь	-39	Октябрь	+14
Март	+6	Июль	-22	Ноябрь	+27
Апрель	-4	Август	-20	Декабрь	?

Уравнение тренда выглядит следующим образом: $T = 300 + 1{,}4t$,

(при расчете параметров тренда для моделирования переменной времени использовались натуральные числа $t=1,\ldots 36$).

Задание:

1. Определите значение сезонной компоненты за декабрь.

2. На основе построенной модели дайте точечный прогноз ожидаемого потребления электроэнергии в течение первого квартала следующего года.

Задача 39. В таблице приводятся данные о производстве и запасах сахара-рафинада за семь лет.

Годы	1	2	3	4	5	6	7
Запасы сахара на нач. года, тыс. тонн	300	310	325	340	350	370	385
Производство сахара, тыс. тонн.	335	340	360	378	400	417	430

Задание:

1. Постройте уравнение линейной регрессии, используя метод первых разностей.

2. Охарактеризуйте тесноту связи между рядами по их уровням, по первым разностям. Сделайте выводы.

Задача 40. Известны данные об уровне дивидендов, выплачиваемых по обыкновенным акциям (в процентах), и среднегодовой стоимости основных фондов компании (х, млн руб.) в сопоставимых ценах за последние девять лет.

Показатель	1	2	3	4	5	6	7	8	9
Среднегодовая стоимость основных	72	75	77	77	79	80	78	79	80
Дивиденды по обыкновенным акциям	4,2	3,0	2,4	2,0	1,9	1,7	1,8	1,6	1,7

Задание:

1. Определите параметры уравнения регрессии по первым разностям и дайте их интерпретацию. В качестве зависимой переменной используйте показатель дивидендов по обыкновением акциям.
2. В чем состоит причина построения уравнения регрессии по первым разностям, а не по исходным уровням рядов?

Задача 41.

Модель денежного рынка:

$$\begin{cases} P=a_{10}+a_{11}D+a_{13}V\ , \\ V=a_{20}+a_{21}P+a_{23}I\ , \end{cases}$$

где P – процентная ставка; D – денежная масса; V – ВВП; I – внутренние инвестиции.

Задание.

1. Применив необходимые и достаточное условия идентификации, определить, идентифицировано ли каждое уравнение модели и система в целом.
2. Запишите приведенную форму модели.

Задача 42.

Строится модель вида:

$y_1=a_{10}+a_{11}y_2+a_{12}x_1$

$y_2=a_{20}+a_{21}y_1+a_{22}x_2$

Задание. Определите структурные коэффициенты, учитывая, что $\Sigma y_1x_1=2600$, $\Sigma y_1x_2=4350$, $\Sigma y_1=350$, $\Sigma y_2=25$, $\Sigma x_1=750$, $\Sigma x_2=350$, $\Sigma x_1^2=1200$, $\Sigma x_2^2=1800$, $n=30$, $\Sigma x_1x_2=1500$, а также $y_2=2x_1+3x_2$.

Задача 43.

Пусть известные следующие темпы прироста:

Период		% безработ-

времени	заработной платы, y_1	цен, y_2	дохода, y_3	цен на импорт, x_2	цен на экспорт, x_3	ных, x_1
1	2	6	10	2	1	1
2	3	7	12	3	2	2
3	4	8	11	1	5	3
4	5	5	15	4	3	2
5	6	4	14	2	3	3
6	7	9	16	2	4	4
7	8	10	18	3	4	5

Задание.

Определите параметры структурной модели следующего вида:

$y_1=a_{11}y_2+a_{12}x_1+a_{13}x_2$

$y_2=a_{21}y_1+a_{22}x_2+a_{23}x_3$

$y_3=a_{31}y_1+a_{32}x_2+a_{33}x_3$

Задача 44. Макроэкономическая модель имеет вид:

$$\begin{cases} C_t = a_{10}+a_{11}Y_t + a_{12}T_t, \\ I_t = a_{20}+a_{21}Y_t + a_{23}K_{t-1}, \\ Y_t = C_t + I_t, \end{cases}$$

где C – конечное потребление; Y – доход; T – налоги; I – инвестиции; K – запас капитала; t – текущий период; t-1 – предыдущий период.

Задание.

1. Применив необходимые и достаточное условия идентификации, определить, идентифицировано ли каждое уравнение модели и система в целом.
2. Запишите приведенную форму модели и определите метод оценки параметров модели.

Приложение А

1. Таблица значений F-критерия Фишера при уровне значимости **$\alpha = 0,05$**.

$к_1$ / $к_2$	1	2	3	4	5	6	8	12	24	∞
1	161,45	199,50	215,72	224,57	230,17	233,97	238,89	243,91	249,04	254,32
2	18,51	19,00	19,16	19,25	19,30	19,33	19,37	19,41	19,45	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,84	8,74	8,64	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,04	5,91	5,77	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,82	4,68	4,53	4,36
6	5,99	5,14	4,76	4,53	4,39	4,28	4,15	4,00	3,84	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,73	3,57	3,41	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,44	3,28	3,12	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,23	3,07	2,90	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3.07	2,91	2,74	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	2,95	2,79	2,61	2,40
12	4,75	3,88	3,49	3,26	3,11	3,00	2,85	2,69	2,50	2,30
13	4,67	3,80	3,41	3,18	3,02	2,92	2,77	2,60	2,42	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,70	2,53	2,35	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,64	2,48	2,29	2,07
16	4,49	3,63	3,24	3.01	2,85	2,74	2,59	2,42	2,24	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,55	2,38	2,19	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,51	2,34	2,15	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,48	2,31	2,11	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,45	2,28	2,08	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,42	2,25	'2,05	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,40	2,23	2,03	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,38	2,20	2,00	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,36	2,18	1,98	1,73
25	4,24	3,38	2,99	2,76	2,60	2,49	2,34	2,16	1,96	1,71
26	4,22	3,37	2,98	2.74	2,59	2,47	2,32	2,15	1,95	1,69
27	4,21	3,35	2,96	2,73	2,57	2,46	2,30	2,13	1,93	1,67
28	4,20	3,34	2,95	2,71	2,56	2,44	2,29	2,12	1,91	1,65
29	4,18	3,33	2,93	2,70	2,54	2,43	2,28	2,10	1,90	1,64
30	4,17	3,32	2,92	2,69	2,53	2,42	2,27	2,09	1,89	1,62
35	4,12	3,26	2,87	2,64	2,48	2,37	2,22	2,04	1,83	1,57
40	4,08	3,23	2,84	2,61	2,45	2,34	2,18	2,00	1,79	1,51
45	4,06	3,21	2,81	2,58	2,42	2,31	2,15	1,97	i,76	1,48

Продолжение

k_2 \ k_1	1	2	3	4	5	6	8	12	24	∞
50	4,03	3,18	2,79	2,56	2,40	2,29	2,13	1,95	1,74	1,44
60	4,00	3,15	2,76	2,52	2,37	2,25	2,10	1,92	1,70	1,39
70	3,98	3,13	2,74	2,50	2,35	2,23	2,07	1,89	1,67	1,35
80	3,96	3,11	2,72	2,49	2,33	2,21	2,06	1.88	1,65	1,31
90	3,95	3,10	2,71	2,47	2,32	2,20	2,04	1,86	1,64	1,28
100	3,94	3,09	2,70	2,46	2,30	2,19	2,03	1,85	1,63	1,26
125	3,92	3.07	2,68	2,44	2,29	2,17	2,01	1,83	1,60	1,21
150	3,90	3,06	2,66	2,43	2,27	2,16	2,00	1,82	1,59	1,18
200	3,89	3,04	2,65	2,42	2,26	2,14	1,98	1,80	1,57	1,14
300	3,87	3,03	2,64	2,41	2,25	2,13	1,97	1,79	1,55	1,10
400	3,86	3,02	2,63	2,40	2,24	2,12	1,96	1,78	1,54	1,07
500	3,86	3,01	2,62	2,39	2,23	2,11	1,96	1,77	1,54	1,06
1000	3,85	3,00	2,61	2,38	2,22	2,10	1,95	1,76	1,53	1,03
∞	3,84	2.99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

2. Критические значения t-критерия Стьюдента при уровне значимости 0,10, 0,05, 0,01 (двухсторонний)

Число степеней свободы d.f.	α			Число степеней свободы d.f.	α		
	0,10	0,05	0,01		0,10	0,05	0,01
1	6,3138	12,706	63,657	18	1,7341	2,1009	2,8784
2	2,9200	4,3027	9,9248	19	1,7291	2,0930	2,8609
3	2,3534	3,1825	5,8409	20	1,7247	2,0860	2,8453
4	2,1318	2,7764	4,6041	21	1,7207	2,0796	2,8314
5	2,0150	2,5706	4,0321	22	1,7171	2,0739	2,8188
6	1,9432	2,4469	3,7074	23	1,7139	2,0687	2,8073
7	1,8946	2,3646	3,4995	24	1,7109	2,0639	2,7969
8	1,8595	2,3060	3,3554	25	1,7081	2,0595	2,7874
9	1,8331	2,2622	3,2498	26	1,7056	2,0555	2,7787
10	1,8125	2,2281	3,1693	27	1,7033	2,0518	2,7707
11	1,7959	2,2010	3,1058	28	1,7011	2,0484	2,7633
12	1,7823	2,1788	3,0545	29	1,6991	2,0452	2,7564
13	1,7709	2,1604	3,0123	30	1,6973	2,0423	2,7500
14	1,7613	2,1448	2,9768	40	1,6839	2,0211	2,7045
15	1,7530	2,1315	2,9467	60	1,6707	2,0003	2,6603
16	1,7459	2,1199	2,9208	120	1,6577	1,9799	2,6174
17	1,7396	2,1098	2,8982	00	1,6449	1,9600	2,5758

Значения статистик Дарбина – Уотсона d_L d_u при 5%-ном уровне значимости

n	**к=1**		***к=2***		**к=3**		***k=4***		***к*=5**	
	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u	d_L	d_u
6	0,61	1,40	-	-	-	-				
7	0,70	1,36	0,47	1,90	-	-				
8	0,76	1,33	0,56	1,78	0,37	2,29				
9	0,82	1,32	0,63	1,70	0,46	2,13				
10	0,88	1,32	0,70	1,64	0,53	2,02				
11	0,93	1,32	0,66	1,60	0,60	1,93				
12	0,97	1,33	0,81	1,58	0,66	1,86				
13	1,01	1,34	0,86	1,56	0,72	1,82				
14	1,05	1,35	0,91	1,55	0,77	1,78				
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1/65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83

ЛИТЕРАТУРА.

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрии: учебник. – М.: ЮНИТИ, 1998. – 143 с.
2. Афанасьев В.Н., Юзбашев М.М., Гуляева Т.И. Эконометрика: учебник /под ред. В.Н. Афанасьева. – М.: Финансы и статистика, 2005. — 256 с.
3. Белокопытов А.В., Смирнов В.Д. Методы корреляционно-регрессионного анализа в эконометрических исследованиях. - Смоленск: ООО «Принт-экспресс», 2004. – 150 с.
4. Бородич С.А. Эконометрика: Учебное пособие. – Минск: Новое знание, 2001. – 408 с.
5. Кремер Н.Ш., Путько Б.А. Эконометрика:учебник. – М., 2002. – 311 с.
6. Курзенев В.А. Вероятность и статистика в управлении (с примерами и задачами). - СПб.: Изд-во Северо-Западной академии государственной службы, 1998. - 160 с.
7. Магнус Я. Катышев П., Пересецкий А. Эконометрика. Начальный курс. – М., 1998. – 248 с.
8. Мардас А.Н. Эконометрика. – СПб.: изд. «Питер», 2001. – 144 с.
9. Практикум по эконометрике. /под ред. И.И. Елисеевой. – М, Финансы и статистика, 2003. – 192 с.
10. Просветов Г.И. Эконометрика. Задачи и решения: учебно-методическое пособие. – 2-е изд. – М., 2005. – 104 с.
11. Свириденкова М.А. Эконометрика: методические указания. – Смоленск: СИБП, 2001. – 41 с.
12. Смирнов В.Д. Методы статистического прогнозирования: учебно-методическое пособие. – Смоленск, 2003. – 67 с.
13. Статистика: учебное пособие. /под ред. Харченко Л.П. и др. – М., 2001. – 241 с.
14. Эконометрика: учебник./под ред. И.И. Елисеевой. – М, Финансы и статистика, 2003. – 344 с.

Содержание