Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Уфимский государственный авиационный технический университет»

Ю. Т. МАНСУРОВА, П. А. ТУКТАРОВА

ЭКОНОМЕТРИКА

Уфа 2022

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Уфимский государственный авиационный технический университет»

Ю. Т. МАНСУРОВА, П. А. ТУКТАРОВА

ЭКОНОМЕТРИКА

*Допущено Редакционно-издательским советом УГАТУ
в качестве учебного пособия для студентов очной и заочной форм обучения,
обучающихся по направлению подготовки специалистов
38.05.01 Экономическая безопасность*

Учебное электронное издание сетевого доступа

ISBN 978-5-4221-1622-5

Уфа 2022

Рецензенты:
начальник управления перспективного развития филиала
ПАО АНК «Башнефть» «Башнефть-Уфанефтехим»
канд. техн. наук А. Р. Галиакбиров;
профессор кафедры газохимии и моделирования химико-технологических процессов УГНТУ д-р техн. наук Т. Г. Умергалин

Мансурова Ю. Т., Туктарова П. А.
Эконометрика : курс лекций : учебное пособие [Электронный ресурс] / Уфимск. гос. авиац. техн. ун-т. – Уфа : УГАТУ, 2022. – URL: https://www.ugatu.su/media/uploads/MainSite/Ob%20universitete/Izdateli/El_izd/2022-173.pdf

Изложены основные теоретические аспекты моделирования парной, множественной регрессий, систем эконометрических уравнений, а также временных рядов. Каждый раздел сопровождается тестовыми заданиями и контрольными вопросами.

Предназначено для студентов, изучающих дисциплину «Эконометрика». Может быть использовано в качестве опорного материала при подготовке выпускной квалификационной работы, а также при написании научных статей с применением описанных эконометрических методов и моделей.

При подготовке электронного издания использовались следующие программные средства:

- Adobe Acrobat – текстовый редактор;
- Microsoft Word – текстовый редактор.

Авторы: *Мансурова Юлия Талгатовна*
Туктарова Полина Андреевна

Корректура и верстка *Л. А. Вяземская*
Программирование и компьютерный дизайн *О. М. Толкачёва*

Подписано к использованию: 23.08.2022
Объем: 4,89 Мб.

ФГБОУ ВО «Уфимский государственный авиационный технический университет»
450008, Уфа, ул. К. Маркса, 12.
Тел.: +7-908-35-05-007
e-mail: rik@ugatu.su

ВВЕДЕНИЕ

Эконометрика есть единство трех составляющих – статистики, экономической теории и математики. Иными словами, эконометрика объединяет совокупность методов и моделей, позволяющих на базе экономической теории, экономической статистики и математического инструментария придавать количественные выражения качественным закономерностям.

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 статистических данных и экономических показателей, а эконометрика, используя традиционные математические методы, проводит анализ качественных взаимосвязей между этими показателями.

Многие базовые понятия эконометрики имеют два определения – экономическое и математическое. При этом экономическая составляющая является первичной. Экономика определяет постановку задачи и исходные предпосылки, а результат, формируемый на математическом языке, представляет интерес в том случае, если удается его экономическая интерпретация. В то же время многие экономические результаты носят характер математических утверждений.

Предметом эконометрии являются факторы, формирующие развитие экономических явлений и процессов. Предпосылки, на которых основываются оценки факторов развития экономики, связаны с риском. Для уменьшения ошибок необходимо включить в эконометрические расчеты все без исключения факторы и выбрать наиболее эффективные методы их оценки, которые обеспечили бы их достоверность.

Развитие экономики, усложнение экономических процессов и повышение требований к принимаемым управленческим решениям в области макро- и микроэкономики потребовало более тщательного и объективного анализа реально протекающих процессов на основе привлечения современных математических и статистических методов.

С другой стороны, проблема нарушения предпосылок классических статистических методов при решении реальных экономических задач привела к необходимости развития и совершенствования классических методов математической статистики и уточнения постановок соответствующих задач.

В результате этих процессов осуществилось выделение и формирование новой отрасли знания под названием «эконометрика», связанной с разработкой и применением методов количественной оценки экономических явлений и процессов и их взаимосвязей.

В данном учебном пособии рассмотрены следующие разделы по дисциплине «Эконометрика»: парная линейная и нелинейная модели регрессии, множественный регрессионный анализ, построение систем одновременных эконометрических уравнений регрессии, моделирование структуры временного ряда, анализ взаимодействия двух временных рядов.

Изучение курса «Эконометрика» студентами специальности 38.05.01 Экономическая безопасность способствует формированию следующих компетенций:

– способен рассчитывать экономические показатели, характеризующие деятельность хозяйствующих субъектов (ОПК-3);

– способен использовать современные информационные технологии и программные средства для решения профессиональных задач (ОПК-6);

– способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности (ОПК-7).

Изучение данной дисциплины является основополагающим для понимания таких дисциплин, как: «Актуарная математика», «Экономические и финансовые риски», «Экономико-математическое моделирование», преддипломная и производственная практика, ГИА.

1. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Основная задача эконометрики заключается в исследовании и количественной оценке объективно существующих взаимосвязей и зависимостей между экономическими явлениями. Наибольший интерес для исследователя представляют причинно-следственные отношения между явлениями, что позволяет выявлять факторы, оказывающие основное влияние на вариацию изучаемых явлений и процессов.

Причинно-следственное отношение – это такая связь между явлениями, при которой изменение одного из них, называемого причиной, ведет к изменению другого, называемого следствием. Следовательно, причина всегда предшествует следствию.

Причинно-следственные связи в социально-экономических явлениях обладают следующими особенностями.

Во-первых, причина x и следствие y взаимодействуют не непосредственно, а через промежуточные факторы, которые, как правило, при анализе опускаются.

Во-вторых, социально-экономические явления развиваются и формируются в результате одновременного воздействия большого числа факторов. Поэтому одной из главных проблем при изучении этих явлений становится задача выявления существенных причин и абстрагирование от второстепенных.

Признаки по их роли в изучаемой взаимосвязи делятся на два класса: факторные и результативные.

Факторными признаками (факторами) называются признаки, обусловливающие изменения других, связанных с ними признаков. Факторные признаки называются также независимыми, объясняющими, или входными переменными.

Результативными называются признаки, изменяющиеся под действием факторных признаков. Результативные признаки называются также зависимыми, объясняемыми, или выходными переменными.

По направлению изменения связи подразделяются на прямые (когда изменение результативного и факторного признаков происходит в одном направлении) и обратные (когда изменение результативного и факторного признаков происходит в противоположных направлениях).

По характеру проявления различают функциональную связь и стохастическую зависимость.

Функциональной называют такую связь, при которой определенному значению факторного признака соответствует одно и только одно значение результативного признака.

Функциональная связь проявляется во всех случаях наблюдения и для каждой конкретной единицы исследуемой совокупности. Такие связи изучаются в основном в естественных науках.

В эконометрике в основном изучаются причинные зависимости, которые проявляется не в каждом отдельном случае, а в общем, среднем при большом числе наблюдений. То есть одним и тем же значениям факторных признаков, как правило, соответствуют различные значения результативного признака.

Но тем не менее, рассматривая всю совокупность наблюдений, можно отметить наличие определенной зависимости между значениями признаков. Такие причинные зависимости называются стохастическими.

Частным случаем стохастической связи является корреляционная связь, при которой изменение среднего значения результативного признака обусловлено изменением факторных признаков.

По аналитическому выражению выделяют связи линейные и нелинейные.

Линейной называется связь, в которой изменение результативного признака прямо пропорционально изменению факторных признаков. В противном случае связь называется нелинейной. Аналитически линейная стохастистическая связь между явлениями может быть представлена уравнением прямой линии на плоскости либо уравнением гиперплоскости в n-мерном пространстве (при наличии n факторных переменных).

Для изучения различных экономических явлений эконометристы используют их упрощенные формальные описания, называемые эконометрическими моделями. При построении модели выделяются существенные факторы, определяющие исследуемое явление, и отбрасываются детали, несущественные для решения поставленной проблемы.

Чтобы получить достоверные данные о каком-либо явлении, необходимо иметь выборку достаточно большого объема. Выборка наблюдений зависимой переменной и объясняющих переменных является отправной точкой любого эконометрического исследования. Обычно рассматриваются два типа выборочных данных.

Пространственная выборка или пространственные данные. В экономике под пространственной выборкой понимают набор показателей экономических переменных, полученных в данный момент времени. О пространственной выборке имеет смысл говорить в том случае, если все наблюдения получены примерно в неизменных условиях, то есть представляют собой набор независимых выборочных данных из некоторой генеральной совокупности.

Временной (динамический) ряд. Временным (динамическим) рядом называется выборка наблюдений, в которой важны не только сами наблюдаемые значения случайных величин, но и порядок их следования друг за другом. Обычно упорядоченность обусловлена тем, что экспериментальные данные представляют собой серию наблюдений одной и той же случайной величины в последовательные моменты времени.

Можно выделить шесть основных этапов эконометрического моделирования.

Этап I (постановочный). Формируется цель исследования, набор участвующих в модели экономических переменных.

Этап II (априорный). Проводится анализ сущности изучаемого объекта, формирование и формализация априорной (известной до начала моделирования) информации.

Этап III (параметризация). Осуществляется непосредственное моделирование, то есть выбор общего вида модели, выявление входящих в нее связей.

Этап IV (информационный). Осуществляется сбор необходимой статистической информации – наблюдаемых значений экономических переменных.

Этап V (идентификация модели). Осуществляется статистический анализ модели и оценка ее параметров.

Этап VI (верификация модели). Проводится проверка истинности, адекватности модели.

Существует несколько подходов к классификации эконометрических моделей. Наиболее существенное значение при практическом использовании имеют следующие подходы:

– по направлению и сложности причинных связей эконометрические модели могут быть разделены на регрессионные и структурные уравнения. В свою очередь регрессионные делят на уравнения парной и множественной регрессии, а структурные уравнения – на системы взаимозависимых уравнений и рекурсивные системы;

– по аналитической форме зависимости модели бывают линейные, нелинейные, степенные и другие;

– по учету фактора времени многофакторные модели делят на статистические и динамические. Эконометрическая модель считается динамической, если в ее уравнения вводят переменные с временным запаздыванием или если время играет в модели роль самостоятельной переменной. В противном случае модель считается статистической.

Основная цель эконометрического регрессионного анализа – дать количественное описание взаимосвязей между экономическими переменными.

В зависимости от количества факторов, включенных в уравнение регрессии, принято различать простую (парную) и множественную регрессии.

Парная регрессия достаточна, если имеется доминирующий фактор, который и используется в качестве объясняющей переменной. Нужно отметить, что не следует ожидать получения точного соотношения между какими-либо двумя экономическими показателями, за исключением тех случаев, когда оно существует по определению. В статистическом анализе факт неточности соотношения признается путем явного включения в него случайного фактора, описываемого случайным остаточным членом.

$$y = \alpha + \beta x + \varepsilon. \tag{1.1}$$

Величина y, рассматриваемая как зависимая переменная, состоит из двух составляющих:

1) неслучайной составляющей $\alpha + \beta x$, где x выступает как объясняющая переменная, а постоянные α и β – как параметры уравнения;

2) случайного члена ε.

На рис. 1.1 показано, как комбинация этих двух составляющих определяет величину y. Показатели x_1, x_2, x_3, x_4 – это четыре гипотетических значения независимой переменной. Если бы соотношение между y и x было точным, то соответствующие значения y были бы представлены точками Q_1, Q_2, Q_3, Q_4 на прямой. Наличие случайного члена (возмущения) приводит к тому, что в действительности значение y получается другим. Случайный член положителен в первом и четвертом наблюдениях и отрицателен в двух других. Если отметить на графике реальные значения y при соответствующих значениях x, то получим точки P_1, P_2, P_3, P_4.

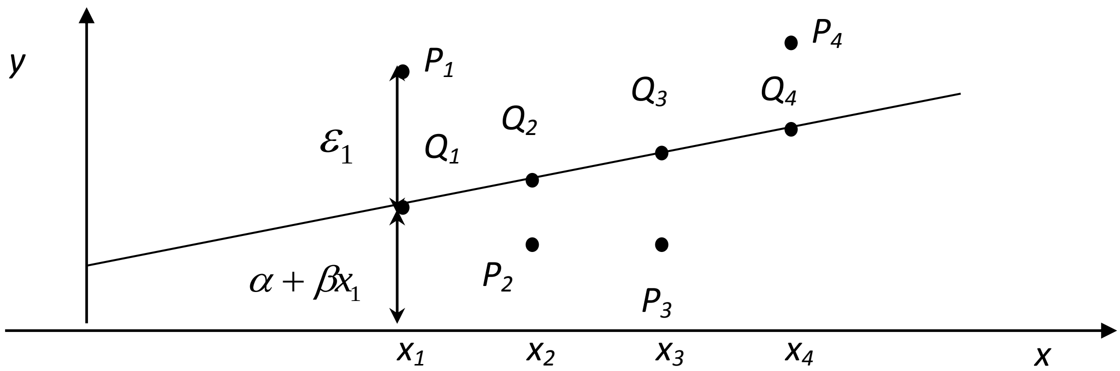


Рис. 1.1. Истинная зависимость между y и x

Нужно отметить, что точки P – это единственные точки, отражающие реальные значения переменных на рисунке. Фактические значения α и β, следовательно, положения точек Q неизвестны, так же, как и фактические значения случайного члена. Задача регрессионного анализа состоит в получении оценок α и β, в определении положения прямой по точкам P.

К основным причинам появления случайной величины относятся:

– невключение независимых переменных. Существует огромное множество факторов, в той или иной мере влияющих на y. Обычно некоторые из них не учитываются и не вводятся в качестве независимых переменных x;

– агрегирование переменных. Рассматриваемая зависимость – попытка объединить вместе несколько микроэкономических соотношений. Например, функция суммарного потребления – попытка общего выражения совокупности решений отдельных индивидов о расходах. Так как отдельные соотношения имеют различные параметры, то попытка объединения лишь аппроксимация, поэтому расхождения неизбежны;

– неправильное описание структуры модели. Например, если зависимость относится к данным о временном ряде, то значение y

может зависеть не от фактического значения x, а от значения, которое ожидалось в предыдущем периоде;

– неправильная функциональная спецификация. Например, истинная зависимость может являться не линейной, а быть более сложной;

– ошибки измерения. Если в измерении одной или более взаимосвязанных переменных имеются ошибки, то наблюдаемые значения будут иметь отклонения.

Зависимость y от x в парной регрессии не обязательно характеризуется линейной парной регрессией, возможны и другие соотношения, например степенная или экспоненциальная зависимости.

Поэтому от правильно выбранной спецификации модели зависит величина случайных ошибок: они тем меньше, чем в большей мере теоретические значения результативного признака y_x подходят к фактическим данным y.

Предположим, что для оценки параметров регрессии взята выборка, содержащая n пар значений переменных (x_i, y_i), где $i = 1,2,...,n$. В этом случае модель (1.1) примет вид

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad (1.2)$$

Отметим основные предпосылки регрессионного анализа, которые известны как условия Гаусса – Маркова.

1-е условие Гаусса – Маркова: математическое ожидание возмущения ε_i равно нулю в любом наблюдении:

$$M(\varepsilon_i) = 0; \qquad (1.3)$$

2-е условие Гаусса – Маркова: дисперсия возмущения ε_i (или зависимой переменной y_i) постоянна для любого i:

$$D(\varepsilon_i) = \sigma^2 \qquad (1.4)$$

(или $D(y_i) = \sigma^2$) – условие равноизменчивости возмущения (зависимой переменной);

3-е условие Гаусса – Маркова: значения возмущения ε_i и ε_j в любых наблюдениях (или переменные y_i и y_j) не коррелируют между собой:

$$\mathrm{Cov}(\varepsilon_i \varepsilon_j) = 0,(i \neq j).$$

Это условие с учетом того, что $M(\varepsilon_i) = M(\varepsilon_j) = 0$ принимает вид

$$M(\varepsilon_i \varepsilon_j) = 0,\ (i \neq j); \qquad (1.5)$$

4-е условие Гаусса – Маркова: случайная компонента должна быть распределена независимо от объясняющих переменных x_i

$$\mathrm{Cov}(x_i, \varepsilon_i) = M(x_i, \varepsilon_i) = 0. \quad (1.6)$$

Следует отметить, что последнее условие заведомо выполняется, если объясняющие переменные x_i считаются детерминированными величинами.

Выполнение 4-го условия Гаусса – Маркова обеспечивает несмещенность оценки параметра b.

Выполнение 1-го и 4-го условий Гаусса – Маркова обеспечивает несмещенность оценки параметра a.

Нарушение одного из условий Гаусса – Маркова приводит к нарушению эффективности оценок, т. е. в классе несмещенных оценок можно найти такие, которые имеют меньшую дисперсию.

В регрессионном анализе обычно делается еще одна предпосылка о нормальности распределения случайного члена, что позволяет выполнить количественную оценку точности полученных оценок параметров.

После построения модели необходимо вычислить значения остатков ε_i и проверить выполнение условий Гаусса – Маркова, так как их нарушение снижает качество модели. Если условия нарушаются, то следует модернизировать модель соответствующим образом.

При выполнении условий Гаусса – Маркова модель (1.2) называется классической нормальной регрессионной моделью.

Линейная регрессия сводится к нахождению уравнения вида $y_x = a + bx$. Такое уравнение позволяет по заданным значениям фактора x иметь теоретические значения результативного признака, подставляя в него фактические значения фактора x. Построение линейной регрессии сводится к оценке ее параметров – a и b. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). Он позволяет получить такие оценки параметров a и b, при которых сумма квадратов отклонений фактических значений результативного признака (y) от расчетных (теоретических (y_x)) минимальна:

$$S = \sum_i \left(y_i - y_{x_i}\right)^2 \to \min. \quad (1.7)$$

Чтобы найти минимум функции, нужно вычислить частные производные по каждому из параметров a и b и приравнять их к нулю:

$$S = \sum (y_i - y_x)^2 = \sum (y - a - bx)^2,$$
$$\begin{aligned} \frac{dS}{da} &= -2\sum y + 2na + 2b\sum x = 0, \\ \frac{dS}{db} &= -2\sum yx + 2a\sum x + 2b\sum x^2 = 0. \end{aligned} \tag{1.8}$$

Получаем систему нормальных уравнений

$$\begin{cases} na + b\sum x = \sum y \\ a\sum x + b\sum x^2 = \sum yx. \end{cases} \tag{1.9}$$

Решая систему, можно найти искомые оценки параметров a и b. Однако при построении линейной регрессионной модели можно воспользоваться готовыми формулами:

$$a = \bar{y} - b\bar{x}; \tag{1.10}$$

$$b = \frac{\text{cov}(x,y)}{\sigma_x^2}, \tag{1.11}$$

где $\text{cov}(x, y)$ – ковариация признаков, $\text{cov}(x, y) = \overline{y \cdot x} - \bar{y} \cdot \bar{x}$; σ_{x}^2 – дисперсия признака x, $\sigma_x^2 = \overline{x^2} - \bar{x}^2$.

Параметр b называется коэффициентом регрессии. Его величина показывает среднее изменение результата с изменением фактора на одну единицу. Возможность четкой экономической интерпретации коэффициента регрессии сделала линейное уравнение регрессии распространенным в эконометрических расчетах.

Уравнения регрессии дополняются показателями тесноты связи. При использовании линейной регрессии такой показатель – коэффициент линейной корреляции

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\text{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\sigma_y}. \tag{1.12}$$

Линейный коэффициент корреляции находится в пределах $-1 \le r_{xy} \le 1$. Если x и y независимы, то коэффициент корреляции равен нулю. Если между переменными существует положительная зависимость, то данный коэффициент будет положительным. Если существует строгая положительная зависимость, коэффициент корреляции примет максимальное значение, равное единице. Аналогичным образом при отрицательной зависимости коэффициент корреляции будет отрицательным с минимальным значением –1.

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции, называемый коэффициентом детерминации. Он характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака

$$r_{yx}^2 = \frac{\sigma_{y_{объясн}}^2}{\sigma_{y_{общ}}^2} \qquad (1.13)$$

Соответственно величина $(1 - r^2)$ характеризует долю дисперсии y, вызванную влиянием остальных не учтенных в модели факторов. Если в выборке отсутствует видимая связь между x и y, то коэффициент будет близок к нулю. При прочих равных условиях желательно, чтобы коэффициент r^2 был как можно больше.

Оценка значимости уравнения и параметров регрессии. После того как найдено уравнение линейной регрессии, проводится оценка значимости как уравнения в целом, так и отдельных его параметров. Проверить значимость уравнения регрессии означает установить соответствие математической модели, выражающей зависимость между переменными и экспериментальным данным, а также проверить достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной.

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

Непосредственному расчету F-критерия предшествует анализ дисперсии. Центральное место в нем занимает разложение общей суммы квадратов отклонений переменной y от среднего значения $\bar{y}$ на две части – «объясненную» и «необъясненную».

Любая сумма квадратов отклонений связана с числом степеней свободы, то есть с числом свободы независимого варьирования признака.

Число степеней свободы связано с числом единиц совокупности (n) и с числом определяемых по ней констант. Имеем два равенства:

$$\sum(y - \bar{y})^2 = \sum(\hat{y}_x - \bar{y})^2 + \sum(y - \hat{y}_x)^2, \qquad (1.14)$$

$$n - 1 = 1 + (n - 2) \qquad (1.15)$$

Разделив каждую сумму квадратов на соответствующее ей число степеней свободы, получим средний квадрат отклонений, или, что то же самое, дисперсию на одну степень свободы.

$$D^{(1)}{}_{общ} = \frac{\sum(y-\bar{y})^2}{n-1}, \qquad (1.16)$$

$$D^{(1)}{}_{факт} = \frac{\sum(\hat{y}_x-\bar{y})^2}{1}, \qquad (1.17)$$

$$D^{(1)}{}_{ост} = \frac{\sum(y-\hat{y}_x)^2}{n-2}. \qquad (1.18)$$

Определение дисперсии на одну степень свободы приводит различные виды дисперсии к сравниваемому виду.

При расчете объясненной или факторной суммы квадратов $\sum(\hat{y}_x - \bar{y})^2$ используются теоретические (расчетные) значения результативного признака, найденные по линии регрессии. В линейной регрессии справедлива формула

$$D_{факт} = \sum(\hat{y}_x - \bar{y})^2 = b^2 \sum(x - \bar{x})^2. \qquad (1.19)$$

Определяя отношение факторной и остаточной дисперсии в расчете на одну степень свободы, получим величину F-отношения (F-критерий):

$$F = \frac{D^{(1)}_{факт}}{D^{(1)}_{ост}}, \qquad (1.20)$$

где F – критерий для проверки нулевой гипотезы H_0: $D_{факт} = D_{ост}$.

Английским статистиком Снедекором разработаны таблицы критических значений F-отношений при различных уровнях существенности нулевой гипотезы и различном числе степеней свободы. Вычисленное значение F-отношения признается достоверным (отличным от единицы), если оно больше табличного. В этом случае нулевая гипотеза об отсутствии связи признаков отклоняется и делается вывод о существенной связи. Если же величина окажется меньше табличной, то вероятность наблюдать имеющуюся выборку при верности нулевой гипотезы выше заданного уровня, и она не может быть отклонена без риска сделать неправильный вывод о наличии связи. Уравнение регрессии считается статистически незначимым.

В линейной регрессии оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров определяется его стандартная ошибка. Стандартная ошибка коэффициента регрессии:

$$m_b = \sqrt{\frac{\sum(y-\hat{y}_x)^2/(n-2)}{\sum(x-\bar{x})^2}} = \sqrt{\frac{D^{(1)}_{ост}}{\sum(x-\bar{x})^2}}, \qquad (1.21)$$

где $D^{(1)}_{ост}$– остаточная дисперсия на одну степень свободы.

Для оценки существенности коэффициента регрессии его величина сравнивается с его стандартной ошибкой, т. е. определяется фактическое значение t-критерия Стьюдента:

$$t_b = \frac{b}{m_b}, \tag{1.22}$$

которое затем сравнивается с табличным значением при определенном уровне значимости α и числе степеней свободы *n*-2.

Стандартная ошибка параметра *a* определяется:

$$m_a = \sqrt{\frac{\sum(y-\hat{y}_x)^2}{n-2} * \frac{\sum x^2}{n\sum(x-\bar{x})^2}} = \sqrt{D_{ост}^{(1)} \frac{\sum x^2}{n\sum(x-\bar{x})^2}}. \tag{1.23}$$

Процедура оценивания такая же: вычисляется *t*-критерий $t_a = \frac{a}{m_a}$, его величина сравнивается с табличным значением.

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

$$m_r = \sqrt{\frac{1-r^2}{n-2}}. \tag{1.24}$$

Фактическое значение *t*-критерия Стьюдента определяется:

$$t_r = \frac{r}{\sqrt{1-r^2}} * \sqrt{n-2}. \tag{1.25}$$

Коэффициент корреляции значим на уровне α (иначе – гипотеза H_0 о равенстве генерального коэффициента корреляции нулю отвергается), если выполняется условие

$$|t_r| = \frac{|r|}{\sqrt{1-r^2}} * \sqrt{n-2} > t_{1-\alpha;n-2}, \tag{1.26}$$

где $t_{1-\alpha;n-2}$ – табличное значение *t*-критерия Стьюдента, определенное на уровне значимости α при числе степеней свободы $n-2$.

Интервальная оценка уравнения регрессии. В прогнозных расчетах по уравнению регрессии определяется предсказываемое (y_p) значение как точечный прогноз $\hat{y}_x$ при $x_p = x_k$, т. е. путем подстановки в уравнение регрессии соответствующего значения *x*. Однако точечный прогноз не реален. Поэтому он дополняется расчетом стандартной ошибки ($m_{\hat{y}x}$) и, соответственно, интервальной оценкой прогнозного значения (y)

$$\hat{y}_x - t_{1-\alpha;n-2} m_{\hat{y}x} \le y^* \le \hat{y}_x + t_{1-\alpha;n-2} m_{\hat{y}x}, \tag{1.27}$$

где $t_{1-\alpha;n-2}$ – значения *t*-критерия Стьюдента при уровне доверительной вероятности (1 – *α*) и числе степеней свободы (*n* – 2).

Формула расчета стандартной ошибки:

$$m_{yx}^2 = D_{ост}^{(1)}\left(\frac{1}{n} + \frac{(x_k-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right), \tag{1.28}$$

соответственно

$$m_{yx} = \sqrt{D_{ост}^{(1)}(\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum(x_i - \bar{x})^2})}. \qquad (1.29)$$

Из формул (1.27) и (1.29) видно, что величина (длина) доверительного интервала зависит от значения объясняющей переменной x: при $x = \bar{x}$ минимальна, а по мере удаления x от $\bar{x}$ величина доверительного интервала увеличивается. Таким образом, прогноз значений (определение неизвестных значений) зависимой переменной y по уравнению регрессии оправдан, если значение объясняющей переменной x не выходит за диапазон ее значений по выборке (причем тем более точный, чем ближе x к $\bar{x}$).

Кроме того, для расчета доверительного интервала можно определить предельную ошибку для каждого показателя

$$\Delta a = t_{табл} m_a, \qquad (1.30)$$

$$\Delta b = t_{табл} m_b. \qquad (1.31)$$

Формулы для расчета доверительных интервалов в этом случае имеют вид

$$\gamma_a = a \pm \Delta a; \quad \gamma_{a\min} = a - \Delta a; \quad \gamma_{a\max} = a + \Delta a; \qquad (1.32)$$

$$\gamma_b = b \pm \Delta b; \quad \gamma_{b\min} = b - \Delta b; \quad \gamma_{b\max} = b = \Delta b. \qquad (1.33)$$

Средняя ошибка аппроксимации. Величина отклонений фактических и расчетных значений результативного признака $(y - y_x)$ по каждому наблюдению представляет собой ошибку аппроксимации. Поскольку $(y - y_x)$ может быть как величиной положительной, так и отрицательной, то ошибки аппроксимации для каждого наблюдения принято определять в процентах по модулю.

Чтобы иметь общее суждение о качестве модели из относительных отклонений по каждому наблюдению, определяют среднюю ошибку аппроксимации как среднюю арифметическую простую:

$$A = \frac{1}{n}\sum\left|\frac{(y - y_x)}{y}\right| * 100 \text{ или } A = \frac{100}{\bar{y}}\sqrt{\frac{\sum(y - y_x)^2}{n}}. \qquad (1.34)$$

При выборе вида модели при прочих равных условиях предпочтение отдается той функциональной зависимости, для которой ошибка аппроксимации будет меньше.

Задача 1

Выявить, в какой степени и как различные факторы воздействуют на уровень коррумпированности.

В качестве исходных данных для расчета модели использовался некий набор величин, представленный в табл. 1.1, где У – уровень коррупции (индекс восприятия коррупции в России 0-100 баллов); X1 – уровень экономического благополучия государства (ВВП на душу населения; доллар США); X2 – уровень з/п должностных лиц (средняя заработная плата в месяц; тысячи рублей на человека);

Таблица 1.1

Исходные данные для расчета регрессионной модели

n	Уровень экономического благополучия государства	Уровень з/п должностных лиц	Уровень коррупции
1	22049	19,384	25
2	23958	21,859	23
3	25068	24,968	21
4	24958	23,657	22
5	25472	24,685	21
6	22095	20,474	24
7	19374	17,947	28
8	19857	17,004	28
9	20475	18,944	27
10	16639	15,047	29
11	16394	15,593	29
12	22958	20,848	24
13	21947	18,967	26
14	25847	24,487	21
15	20384	18,473	27
16	16839	15,686	29
17	21958	19,048	25
18	20748	18,857	27
19	23950	21,947	23
20	20363	18,389	27
21	19287	17,494	28
22	16392	15,696	29

Построим уравнение регрессии и проведем анализ влияния уровня экономического благополучия государства на уровень коррупции. Все расчеты будем проводить в Microsoft Excel.

Проведем промежуточные расчеты для построения уравнения регрессии по уровню экономического благополучия государства и уровню коррупции (рис. 1.2).

n	x	y	xy	y^2	x^2	$(x-\overline{x})^2$	$(y-\overline{y})^2$
1	22049	25	551225	625	486158401	674339,58	0,349
2	23958	23	551034	529	573985764	7453892,8	6,713
3	25068	21	526428	441	628404624	14746996	21,08
4	24958	22	549076	484	622901764	13914256	12,89
5	25472	21	534912	441	648822784	18013079	21,08
6	22095	24	530280	576	488189025	752004,31	2,531
7	19374	28	542472	784	375351876	3436641,9	5,804
8	19857	28	555996	784	394300449	1879142,5	5,804
9	20475	27	552825	729	419225625	566735,21	1,986
10	16639	29	482531	841	276856321	21057252	11,62
11	16394	29	475426	841	268763236	23365798	11,62
12	22958	24	550992	576	527069764	2993529,1	2,531
13	21947	26	570622	676	481670809	517222,49	0,167
14	25847	21	542787	441	668067409	21336841	21,08
15	20384	27	550368	729	415507456	712029,12	1,986
16	16839	29	488331	841	283551921	19261725	11,62
17	21958	25	548950	625	482153764	533165,49	0,349
18	20748	27	560196	729	430479504	230225,49	1,986
19	23950	23	550850	529	573602500	7410273,9	6,713
20	20363	27	549801	729	414651769	747910,49	1,986
21	19287	28	540036	784	371988369	3766775,2	5,804
22	16392	29	475368	841	268697664	23385137	11,62
сумма	**467012**	**563**	**1,2E+07**	**14575**	**1,01E+10**	**1,87E+08**	**167**
средн	**21227,818**	**25,5909**	**535478**	**662,5**	**459109127**	**8488862**	**7,61**

Рис. 1.2. Таблица промежуточных расчетов

Определим количество факторов (m) и объем выборки (n):

m = 1, так как в модели 1 фактор x;

n = 22, так как выборка состоит из 22 наблюдений.

1. Вычислить коэффициент корреляции и сделать вывод о тесноте и направлении связи (рис. 1.3).

1. Вычислить коэффициент корреляции и сделать вывод о тесноте и направлении связи.

Показатель парной линейной корреляции характеризует тесноту связи рассматриваемогофактора с результатом изменяется в границах от -1 до 1, чем ближе к ±1 тем связь теснее, чам ближе к 0 тем связь слабее

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}, \qquad \sigma x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\sigma_y}, \qquad \sigma y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

r	-0,9660	связь между x и y обратная, сильная	по первой формуле
r	-0,9660	связь между x и y обратная, сильная	по второй формуле
σx	2913,5652	есть корреляция близка к 1 или -1 связь сильная	
σy	2,7578	есть корреляция близка к 0 связь слабая	
cov	-7761,6198	есть корреляция +: при росте x, y ростет; при убывании x, y убывает	
		есть корреляция -: при росте x, y убывает; при убывании x, y ростет	

Рис. 1.3. Определение коэффициента корреляции

Так как связь фактора и результата сильная, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза. Делаем вывод, что уровень экономического благополучия государства влияет на уровень коррупции.

2. На уровне значимости α = 0,05 проверить гипотезу о значимости коэффициента корреляции (рис. 1.4).

2. На уровне значимости α = 0,05 проверить гипотезу о значимости коэффициента корреляции.

Используем t-статистику Стьюдента $t = r_{xy} * \sqrt{\frac{n-2}{1-r_{xy}^2}}$,

Если фактическое значение > критического (табличного), параметр значим

Если фактическое значение < критического (табличного), параметр не значим

Оценка производится по модулю.

t критическое определяется из таблицы Стьюдента

Номер строки - (n-m-1)

Номер столбца - уровень значимости (обычно 0,05)

tтабл	2,086	из таблицы Стьюдента 20 строка, столбец 0,05		
t	-16,7038	фактическое		
t	>	tкрит		делаем вывод о значимости коэффициента корреляции.

Рис. 1.4. Оценка значимости коэффициента корреляции

Так как коэффициент корреляции значим, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

3. Составить уравнение парной регрессии $y = a + b*x$ (рис. 1.5).

3. Составить уравнение парной регрессии y = a + b*x

$$a = \overline{y} - b\overline{x},$$

$$b = \frac{\mathrm{cov}(x,y)}{\sigma_x^2} = \frac{\overline{yx} - \overline{y}\cdot\overline{x}}{\overline{x^2} - (\overline{x})^2}.$$

b	-0,0009
a	45,0001

ŷ=45-0,0009*x   при изменении x на 1 еденицу, у изменится на -0,0009 едениц

Рис. 1.5. Построение уравнения линейной парной регрессии

Рассчитаем прогнозное значение $\hat{y}$ (рис. 1.6).

n	x	y	$\hat{y}$	$(\hat{y}-y)^2$	$(\hat{y}-\overline{y})^2$
1	22049	25	24,84	0,025575	0,564
2	23958	23	23,09	0,008953	6,231
3	25068	21	22,08	1,165788	12,33
4	24958	22	22,18	0,032505	11,63
5	25472	21	21,71	0,504565	15,06
6	22095	24	24,8	0,636834	0,629
7	19374	28	27,29	0,509924	2,873
8	19857	28	26,84	1,335668	1,571
9	20475	27	26,28	0,519505	0,474
10	16639	29	29,79	0,618743	17,6
11	16394	29	30,01	1,021339	19,53
12	22958	24	24,01	8,01E-05	2,503
13	21947	26	24,93	1,137764	0,432
14	25847	21	21,37	0,135022	17,84
15	20384	27	26,36	0,406486	0,595
16	16839	29	29,6	0,364498	16,1
17	21958	25	24,92	0,005886	0,446
18	20748	27	26,03	0,941635	0,192
19	23950	23	23,1	0,010391	6,195
20	20363	27	26,38	0,382371	0,625
21	19287	28	27,37	0,402645	3,149
22	16392	29	30,01	1,025038	19,55
сумма	**467012**	**563**	**563**	**11,1912**	**156**
средн	**21227,818**	**25,5909**	**25,6**	**0,50869**	**7,1**

Рис. 1.6. Определение значения $\hat{y}$

4. Нанести данные на чертеж и изобразить прямую регрессии (рис. 1.7).

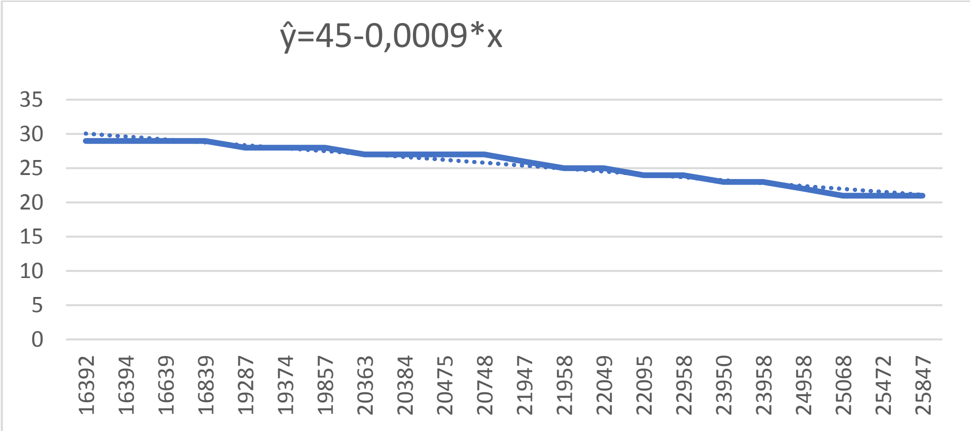


Рис. 1.7. График линейной парной регрессии

5. С помощью коэффициента детерминации R^2 оценить качество построенной модели (рис. 1.8).

5. С помощью коэффициента детерминации R^2 оценить качество построенной модели.

Характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(y_i - \overline{Y})^2} = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(y_i)^2 - n(\overline{Y})^2},$$

R^2	0,9331	
R^2	0,9331	93,3114 %

Построенное уравнение регрессии на 93,31% объясняет зависимость переменной y от переменной x.

Рис. 1.8. Определение коэффициента детерминации

Так как детерминация 93,31%, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

6. Расчет остаточной дисперсии (рис. 1.9).

6. Рассчет остаточной дисперии

Качество подгонки определяется через дисперсию

$$D_{ост} = \frac{\Sigma(y - \hat{y})^2}{n - 2}$$

Dост 0,5596

Чем меньше Dост, тем лучше модель

Рис. 1.9. Определение остаточной дисперсии

На основании только одной остаточной дисперсии нельзя сделать вывод об уравнении регрессии.

7. Определение значимости параметров уравнения с помощью t-критерия Стьюдента (рис. 1.10).

7. Определение значимости параметров уравнения с помощью t критерия Стьюдента

Если фактическое значение > критического (табличного), параметр значим
Если фактическое значение < критического (табличного), параметр не значим

t критическое определяется из таблицы Стьюдента
Номер строки - (n-m-1)
Номер столбца - уровень значимости (обычно 0,05)

tтабл 2,086 из таблицы Стьюдента 20 строка, столбец 0,05

$$m_b = \sqrt{\frac{\sum (y-\hat{y}_x)^2 / (n-2)}{\sum (x-\bar{x})^2}} = \sqrt{\frac{D_{ост}^{(1)}}{\sum (x-\bar{x})^2}}$$

$$m_a = \sqrt{\frac{\sum (y-\hat{y}_x)^2}{n-2} * \frac{\sum x^2}{n\sum (x-\bar{x})^2}} = \sqrt{D_{ост}^{(1)} \frac{\sum x^2}{n\sum (x-\bar{x})^2}}$$

$$t_b = \frac{b}{m_b} \qquad t_a = \frac{a}{m_a}$$

оценивается по модулю

mb	0,0001	
ma	1,1729	
ta	38,3679	> tкрит, значит параметр a значим
tb	-16,7038	> tкрит, значит параметр b значим

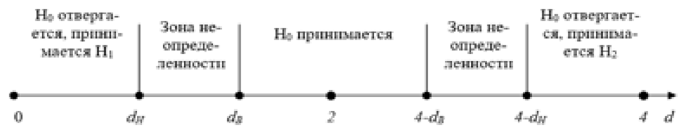
Рис. 1.10. Определение значимости параметров

Так как оба параметра уравнения регрессии значимы, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

8. Для определения наличия автокорреляции остатков между соседними членами ряда используется критерий (тест) Дарбина – Уотсона (рис. 1.11).

8. Для определения наличия автокорреляции остатков между соседними членами ряда используется критерий (тест) Дарбина-Уотсона

Остатки ε должны быть случайными.
Однако иногда каждое следующее значение остатков зависит от предшествующих.
В этом случае имеет место автокорреляция в остатках.



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

d 1,746 фактическое значение критерия (теста) Дарбина-Уотсона
определяется из столбца остатки на листе "Регрессия нет мультик"
с помощью "Анализ данных" → "Регрессия"
созданные регрессионный анализ назвать "Регрессия три х"

гипотеза H0 об отсутствии автокорреляции в остатках
гипотеза H1 – гипотеза о положительной автокорреляции в остатках
гипотеза H2 – гипотеза об отрицательной автокорреляции в остатках.

d табличное определяется из таблицы Дарбина-Уотсона
Столбец - m
Строка - n

dн	1,24	из таблицы Дарбина-Уотсона 22 строка, столбец 1	4-dн	2,76
dв	1,43	из таблицы Дарбина-Уотсона 22 строка, столбец 1	4-dв	2,57

Поскольку фактическое значение d=1,746 попадает в промежуток от 1,43 до 2,57, то принимаем гипотезу H0 об отсутствии автокорреляции в остатках. Значит, с уверенностью (100-α) в 95 % можно утверждать, что нет автокорреляция в остатках.

Модель можно считать адекватной, если автокорреляции в остатках нет

Рис. 1.11. Критерий Дарбина – Уотсона

Так как нет автокорреляции в остатках, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза (рис. 1.12).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,965978347
R-квадрат	0,933114167
Нормированный R-квадрат	0,929769876
Стандартная ошибка	0,748037963
Наблюдения	22

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	1	156,1269659	156,1269659	279,016985	3,25813E-13
Остаток	20	11,19121589	0,559560794		
Итого	21	167,3181818			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	45,00013573	1,172858216	38,36792472	3,3189E-20	42,55359636	47,4466751	42,55359636	47,4466751
Переменная X 1	-0,00091433	5,47378E-05	-16,70380152	3,2581E-13	-0,00102851	-0,00080015	-0,001028511	-0,000800149

ВЫВОД ОСТАТКА

Наблюдение	Предсказанное Y	Остатки	$(\varepsilon_t - \varepsilon_{t-1})^2$	ε^2
1	24,84007809	0,159921912		0,02557502
2	23,09462251	-0,09462251	0,064792863	0,00895342
3	22,07971644	-1,079716438	0,970410047	1,16578759
4	22,18029272	-0,180292715	0,808963033	0,03250546
5	21,7103272	-0,710327201	0,280936556	0,50456473
6	24,79801892	-0,798018917	0,007689837	0,63683419
7	27,28591029	0,714089712	2,286472505	0,50992412
8	26,844289	1,155711002	0,195029365	1,33566792
9	26,27923318	0,720766815	0,189176446	0,5195048
10	29,78660228	-0,786602277	2,27216158	0,61874314
11	30,01061308	-1,010613077	0,050180838	1,02133879
12	24,0089523	-0,008952304	1,003324302	8,0144E-05
13	24,93333973	1,066660273	1,156942417	1,13776414
14	21,36745353	-0,367453528	2,056682394	0,13502209
15	26,3624372	0,637562804	1,010057827	0,40648633
16	29,60373632	-0,603736318	1,54082351	0,36449754
17	24,9232821	0,076717901	0,463017944	0,00588564
18	26,02962115	0,970378849	0,798629891	0,94163511
19	23,10193715	-0,101937148	1,149861599	0,01039118
20	26,38163812	0,618361878	0,518830688	0,38237141
21	27,36545698	0,634543019	0,000261829	0,40264484
22	30,01244174	-1,012441736	2,712558785	1,02503827
			19,53680426	11,1912159

Рис. 1.12. Анализ данных для критерия (теста) Дарбина – Уотсона

9. Тест Голдфелда – Квандта определения гетероскедастичности остатков (рис. 1.13–1.16).

9. Тест Голфелда-Кванда определения гетероскедастичности остатков

Если для каждого значения фактора остатки имеют одинаковую дисперсию, это называется гомоскедастичной. Если это условие не соблюдается, то имеет место гетероскедастичность.

Упорядочим по x исходные данные по возрастанию, затем упорядочиваем значения y (если одинаковые x) по возрастанию

x	y		x	y	
16392	29		16392	29	
16394	29		16394	29	
16639	29		16639	29	
16839	29		16839	29	
19287	28		19287	28	верхняя регрессия
19374	28		19374	28	
19857	28		19857	28	
20363	27		20363	27	
20384	27		20384	27	
20475	27		20475	27	удаляем
20748	27		20748	27	удаляем
21947	26		21947	26	удаляем
21958	25		21958	25	удаляем
22049	25		22049	25	
22095	24		22095	24	
22958	24		22958	24	
23950	23		23950	23	
23958	23		23958	23	верхняя регрессия
24958	22		24958	22	
25068	21		25068	21	
25472	21		25472	21	
25847	21		25847	21	

Удаляем n/6 наблюдений 3,6666667 удаляем 4 средних наблюдений

строим 2 регрессионные модели:

9 верхним данным

9 нижним данным

Данные → Анализ данных → Регрессия → Выделяем данные по верхней регрессии и выводим на этот лист на ячейку A224 (не забыть галочку на остатки)

Данные → Анализ данных → Регрессия → Выделяем данные по нижней регрессии и выводим на этот лист на ячейку A263 (не забыть галочку на остатки)

Верхняя регрессия

Рис. 1.13. Тест Голдфелда – Квандта

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множеств	0,951474702
R-квадрат	0,905304108
Нормиро	0,891776124
Стандартн	0,274144979
Наблюде	9

Дисперсионный анализ

Страница 9

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	5,02946727	5,02946727	66,92084	7,90236E-05
Остаток	7	0,52608829	0,07515547		
Итого	8	5,55555556			

	Коэффициенты	*ндартная ошиб*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*ижние 95,0*	*рхние 95,0%*
Y-пересеч	36,43210031	1,00774106	36,1522437	3,22E-09	34,04917136	38,81502926	34,0492	38,815
Перемен	-0,00044638	5,4566E-05	-8,18051596	7,9E-05	-0,000575409	-0,000317352	-0,0006	-0,0003

ВЫВОД ОСТАТКА

аблюдени	*Предсказанное Y*	*Остатки*	
1	29,11503264	-0,1150326	0,01323251
2	29,11413988	-0,1141399	0,01302791
3	29,00477668	-0,0047767	2,2817E-05
4	28,9155006	0,0844994	0,00714015
5	27,82276135	0,17723865	0,03141354
6	27,78392626	0,21607374	0,04668786
7	27,56832452	0,43167548	0,18634372
8	27,34245603	-0,342456	0,11727613
9	27,33308204	-0,333082	0,11094365
			0,52608829

Рис. 1.14. Анализ данных для теста Голдфелда – Квандта (верхняя регрессия)

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множеств	0,965615278
R-квадрат	0,932412865
Нормиро	0,92275756
Стандарт	0,416887863
Наблюде	9

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	1	16,7834316	16,7834316	96,57001	2,39998E-05
Остаток	7	1,21656843	0,17379549		
Итого	8	18			

	Коэффициенты	ндартная оши	t-статистика	P-Значение	Нижние 95%	Верхние 95%	ижние 95,0	рхние 95,0%
Y-пересеч	47,17450485	2,49779631	18,8864499	2,9E-07	41,26815511	53,08085459	41,2682	53,0809
Перемен	-0,001019484	0,00010374	-9,82700393	2,4E-05	-0,001264798	-0,000774171	-0,0013	-0,0008

ВЫВОД ОСТАТКА

аблюдени	Предсказанное Y	Остатки	
1	24,69589369	0,30410631	0,09248065
2	24,64899741	-0,6489974	0,42119764
3	23,76918239	0,23081761	0,05327677
4	22,75785388	0,24214612	0,05863474
5	22,74969801	0,25030199	0,06265109
6	21,73021362	0,26978638	0,07278469
7	21,61807034	-0,6180703	0,38201095
8	21,20619865	-0,2061987	0,04251788
9	20,82389201	0,17610799	0,03101403
			1,21656843

Рис. 1.15. Анализ данных для теста Голдфелда – Квандта (нижняя регрессия)

По верхней и нижней регрессии рассчитываем столбец "Остаток²" и находим сумму				
S1	0,53	сумма остатков верхней регрессии		
S2	1,22	сумма остатков нижней регрессии		
$F = \dfrac{\max(S1, S2)}{\min(S1, S2)}$		F	2,3125	фактическое значение теста Голфелда-Кванда
F табличное определяется из таблицы Фишера				
n	9	так как регрессии строились исходя из данных сокращенных		
m	1			
Номер строки - (n-m-1)				
Номер столбца - m				
Fкрит	5,59	из таблицы Фишера 7 строка, 1 столбец		
Если фактическое значение < критического, то нет наличия гетероскедастичности				
Если фактическое значение > критического, то есть гетероскедастичность				
F	<	Fкрит	нет наличия гетероскедастичности остатков	
Модель можно считать адекватной, если гетероскедастичности нет				

Рис. 1.16. Тест Голдфелда – Квандта

Так как нет наличия гетероскедастичности, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

10. Оценка значимости уравнения с помощью *F*-критерия Фишера (рис. 1.17).

10. Оценка занчимости уравнения с помощью F-критерия Фишера

Оценка значимости уравнения множественной регрессии в целом осуществляется путем проверки основной гипотезы H0 о статистической незначимости уравнения регрессии.

Находим F табличное из таблицы Фишера.
F табличное определяется из таблицы Фишера
Номер строки - (n-m-1)
Номер столбца - m

m	1	столбец
n	22	
n-m-1	20	строка

Fкрит	4,35	из таблицы Фишера 20 строка, 1 столбец
Dост	0,5596	$Dфакт = \frac{\sum(\hat{y}-\bar{y})^2}{1}$
Dфакт	156,126966	$F = \frac{Dфакт}{Dост}$
F	279,016985	

Если фактическое значение > критического (табличного), то гипотеза H0 о статистической незначимости уравнения регрессии отвергается и принимается гипотеза H1 о статистичекой значимости уравнения регрессии.
Если фактическое значение < критического (табличного), то гипотеза H0 о статистической незначимости уравнения регрессии принимается

F	>	Fкрит	делаем вывод о значимости уравнения регрессии.

Если уравнение регрессии не значимо, то нет смысла в прогнозе

Рис. 1.17. Критерий Фишера

Так как уравнение регрессии значимо, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

11. При уровне значимости α = 0,05 построить доверительные интервалы для оценки параметров регрессии *a*, *b* и сделать вывод об их значимости (рис. 1.18).

11. При уровне значимости α = 0,05 построить доверительные интервалы для оценки параметров регрессии a, b и сделать вывод об их значимости.

$$S(b) = \frac{Dост}{\sqrt{\sum(x-\bar{x})^2}} \qquad S(a) = \frac{Dост}{\sqrt{\sum(x-\bar{x})^2}}\sqrt{x^2}$$

		$b - t_{крит} * S(b); b + t_{крит} * S(b)$		$a - t_{крит} * S(a); a + t_{крит} * S(a)$	
S(b)	0,0001				
S(a)	1,1729	-0,0010	-0,0008	42,5536	47,4467

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

Интервал для параметра b не включает нулевое значение, делаем вывод о значимом отличии от нуля b.
Интервал для параметра a не включает нулевое значение, делаем вывод о значимом отличии от нуля a.

Рис. 1.18. Интервальная оценка

Так как оба интервала не содержат нулевых значений, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

Проведем промежуточные расчеты для построения уравнения регрессии по уровню з/п должностных лиц (средняя заработная плата в месяц; тысячи рублей на человека) и уровню коррупции (рис. 1.19).

n	x	y	xy	y^2	x^2	$(x-\overline{x})^2$	$(y-\overline{y})^2$
1	19,384	25	484,6	625,0	375,7	0,0	0,3
2	21,859	23	502,8	529,0	477,8	5,5	6,7
3	24,968	21	524,3	441,0	623,4	29,7	21,1
4	23,657	22	520,5	484,0	559,7	17,1	12,9
5	24,685	21	518,4	441,0	609,3	26,7	21,1
6	20,474	24	491,4	576,0	419,2	0,9	2,5
7	17,947	28	502,5	784,0	322,1	2,5	5,8
8	17,004	28	476,1	784,0	289,1	6,3	5,8
9	18,944	27	511,5	729,0	358,9	0,3	2,0
10	15,047	29	436,4	841,0	226,4	20,0	11,6
11	15,593	29	452,2	841,0	243,1	15,4	11,6
12	20,848	24	500,4	576,0	434,6	1,8	2,5
13	18,967	26	493,1	676,0	359,7	0,3	0,2
14	24,487	21	514,2	441,0	599,6	24,7	21,1
15	18,473	27	498,8	729,0	341,3	1,1	2,0
16	15,686	29	454,9	841,0	246,1	14,7	11,6
17	19,048	25	476,2	625,0	362,8	0,2	0,3
18	18,857	27	509,1	729,0	355,6	0,4	2,0
19	21,947	23	504,8	529,0	481,7	5,9	6,7
20	18,389	27	496,5	729,0	338,2	1,3	2,0
21	17,494	28	489,8	784,0	306,0	4,1	5,8
22	15,696	29	455,2	841,0	246,4	14,6	11,6
сумма	**429,454**	**563**	**10813,6**	**14575,0**	**8576,7**	**193,5**	**167,3**
средн	**19,52064**	**25,5909**	**491,5**	**662,5**	**389,9**	**8,8**	**7,6**

Рис. 1.19. Таблица промежуточных расчетов

Определим количество факторов (m) и объем выборки (n):
$m = 1$, так как в модели 1 фактор x;
$n = 22$, так как выборка состоит из 22 наблюдений.

1. Вычислить коэффициент корреляции и сделать вывод о тесноте и направлении связи.

1. Вычислить коэффициент корреляции и сделать вывод о тесноте и направлении связи.

Показатель парной линейной корреляции характеризует тесноту связи рассматриваемогофактора с результатом изменяется в границах от -1 до 1, чем ближе к ±1 тем связь теснее, чам ближе к 0 тем связь слабее

$$r_{xy} = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\sqrt{n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}}, \qquad σx=\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$$

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = \frac{\operatorname{cov}(x,y)}{\sigma_x\sigma_y} = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\sigma_x\sigma_y}, \qquad σy=\sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

r	-0,9809	связь между x и y обратная, сильная	по первой формуле
r	-0,9809	связь между x и y обратная, сильная	по второй формуле
σx	2,9660	есть корреляция близка к 1 или -1 связь сильная	
σy	2,7578	есть корреляция близка к 0 связь слабая	
cov	-8,0235	есть корреляция +: при росте x, y ростет; при убывании x, y убывает	
		есть корреляция -: при росте x, y убывает; при убывании x, y ростет	

Рис. 1.20. Определение коэффициента корреляции

Так как связь фактора и результата сильная, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза. Делаем вывод, что уровень з/п должностных лиц влияет на уровень коррупции.

2. На уровне значимости α = 0,05 проверить гипотезу о значимости коэффициента корреляции (рис. 1.21).

2. На уровне значимости α = 0,05 проверить гипотезу о значимости коэффициента корреляции.

Используем t-статистику Стьюдента $t = r_{xy} * \sqrt{\frac{n-2}{1-r_{xy}^2}},$

Если фактическое значение > критического (табличного), параметр значим

Если фактическое значение < критического (табличного), параметр не значим

Оценка производится по модулю.

t критическое определяется из таблицы Стьюдента

Номер строки - (n-m-1)

Номер столбца - уровень значимости (обычно 0,05)

tтабл	2,086	из таблицы Стьюдента 20 строка, столбец 0,05	
t	-22,5702	фактическое	
t	>	tкрит	делаем вывод о значимости коэффициента корреляции.

Рис. 1.21. Оценка значимости коэффициента корреляции

Так как коэффициент корреляции значим, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

3. Составить уравнение парной регрессии $y = a + b*x$ (рис. 1.22).

3. Составить уравнение парной регрессии y = a + b*x

$$a = \overline{y} - b\overline{x},$$

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

b	-0,9121
a	43,3952

ŷ=43,4-0,91*x при изменении x на 1 еденицу, y изменится на -0,91 едениц

Рис. 1.22. Построение уравнения линейной парной регрессии

Рассчитаем прогнозное значение $\hat{y}$ (рис. 1.23).

n	$\hat{y}$	$(\hat{y}-y)^2$	$(\hat{y}-\overline{y})^2$
1	25,7	0,5	0,0
2	23,5	0,2	4,5
3	20,6	0,1	24,7
4	21,8	0,0	14,2
5	20,9	0,0	22,2
6	24,7	0,5	0,8
7	27,0	0,9	2,1
8	27,9	0,0	5,3
9	26,1	0,8	0,3
10	29,7	0,5	16,6
11	29,2	0,0	12,8
12	24,4	0,1	1,5
13	26,1	0,0	0,3
14	21,1	0,0	20,5
15	26,5	0,2	0,9
16	29,1	0,0	12,2
17	26,0	1,0	0,2
18	26,2	0,6	0,4
19	23,4	0,1	4,9
20	26,6	0,1	1,1
21	27,4	0,3	3,4
22	29,1	0,0	12,2
сумма	**563,0**	**6,3**	**161,0**
средн	**25,6**	**0,3**	**7,3**

Рис. 1.23. Определение значения $\hat{y}$

4. Нанести данные на чертеж и изобразить прямую регрессии (рис. 1.24).

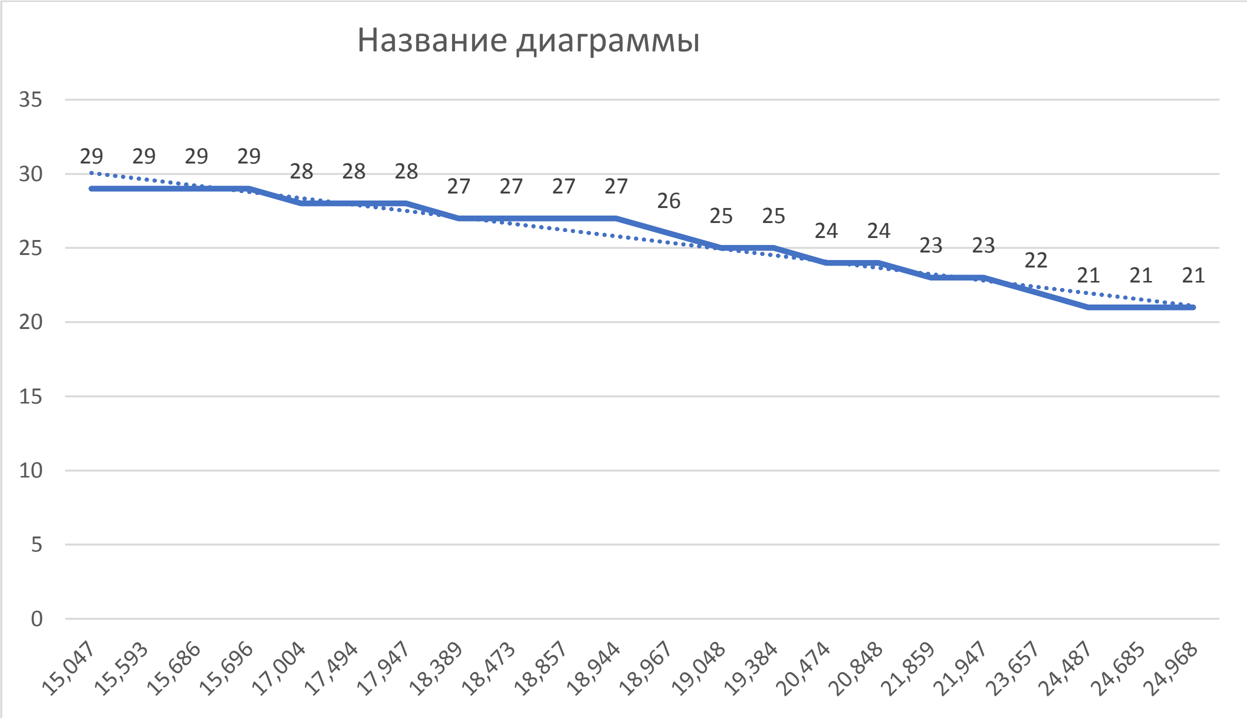


Рис. 1.24. График линейной парной регрессии

5. С помощью коэффициента детерминации R^2 оценить качество построенной модели (рис. 1.25).

5. С помощью коэффициента детерминации R² оценить качество построенной модели.

Характеризует долю дисперсии результативного признака y, объясняемую регрессией, в общей дисперсии результативного признака

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\overline{Y})^2}{\sum_{i=1}^{n}(y_i-\overline{Y})^2}=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(y_i)^2-n(\overline{Y})^2},$$

R^2	0,9622	
R^2	0,9622	96,2222 %

Построенное уравнение регрессии на 79,82% объясняет зависимость переменной y от переменной x.

Рис. 1.25. Определение коэффициента детерминации

Так как детерминация 96,22 %, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

6. Расчет остаточной дисперсии (рис. 1.26).

6. Рассчет остаточной дисперии

Качество подгонки определяется через дисперсию $\quad D_{ост} = \frac{\sum(y-\hat{y})^2}{n-2}$

Dост	0,3160

Чем меньше Dост, тем лучше модель

Рис. 1.26. Определение остаточной дисперсии

На основании только одной остаточной дисперсии нельзя сделать вывод об уравнении регрессии.

7. Определение значимости параметров уравнения с помощью *t*-критерия Стьюдента (рис. 1.27).

7. Определение значимости параметров уравнения с помощью t критерия Стьюдента

Если фактическое значение > критического (табличного), параметр значим
Если фактическое значение < критического (табличного), параметр не значим

t критическое определяется из таблицы Стьюдента
Номер строки - (n-m-1)
Номер столбца - уровень значимости (обычно 0,05)

tтабл 2,086 из таблицы Стьюдента 20 строка, столбец 0,05

$$m_b = \sqrt{\frac{\sum(y-\hat{y}_x)^2/(n-2)}{\sum(x-\bar{x})^2}} = \sqrt{\frac{D^{(1)}_{ост}}{\sum(x-\bar{x})^2}} \qquad m_a = \sqrt{\frac{\sum(y-\hat{y}_x)^2}{n-2} * \frac{\sum x^2}{n\sum(x-\bar{x})^2}} = \sqrt{D^{(1)}_{ост}\frac{\sum x^2}{n\sum(x-\bar{x})^2}}$$

$$t_b = \frac{b}{m_b} \qquad t_a = \frac{a}{m_a}$$ оценивается по модулю

mb	0,0404	
ma	0,7979	
ta	54,3870	> tкрит, значит параметр a значим
tb	-22,5702	> tкрит, значит параметр b значим

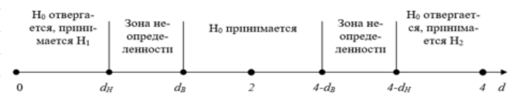
Рис. 1.27. Определение значимости параметров

Так как оба параметра уравнения регрессии значимы, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

8. Для определения наличия автокорреляции остатков между соседними членами ряда используется критерий (тест) Дарбина – Уотсона (рис. 1.28).

8. Для определения наличия автокорреляции остатков между соседними членами ряда используется критерий (тест) Дарбина-Уотсона

Остатки ε должны быть случайными.
Однако иногда каждое следующее значение остатков зависит от предшествующих.
В этом случае имеет место автокорреляция в остатках.



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

d 2,481 фактическое значение критерия (теста) Дарбина-Уотсона
определяется из столбца остатки на листе "Регрессия нет мультик" с помщью "Анализ данных" → "Регрессия"
созданные регрессионный анализ назвать "Регрессия три х"

гипотеза H0 об отсутствии автокорреляции в остатках
гипотеза H1 – гипотеза о положительной автокорреляции в остатках
гипотеза H2 – гипотеза об отрицательной автокорреляции в остатках.

d табличное определяется из таблицы Дарбина-Уотсона
Столбец - m
Строка - n

dн	1,24	из таблицы Дарбина-Уотсона 22 строка, столбец 1	4-dн	2,76
dв	1,43	из таблицы Дарбина-Уотсона 22 строка, столбец 1	4-dв	2,57

Поскольку фактическое значение d=2,481 попадает в промежуток от 1,43 до 2,57, то принимаем гипотезу H0 об отсутствии автокорреляции в остатках. Значит, с уверенностью (100-α) в 95 % можно утверждать, что нет автокорреляция в остатках.

Рис. 1.28. Критерий Дарбина – Уотсона

Так как нет автокорреляции в остатках, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза (рис. 1.29).

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,980929276
R-квадрат	0,962222244
Нормированный R-квадрат	0,960333356
Стандартная ошибка	0,562179038
Наблюдения	22

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	1	160,9972764	160,9972764	509,4120727	1,06157E-15
Остаток	20	6,320905413	0,316045271		
Итого	21	167,3181818			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	43,39522501	0,797896771	54,38701672	3,29957E-23	41,73084151	45,0596085	41,73084151	45,0596085
Переменная X 1	-0,912076614	0,040410731	-22,5701589	1,06157E-15	-0,996371921	-0,827781307	-0,996371921	-0,827781307

ВЫВОД ОСТАТКА

Наблюдение	Предсказанное Y	Остатки		
1	25,71553192	-0,715531923		0,511985933
2	23,4581423	-0,458142303	0,066249416	0,20989437
3	20,62249611	0,377503889	0,698304559	0,142509186
4	21,81822855	0,181771448	0,038311188	0,033040859
5	20,88061379	0,119386208	0,003891918	0,014253067
6	24,72136841	-0,721368414	0,706868333	0,520372388
7	27,02618602	0,973813983	2,873643358	0,948313674
8	27,88627426	0,113725736	0,739751792	0,012933543
9	26,11684563	0,883154367	0,592020418	0,779961636
10	29,6712082	-0,671208197	2,416042982	0,450520444
11	29,17321437	-0,173214366	0,247997856	0,030003217
12	24,38025176	-0,38025176	0,042864482	0,144591401
13	26,09586787	-0,095867871	0,080874196	0,009190649
14	21,06120496	-0,061204962	0,001201517	0,003746047
15	26,54643372	0,453566282	0,264989434	0,205722372
16	29,08839124	-0,088391241	0,293717957	0,007813011
17	26,02198967	-1,021989665	0,871606017	1,044462875
18	26,1961963	0,803803702	3,333521418	0,646100391
19	23,37787956	-0,377879561	1,396375334	0,142792963
20	26,62304815	0,376951846	0,569770454	0,142092694
21	27,43935672	0,560643277	0,033742542	0,314320884
22	29,07927047	-0,079270475	0,40948961	0,006283808
			15,68123478	6,320905413

Рис. 1.29. Анализ данных для критерия (теста) Дарбина – Уотсона

9. Тест Голдфелда – Квандта определения гетероскедастичности остатков (рис. 1.30–1.33).

9. Тест Голфелда-Кванда определения гетероскедастичности остатков

Если для каждого значения фактора остатки имеют одинаковую дисперсию, это называется гомоскедастичной. Если это условие не соблюдается, то имеет место гетероскедастичность.

Упорядочим по x исходные данные по возрастанию, затем упорядочиваем значения y (если одинаковые x) по возрастанию

x	y		x	y	
15,05	29		15,047	29	верхняя регрессия
15,59	29		15,593	29	
15,69	29		15,686	29	
15,7	29		15,696	29	
17	28		17,004	28	
17,49	28		17,494	28	
17,95	28		17,947	28	
18,39	27		18,389	27	
18,47	27		18,473	27	
18,86	27		18,857	27	удаляем
18,94	27		18,944	27	удаляем
18,97	26		18,967	26	удаляем
19,05	25		19,048	25	удаляем
19,38	25		19,384	25	верхняя регрессия
20,47	24		20,474	24	
20,85	24		20,848	24	
21,86	23		21,859	23	
21,95	23		21,947	23	
23,66	22		23,657	22	
24,49	21		24,487	21	
24,69	21		24,685	21	
24,97	21		24,968	21	

Удаляем n/6 наблюдений 3,6666667 удаляем 4 средних наблюдений

строим 2 регрессионные модели:

9 верхним данным — Данные → Анализ данных → Регрессия → Выделяем данные по верхней регрессии и выводим на этот лист на ячейку A224 (не забыть галочку на остатки)

9 нижним данным — Данные → Анализ данных → Регрессия → Выделяем данные по нижней регрессии и выводим на этот лист на ячейку A263 (не забыть галочку на остатки)

Рис. 1.30. Тест Голдфелда – Квандта

Верхняя регрессия

ВЫВОД ИТОГОВ

егрессионная статистика	
Множеств	0,955135326
R-квадрат	0,91228349
Нормиро	0,89975256
Стандарт	0,263848968
Наблюде	9

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	1	5,06824161	5,06824161	72,80254	6E-05
Остаток	7	0,48731394	0,06961628		
Итого	8	5,55555556			

	Коэффициенты	*ндартная оши*	*t-статистика*	*P-Значение*	*Нижние 95*	*Верхние 95%*	*жние 95,0*	*рхние 95,0%*
Y-пересеч	38,28509366	1,18264099	32,3725407	6,94E-09	35,4886	41,08159523	35,4886	41,0816
Перемен	-0,598469843	0,07014053	-8,53244026	6,03E-05	-0,76433	-0,432613856	-0,7643	-0,4326

ВЫВОД ОСТАТКА

аблюдени	*Предсказанное*	*Остатки*	
1	29,27991793	-0,2799179	0,07835404
2	28,95315339	0,04684661	0,0021946
3	28,8974957	0,1025043	0,01050713
4	28,891511	0,108489	0,01176986
5	28,10871244	-0,1087124	0,0118184
6	27,81546222	0,18453778	0,03405419
7	27,54435538	0,45564462	0,20761202
8	27,27983171	-0,2798317	0,07830579
9	27,22956024	-0,2295602	0,0526979
			0,48731394

Рис. 1.31. Анализ данных для теста Голдфелда – Квандта (верхняя регрессия)

Нижняя регрессия

ВЫВОД ИТОГОВ

егрессионная статистика	
Множеств	0,995447917
R-квадрат	0,990916556
Нормиро	0,989618921
Стандарт	0,152831372
Наблюде	9

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*ачимость F*
Регрессия	1	17,836498	17,836498	763,6328	2,1E-08
Остаток	7	0,163502	0,02335743		
Итого	8	18			

	Коэффициенты	*ндартная оши*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*ижние 95,0*	*рхние 95,0%*
Y-пересеч	39,09156322	0,59655396	65,5289643	5,06E-11	37,6809	40,50218919	37,6809	40,5022
Перемен	-0,730684591	0,0264416	-27,6339064	2,09E-08	-0,79321	-0,668160152	-0,7932	-0,6682

ВЫВОД ОСТАТКА

аблюдени	*Предсказанное*	*Остатки*	
1	24,9279731	0,0720269	0,00518787
2	24,1315269	-0,1315269	0,01729932
3	23,85825086	0,14174914	0,02009282
4	23,11952874	-0,1195287	0,01428712
5	23,05522849	-0,0552285	0,00305019
6	21,80575784	0,19424216	0,03773002
7	21,19928963	-0,1992896	0,03971636
8	21,05461408	-0,0546141	0,0029827
9	20,84783034	0,15216966	0,0231556
			0,163502

Рис. 1.32. Анализ данных для теста Голдфелда – Квандта (нижняя регрессия)

По верхней и нижней регрессии рассчитываем столбец "Остаток²" и находим сумму

S1	0,49	сумма остатков верхней регрессии
S2	0,16	сумма остатков нижней регрессии

$$F = \frac{\max(S1, S2)}{\min(S1, S2)}$$

F 2,9805 фактическое значение теста Голфелда-Кванда

F табличное определяется из таблицы Фишера

n	9	так как регрессии строились исходя из данных сокращенных
m	1	

Номер строки - (n-m-1)
Номер столбца - m

Fкрит	5,59	из таблицы Фишера 7 строка, 1 столбец

Если фактическое значение < критического, то нет наличия гетероскедастичности
Если фактическое значение > критического, то есть гетероскедастичность

F < Fкрит нет гетероскедастичности остатков

Модель можно считать адекватной, если гетероскедастичности нет

Рис. 1.33. Тест Голдфелда – Квандта

Так как нет наличия гетероскедастичности, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

10. Оценка значимости уравнения с помощью F-критерия Фишера (рис. 1.34).

10. Оценка значимости уравнения с помощью F-критерия Фишера

Оценка значимости уравнения множественной регрессии в целом осуществляется путем проверки основной гипотезы H0 о статистической незначимости уравнения регрессии.

Находим F табличное из таблицы Фишера.
F табличное определяется из таблицы Фишера
Номер строки - (n-m-1)
Номер столбца - m

m	1	столбец
n	22	
n-m-1	20	строка
Fкрит	4,35	из таблицы Фишера 20 строка, 1 столбец
Dост	0,3160	
Dфакт	160,99728	
F	509,41207	

$$D_{факт} = \frac{\sum(\hat{y} - \bar{y})^2}{1}$$

$$F = \frac{D_{факт}}{D_{ост}}$$

Если фактическое значение > критического (табличного), то гипотеза H0 о статистической незначимости уравнения регрессии отвергается
и принимается гипотеза H1 о статистичекой значимости уравнения регрессии.
Если фактическое значение < критического (табличного), то гипотеза H0 о статистической незначимости уравнения регрессии принимается

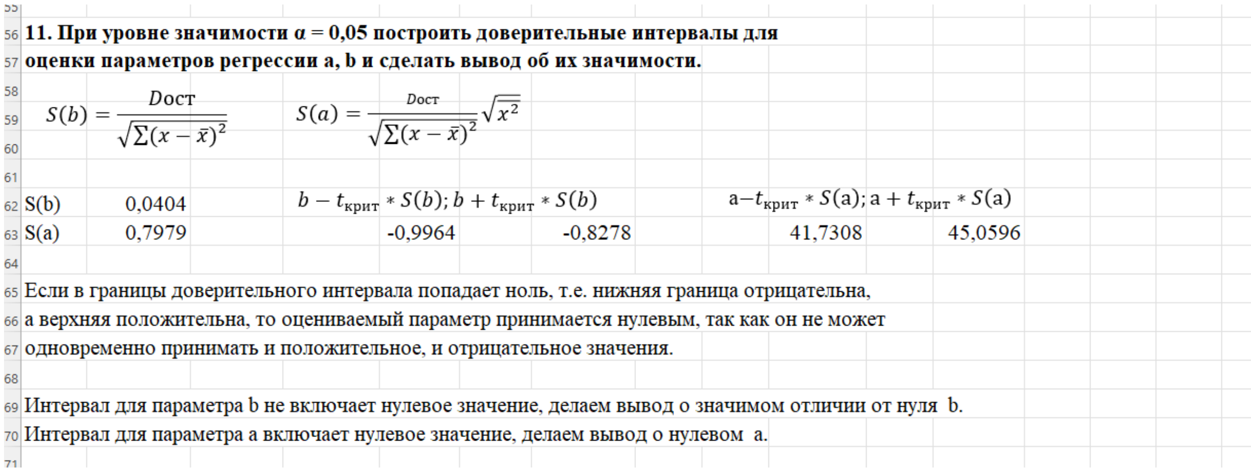
F > Fкрит делаем вывод о значимости уравнения регрессии.

Если уравнение регрессии не значимо, то нет смысла в прогнозе

Рис. 1.34. Критерий Фишера

Так как уравнение регрессии значимо, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

11. При уровне значимости α = 0,05 построить доверительные интервалы для оценки параметров регрессии a, b и сделать вывод об их значимости (рис. 1.35).



11. При уровне значимости α = 0,05 построить доверительные интервалы для оценки параметров регрессии a, b и сделать вывод об их значимости.

$$S(b)=\frac{Dост}{\sqrt{\sum(x-\bar{x})^2}} \qquad S(a)=\frac{Dост}{\sqrt{\sum(x-\bar{x})^2}}\sqrt{\overline{x^2}}$$

S(b)	0,0404	$b-t_{крит}*S(b); b+t_{крит}*S(b)$		$a-t_{крит}*S(a); a+t_{крит}*S(a)$	
S(a)	0,7979	-0,9964	-0,8278	41,7308	45,0596

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна, а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может одновременно принимать и положительное, и отрицательное значения.

Интервал для параметра b не включает нулевое значение, делаем вывод о значимом отличии от нуля b.
Интервал для параметра a включает нулевое значение, делаем вывод о нулевом a.

Рис. 1.35. Интервальная оценка

Так как оба интервала не содержат нулевых значений, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

Проведем сравнение двух уравнений регрессии (табл. 1.2).

Таблица 1.2

Сравнение уравнений регрессий с целью выбора наиболее точного прогноза

	Первая регрессия Влияние уровня экономического благополучия государства на уровень коррупции	Вторая регрессия Влияние средней заработной платы в месяц на уровень коррупции	Вывод
1	*2*	*3*	*4*
Уравнение регрессии	$\hat{y}$ = 45 – 0,0009*x	$\hat{y}$ = 43,4 – 0,91*x	
Корреляция	связь между x и y обратная, сильная – 0,966	связь между x и y обратная, сильная – 0,98	вторая регрессия
Значимость корреляции t-критерий Стьюдента	значима	значима	одинаково

Окончание табл. 1.2

1	*2*	*3*	*4*
Детерминация	93,31 %	96,22 %	вторая регрессия
Остаточная дисперсия	0,559560794	0,316045271	вторая регрессия
Значимость параметров t-критерий Стьюдента	параметр a значим, параметр b значим	параметр a значим, параметр b значим	одинаково
Критерий (тест) Дарбина – Уотсона	отсутствие автокорреляции в остатках	отсутствие автокорреляции в остатках	одинаково
Тест Голдфелда – Квандта	нет наличия гетероскедастичности остатков	нет наличия гетероскедастичности остатков	одинаково
F-критерий Фишера	уравнение значимо	уравнение значимо	одинаково
Доверительный интервал	параметр a значим, параметр b значим	параметр a значим, параметр b значим	одинаково

Исходя из сравнения, делаем вывод, что лучшую оценку даст 2 регрессия, а это значит, что уровень коррупции зависит от уровня заработной платы должностных лиц сильнее, чем от уровня экономического благополучия государства

Контрольные вопросы

1. Какие особенности эконометрических методов можно выделить?
2. Что подразумевается под спецификацией модели парной регрессии?
3. Какие выделяют предпосылки регрессионного анализа?
4. Назовите причины возникновения случайной величины.
5. Какие существуют методы выбора вида математической функции для модели регрессии?
6. Перечислите этапы метода наименьших квадратов.
7. В чем заключается экономическая интерпретация параметра линейной регрессии?
8. Что такое «линейный дисперсионный анализ»?
9. Что показывает показатель парной линейной корреляции?

10. Как проводится оценка существенности уравнения линейной регрессии?

11. Каким образом можно оценить существенность параметров линейной регрессии и корреляции?

Тест

1. В регрессионном анализе зависимая переменная (y) состоит из двух составляющих:

а) статистической и структурной;

б) закономерной и случайной;

в) динамической и стохастической.

2. К причинам появления случайной величины относятся:

а) наличие корреляции;

б) агрегирование переменных;

в) дискретный характер фактора;

г) неправильная функциональная спецификация модели.

3. Классический подход к оцениванию параметров регрессии основан на:

а) методе максимального правдоподобия;

б) методе дисперсионного анализа;

в) методе наименьших квадратов.

4. Система нормальных уравнений – это:

а) система уравнений, решением которой является значения частной корреляции;

б) система, позволяющая определить коэффициенты эластичности;

в) система, позволяющая определить оценки параметров регрессионной модели.

5. Величина какого показателя характеризует среднее изменение результата с изменением фактора на одну единицу:

а) коэффициента эластичности;

б) коэффициента детерминации;

в) коэффициента корреляции;

г) коэффициента регрессии.

6. Установить соответствие математической модели экспериментальным данным означает:

а) провести корреляционный анализ;

б) проверить значимость уравнения регрессии;

в) построить закон распределения случайной величины.

7. Какой метод позволяет построить такое уравнение регрессии, при котором сумма квадратов отклонений фактических значений результативного признака от расчетных минимальна:

а) метод скользящих средних;

б) метод максимального правдоподобия;

в) метод Крамера;

г) метод наименьших квадратов.

8. F-критерий Фишера позволяет провести оценку значимости:

а) отдельных факторов модели;

б) связи между независимыми факторами;

в) уравнения в целом.

9. Дисперсионный анализ включает в себя:

а) расчет F-критерия Фишера;

б) определение наличия мультиколлинеарности;

в) разложение общей суммы квадратов отклонений на объясненную и необъясненную части;

г) построение корреляционной матрицы.

10. Нулевая гипотеза в критерии Фишера формулируется следующим образом:

а) между результативным признаком и независимым фактором имеет место отрицательная линейная зависимость;

б) результативный признак зависит от среднего изменения независимого фактора;

в) между результативным признаком и независимым фактором отсутствует связь.

11. Нулевая гипотеза в критерии Фишера отклоняется в том случае, если:

а) табличное значение больше фактического;

б) табличное значение меньше фактического;

в) табличное значение равно фактическому.

12. Оценку значимости отдельных параметров (a и b) уравнения регрессии можно провести с помощью:

а) критерия Дарбина – Уотсона;

б) критерия надежности результатов;

в) F-критерия Фишера;

г) t-критерия Стьюдента.

13. Коэффициент корреляции характеризует собой:
а) долю объясненной дисперсии в общей;
б) меру тесноты связи между факторами;
в) долю остаточной дисперсии в общей.

2. ПАРНАЯ НЕЛИНЕЙНАЯ РЕГРЕССИЯ

До сих пор мы рассматривали линейные регрессионные модели, в которых переменные имели первую степень (модели, линейные по переменным), а параметры выступали в виде коэффициентов при этих переменных (модели линейные по параметрам). Однако соотношение между социально-экономическими явлениями и процессами часто невозможно описать линейными функциями. Если между экономическими явлениями существуют нелинейные соотношения, то они выражаются с помощью соответствующих нелинейных функций.

Различают два класса нелинейных регрессий:

1) регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам (например, полином $y = a + bx + cx^2 + \varepsilon$);

2) регрессии, нелинейные по оцениваемым параметрам (например, степенная $y = a * x^b * \varepsilon$, экспоненциальная $y = e^{a+bx}\varepsilon$).

Для оценки параметров нелинейных моделей используются два подхода. Первый подход основан на линеаризации модели и заключается в том, что с помощью подходящих преобразований исходных переменных исследуемую зависимость представляют в виде линейного соотношения между преобразованными переменными. Второй подход обычно применяется в случае, когда подобрать соответствующее линеаризующее преобразование не удается. В этом случае применяются методы нелинейной оптимизации на основе исходных переменных.

Нелинейная регрессия по включенным в нее переменным определяется, как и в линейной, методом наименьших квадратов, так как эти функции линейны по параметрам. Так, для полиномов различных степеней применяется замена переменных ($x = x_1, x^2 = x_2, x^3 = x_3, \ldots, x^k = x_k$). Следовательно, полином любого порядка сводится к линейной регрессии с ее методами оценивания параметров и проверки гипотез.

Также метод наименьших квадратов хорошо оценивает гиперболу $y = a + \frac{b}{x} + \varepsilon$. Заменив $\frac{1}{x}$ на z, получим линейное уравнение регрессии $y = a + bz + \varepsilon$. Полулогарифмическая кривая $y = a + b\ln x + \varepsilon$ также может принять линейный вид, в случае замены $\ln x$ на z получим $y = a + bz + \varepsilon$.

Возможны и иные модели, нелинейные по объясняющим переменным, например $y = a + b \cdot \sqrt{x} + \varepsilon$.

В моделях, нелинейных по оцениваемым параметрам, но приводимым к линейному виду, метод наименьших квадратов применяется к преобразованным уравнениям. Если в линейной модели и моделях, нелинейных по переменным, при оценке параметров исходят из критерия $\sum(y - y_x)^2 \to \min$, то в моделях, нелинейных по оцениваемым параметрам, требование метода наименьших квадратов применяется не к исходным данным результативного признака, а к их преобразованным величинам.

Регрессии, нелинейные по оцениваемым параметрам, подразделяются на два типа:

– нелинейные модели внутренне линейные (модель с помощью соответствующих преобразований может быть приведена к линейному виду);

– нелинейные модели внутренне нелинейные (модель не может быть сведена к линейной).

К первому типу относятся: степенная $y = ax^b\varepsilon$, так как логарифмирование уравнения по основанию e приводит к линейному виду $\ln y = \ln a + b\ln x + \ln\varepsilon$. Здесь предполагается, что случайная ошибка ε мультипликативно связана с x. Если представить данную модель в виде $y = ax^b + \varepsilon$, то она становится внутренне нелинейной, так как ее невозможно привести к линейному виду.

Также внутренне нелинейные: $y = a + bx^c + \varepsilon$ и $y = a\left(1 - \frac{1}{1-x^b}\right) + \varepsilon$.

Внутренне линейной является экспоненциальная модель $y = e^{a+bx}$, поскольку, логарифмируя, получим $\ln y = a + bx + \ln\varepsilon$.

В моделях, нелинейных по оцениваемым параметрам, но приводимым к линейному виду, МНК применяется к преобразованным уравнениям. Если в линейной модели и моделях, нелинейных по переменным, при оценке параметров исходят из критерия $\sum(y - y_x)^2 \to \min$, то в моделях, нелинейных по оцениваемым параметрам, требование МНК применяется не к исходным данным результативного признака, а к преобразованным величинам, т. е. к $\ln y$, $\frac{1}{y}$ и т. д.

Коэффициент эластичности. На практике часто бывает необходимо сравнить влияние на зависимую переменную различных объясняющих переменных, когда последние выражаются разными единицами измерения. В экономике широкое распространение получил коэффициент эластичности, который представляет собой показатель силы связи фактора x с результатом y. Он характеризует процентное изменение результата, если фактор изменится на 1 %. Обычно применяют следующую формулу расчета коэффициента эластичности:

$$Э = f'(x)\frac{x}{y}, \tag{2.1}$$

где $f'(x)$ – первая производная, характеризующая соотношение приростов результатов и фактора для соответствующей формы связи. Формулы расчета коэффициентов эластичности для наиболее распространенных типов уравнений регрессии приведены в прил. 1.

При рассмотрении различных зависимостей между x и y формулы расчета коэффициента эластичности будут меняться, и в каждом случае данный коэффициент будет зависеть от значения фактора x. Только для степенной функции $y = a \cdot x^b$ коэффициент эластичности представляет собой постоянную величину, равную параметру b. Параметр b в такой функции имеет четкую экономическую интерпретацию – он показывает процентное изменение результата при увеличении фактора на 1 %.

В силу того, что коэффициент эластичности в большинстве случаев не является величиной постоянной, обычно рассчитывается средний показатель эластичности по формуле

$$\bar{Э} = b \cdot \frac{\bar{x}}{\bar{y}}. \tag{2.2}$$

Корреляция для нелинейной регрессии. Уравнение нелинейной регрессии дополняется показателями корреляции. Наиболее часто используемый – коэффициент (индекс) корреляции:

$$R = \left(1 - \frac{\sigma^2_{ост}}{\sigma^2_y}\right)^{1/2}, \tag{2.3}$$

где σ^2_y – общая дисперсия результативного признака y; $\sigma^2_{ост}$ – остаточная дисперсия, определяемая исходя из уравнения регрессии.

Поскольку формулы расчета общей и остаточной дисперсии известны, то можно записать индекс корреляции в ином виде:

$$R = \sqrt{1 - \frac{\sum(y - y_x)^2}{\sum(y - \bar{y})^2}}, \tag{2.4}$$

причем $0 \leq R \leq 1$, чем ближе значение индекса корреляции к единице, тем теснее связь между рассматриваемыми параметрами.

Кроме того, при расчете тесноты связи для нелинейного уравнения регрессии применяют индекс детерминации (R^2), который равен квадрату индекса корреляции.

Индекс детерминации используется для проверки существенности уравнения нелинейной регрессии в целом F-критерием Фишера:

$$F = \frac{R^2}{1-R^2} * \frac{n-m-1}{m}, \quad (2.5)$$

где R^2– индекс детерминации; n – число наблюдений; m – число параметров при переменных x.

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

Задача 2

Проведем анализ экономических преступлений в сфере субсидирования в Республике Башкортостан за 2001–2021 гг. и обоснование прогноза количества экономических преступлений в сфере субсидирования на 2022 г.

Результирующей переменной будем считать «количество преступлений (y)», а факторной переменной объем инвестиций на душу населения (x). Объем инвестиций на душу населения характеризует инвестиционный климат и уровень инвестиционной активности в регионе. Другими словами, данный показатель характеризует предпосылки экономического развития региона и районов.

Таблица 2.1

Исходные данные

Год	Количество преступлений	Объем инвестиций на душу населения, руб.
1	*2*	*3*
2001	83	50 897
2002	65	54 318
2003	68	55 956
2004	78	57 317
2005	69	65 065
2006	34	70 141

Окончание табл. 2.1

1	*2*	*3*
2007	37	78 046
2008	33	88 714
2009	36	89 644
2010	38	89 742
2011	75	45 807
2012	59	48 886
2013	61	50 360
2014	70	51 585
2015	62	58 559
2016	31	63 127
2018	33	70 241
2019	30	79 843
2020	32	80 680
2021	34	80 768

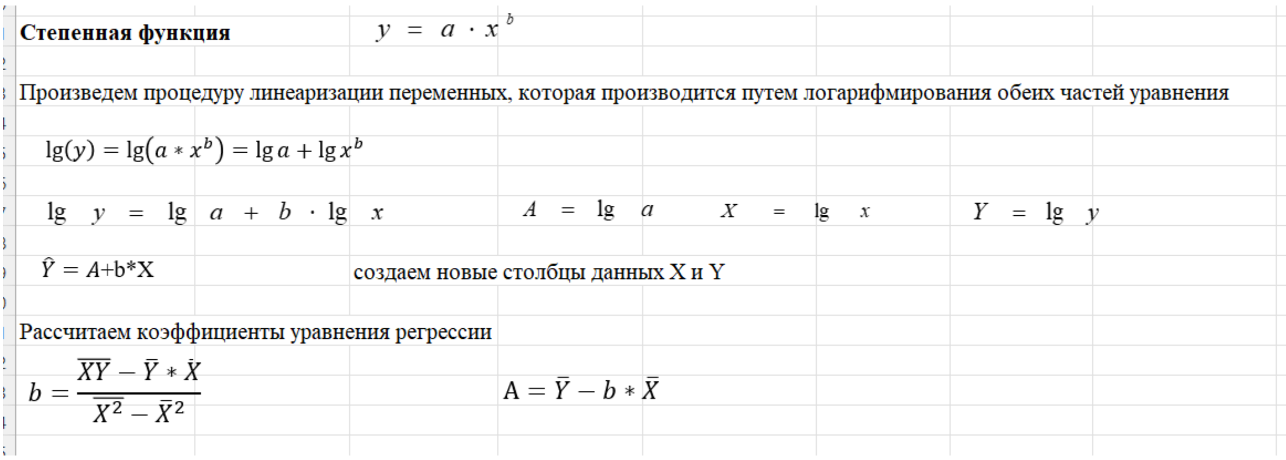
Построим уравнение парной нелинейной регрессии в 4 формах: степенная, экспоненциальная, логарифмическая, гиперболическая. Выберем наилучшую модель, более точно описывающую взаимосвязь между количеством преступлений (y)» и объемом инвестиций на душу населения (x). Все расчеты будем проводить в Microsoft Excel.

Определим количество факторов (m) и объем выборки (n):

$m = 1$, так как в модели 1 фактор x;

$n = 20$, так как выборка состоит из 20 наблюдений.

Степенная функция. Проведем процедуру линеаризации модели (рис. 2.1).



Степенная функция $y = a \cdot x^b$

Произведем процедуру линеаризации переменных, которая производится путем логарифмирования обеих частей уравнения

$\lg(y) = \lg(a * x^b) = \lg a + \lg x^b$

$\lg y = \lg a + b \cdot \lg x$ $A = \lg a$ $X = \lg x$ $Y = \lg y$

$\hat{Y} = A + b * X$ создаем новые столбцы данных X и Y

Рассчитаем коэффициенты уравнения регрессии

$b = \dfrac{\overline{XY} - \bar{Y} * \bar{X}}{\overline{X^2} - \bar{X}^2}$ $A = \bar{Y} - b * \bar{X}$

Рис. 2.1. Линеаризация степенной функции

Выполним расчеты с помощью встроенной функции «Анализ данных» (рис. 2.2).

Данные → Анализ данных → Регрессия → Выделяем данные X и Y, выводим на этот лист на ячейку А48

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,826369369
R-квадрат	0,682886334
Нормированный R-кв	0,665268908
Стандартная ошибка	0,094468062
Наблюдения	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	0,345920246	0,345920246	38,76198138	7,11918E-06
Остаток	18	0,160635866	0,008924215		
Итого	19	0,506556112			

	Коэффициенты	*андартная ошиб*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	8,369509804	1,074334014	7,790416851	3,5693E-07	6,112417796	10,62660181	6,112417796	10,62660181
Переменная X 1	-1,38961727	0,223198987	-6,225912092	7,11918E-06	-1,85854094	-0,920693599	-1,85854094	-0,920693599

Рис. 2.2. Анализ данных для степенной функции

Построим уравнение нелинейной парной регрессии в степенной форме (рис. 2.3).

A	8,37	a	234158433,29
b	-1,39		

$\hat{y} = a * x^b = 234158433{,}29 * x^{-1{,}39}$ уравнение нелинейной парной регрессии

Рассчитаем коэффициент корреляции и детерминации

R	0,83	связь y и x сильная
R^2	0,68	68% изменения y зависит от изменения x

Коэффициент эластичности

$Э = b$

Э	-1,39	%	если x изменится на 1%, то y изменится на 1,39%

Рис. 2.3. Уравнение нелинейной парной регрессии в степенной форме

Экспоненциальная функция. Проведем процедуру линеаризации модели (рис. 2.4).

Экспоненциальная функция $y = e^{a + bx}$

Произведем процедуру линеаризации переменных, которая производится путем логарифмирования обеих частей уравнения

$$\ln(y) = \ln(e^{a+bx}) = (a + b * x) * \ln e$$

$\ln y = a + bx$ $\quad Y = \ln y$

$Y = a + bx$ создаем новый столбец данных Y

Рассчитаем коэффициенты уравнения регрессии

$$b = \frac{\overline{xY} - \bar{Y} * \bar{x}}{\overline{x^2} - \bar{x}^2} \qquad a = \bar{Y} - b * \bar{x}$$

Рис. 2.4. Линеаризация экспоненциальной функции

Выполним расчеты с помощью встроенной функции «Анализ данных» (рис. 2.5).

Данные → Анализ данных → Регрессия → Выделяем данные x и Y, выводим на этот лист на ячейку A106

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,814646935
R-квадрат	0,663649628
Нормированный R-ква	0,644963496
Стандартная ошибка	0,224021227
Наблюдения	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	1,782369707	1,782369707	35,5156239	1,2246E-05
Остаток	18	0,903339184	0,05018551		
Итого	19	2,685708891			

	Коэффициенты	*андартная ошиб*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	5,24314081	0,235275699	22,28509286	1,47657E-14	4,748844907	5,737436712	4,748844907	5,737436712
Переменная X 1	-2,06059E-05	3,45765E-06	-5,959498628	1,2246E-05	-2,78701E-05	-1,33416E-05	-2,78701E-05	-1,33416E-05

Рис. 2.5. Анализ данных для экспоненциальной функции

Построим уравнение нелинейной парной регрессии в экспоненциальной форме (рис. 2.6).

6				
7	a	5,24		
8	b	-0,000021		
9				
0	$\hat{y} = e^{a+b*x} = e^{5,24+0,000021*x}$			уравнение нелинейной парной регрессии
1				
2	Рассчитаем коэффициент корреляции и детерминации			
3				
4	R	0,81		связь y и x сильная
5	R^2	0,66		66% изменения y зависит от изменения x
6				
7	Коэффициент эластичности			
8				
9	$Э = b \cdot x$			
0				
1	Э	-1,37	%	если x изменится на 1%, то y изменится на 1,37%

Рис. 2.6. Уравнение нелинейной парной регрессии в экспоненциальной форме

Логарифмическая функция. Проведем процедуру линеаризации модели (рис. 2.7).

Логарифмическая функция		$y = a + b \ln x$	
Произведем процедуру линеаризации переменных, которая производится путем замены			
$\hat{y}$ = a+b*X	$X = \ln x$		создаем новый столбец данных X
Рассчитаем коэффициенты уравнения регрессии			
$b = \dfrac{\overline{Xy} - \bar{y} * \bar{X}}{\overline{X^2} - \bar{X}^2}$	$a = \bar{y} - b * \bar{X}$		

Рис. 2.7. Линеаризация логарифмической функции

Выполним расчеты с помощью встроенной функции «Анализ данных» (рис. 2.8).

53									
54	Данные → Анализ данных → Регрессия → Выделяем данные X и y, выводим на этот лист на ячейку A160								
55									
56	ВЫВОД ИТОГОВ								
57									
58	*Регрессионная статистика*								
59	Множественный R	0,824281194							
60	R-квадрат	0,679439487							
61	Нормированный R-кв	0,66163057							
62	Стандартная ошибка	10,9940895							
63	Наблюдения	20							
64									
65	Дисперсионный анализ								
66		*df*	*SS*	*MS*	*F*	*Значимость F*			
67	Регрессия	1	4611,389431	4611,389431	38,15164461	7,86389E-06			
68	Остаток	18	2175,660069	120,8700038					
69	Итого	19	6787,0495						
70									
71		*Коэффициенты*	*андартная ошиб*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
72	Y-пересечение	823,5175817	125,0298147	6,586569642	3,46769E-06	560,8396884	1086,195475	560,8396884	1086,195475
73	Переменная X 1	-69,67988471	11,28108292	-6,176701758	7,86389E-06	-93,38056045	-45,97920896	-93,38056045	-45,97920896

Рис. 2.8. Анализ данных для логарифмической функции

Построим уравнение нелинейной парной регрессии в логарифмической форме (рис. 2.9).

a	823,52		
b	-69,68		
$\hat{y} = a + b * lnx$ = 823,52–69,68*lnx			уравнение нелинейной парной регрессии
Рассчитаем коэффициент корреляции и детерминации			
R	0,82		связь y и x сильная
R^2	0,68		68% изменения y зависит от изменения x
Коэффициент эластичности			
$Э = \dfrac{b}{a + b\ln x}$			
Э	-1,40	%	если x изменится на 1%, то y изменится на 1,4%

Рис. 2.9. Уравнение нелинейной парной регрессии в логарифмической форме

Гиперболическая функция. Проведем процедуру линеаризации модели (рис. 2.10).

Гиперболическая функция $y = a + \frac{b}{x}$

Произведем процедуру линеаризации переменных, которая производится путем замены

$\hat{y} = a + b*X$ $\quad X = \frac{1}{x}$ $\quad$ создаем новый столбец данных X

Рассчитаем коэффициенты уравнения регрессии

$$b = \frac{\overline{Xy} - \bar{y} * \bar{X}}{\overline{X^2} - \bar{X}^2} \qquad a = \bar{y} - b * \bar{X}$$

Рис. 2.10. Линеаризация гиперболической функции

Выполним расчеты с помощью встроенной функции «Анализ данных» (рис. 2.11).

Данные → Анализ данных → Регрессия → Выделяем данные X и y, выводим на этот лист на ячейку A214

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,826622811
R-квадрат	0,683305271
Нормированный R-ква	0,665711119
Стандартная ошибка	10,92759708
Наблюдения	20

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	4637,626698	4637,626698	38,83706849	7,03311E-06
Остаток	18	2149,422802	119,4123779		
Итого	19	6787,0495			

	Коэффициенты	*андартная ошиб*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-19,52038756	11,63873313	-1,677191782	0,110784517	-43,97245852	4,931683393	-43,97245852	4,931683393
Переменная X 1	4497296,197	721652,7503	6,231939385	7,03311E-06	2981160,028	6013432,365	2981160,028	6013432,365

Рис. 2.11. Анализ данных для гиперболической функции

Построим уравнение нелинейной парной регрессии в гиперболической форме (рис. 2.12).

30								
31	a	-19,52						
32	b	4497296,20						
33								
34	$\hat{y} = a + \frac{b}{x} = -19{,}52 + 4497296{,}2 * \frac{1}{x}$				уравнение нелинейной парной регрессии			
35								
36	Рассчитаем коэффициент корреляции и детерминации							
37								
38	R	0,83		связь y и x сильная				
39	R^2	0,68		68% изменения y зависит от изменения x				
40								
41	Коэффициент эластичности							
42–44	$Э = \frac{-b}{ax+b}$							
45	Э	-1,41	%	если x изменится на 1%, то y изменится на 1,41%				

Рис. 2.12. Уравнение нелинейной парной регрессии в гиперболической форме

Проведем сравнение четырех уравнений регрессии (табл. 2.2).

Таблица 2.2

Сравнение по критерию Фишера

Модель	F	Вывод
Степенная функция	38,8	> Fкрит делаем вывод о значимости уравнения регрессии
Экспоненциальная функция	35,5	> Fкрит делаем вывод о значимости уравнения регрессии
Логарифмическая функция	38,2	> Fкрит делаем вывод о значимости уравнения регрессии
Гиперболическая функция	38,8	> Fкрит делаем вывод о значимости уравнения регрессии

F табличное определяется из таблицы Фишера.

Номер строки – (n-m-1), где n – количество выборки, m – количество неизвестных.

Номер столбца – m.

Fкрит = 4,41, так как строка 18, столбец 1.

Все уравнения значимы, нельзя выбрать лучшее или исключить по незначимости (табл. 2.3).

Таблица 2.3

Сравнение по коэффициентам корреляции и детерминации

Модель	R	R^2
Степенная функция	0,826	0,6829
Экспоненциальная функция	0,815	0,6636
Логарифмическая функция	0,824	0,6794
Гиперболическая функция	0,827	0,6833

Самая лучшая связь у гиперболической функции, так как коэффициенты корреляции и детерминации самые высокие.

Также нельзя исключить ни одного уравнения, так как все коэффициенты корреляции показывают наличие сильной связи.

Таблица 2.4

Сравнение по критерию Стьюдента

Модель	a	b	Вывод
Степенная функция	234158433,3	–1,4	параметр a значим, параметр b не значим
Экспоненциальная функция	5,2	–0,000021	параметр a значим, параметр b не значим
Логарифмическая функция	823,5	–69,7	параметр a значим, параметр b значим
Гиперболическая функция	-19,5	4497296,2	параметр a значим, параметр b значим

t критическое определяется из таблицы Стьюдента.

Номер строки – (n-m-1), где n – количество выборки, m – количество неизвестных.

Номер столбца – уровень значимости (обычно 0,05).

tкрит = 2,1, так как строка 18, столбец 0,05.

Уравнения степенной и экспоненциальной зависимости исключаем из анализа, так как не значим параметр b.

Выбираем либо логарифмическую зависимость, либо гиперболическую, так как они наиболее точно будут отражать зависимость между количеством преступлений (y)» и объемом инвестиций на душу населения (x).

Контрольные вопросы

1. Какие виды нелинейных регрессий можно выделить?
2. Что подразумевается под процедурой линеаризации?
3. Что показывает коэффициент эластичности?
4. Каким образом производится расчет коэффициента корреляция для нелинейной регрессии?
5. Как проводится оценка существенности уравнения нелинейной регрессии?

Тест

1. К линейному виду не может быть сведена модель:

а) нелинейная относительно параметров;

б) нелинейная относительно факторов;

в) нелинейная модель внутренне линейная;

г) нелинейная модель внутренне нелинейная.

2. Величина какого показателя характеризует процентное изменение результата, если фактор изменится на 1 %:

а) коэффициента регрессии;

б) коэффициент детерминации;

в) коэффициента корреляции;

г) коэффициента эластичности.

3. Какие из регрессий могут быть приведены к линейному виду:

а) $y = a + bx^c + \varepsilon$;

б) $y = ax^b\varepsilon$;

в) $y = \alpha + b \cdot \sqrt{x} + \varepsilon$;

г) $y = ax^b + \varepsilon$.

4. Показатель, позволяющий проранжировать факторы по силе влияния на результат:

а) параметр регрессии;

б) критерий Фишера;

в) коэффициент эластичности;

г) критерий Стьюдента.

5. Для какой функции коэффициент эластичности представляет собой постоянную величину, равную параметру b:

а) линейной;

б) показательной;

в) степенной;
г) экспоненциальной.

3. МНОЖЕСТВЕННЫЙ РЕГРЕССИОННЫЙ АНАЛИЗ

Множественный регрессионный анализ является развитием парного регрессионного анализа. Экономические явления, как правило, определяются большим числом одновременно и совокупно действующих факторов. В связи с этим часто возникает задача исследования зависимости одной переменной Y от нескольких объясняющих переменных $X_1, X_2, ..., X_n$.

Обозначим i-е наблюдение зависимой переменной через y_i, а объясняющих переменных через $x_{i1}, x_{i2}, ..., x_{ip}$. Тогда модель множественной линейной регрессии можно представить в виде

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i, \qquad (3.1)$$

где $i = 1,2, ..., n$.

Модель (3.1), в которой зависимая переменная y_i, случайная ошибка ε_i и объясняющие переменные $x_{i1}, x_{i2}, ..., x_{ip}$ удовлетворяют приведенным в п. 2.1 предпосылкам регрессионного анализа, называется классической нормальной линейной моделью множественной регрессии.

Включение в регрессионную модель новых объясняющих переменных приводит к целесообразности использования матричных обозначений. Тогда вектор значений зависимой переменной, матрица значений объясняющих переменных, вектор параметров и вектор случайных ошибок примут вид:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}; X = \begin{pmatrix} 1 & x_{11} & x_{12} & ... & x_{1p} \\ 1 & x_{21} & x_{22} & ... & x_{2p} \\ 1 & ... & ... & ... & ... \\ 1 & x_{n1} & x_{n2} & ... & x_{\text{пр}} \end{pmatrix}; \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ ... \\ \beta_p \end{pmatrix}; \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{pmatrix} \qquad (3.2)$$

Тогда в сокращенной матричной форме модель (3.1) будет выглядеть следующим образом:

$$Y = X\beta + \varepsilon. \qquad (3.3)$$

Приведенные выше уравнения (3.3) представляют собой систему уравнений наблюдения объекта в рамках модели множественной регрессии в виде

$$y = a + b_1 x_1 + b_2 x_2 + ... + b_p x_p + \varepsilon. \qquad (3.4)$$

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства, в макроэкономических расчетах. Основная цель множественной регрессии – построить модель с большим числом факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на моделируемый показатель.

Под спецификацией модели подразумевается формулировка вида модели исходя из соответствующей теории связи между переменными. Иными словами, суть проблемы спецификации включает в себя два круга вопросов: отбор факторов и выбор вида уравнения регрессии.

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 сильно коррелировать друг с другом, тем более находиться в строгой функциональной связи (т. е. они не должны быть интеркоррелированы). Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результативный показатель и параметры уравнения регрессии оказываются неинтерпретируемыми.

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

Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. Если строится модель с набором p факторов, то для нее рассчитывается показатель детерминации R^2, который фиксирует долю объясненной вариации результативного признака. Влияние других не учтенных в модели факторов оценивается как $(1 - R^2)$ с соответствующей остаточной дисперсией S^2. При дополнительном включении в регрессию $(p + 1)$-го фактора коэффициент детерминации должен возрастать, а остаточная дисперсия уменьшаться:

$$R_{p+1}^2 \geq R_p^2 \text{ и } S_{p+1}^2 \geq S_p^2. \qquad (3.5)$$

Если же этого не происходит и данные показатели практически мало отличаются друг от друга, то включаемый в анализ фактор x_{p+1} не улучшает модель и является лишним фактором. Отбор факторов осуществляется в две стадии: на первой подбираются факторы исходя из сущности проблемы; на второй отбираются наиболее значимые на основе матрицы показателей корреляции и t-статистики Стьюдента для параметров регрессии.

Так как необходима независимость действия факторов, то коллинеарность факторов нарушает это условие. Если факторы явно коррелируют между собой, т. е. $R_{xixj} \geq 0{,}7$, то они дублируют друг друга и один из них лучше 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 незначимости друг от друга.

При определении оптимального набора факторов могут использоваться два метода:

– метод включения;

– метод исключения.

Согласно методу включения сначала строится уравнение регрессии с одним наиболее влияющим фактором (фактор, для которого значение парного коэффициента корреляции с результативным признаком больше по модулю). Затем в него последовательно вводятся следующие факторы и определяется пара наиболее влияющих факторов. На следующем к первым двум добавляется еще по одному фактору и определяется наилучшая тройка факторов и т. д.

На каждом шаге строится модель регрессии и проверяется значимость факторов. В модель включают только значимые факторы. Для проверки значимости фактора могут использоваться либо критерий Стьюдента, либо частный критерий Фишера. Процесс заканчивается, когда не остается факторов, которые следует включить в модель.

Согласно методу исключения сначала строится уравнение регрессии с полным набором факторов, из числа которых затем последовательно исключаются незначимые (наименее значимые) факторы. На каждом шаге исключается только один фактор, так как после исключения какого-либо фактора другой фактор, бывший до этого незначимым, может стать значимым. Процесс заканчивается, когда не остается факторов, которые следует исключить из модели.

Методы включения и исключения не гарантируют определение оптимального набора факторов, но в большинстве случаев дают результаты либо оптимальные, либо близкие к ним.

Не рекомендуется включать в модель очень большое число факторов, так как это может затруднить выявление качественных закономерностей и возрастает опасность включения в модель несущественных случайных факторов.

При отборе факторов рекомендуется пользоваться правилом: число включаемых факторов в 6–7 раз меньше объема совокупности, по которой строится регрессия. Если это соотношение нарушено, то число степеней свободы остаточной вариации (n-m-1) очень мало.

При выборе формы уравнения возникает проблема предпочтения линейной и нелинейной регрессионной зависимости.

В линейной множественной регрессии, которая имеет вид

$$y = a + b_1x_1 + b_2x_2 + \dots + b_px_p, \tag{3.6}$$

параметры при 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 степенной функции вида

$$y_x = ax_{1^{b1}}x_{2^{b2}} + \dots x_{p^{bp}} \tag{3.7}$$

Коэффициенты b_i являются коэффициентами эластичности. Они показывают, на сколько процентов изменяется в среднем результат с изменением соответствующего фактора на 1 % при неизменности действия других факторов. Этот вид уравнения регрессии получил наибольшее распространение в производственных функциях, в исследованиях спроса и потребления.

Возможны и другие линеаризуемые функции для построения уравнения множественной регрессии:

– экспонента $y = e^{\alpha+\beta 1x1+\beta 2x2+\dots+\beta pxp+\varepsilon}$;

– показательная $y = ab_{1^{x1}}b_2^{x2}\dots b_p^{xp}$ и т. д.

Если недостаточно стандартного набора функций регрессии, то можно использовать любые другие функции.

Мультиколлинеарность и ее последствия. 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.

Мультиколлинеарность факторов может привести к следующим нежелательным последствиям:

– оценки параметров становятся ненадежными. Данная ненадежность выражается через большие стандартные ошибки, в то время как модель в целом представляется значимой, т. е. значение множественного коэффициента корреляции завышено;

– небольшое изменение исходных данных приводит к существенному изменению оценок параметров модели (показатель неустойчивости);

– оценки параметров модели имеют неоправданно большие значения, что делает модель непригодной для анализа и прогнозирования;

– становится невозможным определить изолированное влияние факторов на результативный показатель.

Данная проблема является обычной для регрессий временных рядов. Если независимые переменные имеют ярко выраженный временной тренд, то они будут тесно коррелированы.

На практике о наличии мультиколлинеарности судят по матрице парных линейных коэффициентов корреляции (корреляционной матрице). Определитель матрицы для уравнения, включающего три объясняющие переменные, выглядит так:

$$\mathrm{Det}|R| = \begin{vmatrix} r_{x1x1} & r_{x2x1} & r_{x3x1} \\ r_{x1x2} & r_{x2x2} & r_{x3x2} \\ r_{x1x3} & r_{x2x3} & r_{x3x3} \end{vmatrix}. \qquad (3.8)$$

Чем ближе к нулю определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И, наоборот, чем ближе к единице определитель матрицы, тем меньше мультиколлинеарность факторов.

Коэффициент корреляции, измеряющий связь признака с самим собой, равен единице, так как в этом случае имеет место максимально тесная связь.

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

Оценка параметров множественной регрессионной модели. Как и в случае парной регрессии, для множественной регрессионной модели выбираются такие значения коэффициентов регрессии, чтобы обеспечить наилучшее соответствие наблюдениям. Эту задачу решает метод наименьших квадратов.

Так, для уравнения $y = a + b_1x_1 + b_2x_2 + \dots + b_px_p + \varepsilon$ необходимо минимизировать функционал

$$S = \sum_i (y_i - y_{xi})^2 = \sum_i (y_i - (a + b_1x_{1i} + b_2x_{2i} + \dots \\ + b_px_{pi})^2) \to \min \tag{3.9}$$

Чтобы найти экстремум функции нескольких переменных, нужно взять ее частные производные и приравнять их к нулю:

$$\frac{\partial S}{\partial a} = 0, \frac{\partial S}{\partial b_1} = 0, \frac{\partial S}{\partial b_2} = 0, \dots, \frac{\partial S}{\partial b_p} = 0. \tag{3.10}$$

Результатом дифференцирования станет система нормальных уравнений:

$$\begin{cases} \sum y = na + b_1 \sum x_1 + b_2 \sum x_2 + \dots + b_p \sum x_p \\ \sum yx_1 = a \sum x_1 + b_1 \sum x_1^2 + b_2 \sum x_1x_2 + \dots + b_p \sum x_px_1 \\ \sum yx_p = a \sum x_p + b_1 \sum x_1x_p + b_2 \sum x_2x_p + \dots + b_p \sum x_p^2 \end{cases}. \tag{3.11}$$

Решение этой системы возможно любым методом (метод Крамера, Гаусса и т. д.). Полученные в результате коэффициенты позволяют построить уравнение регрессии.

Решение системы (3.11) удобно записать с помощью матричных обозначений. Обозначим

$$B = \begin{bmatrix} a \\ b_1 \\ \dots \\ b_p \end{bmatrix}, Y = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \dots & x_{p1} \\ 1 & x_{12} & \dots & x_{p2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{pn} \end{bmatrix}$$

В этих обозначениях система (3.11) примет вид

$$(X'X)B = X'Y, \tag{3.12}$$

где X' – транспонированная матрица X. Матрица $X'X$ является неособенной квадратной размерности ($p + 1 \times p + 1$) при условии, что столбцы матрицы X линейно независимы.

Решение системы (3.12) определяется соотношением

$$B = (X'X)^{-1}X'Y \tag{3.13}$$

Независимые переменные имеют различный экономический смысл, разные единицы измерения и масштаб. Если нужно определить степень относительного влияния отдельных факторов на изменение результативной переменной y, то переменные следует привести к сопоставимому виду. Это можно осуществить, вводя так называемые стандартизованные переменные.

Для этого на основе матрицы парных коэффициентов корреляции строится уравнение регрессии в стандартизованном масштабе:

$$t_y = \beta_1 t_{x1} + \beta_2 t_{x2} + \dots + \beta_p t_{xp} + \varepsilon, \qquad (3.14)$$

где β_i – стандартизированные коэффициенты регрессии, $t_{y,x1_{xp}}$ – стандартизованные переменные, которые вычисляются по формулам

$$t_y = \frac{y-\bar{y}}{\sigma_y}, \qquad (3.15)$$

$$t_{xi} = \frac{x_i-\bar{x}_i}{\sigma_{xi}}. \qquad (3.16)$$

Для стандартизованных переменных выполняются два условия: среднее значение равно нулю; среднее квадратическое отклонение равно единице.

Применяя МНК к уравнению множественной регрессии в стандартизованном масштабе, после соответствующих преобразований получим систему нормальных уравнений вида

$$\begin{cases} r_{yx1} = \beta_1 + \beta_2 r_{x2\text{x}1} + \beta_3 r_{x3\text{x}1} + \dots + \beta_p t_{xpx1} \\ r_{yx2} = \beta_1 r_{x2\text{x}1} + \beta_2 + \beta_3 r_{x3\text{x}2} + \dots + \beta_p t_{xpx2} \\ \dots \\ r_{yxp} = \beta_1 r_{x\text{px}1} + \beta_2 r_{x\text{px}2} + \beta_3 r_{x3\text{xp}} + \dots + \beta_p \end{cases} \qquad (3.17)$$

Решая ее методом определителей, найдем параметры – стандартизованные коэффициенты регрессии (β-коэффициенты).

Стандартизованные коэффициенты регрессии β_i сравнимы между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат. В этом основное достоинство стандартизованных коэффициентов перед коэффициентами чистой регрессии, которые несравнимы между собой.

Рассмотренный смысл стандартизованных коэффициентов регрессии позволяет использовать их при отсеве факторов – из модели исключаются факторы с наименьшим значением β_i.

Коэффициенты чистой регрессии b_i связаны со стандартизованными коэффициентами регрессии β_i следующим соотношением:

$$b_i = \beta_i \frac{\sigma_y}{\sigma_{xi}}. \qquad (3.18)$$

Это позволяет от уравнения регрессии в стандартизованном масштабе ($t_y = \beta_1 t_{x1} + \beta_2 t_{x2} + \dots + \beta_p t_{xp}$) переходить к уравнению регрессии в натуральном масштабе переменных ($y = a + b_1x_1 +$ $+b_2x_2 + \dots + b_px_p$). При таком переходе параметр a определяется как

$$a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_p\bar{x}_p. \qquad (3.19)$$

Линейные регрессионные модели с гетероскедастичными остатками. При оценке параметров уравнения регрессии применяется метод наименьших квадратов (МНК). При этом делаются определенные предпосылки относительно случайной составляющей ε. Случайная составляющая представляет собой ненаблюдаемую величину. После того как произведена оценка параметров модели, рассчитывая разности фактических и теоретических значений результативного признака y, можно определить оценки случайной составляющей $y - y$. Поскольку они не являются реальными случайными остатками, их можно считать некоторой выборочной реализацией неизвестного остатка заданного уравнения.

При изменении спецификации модели, добавлении в нее новых наблюдений выборочные оценки остатков могут меняться, поэтому в задачу регрессионного анализа входит не только построение самой модели, но и исследование случайных отклонений ε, т. е. остаточных величин.

При использовании критериев Фишера и Стьюдента делаются предположения относительно поведения остатков – остатки представляют собой независимые случайные величины и их среднее значение равно 0; они имеют одинаковую (постоянную) дисперсию и подчиняются нормальному распределению.

Статистические проверки параметров регрессии, показателей корреляции основаны на непроверяемых предпосылках распределения случайной составляющей. Они носят лишь предварительный характер.

После построения уравнения регрессии проводится проверка наличия у случайных остатков тех свойств, которые предполагались.

Связано это с тем, что оценки параметров регрессии должны отвечать определенным критериям. Они должны быть несмещенными, состоятельными и эффективными. Эти свойства оценок, полученных по МНК, имеют чрезвычайно важное практическое значение в использовании результатов регрессии и корреляции.

Несмещенность оценки означает, что математическое ожидание остатков равно нулю. Если оценки обладают свойством несмещенности, то их можно сравнивать по разным исследованиям.

Оценки считаются эффективными, если они характеризуются наименьшей дисперсией. В практических исследованиях это означает возможность перехода от точечного оценивания к интервальному.

Состоятельность оценок характеризует увеличение их точности с увеличением объема выборки. Большой практический интерес представляют те результаты регрессии, для которых доверительный интервал ожидаемого значения параметра регрессии b_i имеет предел значений вероятности, равный единице. Иными словами, вероятность получения оценки на заданном расстоянии от истинного значения параметра близка к единице.

Указанные критерии оценок (несмещенность, состоятельность и эффективность) обязательно учитываются при разных способах оценивания. Метод наименьших квадратов строит оценки регрессии на основе минимизации суммы квадратов остатков, поэтому очень важно исследовать поведение остаточных величин регрессии.

Условия, необходимые для получения несмещенных, состоятельных и эффективных оценок, представляют собой предпосылки МНК, соблюдение которых желательно для получения достоверных результатов регрессии.

Исследования остатков предполагают проверку наличия следующих пяти предпосылок МНК:

1) случайный характер остатков;

2) нулевая средняя величина остатков, не зависящая от x_i;

3) гомоскедастичность остатков – дисперсия каждого отклонения ε_i, одинакова для всех значений x_i;

4) отсутствие автокорреляции остатков – значения остатков распределены независимо друг от друга;

5) остатки подчиняются нормальному распределению.

Если распределение случайных остатков не соответствует некоторым предпосылкам МНК, то следует корректировать модель.

Прежде всего, проверяется случайный характер остатков – первая предпосылка МНК. С этой целью стоится график зависимости остатков от теоретических значений результативного признака. Если на графике получена горизонтальная полоса, то остатки представляют собой случайные величины и МНК оправдан, теоретические значения $\hat{y}$ хорошо аппроксимируют фактические значения y.

Возможны следующие случаи, если ε_i зависит от y, то:

1) остатки не случайны;

2) остатки не имеют постоянной дисперсии;

3) остатки носят систематический характер.

В этих случаях необходимо либо применять другую функцию, либо вводить дополнительную информацию и заново строить уравнение регрессии до тех пор, пока остатки не будут случайными величинами.

Вторая предпосылка МНК относительно нулевой средней величины остатков означает, что $\sum(y - y) = 0$. Это выполнимо для линейных моделей и моделей, нелинейных относительно включаемых переменных.

Вместе с тем несмещенность оценок коэффициентов регрессии, полученных МНК, зависит от независимости случайных остатков и величин x, что также исследуется в рамках соблюдения второй предпосылки МНК. С этой целью наряду с изложенным графиком зависимости остатков от теоретических значений результативного признака строится график зависимости случайных остатков от факторов, включенных в регрессию x_i.

Если остатки на графике расположены в виде горизонтальной полосы, то они независимы от значений x_i. Если же график показывает наличие зависимости, то модель неадекватна. Причины неадекватности могут быть разные. Возможно, что нарушена третья предпосылка МНК и дисперсия остатков не постоянна для каждого значения фактора x_i. Может быть неправильна спецификация модели, и в нее необходимо ввести дополнительные члены. Скопление точек в определенных участках значений фактора x_i говорит о наличии систематической погрешности модели.

Пятая предпосылка о нормальном распределении остатков позволяет проводить проверку параметров регрессии и корреляции с помощью F- и t-критериев. Вместе с тем оценки регрессии, найденные с применением МНК, обладают хорошими свойствами даже при отсутствии нормального распределения остатков, т. е. при нарушении пятой предпосылки МНК.

Совершенно необходимым для получения по МНК состоятельных оценок параметров регрессии является соблюдение третьей и четвертой предпосылок.

В соответствии с третьей предпосылкой МНК требуется, чтобы дисперсия остатков была гомоскедастичной. Это значит, что для каждого значения фактора x_i остатки имеют одинаковую дисперсию.

Если это условие применения МНК не соблюдается, то имеет место гетероскедастичность

При построении регрессионных моделей чрезвычайно важно соблюдение четвертой предпосылки МНК – отсутствие автокорреляции остатков, т. е. значения остатков распределены независимо друг от друга. Автокорреляция остатков означает наличие корреляции между остатками текущих и предыдущих (последующих) наблюдений.

Отсутствие автокорреляции остаточных величин обеспечивает состоятельность и эффективность оценок коэффициентов регрессии.

Особенно актуально соблюдение данной предпосылки МНК при построении регрессионных моделей по рядам динамики, где ввиду наличия тенденции последующие уровни динамического ряда, как правило, зависят от своих предыдущих уровней.

При несоблюдении основных предпосылок МНК приходится корректировать модель, изменяя ее спецификацию, добавлять (исключать) некоторые факторы, преобразовывать исходные данные для того, чтобы получить оценки коэффициентов регрессии, которые обладают свойством несмещенности, имеют меньшее значение дисперсии остатков и обеспечивают в связи с этим более эффективную статистическую проверку значимости параметров регрессии.

При нарушении гомоскедастичности и наличии автокорреляции ошибок рекомендуется традиционный метод наименьших квадратов заменять обобщенным методом, т. е. методом GLS (Generalized Least Squares).

Обобщенный метод наименьших квадратов применяется к преобразованным данным и позволяет получать оценки, которые обладают не только свойством несмещенности, но и имеют меньшие выборочные дисперсии.

Тест Голдфелда – Квандта. Вероятно, наиболее популярным формальным критерием является критерий, предложенный С. Голдфелдом и Р. Квандтом (Goldfeld, Quandt, 1965). При проведении проверки по этому критерию предполагается, что стандартное отклонение распределения вероятностей случайного члена в наблюдении i пропорционально значению x_i. Предполагается также, что случайный член нормально распределен и удовлетворяет другим условиям Гаусса – Маркова.

Все n наблюдений в выборке упорядочиваются по величине X, после чего оцениваются отдельные регрессии для первых и для последних наблюдений; средние наблюдений отбрасываются. Если имеет место гетероскедастичность и если предположение относительно ее природы верно, то дисперсия в последних наблюдениях будет больше, чем в первых, и это будет отражено в сумме квадратов остатков в двух указанных «частных» регрессиях.

Этапы теста Голдфелда – Квандта:

1) все n наблюдений упорядочиваются по величине x;

2) вся упорядоченная выборка после этого разбивается на три подвыборки размерностей k, (n-2k), k соответственно;

3) оцениваются отдельные регрессии для первой подвыборки (k первых наблюдений) и для третьей подвыборки (k последних наблюдений);

4) для каждой подвыборки рассчитывается $S_i = \sum \varepsilon^2$;

5) для сравнения соответствующих дисперсий строится соответствующая F-статистика $F = \frac{\max(S1,S2)}{\min(S1,S2)}$;

6) сравниваем с табличным значением, если фактическое значение меньше критического, то нет наличия гетероскедастичности, если фактическое значение больше критического, то имеет место гетероскедастичность.

Частные уравнения регрессии. На основе линейного уравнения множественной регрессии $y = a + b_1x_1 + b_2x_2 + \ldots + b_px_p + \varepsilon$ могут быть найдены частные уравнения регрессии:

$$\begin{cases} y_{x1\cdot x2,x3,\ldots xp} = f(x_1) \\ y_{x2\cdot x1,x3,\ldots xp} = f(x_2) \\ \cdots \\ y_{xp\cdot x1,x2,\ldots xp-1} = f(x_p) \end{cases}, \qquad (3.20)$$

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 вид:

$$\begin{gathered} y_{x1\cdot x2,x3,\ldots,xp} = a + b_1x_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_p\bar{x}_p + \varepsilon, \\ y_{x2\cdot x1,x3,\ldots,xp} = a + b_1\bar{x}_1 + b_2x_2 + b_3\bar{x}_3 + \ldots + b_p\bar{x}_p + \varepsilon, \\ \ldots \\ y_{xp\cdot x1,x2,\ldots,xp} = a + b_1\bar{x}_1 + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_{p-1}\bar{x}_{p-1} + b_px_p + \varepsilon. \end{gathered} \qquad (3.21)$$

При подстановке в эти уравнения средних значений соответствующих факторов они принимают вид парных уравнений линейной регрессии:

$$\begin{cases} y_{x1\cdot x2x3\ldots xp} = A_1 + b_1 x_1 \\ y_{x2\cdot x1x3\ldots xp} = A_2 + b_2 x_2 \\ \qquad\qquad \ldots \\ y_{xp\cdot x1x2\ldots xp} = A_p + b_p x_p \end{cases}, \qquad (3.22)$$

где $\begin{cases} A_1 = a + b_2\bar{x}_2 + b_3\bar{x}_3 + \ldots + b_p\bar{x}_p \\ A_2 = a + b_1\bar{x}_1 + b_3\bar{x}_3 + \ldots + b_p\bar{x}_p \\ A_p = a + b_1\bar{x}_1 + b_2\bar{x}_2 + \ldots b_{p-1}\bar{x}_{p-1} \end{cases}$.

В отличие от парной регрессии, частные уравнения регрессии характеризуют изолированное влияние фактора на результат, так как другие факторы закреплены на неизвестном уровне. На основе частных уравнений регрессии можно определять частные коэффициенты эластичности:

$$Э_{y_{xi}} = b_i \frac{x_i}{y_{xi\cdot x1x2\ldots xi-1xi+1\ldots xp}}, \qquad (3.23)$$

где b_i – коэффициенты регрессии для фактора x_i в уравнении множественной регрессии.

На практике более широкое применение получили средние частные коэффициенты эластичности, которые рассчитываются по формуле

$$\bar{Э}_{y_{xi}} = b_i \frac{\bar{x}_i}{\bar{y}_{xi}}. \qquad (3.24)$$

Множественная и частная корреляция. Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком.

Индекс (коэффициент) множественной корреляции может быть найден по аналогии с парной регрессией:

$$R_{yx1x2\ldots xp} = \sqrt{1 - \frac{\sigma^2_{ост}}{\sigma^2_y}}, \qquad (3.25)$$

где σ^2_y – общая дисперсия результативного признака; $\sigma^2_{ост}$ – остаточная дисперсия для уравнения.

Принимая во внимание формулы расчета остаточной и общей дисперсий, можно перейти к виду

$$R_{yx1x2\ldots xp} = \sqrt{1 - \frac{\sum(y - y_{x1x2\ldots xp})^2}{\sum(y - \bar{y})^2}}. \qquad (3.26)$$

Для линейного уравнения регрессии данный показатель может быть рассчитан через β-коэффициенты:

$$R_{yx1x2...xp} = \sqrt{\sum \beta_{xi} \cdot r_{yxi}}, \tag{3.27}$$

где β_{xi} – стандартизованные коэффициенты регрессии; r_{yxi} – парные коэффициенты корреляции результата с каждым фактором.

Возможен третий метод расчета индекса множественной корреляции для линейной зависимости – через определители матриц корреляции:

$$R_{yx1x2...xp} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}, \tag{3.28}$$

где Δr – определитель матрицы парных коэффициентов корреляции; Δr_{11} – определитель матрицы межфакторной корреляции.

Для уравнения $y = a + b_1 x_1 + b_2 x_2 + \ldots + b_p x_p + \varepsilon$ определитель матрицы парных коэффициентов корреляции:

$$\Delta r = \begin{vmatrix} 1 & r_{yx1} & r_{yx2} & \ldots & r_{yxp} \\ r_{yx1} & 1 & r_{x1x2} & \ldots & r_{x1xp} \\ r_{yx2} & r_{x2x1} & 1 & \ldots & r_{x2xp} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ r_{yxp} & r_{xpx1} & r_{xpx2} & \ldots & 1 \end{vmatrix}. \tag{3.29}$$

Определитель матрицы межфакторной корреляции получается вычеркиванием из Δr первой строки и первого столбца:

$$\Delta r_{11} = \begin{vmatrix} 1 & r_{x1x2} & \ldots & r_{x1xp} \\ r_{x2x1} & 1 & \ldots & r_{x2xp} \\ \ldots & \ldots & \ldots & \ldots \\ r_{xpx1} & r_{xpx2} & \ldots & 1 \end{vmatrix}. \tag{3.30}$$

Рассмотренная формула позволяет определять совокупный коэффициент корреляции, не обращаясь к уравнению множественной регрессии, а используя парные коэффициенты корреляции.

Величина индекса множественной корреляции должна быть больше или равна максимальному парному индексу корреляции:

$$R_{yx1x2...xp} \geq R_{yxi(\max)} \; (i = \overline{1,p}). \tag{3.31}$$

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 (различия в третьих, четвертых знаках). А значит, сравнивая индексы множественной и парной корреляции, можно сделать вывод о целесообразности включения в уравнение регрессии того или иного фактора.

Кроме того, для оценки тесноты связи применяют множественный индекс (коэффициент) детерминации, который равен квадрату индекса корреляции.

Частные коэффициенты (индексы) корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в уравнение регрессии. Показатели частной корреляции представляют собой отношение сокращения остаточной дисперсии за счет дополнительного включения в анализ нового фактора к остаточной дисперсии, имевшей место до введения его в модель.

В общем виде при наличии p факторов для уравнения $y = a + b_1x_1 + b_2x_2 + \dots + b_px_p + \varepsilon$ коэффициент частной корреляции, измеряющий влияние на y фактора x_i при неизменном уровне других факторов, определяется по формуле:

$$r_{yxi\cdot x1x2\dots xi-1xi+1\dots xp} = \sqrt{1 - \frac{1-R^2_{yx1x2\dots xi\dots xp}}{1-R^2_{yx1x2\dots xi-1xi+1\dots xp}}}, \qquad (3.32)$$

где $R^2_{yx1x2\dots xi\dots xp}$ – множественный коэффициент детерминации всего комплекса p-факторов с результатом; $R^2_{yx1x2\dots xi-1xi+1\dots xp}$ – тот же показатель детерминации без введения в модель фактора x_i.

Частный коэффициент корреляции может быть найден иным образом по формуле:

$$r_{yxi\cdot x1x2\dots xi-1xi+1\dots xp} = -\frac{R_{yxi}}{\sqrt{R_{yy}\cdot R_{xixi}}}, \qquad (3.33)$$

где $R_{yxi}, R_{yy}, R_{xixi}$ – алгебраические дополнения к соответствующим элементам матрицы парной корреляции.

Порядок частного коэффициента корреляции определяется количеством факторов, влияние которых исключается. $r_{yx1\cdot x2}$ – коэффициент частной корреляции первого порядка, следовательно, коэффициенты парной корреляции – коэффициенты нулевого порядка. Коэффициенты более высоких порядков можно определить через коэффициенты более низких порядков по рекуррентной формуле:

$$r_{yxi\cdot x1x2\ldots xp} = -\frac{r_{yxi\cdot x1x2\ldots xp-1} - r_{yxp\cdot x1x2\ldots xp-1} \cdot r_{xixp\cdot x1x2\ldots xp-1}}{\sqrt{(1-r^2_{yxp\cdot x1x2\ldots xp-1})(1-r^2_{xixp\cdot x1x2\ldots xp-1})}}. \qquad (3.34)$$

Сравнивая друг с другом частные коэффициенты корреляции, можно ранжировать факторы по тесноте их связи с результатом. Частные коэффициенты корреляции на основе стандартизованных коэффициентов регрессии (β-коэффициентов) дают конкретную меру тесноты связи каждого фактора с результатом в чистом виде. В двухфакторном анализе справедливы формулы:

$$r_{yx1\cdot x2} = \beta_{x1}\sqrt{\frac{1-r^2_{x1x2}}{1-r^2_{yx2}}} \qquad r_{yx2\cdot x1} = \beta_{x2}\sqrt{\frac{1-r^2_{x1x2}}{1-r^2_{yx1}}}\ . \qquad (3.35)$$

Частные коэффициенты корреляции используются на стадии формирования модели, в процедуре отсева факторов. Строя многофакторную модель:

– на первом шаге определяют уравнение регрессии с полным набором факторов и рассчитывают матрицу частных коэффициентов корреляции;

– на втором отбирают фактор с наименьшей и несущественной по t-критерию Стьюдента величиной показателя частной корреляции. Исключив его из модели, строят новое уравнение регрессии.

Процедура продолжается до тех пор, пока не окажется, что все частные коэффициенты корреляции существенно отличаются от нуля. Зная частные коэффициенты корреляции (последовательно первого, второго и более высокого порядка), можно определить совокупный коэффициент корреляции по формуле:

$$R_{yx1x2\ldots xp} = \sqrt{1-(1-r^2_{yx1})\cdot(1-r^2_{yx2\cdot x1})\cdot(1-r^2_{yx3\cdot x1x2})\ldots(1-r^2_{yxp\cdot x1x2\ldots xp-1})}. \qquad (3.36)$$

Оценка надежности результатов множественной регрессии и корреляции. Оценка значимости уравнения множественной регрессии в целом осуществляется путем проверки основной гипотезы H_0 о статистической незначимости уравнения регрессии. Для проверки этой гипотезы используют общий F-критерий Фишера:

$$F = \frac{D^{(1)}_{факт}}{D^{(1)}_{ост}} = \frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m}, \qquad (3.37)$$

где $D_{\text{факт}}$ – факторная сумма квадратов отклонений на одну степень свободы; $D_{\text{ост}}$ – остаточная сумма квадратов отклонений на одну степень свободы; R^2– коэффициент множественной детерминации; m – число параметров при переменных x; n – число наблюдений.

Оценивается значимость не только уравнения в целом, но и фактора, дополнительно включенного в модель. Для этих целей служит частный F-критерий. Он построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительно включенного в модель фактора, с остаточной дисперсией на одну степень свободы по регрессионной модели в целом. Для x_i:

$$F_{xi} = \frac{R^2_{yx1\ldots x1\ldots xp} - R^2_{yx1\ldots xi-1xi+1\ldots xp}}{1 - R^2_{yx1\ldots xi\ldots xp}} \cdot \frac{n-m-1}{1}, \tag{3.38}$$

где $R^2_{yx1\ldots xi\ldots xp}$ – коэффициент множественной детерминации для модели с полным набором факторов; $R^2_{yx1\ldots xi-1xi+1\ldots xp}$ – тот же показатель, но без включения в модель фактора x_i.

В числителе показан прирост доли объясненной вариации y за счет дополнительного включения в модель соответствующего фактора. Если величина частного F-критерия оказывается меньше табличного значения, то дополнительное включение в модель того или иного фактора нецелесообразно.

Зная величину частного F-критерия, можно определить t-критерий для коэффициента регрессии при i-м факторе:

$$t_{bi} = \sqrt{F_{xi}}\ . \tag{3.39}$$

или

$$t_{bi} = \frac{b_i}{m_{bi}}, \tag{3.40}$$

где b_i – коэффициент регрессии при факторе x_i; m_{bi} – средняя квадратическая ошибка коэффициента регрессии b_i, которая может быть определена по формуле:

$$m_{bi} \frac{\sigma_y \sqrt{1 - R^2_{yx1\ldots xp}}}{\sigma_{xi} \sqrt{1 - R^2_{xix1\ldots xp}}} \frac{1}{\sqrt{n-m-1}}, \tag{3.41}$$

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; $R^2_{xix1\ldots xp}$ – коэффициент детерминации для зависимости фактора x_i со всеми другими факторами уравнения множественной регрессии; n-m-1 – число степеней свободы остаточной суммы квадратов отклонений.

Отсев факторов при построении уравнения регрессии методом исключения можно осуществлять не только по частным коэффициентам корреляции, но и по величинам частного F-критерия Фишера и t-критерия Стьюдента.

Задача 3

По 50 предприятиям региона изучается зависимость прибыли предприятия от выработки продукции на одного работника, объема продаж и затрат на производство (табл. 3.1).

1. Построить уравнение множественной регрессии в стандартизованной форме.
2. Построить уравнение множественной регрессии в линейной форме.
3. Рассчитать коэффициенты множественной корреляции и детерминации.
4. Рассчитать коэффициенты частной корреляции.
5. Провести оценку значимости уравнения множественной регрессии с помощью F-критерия Фишера.
6. Провести оценку значимости факторов множественной регрессии с помощью частного F-критерия Фишера.
7. Провести оценку значимости параметров множественной регрессии с помощью t-критерия Стьюдента.
8. Определить наличия автокорреляции остатков между соседними членами ряда с помощью критерия (теста) Дарбина – Уотсона.
9. Провести тест Голдфелда – Квандта определения гетероскедастичности остатков.
10. Определить доверительные интервалы оценок коэффициентов регрессии, сделать вывод об их значимости.

Таблица 3.1

Исходные данные для расчета регрессионной модели

№ предприятия	y	$x1$	$x2$	$x3$
1	*2*	*3*	*4*	*5*
ПРЕДП 1	1917403	19	1073255,000	203918,450
ПРЕДП 2	426484	25	189842,000	47460,500
ПРЕДП 3	362532	38	207118,000	78704,840
ПРЕДП 4	186700	30	138518,000	41555,400
ПРЕДП 5	157286	27	90757,000	24504,390
ПРЕДП 6	151849	55	62388,000	34313,400
ПРЕДП 7	127440	9	4142,000	372,780
ПРЕДП 8	111285	25	51731,000	12932,750
ПРЕДП 9	104372	62	48400,000	30008,000
ПРЕДП 10	96809	42	46393,000	19485,060
ПРЕДП 11	85365	29	45580,000	13218,2
ПРЕДП 12	81296	46	33339,000	15335,940
ПРЕДП 13	76617	19	43073,000	8183,870
ПРЕДП 14	67649	7	60154,000	4210,780
ПРЕДП 15	54848	46	32761,000	15070,060
ПРЕДП 16	53701	37	23053,000	8529,610
ПРЕДП 17	52473	17	28511,000	4846,870
ПРЕДП 18	50666	36	25412,000	9148,320
ПРЕДП 19	46086	1	3599,000	35,990
ПРЕДП 20	41332	46	18506,000	8512,760
ПРЕДП 21	40521	38	23841,000	9059,580
ПРЕДП 22	40057	20	29420,000	5884,000
ПРЕДП 23	38245	22	18114,000	3985,080
ПРЕДП 24	36946	23	13117,000	3016,910
ПРЕДП 25	34762	23	19135,000	4401,050
ПРЕДП 26	34032	20	15047,000	3009,400
ПРЕДП 27	33062	34	15507,000	5272,380
ПРЕДП 28	32948	35	24980,000	8743,000
ПРЕДП 29	30713	30	20665,000	6199,500
ПРЕДП 30	30596	16	11556,000	1848,960
ПРЕДП 31	29970	64	7593,000	4859,520
ПРЕДП 32	26724	21	18158,000	3813,180
ПРЕДП 33	26388	26	28,000	7,280
ПРЕДП 34	26015	24	2014,000	483,360

Окончание табл. 3.1

1	*2*	*3*	*4*	*5*
ПРЕДП 35	25691	66	11044,000	7289,040
ПРЕДП 36	25206	62	5415,000	3357,300
ПРЕДП 37	24639	52	14216,000	7392,320
ПРЕДП 38	24027	16	10408,000	1665,280
ПРЕДП 39	23863	23	13953,000	3209,190
ПРЕДП 40	22894	37	3254,000	1203,980
ПРЕДП 41	22570	63	8,000	5,040
ПРЕДП 42	20905	24	15405,000	3697,200
ПРЕДП 43	19513	26	2584,000	671,840
ПРЕДП 44	19346	23	8586,000	1974,780
ПРЕДП 45	18991	24	6811,000	1634,640
ПРЕДП 46	18389	38	11911,000	4526,180
ПРЕДП 47	18090	55	11672,000	6419,600
ПРЕДП 48	17674	25	11422,000	2855,500
ПРЕДП 49	16965	34	11788,000	4007,920
ПРЕДП 50	16910	36	1169,000	420,840

Построим уравнение множественной парной регрессии. Все расчеты будем проводить в Microsoft Excel.

Определим количество факторов (m) и объем выборки (n):

m = 3, так как в модели 3 фактора x;

n = 50, так как выборка состоит из 50 наблюдений.

Проведем промежуточные расчеты для построения уравнения регрессии, для этого рассчитаем столбцы y^2, x_1^2, x_2^2, x_3^2, x_1*x_2, x_1*x_2, x_1*x_3, x_3*x_2, $y*x_1$, $y*x_2$, $y*x_3$, $(x_1-\bar{x}_1)^2$, $(x_2-\bar{x}_2)^2$, $(x_3-\bar{x}_3)^2$, $(y-\bar{y})^2$.

1. Построить уравнение множественной регрессии в стандартизованной форме (рис. 3.1, 3.2).

1) построить уравнение множественной регрессии в стандартизованной форме.

Независимые переменные имеют различный экономический смысл, разные единицы измерения и масштаб. Если нужно определить степень относительного влияния отдельных факторов на изменение результативной переменной у, то переменные следует привести к сопоставимому виду. Это можно осуществить, вводя, так называемые, «стандартизованные» переменные.

$$t_y = \beta_1 t_{x1} + \beta_2 t_{x2} + \dots + \beta_p t_{xp} + \varepsilon$$

Для построения уравнения в стандартизированной форме необходимо найти коэффициенты β, используя систему уравнений

$$\begin{cases} r_{yx1} = \beta_1 + \beta_2 r_{x2x1} + \beta_3 r_{x3x1} + \dots + \beta_p r_{xpx1} \\ r_{yx_2} = \beta_1 r_{x2x1} + \beta_2 + \beta_3 r_{x3x2} + \dots + \beta_p r_{xpx2} \\ \dots \\ r_{yxp} = \beta_1 r_{xpx1} + \beta_2 r_{xpx2} + \beta_3 r_{x3xp} + \dots + \beta_p \end{cases}$$

σx1	15,34
σx2	151897,49
σx3	30657,52
σy	271054,95

$$\sigma x = \sqrt{\frac{\sum(x-\bar{x})^2}{n}} \qquad \sigma y = \sqrt{\frac{\sum(y-\bar{y})^2}{n}}$$

Рассчитаем коэффициенты парной корреляции между факторами (x1, x2, x3)

$$r_{x1x2} = \frac{\overline{x1*x2} - \overline{x1}*\overline{x2}}{\sigma x1*\sigma x2} \qquad r_{x1x3} = \frac{\overline{x1*x3} - \overline{x1}*\overline{x3}}{\sigma x1*\sigma x3} \qquad r_{x3x2} = \frac{\overline{x3*x2} - \overline{x3}*\overline{x2}}{\sigma x3*\sigma x2}$$

rx1x2	-0,124	связь между фактором x1 и x2 обратная слабая	rx1x1	1	так как сам фактор x1 связан с самим собой самой сильной связью
rx1x3	-0,015	связь между фактором x1 и x3 обратная слабая	rx2x2	1	так как сам фактор x2 связан с самим собой самой сильной связью
rx2x3	0,972	связь между фактором x2 и x3 прямая сильная	rx3x3	1	так как сам фактор x3 связан с самим собой самой сильной связью

Произведем расчет коэффициентов парной корреляции между факторами (x1, x2, x3) и результатом (y)

$$r_{yxi} = \frac{\overline{y*xi} - \bar{y}*\overline{xi}}{\sigma y*\sigma xi}$$

ryx1	-0,127	связь между результатом у и фактором x1 обратная слабая
ryx2	0,996	связь между результатом у и фактором x2 прямая сильная
ryx3	0,968	связь между результатом у и фактором x3 прямая сильная

Рис. 3.1. Определение парных коэффициентов корреляции

Для трехфактороной модели система уравнений будет выглядит следующим образом

Для задачи имеет вид:
$$\begin{cases} r_{yx1} = \beta_1 + \beta_2 * r_{x1x2} + \beta_3 * r_{x1x3} \\ r_{yx2} = \beta_1 * r_{x2x1} + \beta_2 + \beta_3 * r_{x2x3} \\ r_{yx3} = \beta_1 * r_{x3x1} + \beta_2 * r_{x3x2} + \beta_3 \end{cases}$$

Для решения системы линейных уравнений используем метод Крамера

$$\beta 1 = \frac{\Delta 1}{\Delta} \qquad \beta 2 = \frac{\Delta 2}{\Delta} \qquad \beta 3 = \frac{\Delta 3}{\Delta}$$

где Δ - определитель матрицы парных коэффициентов корреляции из системы уравнений

Δi - определитель матрицы парных коэффициентов корреляции из системы уравнений, при замене i столбца на столбец свободных членов уравнений

β1	-0,01
β2	0,98
β3	0,02

1,000	-0,124	-0,015	-0,127
-0,124	1,000	0,972	0,996
-0,015	0,972	1,000	0,968

Δ 0,04

1,000	-0,124	-0,127
-0,124	1,000	0,996
-0,015	0,972	0,968

Δ3 0,00069

-0,127	-0,124	-0,015
0,996	1,000	0,972
0,968	0,972	1,000

Δ1 -0,00023

1,000	-0,127	-0,015
-0,124	0,996	0,972
-0,015	0,968	1,000

Δ2 0,04

ty=-0,01*tx1+0,98*tx2+0,02*tx3 уравнение в стандартизированном виде

Стандартизованные коэффициенты регрессии сравнимы между собой. Сравнивая их друг с другом, можно ранжировать факторы по силе их воздействия на результат.
Сравниваются β коэффициенты по модулю. Наибольше воздействие на результат оказывает фактор, у которого β коэффициент самый большой
Наибольшее влияние на результат имеет x2.

Стандартизованные коэффициенты регрессии можно использовать их при отсеве факторов из модели - исключаются факторы с наименьшим значением .

Рис. 3.2. Стандартизированное уравнение регрессии

Наибольшее влияние на результат имеет x_2. Таким образом, можно сказать, что прибыль предприятия сильнее всего зависит от объема продаж.

2. Построить уравнение множественной регрессии в линейной форме (рис. 3.3).

2) построить уравнение множественной регрессии в «чистом» виде.

$$\hat{y} = a + b_1x_1 + b_2x_2 + \ldots + b_px_p \qquad b_i = \beta_i \frac{\sigma_y}{\sigma_{xi}} \qquad a = \bar{y} - b_1\bar{x}_1 - b_2\bar{x}_2 - \ldots - b_p\bar{x}_p$$

$$b_1 = \beta_1 \frac{\sigma_y}{\sigma_{x1}} \qquad b_2 = \beta_2 \frac{\sigma_y}{\sigma_{x2}} \qquad b_3 = \beta_3 \frac{\sigma_y}{\sigma_{x3}} \qquad a = \bar{y} - b_1 * \overline{x_1} - b_2 * \overline{x_2} - b_3 * \overline{x_3}$$

b1	-91,90
b2	1,75
b3	0,14
a	11634,71

ŷ=11634,71-91,90*x1+1,75*x2+0,14*x3 уравнение регрессии

Рис. 3.3. Уравнение линейной множественной регрессии

3. Рассчитать коэффициенты множественной корреляции и детерминации (рис. 3.4).

3) множественная корреляция и детерминация

$$R_{yx1x2...xp} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}}$$

Показатель множественной корреляции характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком. изменяется в границах от 0 до 1, чем ближе к 1 тем связь теснее, чам ближе к 0 тем связь слабее

Δr – определитель матрицы парных коэффициентов корреляции

Δr11 – определитель матрицы парных коэффициентов корреляции без 1 строки и 1 столбца

$$\Delta r = \begin{matrix} r_{yy} & r_{yx1} & r_{yx2} & r_{yx3} \\ r_{yx1} & r_{x1x1} & r_{x1x2} & r_{x1x3} \\ r_{yx2} & r_{x2x1} & r_{x2x2} & r_{x2x3} \\ r_{yx3} & r_{x3x1} & r_{x3x2} & r_{x3x3} \end{matrix} \qquad \Delta r_{11} = \begin{matrix} r_{x1x1} & r_{x1x2} & r_{x1x3} \\ r_{x2x1} & r_{x2x2} & r_{x2x3} \\ r_{x3x1} & r_{x3x2} & r_{x3x3} \end{matrix}$$

	y	x1	x2	x3
y	1	-0,127	0,996	0,968
x1	-0,127	1	-0,124	-0,015
x2	0,996	-0,124	1	0,972
x3	0,968	-0,015	0,972	1

R 0,996

Связь между факторами и результатом сильная

R² 0,991

99,1% изменения y зависти от изменения x

Δr	0,00038
Δr11	0,04

Если множественная корреляция имеет слабую связь, то нет смысла в построении уравнения регрессии

Рис. 3.4. Коэффициенты множественной корреляции и детерминации

Так как связь факторов и результата сильная, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза. Делаем вывод, что прибыль предприятия зависит от выработки продукции на одного работника, объема продаж и затрат на производство.

4. Рассчитать коэффициенты частной корреляции (рис. 3.5, 3.6).

4) частная корреляция

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при устранении влияния других факторов, включенных в уравнение регрессии.

$r_{yxi\cdot x1x2...xi-1xi+1...xp} = -\dfrac{R_{yxi}}{\sqrt{R_{yy}\cdot R_{xixi}}}$ $R_{yxi}, R_{yy}, R_{xixi}$ – алгебраические дополнения к соответствующим элементам матрицы парной корреляции.

Ryy	0,044	1,000	-0,124	-0,015	$R_{yy} = (-1)^{1+1}\Delta 11$	Ryx1	0,00023	-0,127	-0,124	-0,015	$R_{yx1} = (-1)^{1+2}\Delta 12$
		-0,124	1,000	0,972				0,996	1,000	0,972	
		-0,015	0,972	1,000				0,968	0,972	1,000	
Rx1x1	0,00049	1,000	0,996	0,968	$R_{x1x1} = (-1)^{2+2}\Delta 22$	Ryx2	-0,043	-0,127	1,000	-0,015	$R_{yx2} = (-1)^{1+3}\Delta 13$
		0,996	1,000	0,972				0,996	-0,124	0,972	
		0,968	0,972	1,000				0,968	-0,015	1,000	
Rx2x2	0,051	1,000	-0,127	0,968	$R_{x2x2} = (-1)^{3+3}\Delta 33$	Ryx3	-0,001	-0,127	1,000	-0,124	$R_{yx3} = (-1)^{1+4}\Delta 14$
		-0,127	1,000	-0,015				0,996	-0,124	1,000	
		0,968	-0,015	1,000				0,968	-0,015	0,972	
Rx3x3	0,009	1,000	-0,127	0,996	$R_{x3x3} = (-1)^{4+4}\Delta 44$						
		-0,127	1,000	-0,124							
		0,996	-0,124	1,000							

Рис. 3.5. Алгебраические дополнения для расчета частной корреляции

$r_{yx1\cdot x2x3} = -\dfrac{R_{yx1}}{\sqrt{R_{yy}R_{x1x1}}}$ $r_{yx2\cdot x1x3} = -\dfrac{R_{yx2}}{\sqrt{R_{yy}R_{x2x2}}}$ $r_{yx3\cdot x1x2} = -\dfrac{R_{yx3}}{\sqrt{R_{yy}R_{x3x3}}}$

$r_{yx1\cdot x2x3}$	-0,049
$r_{yx2\cdot x1x3}$	0,910
$r_{yx3\cdot x1x2}$	0,035

Сравнивая друг с другом частные коэффициенты корреляции, можно ранжировать факторы по тесноте их связи с результатом.
Чем выше частная корреляция по модулю, тем сильнее фактор оказывает влияние на у.
Самый сильно влияющий фактор х2.

Рис. 3.6. Частная корреляция

Самый сильно влияющий фактор $x2$. Делаем вывод, что прибыль предприятия зависит от выработки продукции на одного работника, объема продаж и затрат на производство.

5. Провести оценку значимости уравнения множественной регрессии с помощью F-критерия Фишера (рис. 3.7)

5) оценка значимости уравнения множественной регрессии с помощью F-критерия Фишера

$F = \dfrac{R^2}{1-R^2} * \dfrac{n-m-1}{m}$ F 1731,090 фактическое рассчитанное по формуле

R² – коэффициент множественной детерминации
m – число переменных х
n – число наблюдений

Оценка значимости уравнения множественной регрессии в целом осуществляется путем проверки основной гипотезы Н0 о статистической незначимости уравнения регрессии.

Находим F табличное из таблицы Фишера.
F табличное определяется из таблицы Фишера
Номер строки - (n-m-1)
Номер столбца - m

m	3	столбец
n	50	
n-m-1	46	строка

Fкрит	2,81	из таблицы Фишера 46 строка, 3 столбец

Если фактическое значение > критического (табличного), то гипотеза Н0 о статистической незначимости уравнения регрессии отвергается
и принимается гипотеза Н1 о статистичекой значимости уравнения регрессии.
Если фактическое значение < критического (табличного), то гипотеза Н0 о статистической незначимости уравнения регрессии принимается

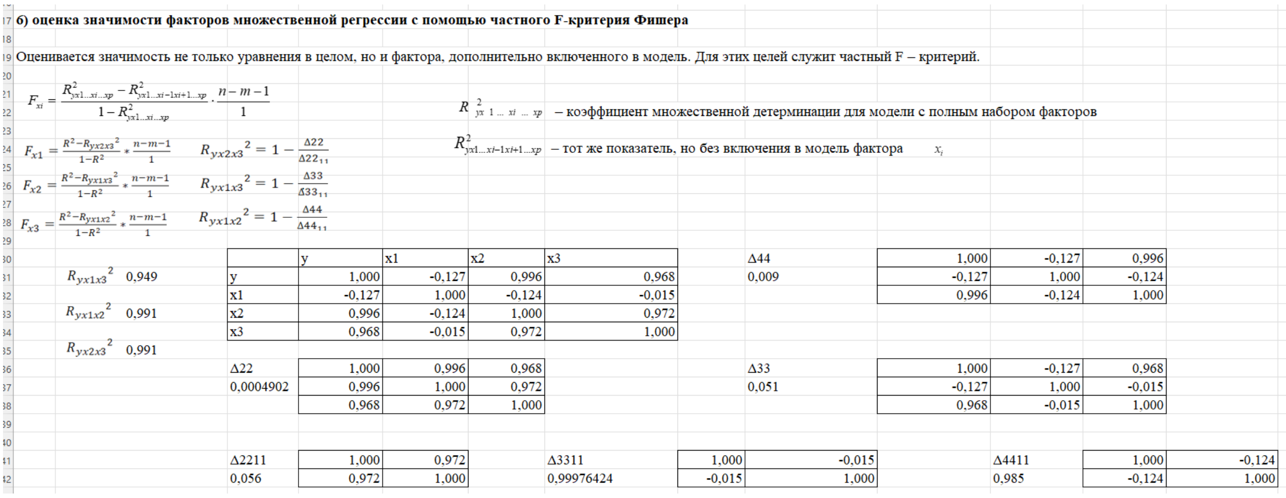
F > Fкрит делаем вывод о значимости уравнения регрессии.

Если уравнение регрессии не значимо, то нет смысла в прогнозе

Рис. 3.7. Оценка значимости уравнения с помощью критерия Фишера

Так как уравнение регрессии значимо, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

6. Провести оценку значимости факторов множественной регрессии с помощью частного F-критерия Фишера (рис. 3.8, 3.9).



б) оценка значимости факторов множественной регрессии с помощью частного F-критерия Фишера

Оценивается значимость не только уравнения в целом, но и фактора, дополнительно включенного в модель. Для этих целей служит частный F – критерий.

$$F_{xi} = \frac{R^2_{yx1...xi...xp} - R^2_{yx1...xi-1xi+1...xp}}{1 - R^2_{yx1...xi...xp}} \cdot \frac{n-m-1}{1}$$

$R^2_{yx\,1\,...\,xi\,...\,xp}$ – коэффициент множественной детерминации для модели с полным набором факторов

$R^2_{yx1...xi-1xi+1...xp}$ – тот же показатель, но без включения в модель фактора x_i

$$F_{x1} = \frac{R^2 - R_{yx2x3}{}^2}{1-R^2} * \frac{n-m-1}{1} \qquad R_{yx2x3}{}^2 = 1 - \frac{\Delta 22}{\Delta 22_{11}}$$

$$F_{x2} = \frac{R^2 - R_{yx1x3}{}^2}{1-R^2} * \frac{n-m-1}{1} \qquad R_{yx1x3}{}^2 = 1 - \frac{\Delta 33}{\Delta 33_{11}}$$

$$F_{x3} = \frac{R^2 - R_{yx1x2}{}^2}{1-R^2} * \frac{n-m-1}{1} \qquad R_{yx1x2}{}^2 = 1 - \frac{\Delta 44}{\Delta 44_{11}}$$

$R_{yx1x3}{}^2$ 0,949

$R_{yx1x2}{}^2$ 0,991

$R_{yx2x3}{}^2$ 0,991

	y	x1	x2	x3
y	1,000	-0,127	0,996	0,968
x1	-0,127	1,000	-0,124	-0,015
x2	0,996	-0,124	1,000	0,972
x3	0,968	-0,015	0,972	1,000

Δ22	1,000	0,996	0,968
0,0004902	0,996	1,000	0,972
	0,968	0,972	1,000

Δ44	1,000	-0,127	0,996
0,009	-0,127	1,000	-0,124
	0,996	-0,124	1,000

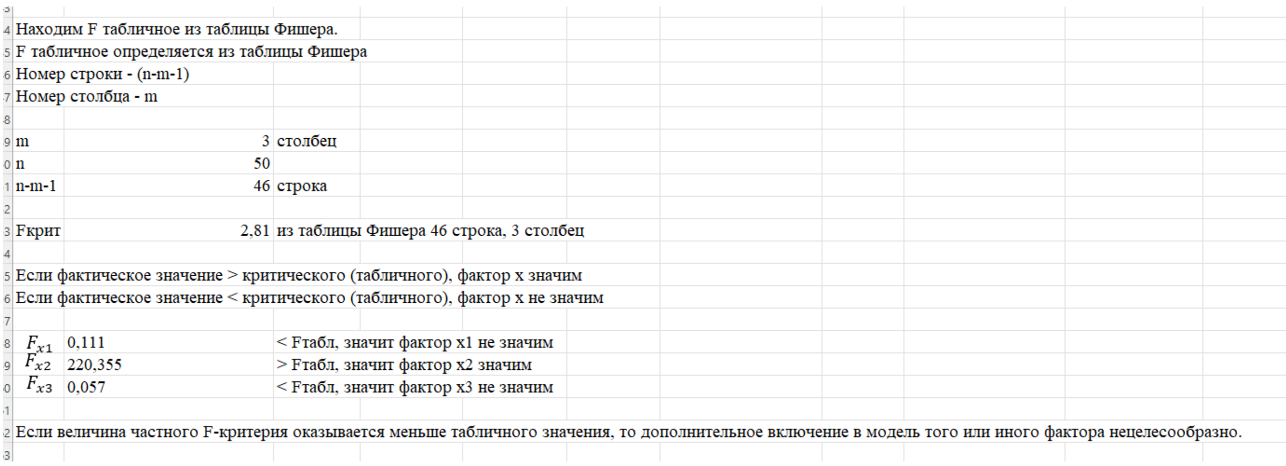
Δ33	1,000	-0,127	0,968
0,051	-0,127	1,000	-0,015
	0,968	-0,015	1,000

Δ2211	1,000	0,972
0,056	0,972	1,000

Δ3311	1,000	-0,015
0,99976424	-0,015	1,000

Δ4411	1,000	-0,124
0,985	-0,124	1,000

Рис. 3.8. Расчет частного критерия Фишера



Находим F табличное из таблицы Фишера.
F табличное определяется из таблицы Фишера
Номер строки - (n-m-1)
Номер столбца - m

m	3	столбец
n	50	
n-m-1	46	строка
Fкрит	2,81	из таблицы Фишера 46 строка, 3 столбец

Если фактическое значение > критического (табличного), фактор x значим
Если фактическое значение < критического (табличного), фактор x не значим

F_{x1}	0,111	< Fтабл, значит фактор x1 не значим
F_{x2}	220,355	> Fтабл, значит фактор x2 значим
F_{x3}	0,057	< Fтабл, значит фактор x3 не значим

Если величина частного F-критерия оказывается меньше табличного значения, то дополнительное включение в модель того или иного фактора нецелесообразно.

Рис. 3.9. Оценка значимости факторов с помощью частного критерия Фишера

Так как фактор x_1 и фактор x_3 не значимы, в адекватности модели множественной линейной регрессии есть сомнения.

7. Провести оценку значимости параметров множественной регрессии с помощью t-критерия Стьюдента (рис. 3.10).

7) оценка значимости параметров множественной регрессии с помощью t критерия Стьюдента

$$t_{bi} = \sqrt{F_{xi}}$$

Если фактическое значение > критического (табличного), параметр b значим
Если фактическое значение < критического (табличного), параметр b не значим

t критическое определяется из таблицы Стьюдента
Номер строки - (n-m-1)
Номер столбца - уровень значимости (обычно 0,05)

tтабл	2,015	из таблицы Стьюдента 46 строка, столбец 0,05
tb1	0,334	< tтабл, значит параметр b1 не значим
tb2	14,844	> tтабл, значит параметр b2 значим
tb3	0,239	< tтабл, значит параметр b3 не значим

Если величина критерия Стьюдента оказывается меньше табличного значения, то дополнительное включение в модель того или иного фактора при соответствующем параметре нецелесообразно.

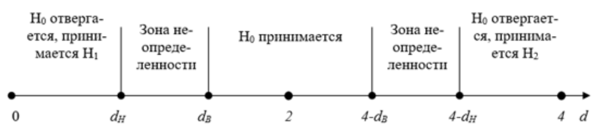
Рис. 3.10. Оценка значимости параметров уравнения с помощью критерия Стьюдента

Так как параметры b_1 и b_3 не значимы, в адекватности модели множественной линейной регрессии есть сомнения.

8. Определить наличие автокорреляции остатков между соседними членами ряда с помощью критерия (теста) Дарбина – Уотсона (рис. 3.11, 3.12).

8) определить наличия автокорреляции остатков между соседними членами ряда с помощью критерия (теста) Дарбина-Уотсона

Остатки ε должны быть случайными.
Однако иногда каждое следующее значение остатков зависит от предшествующих.
В этом случае имеет место автокорреляция в остатках.



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

d	1,600	фактическое значение критерия (теста) Дарбина-Уотсона
		определяется из столбца остатки на листе "Регрессия нет мультик" с помщью "Анализ данных" → "Регрессия"
		созданные регрессионный анализ назвать "Регрессия три х"

гипотеза H0 об отсутствии автокорреляции в остатках
гипотеза H1 – гипотеза о положительной автокорреляции в остатках
гипотеза H2 – гипотеза об отрицательной автокорреляции в остатках.

d табличное определяется из таблицы Дарбина-Уотсона
Столбец - m
Строка - n

dн	1,42	из таблицы Дарбина-Уотсона 50 строка, столбец 3	4-dн	2,58
dв	1,67	из таблицы Дарбина-Уотсона 50 строка, столбец 3	4-dв	2,33

Поскольку фактическое значение d=1,6 попадает в промежуток от 1,42 до 1,67, то нет возможности принять или отклонить ни одну из трех гипотез, поскольку данный промежуток соответствует зоне неопределенности. Значит, с уверенностью в 95 % нельзя утверждать, что имеет или не имеет место положительная или отрицательная автокорреляция в остатках.

Рис. 3.11. Критерий Дарбина – Уотсона

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,995600397
R-квадрат	0,991220151
Нормированный R-квадрат	0,990647552
Стандартная ошибка	26479,32714
Наблюдения	50

Дисперсионный анализ

	df	SS	MS	F	Значимость F
Регрессия	3	3,64129E+12	1,21376E+12	1731,090022	2,74578E-47
Остаток	46	32253119221	701154765,7		
Итого	49	3,67354E+12			

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	ıжние 95,(	Верхние 95,0%
Y-пересечение	11634,71361	9229,250071	1,260634778	0,213797472	-6942,803237	30212,23046	-6942,8	30212,23046
Переменная X 1	-91,90488365	275,2592097	-0,333884863	0,739983819	-645,9729354	462,1631681	-645,973	462,1631681
Переменная X 2	1,748293879	0,117775006	14,84435401	3,67567E-19	1,511225088	1,98536267	1,511225	1,98536267
Переменная X 3	0,138400218	0,579108023	0,238988604	0,812176008	-1,027283773	1,30408421	-1,02728	1,30408421

ВЫВОД ОСТАТКА

Наблюдение	Предсказанное Y	Остатки	$(\varepsilon_t - \varepsilon_{t-1})^2$	ε^2
1	1914476,026	2926,97421		8567178,025
2	347805,2416	78678,75837	5738332803	6190347018
3	381138,2267	-18606,22667	9464368313	346191670,8
4	256799,015	-70099,01505	2651507255	4913871910
5	171214,6022	-13928,60224	3155115274	194005960,4
6	120401,4856	31447,51443	2058991964	988946163,7
7	18100,59574	109339,4043	6067146502	11955105324
8	101567,9776	9717,022409	9924618966	94420524,49

Рис. 3.12. Анализ данных для критерия Дарбина – Уотсона

9. Провести тест Голдфелда – Квандта определения гетероскедастичности остатков (рис. 3.13, 3.14).

9) тест Голфелда-Кванда определения гетероскедастичности остатков

Если для каждого значения фактора остатки имеют одинаковую дисперсию, это называется гомоскедастичной. Если это условие не соблюдается, то имеет место гетероскедастичность.

Упорядочим по х исходные данные по возрастанию фактора котрый имеет наибольшее влияние на у.
Затем упорядочиваем значения у (если одинаковые х) по возрастанию

ryx1	-0,127
ryx2	0,996
ryx3	0,968

Самое сильное влияние имеет х2, так наибольший коэффициент линейной парной корреляции

y	x1	x2	x3
22570	63	8,000	5,040
26388	26	28,000	7,280
16910	36	1169,000	420,840
26015	24	2014,000	483,360
19513	26	2584,000	671,840
22894	37	3254,000	1203,980
46086	1	3599,000	35,990
127440	9	4142,000	372,780
25206	62	5415,000	3357,300
18991	24	6811,000	1634,640
29970	64	7593,000	4859,520
19346	23	8586,000	1974,780
24027	16	10408,000	1665,280

33062	34	15507,000	5272,380
38245	22	18114,000	3985,080
26724	21	18158,000	3813,180
41332	46	18506,000	8512,760
34762	23	19135,000	4401,050
30713	30	20665,000	6199,500
53701	37	23053,000	8529,610
40521	38	23841,000	9059,580
32948	35	24980,000	8743,000
50666	36	25412,000	9148,320
52473	17	28511,000	4846,870
40057	20	29420,000	5884,000
54848	46	32761,000	15070,060
81296	46	33339,000	15335,940
76617	19	43073,000	8183,870
85365	29	45580,000	13218,200
96809	42	46393,000	19485,06
104372	62	48400,000	30008,000
111285	25	51731,000	12932,750
67649	7	60154,000	4210,780
151849	55	62388,000	34313,400
157286	27	90757,000	24504,390
186700	30	138518,000	41555,400
426484	25	189842,000	47460,500
362532	38	207118,000	78704,840
1917403	19	1073255,000	203918,450

Удаляем n/6 наблюдений 8,3 удаляем 8 средних наблюдений

строим 2 регрессионные модели:

21 верхним данным Данные → Анализ данных → Регрессия → Выделяем данные по верхней регрессии и выводим на этот лист на ячейку A396 (не забыть галочку на остатки)

21 нижним данным Данные → Анализ данных → Регрессия → Выделяем данные по нижней регрессии и выводим на этот лист на ячейку A446 (не забыть галочку на остатки)

Рис. 3.13. Выборка для критерия Голдфелда – Квандта

По верхней и нижней регрессии рассчитываем столбец "Остаток²" и находим сумму

S1 8896770984 сумма остатков верхней регрессии

S2 14556966963 сумма остатков нижней регрессии

$F = \dfrac{\max(S1, S2)}{\min(S1, S2)}$ F 1,64 фактическое значение теста Голфелда-Кванда

F табличное определяется из таблицы Фишера

n 21 так как регрессии строились исходя из данных сокращенных

m 3

Номер строки - (n-m-1)

Номер столбца - m

Fкрит 3,2 из таблицы Фишера 17 строка, 3 столбец

Если фактическое значение < критического, то нет наличия гетероскедастичности

Если фактическое значение > критического, то есть гетероскедастичность

F < Fкрит нет наличия гетероскедастичности остатков

Модель можно считать адекватной, если гетероскедастичности нет

Рис. 3.14. Критерий Голдфелда – Квандта

Так как нет наличия гетероскедастичности, то есть смысл в построении уравнения регрессии с целью дальнейшего прогноза.

10. Определить доверительные интервалы оценок коэффициентов регрессии, сделать вывод об их значимости (рис. 3.15).

10) определить доверительные интервалы оценок коэффициентов регрессии, сделать вывод об их значимости.

Доверительные интервалы оценок коэффициентов регрессии скопируем из листа "Регрессия три х"

Если в границы доверительного интервала попадает ноль, т.е. нижняя граница отрицательна,
а верхняя положительна, то оцениваемый параметр принимается нулевым, так как он не может
одновременно принимать и положительное, и отрицательное значения.

	Нижние 95'	Верхние 95%	
Y-пересечение (а)	-6942,80	30212,23	Поскольку интервал включает нулевое значение, делаем вывод о не значимом отличии от нуля праметра а
Переменная X 1 (b1)	-645,97	462,16	Поскольку интервал включает нулевое значение, делаем вывод о не значимом отличии от нуля праметра b1
Переменная X 2 (b2)	1,51	1,99	Поскольку интервал не включает нулевое значение, делаем вывод о значимом отличии от нуля праметра b2
Переменная X 3 (b3)	-1,03	1,30	Поскольку интервал включает нулевое значение, делаем вывод о не значимом отличии от нуля праметра b3

Рис. 3.15. Доверительные интервалы оценок коэффициентов регрессии

Подводя итог, можно сделать вывод, что модель является ненадежной, так как присутствуют незначимые факторы и незначимые параметры.

Контрольные вопросы

1. Какой круг проблем необходимо охватить при спецификации модели множественной регрессии?
2. Как производится отбор факторов при построении множественной регрессии?
3. Что такое «мультиколлинеарность» и как доказать ее наличие в модели?
4. Какой метод применяется для оценки параметров уравнения множественной регрессии?
5. Назовите вид уравнения множественной регрессии в стандартизованной форме и его преимущества.
6. Что представляют собой частные уравнения регрессии?
7. Что характеризует коэффициент множественной корреляции?
8. Перечислите коэффициенты частной корреляции, методы расчета.
9. Как проводится оценка надежности результатов множественной регрессии?

Тест

1. Факторы, включаемые в модель множественной регрессии должны отвечать требованиям:

а) они должны агрегировать друг с другом;

б) каждый фактор должен быть достаточно тесно связан с результатом;

в) факторы не должны сильно коррелировать друг с другом;

г) отвечать спецификации модели.

2. К методам построения уравнения регрессии относятся:

а) метод наименьших квадратов;

б) метод стандартизованных коэффициентов;

в) шаговый регрессионный анализ.

3. Мультиколлинеарность факторов возникает тогда, когда:

а) более чем два фактора связаны между собой линейной зависимостью;

б) F-критерий Фишера меньше критического значения;

в) имеет место совокупное воздействие факторов друг на друга.

4. К последствиям мультиколлинеарности относятся:

а) неправильная функциональная спецификация модели;

б) становится невозможно определить изолированное влияние факторов на результативный показатель;

в) факторы измеряются в разных единицах.

5. В случае наличия мультиколлинеарности:

а) необходимо провести дисперсионный анализ;

б) проверить нулевую гипотезу;

в) из модели исключается фактор, который теснее связан с другими факторами модели.

6. Оценка параметров множественной регрессионной модели проводится с помощью:

а) метода включения;

б) метода стандартизованных коэффициентов;

в) метода максимального правдоподобия;

г) метода наименьших квадратов.

7. При решении вопроса о включении в модель дополнительного фактора используются методы:

а) метод скользящих средних;

б) метод анализа остаточной и факторной дисперсий;

в) метод стандартизованных коэффициентов;

г) метод автокорреляции в остатках.

8. К свойствам матрицы парных коэффициентов корреляции относятся:

а) определитель всегда отличен от нуля;

б) строится только для линейной регрессии;

в) она симметрична относительно главной диагонали;

г) по диагонали располагаются единицы.

9. Показатель, характеризующий влияние отдельного фактора на результативный признак при неизменном уровне других факторов:

а) частное уравнение регрессии;

б) индекс детерминации;

в) коэффициент частной корреляции.

10. В случае, если частный F-критерий Фишера больше табличного значения:

а) принимается нулевая гипотеза о статистической незначимости уравнения в целом;

б) делается вывод о нецелесообразности включения в модель определенного фактора;

в) отвергается нулевая гипотеза о статистической незначимости отдельного фактора;

г) отвергается нулевая гипотеза о статистической незначимости уравнения в целом.

4. СИСТЕМЫ ЭКОНОМЕТРИЧЕСКИХ УРАВНЕНИЙ

При использовании отдельных уравнений регрессии предполагается, что факторы можно изменять независимо друг от друга. Однако практически изменение одной переменной не может происходить при абсолютной неизменности других. Следовательно, отдельно взятое уравнение множественной регрессии не может характеризовать истинное влияние отдельных признаков на вариацию результатирующей переменной. Это означает, что рассматриваемая модель не полна: ее следует дополнить уравнениями, в которых результатирующими переменными выступали бы сами факторы. Таким образом, возникает необходимость рассмотрения системы одновременных или регрессионных уравнений.

Наибольшее распространение в эконометрических расчетах 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.

$$\begin{cases} y_1 = b_{12}y_2 + b_{13}y_3 + \ldots + b_{1n}y_n + a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m + \varepsilon_1 \\ y_2 = b_{21}y_1 + b_{23}y_3 + \ldots + b_{2n}y_n + a_{21}x_1 + a_{22}x_2 + \ldots + a_{2m}x_m + \varepsilon_2 \\ \ldots \\ y_n = b_n y_1 + b_{n2}y_2 + \ldots + b_{nn-1}y_{n-1} + a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m + \varepsilon_n \end{cases}. \quad (4.1)$$

Эта система уравнений называется также структурной формой модели.

Выделяют эндогенные и экзогенные переменные. Эндогенные часто обозначаются через y. Это зависимые переменные, число которых равно числу уравнений в системе. Экзогенные обозначаются через x. Это предопределенные переменные, влияющие на эндогенные переменные, но не зависящие от них.

Структурная форма модели в правой части содержит при эндогенных и экзогенных переменных коэффициенты b_i и a_j, которые называются структурными коэффициентами модели. Для определения структурных коэффициентов модели структурная форма модели преобразуется в приведенную форму. Приведенная форма модели – система линейных функций эндогенных переменных от экзогенных.

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \ldots + \delta_{1m}x_m \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \ldots + \delta_{2m}x_m \\ \ldots \\ y_n = \delta_{n1}x_1 + \delta_{n2}x_2 + \ldots + \delta_{nm}x_m \end{cases}, \quad (4.2)$$

где δ_i – коэффициенты приведенной формы модели.

Применяя МНК, можно оценить δ, а затем оценить значения эндогенных переменных через экзогенные.

Для структурной модели вида $\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 \\ y_2 = b_{21}y_1 + a_{22}x_2 \end{cases}$ приведенная форма $\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 \end{cases}$, $y_2 = \frac{y_1 - a_{11}x_1}{b_{12}}$. Здесь y_2 выражено из первого уравнения модели.

Для первого уравнения $\delta_{11} = \frac{a_{11}}{1 - b_{12}b_{21}}$ и $\delta_{12} = \frac{a_{22}b_{12}}{1 - b_{12}b_{21}}$,

для второго уравнения $\delta_{21} = \frac{a_{11}b_{21}}{1 - b_{12}b_{21}}$ и $\delta_{22} = \frac{a_{22}}{1 - b_{12}b_{21}}$.

Приведенная форма модели позволяет получить значения эндогенной переменной через значения экзогенных переменных. Однако она аналитически уступает структурной форме модели, так как в ней отсутствуют оценки взаимосвязи между эндогенными переменными.

Проблема идентификации. Идентификация – единственность соответствия между приведенной и структурной формами модели.

Структурная модель в полном виде, состоящая в каждом уравнении системы из n эндогенных и m экзогенных переменных, содержит $n(n - 1 + m)$ параметров. Приведенная форма модели в полном виде содержит nm параметров. В полном виде структурная модель содержит большее число параметров, чем приведенная форма модели. То есть $n(n - 1 + m)$ параметров структурной модели не могут быть однозначно определены из nm параметров приведенной формы. Чтобы получить единственно возможное решение, необходимо предположить, что некоторые из структурных коэффициентов равны нулю (уменьшится число структурных коэффициентов).

Структурные модели можно подразделить на:

– идентифицируемые;

– неидентифицируемые;

– сверхидентифицируемые.

Модель идентифицируема, если все структурные коэффициенты определяются однозначно по коэффициентам приведенной формы модели, т. е. если число параметров структурной модели равно числу параметров приведенной модели.

Модель неидентифицируема, если число приведенных коэффициентов меньше числа структурных коэффициентов, и значит структурные коэффициенты не могут быть оценены через коэффициенты приведенной формы модели.

Модель сверхидентифицируема, если число приведенных коэффициентов больше числа структурных коэффициентов, в этом случае на основе коэффициентов приведенной формы можно получить два или более значений одного структурного коэффициента.

Модель считается идентифицируемой, если каждое уравнение системы идентифицируемо. Выполнение условия идентифицируемости модели проверяется для каждого уравнения системы.

Правило определения идентификации уравнения:

$D + 1 = H$ – уравнение идентифицируемо;

$D + 1 < H$ – уравнение неидентифицируемо;

$D + 1 > H$ – уравнение сверхидентифицируемо,

где D – число экзогенных переменных, которые содержатся в системе, но не входят в данное уравнение, H – число эндогенных переменных в i-ом уравнении системы.

Рассмотренное правило представляет необходимое, но не достаточное условие идентификации. Более точно условия идентификации определяются, если накладываются ограничения на коэффициенты матриц параметров структурной модели. Уравнение идентифицируемо, если по отсутствующим в нем переменным (эндогенным и экзогенным) можно из коэффициентов при них в других уравнениях системы получить матрицу, определитель которой не равен нулю, а ранг матрицы не меньше, чем число эндогенных переменных в системе без одного.

Для решения идентифицируемого уравнения применяется косвенный метод наименьших квадратов (КМНК), для решения сверхидентифицируемого – двухшаговый метод наименьших квадратов (ДМНК).

Алгоритм *косвенного метода наименьших квадратов* состоит из двух этапов.

Этап 1. Составляют приведенную форму модели и определяют численные значения параметров каждого ее уравнения обычным МНК.

Этап 2. Путем алгебраических преобразований переходят от приведенной формы к уравнениям структурной формы модели, получая тем самым численные оценки структурных параметров.

Двухшаговый метод наименьших квадратов состоит из выполнения трех этапов.

Этап 1. Составляют приведенную форму модели и определяют численные значения параметров каждого ее уравнения обычным МНК (этот этап полностью совпадает с первым этапом КМНК).

Этап 2. Выявляют эндогенные переменные, находящиеся в правой части структурного уравнения, и находят расчетные значения по соответствующим уравнениям приведенной формы модели.

Этап 3. Обычным МНК определяют параметры структурного уравнения, используя в качестве исходных данных фактические значения экзогенных переменных, и расчетные значения эндогенных переменных, стоящих в правой части данного структурного уравнения.

Задача 4

Имеются данные различных показателей экономической безопасности предприятия за семь последовательных лет (табл. 4.1). Необходимо построить модель функциональной зависимости темпов прироста коэффициентов финансовой безопасности, финансовой устойчивости и финансовой зависимости не только от трех независимых факторов (доли заемного капитала, доли собственного капитала и текущей ликвидности), но и отражающую влияние зависимых факторов друг на друга. Для решение подобной задачи оптимальным является построение системы эконометрических уравнений.

Таблица 4.1

Исходные данные задачи 4

№	Темп прироста					Текущая ликвидность x_1
	коэффициента			доли		
	финансовой безопасности, y_1	финансовой устойчивости, y_2	финансовой зависимости, y_3	заемного капитала x_2	Собственного капитала x_3	
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	1	5

Возможно построение двух видов систем. Для первой необходимо определить параметры структурной модели вида:

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

Решение.

Определение идентификации.

1-е уравнение $y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2$

H = 2 (y_1, y_2) (число эндогенных переменных в уравнении);

D = 1 (x_3) (число экзогенных переменных, отсутствующих в уравнении).

Необходимое условие: D + 1 = H, 1 + 1 = 2 – уравнение идентифицируемо.

Однозначный ответ на вопрос идентификации уравнения позволяет получить достаточное условие.

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

Построим матрицу.

Уравнение	Отсутствующие переменные	
	y_3	x_3
Второе	0	a_{23}
Третье	–1	a_{33}

Det $A = a_{33\_}a_{23} * 0 * (-1) = a_{23}$, если коэффициент a_{23} будет отличен от нуля, то достаточное условие выполняется, поскольку определитель матрицы отличен от нуля, и ее ранг равен числу эндогенных переменных (3 = y_1, y_2, y_3) без одного: rang A = 3 – 1 = 2.

Вывод: первое уравнение точно идентифицируемо. Значит, для определения структурных коэффициентов этого уравнения необходимо применить косвенный метод наименьших квадратов.

2-е уравнение $y_2 = b_{21}y_1 + b_{22}x_2 + a_{23}x_3$.

Необходимое условие: H = 2, D = 1, D + 1 = H – уравнение идентифицируемо.

Для проверки достаточного условия составим матрицу

Уравнение	Отсутствующие переменные	
	y_3	x_1
Первое	0	a_{11}
Третье	–1	a_{31}

$\mathrm{Det}\, A = 0 * a_{31} - a_{11} * (-1) = a_{11}$, если a_{11} будет отличен от нуля, то достаточное условие выполняется, поскольку определитель матрицы отличен от нуля, и ее ранг равен числу эндогенных переменных (3 = y_1, y_2, y_3) без одного: rang A = 3 – 1 = 2.

Вывод: второе уравнение точно идентифицируемо. Значит, для определения структурных коэффициентов этого уравнения необходимо применить косвенный метод наименьших квадратов.

3-е уравнение $y_3 = b_{31}y_1 + a_{31}x_1 + a_{33}x_3$.

Необходимое условие: H = 2, D = 1, D + 1 = H – уравнение идентифицируемо.

Достаточное условие проверяется по матрице

Уравнение	Отсутствующие переменные	
	y_2	x_2
Первое	0	a_{12}
Второе	–1	a_{22}

$\mathrm{Det}\, A = 0 * a_{22} - a_{12} * (-1) = a_{12}$, если a_{12} будет отличен от нуля, то достаточное условие выполняется, поскольку определитель матрицы отличен от нуля, и ее ранг равен числу эндогенных переменных $(3 = y_1, y_2, y_3)$ без одного: rang A = 3 – 1 = 2.

Вывод: третье уравнение точно идентифицируемо. Таким образом, для определения структурных коэффициентов этого уравнения необходимо применить косвенный метод наименьших квадратов.

Поскольку все три уравнения идентифицируемы, то и вся система идентифицируема.

Определение структурных коэффициентов. Для всех трех уравнений системы применяем косвенный метод наименьших квадратов.

Этап 1. Составляют приведенную форму модели и определяют численные значения параметров каждого ее уравнения обычным МНК.

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \ldots + \delta_{13}x_3 \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \ldots + \delta_{23}x_3 \\ y_3 = \delta_{31}x_1 + \delta_{32}x_2 + \ldots + \delta_{33}x_3 \end{cases}$$ – приведенная форма данной модели.

$$\begin{cases} y_1 = -1{,}099 + 1{,}482x_1 + 0{,}604x_2 + 0{,}146x_3 \\ y_2 = 5{,}61 + 1{,}126x_1 - 0{,}44x_2 - 0{,}28x_3 \\ y_3 = 4{,}467 + 1{,}796x_1 + 1{,}61x_2 + 0{,}076x_3 \end{cases}$$ – полученная с помощью обычного МНК приведенная форма с приведенными коэффициентами.

Этап 2. Путем алгебраических преобразований переходят от приведенной формы к уравнениям структурной формы модели, получая тем самым численные оценки структурных параметров.

Для того чтобы преобразовать приведенную форму в структурную, необходимо определить, какие переменные входят в каждое уравнение структурной формы.

$$\begin{cases} y_1 = b_{12}y_2 + a_{11}x_1 + a_{12}x_2 & y_1, y_2, x_1, x_2 \\ y_2 = b_{21}y_1 + a_{22}x_2 + a_{23}x_3 \text{ нужны } & y_1, y_2, x_2, x_3. \\ y_3 = b_{31}y_1 + a_{31}x_1 + a_{33}x_3 & y_1, y_3, x_1, x_3 \end{cases}$$

Выразим из первого уравнения приведенной формы:

$$x_1 = \frac{y_1 + 1{,}099 - 0{,}604x_2 - 0{,}146x_3}{1{,}482}.$$

Это выражение содержит y_1, x_2, x_3, его целесообразно подставить во второе приведенное уравнение:

$$y_2 = 5{,}61 + 1{,}126\frac{y_1 + 1{,}099 - 0{,}604x_2 - 0{,}146x_3}{1{,}482} - 0{,}44x_2 - 0{,}28x_3 = \\ = 6{,}455 + 0{,}76y_1 - 0{,}899x_2 - 0{,}391x_3.$$

Выразим из второго уравнения приведенной формы:

$$x_3 = \frac{5{,}61 + 1{,}126x_1 - 0{,}44x_2 - y_2}{0{,}28}.$$

Это выражение содержит y_2, x_1, x_2, его целесообразно подставить в первое приведенное уравнение:

$$y_1 = -1{,}099 + 1{,}482x_1 + 0{,}604x_2 + 0{,}146\frac{5{,}61 + 1{,}126x_1 - 0{,}44x_2 - y_2}{0{,}28} = \\ = 1{,}826 - 0{,}521y_2 + 2{,}069x_1 + 0{,}375x_2.$$

Выразим из первого уравнения приведенной формы:

$$x_2 = \frac{y_1 + 1{,}099 - 1{,}482x_1 - 0{,}146x_3}{0{,}604}.$$

Это выражение содержит y_1, x_1, x_3, его целесообразно подставить в третье приведенное уравнение:

$$y_3 = 4{,}467 + 1{,}796x_1 + 1{,}61\,\frac{y_1 + 1{,}099 - 1{,}482x_1 - 0{,}146x_3}{0{,}604} + 0{,}076x_3 =$$
$$= 7{,}396 + 2{,}665y_1 - 1{,}05x_1 - 0{,}313x_3.$$

Структурная форма модели:

$$\begin{cases} y_1 = 1{,}826 - 0{,}521y_2 + 2{,}069x_1 + 0{,}375x_2 \\ y_2 = 6{,}455 + 0{,}76y_1 - 0{,}899x_2 - 0{,}391x_3 \\ y_3 = 7{,}396 + 2{,}665y_1 - 1{,}05x_1 - 0{,}313x_3 \end{cases}.$$

В зависимости от целей исследования можно построить структурную модель второго вида

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

Поскольку первые два уравнения анализируемой системы идентичны первым двум уравнениям системы в задаче выше, то необходимо проверить идентификацию только третьего уравнения.

3-е уравнение $y_3 = b_{31}y_1 + a_{33}x_3$.

Необходимое условие: H = 2, D = 2, D + 1 > H – уравнение сверхидентифицируемо.

Для достаточного условия построим матрицу

Уравнение	Отсутствующие переменные		
	y_2	x_1	x_2
Первое	b_{12}	a_{11}	a_{12}
Второе	–1	0	a_{22}

$\mathrm{Det}\,A = b_{12} * a_{22} - a_{12} * (-1)$, если коэффициенты будут отличны от нуля, то достаточное условие выполняется, поскольку определитель матрицы отличен от нуля, и ее ранг равен числу эндогенных переменных (3 = y_1, y_2, y_3) без одного: rang A = 3 – 1 = 2.

Вывод: третье уравнение сверхидентифицируемо.

Значит, для определения структурных коэффициентов этого уравнения необходимо применить двухшаговый метод наименьших квадратов.

Этап 1. Составляют приведенную форму модели и определяют численные значения параметров каждого ее уравнения обычным МНК.

$$\begin{cases} y_1 = \delta_{11}x_1 + \delta_{12}x_2 + \delta_{13}x_3 \\ y_2 = \delta_{21}x_1 + \delta_{22}x_2 + \delta_{23}x_3 \\ y_3 = \delta_{31}x_1 + \delta_{32}x_2 + \delta_{33}x_3 \end{cases}$$ – приведенная форма данной модели.

$$\begin{cases} y_1 = -1{,}099 + 1{,}482x_1 + 0{,}604x_2 + 0{,}146x_3 \\ y_2 = 5{,}61 + 1{,}126x_1 - 0{,}44x_2 - 0{,}28x_3 \\ y_3 = 4{,}467 + 1{,}796x_1 + 1{,}61x_2 + 0{,}076x_3 \end{cases}$$ – полученная с помощью МНК приведенная форма с приведенными коэффициентами.

Этап 2. Выявляют эндогенные переменные, находящиеся в правой части структурного уравнения, и находят расчетные значения по соответствующим уравнениям приведенной формы модели.

В третьем уравнении в правой части структурного уравнения находится одна эндогенная переменная – y_1.

Найдем ее расчетные значения по уравнению регрессии $y_1 = -1{,}099 + 1{,}482x_1 + 0{,}604x_2 + 0{,}146x_3$.

$y_1 = 1{,}737$, $y_2 = 3{,}969$, $y_3 = 4{,}681$, $y_4 = 4{,}719$, $y_5 = 4{,}993$ $y_6 = 6{,}621$, $y_7 = 8{,}269$.

Этап 3. Обычным МНК определяют параметры структурного уравнения, используя в качестве исходных данных фактические значения экзогенных переменных и расчетные значения эндогенных переменных, стоящих в правой части данного структурного уравнения.

В нашей задаче решается уравнение вида $y_3 = \alpha + \beta_1 y_1 + \beta_2 x_3$ по данным представленным в табл. 4.2

Таблица 4.2

Исходные данные для решения уравнения

№	y_3	y_1	x_3
1	10	1,737	1
2	12	3,969	2
3	11	4,681	5
4	15	4,719	3
5	14	4,993	3
6	16	6,621	4
7	18	8,269	1

Полученное уравнение: $y_3 = 8{,}241 + 1{,}304$.

Структурные параметры второго и первого уравнений определяются КМНК.

Система эконометрических уравнений в структурной форме имеет вид

$$\begin{cases} y_1 = 1{,}826 - 0{,}521y_2 + 2{,}069x_1 + 0{,}375x_2 \\ y_2 = 6{,}455 + 0{,}76y_1 - 0{,}899x_2 - 0{,}391x_3 \\ y_3 = 8{,}241 + 1{,}304y_1 - 0{,}386x_3 \end{cases}$$

Полученная система эконометрических уравнений позволяет проанализировать функциональную зависимость темпов прироста коэффициентов финансовой безопасности, финансовой устойчивости и финансовой зависимости не только от доли заемного капитала, доли собственного капитала и текущей ликвидности, но и от взаимного влияния зависимых факторов друг на друга, поскольку они являются независимыми переменными в отдельных уравнениях системы.

Контрольные вопросы

1. Что представляют собой системы уравнений, используемые в эконометрических расчетах?
2. Назовите структурную и приведенную формы модели эконометрических уравнений.
3. В чем заключается проблема идентификации систем эконометрических уравнений?
4. Как проводится процедура оценивания параметров в структурной модели эконометрических уравнений?
5. Перечислите этапы косвенного метода наименьших квадратов и двухшагового метода наименьших квадратов.

Тест

1. С помощью какого метода производится оценка параметров структурной сверхидентифицированной системы эконометрических уравнений:

а) МНК;

б) шаговый регрессионный анализ;

в) КМНК;

г) ДМНК.

2. Зависимые переменные в системе эконометрических уравнений называются:

а) структурными;

б) приведенными;

в) эндогенными;

г) экзогенными;

д) идентифицируемыми.

3. Достаточное условие идентифицируемости уравнения выполняется, если:

а) выполняется необходимое условие идентифицируемости;

б) количество эндогенных переменных равно количеству экзогенных;

в) число экзогенных переменных, которые содержатся в системе, но не входят в данное уравнение, равно числу эндогенных переменных в этом уравнении;

г) по отсутствующим в уравнении переменным можно из коэффициентов при них в других уравнениях системы получить матрицу, определитель которой не равен нулю, а ранг матрицы не меньше, чем число эндогенных переменных в системе без одного.

4. Косвенный метод наименьших квадратов используется для:

а) определения приведенных коэффициентов сверхидентифицированной системы;

б) определения структурных коэффициентов сверхидентифицированной системы;

в) проверки идентифицируемости системы;

г) определения приведенных коэффициентов идентифицированной системы;

д) определения структурных коэффициентов идентифицированной системы.

5. Система эконометрических уравнений неидентифицируема, если:

а) хотя бы одно уравнение системы сверхидентифицируемо;

б) не выполняется необходимое условие идентифицируемости;

в) все уравнения системы неидентифицируемы;

г) число эндогенных переменных в системе больше числа экзогенных переменных;

д) хотя бы одно уравнение системы неидентифицируемо.

5. МОДЕЛИРОВАНИЕ ОДНОМЕРНЫХ ВРЕМЕННЫХ РЯДОВ

При рассмотрении классической модели регрессии характер экспериментальных данных, как правило, не имеет принципиального значения. Однако это становится неверно, если условия классической модели нарушены.

Можно построить эконометрическую модель, используя два типа исходных данных:

– данные, характеризующие совокупность различных объектов в определенный момент (период) времени. Модели, построенные по данным этого типа, называются пространственными моделями;

– данные, характеризующие один объект за ряд последовательных моментов (периодов) времени. Модели, построенные по данным этого типа, – модели временных рядов.

Например, временными рядами будут серия ежедневных наблюдений в течение некоторого периода за ценами товара при закрытии торгов на бирже, дневные объемы выпуска товара, месячные показатели инфляции или индекса потребительских цен, ежеквартальные оценки валового национального продукта (принятые в США) или средних зарплат (принятые в России для ежеквартального индексирования пенсий), ежегодные данные об объеме, выручке и прибыли компании. Временные ряды, естественно, не ограничиваются исключительно экономическими величинами; известно их использование при анализе процессов в энергосистемах, атомной промышленности, химических и нефтехимических производствах, причем в этом случае часто используются более мелкие дискретности времени, чем в экономике – минуты и даже секунды при обработке данных о быстропротекающих процессах в атомной энергетике или при исследовании переходных процессов в химической кинетике. Известно даже успешное применение анализа временных рядов при слежении за подводными лодками «вероятного противника» в 70–80-е гг. и при обработке данных наблюдений в системах ПВО, при прогнозах проходимости радиосигналов в атмосфере и ионосфере, при моделировании транспортных потоков на автотрассах.

Основным положением, на котором базируется использование временных рядов для прогнозирования, является то, что факторы, влияющие на отклик изучаемой системы, действовали некоторым образом в прошлом и настоящем, и ожидается, что они будут действовать схожим образом и в не слишком далеком будущем. Поэтому основной целью анализа временных рядов будет оценка и вычленение этих влияющих факторов с целью прогноза дальнейшего поведения системы и выработки рациональных управленческих решений.

Методы исследования пространственных моделей и моделей временных рядов существенно отличаются, так как, в отличие от пространственных выборок, наблюдения во временных рядах нельзя считать независимыми.

Под временным рядом (динамическим рядом или рядом динамики) подразумевается последовательность наблюдений некоторого признака (случайной величины) Y в последовательные моменты времени. Отдельные наблюдения называются уровнями ряда, которые будем обозначать y_t $(t = 1,2,\ldots,n)$, где n – число уровней.

Цели анализа временных рядов могут быть различными. Можно, например, стремиться предсказать будущее на основании знаний прошлого, пытаться выяснить механизм, лежащий в основе процесса, и управлять им. Необходимо уметь освобождать временной ряд от компонент, которые затемняют его динамику. Часто требуется сжато представлять характерные особенности ряда.

Основной чертой, выделяющей анализ временных рядов среди других видов статистического анализа, является существенность порядка, в котором производятся наблюдения.

Различают два вида временных рядов.

Измерение некоторых величин (температуры, напряжения и т. д.) производится непрерывно, по крайней мере, теоретически. При этом наблюдения можно фиксировать в виде графика. Но даже в том случае, когда изучаемые величины регистрируются непрерывно, практически при их обработке используются только те значения, которые соответствуют дискретному множеству моментов времени. Следовательно, если время измеряется непрерывно, временной ряд называется непрерывным, если же время фиксируется дискретно, т. е. через фиксированный интервал времени, то временной ряд дискретен.

В дальнейшем мы будем иметь дело только с дискретными временными рядами.

Дискретные временные ряды получаются двумя способами:

– выборкой из непрерывных временных рядов через регулярные промежутки времени (например, численность населения, величина собственного капитала фирмы, объем денежной массы, курс акции), такие временные ряды называются моментными;

– накоплением переменной в течение некоторого периода времени (например, объем производства какого-либо вида продукции, количество осадков, объем импорта), в этом случае временные ряды называются интервальными.

В эконометрике принято моделировать временной ряд как случайный процесс, называемый также стохастическим процессом, под которым понимается статистическое явление, развивающееся во времени согласно законам теории вероятностей.

Случайный процесс – это случайная последовательность. Обычно предполагают, что эта последовательность идет от минус до плюс бесконечности. Обычно стоит задача по данному ряду сделать какие-то заключения о свойствах лежащего в его основе случайного процесса, оценить параметры, сделать прогнозы и т. п. В литературе по временным рядам существует некоторая неоднозначность, и иногда временным рядом называют сам случайный процесс, а иногда статистическую модель, которая порождает данный случайный процесс. В дальнейшем мы не будем в явном виде посредством особых обозначений различать случайный процесс и его реализацию.

Для удобства можно провести классификацию случайных процессов и соответствующих им временных рядов на детерминированные и случайные процессы (временные ряды). Детерминированным называют процесс, который принимает заданное значение с вероятностью единица. Когда же мы будем говорить о случайном процессе и случайном временном ряде, то, как правило, будем подразумевать, что он не является детерминированным.

Стохастические процессы подразделяются на стационарные и нестационарные. Стохастический процесс является стационарным, если он находится в статистическом равновесии, т. е. его свойства с вероятностной точки зрения не зависят от времени. Процесс нестационарен, если эти условия нарушаются.

Важное теоретическое значение имеют гауссовские процессы. Это такие процессы, в которых любой набор наблюдений имеет совместное нормальное распределение. Как правило, термин «временной ряд» сам по себе подразумевает, что этот ряд является одномерным (скалярным).

При анализе экономических временных рядов традиционно различают разные виды эволюции (динамики). Эти виды динамики могут, вообще говоря, комбинироваться. Тем самым задается разложение временного ряда на составляющие, которые с экономической точки зрения несут разную содержательную нагрузку.

Перечислим наиболее важные:

– тенденция – соответствует медленному изменению, происходящему в некотором направлении, которое сохраняется в течение значительного промежутка времени (например, рост населения, экономическое развитие, изменение структуры потребления и т. д.). Тенденцию называют также трендом или долговременным движением;

– циклические колебания – это более быстрая, чем тенденция, динамика, в которой есть фаза возрастания и фаза убывания (например, влияние волн экономической активности Кондратьева, демографических ям, циклов солнечной активности). Обычно эта компонента может изменяться по длине периода и своей интенсивности и хорошо коррелирует с циклом деловой активности. На подъеме деловой активности значения отклика оказываются выше чисто трендовых, а в периоды спада и стагнации оказываются заметно ниже ожидаемых по тренду;

– сезонные колебания – соответствуют изменениям, которые происходят регулярно в течение года, недели или суток. Эта компонента временного ряда определяет короткопериодические колебания, связанные именно с изменениями внутригодовой активности, и повторяющиеся через более или менее фиксированные моменты времени; отслежены они, естественно, могут быть при ежеквартальных, ежемесячных и более частых наблюдениях. Естественно связать сезонную компоненту с влиянием традиций (сезонные и рождественские распродажи), социальных привычек (высокая активность в курортном бизнесе в летнее время и существование «мертвых сезонов» в иные периоды) и даже плохо предсказуемой погоды (продажи мороженого и прохладительных напитков, деятельность горнолыжных курортов);

– календарные эффекты – это отклонения, связанные с определенными предсказуемыми календарными событиями – такими, как праздничные дни, количество рабочих дней за месяц, високосность года и т. п.;

– случайные флуктуации – беспорядочные движения относительно большой частоты. Они порождаются влиянием разнородных событий на изучаемую величину (несистематический или случайный эффект). Часто такую составляющую называют шумом (этот термин пришел из технических приложений);

– выбросы – это аномальные движения временного ряда, связанные с редко происходящими событиями, которые резко, но лишь очень кратковременно отклоняют ряд от общего закона, по которому он движется;

– структурные сдвиги – это аномальные движения временного ряда, связанные с редко происходящими событиями, имеющие скачкообразный характер и меняющие тенденцию.

Некоторые экономические ряды можно считать представляющими те или иные виды таких движений почти в чистом виде. Но большая часть их имеет очень сложный вид. В них могут проявляться, например, как общая тенденция возрастания, так и сезонные изменения, на которые могут накладываться случайные флуктуации. Часто для анализа временных рядов оказывается полезным изолированное рассмотрение отдельных компонент.

В общем виде при исследовании временного ряда наибольшее значение имеют следующие составляющие:

$$y_t = u_t + v_t + c_t + \varepsilon_t, \qquad (t = 1,2,\dots,n), \tag{5.1}$$

где u_t – тренд, плавно меняющаяся компонента, описывающая чистое влияние долговременных факторов, длительную тенденцию изменения признака; v_t – сезонная компонента, отражающая повторяемость экономических процессов в течение не очень длительного периода; c_t – циклическая компонента, отражающая повторяемость экономических процессов в течение длительных периодов; ε_t – случайная компонента, отражающая влияние не поддающихся учету и регистрации случайных факторов. В терминах статистики эту компоненту можно считать ошибкой наблюдения и обрабатывать аналогично случайным ошибкам измерений в статистике. Связывают ее обычно со случайными явлениями внешнего мира – ураганами, наводнениями, забастовками, влиянием политических процессов, таких как выборы или неопределенность их исхода, переворотами и мятежами.

В отличие от ε_t, первые три составляющие временного ряда u_t, v_t, c_t являются закономерными.

Модель, в которой временной ряд представлен как сумма перечисленных составляющих, называется аддитивной моделью временного ряда

$$y_t = u_t + v_t + c_t + \varepsilon_t, \quad (t = 1,2,\dots,n). \tag{5.2}$$

Модель, в которой временной ряд представлен как произведение компонент, – мультипликативная модель

$$y_t = u_t \cdot v_t \cdot c_t \cdot \varepsilon_t, \quad (t = 1,2 \dots,n). \tag{5.3}$$

Важнейшей классической задачей при исследовании экономических временных рядов является выявление и статистическая оценка основной тенденции развития изучаемого процесса и отклонений от нее.

Основные этапы анализа временных рядов:

– графическое представление и описание поведения временного ряда;

– выделение и удаление закономерных (неслучайных) составляющих временного ряда;

– сглаживание и фильтрация (удаление низко- или высокочастотных составляющих временного ряда);

– исследование случайной составляющей временного ряда, построение и проверка адекватности математической модели для ее описания;

– прогнозирование развития изучаемого процесса на основе изучаемого временного ряда;

– исследование взаимосвязи между различными временными рядами.

Автокорреляция уровней временного ряда. При наличии во временном ряде тенденции и циклических колебаний значения каждого последующего уровня ряда зависят от предыдущих. Степень тесноты связи между последовательными уровнями временных рядов $y_1, y_2,\dots, y_n$ и $y_{1-\tau}, y_{2-\tau},\dots, y_{n-\tau}$ (временной ряд, состоящий из уровней исходного временного ряда, сдвинутых на τ единиц или с лагом τ) может быть определена с помощью линейного коэффициента корреляции. Поскольку такой линейный коэффициент корреляции измеряет корреляцию между уровнями одного и того же ряда, сдвинутыми относительно друг друга, его называют коэффициентом автокорреляции. Статистической оценкой линейного коэффициента автокорреляции является выборочный коэффициент автокорреляции

$$r(\tau)=\frac{(n-\tau)\sum_{t=\tau+1}^{n}y_t y_{t-\tau}-\sum_{t=\tau+1}^{n}y_t\sum_{t=\tau+1}^{n}y_{t-\tau}}{\sqrt{(n-\tau)\sum_{t=\tau+1}^{n}y_t^2-\left(\sum_{t=\tau+1}^{n}y_t\right)^2}\sqrt{(n-\tau)\sum_{t=\tau+1}^{n}y_{t-\tau}^2-\left(\sum_{t=\tau+1}^{n}y_{t-\tau}\right)^2}}. \quad (5.4)$$

Свойства коэффициента автокорреляции:

1) 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 временных рядов, имеющих сильную нелинейную тенденцию, коэффициент автокорреляции уровней исходного ряда может приближаться к нулю;

2) 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.

При расчете коэффициента автокорреляции следует помнить, что с увеличением τ число $(n-\tau)$ пар наблюдений y_t, $y_{t-\tau}$ уменьшается, поэтому лаг τ должен быть таким, чтобы число $(n-\tau)$ было достаточным. Обычно ориентируются на соотношение

$$\tau \le \frac{n}{4}. \quad (5.5)$$

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

Анализ автокорреляционной функции и коррелограммы позволяет определить 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 τ, ряд содержит циклические колебания с периодичностью в τ моментов времени. Если ни один из коэффициентов автокорреляции не является значимым, можно сделать одно из предположений:

– ряд не содержит ни тенденции, ни циклические колебания;

– ряд содержит сильную нелинейную тенденцию.

Моделирование тенденций временного ряда. Одной из важнейших задач исследования экономического временного ряда является выявление основной тенденции изучаемого процесса, выраженной неслучайной составляющей (тренда или тренда с циклической или/и сезонной компонентой). Одним из способов моделирования тенденций временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени или тренда. Этот способ называется аналитическим выравниванием временного ряда.

Для построения трендов чаще всего применяют следующие функции:

– линейный тренд $y_t = a + bt$;

– гипербола $y_t = a + \frac{b}{t}$;

– экспоненциальный тренд $y_t = e^{a+bt}$;

– тренд в форме степенной функции $y_t = at^b$;

– парабола второго и более высоких порядков.

$$y_t = a + b_1 t + b_2 t^2 + \dots + b_k t^k.$$

При выборе соответствующей функции используют содержательный анализ и визуальные наблюдения (на основе графического изображения временного ряда).

Из двух функций предпочтение обычно отдается той, при которой меньше сумма квадратов отклонений фактических данных от расчетных на основе этих функций. Но этот принцип нельзя доводить до абсурда: для любого ряда из n точек можно подобрать полином $(n-1)$-й степени, проходящий через все точки и, соответственно, с минимальной – нулевой – суммой квадратов отклонений. Поэтому при прочих равных условиях предпочтение следует отдавать более простым функциям.

В этих же целях можно использовать и коэффициенты автокорреляции уровней ряда. Тип тенденции можно определить путем сравнения коэффициентов автокорреляции первого порядка, рассчитанных по исходным и преобразованным уровням 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 исходным уровням ряда.

Выбор наилучшего уравнения в случае, если ряд содержит нелинейную тенденцию, можно осуществить путем перебора основных форм тренда, расчета по каждому уравнению скорректированного коэффициента детерминации $\bar{R}^2$ и выбора уравнения тренда с максимальным значением $\bar{R}^2$.

Для выявления основных тенденций чаще всего используется МНК. Значения временного ряда y_t рассматриваются как зависимая переменная, а время t – как объясняющая. Для нелинейных трендов предварительно проводят стандартную процедуру линеаризации. Согласно МНК параметры прямой $y_t = a + bt$ находятся из системы нормальных уравнений:

$$\begin{cases} an + b\sum_{t=1}^{n} t = \sum_{t=1}^{n} y_t \\ a\sum_{t=1}^{n} t + b\sum_{t=1}^{n} t^2 = \sum_{t=1}^{n} y_t t \end{cases}. \qquad (5.6)$$

Учитывая, что значения переменной $t = 1,2,\dots,n$ образуют натуральный ряд чисел от 1 до n, то справедливы формулы:

$$\sum_{t=1}^{n} t = \frac{n(n+1)}{2}; \quad \sum_{t=1}^{n} t^2 = \frac{n(n+1)(2n+1)}{6}. \qquad (5.7)$$

Наиболее простую экономическую интерпретацию имеют параметры линейного и экспоненциального трендов. Параметры линейного тренда можно интерпретировать так: a – начальный уровень временного ряда в момент времени $t = 0$, b – средний за период абсолютный прирост уровней ряда.

Методы сглаживания временного ряда. Основной целью сглаживания ряда является выделение трендовой компоненты процесса. При сглаживании временного ряда в большей или меньшей степени нивелируется влияние нерегулярной составляющей отклика, так что сглаженный ряд фактически оказывается (в рамках классической модели) суперпозицией тренда и циклической (и возможно сезонной) составляющих процесса, что облегчает их дальнейшее исследование. Обычно используется метод скользящего среднего или метод экспоненциального сглаживания.

Метод скользящих средних. Он основан на переходе от начальных значений членов ряда к их средним значениям на интервале времени, длина которого определена заранее. При этом сам выбранный интервал времени «скользит» вдоль ряда.

Метод крайне субъективен, и результаты сглаживания очень подвержены влиянию длины периода сглаживания. С одной стороны, при небольших периодах не удается выявить трендовую компоненту сильно зашумленного процесса, при больших же периодах происходят значительные потери данных на концах анализируемого интервала.

Скользящее среднее порядка L – это временной ряд, состоящий из средних арифметических L соседних значений y_i, по всем возможным значениям времени. При анализе динамики финансовых процессов в качестве L выбирается обычно нечетное число 3, 5 или 7, и эти схемы называют трехточечной, пятиточечной и т. д.

В общем случае новый временной ряд, выровненный с помощью метода скользящих средних, определяется по формуле:

$$MA_i(L) = \frac{\sum_{t=0}^{L} y_{i+t}}{L} \qquad (5.8)$$

где MA_i – значение скользящих средних по L-точечной схеме в i-м элементе ряда.

Как следует из формулы, в начале и на концах исходного интервала будет теряться по $L / 2$ или $(L – 1) / 2$ исходных точек; так, для 5-точечного сглаживания будут потеряны два первых и два последних значения.

Получаемый таким образом ряд скользящих средних ведет себя более гладко, чем исходный ряд, из-за усреднения отклонений ряда. Действительно, если индивидуальный разброс значений члена временного ряда y_t около своего среднего (сглаженного) значения a характеризуется дисперсией σ^2, то разброс средней из L членов временного ряда около того же значения будет характеризоваться существенно меньшей величиной дисперсии, равной $\frac{\sigma^2}{L}$.

Понятие адаптивных методов прогнозирования. При анализе временных рядов часто более важной бывает текущая тенденция (тенденция в данный момент времени, определяемая несколькими последними наблюдениями), а не тенденция, сложившая на длительном интервале времени. Соответственно, наиболее ценной является информация последнего периода. Исходя из этого в последнее время особую важность получили так называемые адаптивные методы прогнозирования.

Адаптивными называются методы прогнозирования, позволяющие строить самокорректирующиеся (самонастраивающиеся) экономико-математические модели, которые способны оперативно реагировать на изменение условий путем учета результата прогноза, сделанного на предыдущем шаге, и учета различной информационной ценности уровней ряда.

Особенностями адаптивных методов прогнозирования является:

– способность учитывать информационную ценность уровней временного ряда (с помощью системы весов, придаваемых этим уровням);

– использование рекуррентных процедур уточнения параметров модели по мере поступления новых данных наблюдений и тем самым адаптация модели применительно к новым условиям развития явления.

Скорость (быстроту) реакции модели на изменения в динамике процесса характеризует так называемый параметр адаптации. Параметр адаптации должен быть выбран таким образом, чтобы обеспечивалось адекватное отображение тенденции при одновременной фильтрации случайных отклонений. Значение параметра адаптации может быть определено на основе эмпирических данных, выведено аналитическим способом или получено на основе метода проб.

В качестве критерия оптимальности при выборе параметра адаптации обычно принимают минимум среднего квадрата ошибок прогнозирования.

Благодаря указанным свойствам адаптивные методы особенно удачно используются при краткосрочном прогнозировании (при прогнозировании на один или несколько шагов вперед).

Адаптивные методы, как правило, основаны на использовании процедуры экспоненциального сглаживания.

Экспоненциальное сглаживание. Метод экспоненциального сглаживания также дает возможность оценить степень воздействия трендовой и/или циклической компоненты на отклик системы, но в отличие от метода скользящих средних, еще и может быть использован для краткосрочных прогнозов будущей тенденции на один период вперед. Именно поэтому метод обладает явным преимуществом над ранее рассмотренным.

Название метода происходит из того факта, что на самом деле при его применении получаются экспоненциально взвешенные скользящие средние по всему временному ряду; а из этого следует, что сглаженное значение в любой точке ряда является некоторой функцией всех предшествующих наблюдаемых значений. В методе скользящих средних при расчете не учитывается влияние наблюдений, отстоящих более чем на $(L-1)/2$ периодов от рассматриваемого. При экспоненциальном сглаживании учитываются все предшествующие наблюдения – предыдущее учитывается с максимальным весом, предшествующее ему – с несколько меньшим, самое «старое» наблюдение влияет на результат с минимальным статистическим весом.

Алгоритм расчета экспоненциально сглаженных значений в любой точке ряда i основан на трех величинах: наблюдаемом значении y_i в данной точке ряда i, рассчитанном сглаженном значении для предшествующей точки ряда E_{i-1} и некотором заранее заданном коэффициенте сглаживания W, постоянном по всему ряду. Понятно, что в первой точке ряда нет сглаженного значения для предшествующей точки (нет самой такой точки), и сглаженным значением E_1 считается сама наблюдаемая в этой точке величина отклика y_1. Для всех следующих точек действует простое правило вычислений:

$$E_i = w \cdot y_i + (1-w) \cdot E_{i-1}. \qquad (5.9)$$

Некоторой проблемой является выбор коэффициента сглаживания W, который в значительной степени влияет на результаты. К сожалению, объективного критерия при его выборе не существует. При равной степени сглаживания с использованием метода экспоненциального сглаживания и метода скользящего среднего коэффициент W связан с интервалом L простым соотношением

$$w = \frac{2}{L+1}. \qquad (5.10)$$

Таким образом, сглаживание по 5-точечной схеме эквивалентно по своему воздействию на исходные данные экспоненциальному сглаживанию с коэффициентом $W = 0{,}33$. Хотя в принципе W может принимать любые значения из диапазона $0 < W < 1$, обычно ограничиваются интервалом от 0,2 до 0,5. При высоких значениях W

в большей степени учитываются мгновенные текущие наблюдения отклика и наоборот, при низких его значениях сглаженная величина определяется в большей степени прошлой тенденцией развития, нежели текущим состоянием отклика системы.

Аналитики большинства фирм при обработке рядов используют «свои» традиционные значения W. Так, в аналитическом отделе Kodak традиционно используют значение 0,38, а на фирме Ford Motors – 0,28 или 0,3. Вообще для динамично развивающихся фирм и рынков характерны более высокие значения W, чем для более консервативных компаний и стабильных рынков; для прогнозов используют более высокие величины, чем для анализа предшествующих тенденций. Тем не менее реальность такова, что выбор коэффициента W был и останется крайне субъективным.

Понятие об авторегрессионных моделях и моделях скользящего среднего. Для рассматриваемого временного ряда далеко не всегда удается подобрать адекватную модель, для которой ряд возмущений ε_t будет удовлетворять основным предпосылкам регрессионного анализа. В классическом представлении рассматривается модель вида

$$y_t = f(t) + \varepsilon_t,$$

в которой в качестве объясняющей переменной выступала переменная t – время. В эконометрике распространены и другие модели, в которых объясняющими переменными выступают лаговые переменные, т. е. переменные, влияние которых в эконометрической модели характеризуется некоторым запаздыванием.

Авторегрессионная модель p-го порядка (или модель $AR(p)$) имеет вид:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \ldots + \beta_p y_{t-p} + \varepsilon_t \quad (t = 1,2,\ldots,n), \qquad (5.11)$$

где $\beta_0, \beta_1, \ldots, \beta_p$ – некоторые константы.

Эта модель описывает изучаемый процесс в момент t в зависимости от его значений в предыдущие моменты t-1, t-2, …, t-p.

Если исследуемый процесс y_t в момент t определяется его значениями только в предшествующий период t-1, то рассматривают авторегрессионную модель 1-го порядка (или модель $AR(1)$) – марковский случайный процесс:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t \qquad (t = 1,2,\ldots,n). \qquad (5.12)$$

Наряду с авторегрессионными моделями временным рядов в эконометрике рассматриваются также модели скользящего среднего, в которой моделируемая величина задается линейной функцией от возмущений (ошибок) в предыдущий момент времени.

Модель скользящего среднего q-го порядка (модель $MA(q)$), имеет вид:

$$y_t = \varepsilon_t + \gamma_1 \varepsilon_{t-1} + \gamma_2 \varepsilon_{t-2} + \dots + \gamma_q \varepsilon_{t-q} \,. \qquad (5.13)$$

В эконометрике используются также комбинированные модели временных рядов *AR* и *MA*.

Авторегрессионная модель скользящего среднего порядков p и q (модель $ARMA(p, q)$) имеет вид:

$$y_t = \beta_0 + \beta_1 y_{t-1} + \dots + \beta_p y_{t-p} + \varepsilon + \gamma_1 \varepsilon_{t-1} + \dots + \gamma_q \varepsilon_{t-q}. \quad (5.14)$$

Использование соответствующих авторегрессионных моделей для прогнозирования экономических показателей, т. е. автопрогноз на базе рассмотренных моделей, может оказаться эффективным (как правило, в краткосрочной перспективе).

Моделирование сезонных и циклических колебаний. Основные характеристики, отражающие влияние на результативный признак как сезонной, так и циклической компонент, одинаковы. Поэтому методы моделирования сезонных компонент справедливы и для циклических компонент.

Выше отмечалось, что существует два класса моделей временного ряда: аддитивная и мультипликативная. Выбор одной из двух моделей осуществляется на основе анализа структуры сезонных колебаний. Если амплитуда колебаний приблизительно постоянна, строится аддитивная 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.

Процесс построения модели включает в себя шесть шагов.

Поскольку данный алгоритм применяется для временного ряда, содержащего сезонную компоненту, то обязательно наиболее высоким должен оказаться коэффициент автокорреляции порядка больше, чем первый, так как только в этом случае можно с уверенностью утверждать, что данный временной ряд содержит циклические или сезонные колебания с периодичностью в τ моментов времени.

Шаг 1. Выравнивание исходного ряда методом скользящей средней. На этом шаге L определяется равным τ (наиболее высокий коэффициент автокорреляции порядка τ), так как ряд содержит сезонные колебания с периодичностью в τ моментов времени. Полученные таким образом выровненные значения уже не содержат сезонной компоненты. Далее находятся центрированные скользящие средние – средние значения из двух последовательных скользящих средних.

Шаг 2. Расчет значений сезонной компоненты. Для аддитивной модели: найдем оценки сезонной компоненты как разность между фактическими уровнями ряда и центрированными скользящими средними. Для расчета сезонной компоненты обычно используется форма, представленная в табл. 5.1.

Таблица 5.1

Форма для расчета значений сезонной компоненты

Показатели	Номер периода $m = \frac{n}{\tau}$	Номер единицы времени, входящей в период колебания, i			
		1	2	…	τ
	1 2 … m				
Итого за i-ю единицу времени					
Средняя оценка сезонной компоненты для i-й единицы времени $\bar{S}_i$		$\bar{S}_1$	$\bar{S}_2$	…	$\bar{S}_\tau$
Скорректированная сезонная компонента: для аддитивной $S_i = \bar{S}_i - k$; для мультипликативной $S_i = \bar{S}_i \cdot k$, где $k = \frac{\bar{S}_1 + \bar{S}_2 + \ldots + \bar{S}_\tau}{\tau}$		S_1	S_2	…	S_τ

В моделях с сезонной компонентой обычно предполагается, что сезонные воздействия за период взаимопогашаются. В аддитивной модели это выражается в том, что сумма значений сезонной компоненты по всем единицам времени, входящим в период колебания, должна быть равна нулю, т. е. $\sum S_i = 0$.

Для мультипликативной модели: найдем оценки сезонной компоненты как частное от деления фактических уровней ряда на центрированные скользящие средние. Взаимопогашаемость сезонных воздействий выражается в том, что сумма значений сезонной компоненты по всем единицам времени, входящим в период колебания, должна быть равна числу периодов в цикле, т. е. $\sum S_i = \tau$.

Шаг 3. Устранение сезонной компоненты из исходных уровней ряда и получение выравненных данных $u_t + \varepsilon_t$ в аддитивной или $u_t \cdot \varepsilon_t$ в мультипликативной модели.

Для аддитивной модели: устраним влияние сезонной компоненты, вычитая ее значения из каждого уровня исходного временного ряда. Так как исходное уравнение имеет вид $y_t = u_t + v_t + \varepsilon_t \quad (t = 1,2, \dots, n)$, то справедливо $y_t - v_t = u_t + \varepsilon_t$. Таким образом, можно выделить значения, содержащие только тенденцию и случайную компоненту.

Для мультипликативной модели: разделим каждый уровень исходного ряда на соответствующее значение сезонной компоненты. Получим величины $u_t \cdot \varepsilon_t = \frac{y_t}{v_t}$.

Шаг 4. Аналитическое выравнивание уровней $u_t + \varepsilon_t$ или $u_t \cdot \varepsilon_t$ и расчет значений u_t с использованием полученного уравнения тренда. Строим линейный тренд с помощью МНК по уровням $u_t + \varepsilon_t$ или $u_t \cdot \varepsilon_t$, получим линейный тренд $u_t = a + bt$. Подставляя в это уравнение значения $(t = 1,2,\dots, n)$, найдем уровни u_t для каждого момента времени.

Шаг 5. Расчет полученных по модели значений $u_t + v_t$ или $u_t \cdot v_t$. Для аддитивной модели: прибавим к уровням u_t значения сезонной компоненты для соответствующей единицы времени. Для мультипликативной модели: умножим уровни u_t на значения сезонной компоненты для соответствующей единицы времени.

Шаг 6. Расчет абсолютных и/или относительных ошибок. Если полученные значения ошибок не содержат автокорреляции, ими можно заменить исходные уровни ряда и в дальнейшем использовать временной ряд ошибок ε_t для анализа взаимосвязи исходного ряда и других временных рядов. Для аддитивной модели: расчет ошибки можно произвести следующим образом: $\varepsilon_t = y_t - (u_t + v_t) \quad (t = = 1,2, \dots, n)$, а для мультипликативной модели: $\varepsilon_t = \frac{y_t}{u_t \cdot v_t} \quad (t = 1,2, \dots, n)$.

Выявление и устранение сезонного эффекта в практической деятельности используется в двух направлениях:

– воздействие сезонных колебаний следует устранять на этапе предварительной обработки исходных данных при изучении взаимосвязи нескольких временных рядов;

– в процессе анализа структуры одномерных временных рядов с целью прогнозирования уровней ряда в будущие моменты времени.

Задача 5

В табл. 5.2 приведены данные, отражающие изменения индекса экономической безопасности за тридцатилетний период. По этим данным произвести моделирование всех составляющих временного ряда.

Таблица 5.2

Исходные данные для задачи

Год	Месяц	Индекс экономической безопасности
1	*2*	*3*
2017	январь	237
2017	февраль	241
2017	март	274
2017	апрель	228
2017	май	222
2017	июнь	193
2017	июль	217
2017	август	226
2017	сентябрь	295
2017	октябрь	280
2017	ноябрь	274
2017	декабрь	268
2018	январь	303
2018	февраль	318
2018	март	353
2018	апрель	306
2018	май	310
2018	июнь	279
2018	июль	319
2018	август	327

Окончание табл. 5.2

1	2	3
2018	сентябрь	365
2018	октябрь	323
2018	ноябрь	321
2018	декабрь	296
2019	январь	323
2019	февраль	336
2019	март	411
2019	апрель	351
2019	май	342
2019	июнь	330

Рис. 5.1. График временного ряда

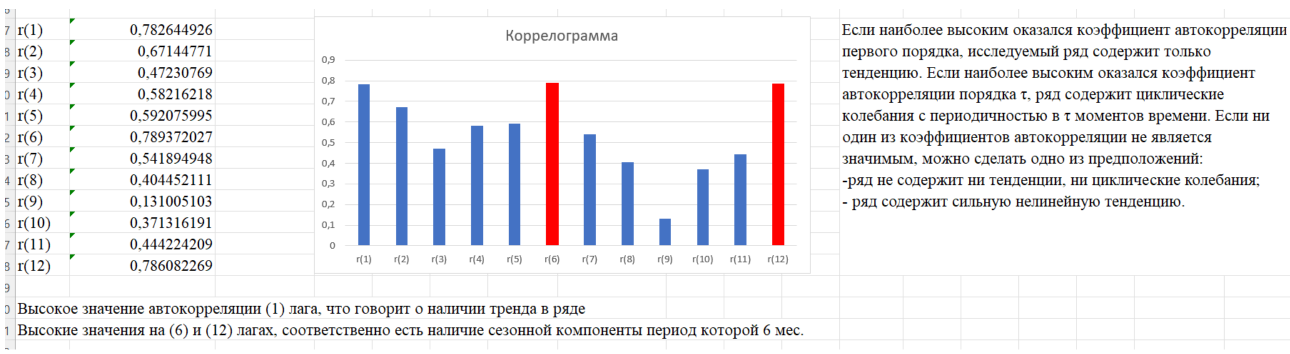


Рис. 5.2. Автокорреляционная функция

Моделирование временного ряда

Одной из важнейших задач исследования экономического временного ряда является выявление основной тенденции изучаемого процесса, выраженной неслучайной составляющей (тренда или тренда с циклической или/и сезонной компонентой). Одним из способов моделирования тенденций временного ряда является построение аналитической функции, характеризующей зависимость уровней ряда от времени, или тренда.

Удалим линейный тренд. Для этого построим уравнение регрессии, где x это t.

Запишем уравнение регрессии

a	222,7310345
b	4,701223582

ŷ=222,7+4,701*x

Рис. 5.3. Удаление тренда из временного ряда

F	56,74596202		

F табличное определяется из таблицы Фишера

Номер строки - (n-m-1), где n - количество выборки, m - количество неизвестных

Номер столбца - m

Fкрит	4,2	так как строка 28, столбец 1	
F	>	Fкрит	делаем вывод о значимости уравнения регрессии.
ta	20,10329152	> tкрит, параметр a значим	
tb	7,532991572	> tкрит, параметр b значим	

t критическое определяется из таблицы Стьюдента

Номер строки - (n-m-1), где n - количество выборки, m - количество неизвестных

Номер столбца - уровень значимости (обычно 0,05)

tкрит	2,05	так как строка 28, столбец 0,05

Рис. 5.4. Оценка линейного уравнения

Оценим сезонную компоненту

Для этого с листа "временной ряд" скопируем столбец остатков

Рассчитаем скорректированную сезонную компоненту (S) методом сезонных индексов

t	y	Ряд без тренда		$\bar{S}$	$\bar{S}$	S (оц.сез.комп.)	Остатки
1	237	9,57	I	-3,37	-3,37	-3,37	12,94
2	241	8,87	II	0,88	0,88	0,88	7,99
3	274	37,17	III	38,63	38,63	38,63	-1,46
4	228	-13,54	IV	-0,29	-0,29	-0,29	-13,24
5	222	-24,24	V	-7,38	-7,38	-7,38	-16,86
6	193	-57,94	VI	-28,46	-28,46	-28,46	-29,48
7	217	-38,64	I		-3,37	-3,37	-35,27
8	226	-34,34	II		0,88	0,88	-35,22
9	295	29,96	III		38,63	38,63	-8,67
10	280	10,26	IV		-0,29	-0,29	10,55
11	274	-0,44	V		-7,38	-7,38	6,93
12	268	-11,15	VI		-28,46	-28,46	17,32
13	303	19,15	I		-3,37	-3,37	22,53
14	318	29,45	II		0,88	0,88	28,58
15	353	59,75	III		38,63	38,63	21,13
16	306	8,05	IV		-0,29	-0,29	8,34
17	310	7,35	V		-7,38	-7,38	14,72
18	279	-28,35	VI		-28,46	-28,46	0,11
19	319	6,95	I		-3,37	-3,37	10,32
20	327	10,24	II		0,88	0,88	9,37
21	365	43,54	III		38,63	38,63	4,92
22	323	-3,16	IV		-0,29	-0,29	-2,87
23	321	-9,86	V		-7,38	-7,38	-2,48
24	296	-39,56	VI		-28,46	-28,46	-11,10
25	323	-17,26	I		-3,37	-3,37	-13,89
26	336	-8,96	II		0,88	0,88	-9,84
27	411	61,34	III		38,63	38,63	22,71
28	351	-3,37	IV		-0,29	-0,29	-3,07
29	342	-17,07	V		-7,38	-7,38	-9,69
30	330	-33,77	VI		-28,46	-28,46	-5,31

Рис. 5.5. Оценка сезонной компоненты

В моделях с сезонной компонентой обычно предполагается, что сезонные воздействия за период взаимопогашаются. В аддитивной модели это выражается в том, что сумма значений сезонной компоненты по всем единицам времени, входящим в период колебания, должна быть равна нулю, т.е. $\sum S_i = 0$

ΣS	-0,0000000000001	если нет, то считаем скорректированнй коэффициент	$k = \frac{\bar{S}_1 + \bar{S}_2 + ... + \bar{S}_\tau}{\tau}$
к	-2,30926E-14		

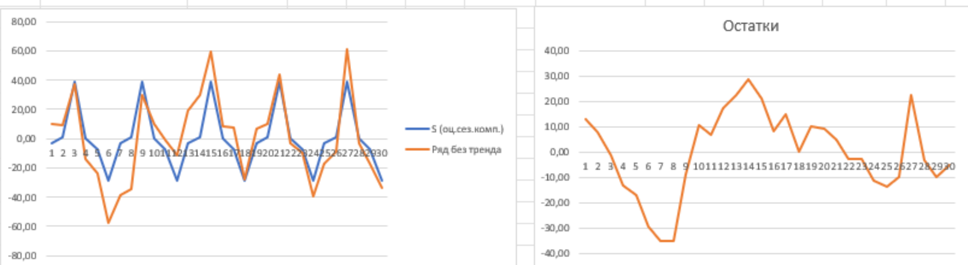


Рис. 5.6. Моделирование временного ряда

Контрольные вопросы

1. Какие элементы временного ряда можно выделить?
2. Какие этапы выделяются при моделировании временного ряда?
3. Охарактеризуйте понятие авторегрессионной модели и модели скользящего среднего.
4. Что характеризует автокорреляция уровней временного ряда?
5. Как проводится процедура выявления структуры временного ряда?
6. Перечислите методы моделирования тенденций временного ряда.
7. Опишите алгоритм моделирования сезонных и циклических колебаний временного ряда.
8. Назовите различия в построении аддитивной и мультипликативной модели временного ряда.
9. Опишите алгоритм построения аддитивной и мультипликативной модели временного ряда.

Тест

1. Пространственными моделями называются:

а) модели, построенные по данным, характеризующим один объект за ряд последовательных моментов (периодов) времени;

б) модели, которые учитывают совокупное влияние множества факторов;

в) модели, построенные по данным, характеризующим совокупность различных объектов в определенный момент (период) времени.

2. Компонента временного ряда, отражающая повторяемость экономических процессов в течение длительных периодов называется:

а) тенденцией;

б) случайной;

в) трендовой;

г) сезонной;

д) циклической.

3. Модели, построенные по данным, характеризующим один объект за ряд последовательных моментов (периодов) времени, называются:

а) моделями системы эконометрических уравнений;

б) моделями множественной регрессии;

в) пространственными моделями;

г) моделями временных рядов.

4. Модель, в которой объясняющими переменными выступают лаговые переменные, т. е. переменные, влияние которых в эконометрической модели характеризуется некоторым запаздыванием, называется:

а) аддитивной;

б) мультипликативной;

в) авторегрессионной.

5. Коэффициент, измеряющий тесноту связи между уровнями одного и того же ряда, называется:

а) линейным коэффициентом корреляции;

б) коэффициентом автокорреляции;

в) множественным коэффициентом корреляции.

6. Коррелограмма – это:

а) график зависимости уровней ряда друг от друга;

б) график зависимости уровней ряда от лага;

в) график зависимости автокорреляционной функции от тенденции;

г) график зависимости автокорреляционной функции от лага.

7. Если ни один из коэффициентов автокорреляции не является значимым, то:

а) можно выдвинуть гипотезу о статистической незначимости уравнения;

б) ряд содержит сильную нелинейную тенденцию;

в) ряд не содержит ни тенденции, ни сезонные колебания.

8. Автокорреляция в остатках имеет место, если:

а) остатки имеют случайный закон распределения;

б) каждое следующее значение остатков зависит от предшествующих;

в) уровни ряда содержат тенденцию.

9. Наличие автокорреляции в остатках первого порядка между соседними членами ряда позволяет определить:

а) F-критерий Фишера;
б) t-критерий Стьюдента;
в) критерий Энгеля – Грангера;
г) критерий Дарбина – Уотсона.

6. ИЗУЧЕНИЕ ВЗАИМОСВЯЗЕЙ ПО ВРЕМЕННЫМ РЯДАМ

Предположим, изучается зависимость между временными рядами $x(t)$ и $y(t)$. Для количественной характеристики этой зависимости используется линейный коэффициент корреляции. Если рассматриваемые временные ряды имеют тенденцию, то коэффициент корреляции по абсолютной величине будет высоким (положительным в случае совпадения и отрицательным в случае противоположной направленности тенденций рядов $x(t)$ и $y(t)$). Однако из этого еще нельзя делать вывод о том, что x причина y или наоборот. Для того чтобы получить коэффициенты корреляции, характеризующие причинно-следственную связь между изучаемыми рядами, следует избавиться от ложной корреляции, вызванной наличием тенденции в каждом ряде.

По двум временным рядам $x(t)$ и $y(t)$ строится уравнение парной линейной регрессии вида $y_t = a + bx_t + \varepsilon_t$. Наличие тенденции в каждом из этих временных рядов означает, что на зависимую $y(t)$ и независимую$x(t)$ переменные модели оказывает воздействие фактор времени, который непосредственно в модели не учтен. Влияние фактора времени будет выражено в корреляционной зависимости между значениями остатков ε_t за текущий и предыдущие моменты времени, которая получила название автокорреляции в остатках.

Таким образом, при изучении взаимодействия временных рядов основная цель заключается в том, чтобы устранить или зафиксировать воздействие фактора времени на формирование уровней ряда. Существует две группы методов:

– методы, основанные на преобразовании уровней исходного ряда в новые переменные, не содержащие тенденции. Эти методы предполагают непосредственное устранение трендовой компоненты из каждого уровня временного ряда (метод последовательных разностей и метод отклонений от тренда);

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

Метод отклонений от тренда. Пусть имеются два временных ряда $x(t)$ и $y(t)$, каждый из которых содержит трендовую и случайную компоненты. Проведение аналитического выравнивания по каждому из этих рядов позволяет найти параметры соответствующих уравнений трендов $y_t = a_1 + b_1 t$ и $x_t = a_2 + b_2 t$, а также определить расчетные по тренду уровни y_t, x_t. Эти значения можно принять за оценку трендовой компоненты каждого ряда. Поэтому влияние тенденции можно устранить путем вычитания расчетных значений уровней ряда от фактических $(y_t - y_t), (x_t - x_t)$. Дальнейший анализ взаимосвязи рядов проводят с использованием не исходных уровней, а отклонений от тренда $(y_t - y_t), (x_t - x_t)$.

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 $y_t = y_t + \varepsilon_t$, где $y_t = a + bt$, тогда первые разности определяются:

$$\Delta_t = y_t - y_{t-1} = a + bt + \varepsilon_t - (a + b(t-1) + \varepsilon_{t-1}) = b + (\varepsilon_t - \varepsilon_{t-1}).$$

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

Если временной ряд содержит тенденцию в форме параболы второго порядка, то для ее устранения можно заменить исходные уровни ряда на вторые разности.

Пусть $y_t = y_t + \varepsilon_t$, где $y_t = a + b_1 t + b_2 t^2$, тогда первые разности определяются:

$$\begin{aligned} \Delta_t &= y_t - y_{t-1} = a + b_1 t + b_2 t^2 + \varepsilon_t - (a + b_1(t-1) + b_2(t-1)^2 + \varepsilon_{t-1}) \\ &= b_1 - b_2 + 2b_2 t + (\varepsilon_t - \varepsilon_{t-1}) \end{aligned}.$$

Как показывает это соотношение, первые разности Δ_t непосредственно зависят от фактора времени t и, следовательно, содержат тенденцию. Поэтому определим вторые разности:

$$\Delta_t^2 = \Delta_t - \Delta_{t-1} = b_1 - b_2 + 2b_2 t + (\varepsilon_t - \varepsilon_{t-1}) -$$
$$(b_1 - b_2 + 2b_2(t-1) + \varepsilon_{t-1} - \varepsilon_{t-2})) = 2b_2 + (\varepsilon_t - 2\varepsilon_{t-1} + \varepsilon_{t-2}).$$

Очевидно, что вторые разности Δ_t^2 не содержат тенденции, поэтому при наличии в исходных уровнях тренда в форме параболы второго порядка их можно использовать для дальнейшего анализа.

Если тенденции временного ряда соответствует экспоненциальный или степенной тренд, метод последовательных разностей следует применять не к исходным уровням ряда, а к их логарифмам. Этот метод имеет два недостатка:

– его применение связано с сокращением числа пар наблюдений, по которым строится уравнение регрессии, и, следовательно, с потерей числа степеней свободы;

– использование вместо исходных уровней временных рядов их прироста или ускорения приводит к потере информации, содержащейся в исходных данных.

Метод включения в модель регрессии фактора времени. Устранить воздействие какого-либо фактора можно, если зафиксировать воздействие этого фактора на результат и другие включенные в модель факторы. Во временных рядах тенденция фиксируется через включение фактора времени в модель в качестве независимой переменной.

Модель вида $y_t = a + b_1 x_t + b_2 t + \varepsilon_t$ относится к группе моделей, включающих фактор времени. Иными словами, в эту модель время включено как отдельная независимая переменная, а значит, при построении уравнения регрессии с помощью МНК по t будет браться частная производная. Значит, система нормальных уравнений примет вид

$$\begin{cases} \sum y = na + b_1 \sum x + b_2 \sum t \\ \sum yx = a \sum x + b_1 \sum x^2 + b_2 \sum x \cdot t. \\ \sum yt = a \sum t + b_1 \sum x \cdot t + b_2 \sum t^2 \end{cases} \qquad (6.1)$$

Преимущество данной модели по сравнению с двумя предыдущими методами в том, что она позволяет учесть всю информацию, содержащуюся в исходных данных, поскольку значения y_t, x_t есть уровни исходных временных рядов.

Критерий Дарбина – Уотсона. Пусть дано уравнение регрессии вида

$$y_t = a + \sum_{j=1}^{k} b_j x_{jt} + \varepsilon_t, \qquad (6.2)$$

где k – число независимых переменных модели.

Для каждого момента времени $t = 1,\dots,n$ значение компоненты ε_t определяется как $\varepsilon_t = y_t - y_t$или $\varepsilon_t = y_t - (a + \sum_j b_j x_{jt})$.

В соответствии с предпосылками МНК остатки ε_t должны быть случайными. Однако иногда каждое следующее значение остатков зависит от предшествующих. В этом случае имеет место автокорреляция в остатках.

Для определения наличия автокорреляции остатков между соседними членами ряда используется критерий (тест) Дарбина – Уотсона.

В этом тесте для оценки корреляции используется статистика:

$$d = \frac{\sum_{t=2}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n} \varepsilon_t^2}. \tag{6.3}$$

Производя ряд преобразований, можно получить, что статистика Дарбина – Уотсона связана с коэффициентом автокорреляции остатков первого порядка следующим образом:

$$d \approx 2(1 - r_1^{\varepsilon}). \tag{6.4}$$

Таким образом, если в остатках существует полная положительная автокорреляция и $r_1^{\varepsilon} = 1$, то $d = 0$. Если в остатках полная отрицательная автокорреляция, то $r_1^{\varepsilon} = -1$ и, следовательно, $d = 4$. Если автокорреляция остатков отсутствует, то $r_1^{\varepsilon} = 0$ и $d = 2$. Следовательно, d-статистика может изменяться в пределах $0 \le d \le 4$.

Пусть выдвигается гипотеза H_0 об отсутствии автокорреляции в остатках, а также две альтернативные гипотезы H_1 – гипотеза о положительной автокорреляции в остатках и H_2 – гипотеза об отрицательной автокорреляции в остатках. Существуют два пороговых значения d_B и d_H, зависящие только от числа наблюдений, числа регрессоров и уровня значимости, такие, что выполняются следующие условия:

а) $d_{\text{В}} < d < 4 - d_{\text{В}}$ – гипотеза H_0 принимается;

б) $d_{\text{Н}} < d < d_{\text{В}}$ или $4 - d_{\text{В}} < d < 4 - d_{\text{Н}}$ – вопрос об отвержении или принятии гипотезы H_0 остается открытым (область неопределенности критерия);

в) $0 < d < d_{\text{Н}}$ – принимается альтернативная гипотеза H_1 о положительной автокорреляции;

г) $4 - d_{\text{Н}} < d < 4$ – принимается альтернативная гипотеза H_2 об отрицательной автокорреляции.

Результат критерия Дарбина – Уотсона представлен на рис. 6.1.

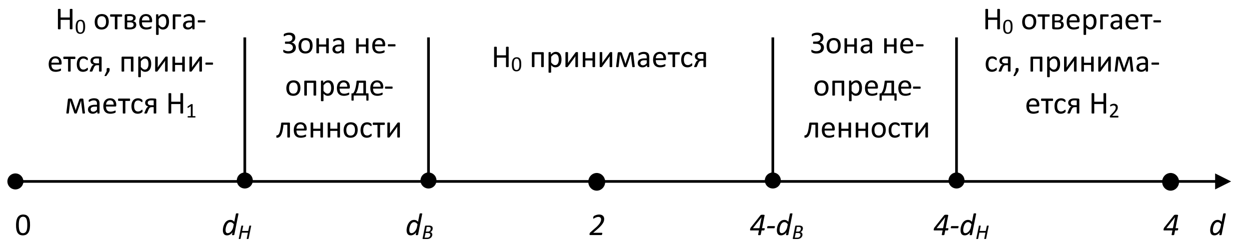


Рис. 6.1. Графическая интерпретация критерия Дарбина – Уотсона

Для статистики Дарбина – Уотсона найдены верхняя d_B и нижняя d_H границы на уровнях значимости α = 0,01; α = 0,025; α = 0,05.

Недостатками критерия Дарбина – Уотсона является наличие области неопределенности критерия, а также то, что критические значения d-статистики, представленные в приложении, определены для объемов выборки не менее 6. Кроме того, этот критерий направлен только на выявление автокорреляции остатков первого порядка.

Коинтеграция временных рядов. В ряде случаев наличие в одном из временных рядов тенденции может быть следствием именно того факта, что другой ряд, включенный в модель, тоже содержит тенденцию, а не просто является результатом прочих случайных причин. Начиная с 70-х гг. XX в. была выдвинута теория коинтеграции временных рядов. Коинтеграция – причинно-следственная зависимость в уровнях двух (или более) временных рядов, которая выражается в совпадении или противоположной направленности их тенденций и случайной колеблемости. Между двумя временными рядами существует коинтеграция в случае, если линейная комбинация этих временных рядов есть стационарный временной ряд (ряд, содержащий только случайную компоненту и имеющий постоянную дисперсию на длительном промежутке времени).

Не всегда наличие тенденции во временных рядах $x(t)$ и $y(t)$ приводит к недостоверности оценок параметров регрессии $y_t = a + bx_t + \varepsilon_t$. Если нестационарные временные ряды $x(t)$ и $y(t)$ являются коинтегрируемыми, то оценки параметров регрессии оказываются состоятельными. Таким образом, наличие коинтеграции нестационарных временных рядов позволяет при построении регрессионной модели использовать их исходные уровни $x(t)$ и $y(t)$.

Через коинтеграцию, к примеру, подтверждаются зависимости между уровнем инфляции, ВВП и денежной массой, ценами на акции и их доходностью, потреблением и уровнем дохода и многие другие экономические зависимости с шумящими переменными.

Но следует отметить, что такой подход применим только к временным рядам, охватывающим достаточно длительные промежутки времени.

Критерий *Энгеля – Грангера* позволяет проверить наличие коинтеграции. Алгоритм критерия Энгеля – Грангера включает в себя следующие этапы.

Шаг 1. Выдвигается гипотеза об отсутствии коинтеграции между рядами y_t и x_t.

Шаг 2. Рассчитывают параметры уравнения регрессии вида $\Delta\varepsilon_t = a + b\varepsilon_{t-1}$, где $\Delta\varepsilon_t$ – первые разности остатков.

Шаг 3. Определяют фактическое значение t-критерия Стьюдента для коэффициента регрессии a.

Шаг 4. Сравнивают полученное значение t-критерия с критическим значением статистики τ. Критические значения τ для уровня значимости 1 %, 5 % и 10 % составляют 2,5899; 1,9439 и 1,6177 соответственно. Если фактическое значение t больше критического значения τ для заданного уровня значимости α, гипотезу об отсутствии коинтеграции исследуемых временных рядов отклоняют и с вероятностью $(1-\alpha)$ принимают альтернативную гипотезу о том, что между рядами существует коинтеграция. В противоположном случае гипотезу об отсутствии коинтеграции не отклоняют.

Задача 6

Для отдельного региона известны индекс экономической безопасности $y(t)$ и доля мошенничества в сфере предпринимательской деятельности $x(t)$ за последние 10 лет. Данные показатели с экономической точки зрения имеют общие предпосылки, для принятия или опровержения данной гипотезы целесообразно произвести проверку наличия коинтеграции временных рядов с помощью критерия Энгеля – Грангера (табл. 6.1).

Таблица 6.1

Исходные данные задачи 6

t	1	2	3	4	5	6	7	8	9	10
$y(t)$	16	14	17	10	18	17	14	16	22	20
$x(t)$	15	9	10	12	12	9	10	14	12	8

Решение.

На первом этапе необходимо из исходных временных рядов исключить тенденцию, например, методом последовательных разностей. Новые уровни временных рядов представлены в табл. 6.2.

Таблица 6.2

Уровни временных рядов без влияния тенденции

t	$y(t)$	$x(t)$	$\Delta_y = y_t - y_{t-1}$	$\Delta_x = x_t - x_{t-1}$
1	16	15	–	–
2	14	9	–2	–6
3	17	10	3	1
4	10	12	–7	2
5	18	12	8	0
6	17	9	–1	–3
7	14	10	–3	1
8	16	14	2	4
9	22	12	6	–2
10	20	8	–2	–4

Далее по новым уровням ряда строится уравнение парной регрессии, для этого решаем систему нормальных уравнений:

$$\begin{cases} 9a + (-7)b = 4 \\ -7a + 87b = 5 \end{cases}.$$

Полученное уравнение парной регрессии $\Delta y_t = -0{,}52176 +$ $+0{,}0994 \cdot \Delta x_t$.

Теоретические значения Δy_t, остаточные ошибки ε_t и промежуточные расчеты критерия Дарбина – Уотсона приведены в табл. 6.3.

Таблица 6.3

Расчет критерия Дарбина – Уотсона

t	Δ_y	Δ_x	Δy_t	ε_t	ε_t^2	$\varepsilon_t - \varepsilon_{t-1}$	$(\varepsilon_t - \varepsilon_{t-1})^2$
1	–	–	–	–	–	–	–
2	–2	–6	–1,1182	–0,8818	0,7776	–	–
3	3	1	–0,4224	3,4224	11,7125	4,3042	18,5261
4	–7	2	–0,3230	–6,6770	44,5829	–10,0994	101,9979
5	8	0	–0,5218	8,5218	72,6204	15,1988	231,0035
6	–1	–3	–0,8200	–0,1800	0,0324	–8,7018	75,7213
7	–3	1	–0,4224	–2,5776	6,6442	–2,3976	5,7485
8	2	4	–0,1242	2,1242	4,5121	4,7018	22,1069
9	6	–2	–0,7206	6,7206	45,1659	4,5964	21,1269
10	–2	–4	–0,9194	–1,0806	1,1678	–7,8012	60,8587
Сумма					187,216		537,0899

Согласно формуле (6.3) d-статистика равна:

$$d = \frac{\sum_{t=2}(\varepsilon_t - \varepsilon_{t-1})^2}{\sum_{t=1}^{n}\varepsilon_t^2} = \frac{537{,}0899}{187{,}216} = 2{,}869.$$

По таблице, представленной в приложении, определим для числа наблюдений $n = 9$ и числа независимых переменных $p = 1$ критические значения $d_H = 0{,}82$ и $d_B = 1{,}32$. Получим следующие промежутки внутри интервала [0;4], представленные на рис. 6.2.

Поскольку фактическое значение $d = 2{,}869$ попадает в промежуток от 2,68 до 3,18, то нет возможности принять или отклонить ни одну из трех гипотез, поскольку данный промежуток соответствует зоне неопределенности. Значит, с уверенностью в 95 % нельзя утверждать, что имеет или не имеет место положительная или отрицательная автокорреляция в остатках.

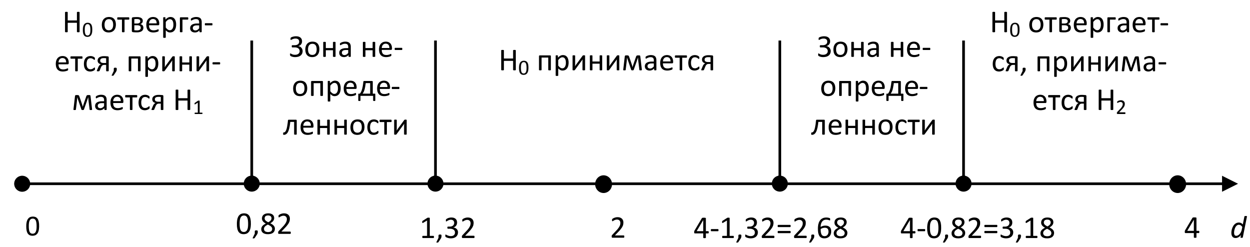


Рис. 6.2. Графическая интерпретация решения задачи 6

В табл. 6.3 представлен расчет случайной компоненты. Эти результаты будем использовать в дальнейшей работе.

Выдвигаем нулевую гипотезу об отсутствии коинтеграции между временными рядами.

Для построения уравнения парной регрессии вида $\Delta\varepsilon_t = a + {} +b\varepsilon_{t-1}$ воспользуемся системой нормальных уравнений:

$$\begin{cases} na + b\sum \varepsilon_{t-1} = \sum \Delta\varepsilon_t \\ a\sum \varepsilon_{t-1} + b\sum \varepsilon_{t-1}^2 = \sum \varepsilon_{t-1}\cdot\Delta\varepsilon_t \end{cases}.$$

Промежуточные расчеты, необходимые для определения параметров уравнения регрессии, представлены в табл. 6.4.

Таблица 6.4

Промежуточные расчеты, необходимые для нахождения параметров парной регрессии

t	ε_t	ε_{t-1}	$\Delta\varepsilon_t$	$\varepsilon_{t-1}\cdot\Delta\varepsilon_t$	ε_{t-1}^2
1	–	–	–	–	–
2	–0,8818	–	–	–	–
3	3,4224	–0,8818	4,3042	–3,79544	0,777571
4	–6,6770	3,4224	–10,0994	–34,5642	11,71282
5	8,5218	–6,677	15,1988	–101,482	44,58233
6	–0,1800	8,5218	–8,7018	–74,155	72,62108
7	–2,5776	–0,18	–2,3976	0,431568	0,0324
8	2,1242	–2,578	4,7018	–12,1194	6,644022
9	6,7206	2,1242	4,5964	9,763673	4,512226
10	–1,0806	6,7206	–7,8012	–52,4287	45,16646
Итого		10,4726	-0,1988	–268,35	186,0489

Тогда система нормальных уравнений примет вид

$$\begin{cases} 8a + 10{,}4726\cdot b = -0{,}1988 \\ 10{,}4726\cdot a + 186{,}0489\cdot b = -268{,}35 \end{cases}.$$

И ее решение позволяет построить искомое уравнение парной регрессии $\Delta\varepsilon_t = 2{,}0115 - 1{,}5556\cdot\varepsilon_{t-1}$.

Согласно третьему шагу алгоритма необходимо рассчитать фактическое значение t-критерия Стьюдента для коэффициента регрессии a. Расчет фактического значения t-критерия Стьюдента производится по известной формуле

$$t_a = \frac{a}{m_a}, \text{ где } m_a = \sqrt{\frac{\sum(\Delta\varepsilon_t - \Delta\varepsilon_t)^2}{n-2}\cdot\frac{\sum\varepsilon_{t-1}^2}{n\sum(\varepsilon_{t-1}-\bar{\varepsilon}_{t-1})^2}}$$

Для нашего примера среднее значение независимого фактора ($\bar{\varepsilon}_{t-1}$) равно $\bar{\varepsilon}_{t-1} = \frac{10,4726}{8} = 1,3091.$

Для нахождения стандартной ошибки воспользуемся расчетами, представленными в табл. 6.5.

Таблица 6.5

Промежуточные расчеты, необходимые для определения стандартной ошибки

t	ε_t	ε_{t-1}	$\Delta\varepsilon_t$	$\Delta\varepsilon_t$	$(\Delta\varepsilon_t - \Delta\varepsilon_t)^2$	$(\varepsilon_{t-1} - \bar{\varepsilon}_{t-1})^2$
1	–	–	–	–	–	–
2	–0,8818	–	–	–	–	–
3	3,4224	–0,8818	4,3042	3,3832	0,8482	4,8000
4	–6,6770	3,4224	–10,0994	–3,3124	46,0636	4,4660
5	8,5218	–6,677	15,1988	12,3982	7,8431	63,7778
6	–0,1800	8,5218	–8,7018	–11,2450	6,4679	52,0230
7	–2,5776	–0,18	–2,3976	2,2915	21,9877	2,2174
8	2,1242	–2,578	4,7018	6,0212	1,7409	15,1064
9	6,7206	2,1242	4,5964	–1,2929	34,6839	0,6644
10	–1,0806	6,7206	–7,8012	–8,4431	0,4120	29,2843
Итого		10,4726	–0,1988	–0,1992	120,0473	172,3395

Тогда стандартная ошибка параметра a равна:

$$m_a = \sqrt{\frac{120,0473 \cdot 186,0489}{6 \cdot 8 \cdot 172,3395}} = 1,6431.$$

Следовательно, фактическое значение t-критерия Стьюдента для параметра a равно: $t_a = \frac{a}{m_a} = \frac{2,0115}{1,6431} = 1,2242.$

Сравниваем полученное значение t-критерия с критическим значением статистики τ. Для всех уровней значимости полученное фактическое значение меньше критических ($1,2242 < 2,5899$, $1,2242 < 1,9439$, $1,2242 < 1,6177$), значит, нулевую гипотезу принимаем и делаем вывод об отсутствии коинтеграции между данными временными рядами.

Значит, предположение о наличии причинно-следственной зависимости между индексом экономической безопасности и долей мошенничества в сфере предпринимательской деятельности следует отклонить.

Контрольные вопросы

1. В чем заключается специфика статистической оценки взаимосвязи двух временных рядов?
2. Какие существуют методы исключения тенденции?
3. Что представляет собой автокорреляция в остатках?
4. В чем заключается критерий Дарбина – Уотсона?
5. Охарактеризуйте понятие коинтеграции временных рядов.
6. Что позволяет оценить критерий Энгеля – Грангера.

Тест

1. Модель, в которой объясняющими переменными выступают лаговые переменные, т. е. переменные, влияние которых в эконометрической модели характеризуется некоторым запаздыванием, называется:

а) аддитивной;

б) мультипликативной;

в) авторегрессионной.

2. Коэффициент, измеряющий тесноту связи между уровнями одного и того же ряда, называется:

а) линейным коэффициентом корреляции;

б) коэффициентом автокорреляции;

в) множественным коэффициентом корреляции.

3. Коррелограмма – это:

а) график зависимости уровней ряда друг от друга;

б) график зависимости уровней ряда от лага;

в) график зависимости автокорреляционной функции от тенденции;

г) график зависимости автокорреляционной функции от лага.

4. Если ни один из коэффициентов автокорреляции не является значимым, то:

а) можно выдвинуть гипотезу о статистической незначимости уравнения;

б) ряд содержит сильную нелинейную тенденцию;

в) ряд не содержит ни тенденции, ни сезонные колебания.

5. Автокорреляция в остатках имеет место, если:

а) остатки имеют случайный закон распределения;

б) каждое следующее значение остатков зависит от предшествующих;

в) уровни ряда содержат тенденцию.

6. Наличие автокорреляции в остатках первого порядка между соседними членами ряда позволяет определить:

а) F-критерий Фишера;

б) t-критерий Стьюдента;

в) критерий Энгеля – Грангера;

г) критерий Дарбина – Уотсона.

7. Под коинтеграцией временных рядов понимается:

а) зависимость между временными рядами, обусловленная наличием тенденции;

б) зависимость между остатками различных уровней;

в) причинно-следственная зависимость в уровнях временных рядов.

8. Сколько гипотез выдвигается при проверке критерия Дарбина – Уотсона:

а) одна;

б) две;

в) три.

9. Стационарный ряд – это:

а) ряд, содержащий только тенденцию и имеющий постоянное ожидание;

б) ряд, содержащий только случайную компоненту и имеющий постоянную дисперсию на длительном промежутке времени;

в) ряд, состоящий только из циклической и сезонной составляющих.

ЗАКЛЮЧЕНИЕ

В пособии рассмотрены методы эконометрики в соответствии с требованиями Государственного стандарта.

Первая глава посвящена вопросам построения парных линейных регрессионных моделей: постановка задачи, спецификация и оценка параметров моделей, оценка качества полученных моделей, получение точечного и интервального прогнозных значений, экономическая интерпретация модели. Приведены условия, обеспечивающие эффективность метода наименьших квадратов (теорема Гаусса – Маркова).

Вторая глава раскрывает вопросы построения парных нелинейных регрессий, вводится понятие коэффициента эластичности.

В третьей главе говорится о построении множественных регрессионных моделей. Подробно рассмотрены вопросы спецификации и оценки параметров модели, оценки качества полученной модели и ее статистической значимости. Вводятся понятия мультиколлинеарности и частных уравнений регрессии, а также множественного и частных коэффициентов корреляции.

В четвертой главе представлены системы эконометрических уравнений и алгоритмы косвенного и двухшагового методов наименьших квадратов.

В пятой главе рассмотрены вопросы моделирования одномерных временных рядов и прогнозирования: структура временного ряда, явление автокорреляции, моделирование тенденции и периодической составляющей ряда, прогнозирование уровней ряда.

Шестая глава дает представление о взаимодействии между временными рядами. Вводится понятие коинтеграции временных рядов, а также алгоритмы критериев Дарбина – Уотсона и Энгеля – Грангера.

СПИСОК ЛИТЕРАТУРЫ

1. Агаларов З. С., Орлов А. И. Эконометрика: учебник. М.: Дашков и К, 2021. 380 с.
2. Берндт Э. Р. Практика эконометрики: классика и современность: учебник для студентов вузов. М.: ЮНИТИ-ДАНА, 2005. 863 с.
3. Вакуленко Е. С., Ратникова Т. А., Фурманов К. К. Эконометрика (продвинутый курс). Применение пакета Stata: учебное пособие для вузов. М.: Юрайт, 2022. 246 с.
4. Галочкин В. Т. Эконометрика: учебник и практикум для вузов. М.: Юрайт, 2022. 293 с.
5. Демидова О. А., Малахов Д. И. Эконометрика: учебник и практикум для вузов. М.: Юрайт, 2022. 334 с.
6. Кремер Н. Ш., Путко Б. А. Эконометрика: учебник и практикум для вузов. 4-е изд., испр. и доп. М.: Юрайт, 2022. 308 с.
7. Магнус Я. Р., Катышев П. К., Пересецкий А. А. Эконометрика. Начальный курс: учебник. 9-е изд., испр. М.: Дело : РАНХиГС, 2021. 504 с.
8. Мансурова Ю. Т., Мухтарова П. А. Эконометрический анализ: учебное пособие. 2-е изд., перераб. и доп. Уфа: УГАТУ, 2011. 161 с.
9. Уткин В. Б. Эконометрика: учебник. 2-е изд. М.: Дашков и К, 2017. 564 с.
10. Эконометрика: учебник для вузов / И. И. Елисеева и др.; под ред. И. И. Елисеевой. М.: Юрайт, 2022. 449 с.

ПРИЛОЖЕНИЕ

Таблица П1

Коэффициенты эластичности для ряда математических функций

Вид функции y	Первая производная y'_x	Коэффициент эластичности, $Э = y'_x \frac{x}{y}$
Линейная $y = a + bx + \varepsilon$	b	$Э = \dfrac{bx}{a + bx}$
Парабола 2-го порядка $y = a + bx + cx^2 + \varepsilon$	b+2cx	$Э = \dfrac{(b + 2cx)x}{a + bx + cx^2}$
Гипербола $y = a + \frac{b}{x} + \varepsilon$	$-\dfrac{b}{x^2}$	$Э = \dfrac{-b}{ax + b}$
Показательная $y = ab^x\varepsilon$	ln$b$*a*b^x	$Э = x\ln b$
Степенная $y = ax^b\varepsilon$	abx^{b-1}	$Э = b$
Полулогарифмическая $y = a + b\ln x + \varepsilon$	$\dfrac{b}{x}$	$Э = \dfrac{b}{a + b\ln x}$
Логистическая $y = \frac{a}{1+be^{-cx+\varepsilon}}$	$\dfrac{abce^{-cx}}{(1 + be^{-cx})^2}$	$Э = \dfrac{cx}{\frac{1}{b}e^{cx} + 1}$
Обратная $y = \frac{1}{a+bx+\varepsilon}$	$\dfrac{-b}{(a + bx)^2}$	$Э = \dfrac{-bx}{a + bx}$
Экспонента $y = e^{a+bx}$	$e^{a+bx} \cdot b$	$Э = b \cdot x$

Таблица П2

Значения F-критерия Фишера при уровне значимости α = 0,05

n-m-1	Число независимых факторов									
	1	2	3	4	5	6	8	12	24	∞
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
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

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
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	1,76	1,48
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

Окончание табл. П2

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
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
∞	3,84	2,99	2,60	2,37	2,21	2,09	1,94	1,75	1,52	1,00

Таблица П3

Критические значения *t*-критерия Стьюдента при уровне значимости 0,10; 0,05; 0,01

Число степеней свободы	α			Число степеней свободы	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	∞	1,6449	1,9600	2,5758

Таблица П4

Верхнее и нижнее значения критерия Дарбина – Уотсона на уровне значимости $\alpha = 0{,}05$

n	$p = 1$		$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	$d_{Н}$	$d_{\text{В}}$	$d_{Н}$	$d_{\text{В}}$	$d_{Н}$	$d_{\text{В}}$	$d_{Н}$	$d_{\text{В}}$	$d_{Н}$	$d_{\text{В}}$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
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				
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
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	1,78	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	0,85	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	1,99
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

Окончание табл. П4

1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,50	1,31	1,57	1,34	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78

Реализация типовых заданий с помощью MS Excel

Парная регрессия

Встроенная статистическая функция ЛИНЕЙН определяет параметры линейной регрессии $y = a + bx$. Порядок вычисления следующий:

1) введите исходные данные или откройте существующий файл, содержащий анализируемые данные;

2) выделите область пустых ячеек 5×2 (5 строк, 2 столбца) для вывода результатов регрессионной статистики или область 1×2 – для получения только оценок коэффициентов регрессии;

3) активизируйте Мастер функций любым из способов:

а) в главном меню выберите Вставка/Функция;

б) на панели инструментов Стандартная щелкните по кнопке Вставка функции;

4) в окне Категория выберите Статистические, в окне Функция – ЛИНЕЙН. Щелкните по кнопке ОК;

5) заполните аргументы функции:

Известные_значения_у – диапазон, содержащий данные результативного признака.

Известные значения_х – диапазон, содержащий данные факторов независимого признака;

Константа – логическое значение, которое указывает на наличие или на отсутствие свободного члена в уравнении; если *константа* = 1, то свободный член рассчитывается обычным образом, если *константа* = 0, то свободный член равен 0;

Статистика – логическое значение, которое указывает, выводить дополнительную информацию по регрессионному анализу или нет. Если *статистика* = 1, то дополнительная информация выводится, если *статистика* = 0, то выводятся только оценки параметров уравнения. Щелкните по кнопке ОК.

6) в левой верхней ячейке выделенной области появится первый элемент итоговой таблицы. Чтобы раскрыть всю таблицу, нажмите на клавишу <F2>, а затем на комбинацию клавиш <CTRL>+<SHIFT>+<ENTER>.

Дополнительная регрессионная статистика будет выводиться в порядке, указанном в табл. П5.

Таблица П5

Вывод регрессионной статистики

Значение коэффициента b	Значение коэффициента a
Среднеквадратическое отклонение b	Среднеквадратическое отклонение a
Коэффициент детерминации R^2	Среднеквадратическое отклонение y
F-статистика	Число степеней свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов

2. С помощью инструмента анализа данных Регрессия, помимо результатов регрессионной статистики, дисперсионного анализа и доверительных интервалов, можно получить остатки и графики подбора линии регрессии, остатков и нормальной вероятности. Порядок действий следующий:

1) проверьте доступ к пакету анализа. В главном меню последовательно выберите Сервис /Надстройки. Установите флажок Пакет анализа;

2) в главном меню выберите Сервис/Анализ данных/ Регрессия. Щелкните по кнопке ОК;

3) заполните диалоговое окно ввода данных и параметров вывода:

Входной интервал Y – диапазон, содержащий данные результативного признака;

Входной интервал X – диапазон, содержащий данные факторов независимого признака;

Метки – флажок, который указывает, содержит ли первая строка названия столбцов или нет;

Константа – ноль – флажок, указывающий на наличие или отсутствие свободного члена в уравнении;

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона;

Новый рабочий лист – можно задать произвольное имя нового листа.

Если необходимо получить информацию и графики остатков, установите соответствующие флажки в диалоговом окне. Щелкните по кнопке ОК.

Множественная регрессия

Сводную таблицу основных статистических характеристик для одного или нескольких массивов данных можно получить с помощью инструмента анализа данных Описательная статистика. Для этого выполните следующие шаги:

1) введите исходные данные или откройте существующий файл, содержащий анализируемые данные;

2) в главном меню выберите последовательно пункты Сервис/Анализ данных/Описательная статистика, после чего щелкните по кнопке ОК;

3) заполните диалоговое окно ввода данных и параметров вывода:

Входной интервал – диапазон, содержащий анализируемые данные, это может быть одна или несколько строк (столбцов);

Группирование – по столбцам или по строкам – необходимо указать дополнительно;

Метки – флажок, который указывает, содержит ли первая строка названия столбцов или нет;

Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона;

Новый рабочий лист – можно задать произвольное имя нового листа.

Если необходимо получить дополнительную информацию *Итоговой статистики, Уровня надежности, k-го наибольшего и наименьшего значений,* установите соответствующие флажки в диалоговом окне. Щелкните по кнопке ОК.

В ППП MS Excel нет специального инструмента для расчета линейных коэффициентов частной корреляции. Матрицу парных коэффициентов корреляции переменных можно рассчитать, используя инструмент анализа данных Корреляция.

Для этого:

1) в главном меню последовательно выберите пункты Сервис/Анализ данных/Корреляция. Щелкните по кнопке ОК;

2) заполните диалоговое окно ввода данных и параметров вывода;

3) результаты вычислений – матрица коэффициентов парной корреляции.

3. Вычисление параметров линейного уравнения множественной регрессии.

Эта операция проводится с помощью инструмента анализа данных Регрессия. Она аналогична расчету параметров парной линейной регрессии, описанной в первой лабораторной работе, только, в отличие от парной регрессии, в диалоговом окне при заполнении параметра *входной интервал X* следует указать не один столбец, а все столбцы, содержащие значения факторных признаков.

Временные ряды

Для определения параметров линейного тренда по методу наименьших квадратов используется статистическая функция ЛИНЕЙН, для определения экспоненциального тренда ЛГРФПРИБЛ. В качестве зависимой переменной в данной лабораторной работе выступает время.

Построение графиков осуществляется с помощью Мастера диаграмм.

Порядок построения следующий:

1) введите исходные данные или откройте существующий файл, содержащий анализируемые данные;

2) активизируйте Мастер диаграмм любым из следующих способов:

а) в главном меню выберите Вставка/Диаграмма;

б) на панели инструментов Стандартная щелкните по кнопке Мастер диаграмм;

3) в окне *Тип* выберите График; вид графика выберите в поле рядом со списком типов. Щелкните по кнопке Далее;

4) заполните диапазон данных. Установите флажок размещения данных в столбцах (строках). Щелкните по кнопке Далее;

5) заполните параметры диаграммы на разных закладках: названия диаграммы и осей, значения осей, линии сетки, параметры легенды, таблица и подписи данных. Щелкните по кнопке Далее;

6) укажите место размещения диаграммы на отдельном или на имеющемся листе. Щелкните по кнопке Далее.

Готовая диаграмма, отражающая динамику уровней изучаемого ряда, будет готова.

В MS Excel линия тренда может быть добавлена в диаграмму с областями гистограммы или в график.

Для этого:

1) выделите область построения диаграммы; в главном меню выберите Диаграмма/Добавить линию тренда;

2) в появившемся диалоговом окне выберите вид линии тренда и задайте соответствующие параметры. Для полиномиального тренда необходимо задать степень аппроксимирующего полинома, для скользящего среднего – количество точек усреднения.

В качестве дополнительной информации на диаграмме можно отобразить уравнение регрессии и значение среднеквадратического отклонения, установив соответствующие флажки на закладке Параметры. Щелкните по кнопке ОК.

Сравнивая значения R^2 по разным уравнениям трендов (полиномиальный, экспоненциальный, линейный, степенной и логарифмический), можно выбрать функцию, лучше всего описывающую исходный временной ряд.