Набережночелнинский институт
федерального государственного автономного образовательного
учреждения высшего образования
«КАЗАНСКИЙ (ПРИВОЛЖСКИЙ)
ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»

Кафедра
«Бизнес-информатика и математические методы в экономике»

ЭКОНОМЕТРИКА

Электронный образовательный ресурс

Набережные Челны, 2018

УДК 330.43(075.8)
ББК 65в631я73
Ф33

Печатается по решению редакционно-издательского совета
Набережночелнинского института
федерального государственного автономного образовательного учреждения
высшего образования
«Казанский (Приволжский) федеральный университет»,
от 05.05.2016 г. (протокол № 10)

Рецензенты:

д.ф.-м.н., профессор Набережночелнинского института Казанского федерального университета Габбасов Н.С.

к.п.н., доцент Набережночелнинского филиала Казанского инновационного университета имени В.Г. Тимирясова (ИЭУП) Титова С.В.

Федоров Д.Ф. Эконометрика: Электронный образовательный ресурс / Д.Ф. Федоров, А.К. Розенцвайг, А.Н. Карамышев, И.Ф. Назмиев. – Набережные Челны: Изд-во Набережночелнинского института КФУ, 2018. – 104 с.

Электронный образовательный ресурс содержит последовательное изложение методов количественного анализа и моделирования экономических процессов, а также краткие методические указания по решению типовых практических задач, в том числе с помощью пакета прикладных программ MS Excel.

Предназначен для студентов экономических специальностей, а также для тех, кто изучает эту дисциплину самостоятельно.

Содержание

Предисловие

Экономист, не владеющий методами эконометрики, не может эффективно работать аналитиком. Менеджер, не понимающий значение этих методов, обречен на принятие ошибочных решений.

Курс «Эконометрика» включен в учебный план специальности 080507.65 – «Менеджмент организации» и связывает экономическую теорию, прикладные экономические исследования и практику. Благодаря эконометрике осуществляется обмен информацией между этими взаимодополняющими областями, происходит взаимное обогащение и взаимное развитие теории и практики.

Эконометрика дает методы экономических измерений, а также методы оценки параметров моделей микро- и макроэкономики. При этом экономические зависимости выражаются в виде математических соотношений, а затем проверяются эмпирически на достоверность статистическими методами.

Пособие содержит курс лекций по основным разделам эконометрики: парная и множественная регрессия, решения типовых задач по указанным разделам.

По всем разделам представлены тесты и варианты контрольных работ. Для выполнения контрольных заданий по 31 вариантам рассмотрены типовые задачи.

Учебное пособие предназначено для студентов дневной формы обучения, но может быть полезно студентам заочной и дистанционной форм обучения для самостоятельного изучения дисциплины.

Учебный материал пособия условно разбит на две части. В первой части рассмотрены модели парной регрессии. Во второй части достаточно подробно разбирается модель множественной линейной регрессии.

Введение

Переход к рыночной экономике повышает требования к качеству подготовки экономистов, которые, чтобы быть конкурентоспособными и востребованными на рынке труда, должны владеть количественными методами анализа в экономике. При этом высокий динамизм происходящих в стране социально-экономических процессов приводит к тому, что знания об экономике отстают от потребностей управления. В связи с этим деятельность экономиста должна содержать прогностическую составляющую, обеспечивающую возможность заранее сигнализировать о наступлении тех или иных «особых» ситуаций. Сегодня нужны специалисты, не только владеющие опытом и знаниями предыдущих поколений, но и готовые к встрече с новыми постановками задач, обусловленными спецификой России. Из указанного выше следуют новые требования к статистической, эконометрической подготовке экономистов.

В связи с этим дисциплина «Эконометрика» сегодня входит в учебные планы подготовки экономистов всех специальностей и направлений в качестве базовой, обязательной дисциплины и преподается как во всех ведущих университетах мира, так и в отечественных вузах.

Эконометрика рассматривается как дисциплина, объединяющая совокупность результатов, методов и приемов экономической теории, экономической статистики и математико-статистического инструментария для количественного выражения качественных закономерностей. Курс эконометрики призван научить различным способам выражения связей и закономерностей через эконометрические модели, основанные на данных статистических наблюдений. Эконометрический подход предусматривает анализ соответствия выбранной модели изучаемому объекту, рассмотрение причин, приводящих к необходимости пересмотра моделей на основе более точной системы представлений. Эконометрика занимается, по существу,

статистическими выводами, т. е. использованием выборочных данных, для получения некоторого представления о свойствах генеральной совокупности.

Учебный курс «Эконометрика» опирается на курсы «Микроэкономика», «Макроэкономика», «Теория статистики», «Теория вероятностей и математическая статистика». В свою очередь, «Эконометрика» выступает в качестве базы для курса «Эконометрическое моделирование», а также продвинутых курсов прикладной микро- и макроэкономики. При этом применение методов эконометрики позволяет осуществить проверку справедливости положений экономической теории.

Назначение эконометрики мы видим в придании конкретного количественного выражения общим (качественным) закономерностям экономической теории на базе экономической статистики с использованием математико-статистического инструментария.

При эконометрическом подходе исследователь строит рассуждения и выводы, опираясь в своих модельных построениях на вероятностно-статистический подход и результаты конкретных измерений интересующих его социально-экономических показателей.

Вероятностно-статистическая модель — это математическая модель, имитирующая механизм функционирования гипотетического (не конкретного) реального явления стохастической природы, значения отдельных параметров которой оцениваются по результатам наблюдений, исходным статистическим данным, характеризующим функционирование моделируемого конкретного (а не гипотетического) явления.

Построение и экспериментальная проверка вероятностно-статистической модели обычно основаны на одновременном использовании априорной информации о природе и содержательной сущности анализируемого явления, представленной в виде теоретических закономерностей и исходных статистических данных, характеризующих процесс и результаты функционирования изучаемого явления.

Вероятностно-статистическое моделирование включает следующие этапы:

1. Постановка задачи, а именно определение конечных прикладных целей моделирования; набора показателей, взаимосвязи между которыми нас интересуют, а также группировка этих показателей в рамках поставленной задачи на входные или объясняющие, которые полностью или частично регулируемы и поддаются регистрации и прогнозу, и выходные или объясняемые, которые формируются в процессе функционирования моделируемой системы и обычно трудно поддаются непосредственному прогнозу.

2. Априорный, предмодельный анализ содержательной сущности моделируемого явления состоит в формировании и формализации имеющейся априорной информации об этом явлении в виде ряда гипотез и исходных допущений.

3. Информационно-статистический этап посвящен сбору необходимой статистической информации, т.е. регистрации значений показателей, участвующих в описании модели, различных моментов времени и (или) точек пространства функционирования моделируемой системы.

4. Этап спецификации модели включает в себя получение общего вида модельных соотношений, связывающих между собой интересующие нас входные и выходные переменные. На данном этапе определяют лишь структуру модели, ее аналитическую запись, в которой наряду с известными будут присутствовать величины, содержательный смысл которых определен, а числовые значения - нет. Это параметры модели, неизвестные значения которых подлежат статистическому оцениванию.

Этап идентификации модели предназначен для проведения статистического анализа модели с целью «настройки» значений ее неизвестных параметров на полученные нами статистические данные. При этом предварительно необходимо ответить на вопрос о возможности оценки неизвестных параметров модели по имеющимся исходным статистическим

данным и определенной на этапе спецификации структуре модели. Это вопрос об идентифицируемости модели. После положительного ответа на этот вопрос переходят к задаче идентификации модели, т. е. оцениванию неизвестных значений параметров модели по имеющимся исходным статистическим данным.

Если проблема идентифицируемости решается отрицательно, то возвращаются к этапу 4 и вносят необходимую корректировку в решение задачи спецификации модели.

6. Этап верификации модели, анализа ее точности и адекватности заключается в использовании различных процедур сопоставления модельных заключений, оценок и выводов с реально наблюдаемой действительностью. При отрицательных результатах этого этапа необходимо возвратиться к этапу 4, а иногда и к этапу 1.

Модели, построение которых основано только на априорной информации и не предусматривает проведения этапов 3, 5 и 6, называются экономико-математическими. Построение эконометрической модели требует проведения всех шести этапов.

7. Интерпретация полученных результатов, т.е. перевод их с формализованного языка математики на содержательный язык рекомендаций по принятию управленческих решений.

Однако далеко не всегда целевые установки исследователей подкреплены объективными возможностями их реализации.

Можно выделить три основных типа целей подобных исследований:

1. Установление факта наличия или отсутствия статистически значимой связи между X и Y. При такой постановке задачи статистический вывод сводится к утверждениям: «связь есть» или «связи нет» и сопровождается обычной численной характеристикой степени тесноты исследуемой зависимости. Выбор вида функции $f(X)$ и состава объясняющих переменных X нацелен исключительно на максимизацию величины этого измерителя тесноты связи. Такие задачи решаются методами корреляционного анализа.

2. Построение прогноза неизвестных индивидуальных Y(X) и средних $\bar{Y}(X)$ значений результативных показателей по заданным значениям X объясняющих переменных. При этом статистический вывод включает в себя как точечный, так и интервальный прогноз результативного показателя Y(X) или $\bar{Y}(X)$, который сопровождается величиной доверительной вероятности. Как и в предыдущем случае, выбор формы связи, конкретного вида функции *f*(X) и состава объясняющих переменных X играет вспомогательную роль и нацелен исключительно на минимизацию ошибки получаемого прогноза. Однако в этом случае существенно используются значения функции *f*(X), которые являются отправной точкой при построении прогноза.

3. Выявление причинных связей между объясняющими переменными X и результативными показателями Y, управление значениями Y путем регулирования величин объясняющих переменных X. Такая постановка задачи претендует на проникновение в механизм преобразования объясняющих переменных X и ε в результативные показатели Y. При этом на первый план выходит задача правильного выбора структуры модели (выбора общего вида функции *f*(X)), решение которой обеспечивает возможность количественного измерения эффекта воздействия на Y каждой из объясняющих переменных x_1, x_2,..., x_k в отдельности.

Однако задача правильного выбора общего вида функции *f*(X) и является самым слабым местом во всей технике статистического исследования зависимостей. К сожалению, не существует стандартных методов, строгой теоретической базы для решения этой важнейшей задачи.

С этой целью подбирают класс функций, связывающий результативный показатель *Y* с объясняющими переменными x_1, x_2,..., x_k, отбирают из них наиболее информативные объясняющие переменные, определяют неизвестные значения параметров уравнения связи и анализируют точность получения уравнения.

1. Парная регрессия и корреляция

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

$$\hat{y}_i = M(y / x = x_i) + \varepsilon_i,$$

где y – зависимая переменная (результативный признак); x – независимая, или объясняющая, переменная (признак-фактор). Знак «^» означает, что между переменными x и y нет строгой функциональной зависимости, поэтому практически в каждом отдельном случае величина y складывается из двух слагаемых:

$$y_i = \hat{y}_x + e_i,$$

где y – фактическое значение результативного признака; $\hat{y}_x$ – эмпирическое значение результативного признака, найденное исходя из уравнения регрессии; e_i − случайная величина, характеризующая отклонения реального значения результативного признака от эмпирического, найденного по уравнению регрессии.

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

От правильно выбранной спецификации модели зависит величина случайных ошибок: они тем меньше, чем в большей мере эмпирические значения результативного признака $\hat{y}_x$ совпадают с фактическими данными y.

К ошибкам спецификации относятся неправильный выбор той или иной математической функции для $\hat{y}_x$ и недоучет в уравнении регрессии какого-либо существенного фактора.

Наряду с ошибками спецификации могут иметь место ошибки выборки, которые возникают в силу неоднородности данных в исходной статистической совокупности, извлекаемой случайным образом из генеральной совокупности, что, как правило, бывает при изучении экономических процессов.

Использование временной информации также представляет собой выборку из всего множества хронологических данных. Изменив временной интервал, можно получить другие результаты моделирования.

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

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

Предполагая, что ошибки измерения сведены к минимуму, основное внимание в эконометрических исследованиях уделяется ошибкам спецификации модели.

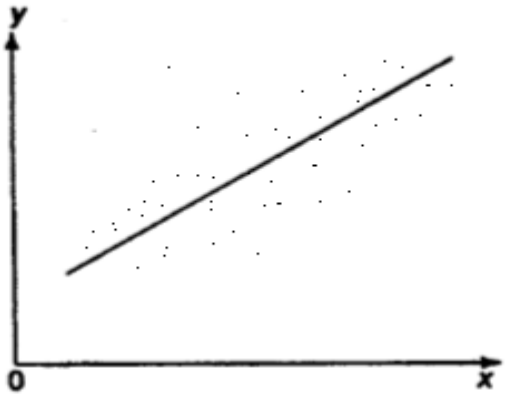
В парной регрессии выбор вида математической функции $\hat{y}_x = f(x)$ может быть осуществлен тремя методами:

1) графическим;

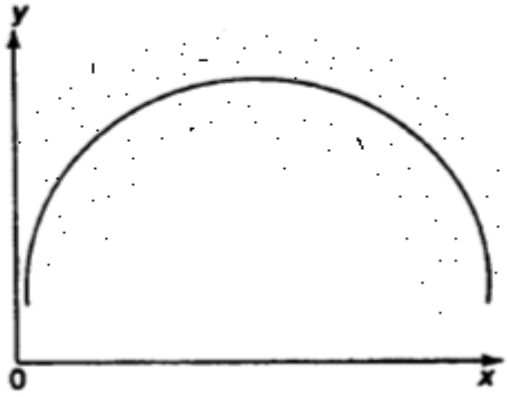
2) аналитическим, т.е. исходя из теории изучаемой взаимосвязи;

3) экспериментальным.

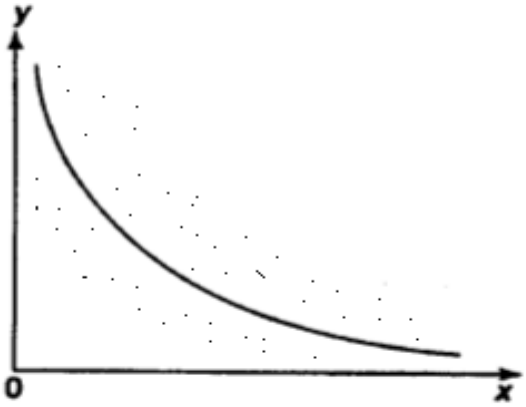
При изучении зависимости между двумя признаками графический метод подбора вида уравнения регрессии достаточно нагляден. Он основан на поле корреляции. Основные типы кривых, используемые при количественной оценке связей, представлены на рис. 1.1:



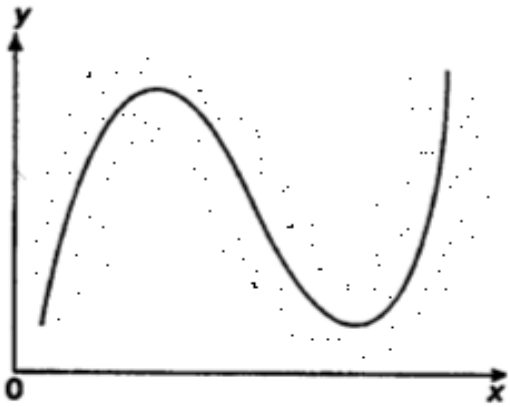
$$\hat{y}_x = a + b \cdot x$$



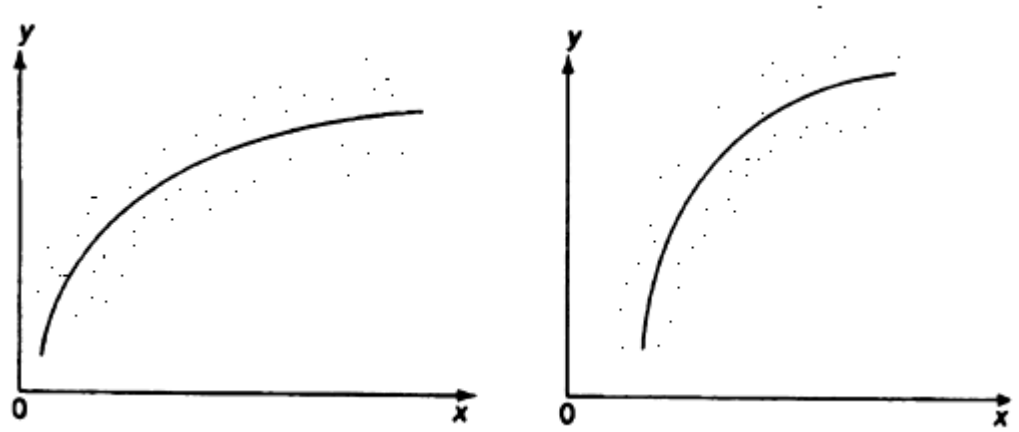
$$\hat{y}_x = a + b \cdot x + c \cdot x^2$$



$$\hat{y}_x = a + b/x$$



$$\hat{y}_x = a + b \cdot x + c \cdot x^2 + d \cdot x^3$$



$\hat{y}_x = a \cdot x^b$ $\hat{y}_x = a \cdot b^x$

Рис. 1.1. Основные типы кривых, используемые при количественной оценке связей между двумя переменными

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

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

Если уравнение регрессии проходит через все точки корреляционного поля, что возможно только при функциональной связи, когда все точки лежат на линии регрессии $\hat{y}_x = f(x)$, то фактические значения результативного признака совпадают с теоретическими $y = \hat{y}_x$, т.е. они полностью обусловлены влиянием фактора x. В этом случае остаточная дисперсия $\sigma^2_{\text{ост}} = 0.$

В практических исследованиях, как правило, имеет место некоторое рассеяние точек относительно линии регрессии. Оно обусловлено влиянием прочих, не учитываемых в уравнении регрессии, факторов. Иными словами,

имеют место отклонения фактических данных от теоретических $(y_i - \bar{y}_{x_i})$. Величина этих отклонений и лежит в основе расчета остаточной дисперсии:

$$\sigma^2_{ост} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{x_i})^2 .$$

Чем меньше величина остаточной дисперсии, тем меньше влияние не учитываемых в уравнении регрессии факторов и тем лучше уравнение регрессии соответствует исходным данным.

Считается, что число наблюдений должно n >3m+1, где n – объем выборки, а m – количество объясняющих переменных *x*. Это означает, что искать линейную регрессию, имея менее 5 наблюдений, вообще не имеет смысла. Если вид функции усложняется, то требуется увеличение объема наблюдений.

1.1. Линейная модель парной регрессии и корреляции

Рассмотрим простейшую модель парной регрессии – линейную регрессию. Линейная регрессия находит широкое применение в эконометрике ввиду явной экономической интерпретации ее параметров.

Линейная регрессия имеет теоретическое уравнение вида:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i . \qquad (1.1)$$

Уравнение вида $\hat{y}_i = b_0 + b_1 x_i$ позволяет по заданным значениям фактора x находить теоретические значения результативного признака, подставляя в него фактические значения фактора x.

Построение линейной регрессии сводится к оценке ее параметров – b_0 и b_1. Классический подход к оцениванию параметров линейной регрессии основан на методе наименьших квадратов (МНК). МНК позволяет получить такие оценки параметров b_0 и b_1, при которых сумма квадратов отклонений

фактических значений результативного признака y от теоретических $\hat{y}_x$ минимальна:

$$\sum_{i=1}^{n}(y_i - \bar{y}_{x_i})^2 = \sum_{i=1}^{n} e_i^{\,2} \to \min \ . \qquad (1.2)$$

Т.е. из всего множества линий линия регрессии на графике выбирается так, чтобы сумма квадратов расстояний по вертикали между точками и этой линией была бы минимальной (рис. 1.2):

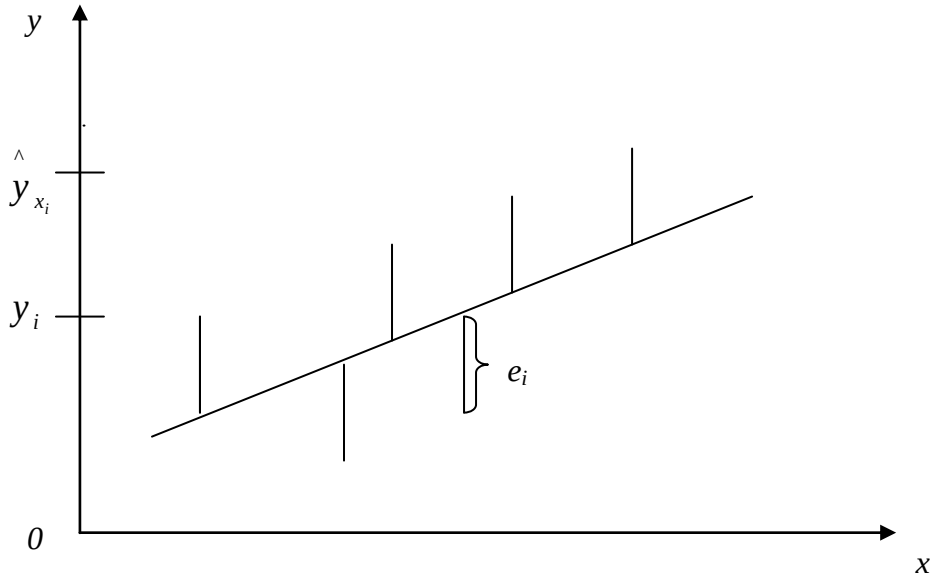


Рис. 1.2. Линия регрессии с минимальной дисперсией остатков

Как известно из курса математического анализа, чтобы найти минимум функции (1.2), надо вычислить частные производные по каждому из параметров b_0 и b_1 и приравнять их к нулю. Обозначим $\sum_{i=1}^{n} e_i^2$ через $Q(b_0, b_1)$, тогда:

$$Q(b_0, b_1) = \sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 \ .$$

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = 0; \\ \dfrac{\partial Q}{\partial b_1} = -2\sum_{i=1}^{n} x_i (y_i - b_0 - b_1 x_i) = 0. \end{cases} \quad (1.3)$$

После несложных преобразований получим следующую систему линейных уравнений для оценки параметров b_0 и b_1:

$$\begin{cases} b_0 n + b_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i . \end{cases} \quad (1.4)$$

Решая систему уравнений (1.4), найдем искомые оценки параметров b_0 и b_1. Можно воспользоваться следующими готовыми формулами, которые следуют непосредственно из решения системы (1.4):

$$b_0 = \bar{y} - b_1 \bar{x},$$

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{\overline{xy} - \bar{x} * \bar{y}}{\overline{x^2} - (\bar{x})^2}, \quad (1.5)$$

где $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$; $\bar{yx} = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$; $\bar{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

$S_{xy} = \overline{y \cdot x} - \bar{y} \cdot \bar{x}$ – ковариация признаков x и y , $S_{xy}^2 = \overline{x^2} \cdot \bar{x}^2$ – дисперсия признака x.

Ковариация – числовая характеристика совместного распределения двух случайных величин, равная математическому ожиданию произведения отклонений этих случайных величин от их математических ожиданий.

Дисперсия – характеристика случайной величины, определяемая как математическое ожидание квадрата отклонения случайной величины от ее математического ожидания.

Математическое ожидание – сумма произведений значений случайной величины на соответствующие вероятности.

Параметр b_1 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 достаточно распространенным в эконометрических исследованиях.

Формально b_0 – значение y при $x = 0$. Если признак-фактор x не может иметь нулевого значения, то вышеуказанная трактовка свободного члена b_0 не имеет смысла, т.е. параметр b_0 может не иметь экономического содержания.

1.2. Нелинейные модели регрессии и их линеаризация

При нелинейной зависимости признаков, приводимой к линейному виду, параметры множественной регрессии также определяются по МНК с той лишь разницей, что он используется не к исходной информации, а к преобразованным данным. Так, рассматривая степенную функцию

$$\hat{y}_i = b_0 \cdot x_i^{b_1},$$

мы преобразовываем ее в линейный вид:

$$\lg \hat{y}_i = \lg b_0 + b_1 \cdot \lg x_i,$$

где переменные выражены в логарифмах.

Далее обработка МНК та же: строится система нормальных уравнений и определяются неизвестные параметры. Потенцируя значение $\lg b_0$, находим параметр b_0 и соответственно общий вид уравнения степенной функции.

Вообще говоря, нелинейная регрессия по включенным переменным не таит каких-либо сложностей в оценке ее параметров. Эта оценка определяется, как и в линейной регрессии, МНК, рассмотренным выше.

$$\hat{y}_i = b_0 + b_1 \cdot x_i .$$

1.3. Показатели качества уравнения парной регрессии

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает линейный коэффициент корреляции r_{xy}, который рассчитывается по следующей формуле:

$$r_{xy} = b_1 \cdot \frac{S_{xy}}{S_x \cdot S_y} \qquad (1.6)$$

Линейный коэффициент корреляции находится в пределах: $-1 \le r_{xy} \le 1$. Чем ближе абсолютное значение r_{xy} к единице, тем сильнее линейная связь между факторами (при $r_{xy} = \pm 1$ имеем строгую функциональную зависимость). Но следует иметь в виду, что близость абсолютной величины линейного коэффициента корреляции к нулю еще не означает отсутствия связи между признаками. При другой (нелинейной) спецификации модели связь между признаками может оказаться достаточно тесной.

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r_{xy}^2, называемый коэффициентом детерминации. Коэффициент детерминации характеризует долю дисперсии результативного признака y, объясняемую уравнением регрессии, в общей дисперсии результативного признака:

$$r_{xy}^2 = 1 - \frac{S_0^2}{S_y^2} \qquad (1.7)$$

где $S_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{x_i})^2$, $S_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2 = \bar{y^2} - (\bar{y})^2$.

Соответственно величина $1 - r_{xy}^2$ характеризует долю дисперсии y , вызванную влиянием остальных, не учтенных в модели, факторов.

После того как оценено уравнение линейной регрессии, проводится проверка значимости как уравнения в целом, так и отдельных его параметров.

Проверить значимость уравнения регрессии – значит установить, соответствует ли математическая модель, выражающая зависимость между переменными, экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных (одной или нескольких) для описания зависимой переменной.

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

$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_{x_i}}{y_i}\right| * 100\% \; . \qquad (1.8)$$

Средняя ошибка аппроксимации не должна превышать 8–10%.

Оценка значимости уравнения регрессии в целом производится на основе F -критерия Фишера, которому предшествует дисперсионный анализ. В математической статистике дисперсионный анализ рассматривается как самостоятельный инструмент статистического анализа. В эконометрике он применяется как вспомогательное средство для изучения качества регрессионной модели.

Согласно основной идее дисперсионного анализа, общая сумма квадратов отклонений переменной y от среднего значения $\bar{y}$ раскладывается на две части – «объясненную» и «необъясненную»:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_{x_i}-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_{x_i})^2,$$

где $\sum_{i=1}^{n}(y_i-\bar{y})^2$ – общая сумма квадратов отклонений; $\sum_{i=1}^{n}(\hat{y}_{x_i}-\bar{y})^2$ – сумма квадратов отклонений, объясненная регрессией (или факторная сумма квадратов отклонений); $\sum_{i=1}^{n}(y_i-\hat{y}_{x_i})^2$ – остаточная сумма квадратов отклонений, характеризующая влияние неучтенных в модели факторов.

Схема дисперсионного анализа имеет вид, представленный в таблице 1.1 (n – число наблюдений, m – число параметров при переменной x).

Таблица 1.1

Компоненты дисперсии	Сумма квадратов	Число степеней свободы	Дисперсия на одну степень свободы
Общая	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$	$S^2_{общ.}=\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n-1}$
Факторная	$\sum_{i=1}^{n}(\hat{y}_{x_i}-\bar{y})^2$	m	$S^2_{факт.}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{m}$
Остаточная	$\sum_{i=1}^{n}(y_i-\hat{y}_{x_i})^2$	$n-m-1$	$S^2_0=\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{n-m-1}$

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получим величину F-критерия Фишера:

$$F_{расч.}=\frac{S^2_{факт.}}{S^2_0}. \qquad (1.9)$$

Расчетное значение F – критерия Фишера (1.9) сравнивается с табличным значением $F_{\text{табл}}(\alpha;k_1;k_2)$ при уровне значимости α

(зафиксированное значение ошибки I рода, состоящей в том, чтобы на основании данных выборочного исследования принять альтернативную гипотезу) и степенях свободы $k_1 = m$ и $k_2 = n - m - 1$. При этом, если фактическое значение F – критерия больше табличного, то признается статистическая значимость уравнения в целом.

Для парной линейной регрессии $m = 1$, поэтому

$$F_{расч.} = \frac{S^2_{факт.}}{S^2_0} = \frac{\sum_{i=1}^{n} (\hat{y}_{x_i} - \bar{y})^2}{\sum_{i=1}^{n} (y_i - \hat{y}_{x_i})^2} * (n-2). \qquad (1.10)$$

Величина F-критерия связана с коэффициентом детерминации r^2_{xy}, и ее можно рассчитать по следующей формуле:

$$F = \frac{r^2_{xy}}{1 - r^2_{xy}} \cdot (n-2). \qquad (1.11)$$

Оценивается значимость не только уравнения в целом, но и отдельных его параметров. С этой целью по каждому из параметров определяется его стандартная ошибка: S_{b0}, S_{b1}.

Стандартная ошибка коэффициента регрессии определяется по формуле:

$$S_{b_1} = \sqrt{\frac{S^2_0}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{S^2_0}{\sigma_x * \sqrt{n}}, \qquad (1.12)$$

где $S^2_0 = \frac{\sum_{i=1}^{n} e_i^{\,2}}{n-2}$ – 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: $t_{b_1} = \frac{b_1}{S_{b_1}}$, которое затем сравнивается с

табличным значением при определенном уровне значимости α и числе степеней свободы $(n-2)$. Доверительный интервал для коэффициента регрессии определяется как $b_1 \pm t_{табл.} \cdot S_{b_1}$. Поскольку знак коэффициента регрессии указывает или на рост результативного признака y при увеличении признака-фактора x ($b>0$), или на уменьшение результативного признака при увеличении признака-фактора ($b<0$), или его независимость от объясняющей переменной ($b=0$) (рис. 1.3), то границы доверительного интервала для коэффициента регрессии не должны содержать противоречивых результатов, например, $-1,5 \le b \le 0,8$. Такого рода запись указывает, что истинное значение коэффициента регрессии одновременно содержит положительные и отрицательные величины и даже ноль, чего не может быть.

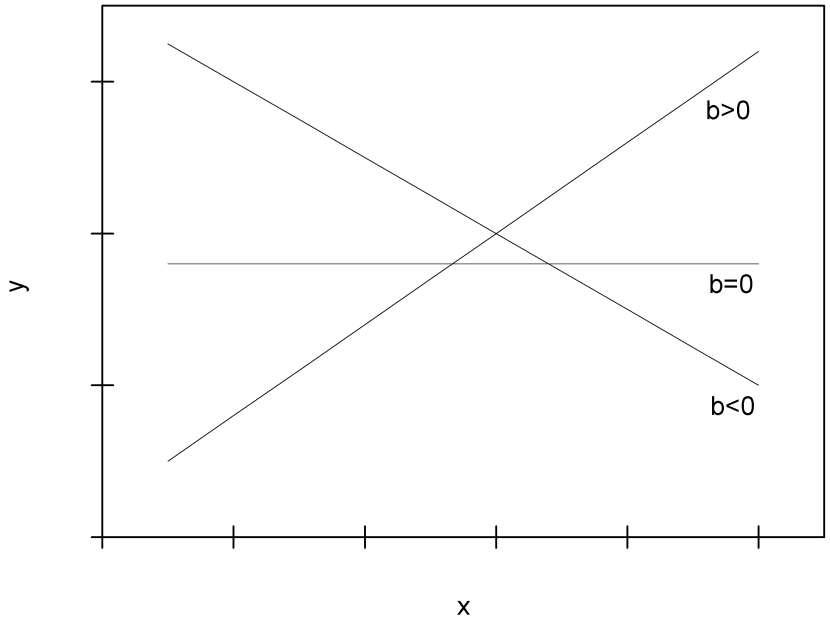


Рис. 1.3**.** Наклон линии регрессии в зависимости от значения параметра b_1

Стандартная ошибка параметра определяется по формуле:

$$S_{b_0} = \sqrt{\frac{S_0^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{S_0^2 \cdot \sum_{i=1}^{n} x_i^2}{\sigma_x \cdot n}. \qquad (1.13)$$

Процедура оценивания существенности данного параметра не отличается от рассмотренной выше для коэффициента регрессии. Вычисляется t-критерий: $t_{b_0} = \frac{b_0}{S_{b_0}}$, его величина сравнивается с табличным значением при $n-2$ степенях свободы.

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

$$S_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}}. \qquad (1.14)$$

Фактическое значение t– критерия Стьюдента определяется как $t_r = \frac{r}{S_r}$.

Существует связь между t – критерием Стьюдента и F -критерием Фишера:

$$t_{b_1} = t_r = \sqrt{F}. \qquad (1.15)$$

В прогнозных расчетах по уравнению регрессии определяется предсказываемое $\bar{y}_p$ значение как точечный прогноз $\hat{y}_x$ при $x_p = x_k$, т.е. путем подстановки в уравнение регрессии $\hat{y}_i = b_0 + b_1 x_i$ соответствующего значения x. Однако точечный прогноз явно не реален. Поэтому он дополняется расчетом стандартной ошибки $\bar{y}_p$, т.е. $S(\hat{Y}_p)$, и соответственно интервальной оценкой прогнозного значения $\bar{y}_p$:

$$\bar{y}_p - S(\hat{Y}_p)_{Y_0} \cdot t_{табл.} \le \hat{y}_p \le \bar{y}_p + S(\hat{Y}_p)_{Y_0} \cdot t_{табл.},$$

где $S(\hat{Y}_p)_{Y_0}$ – средняя ошибка прогнозируемого индивидуального значения:

$$S(\hat{Y}_p)_{Y_0} = S_0 \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot \sigma_x^2}} \,. \qquad (1.16)$$

Доверительный интервал для условного математического ожидания рассчитывается по формуле:

$$\bar{y}_p - S(\hat{Y}_p)_{M(Y/x_p)} \cdot t_{табл.} \le \hat{y}_p \le \bar{y}_p + S(\hat{Y}_p)_{M(Y/x_p)} \cdot t_{табл.},$$

где средняя ошибка прогнозируемого индивидуального значения определяется следующим образом:

$$S(\hat{Y}_p)_{M(Y/x_p)} = S_0 \cdot \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot \sigma_x^2}} \,.$$

1.4. Решение типовых задач

Задача 1.2.1

Для данных из таблицы методом наименьших квадратов вычислить уравнение линейной регрессии:

X_i	1	3	4	6	7
Y_i	2	2	6	8	11

Решение: для расчета параметров b_0 и b_1 линейной регрессии $\hat{y}_i = b_0 + b_1 x_i$ рассчитаем следующую таблицу (используя возможности MS Excel):

№п/п	X	Y	XY	X^2	Y^2	Y(x)	e_i
1	1	2	2	1	4	0,8596	1,1404
2	3	2	6	9	4	3,9474	-1,9474
3	4	6	24	16	36	5,4912	0,5088
4	6	8	48	36	64	8,5789	-0,5789
5	7	11	77	49	121	10,1228	0,8772
Сумма	21	29	157	111	229	29	0,00
Ср. значение	4,20	5,80	31,40	22,2	45,80	5,80	0,00

Затем, используя формулы расчета коэффициентов уравнения регрессии, определяем соответствующие их значения:

$$b_1 = \frac{\bar{yx} - \bar{y} * \bar{x}}{\bar{x^2} - (\bar{x})^2} = \frac{31{,}40 - 5{,}80 * 4{,}20}{22{,}2 - (5{,}80)^2} = 1{,}54;$$

$$b_0 = \bar{y} - b_1 \bar{x} = 5{,}80 - 1{,}54 * 4{,}20 = -0{,}68.$$

Таким образом, уравнение линейной регрессии $\hat{y}_i = b_0 + b_1 x_i$ = *-0,68+1,54x_i*.

Задача 1.2.2

Рассчитайте коэффициент корреляции, если уравнение регрессии

y = 7 + 2x, $\sigma_x = 2$, $\sigma_y = 8$.

Решение: тесноту линейной связи уравнения регрессии $\hat{y}_i = b_0 + b_1 x_i$ оценивает коэффициент корреляции

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = 2 * \frac{2}{8} = 0{,}5.$$

Задача 1.2.3

Получено уравнение регрессии y = 3 +3x. Известны $\sigma_x = 2$, $\sigma_y = 8$ и F = 36. На основании скольких наблюдений (n) получено уравнение?

Решение: количество наблюдений мы можем определить исходя из формулы:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} * (n - 2).$$

Для этого нам необходимо определить значение параметра r_{xy}^2:

$$r_{xy} = b\frac{\sigma_x}{\sigma_y} = 3 * \frac{2}{8} = 0{,}75;$$

$$r_{xy}^2 = 0{,}5625;$$

$$n = \frac{F - Fr_{xy}^2 + 2r_{xy}^2}{r_{xy}^2} = \frac{36 - 36 * 0{,}5625 + 2 * 0{,}5625}{0{,}5625} = \frac{16{,}875}{0{,}5625} = 30.$$

Задача 1.2.5.

По данным проведенного опроса восьми групп семей известны расходы населения на продукты питания и уровни доходов семей.

Расходы на продукты питания, y, тыс. руб.	0,9	1,2	1,8	2,2	2,6	2,9	3,3	3,8
Доходы семьи, x, тыс. руб.	1,2	3,1	5,3	7,4	9,6	11,8	14,5	18,7

Требуется:

1. Построить линейное уравнение парной регрессии y(x);

2. Рассчитать линейный коэффициент парной корреляции и среднюю ошибку аппроксимации.

3. Оценить качество уравнения регрессии в целом с помощью F - критерия Фишера.

4. Оценить статистическую значимость параметров регрессии и корреляции.

5. Выполнить прогноз расходов на продукты питания при прогнозном значении признака-фактора доходов семьи, составляющем 110% от среднего уровня.

6. Оценить точность прогноза, рассчитав ошибку прогноза и его доверительный интервал.

Решение: для удобства дальнейших вычислений составим таблицу.

Таблица 1.3

	x	y	$x \cdot y$	x^2	y^2	$\hat{y}_x$	$y - \hat{y}_x$	$\left(y - \hat{y}_x\right)^2$	A_i, %
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	1,2	0,9	1,08	1,44	0,81	1,038	–0,138	0,0190	15,33
2	3,1	1,2	3,72	9,61	1,44	1,357	–0,157	0,0246	13,08
3	5,3	1,8	9,54	28,09	3,24	1,726	0,074	0,0055	4,11
4	7,4	2,2	16,28	54,76	4,84	2,079	0,121	0,0146	5,50
5	9,6	2,6	24,96	92,16	6,76	2,449	0,151	0,0228	5,81
6	11,8	2,9	34,22	139,24	8,41	2,818	0,082	0,0067	2,83
7	14,5	3,3	47,85	210,25	10,89	3,272	0,028	0,0008	0,85
8	18,7	3,8	71,06	349,69	14,44	3,978	–0,178	0,0317	4,68
Итого	71,6	18,7	208,71	885,24	50,83	18,717	–0,017	0,1257	52,19
Среднее значение	8,95	2,34	26,09	110,66	6,35	2,34	–	0,0157	6,52
σ	5,5345	0,9352	–	–	–	–	–	–	–
σ^2	30,5612	0,8741	–	–	–	–	–	–	–

1. Рассчитаем параметры линейного уравнения парной регрессии $\hat{y}_i = b_0 + b_1 x_i$. Для этого воспользуемся формулами (1.5):

$$b_0 = \bar{y} - b_1 \bar{x} = 2{,}34 - 0{,}1684 \cdot 8{,}95 = 8{,}361,$$

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{26{,}09 - 8{,}95 \cdot 2{,}34}{30{,}56} = 0{,}1684. \qquad (1.5)$$

Получили уравнение:

$$\hat{y}_i = b_0 + b_1 x_i = 0{,}8361 + 0{,}1684 \cdot x_i.$$

т.е. с увеличением дохода семьи на 1000 руб. расходы на питание увеличиваются на 168 руб.

2. Как было указано выше, уравнение линейной регрессии всегда дополняется показателем тесноты связи – линейным коэффициентом корреляции r_{xy}:

$$r_{x_i y_i} = b_1 \cdot \frac{S_{x_i}}{S_{y_i}} = 0{,}1684 \cdot \frac{5{,}5345}{0{,}9352} = 0{,}9942 \ .$$

Близость коэффициента корреляции к 1 указывает на тесную линейную связь между признаками.

Средняя ошибка аппроксимации (находим с помощью столбца 10 таблицы 1.3; $A_i = \left|\frac{y_i - \hat{y}_{x_i}}{y_i}\right| \cdot 100\%$; $\bar{A} = 6,52\%$ говорит о хорошем качестве уравнения регрессии, т.е. свидетельствует о хорошем подборе модели к исходным данным.

3. Коэффициент детерминации $r_{xy}^2 = 0,9872$ (тот же результат получим, если воспользуемся формулой (1.7)) показывает, что уравнением регрессии объясняется 98,7% дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,3%.

Оценим качество уравнения регрессии в целом с помощью F-критерия Фишера. Сосчитаем фактическое значение F-критерия:

$$F = \frac{r_{xy}^2}{1 - r_{xy}^2} \cdot (n-2) = \frac{0,987}{1 - 0,987} \cdot 6 = 455,54.$$

Табличное значение ($k_1 = 1$, $k_2 = n - 2 = 6$, $\alpha = 0,05$): $F_{табл.} = 23$. Так как $F_{\text{факт}} > F_{\text{табл}}$, то признается статистическая значимость уравнения в целом.

4. Для оценки статистической значимости коэффициентов регрессии и корреляции рассчитаем t-критерий Стьюдента и доверительные интервалы каждого из показателей. Рассчитаем случайные ошибки параметров линейной регрессии и коэффициента корреляции

$$S_0^2 = \frac{\sum_{i=1}^{n} e_i^{\,2}}{n-2} = \frac{0,1257}{8-2} = 0,021.$$

$$S_{b_1} = \sqrt{\frac{S_0^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{S_0^2}{\sigma_x * \sqrt{n}} = \frac{\sqrt{0,021}}{5,5345 \cdot \sqrt{8}} = 0,0093,$$

$$S_{b_0} = \sqrt{\frac{S_0^2 \cdot \sum_{i=1}^{n} x_i^2}{n \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2}} = \frac{S_0^2 \cdot \sum_{i=1}^{n} x_i^2}{\sigma_x \cdot n} = \frac{\sqrt{0{,}021 \cdot 885{,}2405}}{5{,}5345 \cdot 8} = 0{,}0975,$$

$$S_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}} = \sqrt{\frac{1 - 0{,}987}{6}} = 0{,}0465$$

Фактические значения t-статистик:

$$t_{b_1} = \frac{b_1}{S_{b_1}} = \frac{0{,}1684}{0{,}0093} = 18{,}0651,$$

$$t_{b_0} = \frac{b_0}{S_{b_0}} = \frac{0{,}8361}{0{,}0975} = 8{,}5742,$$

$t_r = \frac{0{,}994}{0{,}0465} = 21{,}376$. Табличное значение t-критерия Стьюдента при $\alpha = 0{,}05$ и числе степеней свободы $\nu = n - 2 = 6$ есть $t_{\text{табл}} = 2{,}447$. Так как $t_{b_0} > t_{табл.}$ $t_{b_1} > t_{табл.}$ и $t_r > t_{\text{табл}}$, то признаем статистическую значимость параметров регрессии и показателя тесноты связи. Рассчитаем доверительные интервалы для параметров регрессии b_0 и b_1: $b_0 \pm t_{табл.} \cdot S_{b_0}$ и $b_1 \pm t_{табл.} \cdot S_{b_1}$. Получим, что $b_0 \in [0{,}5974; 1{,}0751]$ и $b_1 \in [0{,}1454; 0{,}1912]$.

5. И, наконец, найдем прогнозное значение результативного фактора $\hat{y}_p$ при значении признака-фактора, составляющем 110% от среднего уровня $x_p = 1{,}1 \cdot \bar{x} = 1{,}1 \cdot 8{,}95 = 9{,}845$, т.е. найдем расходы на питание, если доходы семьи составят 9,85 тыс. руб.

$$\hat{y}_p = 0{,}836 + 0{,}168 \cdot 9{,}845 = 2{,}490 \text{ (тыс. руб.)}$$

Значит, если доходы семьи составят 9,845 тыс. руб., то расходы на питание будут 2,490 тыс. руб.

Найдем доверительный интервал прогноза. Ошибка прогноза

$$S_{\hat{y}_p} = S_0 \cdot \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{n \cdot S_x^2}} = \sqrt{0{,}021 + \frac{1}{8} + \frac{(9{,}845 - 8{,}95)^2}{8 \cdot 30{,}56}} = 0{,}1540,$$

а доверительный интервал ($\bar{y}_p - S_{\hat{y}_p} \cdot t_{табл.} \le \hat{y}_p \le \bar{y}_p + S_{\hat{y}_p} \cdot t_{табл.}$):

$$2{,}113 < \hat{y}_p < 2{,}867.$$

Прогноз является статистически надежным.

Теперь на одном графике изобразим исходные данные и линию регрессии:

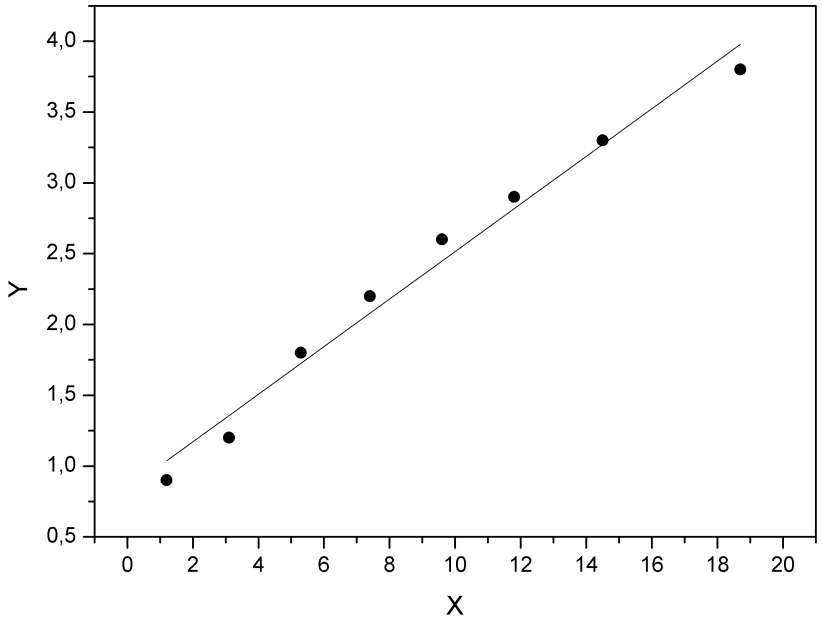


Рис. 1.5

1.5. Упражнения и задачи

Задача 1.3.1

Так называемая кривая Филипса описывает связь темпа роста заработной платы и уровня безработицы. А именно,

$$\delta\omega_t = \beta_1 + \beta_2 \frac{1}{u_t} + \varepsilon_t,$$

где ω_t – уровень заработной платы, $\delta\omega_t = 100(\omega_t - \omega_{t-1})/\omega_{t-1}$ - темп роста заработной платы (в процентах) и u_t – процент безработных в год *t*. Теория предполагает, что $\beta_1 < 0$ и $\beta_2 > 0$.

Используя данные для некоторой страны из таблицы

a) найдите оценки коэффициентов уравнения и проверьте наличие значимой связи между $\delta\omega_t$ и u_t;

b) найдите «естественный уровень безработицы», т.е. такой уровень безработицы, при котором $\delta\omega_t = 0$;

c) когда изменения в уровне безработицы оказывали наибольшее (наименьшее) влияние на темп изменения заработной платы;

d) найдите 95% – доверительные интервалы для β_1 и β_2.

Год	ω_t	u_t	Год	ω_t	u_t
1	1,62	1,0	10	2,66	1,8
2	1,65	1,4	11	1,73	1,9
3	1,79	1,1	12	2,80	1,5
4	1,94	1,5	13	2,92	1,4
5	2,03	1,5	14	3,02	1,8
6	2,12	1,2	15	3,13	1,1
7	2,26	1,0	16	3,28	1,5
8	2,44	1,1	17	3,43	1,3
9	2,57	1,3	18	3,58	1,4

Задача 1.3.2

Для 14 однотипных предприятий (*i* – номер предприятия) имеются данные за год (см. табл.)

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
y_i	20	24	28	30	31	33	34	37	28	40	41	43	45	48
x_i	32	30	36	40	41	47	56	54	60	55	61	67	69	76

y_i – производительность труда, т/ч;

x_i – уровень механизации работ, %.

Требуется:

1. Построить выборочное уравнение линейной парной регрессии (найти значения b_1 и b_0);

2. Рассчитать значение выборочного коэффициента корреляции rxy., среднюю ошибку аппроксимации, выборочный коэффициент детерминации R2 и стандартные отклонения коэффициентов регрессии;

3. На уровне значимости α=0,05 оценить значимость коэффициентов и уравнения регрессии, проверить значимость линейной функции регрессии Найти доверительные интервалы для значимых коэффициентов регрессии и значений yi;

4. Оформить выводы в виде аналитической записки.

Задача 1.3.3

Имеются следующие статистические данные по Республике Татарстан:

Год	2001	2002	2003	2004	2005	2006	2007
Зарегистрировано преступлений, y_i	71266	57632	58866	63529	922324	105105	81251
Общая численность безработных, x_i	116144	99669	125714	137444	126825	107042	108327

Требуется:

5. Определить по МНК оценки коэффициентов уравнения регрессии.

6. Проверить статистическую значимость коэффициентов, входящих в уравнение регрессии.

7. Найти доверительные интервалы для коэффициентов регрессии при

уровне значимости α= 0,05.

8. Рассчитать коэффициент детерминации и на уровне значимости 0,05 проверить значимость линейной функции регрессии с помощью F-критерия Фишера.

9. Найти точечное (с надёжностью 0,95) предсказание зависимой переменной при значении объясняющей переменной, равном максимальному наблюдённому её значению, увеличенному на 10%.

2. Множественная регрессия

Множественная регрессия представляет собой модель вида

$$\hat{y}_i = M(y / x_1, x_2, ..., x_m) + \varepsilon_i ,$$

где y — результативный признак, а x_1, x_2, x,…,x_m — независимые или объясняющие переменные (признаки-факторы), ε_i – случайная ошибка отклонения.

Цель множественной регрессии — определить степень влияния каждого из факторов в отдельности и их совместное воздействие на результативный признак.

Включаемые в модель множественной регрессии факторы должны объяснять вариацию независимой переменной. Как и в случае парной регрессии, для модели множественной регрессии с некоторым набором факторов рассчитывается множественный коэффициент детерминации, определяющий долю объясненной вариации результативного признака за счет факторов, входящих в модель.

Остановимся на теоретической линейной модели множественной регрессии:

$$\hat{y}_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_{im} + \varepsilon_i$$

где b_i — коэффициенты регрессии, каждый из которых показывает, насколько единиц изменится ***y*** с изменением соответствующего признака ***x*** на единицу при условии, что остальные признаки не изменятся;

$\hat{y}_i$ — теоретическое значение, представляющее собой оценку ожидаемого значения ***y*** при фиксированных значениях переменных x_m

Как и в случае парной регрессии по любой конечной выборке нельзя точно получить вектор коэффициентов уравнения $\beta = \beta_0, \beta_1, \beta_2, ..., \beta_m$. Мы можем только рассчитать эмпирическое уравнение регрессии в форме:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_{im} + e_i .$$

В этом случае вектор $b = (b_0, b_1, b_2, ..., b_m)$ является вектором оценки теоретического вектора β, e_i – оценка теоретического отклонения ε_i.

2.1. Спецификация модели. Отбор факторов при построении уравнения множественной регрессии

Построение уравнения множественной регрессии начинается с решения вопроса о спецификации модели. Он включает в себя два круга вопросов: отбор факторов и выбор вида уравнения регрессии.

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

1. Они должны быть количественно измеримы. Если необходимо включить в модель качественный фактор, не имеющий количественного измерения, то ему нужно придать количественную определенность.

2. Факторы не должны быть мультиколлинеарны и тем более находиться в точной функциональной связи.

Включение в модель факторов с высокой мультиколлинеарностью, может привести к нежелательным последствиям – система нормальных уравнений может оказаться плохо обусловленной и повлечь за собой неустойчивость и ненадежность оценок коэффициентов регрессии.

Если между факторами существует высокая корреляция, то нельзя определить их изолированное влияние на результативный показатель и параметры уравнения регрессии оказываются неинтерпретируемыми.

Включаемые во множественную регрессию факторы должны объяснить вариацию независимой переменной. Если строится модель с набором m факторов, то для нее рассчитывается показатель детерминации R^2, который фиксирует долю объясненной вариации результативного признака за счет рассматриваемых в регрессии m факторов. Влияние других, не учтенных в модели факторов, оценивается как $1-R^2$ с соответствующей остаточной дисперсией S^2.

При дополнительном включении в регрессию $m+1$ фактора коэффициент детерминации должен не убывать, а остаточная дисперсия не возрастать:

$$R^2_{m+1} \geq R^2_m \quad \text{и} \qquad S^2_{m+1} \leq S^2_m.$$

Если же этого не происходит и данные показатели практически не отличаются друг от друга, то включаемый в анализ фактор x_{m+1} не улучшает модель и практически является лишним фактором.

Насыщение модели лишними факторами не только не снижает величину остаточной дисперсии и не увеличивает показатель детерминации, но и приводит к статистической незначимости параметров регрессии по критерию Стьюдента.

Таким образом, хотя теоретически регрессионная модель позволяет учесть любое число факторов, практически в этом нет необходимости. Отбор факторов производится на основе качественного теоретико-экономического анализа. Однако теоретический анализ часто не позволяет однозначно

ответить на вопрос о количественной взаимосвязи рассматриваемых признаков и целесообразности включения фактора в модель. Поэтому отбор факторов обычно осуществляется в две стадии: на первой подбираются факторы исходя из сущности проблемы; на второй – на основе матрицы показателей корреляции определяют статистики для параметров регрессии.

Параметры уравнения множественной регрессии находят методом наименьших квадратов.

2.2. Метод наименьших квадратов (МНК). Свойства оценок на основе МНК

Возможны разные виды уравнений множественной регрессии: линейные и нелинейные.

Ввиду четкой интерпретации параметров наиболее широко используется линейная функция. В линейной множественной регрессии $\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_{im}$ 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.

Рассмотрим линейную модель множественной регрессии

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_{im} + e_i . \qquad (2.1)$$

Классический подход к оцениванию параметров линейной модели множественной регрессии основан на методе наименьших квадратов (МНК). Чтобы получить по методу МНК наилучшие оценки должны выполнятся ряд предпосылок относительно случайного отклонения e_i. Эти предпосылки называются предпосылками Гаусса-Маркова или условиями Гаусса-Маркова:

1. Математическое ожидание случайного отклонения в теоретическом уравнении регрессии равно 0 для любых наблюдений, т.е.

$$M(\varepsilon_i) = 0, \ \forall_i$$

Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную.

2. Дисперсия случайного отклонения постоянна, т.е.

$$D(\varepsilon_i) = D(\varepsilon_j) = \sigma^2 , \forall_{i \neq j}$$

Из данного условия следует, что несмотря на то, что при каждом конкретном наблюдении случайное отклонение e_i может быть различным, но не должно быть причин, вызывающих большую ошибку. Выполнимость данной предпосылки называют *гомоскедастичностью*. Если предпосылка не выполняется, то говорят, что в модели присутствует эффект *гетероскедастичности* – изменяющихся отклонений.

3. Наблюдаемые значения случайных отклонений ε_i и ε_j независимы друг от друга. Если данное условие выполняется, то говорят об отсутствии автокорреляции.

4. Случайное отклонение должно быть независимо от объясняющей переменной. Обычно это условие выполняется автоматически, т.к. в эконометрических моделях объясняющие переменные не являются случайными величинами.

5. Регрессионная модель является линейной относительно параметров. Из этого условия следует, что математическое ожидание коэффициентов уравнения регрессии дает несмещенные оценки для коэффициентов. Дисперсии этих оценок уменьшаются при увеличении объема используемой выборки.

6. Наряду с выполнимостью указанных предпосылок при построении линейных регрессионных моделей обычно делаются еще некоторые предположения, а именно:

- случайное отклонение имеет нормальный закон распределения;

- число наблюдений существенно больше числа объясняющих переменных;

- отсутствуют ошибки спецификации;

- отсутствует линейная взаимосвязь между двумя или несколькими объясняющими переменными.

МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от расчетных $\hat{y}$ минимальна:

$$\sum_{i=1}^{n}(y_i - \hat{y}_{x_i})^2 \to \min. \qquad (2.2)$$

Как известно из курса математического анализа, для того чтобы найти экстремум функции нескольких переменных, надо вычислить частные производные первого порядка по каждому из параметров и приравнять их к нулю.

Итак. Имеем функцию $m+1$ аргумента:

$$Q(b_0, b_1, b_2, ..., b_m) = \sum_{i=1}^{n}(y - b_0 - b_1 x_1 - b_2 x_2 - ... - b_m x_{im})^2 \qquad (2.3)$$

Чтобы найти минимум функции (2.3), надо вычислить производные по каждому из параметров и приравнять их к нулю, т.к. равенство нулю производной – необходимое условие экстремума.

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum(y - b_0 - b_1 x_1 - b_2 x_2 - ... - b_m x_{im}) = 0; \\ \dfrac{\partial Q}{\partial b_1} = -2\sum x_1(y - b_0 - b_1 x_1 - b_2 x_2 - ... - b_m x_{im}) = 0; \\ \qquad ..... \\ \dfrac{\partial Q}{\partial b_m} = -2\sum x_{im}(y - b_0 - b_1 x_1 - b_2 x_2 - ... - b_m x_{im}) = 0. \end{cases} \qquad (2.4)$$

После элементарных преобразований приходим к системе линейных нормальных уравнений для нахождения параметров линейного уравнения множественной регрессии (2.1):

Так, для уравнения (2.1.) система нормальных уравнений имеет вид:

$$\begin{cases} \sum y = b_0 \cdot n + b_1 \cdot \sum x_1 + b_2 \sum x_2 + \text{K}\ b_m \cdot \sum x_{im} \\ \sum y \cdot x_1 = b_0 \cdot \sum x_1 + b_1 \cdot \sum x_1^2 + b_2 \sum x_1 \cdot x_2 + \text{K}\ b_m \cdot \sum x_1 \cdot x_{im} \\ \text{K K K K K K K K K K K K K K K K K K K K K K K} \\ \sum y \cdot x_{im} = b_0 \cdot \sum x_{im} + b_1 \cdot \sum x_{im} \cdot x_1 + b_2 \sum x_{im} \cdot x_2 + \text{K}\ b_m \cdot \sum x_{im}^2 \end{cases} \quad (2.5)$$

Решение системы (2.5) может быть осуществлено по одному из известных способов: Метод Гаусса, метод Крамера и т.д.

Пример. По 10-ти предприятиям региона (см. табл.) изучается зависимость выработки продукции на одного работника у (тыс. руб.) от ввода в действие новых основных фондов x_2 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_1 (%). Требуется составить уравнение множественной регрессии.

№ предприятия	1	2	3	4	5	6	7	8	9	10
x_1, (%)	1	2	3	5	7	10	14	16	19	22
x_2, (%)	0	1	3	4	6	8	11	14	16	19
y, (тыс. руб.)	6	11	19	28	31	35	39	42	46	49

Решение:

Предположим, что зависимость выработки продукции на одного работника характеризуется следующим уравнением:

$$\hat{y}_x = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 .$$

На основании исходных данных составляем систему уравнений для определения коэффициентов b_0, b_1 и b_2.

$$\sum y = 306; \sum x_1 = 99; \sum x_2 = 82; \sum y \cdot x_1 = 3962; \sum y \cdot x_2 = 3330;$$

$$\sum x_1^2 = 1485; \ \sum x_2^2 = 1060; \sum y^2 = 11290 ; \sum x_1 \cdot x_2 = 1253.$$

$$\begin{cases} 306 = 10 \cdot b_0 + 99 \cdot b_1 + 82 \cdot b_2 \\ 3962 = 99 \cdot b_0 + 1485 \cdot b_1 + 1253 \cdot b_2 \\ 3330 = 82 \cdot b_0 + 1253 \cdot b_1 + 1060 \cdot b_2 \end{cases}$$

Решим эту систему по методу Крамера. Вычисляем определитель системы:

$$\Delta = \begin{vmatrix} 10 & 99 & 82 \\ 99 & 1485 & 1253 \\ 82 & 1253 & 1060 \end{vmatrix} = 10 \cdot 1485 \cdot 1060 + 82 \cdot 99 \cdot 1253 + 82 \cdot 99 \cdot 1253 - 82 \cdot 1485 \cdot 82 - 1253 \cdot 1253 \cdot 10 - 1060 \cdot 99 \cdot 99 = 10418.$$

Аналогично вычисляем частные определители, заменяя соответствующий столбец столбцом свободных членов:

$$\Delta_1 = \begin{vmatrix} 306 & 99 & 82 \\ 3962 & 1485 & 1253 \\ 3330 & 1253 & 1060 \end{vmatrix} = 141628;$$

$$\Delta_2 = \begin{vmatrix} 10 & 306 & 82 \\ 99 & 3962 & 1253 \\ 82 & 3330 & 1060 \end{vmatrix} = -6612;$$

$$\Delta_3 = \begin{vmatrix} 10 & 99 & 306 \\ 99 & 1485 & 3962 \\ 82 & 1253 & 3330 \end{vmatrix} = 29588.$$

Коэффициенты уравнения определяются по формулам:

$$b_0 = \frac{\Delta_1}{\Delta} = \frac{141628}{10418} \approx 13{,}5946; \quad b_1 = \frac{\Delta_2}{\Delta} = \frac{-6612}{10418} \approx -0{,}6347; \quad b_2 = \frac{\Delta_3}{\Delta} = \frac{29588}{10418} \approx 2{,}84.$$

Таким образом, уравнение имеет вид:

$$\hat{y}_x = 13{,}5946 - 0{,}6347 \cdot x_1 + 2{,}84 \cdot x_2.$$

Возможен и иной подход к определению параметров множественной регрессии, когда уравнение регрессии строится, используя МНК. Для начала необходимо определить значения следующих параметров:

$$\bar{x}_i = \frac{\sum_{k=1}^{n} x_{ik}}{n}, \bar{y} = \frac{\sum_{k=1}^{n} y_k}{n}, \bar{x}_i^{\,2} = \frac{\sum_{k=1}^{n} x_{ik}^2}{n}, \bar{y}_i^{\,2} = \frac{\sum_{k=1}^{n} y_{ik}^2}{n}, \overline{x_i x_j} = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{n}, \quad \overline{yx_i} = \frac{\sum_{k=1}^{n} y_k x_{ik}}{n} \quad (2.6)$$

$$\bar{x}_1 = 9{,}9; \quad \bar{x}_2 = 8{,}2; \quad \bar{y} = 30{,}6; \; \bar{x}_1^{\,2} = 148{,}5; \; \bar{x}_2^{\,2} = 106; \; \bar{y}_2^{\,2} = 1129$$

$$\overline{yx_1} = 396{,}2; \; \overline{yx_2} = 333; \; \overline{x_1 x_2} = 125{,}3.$$

Для вычисления коэффициентов уравнения регрессии необходимо определить значения 6-ти сумм:

1. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2=\left[\bar{x}_1^2-(\bar{x}_1)^2\right]*n=\left[148{,}5-(9{,}9)^2\right]*n=\left[50{,}49\right]*n$

2. $\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2=\left[\bar{x}_2^2-(\bar{x}_2)^2\right]*n=\left[106-(8{,}2)^2\right]*n=\left[38{,}76\right]*n$

3. $\sum_{i=1}^{n}(y_i-\bar{y})^2=\left[\bar{y}^2-(\bar{y})^2\right]*n=\left[1129-(30{,}6)^2\right]*n=\left[192{,}64\right]*n$

4. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_2^{\,2})(y_i-\bar{y})=\left[\bar{x}_1\bar{y}-\bar{x}_1*\bar{y}\right]*n=\left[396{,}2-9{,}9*8{,}2\right]*n=\left[93{,}26\right]*n$

5. $\sum_{i=1}^{n}(x_{i2}-\bar{x}_2^{\,2})(y_i-\bar{y})=\left[\bar{x}_2\bar{y}-\bar{x}_2*\bar{y}\right]*n=\left[333-8{,}2*30{,}6\right]*n=\left[82{,}08\right]*n$

6. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1^{\,2})(x_{i2}-\bar{x}_2^{\,2})=\left[\bar{x}_1\bar{x}_2-\bar{x}_1*\bar{x}_2\right]*n=\left[125{,}5-9{,}9*8{,}2\right]*n=\left[44{,}12\right]*n$

Подставим полученные значения 6-ти сумм в формулы для расчета коэффициентов уравнения регрессии (m=2):

$$b_0=\bar{y}-b_1*\bar{x}_1-b_2*\bar{x}_2$$

$$b_1=\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i2}-x_2)^2-\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i-1}^{n}(x_{i1}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2-(\sum_{i=1}^{n}(x_{i1}-x_1)(x_{i2}-x_2))^2}$$

$$b_2=\frac{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i1}-x_1)^2-\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i-1}^{n}(x_{i1}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2-(\sum_{i=1}^{n}(x_{i1}-x_1)(x_{i2}-x_2))^2}$$

или

$$b_1=\frac{[2]*[4]-[5]*[6]}{[1]*[2]-[6]^2}=\frac{38{,}76*93{,}26-82{,}08*44{,}12}{50{,}49*38{,}76-44{,}12^2}=-0{,}6346;$$

$$b_2=\frac{[1]*[5]-[4]*[6]}{[1]*[2]-[6]^2}=\frac{50{,}49*82{,}08-93{,}26*44{,}12}{50{,}49*38{,}76-44{,}12^2}=2{,}84;$$

$$b_0=\bar{y}-b_1*\bar{x}_1-b_2*\bar{x}_2=30{,}6+0{,}6346*9{,}9-2{,}84*8{,}2=13{,}5946.$$

Таким образом, мы получили эмпирические значения параметров множественной линейной регрессии, которая имеет следующий вид:

$$y = b_0 + b_1 x_1 + b_2 x_2 = 13{,}5946 - 0{,}6347 x_1 + 2{,}84 x_2$$

Сравнивая полученное уравнение с полученным ранее, мы видим хорошее соответствие полученных разными способами результатов.

2.3. Стандартные ошибки коэффициентов уравнений множественной линейной регрессии

Значения стандартных ошибок позволяет оценивать точность эмпирических коэффициентов уравнений регрессии и проверять выдвигаемые относительно них гипотезы.

Выборочные дисперсии эмпирических коэффициентов множественной регрессии можно определить следующим образом:

$$S_{b_j}^2 = S^2 z'_{jj} = \frac{\sum_{i=1}^{n} e_i^2}{n-m-1} z'_{jj},\ j = 1,2,\ldots,m \qquad (2.9)$$

Здесь z'jj - j-тый диагональный элемент матрицы $Z^{-1} = (X^T X)^{-1}$.

Рассмотрим пример с m=1, где m - количество объясняющих переменных.

$$X^T X = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ x_{11} & x_{21} & \ldots & x_{n1} \end{pmatrix} * \begin{pmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \ldots & \ldots \\ 1 & x_{n1} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 \end{pmatrix}$$

Найдем величину обратной матрицы Z^{-1}. Она будет иметь следующий вид:

$$Z^{-1} = (X^T X)^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \cdot \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix} =$$

$$= \begin{pmatrix} \dfrac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} & -\dfrac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ -\dfrac{\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} & \dfrac{n}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{pmatrix}$$

При этом:

$$S_0^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y})^2}{n-m-1} = \frac{\sum_{i=1}^{n} e_i^2}{n-m-1} \qquad (2.10)$$

где m - количество объясняющих переменных модели.

В частности, для уравнения множественной регрессии:

$Y = b_0 + b_1X_1 + b_2X_2$

с двумя объясняющими переменными m=2 используются следующие формулы:

$$S_{b_0}^2 = \frac{1}{n} + \frac{\bar{x}_1^2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 + \bar{x}_2^2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 - 2\bar{x}_1 \bar{x}_2 \sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 - (\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2))^2} \cdot S_0^2$$

$$S_{b_1}^2 = \frac{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 - (\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2))^2} \cdot S_0^2$$

$$S_{b_2}^2 = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2 - (\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2))^2} \cdot S_0^2$$

$$S_{b_0} = \sqrt{S_{b_0}^2},\ S_{b_1} = \sqrt{S_{b_1}^2},\ S_{b_2} = \sqrt{S_{b_2}^2}$$

Здесь S_{bj} - стандартная ошибка коэффициента регрессии; S_0 - стандартная ошибка множественной регрессии (несмещенная оценка).

По аналогии с парной регрессией после определения точечных оценок *bj* коэффициентов *βj (j=1,2,…,m)* теоретического уравнения множественной регрессии могут быть рассчитаны интервальные оценки указанных коэффициентов.

Доверительный интервал, покрывающий с надежностью *(1-α)* неизвестное значение параметра *βj,* определяется как:

$$\left(b_j - t_{\alpha/2,n-m-1} \cdot S_{b_j}; b_j + t_{\alpha/2,n-m-1} \cdot S_{b_j} \right) \qquad (2.11)$$

Далее, как и в случае парной регрессии, статистическая значимость коэффициентов множественной регрессии с *m* объясняющими переменными проверяется на основе t-статистики:

$$t_{b_j} = \frac{b_j}{S_{b_j}} \qquad (2.12)$$

имеющей в данном случае распределение Стьюдента с числом степеней свободы $\nu = n - m - 1$. При требуемом уровне значимости наблюдаемое значение t-статистики сравнивается с критической точной $t_{\alpha/2,n-m-1}$ распределения Стьюдента.

В случае, если $\left|t_{b_j}\right| > t_{\alpha/2,n-m-1}$, то статистическая значимость соответствующего коэффициента множественной регрессии подтверждается. Это означает, что фактор *Xj* линейно связан с зависимой переменной *Y*. Если

же установлен факт незначимости коэффициента *bj*, то рекомендуется исключить из уравнения переменную *Xj*. Это не приведет к существенной потере качества модели, но сделает ее более конкретной.

2.4. Проверка общего качества уравнения регрессии

После проверки значимости каждого коэффициента регрессии обычно проверяется общее качество уравнения регрессии. Для этой цели, как и в случае парной регрессии, используется коэффициент детерминации R^2, который в общем случае рассчитывается по формуле:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \qquad (2.13)$$

Коэффициент детерминации характеризует тесноту связи рассматриваемого набора факторов с исследуемым признаком или, иначе, оценивает тесноту совместного влияния факторов на результат.

Для множественной регрессии коэффициент детерминации является неубывающей функцией от числа объясняющих переменных. Добавление новой объясняющей переменной никогда не уменьшает значение R^2.

Иногда при расчете коэффициента детерминации для получения несмещенных оценок в числителе и знаменателе вычитаемой из единицы дроби делается поправка на число степеней свободы. Вводится так называемый скорректированный (исправленный) коэффициент детерминации:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2 /(n-m-1)}{\sum_{i=1}^{n} (y_i - \hat{y})^2 /(n-1)} \qquad (2.14)$$

Соотношение может быть представлено в следующем виде:

$$\bar{R}^2 = 1-(1-R^2)\frac{n-1}{n-m-m1} \qquad (2.15)$$

Из чего следует, что $\bar{R}^2 < R^2$ для m > 1. С ростом значения m скорректированный коэффициент детерминации $\bar{R}^2$ растет медленнее, чем обычный коэффициент детерминации R^2 . Другими словами, он корректируется в сторону уменьшения с ростом числа объясняющих переменных. Нетрудно заметить, что $\bar{R}^2 = R^2$ только при $R^2 = 1$. $\bar{R}^2$ может принимать и отрицательные значения (например, при $R^2 = 0$).

Доказано, что $\bar{R}^2$ увеличивается при добавлении новой объясняющей переменной тогда и только тогда, когда *t* – статистика для этой переменной по модулю больше единицы. Поэтому добавление в модель новых объясняющих переменных осуществляется до тех пор, пока растет скорректированный коэффициент детерминации.

2.5. Оценка общего качества уравнения множественной регрессии

Значимость уравнения множественной регрессии в целом, так же как и в парной регрессии, оценивается с помощью F -критерия Фишера:

$$F_{расч.} = \frac{R^2}{1-R^2}\cdot\frac{n-m-1}{m}, \qquad (2.16)$$

где R^2 – коэффициент детерминации; m – количество объясняющих переменных *X* (в линейной регрессии совпадает с числом включенных в модель факторов); *n* – число наблюдений.

Частные F -критерии F_{xi}, к примеру в случае m=2, оценивают статистическую значимость присутствия факторов x_1 и x_2 в уравнении множественной регрессии, целесообразность включения в уравнение одного фактора после другого, т.е. F_{x1} оценивает целесообразность включения в уравнение фактора x_1 после того, как в него был включен фактор x_2. Соответственно F_{x2} указывает на целесообразность включения в модель

фактора x_2 после фактора x_1. Необходимость такой оценки связана с тем, что не каждый фактор, вошедший в модель, может существенно увеличивать долю объясненной вариации результативного признака. Кроме того, при наличии в модели нескольких факторов они могут вводиться в модель в разной последовательности. Ввиду корреляции между факторами значимость одного и того же фактора может быть разной в зависимости от последовательности его введения в модель.

Частный F -критерий построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительно включенного фактора, с остаточной дисперсией на одну степень свободы по регрессионной модели в целом. В общем виде для фактора x_i частный F -критерий определится как

$$F_{x_i} = \frac{R^2_{yx_1 \ldots x_i \ldots x_m} - R^2_{yx_1 \ldots x_{i-1} x_{i+1} \ldots x_m}}{1 - R^2_{yx_1 \ldots x_i \ldots x_m}} \cdot \frac{n-m-1}{1}, \qquad (2.17)$$

где $R^2_{yx_1 \ldots x_i \ldots x_m}$ – коэффициент множественной детерминации для модели с полным набором факторов, $R^2_{yx_1 \ldots x_{i-1} x_{i+1} \ldots x_m}$ – тот же показатель, но без включения в модель фактора x_i , n – число наблюдений, m – число параметров в модели (без свободного члена).

Фактическое значение частного F-критерия сравнивается с табличным при уровне значимости α и числе степеней свободы: 1 и $n-m-1$. Если фактическое значение F_{x_i} превышает $F_{\text{табл}}(\alpha, k_1, k_2)$, то дополнительное включение фактора x_i в модель статистически оправданно и коэффициент чистой регрессии b_i при факторе x_i статистически значим. Если же фактическое значение F_{x_i} меньше табличного, то дополнительное включение в модель фактора x_i не увеличивает существенно долю объясненной вариации признака y , следовательно, нецелесообразно его включение в

модель; коэффициент регрессии при данном факторе в этом случае статистически незначим.

Для двухфакторного уравнения частные F-критерии имеют вид:

$$F_{x_1} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot (n-3), \quad F_{x_2} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot (n-3). \quad (2.17a)$$

С помощью частного F-критерия можно проверить значимость всех коэффициентов регрессии в предположении, что каждый соответствующий фактор x_i вводился в уравнение множественной регрессии последним.

2.6. Решение типовых задач

Задача 2.6.1.

Рассмотрим в качестве примера множественной регрессии двухфакторную линейную модель. Исходные данные представлены в таблице 2.6.1.

Таблица 2.6.1

№ группы	Расходы на питание ***(y)***	Душевой доход ***(x_1)***	Размер семьи ***(x_2)***
1	431	626	1,5
2	614	1575	2,1
3	790	2235	2,4
4	898	2657	2,7
5	1111	3699	3,2
6	1303	4794	3,4
7	1486	5924	3,6
8	1643	7279	3,7
9	1912	9348	4
10	2409	18805	3,7

Необходимо:

1. по МНК определить параметры множественной линейной регрессии $\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2$;

2. оценить статистическую значимость найденных эмпирических коэффициентов регрессии b_1, b_2;

3. сравнить влияние факторов на результат при помощи средних коэффициентов эластичности;

4. построить 95-% доверительные интервалы для найденных коэффициентов;

5. вычислить коэффициент детерминации R^2 и оценить его статистическую значимость при $\alpha = 0{,}05$;

6. Проверить качество построенного уравнения регрессии с помощью F-статистики Фишера.

7. Оценить целесообразность включения в уравнение одного фактора после другого с помощью частных F-статистик Фишера.

Решение:

Определим по МНК коэффициенты уравнения регрессии. Для этого нам необходимо рассчитать следующую таблицу:

№ группы	y	x_1	x_2	x_1y	x_2y	x_1x_2	x_1^2	x_2^2	y^2	ỹ	e	e^2	$(e_i-e_{i-1})^2$
1	431	626	1,5	269806	646,5	939,0	391876	2,25	185761	369,4001	61,600	3794,545	
2	614	1575	2,1	967050	1289,4	3307,5	2480625	4,41	376996	643,7322	-29,732	884,006	8341,557
3	790	2235	2,4	1765650	1896,0	5364,0	4995225	5,76	624100	794,2598	-4,260	18,146	648,8435
4	898	2657	2,7	2385986	2424,6	7173,9	7059649	7,29	806404	927,6443	-29,644	878,786	644,3727
5	1111	3699	3,2	4109589	3555,2	11836,8	13682601	10,24	1234321	1174,346	-63,346	4012,706	1135,797
6	1303	4794	3,4	6246582	4430,2	16299,6	22982436	11,56	1697809	1321,877	-18,877	356,348	1977,469
7	1486	5924	3,6	8803064	5349,6	21326,4	35093776	12,96	2208196	1471,929	14,071	197,979	1085,551
8	1643	7279	3,7	11959397	6079,1	26932,3	52983841	13,69	2699449	1603,859	39,141	1532,003	628,5196
9	1912	9348	4	17873376	7648,0	37392,0	87385104	16,00	3655744	1855,877	56,123	3149,813	288,4017
10	2409	18805	3,7	45301245	8913,3	69578,5	353628025	13,69	5803281	2434,075	-25,075	628,749	6593,127
Сумма	**12597**	**56942**	**30,3**	**99681745**	**42232**	**200150**	**580683158**	**97,85**	**19292061**	**12597**	**0,000**	**15453,08**	**21343,64**
Ср.знач.	**1259,7**	**5694,2**	**3,03**	**9968174,5**	**4223,2**	**20015**	**58068316**	**9,785**	**1929206,1**	**1259,7**	**0,000**		

Для вычисления коэффициентов уравнения регрессии необходимо определить значения 6-ти сумм:

1. $\sum_{i=1}^{n}(x_{i1}-\bar{x_1})^2 = \left[\bar{x_1^2}-(\bar{x_1})^2\right]*n = \left[58068316-(5694{,}2)^2\right]*n = \left[2564440{,}16\right]*n$

2. $\sum_{i=1}^{n}(x_{i2}-\bar{x_2})^2 = \left[\bar{x_2^2}-(\bar{x_2})^2\right]*n = \left[9{,}79-(3{,}03)^2\right]*n = \left[0{,}6041\right]*n$

3. $\sum_{i=1}^{n}(y_i-\bar{y})^2=\left[\bar{y^2}-(\bar{y})^2\right]*n=\left[1929206-(1260)^2\right]*n=[34236{,}201]*n$

4. $\sum_{i=1}^{n}(x_{i1}-\bar{x_2}^2)(y_i-\bar{y})=\left[\bar{x_1}\bar{y}-\bar{x_1}*\bar{y}\right]*n = [99681745-5694{,}2*1260]*n = [279519076]*n$

5. $\sum_{i=1}^{n}(x_{i2}-\bar{x_2}^2)(y_i-\bar{y})=\left[\bar{x_2}\bar{y}-\bar{x_2}*\bar{y}\right]*n = [422319-3{,}03*1260]*n = [406{,}299]*n$

6. $\sum_{i=1}^{n}(x_{i1}-\bar{x_1}^2)(x_{i2}-\bar{x_2}^2)=\left[\bar{x_1}\bar{x_2}-\bar{x_1}*\bar{x_2}\right]*n = [20015-5694{,}2*3{,}03]*n = [2761{,}574]*n$

Подставим полученные значения 6-ти сумм в формулы для расчета коэффициентов уравнения регрессии (m=2):

$$b_0=\bar{y}-b_1*\bar{x}_1-b_2*\bar{x}_2$$

$$b_1=\frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i2}-x_2)^2-\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i-1}^{n}(x_{i1}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2-(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}$$

$$b_2=\frac{\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i1}-x_1)^2-\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(y_i-\bar{y})*\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum_{i-1}^{n}(x_{i1}-\bar{x}_1)^2\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2-(\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}$$

или

$$b_1=\frac{[2]*[4]-[5]*[6]}{[1]*[2]-[6]^2}=\frac{0{,}6041*279519076-406{,}299*2761{,}574}{2564440216*0{,}6041-2761{,}574^2}=0{,}072;$$

$$b_2=\frac{[1]*[5]-[4]*[6]}{[1]*[2]-[6]^2}=\frac{2564440216*406{,}299-279519076*2761{,}574}{2564440216*0{,}6041-2761{,}574^2}=343{,}293;$$

$$b_0=\bar{y}-b_1*\bar{x}_1-b_2*\bar{x}_2=1260-0{,}072*5694{,}2-343{,}293*3{,}03=-190{,}63.$$

Таким образом, мы получили эмпирические значения параметров множественной линейной регрессии, которая имеет следующий вид:

$$y=b_0+b_1x_1+b_2x_2=-190{,}63+0{,}072x_1+343{,}293x_2$$

Рассмотрим матричный вид определения вектора оценок коэффициентов регрессии

а. Определим вектор оценок коэффициентов регрессии. Согласно методу наименьших квадратов, вектор получается из выражения:

$B = (X^T X)^{-1} X^T Y$

Матрица X

1	626	1,5
1	1575	2,1
1	2235	2,4
1	2657	2,7
1	3699	3,2
1	4794	3,4
1	5924	3,6
1	7279	3,7
1	9348	4
1	18805	3,7

Матрица Y

431
614
790
898
1111
1303
1486
1643
1912
2409

Матрица X^T

1	1	1	1	1	1	1	1	1	1
626	1575	2235	2657	3699	4794	5924	7279	9348	18805
1,5	2,1	2,4	2,7	3,2	3,4	3,6	3,7	4	3,7

б. Умножаем матрицы, $(X^T X)$

$$X^T X = \begin{vmatrix} 10 & 56942 & 30{,}3 \\ 56942 & 580683158 & 200150 \\ 30{,}3 & 200150 & 97{,}85 \end{vmatrix}$$

в. Умножаем матрицы, $(X^T Y)$

$$X^T Y = \begin{vmatrix} 12597 \\ 99681745 \\ 42231{,}9 \end{vmatrix}$$

г. Находим определитель $\det (X^T X)^T = 7865492387$

д. Находим обратную матрицу $(X^T X)^{-1}$

$$(X^T X)^{-1} = \begin{pmatrix} 2{,}13080424 & 0{,}0006265 & -0{,}78796826 \\ 0{,}00006265 & 0{,}00000001 & -0{,}00003511 \\ -0{,}78796826 & -0{,}00003511 & 0{,}32693683 \end{pmatrix}$$

е. Вектор оценок коэффициентов регрессии равен:

$B = (X^T X)^{-1} X^T Y$

$$y(x) = \begin{pmatrix} 2{,}13080424 & 0{,}0006265 & -0{,}78796826 \\ 0{,}00006265 & 0{,}00000001 & -0{,}00003511 \\ -0{,}78796826 & -0{,}00003511 & 0{,}32693683 \end{pmatrix} * \begin{vmatrix} 12597 \\ 99681745 \\ 42231{,}9 \end{vmatrix} = \begin{vmatrix} -190{,}6301 \\ 0{,}0720 \\ 343{,}2930 \end{vmatrix}$$

Таким образом, мы получили уравнение регрессии: y = -190,6301 + $0{,}072x_1 + 343{,}293x_2$

2. оценим статистическую значимость найденных эмпирических коэффициентов регрессии b_1, b_2 с помощью *t*-статистики Стьюдента. Для этого сначала необходимо определить стандартные ошибки коэффициентов корреляции:

$$S_0^{\,2} = \frac{\sum_{i=1}^{n} e_i^2}{n-m-1} = \frac{1545308}{10-2-1} = 2207{,}582$$

$$S_{b_1}^2 = \frac{\sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 - (\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2))^2} \cdot S_0^2$$

$$S_{b_2}^2 = \frac{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2}{\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n} (x_{i2} - \bar{x}_2)^2 - (\sum_{i=1}^{n} (x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2))^2} \cdot S_0^2$$

или

$$S_{b1}^2 = \frac{S^2 * [2]}{[1] * [2] - [6]^2} = \frac{2207{,}582 * 6{,}041}{786549239} = 0{,}0000169$$

$$S_{b2}^2 = \frac{S^2 * [1]}{[1] * [2] - [6]^2} = \frac{2207{,}582 * 2564440216}{786549239} = 719{,}75304$$

Определим значения t-статистик для каждого из коэффициентов:

$$t_{b1} = \frac{b_1}{S_{b1}} = \frac{0{,}072}{0{,}0041} = 17{,}5$$

$$t_{b2} = \frac{b_2}{S_{b2}} = \frac{343{,}293}{26{,}828} = 12{,}7$$

Сравним полученные расчетные значения t-статистики Стьюдента с соответствующим критическим значением (см. таблица Распределение Стъюдента):

$$t_{критич.} = t_{\frac{\alpha}{2};\nu} = t_{\frac{\alpha}{2};n-m-1} = t_{0{,}025;7} = 2{,}365$$

Так как $t_{b1}, t_{b2} > t_{критич.}$, мы делаем вывод о том, что оба коэффициента сильно значимы для построенной модели.

3. Рассчитаем средние коэффициенты эластичности для коэффициентов , входящих в уравнение множественной регрессии по следующей формуле:

$$\bar{Э}_{yx_j} = b_j \frac{\bar{x}_j}{\bar{y}}$$

$$\bar{Э}_{yx_1} = b_1 \frac{\bar{x}_1}{\bar{y}} = 0{,}072 * \frac{5694{,}2}{1260} = 0{,}326$$

$$\bar{Э}_{yx_2} = b_2 \frac{\bar{x}_2}{\bar{y}} = 343{,}293 * \frac{3{,}03}{1260} = 0{,}826$$

Таким образом, в случае изменения фактора X_1 (доход семьи) на 1% зависимая переменная Y (расходы на питание) изменится на 0,326%, а если фактор X_2 (количество человек в семье) изменится на 1%, то значение параметра Y изменится на 0,826%. Следовательно, большей чувствительностью модель обладает по фактору количество человек в семье.

4. Построим 95-% доверительные интервалы для найденных коэффициентов:

$$b_j - t_{\frac{\alpha}{2};n-m-1} * S_{b_j} < \beta_j < b_j + t_{\frac{\alpha}{2};n-m-1} * S_{b_j}$$

$$b_1 - t_{0,025;7} * S_{b_1} < \beta_1 < b_1 + t_{0,025;7} * S_{b_1}$$

$$0{,}06 < \beta_1 < 0{,}082$$

$$b_2 - t_{0,025;7} * S_{b2} < \beta_2 < b_2 + t_{0,025;7} * S_{b_2}$$

$$279{,}845 < \beta_2 < 406{,}741$$

Таким образом, если по другим выборкам мы получим значение коэффициентов, принадлежащие этим интервалам, то мы можем утверждать, что уравнение регрессии покажет такое же поведение для Y как определенное по выборке.

5. Определим коэффициент детерминации R^2 и оценим его статистическую значимость при α = 0,05

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{1545{,}08}{34236{,}01 * 10} = 0{,}995$$

Следовательно, учтенные в модели факторы на 99,5% определяют поведение зависимой переменной Y.

6. Проверим качество построенного уравнения регрессии с помощью F-статистики Фишера:

$$F = \frac{R^2}{1 - R^2} * \frac{n - m - 1}{m} = \frac{0{,}995}{1 - 0{,}995} * 3{,}5 = \mathbf{696{,}5}$$

Сравним полученные расчетные значения F-статистики Фишера с соответствующим табличным значением (см. таблица Распределение Фишера):

$$F_{табл.} = F_{\alpha;\nu_1;\nu_2} = F_{0{,}05;m;n-m-1} = F_{0{,}05;2;7} = \mathbf{19{,}4}$$

Следовательно, так как $F_{расч} > F_{таблич}$, мы делаем вывод о том, что модель имеет хороший уровень качества.

6. Оценим целесообразность включения в уравнение одного фактора после другого с помощью частных F-статистик Фишера

$$F_{x_{1факт.}} = \frac{R^2_{yx_1x_2} - r^2_{yx_2}}{1 - R^2_{yx_1x_2}} * \frac{n-m-1}{1} = \frac{0{,}9954 - 0{,}8899}{1 - 0{,}9954} * \frac{10-2-1}{1} = 160{,}2173$$

$$r_{yx_2} = \frac{[5]}{\sqrt{[2]*[3]}} = \frac{406{,}299}{\sqrt{0{,}6041*34236{,}01}} = 0{,}8934$$

$$F_{x_{2факт.}} = \frac{R^2_{yx_1x_2} - r^2_{yx_1}}{1 - R^2_{yx_1x_2}} * \frac{n-m-1}{1} = \frac{0{,}9954 - 0{,}7982}{1 - 0{,}9954} * \frac{10-2-1}{1} = 300{,}0869$$

$$r_{yx_1} = \frac{[4]}{\sqrt{[1]*[3]}} = \frac{279519{,}76}{\sqrt{2564440{,}16*34236{,}01}} = 0{,}9433$$

В нашем случае $F_{расч\,x_1}$ = 160,2173; $F_{расч\,x_2}$ = 300,0869; $F_{таблич}$ = 23,68 (для числа степеней свободы 7 и 1 соответственно и уровня значимости 0,05)

Сравнивая значения $F_{расч\,x_2}$ и $F_{таблич}$ (300,0869 > 23,68), делаем вывод о том, что включение в модель фактора X_2 после фактора X_1 улучшает модель.

Сравнение $F_{расч\,x_1}$ и $F_{таблич}$ (160,2173 > 23,86) также показывает, что включение в модель дополнительного фактора X_1 после того, как фактор X_2 уже включен в уравнение улучшает модель, но не на столько как в первом случае. Поэтому приходим к выводу о целесообразности включения фактора X_2 после X_1 и что оба фактора одинаково значимы для построенной модели.

Рассмотрим решение задачи с помощью Excel

Excel позволяет при построении уравнения линейной регрессии большую часть работы сделать очень быстро. Важно понять, как интерпретировать полученные результаты. Воспользуемся надстройкой *Пакет анализа*.

Сервис — Анализ данных — Регрессия — ОК. Появляется диалоговое окно, которое нужно заполнить. В графе Входной интервал Y: указывается ссылка на ячейки, содержащие значения результативного признака y. В графе Входной интервал X: указывается ссылка на ячейки, содержащие значения факторов x_1,..., x_m ($m \leq 16$). Если первые из ячеек содержат пояснительный

текст, то рядом со словом *Метки* нужно поставить «галочку» *Уровень надежности* (доверительная вероятность) по умолчанию предполагается равным 95%. Если исследователя это значение не устраивает, то рядом со словами Уровень надежности нужно поставить «галочку» и указать требуемое значение. Поставив «галочку» рядом со словом *константа-ноль*, исследователь получит b_0 =0 по умолчанию. Если нужны значения остатков e_i и их график, то нужно поставить «галочки» рядом со словами *Остатки* и *График остатков*. ОК. Появляется итоговое окно.

Если число в графе Значимость *F* превышает 1 — *Уровень надежности*, то принимается гипотеза $R^2 = 0$. Иначе принимается гипотеза $R^2 \neq 0$.

P-значение — это значения уровней значимости, соответствующие вычисленным *t*-статистикам. *P*-значение = СТЬЮДРАСП (*t* -статистика; n-m-1) (статистическая функция мастера функций f_x). Если P-значение превышает 1 — *Уровень надежности*, то соответствующая переменная статистически незначима и ее можно исключить из модели.

Нижние 95% и Верхние 95% — это нижние и верхние границы 95-процентных доверительных интервалов для коэффициентов теоретического уравнения линейной регрессии. Если исследователь согласился с принятым по умолчанию значением доверительной вероятности 95%, то последние два столбца будут дублировать два предыдущих столбца. Если исследователь вводил свое значение доверительной вероятности *p*, то последние два столбца содержат значения соответственно нижней и верхней границы *p*-процентных доверительных интервалов.

	1	2
1	ВЫВОД ИТОГОВ	
2		
3	*Регрессионная статистика*	
4	Множественный R	0,997740614
5	R-квадрат	0,995486332
6	Нормированный R-квадрат	0,994196713
7	Стандартная ошибка	46,98492657
8	Наблюдения	10

10	Дисперсионный анализ								
11		*df*	*SS*	*MS*	*F*	*Значимость F*			
12	Регрессия	2	3408167,017	1704083,508	771,9226221	6,17808E-09			
13	Остаток	7	15453,08327	2207,583324					
14	Итого	9	3423620,1						
15									
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
17	Y-пересечение	-190,6301122	68,5851872	-2,77946478	0,02731929	-352,8082	-28,4520	-352,8082	-28,4520
18	Переменная X 1	0,072029818	0,004117656	17,49291973	0,00000049	0,0623	0,0818	0,0623	0,0818
19	Переменная X 2	343,2930441	26,82822156	12,79596724	0,00000413	279,8544	406,7317	279,8544	406,7317

23	ВЫВОД ОСТАТКА		
24			
25	*Наблюдение*	*Предсказанное Y*	*Остатки*
26	1	369,4001199	61,59988006
27	2	643,7322435	-29,73224351
28	3	794,2598365	-4,259836491
29	4	927,6443328	-29,64433284
30	5	1174,345925	-63,34592505
31	6	1321,877184	-18,87718436
32	7	1471,929487	14,07051271
33	8	1603,859195	39,14080519
34	9	1855,876801	56,12319894
35	10	2434,074875	-25,07487465

Таким образом, при определении зависимости расходов на питание (Y) от размера дохода (X_1) и от количества человек в семье (X_2) было обнаружено, что при неизменном значении параметра количество человек в семье и изменения на 1 у.е. размера дохода семьи, расходы на питание вырастут на 0,07 у.е. В то же время, если не измениться доход семьи, а количество человек станет на одного больше, то расходы в семье вырастут на 343, 29 у.е.

Поскольку $F_{расч} > F_{таблич}$ нулевая гипотеза отклоняется, то есть можно сделать вывод о статистической значимости уравнения в целом. Все выше обозначенные выводы позволяют сделать, что двухфакторная модель зависимости расходов на питание от душевого дохода и размера семьи является статистически значимой, надежной и может использоваться для прогнозов.

2.7. Упражнения и задачи

Задача 2.7.1

По 10-ти предприятиям изучается зависимость объема выпуска продукции от численности персонала и расхода материалов. Даны:

Коэффициент детерминации: ?

Множественный коэффициент корреляции: 0,80

Уравнение регрессии: $y = ? + 0{,}48x_1 + 70x_2 + 1{,}2x_3$

Стандартные ошибки параметров (S): 2 0,06 ? 0,24

t-критерий для параметров 1,5 ? 4 ?

Восстановить пропущенные характеристики оценки значимости уравнения.

Задача 2.7.2.

Проверить гипотезу H_0 о статистической незначимости уравнения регрессии, если индекс корреляции $r_{xy}=0{,}4$, число измерений 10, число параметров 3. (По критерию Фишера).

Задача 2.7.3.

По 20-ти предприятиям концерна изучается зависимость прибыли y от выработки продукции на одного работника x_1 и индекса цен x_2. Полученные данные:

$\bar{y} = 250,\ \bar{x}_1 = 47,\ \bar{x}_2 = 112,$

$\sigma = 38,\ \sigma_{x1} = 12,\ \sigma_{x2} = 21,$

$r_{yx1} = 0{,}68, r_{yx2} = 0{,}63, r_{x1x2} = 0{,}42$

Найдите уравнение множественной регрессии в стандартизированном натуральном масштабе.

Задача 2.7.4.

По 20 наблюдениям получены следующие результаты:

$$\sum x_{i1} = 4{,}88;\ \sum x_{i1}^2 = 2{,}518;\ \sum x_{i2} = 26{,}7;$$

$$\sum x_{i2}^2 = 75{,}15;\ \sum y_i = 44{,}7;\ \sum x_{i1}x_{i2} = 13{,}75;$$

$$\sum x_{i1}y_i = 22{,}1;\ \sum x_{i2}y_i = 125{,}75;\ \sum y_i^2 = 210{,}4;\ \sum e_i^2 = 0{,}015.$$

a) Оцените коэффициенты линейной регрессии $y = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \varepsilon$

b) Определите стандартные ошибки коэффициентов;

c) Вычислите R^2 и $\bar{R}^2$;

d) Оцените 95%-е доверительные интервалы для коэффициентов β_1 и β_2;

e) Оцените статистическую значимость коэффициентов регрессии и детерминации при уровне значимости α=0,05;

f) Сделайте выводы по модели.

Задача 2.7.5.

Для оценки коэффициентов уравнения регрессии $y = \beta_0 + \beta_1 X_1 + \beta_2 x_2 + \varepsilon$ вычисления проведены в матричной форме при n = 30.

$$X^T X = \begin{pmatrix} 30 & 646{,}3 & 421.41 \\ 646{,}3 & 16085{,}46 & 8990{,}313 \\ 421{,}41 & 8990{,}313 & 6616{,}081 \end{pmatrix};\ X^T Y = \begin{pmatrix} 606{,}73 \\ 14751{,}27 \\ 8326{,}798 \end{pmatrix}$$

$$\bar{Y} = 20{,}22;\ \bar{X}_1 = 21{,}54\ \bar{X}_2 = 14{,}05$$

a) Определите эмпирические коэффициенты регрессии;

b) Оцените их дисперсию и ковариацию cov(b_1,b_2)

c) С доверительной вероятностью γ = 0,95 оценить значимость коэффициентов регрессии и для значимых коэффициентов определить доверительные интервалы, оценить значимость уравнения регрессии.

Задача 2.7.6

Вычислить эластичность в общем виде и в точке x = 1 для функции:

$y = 3x^2 + 2/x$

где $x_1 = x^2$, $x_2 = 1/x$

Задача 2.7.7.

Предполагается, что объем предложения товара y линейно зависит от цены товара X_1 и зарплаты сотрудников X_2: $y=\beta_0+\beta_1X_1+\beta_2X_2$. Статистические данные собраны за 14 месяцев.

x_1	32	30	36	40	41	47	56	54	60	55	61	67	69	76
x_2	33	31	41	39	46	43	34	38	42	35	39	44	40	41
y	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Найти:

1. Оценить по МНК коэффициенты теоретического уравнения множественной регрессии $\hat{y}_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$;

2. Дайте сравнительную оценку силы связи факторов с результатом с помощью средних (общих) коэффициентов эластичности;

3. Оценить статистическую значимость параметров регрессионной модели с помощью t-критерия; нулевую гипотезу о значимости уравнения и показателей тесноты связи проверьте с помощью F-критерия;

4. Рассчитайте доверительный интервал прогноза для уровня значимости 5% ($\alpha = 0{,}05$);

3. Автокорреляция

3.1. Понятие автокорреляции. Методы ее обнаружения и устранения

Автокорреляция (последовательная корреляция) определяется как корреляция между наблюдаемыми показателями, упорядоченными во времени (временные ряды) или в пространстве (перекрестные данные). Среди основных причин, вызывающих появление автокорреляции, можно выделить ошибки спецификации, инерцию в изменении экономических показателей, эффект паутины, сглаживание данных.

Методы обнаружения автокорреляции остатков.

1. Критерий Дарбина-Уотсона.

При статистическом анализе уравнения регрессии на начальном этапе чаще других проверяют выполнимость одной предпосылки, а именно, условия статистической независимости отклонений между собой. Поскольку значения ε_i теоретического уравнения регрессии $Y = \beta_0 + \beta_1 X + \varepsilon$ остаются неизвестными ввиду неопределенности истинных значений коэффициентов регрессии, то проверяется статистическая незначимость их оценок – отклонений e_i, $i = 1,2, \ldots, n$. При этом обычно проверяется их некоррелированность, являющаяся необходимым, но недостаточным условием независимости. Причем проверяется некоррелированность не любых, а только соседних величин e_i. Соседними обычно считаются соседние во времени (при рассмотрении временных рядов) или по возрастанию объясняющей переменной X (в случае перекрестной выборки) значения e_i. Для этих величин несложно рассчитать коэффициент корреляции, называемый в этом случае коэффициентом автокорреляции первого порядка,

$$r_{e_i e_{-1}} = \frac{\sum_{i=1}^{n}(e_i - M(e_i))(e_{i-1} - M(e_{i-1}))}{\sqrt{\sum_{i=1}^{n}(e_i - M(e_i))^2 \sum_{i=1}^{n}(e_{i-1} - M(e_{i-1}))^2}} = \frac{\sum_{i=1}^{n} e_i e_{i-1}}{\sqrt{\sum_{i=1}^{n} e_i^2 \sum_{i=1}^{n} e_{i-1}^2}} \qquad (3.1)$$

При этом учитывается, что $M(e_i) = 0$, $i = 1,2, \ldots, n$.

На практике для анализа коррелированности отклонений вместо коэффициента корреляции используют тесно с ним связанную статистику Дарбина-Уотсона DW, рассчитываемую по формуле:

$$DW = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^{\,2}} \qquad (3.2)$$

Действительно,

$$\sum_{i=1}^{n}(e_i - e_{i-1})^2 = \sum_{i=1}^{n}(e_i^2 - 2e_i e_{i-1} + e_{i-1}^2) = \sum_{i=1}^{n} e_i^2 - 2\sum_{i=1}^{n} e_i e_{i-1} + \sum_{i=1}^{n} e_{i-1}^2 \approx 2\sum_{i=1}^{n} e_i^2 - 2\sum_{i=1}^{n} e_i e_{i-1}.$$

Здесь сделано допущение, что при больших n выполняется соотношение: $\sum_{i=1}^{n} e_i^2 \approx \sum_{i=1}^{n} e_{i-1}^2$.

Тогда

$$DW \approx \frac{2(\sum_{i=1}^{n} e_i^2 - \sum_{i=1}^{n} e_i e_{i-1})}{\sum_{i=1}^{n} e_i^2} = 2(1 - r_{e_i e_{i-1}}). \qquad (3.3)$$

Нетрудно заметить, что если $e_i = e_{i-1}$ при любом i, то $r_{e_i e_{i-1}} = 1$ и DW = 0. Если $e_i = -e_i$, то $r_{e_i e_{i-1}} = -1$, и DW = 4. Во всех других случаях 0<DW<4.

Согласно формуле 3.1 статистика Дарбина-Уотсона тесно связана с выборочным коэффициентом корреляции $r_{e_i e_{i-1}}$:

$$DW = 2(1 - r_{e_i e_{i-1}}) \qquad (3.4)$$

Таким образом, $0 \le DW \le 4$ и его значения могут указать на наличие либо отсутствие автокорреляции. Действительно, если $r_{e_i e_{i-1}} \approx 0$ (автокорреляция отсутствует), то $DW \approx 2$. Если $r_{e_i e_{i-1}} \approx 1$ (положительная автокорреляция), то $DW \approx 0$. Если $r_{e_i e_{i-1}} \approx -1$(отрицательная автокорреляция), то $DW \approx 4$.

Для более точного определения, какое значение DW свидетельствует об отсутствии автокорреляции, а какое о ее наличии, была построена таблица критических точек распределения Дарбина-Уотсона. По ней для заданного

уровня значимости α, числа наблюдений n и количества объясняющих переменных m определяются два значения: d_L – нижняя граница и d_U – верхняя граница.

Общая схема критерия Дарбина-Уотсона будет следующей:

1. По построенному эмпирическому уравнению регрессии определяются значения отклонений e_i для каждого наблюдения.

2. По формуле 3.2 рассчитывается статистика DW.

3. По таблице критических точек Дарбина-Уотсона определяются два числа d_L и d_U и осуществляют вывод по следующей схеме:

$0 \le DW < d_L$ – существует положительная автокорреляция,

$d_L \le DW < d_U$ – вывод о наличии автокорреляции не определен,

$d_U \le DW < 4$ – автокорреляция отсутствует,

$4- d_U \le DW < 4-d_L$ – вывод о наличии автокорреляции не определен,

$4-d_L \le DW \le 4$ – существует отрицательная автокорреляция.

2. Метод рядов.

Этот метод достаточно прост: последовательно выписываются знаки отклонений e_i. Например,

(-----)(+++++++)(---)(++++)(-),

т.е. 5 «-», 7 «+», 3 «-», 4 «+», 1 «-» при 20 наблюдениях.

Ряд определяется как непрерывная последовательность одинаковых знаков. Количество знаков в ряду называется *длиной ряда*.

Визуальное распределение знаков свидетельствует о неслучайном характере связей между отклонениями. Если рядов слишком мало по сравнению с количеством наблюдений n, то вполне вероятна положительная автокорреляция. Если же рядов слишком много, то вероятна отрицательная автокорреляция. Для более детального анализа предлагается следующая процедура. Пусть

n – объем выборки;

n_1 – общее количество знаков «+» при n наблюдениях (количество положительных отклонений e_i);

n_2 – общее количество знаков «-» при n наблюдениях (количество отрицательных отклонений e_i);

k – количество рядов.

При достаточно большом количестве наблюдений (n_1 >10, n_2 >10) и отсутствии автокорреляции СВ k имеет асимптотически нормальное распределение с

$$M(k)=\frac{2n_1n_2}{n_1+n_2}+1;\ D(k)=\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}.$$

Тогда, если $M(k)-u_{\alpha/2}\cdot D(k)<k<M(k)+u_{\alpha/2}\cdot D(k)$, то гипотеза об отсутствии автокорреляции не отклоняется.

При небольшом числе наблюдений (n_1<20, n_2<20) Свед и Эйзенхарт разработали таблицы критических значений количества рядов при n наблюдениях. Суть таблиц в следующем. На пересечении строки n_1 и столбца n_2 определяется нижнее k_1 и верхнее k_2 значения при уровне значимости α = 0,05.

Если $k_1 < k < k_2$, то говорят об отсутствии автокорреляции.

Если $k \le k_1$, то говорят о положительной автокорреляции остатков.

Если $k \ge k_2$, то говорят об отрицательной автокорреляции остатков.

3.1. Решение типовых задач

На примере задачи 2.6.1 рассмотрим обозначенные выше способы обнаружения автокорреляции остатков.

1. Критерий Дарбина-Уотсона.

$$DW=\frac{\sum_{i=1}^{n}(e_i-e_{i-1})^2}{\sum_{i=1}^{n}e_i^{\,2}}$$

Для расчета данной статистики необходимо рассчитать две суммы: $\sum_{i=1}^{n}(e_i - e_{i-1})^2$ и $\sum_{i=1}^{n} e_i^2$. Значения этих сумм можно получить используя расчетную таблицу (см. выше задача 2.6.1). В данном случае $\sum_{i=1}^{n}(e_i - e_{i-1})^2$ = 21343,64, а $\sum_{i=1}^{n} e_i^2$ = 15453,08, следовательно:

$$DW = \frac{\sum_{i=1}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^{\,2}} = \frac{21343{,}64}{15453{,}08} = 1{,}381$$

По таблице критических точек Дарбина-Уотсона определяются два числа d_L и d_U и осуществляют вывод по следующей схеме:

В данном случае d_L = 0,697, d_U = 1,641. Обозначим, полученные значения на отрезке.

Так как статистика DW = 1,381 попадает в область неопределенности, то нельзя с полной уверенностью сделать вывод о поведении отклонений e_i. Необходимо воспользоваться другим методом обнаружения автокорреляции остатков, например, методом рядов.

2. Метод рядов.

Последовательно определяются знаки отклонений e_i. В данном случае,

(+)(-----)(+++)(-),

n (объем выборки) = 10;

n_1 (общее количество знаков «+» при n наблюдениях) = 4;

n_2 (общее количество знаков «-» при n наблюдениях) = 6;

k (количество рядов) = 4.

По таблицам критических значений количества рядов для определения наличия автокорреляции по методу рядов на пересечении строки n_1 и столбца n_2 определяется нижнее k_1 и верхнее k_2 значения при уровне значимости α = 0,05. В данном случае $k_1 = 2$, верхнее $k_2 = 9$. Следовательно, если $k_1 < k < k_2$, то делаем вывод об отсутствии автокорреляции.

В случае обнаружения автокорреляции в модели, необходимо ее устранить одним из следующих методов:

1. авторегрессионная схема первого AR(1), второго порядка AR(2), третьего AR(3) порядка;

2. метод Кохрана-Оркатта;

3. метод Хилдрета-Лу

4. метод первых разностей.

3.2. Упражнения и задачи

Задача 3.2.1.

По данным задачи 2.7.7. (см. выше) определить наличие в модели автокорреляции остатков, используя статистику Дарбина-Уотсона и метод рядов.

Задача 3.2.2.

Определить наличие автокорреляции методом рядов и проверить ее присутствие с помощью статистики Дарбина-Уотсона.

e_i	8,3	4,26	-12,46	-1,86	-7,38	5,26	-9,66	-2,26	8,34	7,46

4. Гетероскедастичность

4.1. Суть гетероскедастичности

Одной из предпосылок МНК является условие постоянства дисперсий случайных отклонений (гомоскедастичность). Не должно быть априорной причины, вызывающей большую ошибку (отклонение) при одних наблюдениях и меньшую — при других. Невыполнимость данной предпосылки называется гетероскедастичностью.

На практике гетероскедастичность не так уж и редка. Проблема гетероскедастичности характерна для перекрестных данных и довольно редко встречается при рассмотрении временных рядов. Оценки, полученные по МНК, при наличии гетероскедастичности не будут эффективными (то есть они не будут иметь наименьшую дисперсию по сравнению с другими оценками данного параметра). Стандартные ошибки коэффициентов S_{b_i} будут занижены. Поэтому статистики $t_{b_i} = b_i / S_{b_i}$ будут завышены, что может привести к признанию статистически значимыми коэффициентов, которые таковыми не являются. Доверительные интервалы $b_i \pm t_{\alpha;n-m-1} \cdot S_{b_i}$ теоретических коэффициентов уравнения линейной регрессии получаются уже, чем на самом деле.

Как выяснить наличие гетероскедастичности и смягчить ее по следствия?

5.2. Методы обнаружения гетероскедастичности

Тест ранговой корреляции Спирмена

Предполагается, что дисперсии отклонений будут либо увеличиваться, либо уменьшатся с ростом значений *X*. Пусть n – число наблюдений.

Значения переменной *X* и $|e_i|$ ранжируются (упорядочиваются по величине). Обозначим через *d* разность между рангами значений переменной *X* и $|e_i|$, $d_i = N_{x_i} - N_{e_i}$

Коэффициент ранговой корреляции $r_{x,e} = 1 - 6\dfrac{\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$. (4.1)

Зададим доверительную вероятность *p*. $\alpha = (1-p)/2$. По таблицам находим граничную точку $t_{\frac{\alpha}{2};n-m-1}$. Статистика $t_{расч.} = \left| \dfrac{r_{x,e}\sqrt{n-2}}{\sqrt{1-r_{x,e}^2}} \right|$ (4.2)

Если $t_{расч.} < t_{\frac{\alpha}{2};n-m-1}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности. Иначе гипотеза об отсутствии гетероскедастичности отклоняется. В модели содержащей несколько факторов, проверка гипотезы об отсутствии гетероскедастичности проводится с помощью t-статистики для каждого из них отдельно.

Тест Голдфельда-Квандта

Предполагается, что стандартное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению x_i переменной *X* в этом наблюдении, т.е. $\sigma_i^2 = \sigma^2 x_i^2, i = 1,2,...,n$. Также предполагается, что ε_i имеет нормальное распределение и отсутствует автокорреляция остатков. Все *n* наблюдений упорядочиваются по величине *x*. Эта упорядоченная выборка делится на три примерно равные части объемом *k, n-2k и k* соответственно. При n = 30 k = 11, при n = 60 k = 22.

Для каждой из выборок объема k оценивается свое уравнение регрессии и находятся суммы квадратов отклонений $S_1 = \sum_{i=1}^{n} e_i^2$ и $S_3 = \sum_{i=n-k+1}^{n} e_i^2$ соответственно.

Задается доверительная вероятность р. $\alpha = 1 - p$. По F-таблицам находим граничную точку $F_{\alpha;k-m-1;k-m-1}$, где m – число факторов модели. Статистика $F = S_3 / S_1$.

Если $F < F_{\alpha;k-m-1;k-m-1}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности. Иначе гипотеза об отсутствии гетероскедастичности отклоняется. Для множественной регрессии тест обычно проводится для того фактора, который в максимальной степени связан с σ_i. При этом выбирают k > m+1. Если нет уверенности относительно выбора фактора x_j, то данный тест можно осуществить для каждого фактора.

4.3. Смягчение проблемы гетероскедастичности. Метод взвешенных наименьших квадратов

Гетероскедастичность не позволяет получить эффективные оценки коэффициентов уравнения регрессии, что приводит к необоснованным выводам относительно качества этих оценок. Поэтому при обнаружении гетероскедастичности возникает необходимость каких-то преобразований модели в целях ее устранения. Вид преобразований зависит от того, знаем мы поведение дисперсий отклонений или нет.

Корректировка гетероскедастичности также является достаточно серьезной проблемой. Один из возможных методов устранения гетероскедастичности – это метод взвешенных наименьших квадратов (ВНК). Для его применения необходима определенная информация, либо обоснованные предположения о величине дисперсий σ_i^2 отклонений $\varepsilon_i, i = 1,..., n$.

Например, может оказаться целесообразным предположить, что дисперсии σ_i^2 отклонений ε_i пропорциональны значениям x_i (рис.4.3.1, а) или значениям x_i^2 (рис. 4.3.1, б)

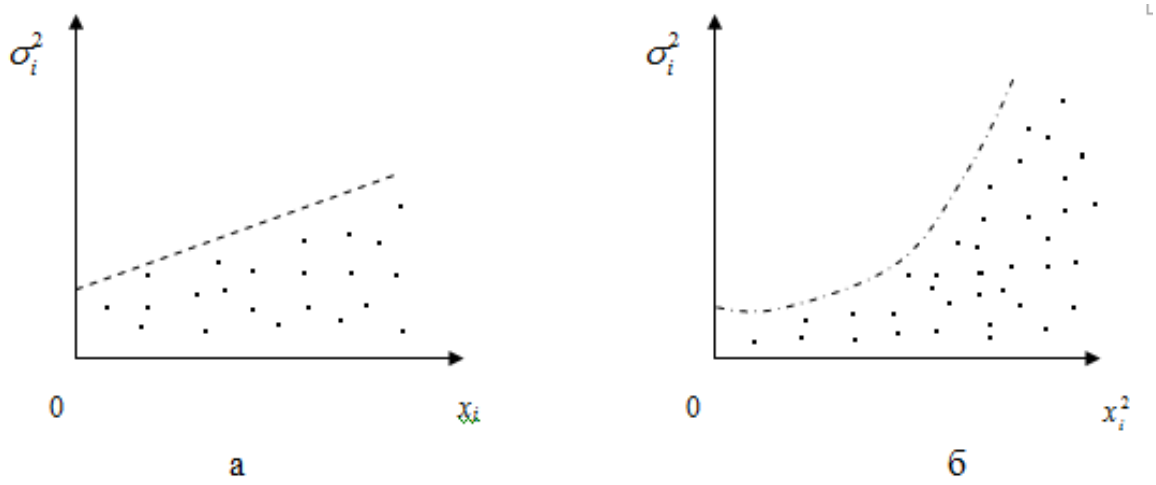


рис.4.3.1

Рассмотрим случай, когда дисперсии отклонений σ_i^2 неизвестны и пропорциональны x_i, т.е. $\sigma_i^2 = \sigma^2 x_i$. Тогда уравнение преобразуется делением его левой и правой частей на $\sqrt{x_i}$:

$$y_i / \sqrt{x_i} = (\beta_0 + \beta_1 x_i + \varepsilon_i) / \sqrt{x_i} \rightarrow y_i / \sqrt{x_i} = \beta_0 / \sqrt{x_i} + \beta_1 \sqrt{x_i} + v_i$$

где $v_i = \varepsilon_i / \sqrt{x_i}$

В случае, когда σ_i^2 неизвестны и пропорциональны x_i^2, в уравнении линейной регрессии $\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ разделим обе части на x_i.

Тогда $y_i / x_i = (\beta_0 + \beta_1 x_i + \varepsilon_i) / x_i \rightarrow y_i / x_i = \beta_0 / x_i + \beta_1 + \varepsilon_i / x_i$

Обозначим $z_i = y_i / x_i, t_i = 1/x_i, v_i = \varepsilon_i / x_i$

Тогда $z_i = \beta_1 + \beta_0 t_i + v_i$.

Для этого уравнения уже выполнено условие гомоскедастичности. Методом наименьших квадратов находим оценки коэффициентов β_0, β_1 и возвращаемся к исходному уравнению $\hat{y}_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. В случае, когда число факторов m > 1, исходное уравнение делится на переменную, которая в максимальной степени связана с σ_i.

4.4. Решение типовых задач

Задача 4.4.1

На примере задачи 2.6.1, где m=2 проверим гипотезу об отсутствии гетероскедастичности в построенной модели по тесту Спирмена. Доверительная вероятность p = 95%.

Решение:

1	2	3	4	5	6	7	8	9	10	11	12
№ гр.	x_{i1}	x_{i2}	e_i	$\|e_i\|$	Ранг x_{i1}	Ранг x_{i2}	Ранг e_i	$d_{i1}=P_{xi1}-P_{ei}$	$d_{i2}=P_{xi2}-P_{ei}$	d_{i1}^2	d_{i2}^2
1	626	1,5	61,60	61,60	1	1	9	-8	-8	64,00	64,00
2	1575	2,1	-29,73	29,73	2	2	6	-4	-4	16,00	16,00
3	2235	2,4	-4,26	4,26	3	3	1	2	2	4,00	4,00
4	2657	2,7	-29,64	29,64	4	4	5	-1	-1	1,00	1,00
5	3699	3,2	-63,35	63,35	5	5	10	-5	-5	25,00	25,00
6	4794	3,4	-18,88	18,88	6	6	3	3	3	9,00	9,00
7	5924	3,6	14,07	14,07	7	7	2	5	5	25,00	25,00
8	7279	3,7	39,14	39,14	8	8,5	7	1	1,5	1,00	2,25
9	9348	4	56,12	56,12	9	10	8	1	2	1,00	4,00
10	18805	3,7	-25,07	25,07	10	8,5	4	6	4,5	36,00	20,25

Заполним таблицу. Модули элементов четвертого столбца запишем в 5-й столбец. В 6-м, 7-м и 8-м столбцах ранжированы по возрастанию элементы 2-го, 3-го и 5-го столбцов соответственно. n = 10 наблюдений.

Коэффициент ранговой корреляции

$$r_{x_{i1},e_i} = 1 - 6\frac{\sum_{i=1}^{n} d_{i1}^2}{n(n^2-1)} = 1 - 6\cdot\frac{180}{10(100-1)} = -0{,}09.$$

По таблицам находим граничную точку $t_{\frac{\alpha}{2};n-m-1} = t_{\frac{0{,}05}{2};10-2-1} = t_{0{,}025;7} =$ 2,365.

Статистика $t_{расч.} = \left|\frac{r_{x_{i1},e_i}\sqrt{n-2}}{\sqrt{1-r_{x_{i1},e_i}^2}}\right| = \left|\frac{-0{,}09\sqrt{10-2}}{\sqrt{1-0{,}0081}}\right| = |-0{,}256| = 0{,}256.$

Таким образом $t_{расч.} < t_{\frac{\alpha}{2};n-m-1}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности по фактору X_1. В модели, содержащей несколько факторов, как уже было сказано, проверка гипотезы

об отсутствии гетероскедастичности проводится с помощью t-статистики для каждого из них отдельно. Следовательно, определим наличие гетероскедастичности по фактору X_2.

Коэффициент ранговой корреляции

$$r_{x_2,e_i} = 1 - 6\frac{\sum_{i=1}^{n} d_{i2}^2}{n(n^2-1)} = 1 - 6 \cdot \frac{170{,}5}{10(100-1)} = -0{,}03.$$

По таблицам находим граничную точку $t_{\frac{\alpha}{2};n-m-1} = t_{\frac{0{,}05}{2};10-2-1} = t_{0{,}025;7} =$ 2,365.

Статистика $t_{расч.} = \left| \frac{r_{x_{i1},e_i}\sqrt{n-2}}{\sqrt{1-r_{x_{i1},e_i}^2}} \right| = \left| \frac{-0{,}03\sqrt{10-2}}{\sqrt{1-0{,}0081}} \right| = |-0{,}085| = 0{,}085.$

Таким образом $t_{расч.} < t_{\frac{\alpha}{2};n-m-1}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности по фактору X_2.

Задача 4.4.2.

Рассматривается регрессионная линейная модель с m=2 факторами. n = 30 наблюдений. Для первых и последних k=11 наблюдений суммы квадратов отклонений S_1=20 и S_3=45 соответственно. С помощью теста Голдфельда-Квандта проверим гипотезу об отсутствии гетероскедастичности. Доверительная вероятность p = 95%.

Решение:

α = 1- p = 1 – 0,95 = 0,05. По F – таблицам Фишера находим граничную точку $F_{\alpha;k-m-1;k-m-1} = F_{0{,}05;11-2-1;11-2-1} = 3{,}44$.

Статистика F = $S_3/S_1 = 45/20 = 2{,}25 < 3{,}44$.

Таким образом, на уровне значимости 5% принимается гипотеза об отсутствии гетероскедастичности.

4.5. Упражнения и задачи

Задача 5.5.1

Рассматривается регрессионная линейная модель с m=2 факторами. n = 30 наблюдений. Для первых и последних k=11 наблюдений суммы квадратов отклонений S_1=18 и S_3=52 соответственно. С помощью теста Голдфельда-Квандта проверим гипотезу об отсутствии гетероскедастичности. Доверительная вероятность p = 99%.

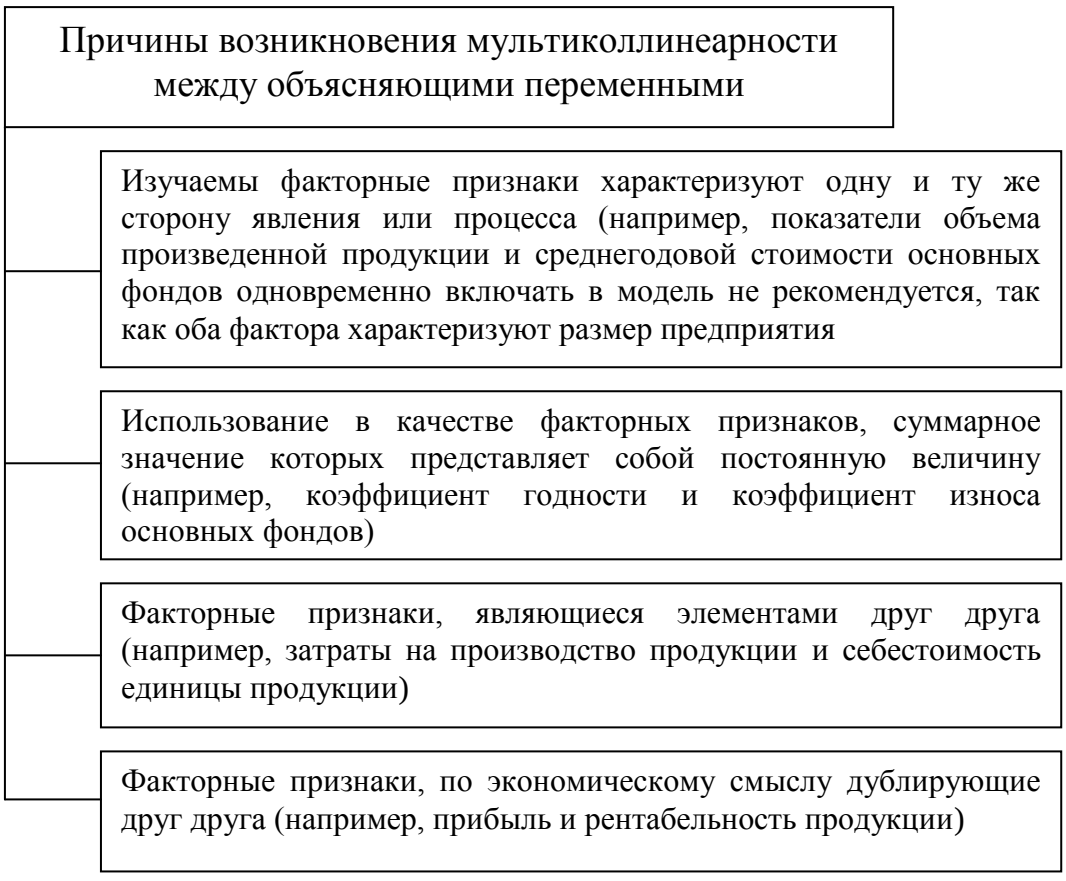
Задача 5.5.2

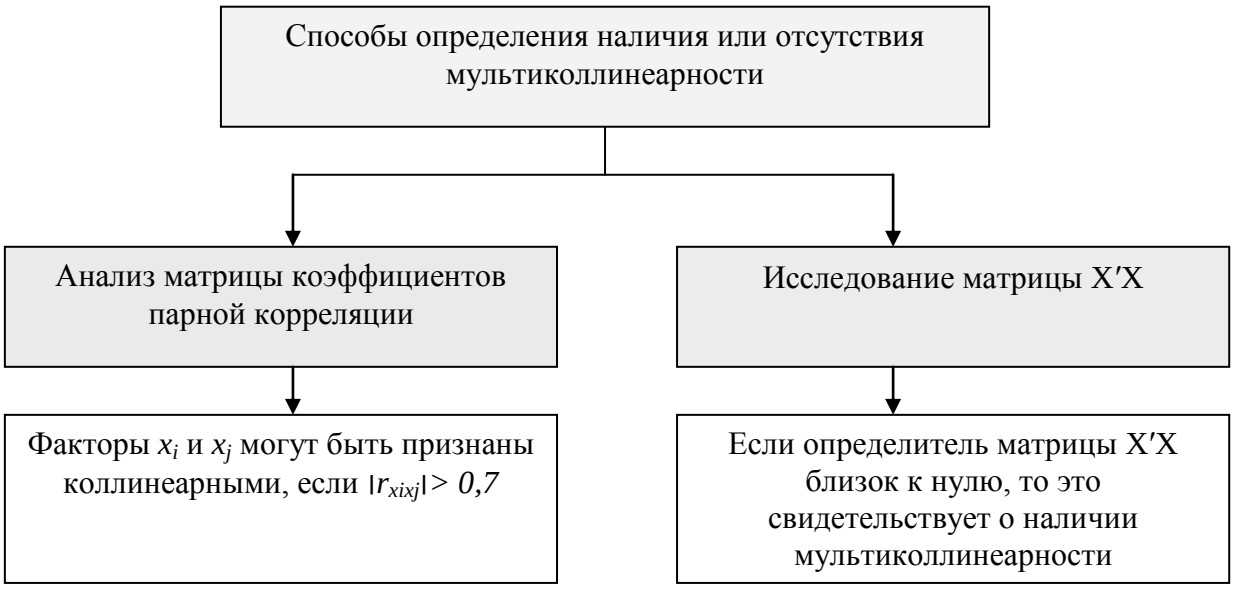
В задаче 2.7.7 определить наличие гетероскедастичности в построенной модели.

5. Мультиколлинеарность

5.1. Понятие мультиколлинеарности. Способы ее обнаружения и методы устранения

Еще одной серьезной проблемой при построении моделей множественной линейной регрессии по МНК является мультиколлинеарность − линейная взаимосвязь двух или нескольких объясняющих переменных. Причем, если объясняющие переменные связаны строгой функциональной зависимостью, то говорят о совершенной мультиколлинеарности. На практике можно столкнуться с очень высокой (или близкой к ней) мультиколлинеарностью − сильной корреляционной зависимостью между объясняющими переменными. Причины мультиколлинеарности и способы ее устранения анализируются ниже.





Устранение мультиколлинеарности возможно посредством исключения из корреляционной модели одного или нескольких линейно связанных факторных признаков или преобразования исходных факторных признаков в новые, укрупненные факторы. Вопрос о том, какой из факторов следует отбросить, решается на основе количественного и логического анализа изучаемого явления.

Описание методов устранения или снижения уровня мультиколлинеарности

Метод	**Суть метода**
Сравнение значений линейных коэффициентов корреляции	При отборе факторов предпочтение отдается тому фактору, который более тесно, чем другие факторы, связан с результативным признаком, причем желательно, чтобы связь данного факторного признака с Y была выше, чем его связь с другим факторным признаком. В данном случае имеет место расчет общих и частных коэффициентов корреляции, по результатам расчетов которых принимается окончательное решение о преобразовании исходной модели. $r_{yx_i} > r_{x_i x_k}, r_{yx_k} > r_{x_i x_k}$ и $r_{x_i x_k} < 0{,}8$
Метод включения факторов	Метод заключается в том, что в модель включаются факторы по одному в определенной последовательности. На первом шаге вводится тот фактор, который имеет наибольший коэффициент корреляции с зависимой переменной. На втором и последующих шагах в модель включается фактор, который имеет наибольший коэффициент корреляции с остатками модели. После включения каждого фактора в модель рассчитывают ее характеристики и модель проверяют на достоверность.

Метод	Суть метода
Метод исключения факторов	Метод состоит в том, что в модель включаются все факторы. Затем после построения уравнения регрессии из модели исключают фактор, коэффициент при котором незначим и имеет наименьшее значение t-статистики. После этого получают новое уравнение регрессии и снова проводят оценку значимости всех оставшихся коэффициентов регрессии. Процесс исключения факторов продолжается до тех пор, пока модель не станет удовлетворять определенным условиям и все коэффициенты регрессии не будут значимы.

В настоящее время при построении корреляционных моделей исходят из условия нормальности многомерного закона распределения генеральной совокупности. Эти условия обеспечивают линейный характер связи между изучаемыми признаками, что делает правомерным использование в качестве показателей тесноты связи парного, частного коэффициентов корреляции и коэффициента множественной корреляции. Частные коэффициенты корреляции характеризуют связи признаков из совокупности признаков при условии, что все связи этих признаков с другими признаками закреплены на условно-постоянном (среднем) уровне.

Частный коэффициент корреляции изменяется в пределах от -1 до +1. Если частный коэффициент корреляции равен ±1, то связь между двумя величинами функциональная, а равенство нулю свидетельствует о линейной независимости этих величин.

5.2. Решение типовых задач

Задача 5.2.1

На примере задачи 2.6.1, где m=2, проверим наличие мультиколлинеарности в построенной модели по следующей формуле:

$$r_{x_1x_2} = \frac{[6]}{\sqrt{[1]*[2]}} = \frac{2761{,}574}{\sqrt{2564440{,}216*0{,}6041}} = 0{,}702.$$

Следовательно, есть подозрение, что в модели присутствует некоторая мультиколлинеарность. Рассчитаем частные коэффициенты корреляции, на основании которых сделаем окончательный вывод о значимости обнаруженной в модели мультиколлинеарности и необходимости ее устранения.

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

$$= \frac{0{,}9433 - 0{,}8934 \cdot 0{,}7016}{\sqrt{(1-0{,}8934^2)(1-0{,}7016^2)}} = \frac{0{,}3165}{0{,}2018 \cdot 0{,}5078} = 0{,}9888.$$

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

$$= \frac{0{,}8934 - 0{,}9433 \cdot 0{,}7016}{\sqrt{(1-0{,}9433^2)(1-0{,}7016^2)}} = \frac{0{,}2316}{\sqrt{0{,}1102 \cdot 0{,}5078}} = 0{,}9789.$$

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

$$= \frac{0{,}7016 - 0{,}9433 \cdot 0{,}8934}{\sqrt{(1-0{,}9433^2)(1-0{,}8934^2)}} = \frac{-0{,}1411}{\sqrt{0{,}1102 \cdot 0{,}2018}} = -0{,}9463,$$

где

$$r_{yx_2} = \frac{[5]}{\sqrt{[2]*[3]}} = \frac{406{,}299}{\sqrt{0{,}6041*34236{,}01}} = 0{,}8934.$$

$$r_{yx_1} = \frac{[4]}{\sqrt{[1]*[3]}} = \frac{279519{,}76}{\sqrt{256444021{,}6*34236{,}01}} = 0{,}9433.$$

Таким образом, полученные величины частных коэффициентов корреляции очень близки по модулю к единице, т.е. теснота связи между расходами на питание и каждым из исследуемых факторов при неизменном значении другого весьма значительна. Возможно, в данной модели наличие мультиколлинеарности не настолько ухудшает ее качество. Иногда

мультиколлинеарность не является таким уж «злом», чтобы прилагать существенные усилия по ее устранению. Все зависит от целей исследования. Если основная задача модели – прогноз будущих значений результативного признака, то при $R^2 \geq 0{,}9$ наличие мультиколлинеарности обычно не сказывается на прогнозных качествах модели.

Итак, можно сказать, что единого метода устранения мультиколлинеарности не существует. Простейшим методом устранения мультиколлинеарности является исключение из модели ряда коррелированных переменных. В прикладных моделях лучше не сокращать число факторов до тех пор, пока мультиколлинеарность не станет серьезной проблемой. Иногда для уменьшения мультиколлинеарности достаточно увеличить объем выборки. Но при этом может усилиться автокорреляция.

Задача 5.2.2

По 18 наблюдениям получены следующие результаты:

$\sum x_{i1} = 11{,}48$; $\sum x_{i1}^2 = 7{,}83$; $\sum x_{i2} = 440{,}10$;

$\sum x_{i2}^2 = 13379{,}25$; $\sum y_i = 62640$; $\sum x_{i1} x_{i2} = 249{,}69$;

$\sum x_{i1} y_i = 44048{,}49$; $\sum x_{i2} y_i = 1303131$; $\sum y_i^2 = 279130200$;

$\sum e_i^2 = 2776879{,}57$.

Требуется:

а) по МНК определить параметры множественной линейной регрессии $\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2$;

б) проверить модель на наличие мультиколлинеарности между объясняющими переменными.

Решение:

а) для вычисления коэффициентов уравнения регрессии необходимо определить значения 6-ти сумм:

1. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2=\left[\bar{x}_1^2-(\bar{x}_1)^2\right]\cdot n = 0{,}03\cdot n.$

2. $\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2=\left[\bar{x}_2^2-(\bar{x}_2)^2\right]\cdot n=145{,}49\cdot n.$

3. $\sum_{i=1}^{n}(y_i-\bar{y})^2=\left[\bar{y}^2-(\bar{y})^2\right]\cdot n= 339686{,}33\cdot n.$

4. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_2)(y_i-\bar{y})=\left[\bar{x}_1\bar{y}-\bar{x}_1\cdot\bar{y}\right]\cdot n=227{,}48\cdot n.$

5. $\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})=\left[\bar{x}_2\bar{y}-\bar{x}_2\cdot\bar{y}\right]\cdot n = -1268{,}83\cdot n.$

6. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)=\left[\bar{x}_1\bar{x}_2-\bar{x}_1\cdot\bar{x}_2\right]\cdot n = -1{,}72\cdot n.$

Подставим полученные значения 6-ти сумм в формулы для расчета коэффициентов уравнения регрессии (m=2):

$a=\bar{y}-b_1*\bar{x}_1-b_2*\bar{x}_2;$

$b_1=\frac{[2]*[4]-[5]*[6]}{[1]*[2]-[6]^2}=9683{,}94;$

$b_2=\frac{[1]*[5]-[4]*[6]}{[1]*[2]-[6]^2}=27{,}48;$

$b_0=\bar{y}-b_1*\bar{x}_1-b_2*\bar{x}_2=$ -3368,71.

Таким образом, мы получили эмпирические значения параметров множественной линейной регрессии, которая имеет следующий вид:

$$\hat{y}_i=b_0+b_1x_1+b_2x_2=\text{ - }3368{,}71+9683{,}94x_1+27{,}48x_2;$$

б) Проверим модель на наличие мультиколлинеарности.

$$r_{x_1x_2}=\frac{[6]}{\sqrt{[1]*[2]}}=\frac{-1{,}72}{\sqrt{0{,}03\cdot 145{,}49}}=-0{,}82.$$

$$\left|r_{x_1x_2}\right|=\left|-0{,}82\right|=0{,}82.$$

Следовательно, в модели между переменными X_1 и X_2 присутствует мультиколлинеарность. Поэтому, если при дальнейшей проверке на гетероскедастичность окажется, что она существует по переменной X_2 и

коэффициент b_2 не значим, то из модели необходимо исключить переменную X_2. Проверим, целесообразно ли это делать без такой проверки. Для этого есть два варианта. Первый – построить два уравнения регрессии $Y(X_1)$ и $Y(X_2)$ и посмотреть, у какой из этих моделей качество лучше. Если подтвердится, что у первой модели качество выше, чем у второй, то сразу без дальнейшей проверки на гетероскедастичность можно исключить из модели фактор X_2. Эту же проверку можно сделать иначе, для чего необходимо рассчитать частные коэффициенты корреляции, на основании которых сделаем окончательный вывод.

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_2}^2)(1-r_{x_1x_2}^2)}} = \frac{0{,}71-(-0{,}57)\cdot(-0{,}82)}{\sqrt{(1-0{,}57^2)(1-0{,}82^2)}} = \frac{0{,}2426}{0{,}82\cdot 0{,}57} = 0{,}6,$$

$$r_{yx_2(x_1)} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1x_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{x_1x_2}^2)}} = \frac{-0{,}57-0{,}71\cdot(-0{,}82)}{\sqrt{(1-0{,}71^2)(1-0{,}82^2)}} = \frac{0{,}012}{0{,}71\cdot 0{,}57} = 0{,}029,$$

$$r_{x_1x_2(y)} = \frac{r_{x_1x_2} - r_{yx_1} \cdot r_{yx_2}}{\sqrt{(1-r_{yx_1}^2)(1-r_{yx_2}^2)}} = \frac{-0{,}82-0{,}71\cdot(-0{,}57)}{\sqrt{(1-0{,}71^2)(1-(-0{,}57)^2)}} = \frac{-0{,}4153}{0{,}71\cdot 0{,}82} = -0{,}71,$$

где

$$r_{yx_2} = \frac{[5]}{\sqrt{[2]*[3]}} = \frac{-1268{,}83}{\sqrt{145{,}49\cdot 339683{,}33}} = -0{,}57,$$

$$r_{yx_1} = \frac{[4]}{\sqrt{[1]*[3]}} = \frac{227{,}48}{\sqrt{0{,}03\cdot 339683{,}33}} = 0{,}71.$$

Следовательно, связь между yx_1 без учета влияния фактора x_2 намного существенней, чем связь между yx_2 без учета фактора x_1. Поэтому без проверки на гетероскедастичность остатков в исходной модели можно сразу предложить изменить ее спецификацию, исключив из модели фактор x_2. Но поскольку коэффициент $\left|r_{x_1x_2(y)}\right| > 0{,}3$, то нельзя просто пренебречь фактором x_2, необходимо преобразовать выборку, перейдя к новой

$\left(\frac{Y}{X_{i2}};\frac{X_{i1}}{X_{i2}}\right)$. В построенной модели мультиколлинеарность будет отсутствовать. При этом она будет учитывать оба объясняющих фактора.

Задача 5.2.3

По 25 наблюдениям получены следующие результаты:

Для вычисления коэффициентов уравнения регрессии необходимо определить значения 6-ти сумм:

1. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2=\left[\bar{x}_1^2-(\bar{x}_1)^2\right]*n=[3{,}4129]\cdot n.$

2. $\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2=\left[\bar{x}_2^2-(\bar{x}_2)^2\right]*n=[2{,}6040]\cdot n.$

3. $\sum_{i=1}^{n}(y_i-\bar{y})^2=\left[\bar{y}^2-(\bar{y})^2\right]*n=[8{,}3748]\cdot n.$

4. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_2)(y_i-\bar{y})=\left[\bar{x}_1\bar{y}-\bar{x}_1*\bar{y}\right]*n=[1{,}7862]\cdot n.$

5. $\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)(y_i-\bar{y})=\left[\bar{x}_2\bar{y}-\bar{x}_2*\bar{y}\right]*n=[3{,}3025]\cdot n.$

6. $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)=\left[\bar{x}_1\bar{x}_2-\bar{x}_1*\bar{x}_2\right]*n=[-0{,}7359]\cdot n.$

Требуется:

а) по МНК определить параметры множественной линейной регрессии $\hat{y}_i=b_0+b_1x_1+b_2x_2$;

б) проверить модель на наличие мультиколлинеарности между объясняющими переменными.

Решение:

а. Подставим полученные значения 6-ти сумм в формулы для расчета коэффициентов уравнения регрессии (m=2):

$a=\bar{y}-b_1*\bar{x}_1-b_2*\bar{x}_2;$

$$b_1 = \frac{[2]*[4]-[5]*[6]}{[1]*[2]-[6]^2} = 0{,}8486;$$

$$b_2 = \frac{[1]*[5]-[4]*[6]}{[1]*[2]-[6]^2} = 1{,}5080;$$

$$b_0 = \bar{y} - b_1 * \bar{x}_1 - b_2 * \bar{x}_2 = -1{,}8388.$$

Таким образом, мы получили эмпирические значения параметров множественной линейной регрессии, которая имеет следующий вид:

$$\hat{y}_i = b_0 + b_1 x_1 + b_2 x_2 = -1{,}8388 + 0{,}8486 x_1 + 1{,}5080 x_2.$$

б) Проверим модель на наличие мультиколлинеарности.

$$r_{x_1 x_2} = \frac{[6]}{\sqrt{[1]*[2]}} = \frac{-0{,}7359}{\sqrt{3{,}4129 \cdot 2{,}6040}} = -0{,}2469.$$

$$\left| r_{x_1 x_2} \right| = \left| -0{,}2469 \right| = 0{,}2469.$$

Следовательно, в модели между переменными X_1 и X_2 мультиколлинеарность отсутствует. Проверим отсутствие мультиколлинеарности по общей схеме. Для этого вычислим частные коэффициенты корреляции:

$$r_{yx_1(x_2)} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}^2)(1 - r_{x_1 x_2}^2)}} =$$

$$\frac{0{,}6177 - 0{,}3805 \cdot (-0{,}2469)}{\sqrt{(1 - 0{,}3805^2)(1 - (-0{,}2469)^2)}} = \frac{0{,}7116}{0{,}8552 \cdot 0{,}9390} = 0{,}8862,$$

$$r_{yx_2(x_1)} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{yx_1}^2)(1 - r_{x_1 x_2}^2)}} =$$

$$\frac{0{,}3805 - 0{,}6177 \cdot (-0{,}2469)}{\sqrt{(1 - 0{,}6177^2)(1 - (-0{,}2469)^2)}} = \frac{0{,}5330}{0{,}6185 \cdot 0{,}9390} = 0{,}9177,$$

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

$$\frac{-0{,}2469 - 0{,}6177 \cdot 0{,}3805}{\sqrt{(1-0{,}6177^2)(1-0{,}3805^2)}} = \frac{-0{,}4819}{0{,}6185 \cdot 0{,}8552} = -0{,}9111,$$

где

$$r_{yx_2} = \frac{[5]}{\sqrt{[2]*[3]}} = \frac{1{,}7862}{\sqrt{2{,}6040 \cdot 8{,}3748}} = 0{,}3805,$$

$$r_{yx_1} = \frac{[4]}{\sqrt{[1]*[3]}} = \frac{3{,}3025}{\sqrt{3{,}4129 \cdot 8{,}3748}} = 0{,}6177.$$

Таким образом, из проведенного исследования можно сделать вывод, что фактор X_2 в большей степени коррелирует с величиной Y, чем объясняющая переменная X_1, так как $r_{yx1} > r_{yx2}$, и соответственно $R_2^2 > R_1^2$. Поэтому при исключении одной из переменных по коэффициентам корреляции в случае обнаружения мультиколлинеарности между объясняющими переменными, необходимо было бы исключить из модели фактор X_1, так как $r_{yx_2(x_1)} < r_{yx_1(x_2)}$. Коэффициент частной корреляции $\left|r_{x_1x_2(y)}\right| \approx 1$, определяющий внутреннюю связь между переменными X_1 и X_2, оказался по модулю близок к 1, следовательно, экономически оправданным является составление зависимости между этими переменными.

Задача 5.2.4

По выборке объема n = 50 для X_1, X_2, X_3 построена следующая корреляционная матрица

$$\begin{bmatrix} 1{,}0 & 0{,}45 & -0{,}35 \\ 0{,}45 & 1{,}0 & 0{,}52 \\ -0{,}35 & 0{,}52 & 1{,}0 \end{bmatrix}$$

1. Найдите и оцените статистическую значимость частных коэффициентов корреляции $r_{12,3}$; $r_{23,1}$; $r_{13,2}$.

2. При рассмотрении какой регрессии будет иметь место мультиколлинеарность.

Решение: для оценки статистической значимости частных коэффициентов корреляции необходимо рассчитать обратную матрицу.

$$|\mathrm{R}| = \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} = \begin{bmatrix} 1{,}0 & 0{,}45 & -0{,}35 \\ 0{,}45 & 1{,}0 & 0{,}52 \\ -0{,}35 & 0{,}52 & 1{,}0 \end{bmatrix} = 0{,}4046.$$

$$\mathrm{R}^{-1} = \frac{1}{|R|} \cdot \begin{bmatrix} 1 & r_{12} & r_{13} \\ r_{21} & 1 & r_{23} \\ r_{31} & r_{32} & 1 \end{bmatrix} = \mathrm{C}^{*} = \begin{pmatrix} C_{11}^{*} & C_{12}^{*} & C_{13}^{*} \\ C_{21}^{*} & C_{22}^{*} & C_{23}^{*} \\ C_{31}^{*} & C_{32}^{*} & C_{33}^{*} \end{pmatrix} = \begin{pmatrix} 3{,}03 & -2{,}62 & 2{,}43 \\ -2{,}62 & 3{,}64 & -2{,}81 \\ 2{,}43 & -2{,}81 & 3{,}31 \end{pmatrix}.$$

$$r_{12,3} = \frac{-C_{12}^{*}}{\sqrt{C_{11}^{*} \cdot C_{22}^{*}}} = \frac{2{,}62}{\sqrt{3{,}03 \cdot 3{,}64}} = 0{,}79.$$

$$r_{23,1} = \frac{-C_{23}^{*}}{\sqrt{C_{22}^{*} \cdot C_{33}^{*}}} = \frac{-2{,}81}{\sqrt{3{,}64 \cdot 3{,}31}} = 0{,}81.$$

$$r_{13,2} = \frac{-C_{13}^{*}}{\sqrt{C_{11}^{*} \cdot C_{33}^{*}}} = \frac{-2{,}43}{\sqrt{3{,}03 \cdot 3{,}31}} = -0{,}77.$$

Таким образом, мультиколлинеарность присутствует во всех трех случаях, поскольку одним из основных признаков ее наличия в модели являются высокие значения частных коэффициентов корреляции. Кроме того, для оценки мультиколлинеарности факторов может использоваться определитель матрицы парных коэффициентов корреляции между факторами. Чем ближе к 0 определитель матрицы межфакторной корреляции, тем сильнее мультиколлинеарность факторов и ненадежнее результаты множественной регрессии. И наоборот, чем ближе к 1 определитель матрицы межфакторной корреляции, тем меньше мультиколлинеарность факторов.

5.3. Упражнения и задачи

Задача 5.3.1

В модели три фактора X_1, X_2, X_3. Коэффициенты корреляции $r_{12} = 0{,}42$, $r_{13} = -0{,}36$, $r_{23} = 0{,}53$. Определить наличие мультиколлинеарности между обозначенными факторами, используя обратную матрицу.

Задача 5.3.2

В задачах 2.6.1 и 2.7.7 определить наличие мультиколлинеарности между объясняющими переменными.

6. Фиктивные переменные в регрессионных моделях

6.1. Необходимость использования в моделях фиктивных переменных

Очень часто в регрессионных моделях в качестве объясняющих переменных используют не только количественные (определяются численно), но и качественные. Обычно в моделях влияние качественного фактора выражается в виде фиктивной переменной, которая отражает два противоположных состояния качественного фактора, которые можно охарактеризовать как: да или нет. Например,

$$D = \begin{cases} 0, \text{ если сотрудник не имеет высшего образования,} \\ 1, \text{ если сотрудник имеет высшее образование.} \end{cases}$$

Модели, в которых объясняющие переменные носят как количественный, так и качественный характер, называются ANCOVA – моделями. Если качественная переменная имеет k альтернативных значений, то при моделировании используют только k – 1 фиктивную переменную. Значения фиктивной переменной можно менять на противоположные. Суть модели от этого не изменится.

6.2. ANCOVA – модель

Пусть рассматривается уравнение $Y = \beta_0 + \beta_1 x + \varepsilon_i$ и в модель решено ввести фиктивную переменную D.

Это можно сделать двумя способами: $Y = \beta_0 + \beta_1 x + \gamma_1 D + \varepsilon_I$ и $Y = \beta_0 + \beta_1 x + \gamma_1 D + \gamma_1 Dx + \varepsilon_i$.

Коэффициенты γ_1 и γ_1 называются *дифференциальным свободным членом* и *дифференциальным угловым коэффициентом* соответственно.

Фиктивная переменная D во втором уравнении используется как в аддитивном ($\gamma_1 D$), так и в мультипликативном виде ($\gamma_1 Dx$), что позволяет

фактически разбивать рассматриваемую зависимость на две части, связанные с периодами изменения рассматриваемой в модели переменной.

6.3. Решение типовых задач

Пример 6.3.1

Исследуется надежность станков трех производителей A, B, C. При этом учитывается возраст станка M (в месяцах) и время H (в часах) безаварийной работы до последней поломки. Выборка из 40 станков дала следующие результаты.

Фирма	Y	X	F	R	Фирма	Y	X	F	R
A	280	23	0	0	A	236	48	0	0
B	230	30	1	0	A	205	59	0	0
C	112	65	1	1	A	240	25	0	0
A	176	69	0	0	B	65	69	1	0
C	90	75	1	1	A	115	71	0	0
A	176	63	0	0	C	200	26	1	1
B	216	25	1	0	B	126	45	0	0
C	110	75	1	1	A	225	40	1	0
B	45	75	1	0	C	210	30	1	1
A	200	52	0	0	B	45	69	0	0
B	265	20	1	0	A	260	30	1	0
C	148	70	1	1	B	220	22	1	0
C	150	62	1	1	B	194	33	1	0
B	176	40	1	0	C	156	48	0	1
A	123	66	0	0	B	100	75	0	0
A	245	20	0	0	A	240	21	1	0
C	176	39	1	1	B	88	56	1	1
B	260	25	1	0	A	120	58	0	1

У уравнения регрессии $H = \beta_0 + \beta_1 M + \varepsilon$ без учета различия станков различных фирм невысокий коэффициент детерминации $R^2 = 0{,}686$. Поэтому нужно учитывать производителя станков.

Качественная переменная «Производитель станков» может принимать k = 3 значения (A, B, C).

Поэтому нужно ввести в модель k-1 = 3-1=2 фиктивные переменные F и R.

$$F = \begin{cases} 0, \text{ если производитель A,} \\ 1, \text{ если производитель B или C.} \end{cases}$$

$$R = \begin{cases} 0, \text{ если производитель A или B,} \\ 1, \text{ если производитель C.} \end{cases}$$

Для производителя A: $F = R = 0$, для производителя B: $F = 1$, $R = 0$, для производителя C: $F = R = 1$.

Теперь нужно оценить коэффициенты уравнения.

$Y = \beta_0 + \beta_1 M + \gamma_1 F + \gamma_2 R + \varepsilon_i$.

Пример 6.3.2

На основе имеющихся данных, исследуется эффективность лекарств Y в зависимости от X (возраст пациента). При этом сравнивается эффективность лекарств A и B.

Вводится фиктивная переменная D:

$$D = \begin{cases} 0, \text{ если лекарство A,} \\ 1, \text{ если лекарство B.} \end{cases}$$

Возможен один из трех вариантов: $Y = \beta_0 + \beta_1 x + \varepsilon_i$, $Y = \beta_0 + \beta_1 x + \gamma_1 D + \varepsilon_i$ и $Y = \beta_0 + \beta_1 x + \gamma_1 D + \gamma_1 D x + \varepsilon_i$.

Лек-во	Y	X	D	Dx	Лек-во	Y	X	D	Dx
A	54	69	0	0	B	30	40	1	40
B	30	48	1	48	B	23	41	1	41
A	58	73	0	0	A	21	55	0	0
B	66	64	1	64	B	43	45	1	45
B	67	60	1	60	A	38	58	0	0
A	64	62	0	0	B	43	58	1	58
A	67	70	0	0	A	43	64	0	0
A	33	52	0	0	B	45	55	1	55
A	33	63	0	0	B	48	57	1	57
B	42	48	1	48	A	48	63	0	0
B	33	46	1	46	A	53	60	0	0
A	28	55	0	0	B	58	62	1	62

6.4. Упражнения и задачи

Задача 6.4.1

В примере 6.3.1 оценить коэффициенты уравнения $Y = \beta_0 + \beta_1 M + \gamma_1 F + \gamma_2 R + \varepsilon_i$.

Задача 6.4.2

В примере 6.3.1 определить, какая из моделей является наиболее предпочтительной для прогнозов.

7. Контрольные задания

7.1. Парная линейная регрессия

1. Определить по МНК оценки коэффициентов уравнения регрессии
2. Найти оценки дисперсий оценок коэффициентов регрессии.
3. Проверить статистическую значимость коэффициентов, входящих в уравнение регрессии.
4. Найти доверительные интервалы для коэффициентов регрессии с доверительной вероятностью $\gamma = 0.9$ для чётных вариантов и 0.95 для нечётных.
5. Найти коэффициент детерминации и на уровне значимости 0.05 проверить значимость линейной функции регрессии с помощью F-критерия Фишера.
6. Найти точечное (с надёжностью 0.9) предсказание зависимой переменной при значении объясняющей переменной, равном максимальному наблюдённому её значению, увеличенному на 10%.

Вариант 1

x_1	32	30	36	40	41	47	56	54	60	55	61	67	69	76
y	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Вариант 2

x_1	55	46	40	39	35	29	31	75	68	66	60	54	59	53
y	33	32	30	29	27	23	19	47	44	42	40	39	37	36

Вариант 3

x_1	48	57	55	61	56	62	68	70	77	42	41	37	31	33
y	34	35	38	39	41	42	44	46	49	32	31	29	25	21

Вариант 4

x_1	52	54	45	39	38	34	28	30	74	67	65	59	53	58
y	35	32	31	29	28	26	22	18	46	43	41	39	38	36

Вариант 5

x_1	43	49	58	56	62	57	63	69	71	78	34	32	38	42
y	33	35	36	39	40	42	43	45	47	50	22	26	30	32

Вариант 6

x_1	52	57	51	53	44	38	37	33	27	29	73	66	64	58
y	37	35	34	31	30	28	27	25	21	17	45	42	40	38

Вариант 7

x_1	39	43	44	50	59	57	63	58	64	70	72	79	35	33
y	31	33	34	36	37	40	41	43	44	46	48	51	23	27

Вариант 8

x_1	63	57	51	56	50	52	43	37	36	32	26	28	72	65
y	39	37	36	34	33	30	29	27	26	24	20	16	44	41

Вариант 9

x_1	64	59	65	71	73	80	36	34	40	44	45	51	60	58
y	42	44	45	47	49	52	24	28	32	34	35	37	38	41

Вариант 10

x_1	46	52	61	59	65	60	66	72	74	81	37	35	41	45
y	36	38	39	42	43	45	46	48	50	53	25	29	33	35

Вариант 11

x_1	62	30	36	50	41	47	56	54	60	55	61	67	69	66
y	5	2	2	3	3	3	4	3	3	4	4	4	4	4

Вариант 12

x_1	45	46	40	39	35	29	61	75	68	66	60	54	59	53
y	3	2	3	9	7	3	9	7	4	2	6	9	7	6

Вариант 13

x_1	33	31	41	39	46	43	34	38	42	35	39	44	40	41
y	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Вариант 14

x_1	55	46	40	39	35	29	31	75	68	66	60	54	59	53
y	36	38	39	42	43	45	46	48	50	53	25	29	33	35

Вариант 15

x_1	48	57	55	61	56	62	68	70	77	42	41	37	31	33
y	5	2	2	3	3	3	4	3	3	4	4	4	4	4

Вариант 16

x_1	52	54	45	39	38	34	28	30	74	67	65	59	53	58
y	31	33	34	36	37	40	41	43	44	46	48	51	23	27

Вариант 17

x_1	43	49	58	56	62	57	63	69	71	78	34	32	38	42
y	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Вариант 18

x_1	52	57	51	53	44	38	37	33	27	29	73	66	64	58
y	38	35	34	31	30	28	27	25	21	17	45	42	40	38

Вариант 19

x_1	39	43	44	50	59	57	63	58	64	70	72	79	35	33
y	36	38	39	42	43	45	46	48	50	53	25	29	33	35

Вариант 20

x_1	45	46	40	39	35	29	61	75	68	66	60	54	59	53
y	39	37	36	34	33	30	29	27	26	24	20	16	44	41

Вариант 21

x_1	64	59	65	71	73	80	36	34	40	44	45	51	60	58
y	3	2	3	9	7	3	9	7	4	2	6	9	7	6

Вариант 22

x_1	62	30	36	50	41	47	56	54	60	55	61	67	69	66
y	36	38	39	42	43	45	46	48	50	53	25	29	33	35

Вариант 23

x_1	52	57	51	53	44	38	37	33	27	29	73	66	64	58
y	5	2	2	3	3	3	4	3	3	4	4	4	4	4

Вариант 24

x_1	46	52	61	59	65	60	66	72	74	81	37	35	41	45
y	3	2	3	9	7	3	9	7	4	2	6	9	7	6

Вариант 25

x_1	32	39	35	31	40	43	36	38	28	30	38	37	41	36
y	39	37	36	34	33	30	29	27	26	24	20	16	44	41

Вариант 26

x_1	44	42	49	46	37	41	45	38	42	47	43	44	36	34
y	42	44	45	47	49	52	24	28	32	34	35	37	38	41

Вариант 27

x_1	46	52	61	59	65	60	66	72	74	81	37	35	41	45
y	39	37	36	34	33	30	29	27	26	24	20	16	44	41

Вариант 28

x_1	62	30	36	50	41	47	56	54	60	55	61	67	69	66
y	5	2	2	3	3	3	4	3	3	4	4	4	4	4

Вариант 29

x_1	52	57	51	53	44	38	37	33	27	29	73	66	64	58
y	3	2	3	9	7	3	9	7	4	2	6	9	7	6

Вариант 30

x_1	33	31	41	39	46	43	34	38	42	35	39	44	40	41
y	39	37	36	34	33	30	29	27	26	24	20	16	44	41

Вариант 31

x_1	55	46	40	39	35	29	31	75	68	66	60	54	59	53
y	42	44	45	47	49	52	24	28	32	34	35	37	38	41

7.2. Множественная линейная регрессия

Вариант 1−12

Предполагается, что объем предложения товара y линейно зависит от цены товара X_1 и зарплаты сотрудников X_2: $y=\beta_0+\beta_1X_1+\beta_2X_2$. Статистические данные собраны за 14 месяцев.

Требуется:

1. Оценить по МНК коэффициенты уравнения регрессии.

2. Дайте сравнительную оценку силы связи факторов с результатом с помощью средних (общих) коэффициентов эластичности.

3. Рассчитать частные коэффициенты эластичности по факторам входящим в модель.

4. Оценить статистическую значимость параметров регрессионной модели с помощью t-критерия; нулевую гипотезу о значимости уравнения и показателей тесноты связи проверьте с помощью F-критерия.

5. Рассчитайте доверительный интервал прогноза для уровня значимости 5% ($\alpha = 0{,}05$).

Вариант 1

x_1	32	30	36	40	41	47	56	54	60	55	61	67	69	76
x_2	33	31	41	39	46	43	34	38	42	35	39	44	40	41
y	20	24	28	30	31	33	34	37	38	40	41	43	45	48

Вариант 2

x_1	55	46	40	39	35	29	31	75	68	66	60	54	59	53
x_2	33	42	45	38	40	30.	32	40	39	43	38	34	41	37
y	33	32	30	29	27	23	19	47	44	42	40	39	37	36

Вариант 3

x_1	48	57	55	61	56	62	68	70	77	42	41	37	31	33
x_2	44	35	39	43	36	40	45	41	42	47	40	42	32	34
y	34	35	38	39	41	42	44	46	49	32	31	29	25	21

Вариант 4

x_1	52	54	45	39	38	34	28	30	74	67	65	59	53	58
x_2	36	32	41	44	37	39	29	31	39	38	42	37	33	40
y	35	32	31	29	28	26	22	18	46	43	41	39	38	36

Вариант 5

x_1	43	49	58	56	62	57	63	69	71	78	34	32	38	42
x_2	48	45	36	40	44	37	41	46	42	43	35	33	43	41
y	33	35	36	39	40	42	43	45	47	50	22	26	30	32

Вариант 6

x_1	52	57	51	53	44	38	37	33	27	29	73	66	64	58
x_2	32	39	35	31	40	43	36	38	28	30	38	37	41	36
y	37	35	34	31	30	28	27	25	21	17	45	42	40	38

Вариант 7

x_1	39	43	44	50	59	57	63	58	64	70	72	79	35	33
x_2	44	42	49	46	37	41	45	38	42	47	43	44	36	34
y	31	33	34	36	37	40	41	43	44	46	48	51	23	27

Вариант 8

x_1	63	57	51	56	50	52	43	37	36	32	26	28	72	65
x_2	40	35	31	38	34	30	39	42	35	37	27	29	37	36
y	39	37	36	34	33	30	29	27	26	24	20	16	44	41

Вариант 9

x_1	64	59	65	71	73	80	36	34	40	44	45	51	60	58
x_2	46	39	43	48	44	45	37	35	45	43	50	47	38	42
y	42	44	45	47	49	52	24	28	32	34	35	37	38	41

Вариант 10

x1	46	52	61	59	65	60	66	72	74	81	37	35	41	45
x2	51	48	39	43	47	40	44	49	45	46	38	36	46	44
y	36	38	39	42	43	45	46	48	50	53	25	29	33	35

Вариант 11

x_1	62	30	36	50	41	47	56	54	60	55	61	67	69	66
x_2	43	51	41	39	46	43	34	38	42	25	39	44	40	41
y	5	2	2	3	3	3	4	3	3	4	4	4	4	4

Вариант 12

x_1	45	46	40	39	35	29	61	75	68	66	60	54	59	53
x_2	63	42	45	38	40	30.	32	40	39	43	38	34	41	37
y	3	2	3	9	7	3	9	7	4	2	6	9	7	6

Вариант 13−25

По 20 предприятиям региона изучается зависимость выработки продукции на одного работника y (тыс. руб.) от ввода в действие новых основных фондов x_1 (% от стоимости фондов на конец года) и от удельного веса рабочих высокой квалификации в общей численности рабочих x_2 (%) (смотри таблицу своего варианта).

Требуется:

1. По МНК определить параметры множественной линейной регрессии $y = a + b_1 x_1 + b_2 x_2$.

2. Оценить статистическую значимость найденных эмпирических коэффициентов регрессии b_1, b_2.

3. Сравнить влияние факторов на результат при помощи средних коэффициентов эластичности.

4. Построить 95-% доверительные интервалы для найденных коэффициентов.

5. Вычислить коэффициент детерминации R^2 и оценить его статистическую значимость при $\alpha = 0{,}05$.

6. Проверить качество построенного уравнения регрессии с помощью F-статистики Фишера.

7. Оценить целесообразность включения в уравнение одного фактора после другого с помощью частных F-статистик Фишера.

Вариант 13

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	6	3,6	9	11	9	6,3	21
2	6	3,6	12	12	11	6,4	22
3	6	3,9	14	13	11	7	24
4	7	4,1	17	14	12	7,5	25
5	7	3,9	18	15	12	7,9	28
6	7	4,5	19	16	13	8,2	30
7	8	5,3	19	17	13	8	30
8	8	5,3	19	18	13	8,6	31

9	9	5,6	20	19	14	9,5	33
10	10	6,8	21	20	14	9	36

Вариант 14

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	6	3,5	10	11	10	6,3	21
2	6	3,6	12	12	11	6,4	22
3	7	3,9	15	13	11	7	23
4	7	4,1	17	14	12	7,5	25
5	7	4,2	18	15	12	7,9	28
6	8	4,5	19	16	13	8,2	30
7	8	5,3	19	17	13	8,4	31
8	9	5,3	20	18	14	8,6	31
9	9	5,6	20	19	14	9,5	35
10	10	6	21	20	15	10	36

Вариант 15

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,7	9	11	11	6,3	22
2	7	3,7	11	12	11	6,4	22
3	7	3,9	11	13	11	7,2	23
4	7	4,1	15	14	12	7,5	25
5	8	4,2	17	15	12	7,9	27
6	8	4,9	19	16	13	8,1	30
7	8	5,3	19	17	13	8,4	31
8	9	5,1	20	18	13	8,6	32
9	10	5,6	20	19	14	9,5	35
10	10	6,1	21	20	15	9,5	36

Вариант 16

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,5	9	11	10	6,3	22
2	7	3,6	10	12	10	6,5	22
3	7	3,9	12	13	11	7,2	24
4	7	4,1	17	14	12	7,5	25
5	8	4,2	18	15	12	7,9	27
6	8	4,5	19	16	13	8,2	30
7	9	5,3	19	17	13	8,4	31
8	9	5,5	20	18	14	8,6	33

9	10	5,6	21	19	14	9,5	35
10	10	6,1	21	20	15	9,6	36

Вариант 17

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,6	9	11	10	6,3	21
2	7	3,6	11	12	11	6,9	23
3	7	3,7	12	13	11	7,2	24
4	8	4,1	16	14	12	7,8	25
5	8	4,3	19	15	13	8,1	27
6	8	4,5	19	16	13	8,2	29
7	9	5,4	20	17	13	8,4	31
8	9	5,5	20	18	14	8,8	33
9	10	5,8	21	19	14	9,5	35
10	10	6,1	21	20	14	9,7	34

Вариант 18

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,5	9	11	10	6,3	21
2	7	3,6	10	12	10	6,8	22
3	7	3,8	14	13	11	7,2	24
4	7	4,2	15	14	12	7,9	25
5	8	4,3	18	15	12	8,1	26
6	8	4,7	19	16	13	8,3	29
7	9	5,4	19	17	13	8,4	31
8	9	5,6	20	18	13	8,8	32
9	10	5,9	20	19	14	9,6	35
10	10	6,1	21	20	14	9,7	36

Вариант 19

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,8	11	11	10	6,8	21
2	7	3,8	12	12	11	7,4	23
3	7	3,9	16	13	11	7,8	24
4	7	4,1	17	14	12	7,5	26
5	7	4,6	18	15	12	7,9	28
6	8	4,5	18	16	12	8,1	30
7	8	5,3	19	17	13	8,4	31
8	9	5,5	20	18	13	8,7	32

9	9	6,1	20	19	13	9,5	33
10	10	6,8	21	20	14	9,7	35

Вариант 20

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,8	9	11	11	7,1	22
2	7	4,1	14	12	11	7,5	23
3	7	4,3	16	13	12	7,8	25
4	7	4,1	17	14	12	7,6	27
5	8	4,6	17	15	12	7,9	29
6	8	4,7	18	16	13	8,1	30
7	9	5,3	20	17	13	8,5	32
8	9	5,5	20	18	14	8,7	32
9	11	6,9	21	19	14	9,6	33
10	10	6,8	21	20	15	9,8	36

Вариант 21

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,9	12	11	11	7,1	22
2	7	4,2	13	12	12	7,5	25
3	7	4,3	15	13	13	7,8	26
4	7	4,4	17	14	12	7,9	27
5	8	4,6	18	15	13	8,1	30
6	8	4,8	19	16	13	8,4	31
7	9	5,3	19	17	13	8,6	32
8	9	5,7	20	18	14	8,8	32
9	10	6,9	21	19	14	9,6	34
10	10	6,8	21	20	14	9,9	36

Вариант 22

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,6	12	11	10	7,2	23
2	7	4,1	14	12	11	7,6	25
3	7	4,3	16	13	12	7,8	26
4	7	4,4	17	14	11	7,9	28
5	7	4,5	18	15	12	8,2	30
6	8	4,8	19	16	12	8,4	31
7	8	5,3	20	17	12	8,6	32
8	8	5,6	20	18	13	8,8	32

9	9	6,7	21	19	13	9,2	33
10	10	6,9	22	20	14	9,6	34

Вариант 23

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,5	9	11	10	6,3	21
2	7	3,6	10	12	10	6,8	22
3	7	3,8	14	13	11	7,2	24
4	7	4,2	15	14	12	7,9	25
5	8	4,3	18	15	12	8,1	26
6	8	4,7	19	16	13	8,3	29
7	9	5,4	19	17	13	8,4	31
8	9	5,6	20	18	13	8,8	32
9	10	5,9	20	19	14	9,6	35
10	10	6,1	21	20	14	9,7	36

Вариант 24

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,8	11	11	10	6,8	21
2	7	3,8	12	12	11	7,4	23
3	7	3,9	16	13	11	7,8	24
4	7	4,1	17	14	12	7,5	26
5	7	4,6	18	15	12	7,9	28
6	8	4,5	18	16	12	8,1	30
7	8	5,3	19	17	13	8,4	31
8	9	5,5	20	18	13	8,7	32
9	9	6,1	20	19	13	9,5	33
10	10	6,8	21	20	14	9,7	35

Вариант 25

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1	7	3,8	9	11	11	7,1	22
2	7	4,1	14	12	11	7,5	23
3	7	4,3	16	13	12	7,8	25
4	7	4,1	17	14	12	7,6	27
5	8	4,6	17	15	12	7,9	29
6	8	4,7	18	16	13	8,1	30
7	9	5,3	20	17	13	8,5	32
8	9	5,5	20	18	14	8,7	32

9	11	6,9	21	19	14	9,6	33
10	10	6,8	21	20	15	9,8	36

Вариант 26−31

Предполагается, что объем предложения товара y линейно зависит от цены товара X_1 и зарплаты сотрудников X_2. Статистические данные собраны за 14 месяцев.

Требуется:

1. По МНК определить параметры множественной линейной регрессии $y = a + b_1 x_1 + b_2 x_2$.

2. Оценить статистическую значимость найденных эмпирических коэффициентов регрессии b_1, b_2.

3. Сравнить влияние факторов на результат при помощи средних коэффициентов эластичности.

4. Построить 95-% доверительные интервалы для найденных коэффициентов.

5. Вычислить коэффициент детерминации R^2 и оценить его статистическую значимость при α = 0,05.

6. Проверить качество построенного уравнения регрессии с помощью F-статистики Фишера.

7. Оценить целесообразность включения в уравнение одного фактора после другого с помощью частных F-статистик Фишера.

Вариант 26

x_1	64	59	65	71	73	80	36	34	40	44	45	51	60	58
x_2	44	42	49	46	37	41	45	38	42	47	43	44	36	34
y	37	35	34	31	30	28	27	25	21	17	45	42	40	38

Вариант 27

x_1	46	39	43	48	44	45	37	35	45	43	50	47	38	42
x_2	44	42	49	46	37	41	45	38	42	47	43	44	36	34
y	31	33	34	36	37	40	41	43	44	46	48	51	23	27

Вариант 28

x_1	63	57	51	56	50	52	43	37	36	32	26	28	72	65

x_2	40	35	31	38	34	30	39	42	35	37	27	29	37	36
y	39	37	36	34	33	30	29	27	26	24	20	16	44	41

Вариант 29

x_1	39	43	44	50	59	57	63	58	64	70	72	79	35	33
x_2	46	39	43	48	44	45	37	35	45	43	50	47	38	42
y	42	44	45	47	49	52	24	28	32	34	35	37	38	41

Вариант 30

x1	46	52	61	59	65	60	66	72	74	81	37	35	41	45
x2	51	48	39	43	47	40	44	49	45	46	38	36	46	44
y	42	44	45	47	49	52	24	28	32	34	35	37	38	41

Вариант 31

x_1	62	30	36	50	41	47	56	54	60	55	61	67	69	66
x_2	43	51	41	39	46	43	34	38	42	25	39	44	40	41
y	45	46	40	39	35	29	61	75	68	66	60	54	59	53

Литература

1. Берндт Э.Р. Практика эконометрики: классика и современность: учебник для студентов вузов. – М.: ЮНИТИ-ДАНА, 2005. – 863 с.
2. Бородич С.А. Эконометрика: учеб. пособие / С.А. Бородич. – 2-е изд., испр. – Мн.: Новое знание, 2004. – 416 с. (Экономическое образование).
3. Введение в эконометрику: учеб. пособие / Л.П. Яновский, А.Г. Буховец. – М.: КНОРУС, 2009. – 256 с.
4. Доугерти К. Введение в эконометрику: пер. с англ. – М.: ИНФРА-М, 1999. – 402 с.
5. Луговская Л.В. Эконометрика в вопросах и ответах: учеб. пособие. – М.: ТК Велби: Проспект, 2006. – 208 с.
6. Кремер Н.Ш. Эконометрика: Учебник для вузов / Н.Ш. Кремер, Б.А. Бутко; под ред. Н.Ш. Кремера. – М.: ЮНИТИ-ДАНА, 2002. – 311 с.
7. П.К. Катышев. Сборник задач к начальному курсу эконометрики / Я.Р. Магнус, А.А. Пересецкий. – М.: Дело, 2002. – 208 с.

8. Я.Р. Магнус. Эконометрика. Начальный курс: учебник / П.К. Катышев, А.А. Пересецкий. – М.: Дело, 2001. – 400 с.

9. Поленова Т.М. Эконометрика: метод. рекомендации к практ. занятиям по курсу. – М.: Изд-во РАГС, 2009. – 64 с.

10. Практикум по эконометрике: учеб. пособие / под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2006. – 344 с.

11. Прикладная статистика. Основы эконометрики: учебник для вузов: В 2 т. – Т. 1. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. – М.: ЮНИТИ-ДАНА, 2001. – 656 с.

12. Практикум по эконометрике: регрессионный анализ средствами Excel / А.И. Приходько. – Ростов н/Д.: Феникс, 2007. – 256 с.

13. Прикладная статистика. Основы эконометрики: учебник для вузов: В 2 т. – Т. 2. Айвазян С.А. Основы эконометрики. – М.: ЮНИТИ-ДАНА, 2001. – 432 с.

14. Сборник задач по эконометрике: учеб. пособие для студентов экон. вузов / сост. Е.Ю. Дорохина, Л.Ф. Преснякова, Н.П. Тихомиров. - М.: Экзамен, 2003. - 224 с.

15. Эконометрика: учебник / под ред. В.С. Мхитаряна. – М.: Проспект, 2008. – 384 с.

16. Эконометрика: учебник / под ред. И.И. Елисеевой. – М.: Финансы и статистика, 2006. – 576 с.

17. Эконометрика: учебник / Н.П. Тихомиров, Е.Ю. Дорохина. – М.: Экзамен, 2003. – 512 с. (Программа курса, Учебник: ч 1 - 4.)

18. Эконометрика: учеб. пособие / А.В. Гладилин, А.Н. Герасимов, Е.И. Громов. – М.: КНОРУС, 2008. – 232 с.

19. Эконометрика: учеб. пособие для вузов / А.И. Орлов – М.: Экзамен, 2002. – 576 с.

20. Эконометрика: учеб. пособие в схемах и табл. / под ред. С.А. Орехова. – М.: Эксмо, 2008. – 224 с.

Отпечатано в Издательско-полиграфическом центре
Набережночелнинского института
Казанского (Приволжского) федерального университета

Подписано в печать 13.052016
Формат 60х84/16. Печать ризографическая
Бумага офсетная. Гарнитура «Nimes New Roman»
Усл. п. л. 5,8. Уч-изд. л. 5,8
Тираж 50 экз. Заказ №738

423810, г. Набережные Челны, Новый город, проспект Мира, 68/19
Тел./факс (8552) 39-69-99 e-mail: ic-nchi-rpfu@mail.ru