Федеральное агентство по образованию
Государственное образовательное учреждение
высшего профессионального образования
Ивановский государственный химико-технологический университет

МНОГОМЕРНЫЕ СТАТИСТИЧЕСКИЕ МЕТОДЫ
РЕГРЕССИОННЫЙ АНАЛИЗ

Методические указания

Составитель Т.А. Баранова

Иваново 2007

Составитель Т.А. Баранова

УДК 519.237.5(072)

Многомерные статистические методы. Регрессионный анализ: Метод. указания / Иван. гос. хим.-технол. ун-т.; Сост. Т.А. Баранова. – Иваново, 2007. – 40с.

Методические указания разработаны в соответствии с Государственным образовательным стандартом и предназначены для студентов ИГХТУ специальности 061800 ″Математические методы в экономике″.

Приведены основные теоретические положения о регрессионном анализе, примеры решения задач и практические задания, предназначенные для самостоятельной работы студентов.

Таблиц 22. Рисунков 3. Библиогр.: 7 назв.

Рецензент
заведующая кафедрой управления и экономико–математического моделирования, доктор экономических наук, профессор А.Н. Ильченко (Ивановский государственный химико–технологический университет)

ВВЕДЕНИЕ

С тех пор, как экономика стала серьезной самостоятельной наукой, исследователи пытаются дать свое представление о возможных путях экономического развития, спрогнозировать ту или иную ситуацию, предвидеть будущие значения экономических показателей, указать инструменты изменения ситуации в желательном направлении. С другой стороны, во многих случаях различные экономисты предлагают разные, а зачастую противоположные методы решения той или иной задачи. Политики, либо управляющие производством, выбирая одну из возможных стратегий решения, получают определенный результат. Плох он или хорош, и можно ли было получить лучший результат, проверить весьма затруднительно. Экономическая ситуация практически никогда не повторяется в точности, следовательно, нет возможности применить две стратегии при одних и тех условиях с целью сравнения конечного результата. Поэтому одной из центральных задач экономического анализа является предсказание, либо прогнозирование развития некоторого экономического объекта при создании тех или иных условий.

Поведение и значение любого экономического показателя зависит практически от бесконечного количества факторов, и все учесть нереально. Обычно лишь ограниченное количество факторов действительно существенно воздействуют на исследуемый экономический показатель. Доля влияния остальных факторов столь незначительна, что их игнорирование не может привести к существенным отклонениям в поведении исследуемого объекта. Выделение и учет в модели лишь ограниченного числа реально доминирующих факторов и является серьезной предпосылкой для качественного анализа, прогнозирования и управления ситуацией.

Экономическая теория выявила и исследовала значительное число устоявшихся и стабильных связей между различными показателями. Например, хорошо изученными являются зависимости спроса или потребления от уровня дохода и цен на товары; зависимость между уровнями безработицы и инфляции; зависимость объема производства от целого ряда факторов (размера основных фондов, их возраста, качества персонала и т.д.); зависимость между производительностью труда и уровнем механизации, а также многие другие зависимости.

Любая экономическая политика заключается в регулировании экономических переменных, и она должна базироваться на знании того, как эти переменные связаны с другими переменными, ключевыми для принимающего решения политика или предпринимателя. Однако, в реальных ситуациях даже устоявшиеся зависимости могут проявляться по–разному.

Можно указать два варианта рассмотрения взаимосвязей между двумя переменными X и Y. В первом случае обе переменные считаются равноценными в том смысле, что они не подразделяются на первичную и вторичную

(независимую и зависимую) переменные. Основным в этом случае является вопрос о наличии и силе взаимосвязи между этими переменными. Например, между ценой товара и объемом спроса на него, между урожаем картофеля и урожаем зерна, между интенсивностью движения транспорта и числом аварий. При исследовании силы линейной зависимости меду такими переменными обращаются к корреляционному анализу, основной мерой которого является коэффициент корреляции. Вполне вероятно, что связь в этом случае не носит направленного характера. Например, урожайность картофеля и зерновых обычно изменяется в одном и том же направлении, однако очевидно, что ни одна из этих переменных не является определяющей.

Другой вариант рассмотрения взаимосвязей выделяет одну из величин как независимую (объясняющую), а другую как зависимую (объясняемую). В этом случае изменение первой из них может служить причиной изменения другой. Например, рост дохода ведет к увеличению потребления; рост цены – к снижению спроса; снижение процентной ставки увеличивает инвестиции; увеличение обменного курса валюты сокращает объем чистого экспорта и т.д. Однако такая зависимость не является однозначной в том смысле, что каждому конкретному значению объясняющей переменной (набору объясняющих переменных) может соответствовать не одно, а множество значений из некоторой области. Другими словами, каждому конкретному значению объясняющей переменной (набору объясняющих переменных) соответствует некоторое вероятностное распределение зависимой переменной (рассматриваемой как случайная величина (далее СВ)). Поэтому анализируют, как объясняющая(ие) переменная(ые) влияет(ют) на зависимую переменную «в среднем». Зависимость такого типа называется функцией регрессии Y на X и выражается соотношением

$$M(Y/x) = f(x), \tag{1}$$

При этом X называется независимой (объясняющей) переменной (регрессором), Y – зависимой (объясняемой) переменной. При рассмотрении зависимости двух СВ говорят о парной регрессии.

Зависимость нескольких переменных, выражаемая функцией

$$M(Y/x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m), \tag{2}$$

называют множественной регрессией.

Термин «регрессия» (лат. «*regression*» – движение назад, возвращение в прежнее состояние) был введен английским психологом и антропологом Фрэнсисом Гальтоном в конце XIX века при анализе зависимости между ростом родителей и ростом детей. Исследователь заметил, что рост детей у очень высоких родителей в среднем меньше, чем средний рост родителей. У очень низких родителей, наоборот, средний рост детей выше. И в том, и в другом случае средний рост детей стремится (возвращается) к среднему росту людей в данном регионе. Выявленная тенденция была названа «регрессией к среднему состоянию».

В настоящее время под регрессией понимается функциональная зависимость между объясняющими переменными и условным математическим ожиданием (средним значением) зависимой переменой, которая строится с целью предсказания (прогнозирования) этого среднего значения при фиксированных значениях первых. Для отражения того факта, что реальные значения зависимой переменной не всегда совпадают с ее условными математическими ожиданиями и могут быть различными при одном и том же значении объясняющей переменной (наборе объясняющих переменных), фактическая зависимость должна быть дополнена некоторым слагаемым ε, которое, по существу, является СВ и указывает на стохастическую суть зависимости. Из этого следует, что связи между зависимой и объясняющей(ими) переменными выражаются соотношениями, называемыми регрессионными моделями (уравнениями)

$$Y = M(Y / x) + \varepsilon, \tag{3}$$

$$Y = M(Y / x_1, x_2, \ldots, x_m) + \varepsilon, \tag{4}$$

Возникает вопрос о причинах обязательного присутствия в регрессионных моделях случайного фактора (отклонения). Среди таких причин наиболее существенными являются:

1. *Невключение в модель всех объясняющих переменных.*

Любая регрессионная модель является упрощением реальной ситуации. Последняя всегда представляет собой сложнейшее переплетение различных факторов, многие из которых в модели не учитываются, что порождает отклонение реальных значений зависимой переменой от ее модельных значений.

2. *Неправильный выбор функциональной формы модели.*

Из–за слабой изученности исследуемого процесса или его переменчивости может быть неверно подобрана функция, его моделирующая. Это, безусловно, скажется на отклонении модели от реальности, что отразится на величине случайного члена.

3. *Агрегирование переменных.*

Во многих моделях рассматриваются зависимости между факторами, которые сами представляют сложную комбинацию других, более простых переменных.

4. *Ошибки измерений.*

Какой бы качественной ни была модель, ошибки измерений переменных отразятся на несоответствии модельных значений эмпирическим данным, что также отразится на величине случайного члена.

5. *Ограниченность статистических данных.*

Зачастую строятся модели, выражаемые непрерывными функциями, но для этого используется набор данных, имеющих дискретную структуру. Это несоответствие находит свое выражение в случайном отклонении.

6. *Непредсказуемость человеческого фактора.*

Эта причина может «испортить» самую качественную модель. Действительно, при правильном выборе формы модели, тщательном подборе объясняющих

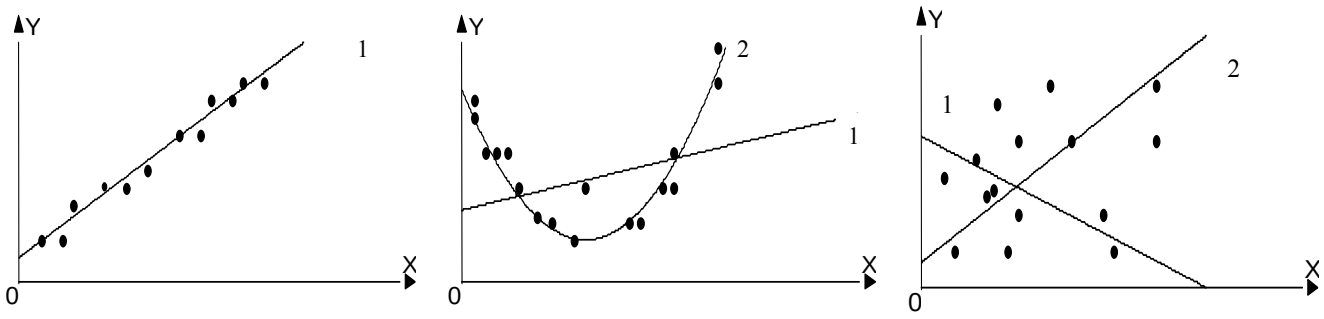
переменных все равно невозможно спрогнозировать поведение каждого индивидуума.

Таким образом, случайный член является отражением влияния всех описанных выше причин и не только их. Этот список может быть дополнен.

Основными этапами регрессионного анализа являются:

1. Выбор вида уравнения регрессии (***спецификация модели***).
2. Выбор независимых переменных, оказывающих существенное влияние на зависимую переменную.
3. Оценка параметров уравнения регрессии (***параметризация модели***).
4. Измерение влияния отдельных факторов на зависимую переменную.
5. Оценка статистической надежности регрессионной модели (***верификация модели***).

Выбор формулы связи переменных называется *спецификацией* уравнения регрессии. В случае парной регрессии выбор формулы обычно осуществляется по графическому изображению реальных статистических данных в виде точек в декартовой системе координат, которая называется *корреляционным полем* (*диаграммой рассеивания*).



а) б) в)

Рис. 1. Типы корреляции

На рис. 1 представлены три ситуации: а) взаимосвязь между X и Y близка к линейной, и прямая 1 достаточно хорошо соответствует эмпирическим точкам. Поэтому в данном случае в качестве зависимости между X и Y целесообразно выбрать линейную функцию $y = b_0 + b_1 x$; б) реальная взаимосвязь между X и Y, скорее всего описывается квадратичной функцией $y = ax^2 + bx + c$ (линия 2). И какую бы мы ни провели прямую (например, линия 1), отклонения точек наблюдений от нее будут существенными и неслучайными; в) явная зависимость между X и Y отсутствует. В данном случае какую бы мы ни выбрали форму связи между переменными, результаты спецификации и *параметризации* (определение коэффициентов уравнения регрессии) модели будут неудачными. В случае множественной регрессии определение подходящего вида зависимости является более сложной задачей.

Основные типы кривых, используемых при количественной оценке связей, представлены на рис. 2.

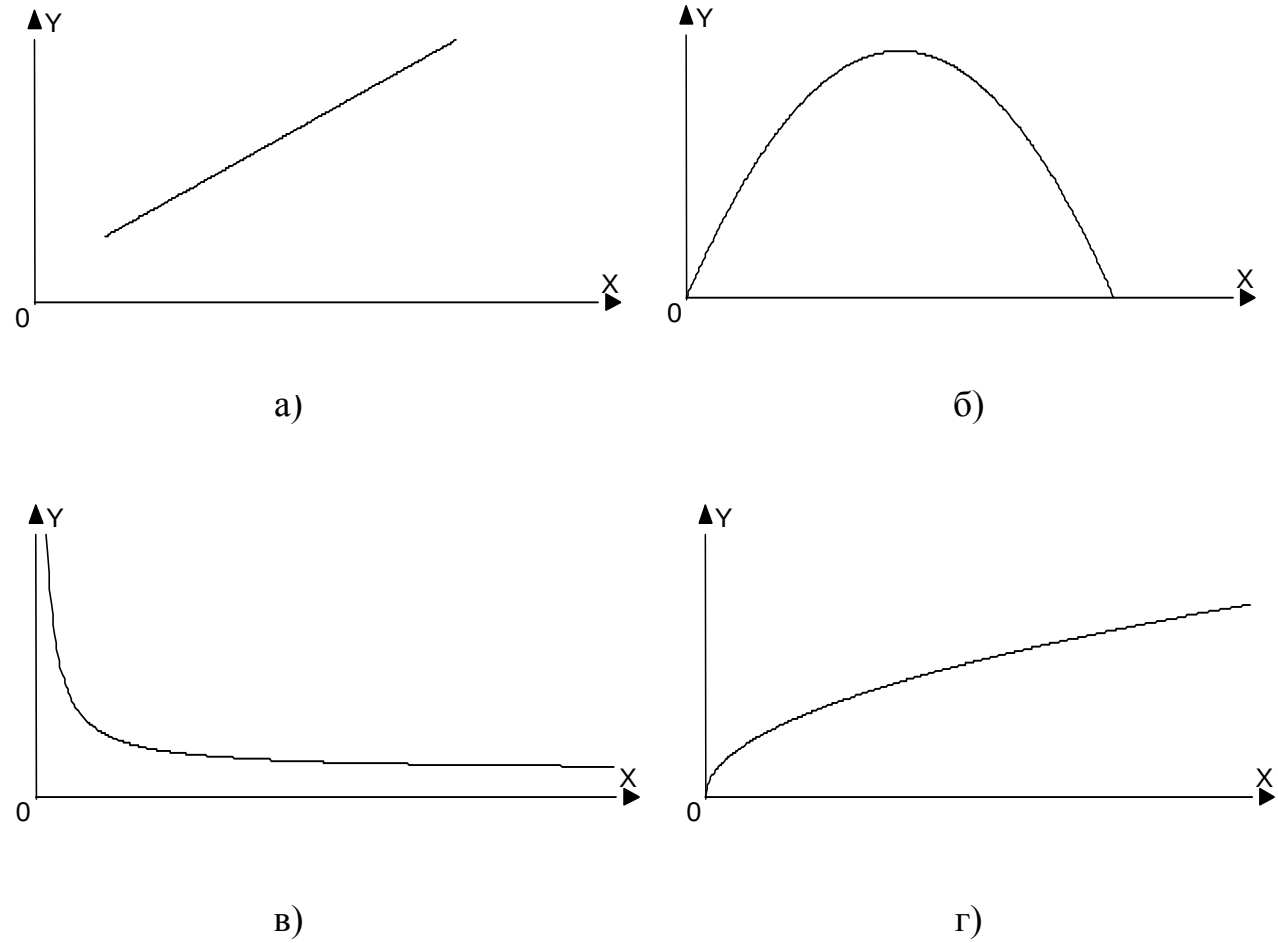


Рис. 2 Основные типы кривых, используемые при количественной оценке связей между двумя переменными: а) $\hat{y}_x = b_0 + b_1 x$; б) $\hat{y}_x = b_0 + b_1 x + b_2 x^2$; в) $\hat{y}_x = b_0 + \frac{b_1}{x}$; г) $\hat{y}_x = b_0 \cdot x^{b_1}$

ДВУМЕРНАЯ ЛИНЕЙНАЯ РЕГРЕССИОННАЯ МОДЕЛЬ

После того, как выбрана форма зависимости между переменными X и Y, начинается следующий этап регрессионного анализа – *параметризация модели.* Задача состоит в том, чтобы по конкретной выборке (x_i, y_i) найти оценки b_0 и b_1 неизвестных параметров β_0 и β_1, так, чтобы построенная линия регрессии являлась бы наилучшей в определенном смысле среди всех других прямых.

Самым распространенным и теоретически обоснованным является метод нахождения коэффициентов уравнения регрессии b_0 и b_1, называемый методом наименьших квадратов (МНК). Этот метод оценки b_0 и b_1 является наиболее простым с вычислительной точки зрения. Кроме того, оценки коэффициентов регрессии, найденные по МНК при определенных предпосылках, обладают рядом оптимальных свойств. Среди других методов определения оценок b_0 и b_1 коэффициентов регрессии применяются метод моментов (ММ) и метод максимального правдоподобия.

Условия Гаусса–Маркова

Использование метода наименьших квадратов для оценки параметров собственно линейных регрессионных моделей дает хорошее приближение оценок к истинным значениям параметров. Доказано, что для получения по МНК наилучших результатов необходимо, чтобы выполнялся ряд предпосылок относительно случайного отклонения. Предпосылки МНК называют условиями Гаусса–Маркова.

1. *Математическое ожидание случайного отклонения ε_i равно нулю:* $M(\varepsilon_i)=0$ *для всех наблюдений.* Данное условие означает, что случайное отклонение в среднем не оказывает влияния на зависимую переменную. Выполнимость $M(\varepsilon_i)=0$ влечет выполнимость $M(Y/X=x_i)=\beta_0+\beta_1 x_i$.
2. *Дисперсия случайных отклонений ε_i постоянна* $M(\varepsilon_i^2)=\sigma^2$.Выполнимость данной предпосылки называется гомоскедастичность (постоянством дисперсии отклонений), невыполнимость данной предпосылки называется гетероскедастичностью (непостоянством дисперсии отклонений).
3. *Случайные отклонения ε_i и ε_j являются независимыми друг от друга для $i\neq j$.* Выполнимость данной предпосылки предполагает, что отсутствует систематическая связь между любыми случайными отклонениями. В данном случае говорят об отсутствии автокорреляции.
4. *Случайное отклонений должно быть независимо от объясняющих переменных.* Обычно это условие выполняется автоматически, если объясняющие переменные не являются случайными в данной модели.
5. *Модель является линейной относительно параметров.*

В англоязычной литературе оценки, полученные при выполнении условий Гаусса–Маркова, называют BLUE (Best Linear Unbiased Estimators) – наилучшие линейные несмещенные оценки.

Нахождение оценок β_0 и β_1 уравнения в случае парной линейной регрессии

Пусть по выборке (x_i,y_i), i=1,2,…,n, требуется определить оценки b_0 и b_1 эмпирического уравнения регрессии

$$\hat{y}_i = b_0 + b_1 x_i, \qquad (5)$$

В этом случае при использовании МНК минимизируется следующая функция (рис. 3)

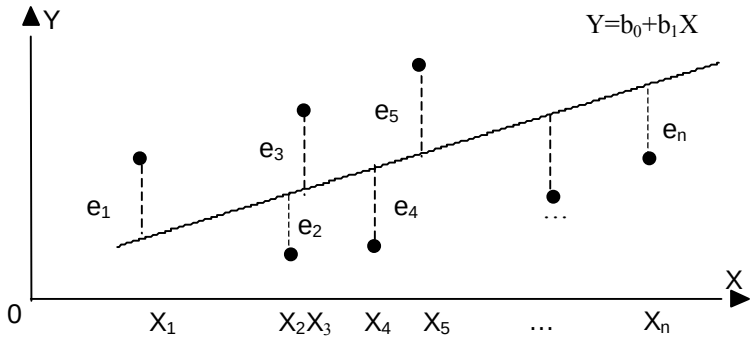


Рис. 3 Линия регрессии с минимальной дисперсией остатков.

$$Q(b_0, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2 , \qquad (6)$$

Нетрудно заметить, что функция Q является квадратичной функцией двух параметров b_0 и b_1 ($Q = Q(b_0, b_1)$), поскольку x_i,y_i, (i=1,2,…,n) – известные данные наблюдений. Так как функция Q непрерывна, выпукла и ограничена снизу ($Q \geq 0$), то она имеет минимум.

Необходимым условием существования минимума функции двух переменных (6) является равенство нулю ее частных производных по неизвестным параметрам b_0 и b_1:

$$\begin{cases} \dfrac{\partial Q}{\partial b_0} = -2\sum (y_i - b_0 - b_1 x_i) = 0 ; \\ \\ \dfrac{\partial Q}{\partial b_1} = -2\sum (y_i - b_0 - b_1 x_i) x_i = 0 . \end{cases} \Rightarrow \qquad (7)$$

Данная система является системой нормальных уравнений. Далее ее необходимо решить относительно параметров b_0 и b_1, которые являются в данном случае переменными

$$\begin{cases} nb_0 + b_1 \sum x_i = \sum y_i ; \\ b_0 \sum x_i + b_1 \sum x_i^2 = \sum x_i y_i . \end{cases} \qquad (8)$$

Разделив оба уравнения системы (8) на n, получим:

$$\begin{cases} b_0 + b_1 \bar{x} = \bar{y}; \\ \\ b_0 \bar{x} + b_1 \overline{x^2} = \overline{xy}. \end{cases} \Rightarrow \qquad \begin{cases} b_1 = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - \bar{x}^2}; \\ b_0 = \bar{y} - b_1 \bar{x}. \end{cases} \qquad (9)$$

Здесь $\bar{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i , \quad \overline{x^2} = \dfrac{1}{n}\sum_{i=1}^{n} x_i^2 , \quad \bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i ,$ (10)

Таким образом, по МНК оценки параметров b_0 и b_1 определяются по формулам (9).

Нетрудно заметить, что b_1 можно вычислить по формуле:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_x^2}, \tag{11}$$

Тогда

$$b_1 = \frac{S_{xy}}{S_x^2} = \frac{S_{xy}}{S_x \cdot S_y} \cdot \frac{S_y}{S_x} = r_{xy} \cdot \frac{S_y}{S_x}, \tag{12}$$

где r_{xy} – выборочный коэффициент корреляции; S_x, S_y – стандартные отклонения. Таким образом, коэффициент регрессии пропорционален ковариации и коэффициенту корреляции, а коэффициенты пропорциональности служат для соизмерения перечисленных разномерных величин.

Итак, если коэффициент корреляции уже рассчитан, то легко может быть найден коэффициент b_1 парной регрессии по формуле (12).

Если, кроме уравнения регрессии Y на X ($\hat{Y} = b_0 + b_x X$), для тех же эмпирических данных найдено уравнение регрессии X на Y ($\hat{X} = c_0 + b_y Y$), то произведение коэффициентов b_x и b_y равно r_{xy}^2:

$$b_x \cdot b_y = r_{yx} \frac{S_y}{S_x} \cdot r_{xy} \frac{S_x}{S_y} = r_{xy}^2, \tag{13}$$

Коэффициенты c_0 и b_y находятся по формулам, аналогичным формулам (9):

$$\begin{cases} b_y = \dfrac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{y^2} - \bar{y}^2}; \\ c_0 = \bar{x} - b_y \bar{y} \end{cases} \tag{14}$$

Матричная форма записи парного линейного уравнения регрессии

Парное линейное уравнение регрессии может быть записано в матричной форме:

$$Y = BX + E, \tag{15}$$

где Y – вектор–столбец размерности (n×1) фактических значений результативного признака; B – вектор–столбец размерности (2×1) подлежащих оценке параметров модели, т.е. коэффициента регрессии «b_1» и свободного члена (параметра «b_0» в уравнении (5)); X=(x_0,x_1) – матрица размерности (n×2) значений факторов. При этом x_0=1 и связано с наличием в уравнении регрессии свободного члена, а x_1 – собственно реальные значения включенного в уравнение регрессии фактора; E – вектор–столбец случайной величины ε_i размерности (n×1).

Матрица исходных данных имеет вид

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \qquad X = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \tag{16}$$

Оценивая параметры линейного уравнения регрессии, найдем вектор B и далее вектор случайной компоненты E, т.е.

$$B = \begin{pmatrix} b_0 \\ b_1 \end{pmatrix}, \qquad E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}, \tag{17}$$

В матричной форме применение МНК записывается так:

$$S = (Y - XB)^T (Y - XB) \to \min, \tag{18}$$

Дифференцируя S по вектору B и приравнивая первые частные производные по B к нулю, получим:

$$\frac{\partial S}{\partial B} = -2X^T Y + 2X^T XB = 0, \tag{19}$$

Отсюда после перестановки членов получаем $(X^T X)B = X^T Y$. Соответственно оценка вектора B составит

$$B = (X^T X)^{-1} X^T Y, \tag{20}$$

Можно показать, что в случае нормального закона распределения случайной величины y оценки метода наименьших квадратов и метода наибольшего правдоподобия совпадают.

Проверка качества уравнения регрессии: проверка гипотез относительно коэффициентов линейного уравнения регрессии

Эмпирическое уравнение регрессии определяется на основе конечного числа статистических данных. Поэтому коэффициенты эмпирического уравнения регрессии являются случайными величинами (СВ), изменяющимися от выборки к выборке. При проведении статистического анализа перед исследователем зачастую возникает необходимость сравнения эмпирических коэффициентов регрессии b_0 и b_1 с некоторыми теоретически ожидаемыми значениями β_0 и β_1 этих коэффициентов. Данный анализ осуществляется по схеме статистической проверки гипотез.

Для проверки гипотезы $H_0 : b_1 = \beta_1$ против $H_1 : b_1 \neq \beta_1$ используется статистика

$$t = \frac{b_1 - \beta_1}{S_{b_1}}, \tag{21}$$

которая при справедливости нулевой гипотезы имеет распределение Стьюдента с числом степеней свободы $\nu = n - 2$, где n – объем выборки; $S_{b_1} = \sqrt{S_{b_1}^2}$ – стандартная ошибка коэффициента регрессии b_1. Для расчета $S_{b_1}^2$ применяется следующая формула

$$S_{b_1}^2 = \frac{S^2}{n \cdot (\overline{x^2} - \overline{x}^2)} = \frac{\sum e_i^2}{n \cdot (n-2)(\overline{x^2} - \overline{x}^2)} = \frac{\sum (y_i - b_0 - b_1 x_i)^2}{n \cdot (n-2)(\overline{x^2} - \overline{x}^2)}, \qquad (22)$$

где $S^2 = \frac{\sum e_i^2}{n-2}$ – необъясненная дисперсия (мера разброса зависимой переменной вокруг линии регрессии), $S = \sqrt{\frac{\sum e_i^2}{n-2}}$ – стандартная ошибка оценки (стандартная ошибка регрессии).

Таким образом, $H_0 : b_1 = \beta_1$ может быть отклонена на основании того, что

$$\left| T_{íàáë} \right| = \left| \frac{b_1 - \beta_1}{S_{b_1}} \right| \geq t_{\frac{\alpha}{2}, n-2} \qquad (23)$$

где α – требуемый уровень значимости. При невыполнении (23) считается, что нет оснований для отклонения H_0.

Наиболее важной на начальном этапе статистического анализа построенной модели все же является задача установления наличия линейной зависимости между Y и X. Эта проблема может быть решена по следующей схеме: $H_0 : b_1 = 0, H_1 : b_1 \neq 0$. Гипотеза в такой постановке называется *гипотезой о статистической значимости коэффициента регрессии.* При этом, если нулевая гипотеза принимается, то есть основания считать, что между величинами Y и X нет линейной зависимости. В этом случае говорят, что коэффициент b_1 *статистически незначим* (он слишком близок к нулю). При отклонении нулевой гипотезы коэффициент b_1 считается *статистически значимым*, что указывает на наличие определенной линейной зависимости между величинами Y и X. В данном случае рассматривается двусторонняя критическая область, так как важным является именно отличие от нуля коэффициента регрессии, и он может быть как положительным, так и отрицательным.

С помощью t–статистики для наблюдаемого значения критерия

$$t_{íàáë} = \frac{b_1}{S_{b_1}} = \frac{b_1}{\sqrt{S_{b_1}^2}} \qquad (24)$$

и таблицы критических точек распределения Стьюдента на основании $|t_{íàáë}| \le t_{êð}$ делается окончательный вывод о справедливости (несправедливости) нулевой гипотезы.

Аналогично проверяется статистическая значимость коэффициента b_0. При выполнимости нулевой гипотезы считается, что коэффициент b_0 является статистически незначимым и в данном случае свободным членом уравнения регрессии можно пренебречь, рассматривая регрессию как $\hat{y}_x = b_1 x$.

При оценке значимости коэффициентов b_0 и b_1 линейной регрессии на начальном этапе можно использовать следующее «*грубое правило*», позволяющее не прибегать к таблицам:

а) если стандартная ошибка коэффициента больше его модуля ($|t| \le 1$), то коэффициент не может быть признан значимым, так как доверительная вероятность при двусторонней альтернативной гипотезе составит менее чем 0,7;

б) если $1 < |t| \le 2$, то найденная оценка может рассматриваться как относительно (слабо) значимая. Доверительная вероятность в этом случае лежит между значениями 0,7 и 0,95;

в) если $2 < |t| \le 3$, это свидетельствует о значимой линейной связи между величинами *Y* и *X*. В этом случае доверительная вероятность колеблется от 0,95 до 0,99;

г) если $|t| > 3$, то это почти гарантия наличия линейной связи.

Конечно, в каждом конкретном случае имеет значение число наблюдений. Чем их больше, тем надежнее при прочих равных условиях выводы о значимости коэффициента. Однако, для $n > 10$ предложенное «*грубое правило*» практически всегда работает.

Интервальные оценки коэффициентов линейного уравнения регрессии

Как отмечалось ранее, для построения уравнения регрессии по МНК используется предположение о нормальном распределении отклонений ε_i с нулевым математическим ожиданием и постоянной дисперсией, т.е. $\varepsilon_i \sim N(0, \sigma^2)$. Естественность этого предположения обосновывается хорошо известной в теории вероятностей *центральной предельной теоремой* (ЦПТ): если случайная величина представляет собой сумму очень большого числа независимых случайных величин, влияние каждой из которых на всю сумму ничтожно мало, то рассматриваемая случайная величина имеет распределение, близкое к нормальному.

Но случайное отклонение ε_i как раз и отражает влияние на независимую величину тех переменных, которые не включены в модель. Таких переменных

обычно очень много, причем их индивидуальное влияние достаточно мало (иначе их необходимо было учесть в модели). Следовательно, при рассмотрении случайных отклонений практически выполняются условия ЦПТ. Тогда можно заключить, что ε_i=1, 2, …, n, имеют нормальное распределение с $M(\varepsilon_i)=0, \sigma^2(\varepsilon_i)=\sigma^2$. Это позволяет получать не только наилучшие линейные несмещенные точечные оценки b_0 и b_1 линейного уравнения регрессии, но и находить их интервальные оценки, что дает определенные гарантии точности.

Как указывалось выше, статистики

$$t_{b_0}=\frac{b_0-\beta_0}{S_{b_0}}, \quad t_{b_1}=\frac{b_1-\beta_1}{S_{b_1}}, \tag{25}$$

имеют распределение Стьюдента с числом степеней свободы $\nu=n-2$. Далее для определения 100(1–α)%–го доверительного интервала с помощью таблиц критических точек распределения Стьюдента по доверительной вероятности $\gamma=1-\alpha$ и числу степеней свободы ν определяют критическое значение $t_{\frac{\alpha}{2},n-2}$, удовлетворяющее условию

$$P\left(|t|<t_{\frac{\alpha}{2},n-2}\right)=1-\alpha, \tag{26}$$

Подставив каждую из формул (25) в (26), получаем

$$P\left(-t_{\frac{\alpha}{2},n-2}<\frac{b_0-\beta_0}{S_{b_0}}<t_{\frac{\alpha}{2},n-2}\right)=1-\alpha$$

$$P\left(-t_{\frac{\alpha}{2},n-2}<\frac{b_1-\beta_1}{S_{b_1}}<t_{\frac{\alpha}{2},n-2}\right)=1-\alpha \tag{27}$$

После преобразований выражений, стоящих в скобках, имеем

$$P\left(b_0-t_{\frac{\alpha}{2},n-2}S_{b_0}<\beta_0<b_0+t_{\frac{\alpha}{2},n-2}S_{b_0}\right)=1-\alpha, \tag{28}$$

$$P\left(b_1-t_{\frac{\alpha}{2},n-2}S_{b_1}<\beta_1<b_1+t_{\frac{\alpha}{2},n-2}S_{b_1}\right)=1-\alpha, \tag{29}$$

Таким образом, соотношения (28), (29) определяют доверительные интервалы

$$b_0-t_{\frac{\alpha}{2},n-2}S_{b_0}<\beta_0<b_0+t_{\frac{\alpha}{2},n-2}S_{b_0} \tag{30}$$

$$b_1 - t_{\frac{\alpha}{2},n-2} S_{b_1} < \beta_1 < b_1 + t_{\frac{\alpha}{2},n-2} S_{b_1} \quad (31)$$

которые с надежностью $(1-\alpha)$ накрывают определяемые параметры β_0 и β_1. Фактически доверительный интервал определяет значения теоретических коэффициентов регрессии β_0 и β_1, которые будут приемлемыми с надежностью $(1-\alpha)$ при найденных оценках b_0 и b_1.

Доверительные интервалы для зависимой переменной

В данном случае возможен двоякий подход: *во–первых*, можно предсказать условное математическое ожидание зависимой переменной при определенных значениях объясняющих переменных (*предсказание среднего значения*), *во–вторых*, можно прогнозировать некоторое конкретное значение зависимой переменной (*предсказание конкретного значения*).

Предсказание среднего значения. Пусть построено уравнение парной регрессии $\hat{y}_i = b_0 + b_1 x_i$, на основе которого необходимо предсказать условное математическое ожидание $M(Y/X = x_p)$ переменной Y при X=x_p. В данном случае значение $\hat{y}_p = b_0 + b_1 x_p$ является оценкой $M(Y/X = x_p)$. Тогда естественным является вопрос, как сильно может уклониться модельное среднее значение $\hat{y}_p$, рассчитанное по эмпирическому уравнению регрессии, от соответствующего условного математического ожидания. Ответ на этот вопрос дается на основе интервальных оценок, построенных с заданной надежностью $(1-\alpha)$ при любом конкретном значении x_p объясняющей переменной.

Для построения доверительного интервала необходимо показать, что случайная величина $\hat{Y}_p$ имеет нормальное распределение с конкретными параметрами.

Доверительный интервал для $M(Y/X = x_p) = \beta_0 + \beta_1 x_p$ имеет вид

$$\left(b_0 + b_1 x_p - t_{\frac{\alpha}{2},n-2} S \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}}; b_0 + b_1 x_p + t_{\frac{\alpha}{2},n-2} S \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}} \right) \quad (32)$$

Для проверки гипотезы $H_0 : M(Y/X = x_p) = y_p$ против $H_1 : M(Y/X = x_p) \neq y_p$ используется статистика

$$T = \frac{M(Y/X = x_p) - y_p}{S\sqrt{\frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}}}, \qquad (33)$$

имеющая распределение Стьюдента с числом степеней свободы $\nu = n - 2$. Поэтому нулевая гипотеза отклоняется, если $\left|T_{íàáë}\right| \geq t_{\frac{\alpha}{2},n-2}$ (α – требуемый уровень значимости).

Предсказание индивидуальных значений зависимой переменной. На практике иногда более важно знать дисперсию *Y*, чем ее средние значения или доверительные интервалы для условных математических ожиданий. Это позволяет определить допустимые границы для конкретного значения *Y*.

Доверительный интервал

$$\left(b_0 + b_1 x_p \pm t_{\frac{\alpha}{2},n-2} S\sqrt{1 + \frac{1}{n} + \frac{(\bar{x} - x_p)^2}{\sum (x_i - \bar{x})^2}} \right), \qquad (34)$$

определяет границы, за пределами которых могут оказаться не более 100α % точек наблюдений при $X = x_p$. Необходимо отметить, что данный интервал шире доверительного интервала для условного математического ожидания (32).

Проверка общего качества уравнения регрессии. Коэффициент детерминации R^2

Суммарной мерой общего качества уравнения регрессии (соответствия уравнения регрессии статистическим данным) является коэффициент детерминации R^2. В случае парной линейной регрессии коэффициент детерминации будет совпадать с квадратом коэффициента корреляции. В общем случае коэффициент детерминации рассчитывается по формуле

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2}, \qquad (35)$$

Оценка значимости уравнения регрессии в целом дается с помощью *F*–критерия Фишера. Непосредственному расчету *F*–критерия предшествует анализ дисперсии. Общая сумма квадратов отклонений переменной *y* от среднего значения $\bar{y}$ может быть разложена на две части – «объясненную» и «необъясненную» (остаточную)

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

Общая суммаквадратов отклонений = Сумма квадратов отклонений, объясненная регрессией + Остаточная сумма квадратов отклонений (36)

Любая сумма квадратов отклонений связана с числом степеней свободы *df*, т. е. с числом свободы независимого варьирования признака. Число степеней свободы связано с числом единиц совокупности *n* и с числом определяемых по ней констант. Применительно к исследуемой проблеме число степеней свободы должно показать, сколько независимых отклонений из *n* возможных $[(y_1-\bar{y}),(y_2-\bar{y}),\cdots,(y_n-\bar{y})]$ требуется для образования данной суммы квадратов. Разделив каждую сумму квадратов на соответствующее ей число степеней свободы, получим дисперсию на одну степень свободы

$$D_{общ}=\frac{\sum (y-\bar{y})^2}{n-1},\quad D_{факт}=\frac{\sum (\hat{y}_x-\bar{y})^2}{1},\quad D_{ост}=\frac{\sum (y-\hat{y}_x)^2}{n-2} \qquad (37)$$

Определение дисперсии на одну степень свободы приводит дисперсии к сравнимому виду. Сопоставляя факторную и остаточную дисперсии в расчете на одну степень свободы, получаем величину *F*–критерия

$$F=\frac{D_{факт}}{D_{ост}}, \qquad (38)$$

Статистика (38) используется для проверки нулевой гипотезы $H_0: D_{ост}=D_{факт}$. Табличное значение *F*–критерия – это максимальная величина отношения дисперсий, которая может иметь место при случайном расхождении их для данного уровня вероятности наличия нулевой гипотезы. Вычисленное значение *F*–отношения признается достоверным (отличным от единицы), если оно больше табличного. В этом случае нулевая гипотеза об отсутствии связи признаков отклоняется и делается вывод о существенности этой связи: $F_{факт}>F_{табл}$. В противном случае нулевая гипотеза не может быть отклонена без риска сделать неправильный вывод о наличии связи.

Пример. По группе предприятий, выпускающих один и тот же вид продукции, рассматривается функция издержек $y=b_0+b_1x+e$. Необходимая для расчета оценок параметров b_0 и b_1 информация представлена в таблице 1.

Исходные данные Таблица 1

Номер предприятия	1	2	3	4	5	6	7
Выпуск продукции, тыс. ед., ***x***	1	2	4	3	5	3	4
Затраты на производство, млн. руб., ***y***	30	70	150	100	170	100	150

Решение

1. Система нормальных уравнений будет иметь вид:

$$\begin{cases} 7 \cdot b_0 + 22 \cdot b_1 = 770 \\ 22 \cdot b_0 + 80 \cdot b_1 = 2820 \end{cases}$$

Решив ее, получим b_0= –5,79; b_1=36,84. Запишем уравнение регрессии $\hat{y} = -5{,}79 + 36{,}84x$. Подставив в данное уравнение значения x, найдем теоретические значения $\hat{y}_i$ (см. седьмой столбец таблицы 2). В данном случае величина параметра b_0 не имеет экономического смысла. В рассматриваемом примере имеем:

$\bar{x} = 3{,}14$; $\sigma_x = 1{,}25$; $V_x = 39{,}8\%$; $\bar{y} = 110$; $\sigma_y = 46{,}29$; $V_y = 42{,}1\%$ (V_x и V_y – коэффициенты вариации по переменным x и y). Расчетные данные помещены в таблице 2.

Расчетные данные Таблица 2

Номер предприятия	x	y	x^2	xy	y^2	$\hat{y}_i$	e_i	e_i^2
1	1	30	1	30	900	31,05	–1,05	1,1025
2	2	70	4	140	4900	67,89	2,11	4,4521
3	4	150	16	600	22500	141,57	8,43	71,0649
4	3	100	9	300	10000	104,73	–4,73	22,3729
5	5	170	25	850	28900	178,41	–8,41	70,7281
6	3	100	9	300	10000	104,73	–4,73	22,3729
7	4	150	16	600	22500	141,57	8,43	71,0649
Итого	22	770	80	2820	99700	769,95	≈0	263,1583
Среднее значение	3,14	110	11,43	402,8571	14242,8571			

2. Применительно к нашему примеру матричный метод определения МНК–оценок сводится к следующему:

а) по правилу умножения матриц

$$X^T X = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 3 & 5 & 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 4 \\ 1 & 3 \\ 1 & 5 \\ 1 & 3 \\ 1 & 4 \end{pmatrix} = \begin{pmatrix} 7 & 22 \\ 22 & 80 \end{pmatrix},$$

в матрице $X^T X$ число 7 получено как сумма произведений элементов первой строки матрицы X^T и первого столбца матрицы X, что соответствует объему

совокупности n; аналогично число 22 получено как сумма произведений элементов первой строки матрицы X^T и второго столбца матрицы X и т.д.;

б) найдем обратную матрицу:

$$\left(X^T X\right)^{-1} = \frac{1}{7 \cdot 80 - (22)^2} \cdot \begin{pmatrix} 80 & -22 \\ -22 & 7 \end{pmatrix} = \begin{pmatrix} 1{,}05263 & -0{,}28947 \\ -0{,}28947 & 0{,}09211 \end{pmatrix};$$

в) $$X^T Y = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 3 & 5 & 3 & 4 \end{pmatrix} \cdot \begin{pmatrix} 30 \\ 70 \\ 150 \\ 100 \\ 170 \\ 100 \\ 150 \end{pmatrix} = \begin{pmatrix} 770 \\ 2820 \end{pmatrix};$$

г) вектор оценок параметров регрессии будет

$$B = \begin{pmatrix} 1{,}05263 & -0{,}28947 \\ 0{,}28947 & 0{,}09211 \end{pmatrix} \cdot \begin{pmatrix} 770 \\ 2820 \end{pmatrix} = \begin{pmatrix} -5{,}79 \\ 36{,}84 \end{pmatrix},$$

где число –5,79 соответствует сумме произведений значений первого столбца матрицы $(X^T X)^{-1}$ на столбец матрицы $X^T Y$ и представляет собой свободный член уравнения регрессии, т.е. b_0= –5,79; число 36,84 получено как сумма произведений второго столбца первой матрицы на столбец матрицы $X^T Y$ и соответствует величине коэффициента регрессии b_1. Тогда уравнение регрессии имеет вид $\hat{y} = -5{,}79 + 36{,}84x$.

После того, как уравнение регрессии получено, начинается проверка качества полученного уравнения регрессии.

Наиболее важной на начальном этапе статистического анализа построенной модели является задача установления наличия линейной зависимости между Y и X. При этом проверяется гипотезы: а) $H_0 : b_1 = 0$ против $H_1 : b_1 \neq 0$ и б) $H_0 : b_0 = 0$ против $H_1 : b_0 \neq 0$ при заданном уровне значимости α=0,05 и объеме выборки n=7. Для вычисления $t_{набл} = \frac{b_1}{S_{b_1}}$ и $t_{набл} = \frac{b_0}{S_{b_0}}$ необходимо вычислить S_{b_1} и S_{b_0}

$$S_{b_1}^2 = \frac{\sum e_i^2}{n \cdot (n-2)(\overline{x^2} - \overline{x}^2)} = \frac{263{,}1583}{7 \cdot 5 \cdot (11{,}4286 - 3{,}1429^2)} = 4{,}8484;\ S_{b_1} = 2{,}202$$

$$S_{b_0}^2 = \frac{S^2 \sum x_i^2}{n^2(\overline{x^2} - \overline{x}^2)} = S_{b_1}^2 \cdot \overline{x^2} = 4{,}8484 \cdot 11{,}4286 = 55{,}413;\ S_{b_0} = 7{,}444$$

Дисперсию оценки b_0 можно найти другим способом, рассчитав сначала значение необъясненной дисперсии S^2:

$$S^2 = \frac{\sum e_i^2}{n-2} = \frac{263{,}1583}{7-2} = 52{,}632$$

$$S_{b_0}^2 = \frac{S^2 \sum x_i^2}{n^2(\overline{x^2} - \overline{x}^2)} = \frac{52{,}632 \cdot 80}{7^2 \cdot (11{,}4286 - 3{,}1429^2)} = \frac{4210{,}56}{7 \cdot 10{,}855} = 55{,}413.$$

Далее находим $t_{b_1} = \frac{36{,}84}{2{,}202} = 16{,}73$ и $t_{b_0} = \frac{-5{,}79}{7{,}44} = -0{,}778$. По таблице распределения Стьюдента находим $t_{êðèò}\,(0{,}025; 7-2) = 2{,}571$ и сравниваем наблюдаемые значения *t*–статистик с критическими. В нашем случае $\left|t_{b_1}\right| = 16{,}73 > 2{,}571 = t_{êðèò}$, следовательно, гипотеза о статистической незначимости коэффициента b_1 отвергается с вероятностью ошибки 0,95. По «грубому» правилу $\left|t_{íàáë}\right| > 3$, что свидетельствует о наличии линейной связи между переменными *X* и *Y*. Сравнивая $\left|t_{íàáë}\right| = 0{,}778 < 2{,}571 = t_{êðèò}$, для коэффициента регрессии b_0, делаем вывод, что гипотеза о статистической незначимости коэффициента b_0 не отклоняется. Это означает, что свободным членом в уравнении регрессии можно пренебречь, рассматривая регрессию как $\hat{y} = b_1 x$, т.е. $\hat{y} = 36{,}84x$.

После проверки значимости каждого коэффициента регрессии обычно проверяется общее качество уравнения регрессии. Для этой цели вычисляется коэффициент детерминации

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{263{,}1583}{15000} = 1 - 0{,}018 = 0{,}982$$

Выполним расчеты коэффициента детерминации по другой формуле $R^2 = r_{xy}^2$, где r_{xy} – коэффициент корреляции между *X* и *Y*.

$$r_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sqrt{\overline{x^2} - \overline{x}^2}\sqrt{\overline{y^2} - \overline{y}^2}} = \frac{402{,}8571 - 3{,}1429 \cdot 110}{\sqrt{11{,}4286 - 3{,}1429^2}\sqrt{14242{,}8571 - 110^2}} = 0{,}991$$

$$R^2 = 0{,}991^2 = 0{,}9826 \approx 0{,}983$$

Далее выдвигаем гипотезы $H_0 : R^2 = 0$ (статистически незначим) против $H_1 : R^2 > 0$ (статистически значим), задаем уровень значимости α=0,95 и объем

выборки в нашем случае n=7. Для проверки нулевой гипотезы используется F–статистика $F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$, где m – число объясняющих переменных. Величина F при выполнении предпосылок МНК и справедливости нулевой гипотезы имеет распределение Фишера. Тогда

$$F_{набл} = \frac{0{,}982}{1-0{,}982} \cdot \frac{7-1-1}{1} = \frac{0{,}982 \cdot 5}{0{,}018} = 272{,}78$$

$$F_{крит}(\alpha, m, n-m-1) = F_{крит}(0{,}05; 1; 5) = 6{,}61 .$$

Так как $|F_{набл}| = 272{,}78 > 6{,}61 = F_{крит}$, то нулевая гипотеза отклоняется. Это равносильно тому, что R^2 статистически значим и качество уравнения регрессии хорошее.

Построим интервальные оценки коэффициентов линейного уравнения регрессии. Используем формулы (30), (31), подставляя в них значения $b_0, b_1, t_{\frac{\alpha}{2}, n-2}, S_{b_0}, S_{b_1}$, получаем

$$-5{,}79 - 2{,}571 \cdot 7{,}444 < \beta_0 < -5{,}79 + 2{,}571 \cdot 7{,}444; \quad -24{,}928 < \beta_0 < 13{,}348$$

$$36{,}84 - 2{,}571 \cdot 2{,}202 < \beta_1 < 36{,}84 + 2{,}571 \cdot 2{,}202 ; \qquad 31{,}179 < \beta_1 < 42{,}501$$

Построим доверительные интервалы для зависимой переменной. Сначала рассчитаем 95%–й доверительный интервал для для условного математического ожидания $M(Y / X = x_p)$ при X=7. Воспользовавшись формулой (32), определим границы доверительного интервала

$$-5{,}79 + 36{,}84 \cdot 7 \pm 2{,}571 \cdot 7{,}255 \sqrt{\frac{1}{7} + \frac{(3{,}1429 - 7)^2}{10{,}858}}$$ или (229,26; 274,92).

Таким образом, доверительный интервал для среднего значения Y при X=7 имеет вид (229,26; 274,92). Другими словами, затраты на производство при выпуске продукции X=7 (тыс. единиц) с вероятностью 95%будут находиться в интервале (229,26; 274,92).

Далее рассчитаем границы интервала, в котором будет сосредоточено не менее 95% возможных средних затрат при неограниченно большом числе наблюдений и объеме выпуска X=7. Для этого воспользуемся формулой (34)

$$-5{,}79 + 36{,}84 \cdot 7 \pm 2{,}571 \cdot 7{,}255 \sqrt{1 + \frac{1}{7} + \frac{(3{,}1429 - 7)^2}{10{,}858}}$$ или (222,527; 281,653).

Таким образом, интервал, в котором будут находиться, по крайней мере, 95% индивидуальных значений затрат на производство при выпуске продукции X=7 имеет вид (222,527; 281,653). Нетрудно заметить, что он включает в себя доверительный интервал для условного среднего затрат на производство.

Уравнение регрессии всегда дополняется показателем тесноты связи. При использовании линейной регрессии в качестве такого показателя выступает линейный коэффициент корреляции r_{xy}. Имеются разные модификации формулы линейного коэффициента корреляции, например: $r_{xy} = b_1 \frac{\sigma_x}{\sigma_y} = \frac{\mu_{xy}}{\sigma_x \cdot \sigma_y} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\sigma_x \cdot \sigma_y}$. Как известно, линейный коэффициент корреляции находится в пределах $-1 \le r_{xy} \le 1$. Если коэффициент регрессии $b_1>0$, то $0 \le r_{xy} \le 1$, и, наоборот, при $b_1<0$ $-1 \le r_{xy} \le 0$. По данным таблицы 1 величина линейного коэффициента корреляции составила $r_{xy} = 0{,}991$, что означает наличие очень тесной зависимости затрат на производство от величины объема выпущенной продукции.

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

Для оценки качества подбора линейной функции рассчитывается квадрат линейного коэффициента корреляции r_{xy}^2, называемый *коэффициентом детерминации.* Он характеризует долю дисперсии результативного признака *y*, объясняемую регрессией, в общей дисперсии результативного признака $r_{xy} = \frac{\sigma^2_{yобъяс.}}{\sigma^2_{yоб.}}$. Соответственно величина $1 - r^2$ характеризует долю дисперсии переменной у, вызванному влиянием остальных, неучтенных в модели факторов. В нашем примере $r^2 = 0{,}982$. Таким образом, уравнением регрессии объясняется 98,2 % дисперсии результативного признака, а на долю прочих факторов приходится лишь 1,8 % ее дисперсии (так называемая остаточная дисперсия).

Величина коэффициента детерминации является одним из критериев оценки качества линейной модели. Чем больше доля объясненной дисперсии, тем соответственно меньше роль прочих факторов и, следовательно, линейная модель хорошо аппроксимирует исходные данные, и ею можно воспользоваться для прогноза значений результативного признака. Таким образом, коэффициент детерминации является мерой, позволяющей определить, в какой степени найденная прямая регрессии дает лучший результат для объяснения поведения зависимой переменной *y*, чем прямая $Y = \overline{y}$. Естественно возникает вопрос, какое значение коэффициента детерминации можно считать удовлетворительным. Точную границу приемлемости (статистической значимости) R^2 для всех случаев сразу указать

невозможно. Нужно обращать внимание на объем выборки, число объясняющих переменных, наличие трендов и содержательную интерпретацию. Коэффициент детерминации может оказаться даже отрицательным. Обычно это случается для линейных уравнений регрессии, в которых отсутствует свободный член. Оценивая такое уравнение по МНК, мы вынуждены рассматривать лишь те прямые (гиперплоскости), которые проходят через начало координат. Значение коэффициента детерминации будет отрицательным и в том случае, если разброс значений зависимой переменной вокруг линии $Y = \bar{y}$ меньше, чем вокруг любой из прямых (гиперплоскостей), проходящих через начало координат.

МНОЖЕСТВЕННОЕ ЛИНЕЙНОЕ УРАВНЕНИЕ РЕГРЕССИИ

На любой экономический показатель чаще всего оказывает влияние не один, несколько факторов. Например, спрос на некоторое благо определяется не только ценой данного блага, но и ценами на замещающие и дополняющие блага, доходом потребителей и многими другими факторами. Для того, чтобы иметь правильное представление о влиянии дохода на потребление, необходимо изучить их корреляцию при неизменном уровне других факторов. Решение такой задачи предполагает отбор единиц совокупности с одинаковыми значениями всех других факторов, кроме дохода. Этот путь приводит к планированию эксперимента – методу, который используется в химических, физических, биологических и других исследованиях. Экономист в отличие от экспериментатора–естественника лишен возможности регулировать другие факторы. Поведение отдельных экономических переменных контролировать нельзя, т.е. не удается обеспечить равенство всех прочих условий для оценки влияния одного исследуемого фактора. В этом случае следует попытаться выявить влияние других факторов, введя их в модель, т.е. построить уравнение множественной регрессии

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \qquad (39)$$

где ε_i – взаимно некоррелированные случайные величины с нулевым математическим ожиданием и дисперсией σ^2. Такого рода уравнение может применяться при изучении потребления. Тогда коэффициенты β_i – частные производные потребления y по соответствующим факторам x_i:

$\beta_1 = \frac{dy}{dx_1}, \beta_2 = \frac{dy}{dx_2}, \ldots, \beta_k = \frac{dy}{dx_k}$ в предположении, что все остальные x_i постоянны.

В 1930–е годы Дж. М. Кейнс сформулировал гипотезу потребительской функции. С того времени исследователи неоднократно обращались к проблеме ее совершенствования. Современная потребительская функция рассматривается

чаще всего как модель вида $C = f(y, P, M, Z)$, где C – потребление; y – доход; P – цена, индекс стоимости жизни; M – наличные деньги; Z – ликвидные активы.

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

Построение модели множественной регрессии включает в себя все этапы построения регрессионной модели в случае парной линейной регрессии.

Оценка параметров уравнения множественной линейной регрессии

Параметры уравнения множественной регрессии оцениваются, как и в парной регрессии, методом наименьших квадратов. При его применении строится система нормальных уравнений, решение которой и позволяет получить оценки параметров регрессии. Так, для уравнения (39) система нормальных уравнений будет иметь вид

$$\begin{cases} \sum y = nb_0 + b_1\sum x_1 + b_2\sum x_2 + \ldots + b_k\sum x_k \\ \sum y \cdot x_1 = b_0\sum x_1 + b_1\sum x_1^2 + b_2\sum x_1x_2 + \ldots + b_k\sum x_kx_1 \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ \sum y \cdot x_k = b_0\sum x_k + b_1\sum x_1x_k + b_2\sum x_2x_k + \ldots + b_k\sum x_k^2 \end{cases} \qquad (40)$$

Ее решение может быть осуществлено методом определителей

$b_0 = \dfrac{\Delta b_0}{\Delta}$, $b_1 = \dfrac{\Delta b_1}{\Delta}$, …, $b_k = \dfrac{\Delta b_k}{\Delta}$, где Δ – определитель системы, составленный из коэффициентов при неизвестных; $\Delta b_0, \Delta b_1, \ldots, \Delta b_k$ – частные определители. При этом

$$\Delta = \begin{vmatrix} n & \sum x_1 & \sum x_2 & \ldots & \sum x_k \\ \sum x_1 & \sum x_1^2 & \sum x_2x_1 & \ldots & \sum x_kx_1 \\ \sum x_2 & \sum x_1x_2 & \sum x_2^2 & \ldots & \sum x_kx_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ \sum x_k & \sum x_1x_k & \sum x_2x_k & \ldots & \sum x_k^2 \end{vmatrix} \qquad (41)$$

Уравнение множественной линейной регрессии в матричной форме имеет вид

$$Y = XB + E \qquad (42)$$

где $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$; $X = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix}$; $B = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{pmatrix}$; $E = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$ (43)

Процедура оценки параметров b_0, b_1, …, b_n такая же, как и в парной линейной регрессии.

Пример. На основе линейной регрессионной модели исследовать зависимость фондоотдачи в процентах на единицу ОПФ (y) от среднечасовой производительности вращающихся печей (x_1) и удельного веса активной части ОПФ (x_2). В таблице 3 приводятся исходные данные для n=15 цементных заводов страны.

Исходные данные Таблица 3

Номер пред–приятия	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Фондо–отдача, y	26	33	24	29	42	24	52	56	26	45	27	54	34	48	45
Средне–часовая производитель–ность печей, x_1	37	33	15	36	26	24	15	33	44	34	63	8	44	43	31
Удель–ный вес актив–ной части ОПФ (%), x_2	39	40	35	48	53	42	54	54	50	53	46	50	43	55	51

Решение Уравнение регрессии будем искать в виде $\hat{y} = b_0 + b_1 x + b_2 x_2 + e$. Для определения вектора оценок $b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$ найдем предварительную симметричную матрицу $X^T X$, которая имеет вид

$$X^T X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} \end{pmatrix} = \begin{pmatrix} n & \sum x_{i1} & \sum x_{i2} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} \\ \sum x_{i2} & \sum x_{i1}x_{i2} & \sum x_{i2}^2 \end{pmatrix}$$

и равна для нашего случая

$$X^T X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 37 & 33 & \dots & 31 \\ 39 & 40 & \dots & 51 \end{pmatrix} \begin{pmatrix} 1 & 37 & 39 \\ 1 & 33 & 40 \\ \vdots & \vdots & \vdots \\ 1 & 31 & 51 \end{pmatrix} = \begin{pmatrix} 15 & 486 & 713 \\ 486 & 18416 & 23132 \\ 713 & 23132 & 34455 \end{pmatrix}$$

Вектор $X^T Y$ имеет вид

$$X^T Y = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{11} & x_{21} & \dots & x_{n1} \\ x_{12} & x_{22} & \dots & x_{n2} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum y_i \\ \sum x_{i1} y_i \\ \sum x_{i2} y_i \end{pmatrix}$$ и для нашего примера равен

$$X^T Y = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 & 1 \\ 32 & 33 & 15 & \dots & 43 & 31 \\ 39 & 40 & 35 & \dots & 55 & 51 \end{pmatrix} \begin{pmatrix} 26 \\ 33 \\ 24 \\ \vdots \\ 48 \\ 45 \end{pmatrix} = \begin{pmatrix} 565 \\ 17513 \\ 27656 \end{pmatrix}$$

Далее необходимо получить матрицу, обратную к $X^T X$. Воспользуемся методом нахождения обратной матрицы, известным из курса «Высшей математики», получим обратную матрицу

$$\left(X^T X\right)^{-1} = \begin{pmatrix} 4{,}4075412 & -0{,}011171 & -0{,}083709 \\ -0{,}011171 & 0{,}000375 & -0{,}000020 \\ -0{,}083709 & -0{,}000020 & 0{,}001775 \end{pmatrix}$$

Далее находим вектор оценок

$$b = \left(X^T X\right)^{-1} X^T Y = \begin{pmatrix} 4{,}4075412 & -0{,}011171 & -0{,}083709 \\ -0{,}011171 & 0{,}000375 & -0{,}000020 \\ -0{,}083709 & -0{,}000020 & 0{,}001775 \end{pmatrix} \begin{pmatrix} 565 \\ 17513 \\ 27656 \end{pmatrix} = \begin{pmatrix} -20{,}4102 \\ -0{,}3118120 \\ 1{,}505803 \end{pmatrix}$$

Таким образом,

$b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} = \begin{pmatrix} -20{,}4102 \\ -0{,}3118120 \\ 1{,}505803 \end{pmatrix}$ и оценка уравнения регрессии имеет вид

$\hat{y} = -20{,}4102 - 0{,}3118120x_1 + 1{,}505803x_2$.

Для проверки значимости уравнения регрессии необходимо найти $Q_{îñò} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ и $Q_R = \sum_{i=1}^{n}\hat{y}_i^2$. Для этих целей составим вспомогательную таблицу (таблица 4).

Вспомогательная таблица Таблица 4

Номер предприятия	y_i	x_{1i}	x_{2i}	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	26	37	39	26,779073	0,6069547
2	33	33	40	29,532124	12,026163
3	24	15	35	27,615725	13,073467
4	29	36	48	40,643112	135,56205
5	42	26	53	51,290247	86,308689
6	24	24	42	35,350038	128,82336
7	52	15	54	56,225982	17,858923
8	56	33	54	50,613366	29,015825
9	26	44	50	41,160222	229,83233
10	45	34	53	48,795751	14,407725
11	27	63	46	29,212582	4,8955191
12	54	8	50	52,385454	2,6067587
13	34	44	43	30,619601	11,427097
14	48	43	55	49,001049	1,0020991
15	45	31	51	46,719581	2,9569588
Итого	565	486	713		690,40385

Из таблицы 3 следует, что $Q_{ост} = 690{,}40385$. Тогда несмещенная оценка остаточной дисперсии σ^2 равна

$\hat{s}^2 = \frac{1}{n-3}Q_{ост} = \frac{1}{12} \cdot 690{,}40385 = 57{,}533654$ и $\hat{s} = 7{,}5850941$. По данным таблицы 3 найдем $Q_R = \sum_{i=1}^{n}\hat{y}_i^2 = 26801{,}319$. Далее на уровне значимости α=0,05 проверим значимость уравнения регрессии в целом, т.е. гипотезу $H_0 : \beta = 0$ (все коэффициенты в уравнении регрессии статистически незначимы) против $H_1 : \beta \neq 0$ (хотя бы один из коэффициентов в уравнении

регрессии является статистически значимым). Для проверки выдвинутой нулевой гипотезы используется статистика

$$F_{набл} = \frac{\frac{1}{k+1} \cdot Q_R}{\frac{1}{n-k-1} \cdot Q_{ост}}, \quad (44)$$

которая при выполнении гипотезы H_0 имеет F–распределение с (k+1) и (n-k-1) степенями свободы. Если уравнение регрессии незначимо, т.е. все коэффициенты уравнения регрессии для генеральной совокупности равны нулю, то на этом анализ уравнения регрессии заканчивается. Если же нулевая гипотеза $H_0 : \beta = 0$ отвергается, то представляет интерес проверка значимости отдельных коэффициентов регрессии и построение интервальных оценок для значимых коэффициентов.

Согласно выражению (44) $F_{набл} = \frac{\frac{1}{3} \cdot 26801{,}319}{\frac{1}{12} \cdot 690{,}40385} = 155{,}27907$.

По таблице F–распределения для α=0,05 и чисел степеней свободы ν_1=3 и ν_2=12 найдем критическое значение $F_{кр}(\alpha;\nu_1;\nu_2)= F_{кр}(0{,}05;\ 3;\ 12)=3{,}49$. Так как $F_{набл}>F_{кр}$, то гипотеза $H_0 : \beta = 0$ отвергается, т.е. хотя бы один элемент вектора $\beta = (\beta_0, \beta_1, \beta_2)^T$ не равен нулю.

Перед проверкой значимости отдельных коэффициентов регрессии найдем оценку ковариационной матрицы вектора b

$$\Sigma(b) = \sigma^2 (X^T X)^{-1} \quad (45)$$

Статистический смысл элементов этой матрицы заключается в следующем: элементы главной диагонали представляют собой дисперсии вектора оценок b. Вне главной диагонали ковариационной матрицы расположены значения коэффициентов ковариации (корреляционные моменты). Например на пересечении *i*–й строки и *j*–го столбца матрицы расположен коэффициент ковариации

$$\mathrm{cov}(b_i b_j) = M(b_i - \beta_i)(b_j - \beta_j) \quad (46)$$

Тогда, умножая элементы обратной матрицы $(X^T X)^{-1}$ на $\hat{s}^2 = 57{,}533654$, будем иметь

$$\hat{S}(b) = \hat{s}^2 (X^T X)^{-1} = \begin{pmatrix} 253{,}582 & -0{,}643 & -4{,}816 \\ -0{,}643 & 0{,}0216 & -0{,}0012 \\ -4{,}816 & -0{,}0012 & 0{,}1021 \end{pmatrix}$$

Из статистического смысла ковариационной матрицы следует, что оценки дисперсии коэффициентов уравнения регрессии b_0, b_1, b_2 соответственно равны $\hat{s}^2_{b_0} = 253{,}582$, $\hat{s}^2_{b_1} = 0{,}0216$, $\hat{s}^2_{b_2} = 0{,}1021$.

Проверим значимость коэффициента β1, т.е. гипотезу $H_0 : \beta_1 = 0$. Вычисляем наблюдаемое значение критерия $t_{b_1} = \dfrac{b_1}{\hat{s}_{b_1}} = \dfrac{-0{,}311812}{\sqrt{0{,}0216}} = -2{,}122$.

По таблице *t*–распределения находим $t_{кр}(\alpha, n-3) = t_{кр}(0{,}05; 12) = 2{,}179$. Так как $|t_{b_1}| < t_{кр}$, гипотеза о том, что $\beta_1=0$ (статистически незначим) не отвергается.

Проверим теперь гипотезу $H_0 : \beta_2 = 0$. Вычисляем наблюдаемое значение критерия $t_{b_2} = \dfrac{b_2}{\hat{s}_{b_2}} = \dfrac{1{,}505803}{\sqrt{0{,}1021}} = 4{,}713$. Так как $|t_{b_2}| > t_{кр}$, то гипотеза $H_0 : \beta_2 = 0$ отвергается, т.е. коэффициент β_2 является в уравнении регрессии статистически значимым.

Далее перейдем к алгоритму пошагового регрессионного анализа и исключим из рассмотрения переменную x_1, которая имеет незначимый коэффициент β_1 уравнения регрессии. Уравнение регрессии будем искать в виде $\hat{y} = \beta_0 + \beta_1 x_2$. Исходные данные для оценки коэффициентов β_0 и β_1 представлены в таблице 5.

Исходные данные Таблица 5

Номер предприятия	y_i	x_{2i}	$\hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	26	39	25,46791	0,2831197
2	33	40	26,8840	37,405456
3	24	35	19,80355	17,610192
4	29	48	38,21272	84,874209
5	42	53	45,29317	10,844968
6	24	42	29,71618	32,674713
7	52	54	46,70926	27,991929
8	56	54	46,70926	86,317849
9	26	50	41,0449	226,34901
10	45	53	45,29317	0,0859486
11	27	46	35,38054	70,23345
12	54	50	41,0449	167,83461
13	34	43	31,13227	8,2238758
14	48	55	43,12535	0,157126
15	45	51	42,46099	6,4465717
Итого	565	713		777,19156

Тогда матрица X^T будет иметь вид

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 39 & 40 & 35 & 48 & 53 & 42 & 54 & 54 & 50 & 53 & 46 & 50 & 48 & 55 & 51 \end{pmatrix}$$

$$\text{и } (X^T X) = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 39 & 40 & \cdots & 51 \end{pmatrix} \begin{pmatrix} 1 & 39 \\ 1 & 40 \\ \vdots & \vdots \\ 1 & 51 \end{pmatrix} = \begin{pmatrix} 15 & 713 \\ 713 & 34455 \end{pmatrix}$$

Обратную матрицу $(X^T X)^{-1}$ вычисляем по формуле

$$(X^T X)^{-1} = \frac{1}{\left|X^T X\right|} \cdot \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}^T, \qquad (47)$$

где A_{ij} – алгебраические дополнения к элементу a_{ij} матрицы $X^T X$. Определитель $\left|X^T X\right| = 15 \cdot 34455 - 713 \cdot 713 = 8456$, тогда обратная матрица

$$\left(X^T X\right)^{-1} = \frac{1}{8456} \cdot \begin{pmatrix} 34455 & -713 \\ -713 & 15 \end{pmatrix} = \begin{pmatrix} 4{,}0746515 & -0{,}0843188 \\ -0{,}0843188 & 0{,}0017738 \end{pmatrix}.$$

Находим вектор $X^T Y = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 39 & 40 & \cdots & 51 \end{pmatrix} \cdot \begin{pmatrix} 26 \\ 33 \\ \vdots \\ 45 \end{pmatrix} = \begin{pmatrix} 565 \\ 27656 \end{pmatrix}$. Тогда вектор оценок параметров b_0 и b_1 уравнения регрессии имеет вид

$$b = (X^T X)^{-1}(X^T Y) = \begin{pmatrix} 4{,}074512 & -0{,}0843188 \\ -0{,}0843188 & 0{,}0017738 \end{pmatrix} \cdot \begin{pmatrix} 565 \\ 27656 \end{pmatrix} = \begin{pmatrix} -29{,}7596 \\ 1{,}416090 \end{pmatrix}$$

Оценка уравнения регрессии имеет вид $\hat{Y} = -29{,}7596 + 1{,}416090 x_2$ (без переменной x_1). Далее необходимо оценить качество полученного уравнения регрессии. Для этого по данным таблицы 5 найдем несмещенную оценку остаточной дисперсии

$$\hat{s}^2 = \frac{1}{n-k-1} \cdot Q_{ост} = \frac{1}{n-k-1} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{13} \cdot 777{,}19156 = 59{,}784$$

Найдем оценку ковариационной матрицы вектора b

$$\hat{S}(b) = \hat{s}^2 (X^T X)^{-1} = \begin{pmatrix} 243{,}597 & -5{,}0409 \\ -5{,}0409 & 0{,}1060 \end{pmatrix},$$

откуда $\hat{s}^2_{b_0} = 243{,}597$; $\hat{s}^2_{b_2} = 0{,}106$.

Для проверки значимости коэффициента β_2, т.е. гипотезы $H_0 : \beta_2 = 0$, найдем $t_{b_2} = \frac{b_2}{\hat{s}_{b_2}} = \frac{1{,}41609}{\sqrt{0{,}106}} = 4{,}349$. Затем по таблице распределения Стьюдента определим критическое значение $t_{кр}(\alpha;\ n–2) = t_{кр}(0{,}05;\ 13)=2{,}16$. Так как $t_{b_2} > t_{êð}$, нулевая гипотеза отвергается (коэффициент β_2 является статистически значимым). Таким образом , окончательно оценка уравнения регрессии со значимыми коэффициентами имеет вид $\hat{y} = -29{,}7596 + 1{,}41609x_2$. Коэффициент регрессии при x_2 показывает, что при росте удельного веса активной части ОПФ на единицу (%) фондоотдача в среднем увеличивается на 1,41609 единиц.

Найдем с доверительной вероятностью (надежностью) γ=0,95 интервальную оценку для коэффициента регрессии β_2

$$\beta_2 - t_\gamma \cdot \hat{s}_{b_2} \le \beta_2 \le \beta_2 + t_\gamma \cdot \hat{s}_{b_2}, \qquad (48)$$

где t_γ=2,16 находим по таблице распределения Стьюдента при α = 0,05 и $\nu = n - 2 = 13$.

Тогда $1{,}41609 - 0{,}70324 \le \beta_2 \le 1{,}41609 + 0{,}70324$; $0{,}71285 \le \beta_2 \le 2{,}11933$.

Знания, умения, навыки по теме «Регрессионный анализ»

Изучив тему «Регрессионный анализ», студент

должен знать:

1. Основные виды уравнений регрессии.
2. Алгоритм отбора факторов для построения регрессионной модели.
3. Процедуры анализа регрессионных моделей.
4. Алгоритм анализа двумерной и трехмерной моделей.

должен уметь:

1. Применять теорию регрессионного анализа для решения задач.
2. Видеть возможности использования регрессионного анализа в профессиональной деятельности.
3. Получить навыки использования статистических пакетов для решения задач регрессионного анализа.

1. Варианты заданий и исходные данные для самостоятельной работы на ЭВМ (парная линейная регрессия)

Вариант 1. При измерении некоторой величины y через равные интервалы x получены следующие результаты

X	–3	–2	–1	0	1	2	3
Y	–1,70	–1,01	–0,21	0,52	0,73	1,30	1,50

Провести регрессионный анализ.

Вариант 2. Осредненные данные по группе хозяйств, характеризующие связь между выходом продукции на 1 га сельскохозяйственных угодий (y) и основными средствами совхозов, также приходящимися на 1 га угодий (x), приведены в таблице

X	11,3	12,9	13,6	16,8	18,8	22,0	22,2	23,7
Y	13,2	15,6	17,2	18,8	20,2	23,3	22,4	23,0

Провести регрессионный анализ.

Вариант 3. Результаты исследования динамики веса поросят приведены в таблице

Возраст (недели), x	0	1	2	3	4	5	6
Вес (кг), y	1,2	2,5	3,9	5,2	6,4	7,7	9,2

Провести регрессионный анализ.

Вариант 4. Себестоимость y (тыс. руб.) одного экземпляра книги в зависимости от тиража x (тыс. экз.) характеризуется данными, собранными издательством

X, тыс. экз.	1	2	3	5	10	20	30	50
Y, тыс. руб.	9,10	5,30	4,11	2,83	2,11	1,62	1,41	1,30

Провести регрессионный анализ.

Вариант 5. На основании данных о динамике процента хронических больных на 1000 жителей, приведенных в таблице, провести регрессионный анализ.

Годы, x	0	1	2	3	4	5	6
Больные (%), y	10	8	7	5	3	2	2

Вариант 6. Изучается зависимость материалоемкости продукции от размера предприятия по 10 однородным заводам (данные приведены в таблице)

Потреблено материалов на единицу продукции, кг	9	6	5	4	3,7	3,6	3,5	6	7	3,5
Выпуск, тыс. единиц	10	20	30	40	50	60	70	150	120	100

Провести регрессионный анализ.

Вариант 7. Имеются данные о цене однокомнатной квартиры и величине ее общей площади по 10 сделкам одного района города

Цена квартиры, тыс. руб.	29	31	35	35	45	46	45	44	38
Площадь, $м^2$	35	35	33	34	38	40	40	39	37

Провести регрессионный анализ.

Вариант 8. Имеются данные о годовой цене программы «Мастер делового администрирования» (MBA) и числе слушателей в образовательном учреждении

Цена программы, тыс. руб., *y*	8	5	4,9	4	3,8	3,5	3,8	3,7	3,6	3,5
Число слушателей, чел., *x*	5	10	12	15	20	22	25	30	35	36

Провести регрессионный анализ.

Вариант 9. Имеются данные по 10 хозяйствам

Урожайность, ц/га, *y*	15	18	17	22	25	20	24	19	23	27
Внесено удобрений, кг/га, *x*	2,1	3,6	3,5	5,0	6,5	4,2	6,3	6,0	6,8	7,5

Провести регрессионный анализ.

Вариант 10. По 10 регионам страны изучается зависимость ежемесячного среднедушевого денежного дохода *y* от удельного веса населения в трудоспособном возрасте в общей численности населения, *x*

Удельный вес населения в трудоспособном возрасте в общей численности населения, %, *x*	60,6	59,6	60,8	59,4	60,4	60,8	60,6	59,3	60,3	62
Среднедушевой ежемесячный денежный доход, тыс. руб., *y*	3,4	3,1	3,7	3,4	3,6	3,3	3,1	3,3	3,6	4,2

Провести регрессионный анализ.

Вариант 11. По 10 регионам страны изучается зависимость ежемесячного среднедушевого денежного дохода y от удельного веса населения в трудоспособном возрасте в общей численности населения, x

Удельный вес населения в трудоспособном возрасте в общей численности населения, %, x	59,3	60,3	62,3	60,2	59	61,4	58,9	59,2	59,8	61
Среднедушевой ежемесячный денежный доход, тыс. руб., y	3,3	3,6	3,8	3,2	3,3	4,1	3,4	3,2	3,4	3,9

Провести регрессионный анализ.

Вариант 12. По 10 регионам страны изучается зависимость ожидаемой продолжительности жизни при рождении (лет) y от уровня заболеваемости детей в возрасте 0–14 лет на тысячу человек, x

Номер региона	Уровень заболеваемости детей в возрасте 0–14 лет на тыс. чел., x	Ожидаемая продолжительность жизни при рождении, лет, y
1	1108,4	67,5
2	1164,4	69,3
3	438,8	75,1
4	618,1	68,7
5	1312,4	66,2
6	982,2	68,1
7	843,0	70,0
8	1233,6	67,3
9	1173,0	67,1
10	1415,5	65,4

Провести регрессионный анализ.

Вариант 13. По 10 регионам страны изучается зависимость розничной продажи телевизоров (y) от среднедушевых денежных доходов в месяц (x)

Среднедушевой денежный доход в месяц, тыс. руб., x	2,8	2,4	2,1	2,6	2,7	2,5	2,4	2,6	2,8	2,6
Розничная продажа телевизоров, тыс. шт, y	28,0	21,3	21,0	23,3	25,8	21,9	20,0	22,0	23,9	26

Провести регрессионный анализ.

Вариант 14. По 10 регионам страны изучается зависимость розничной продажи телевизоров (y) от среднедушевых денежных доходов в месяц (x)

Среднедушевой денежный доход в месяц, тыс. руб., x	2,5	2,7	2,6	2,8	2,9	3,1	3,2	3,3	3,9	4,0
Розничная продажа телевизоров, тыс. шт, y	21	22	23	24	27	28	30	31	32	34

Провести регрессионный анализ.

Вариант 15. По 10 регионам страны изучается зависимость розничной продажи видеомагнитофонов (y) от среднедушевых ежемесячных денежных доходов (x)

Среднедушевой ежемесячный денежный доход, тыс. руб., x	2,4	3,0	2,2	4,0	2,5	5,0	2,3	3,0	3,4	4,0
Розничная продажа видеомагнитофонов, тыс. шт., y	4,8	5,7	5,1	5,5	6,2	4,9	7,0	4,7	4,9	5,5

Провести регрессионный анализ.

2. Варианты заданий и исходные данные для самостоятельной работы на ЭВМ (множественная регрессия)

Варианты заданий 1–25 по регрессионному анализу представлены в таблице 6, а значения показателей производственно–хозяйственной деятельности предприятий машиностроения приведены в таблице 7.

Рассматриваются следующие показатели:
Y_1 – производительность труда;
Y_2 – индекс снижения себестоимости продукции;
Y_3 – рентабельность;
X_4 – трудоемкость единицы продукции;
X_5 – удельный вес рабочих в составе ППП;
X_6 – удельный вес покупных изделий;
X_7 – коэффициент сменности оборудования;
X_8 – премии и вознаграждения на одного работника;
X_9 – удельный вес потерь от брака;
X_{10} – фондоотдача;
X_{11} – среднегодовая численность ППП;
X_{12} – среднегодовая стоимость ОПФ;
X_{13} – среднегодовой фонд заработной платы ППП;
X_{14} – фондовооруженность труда;
X_{15} – оборачиваемость нормируемых оборотных средств;
X_{16} – оборачиваемость ненормируемых оборотных средств;
X_{17} – непроизводственные расходы.

Варианты заданий 1–25 по регрессионному анализу Таблица 6

Номер варианта	Результативный признак, Y	Номера факторных признаков, X
1	1	6, 8, 11, 12, 17
2	1	6, 8, 11, 13, 17
3	1	8, 11, 12, 13, 17
4	1	6, 8, 13, 14, 17
5	1	8, 11, 13, 14, 17
6	1	6, 8, 12, 13,17
7	1	7, 11, 12, 13 17
8	1	7, 9, 12, 13, 17
9	1	8, 11, 12, 13, 17
10	1	8, 9, 13, 14, 17
11	1	5, 6, 7, 9, 17
12	1	5, 7, 9, 11, 17
13	1	5, 6, 12, 13, 17
14	1	5, 7, 10, 14, 17
15	1	5, 6, 10, 14, 17
16	3	8, 10, 15, 16, 17
17	3	5, 6, 10, 15, 17
18	3	5, 6, 7, 11, 12
19	3	8, 9, 10, 11, 17
20	3	8, 9, 10, 12, 17
21	2	4, 5, 6, 8, 9
22	2	4, 5, 6, 7, 9
23	2	5, 6, 7, 8, 9
24	2	4, 5, 8, 9, 17
25	2	4, 5, 7, 9, 17

Таблица исходных данных Таблица 7

Номер предприятия	Y_1	Y_2	Y_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
1	9,26	204,2	13,26	0,23	0,78	0,40	1,37	1,23	0,23	1,45
2	9,38	209,6	10,16	0,24	0,75	0,26	1,49	1,04	0,39	1,30
3	12,11	222,6	13,72	0,19	0,68	0,40	1,44	1,80	0,43	1,37
4	10,81	236,6	12,85	0,17	0,70	0,50	1,42	0,43	0,18	1,65
5	9,35	62,0	10,63	0,23	0,62	0,40	1,35	0,88	0,15	1,91
6	9,87	53,1	9,12	0,43	0,76	0,19	1,39	0,57	0,34	1,68
7	8,17	172,1	25,83	0,31	0,73	0,25	1,16	1,72	0,38	1,94
8	9,12	56,5	23,39	0,26	0,71	0,44	1,27	1,70	0,09	1,89

9	5,88	52,6	14,68	0,49	0,69	0,17	1,16	0,84	0,14	1,94
10	6,30	46,6	10,05	0,36	0,73	0,39	1,25	0,60	0,21	2,06
11	6,22	53,2	13,99	0,37	0,68	0,33	1,13	0,82	0,42	1,96
12	5,49	30,1	9,68	0,43	0,74	0,25	1,10	0,84	0,05	1,02
13	6,50	146,4	10,03	0,35	0,66	0,32	1,15	0,67	0,29	1,85
14	6,61	18,1	9,13	0,38	0,72	0,02	1,23	1,04	0,48	0,88
15	4,32	13,6	5,37	0,42	0,68	0,06	1,39	0,66	0,41	0,62

Продолжение таблицы 7

Номер предприятия	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	X_{16}	X_{17}
1	26006	167,69	47750	6,40	166,32	10,08	17,72
2	23935	186,10	50391	7,80	92,88	14,76	18,39
3	22589	220,45	43149	9,76	158,04	6,48	26,46
4	21220	169,30	41089	7,90	93,96	21,96	22,37
5	7394	39,53	14257	5,35	173,88	11,88	28,13
6	11586	40,41	22661	9,90	162,30	12,60	17,55
7	26609	102,96	52509	4,50	88,56	11,52	21,92
8	7801	37,02	14903	4,88	101,16	8,28	19,52
9	11587	45,74	25587	3,46	166,32	11,52	23,99
10	9475	40,07	16821	3,60	140,76	32,40	21,76
11	10811	45,44	19459	3,56	128,52	11,52	25,68
12	6371	41,08	12973	5,65	177,84	17,28	18,13
13	26761	136,14	50907	4,28	114,48	16,20	25,74
14	4210	42,39	6920	8,85	93,24	13,32	21,21
15	3557	37,39	5736	8,52	126,72	17,28	22,97

ГЛОССАРИЙ

- **Бесповторная выборка** – это выборка, при которой отобранный объект в генеральную совокупность не возвращается.
- **Вероятность** – количественная мера объективной возможности появления события в условиях данного эксперимента.
- **Вероятность доверительная** – вероятность, признаваемая достаточной для суждения о достоверности характеристик, полученных на основе выборочных наблюдений.
- **Вероятность события классическая** – отношение числа исходов эксперимента, благоприятствующих появлению события, к общему числу исходов.
- **Двусторонняя критическая область –** это область, определяемая как $(-\infty, k_{1-\alpha/2}) \cup (k_{\alpha/2}, +\infty)$. Она определяется в случае, когда альтернативная гипотеза имеет вид:$H_1: \Theta \neq \Theta_0$.
- **Дисперсия случайной величины** – математическое ожидание квадрата отклонений случайной величины от ее математического ожидания.
- **Интервальная оценка** – числовой интервал, относительно которого с вероятностью, близкой к единице можно утверждать, что оцениваемый параметр находится внутри него.
- **Конкурирующая (альтернативная) гипотеза** – гипотеза, противоположная нулевой, которая будет верна в том случае, если нулевая гипотеза противоречит опытным данным.
- **Корреляционная зависимость** – зависимость математического ожидания одной случайной величины от вариации других.
- **Критическая область** – подмножество значений выборочной характеристики, составляющей основу статистического критерия, при которых нулевая гипотеза отвергается.
- **Критические точки –** это точки, разделяющие критическую область и область принятия гипотезы.
- **Левосторонняя критическая область** – это область, определяемая как $(-\infty, k_{1-\alpha})$. Она используется в случае, когда альтернативная гипотеза имеет вид: $H_1: \Theta < \Theta_0$.
- **Математическое ожидание** – средняя величина возможных значений случайной величины, взвешенных по их вероятности.
- **Механический отбор** – это отбор, при котором генеральную совокупность «механически» делят на столько групп, сколько объектов должно войти в выборку, а из каждой группы выбирают один объект.
- **Мощность критерия** – вероятность не совершить ошибку второго рода, отвергнуть ложную нулевую гипотезу.

- **Несмещенная оценка** – точечная оценка параметра, математическое ожидание которой равно самому параметру.
- **Нулевая гипотеза** – статистическая гипотеза, которую необходимо проверить.
- **Правосторонняя критическая область** – это область, определяемая как $(k_\alpha, +\infty)$. Она используется в случае, когда альтернативная гипотеза имеет вид: $H_1: \Theta > \Theta_0$.
- **Повторная выборка** – это выборка, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.
- **Простой случайный отбор** – это отбор, при котором объекты извлекают по одному из всей генеральной совокупности.
- **Серийный отбор** – это отбор, при котором объекты отбирают из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию.
- **Состоятельная оценка** – точечная оценка, которая сходится по вероятности к оцениваемому параметру.
- **Статистические выводы –** это заключения о генеральной совокупности (т.е. о законе распределения исследуемой случайной величины и его параметрах либо о наличии и силе связи между исследуемыми переменными) на основе выборки, случайно отобранной из генеральной совокупности.
- **Статистическая гипотеза** – любое предположение либо относительно неизвестного закона распределения, либо относительно неизвестных параметров известного закона распределения.
- **Статистический критерий** – однозначно определенное правило, устанавливающее условия, при которых проверяемую гипотезу следует отвергнуть, либо не отвергать. Основу критерия составляет выборочная характеристика, точное или приближенное распределение которой известно при справедливости нулевой гипотезы. Правила проверки гипотезы определяют, при каких условиях гипотеза будет принята.
- **Типический отбор** – это отбор, при котором объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части.
- **Точечная оценка** – функция результатов наблюдения, значение которой принимается за приближенное значение параметра генеральной совокупности.
- **Уровень значимости** – вероятность не совершить ошибку первого рода, вероятность отвергнуть истинную нулевую гипотезу. С уменьшением вероятности ошибки первого рода увеличивается вероятность ошибки второго рода.
- **Эффективная оценка** – точечная оценка, обладающая наименьшей дисперсией среди всех возможных несмещенных оценок параметра данной генеральной совокупности при фиксированном объеме выборки.

СПИСОК ЛИТЕРАТУРЫ

1. Айвазян, С.А. Классификация многомерных наблюдений /С.А. Айвазян, З.И. Бежаева, О.В. Староверов – М.: Статистика, 1974. – 240 с.
2. Айвазян, С.А. Прикладная статистика. Основы моделирования и первичная обработка данных /С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин – М.: Финансы и статистика, 1983. – 472 с.
3. Айвазян, С.А. Прикладная статистика. Исследование зависимостей /С.А. Айвазян, И.С. Енюков, Л.Д. Мешалкин – М.: Финансы и статистика, 1985. – 488 с.
4. Прикладная статистика. Классификация и снижение размерности /С.А. Айвазян, В.М. Бухштабер, И.С. Енюков и др. – М.: Финансы и статистика, 1989.
5. Андерсон, Т. Введение в многомерный статистический анализ /Т. Андерсон Пер. с англ. – М.: ГИФМЛ, 1963. – 500 с.
6. Болч, Б. Многомерные статистические методы экономики /Б. Болч, К. Хуань Пер. с англ. – М.: Статистика, 1979. – 317 с.
7. Дубров, А.М. Многомерные статистические методы /А.М. Дубров, В.С. Мхитарян, Л.И. Трошин – М.: Финансы и статистика, 2003 . – 352 с.

Составитель
Баранова Татьяна Анатольевна

Многомерные статистические методы.
Регрессионный анализ

Методические указания

Редактор Г.В. Куликова

Подписано в печать 9.03.2007. Формат 60×84 1/16. Печать плоская.
Усл. печ. л. 2,33. Уч. изд. л. 2,58. Тираж 100 экз. Заказ ___
Отпечатано на полиграфическом оборудовании
кафедры экономики и финансов ГОУ ВПО “ИГХТУ”
ГОУ ВПО “Ивановский государственный химико–технологический университет”
153000, г. Иваново, пр. Ф. Энгельса, 7