КАЗАНСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ
ИНСТИТУТ УПРАВЛЕНИЯ, ЭКОНОМИКИ И ФИНАНСОВ
Кафедра экономико-математического моделирования

А.М. ШИХАЛЁВ

РЕГРЕССИОННЫЙ АНАЛИЗ. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Учебно-методическое пособие

Казань – 2015

УДК (075.8)311
ББК 60.6я73

Принято на заседании кафедры экономико-математического моделирования
Протокол № 1 от 18 сентября 2014 года

Рецензенты:
кандидат экономических наук,
доцент кафедры экономико-математического моделирования КФУ **Е.Л. Фесина;**
кандидат экономических наук,
доцент кафедры экономико-математического моделирования КФУ **Е.И. Кадочникова**

Шихалёв А.М.
Регрессионный анализ. Парная линейная регрессия / А.М. Шихалёв. – Казань: Казан. ун-т, 2015. – 46 с.

Учебно-методическое пособие предназначено для подготовки и выполнения студентами самостоятельной работы 3 из раздела «Общая теория статистики».

Регрессионный анализ занимает видное место среди других формализованных методов исследования степени тесноты и направления взаимосвязей социально-экономических явлений. Поскольку аппарат регрессионного анализа служит основой не только для построения последующих суждений о характере взаимосвязи изучаемых процессов, но и для формирования умозаключений, его результаты могут быть использованы не только для описания объекта исследования, но и для выработки точечных экстраполяционных прогнозов.

Ключевые слова: аппроксимация, экстраполяция, функция, аргумент, метод наименьших квадратов, коэффициенты регрессии, ошибка аппроксимации, коэффициент детерминации, нулевая гипотеза.

СОДЕРЖАНИЕ

1. ПОСТАНОВКА ЗАДАЧИ РЕГРЕССИОННОГО АНАЛИЗА

1.1. Возможная мотивация создания теоретической модели наблюдаемого явления

Цели задач статистических исследований можно свести к одной главной: выявление объективных закономерностей исследуемых социально-экономических явлений и процессов. Объективных – значит не зависящих ни от человека, ни от человечества. Множество известных примеров можно дополнить еще одним: с ростом уровня полноценного образования продуктивность и качество результатов труда возрастают. То есть между уровнем образования и результатами труда существует известная всем причинно-следственная связь. Такая связь может быть установлена на содержательном или вербальном уровне (на базе понятий, в том числе и научных) и на формализованном уровне. Тогда причинно-следственная связь между уровнем образования и качеством труда должна быть выражена количественным соотношением, пусть даже в общем виде на языке определенных символов. В таком случае принято записывать примерно как

$$y = f(x), \qquad (1)$$

где y – результаты труда, x – уровень образования работника (работников).

Вид функции f может быть различным для разных видов и условий труда в линейном

$$y = a + b \cdot x \qquad (2)$$

или в любом нелинейном виде, где «a» по смыслу является некоторым начальным значением функции «y», а «b» является скоростью изменения y в зависимости от значений «x», в чем при вычислении производной от выражения (2) по аргументу «x» легко убедиться:

$$\frac{dy}{dx} = 0 + b \cdot \frac{dx}{dx} = 0 + b \cdot 1 = b. \qquad (3)$$

Размерность коэффициента «b» легко установить: размерность комплекса b·x как слагаемого с коэффициентом «a» измеряется в единицах переменной «y». Если в качестве зависимой переменной примем реализованную продукцию в стоимостном выражении тыс. руб. или т.р., а размерность «x» в годах обучения, то размерность всех компонентов уравнения (2) будет такой:

$$y \;=\; a \;+\; b \;\cdot\; x$$

$$[\text{т.р.}] = [\text{т.р.}] + \left[\frac{\text{т.р.}}{\text{год}}\right]\cdot[\text{год}]. \qquad (4)$$

Значит размерность коэффициента «b» как скорости изменения исследуемой функции «y» в контексте задачи будет иметь размерность [тыс. руб. / год] – будущий рост реализованной продукции на каждый год учебы.

Еще проще интерпретировать «b» из школьной физики по разделу равномерного прямолинейного движения. Преодоленный путь S (пусть 12 км) будет складываться из уже пройденного начального отрезка пути S_0 (2 км) и отрезка пути «v·t», где v – скорость движения (пусть 5 км / час) и t - время движения с данной скоростью (2 часа):

$$S = S_0 + \frac{dS}{dt}\cdot t = S_0 + v\cdot t = 2 + 5\cdot 2 = 12 \text{ (км)}. \qquad (5)$$

Таким образом, введя в рассмотрение с примером преодоления расстояния (5) при условии выполнения размерности, как это показано в (4), мы построили модель преодоления заданного расстояния – сначала в общем виде, а подставив на место символов конкретную числовую информацию, получили результат. Если при известных параметрах уравнения (5) S_0 = a - уже пройденного отрезка пути S_0 = 2 км, известной скорости b = v = 5 км / час и *заданным нами* временем нахождения в пути t = 2 часа получили результат y = S = 12 км – по аналогии с уравнением (2).

Уравнение (2) называется простым линейным уравнением, связывающим две переменные (или пару переменных) – зависимую «y» и независимую «x».

При этом мы сознательно не ставим вопроса о происхождении коэффициентов уравнения (5) в обозначениях уравнения (2), где $y = S$, $a = S_0$, $b = v$, $x = t$. Вполне возможно, что параметры уравнения (5) нами были получены в результате проведения собственных наблюдений: пройденные и будущие километры мы могли определить по числу сделанных шагов (если известна приблизительная длина шага), средняя скорость пешехода $v \approx 5$ км / час была нам известна из прошлого опыта преодоления расстояний, а время t наблюдали по своим часам.

Поэтому мы и решили (подозревая о существовании уравнения 5 или не подозревая), что преодолеем оставшееся расстояние в 10 км за 2 часа. Для этого нам хватило простого житейского опыта. Даже более того, зная, что скорость 5 км /час является по своей природе средней (реально – где-то при реальном передвижении выше ее, где-то на участке – ниже), да и расстояние измеряли шагами, постоянство длины которых на двенадцатикилометровом маршруте обеспечить проблематично, уравнение (5) более корректно должно быть записано как-то так: $S \approx S_0 + v \cdot t$.

Располагая уравнением такого вида, мы тем самым формализуем свой жизненный опыт в отношении понимания и прогнозирования разных ситуаций, связанных с передвижением шагом. Так, если начальный участок отсутствует ($S_0 = 0$), а необходимо преодолеть, скажем, те же 10 км, но со скоростью чуть меньшей (примерно 4 км / час), то из уравнения (5) легко определить прогнозное значение требуемого времени: потребуется приблизительно (10 / 4) = 2,5 часа.

Возможно также, что существуют такие области экономического знания, в которых приведенные выше рассуждения окажутся практически достаточными, и параметры уравнения (2) a и b, как и в аналогичном случае параметры уравнения (5) S_0 и v можно было бы оценить *на умозрительном уровне*. Однако в реальной практике такие ситуации встречаются не столь часто, как хотелось бы. Особенно при исследовании конкретных социально-экономических явлений и процессов. Какой реальной исходной информацией мы можем располагать?

Необходимо вспомнить известные нам источники статистической информации, основными из которых являются Росстат, Госкомстат РФ и РТ, ведомственная статистика, научная периодика, Интернет, СМИ, экспертная информация (если отсутствует какая-либо необходимая информация) и *результаты собственных наблюдений*, что рассмотрим подробнее. Предположим, нас заинтересовал вопрос о выяснении закономерности (если таковая в природе имеется) наличия у нас свободных суммы (в тыс. руб., например, что для студента немаловажно) на конец каждого календарного месяца. Да, мы располагаем некоторым практическим опытом, но имеющегося опыта для выявления требуемой нам закономерности (тренда по времени) оказывается явно недостаточным: этот опыт нами до сих пор не фиксировался, не формализовался. Таким образом, для достижения поставленной цели необходимо прежде всего создать статистическую базу из собственных фиксированных наблюдений, а уж потом как-то попытаться формализовать эти наблюдения в удобном для использования виде.

Самый простой и доступный способ создания необходимой статистической базы – непосредственные наблюдения с фиксаций значений сумм (в тыс. руб.), которые остаются у вас на последнее число календарного месяца (вряд ли такой информацией располагает Госкомстат РФ и РТ) в течение нескольких месяцев *при прочих равных условиях*. Последнее означает, что множество действующих факторов и условий вашего бытия за исследуемый период (скажем, с января по май) не претерпит существенных изменений как внешнего, так и внутреннего (личного) порядка.

При организации такого статистического наблюдения мы фиксируем оставшуюся у нас наличность, которую для простоты далее будем именовать «прибылью», лишь на последний день каждого месяца. Поэтому информация будет носить дискретный (моментный) характер, хотя в отдельные дня каждого месяца картина с «прибылью» будет носить неоднозначный характер: в какаието дни наличных сумм будет более, чем достаточно, в какие-то их дефицит (долги, например). В этом смысле ежемесячная дискретная картина от реальной непрерывной (почти непрерывной: в месяце – от 28 до 31 дня) может отличаться весьма заметно. Предположим, что нам это зафиксировать удалось, и наши наблюдения могут быть сведены в табл. 1:

Таблица 1

Результаты ежемесячных наблюдений

Номер наблюдения	Месяцы	Оставшиеся суммы, тыс.руб.
1	Январь	1
2	Февраль	1
3	Март	1
4	Апрель	3
5	Май	4

В первом приближении попытаемся наш опыт, зафиксированный в табл. 1 оценить на умозрительном уровне. Из табл. 1 видно, что первые три месяца сумма была постоянной, а затем стала увеличиваться – и это все, что мы можем сделать в виде умозаключения на момент исследования. Значит для выявления более содержательной закономерности необходимо прибегнуть к использованию каких-то иных инструментальных средств. Для этого обратимся к содержанию табл. 1 еще раз.

Итак, наш опыт зафиксирован, в результате чего мы располагаем двумя статистическими совокупностями, которые необходимо как-то *эксплицировать* или формализовать. Сначала заменим месяцы на соответствующие коды: январь – 1, февраль – 2 и т.д. (а можно и так: январь – 0, февраль – 1 и т.д.). Поскольку требуется выявить причинно-следственную связь между временем (в мес.) и оставшимися суммами на конец месяца (в тыс. руб.), присвоим им имена совокупностей $X = \{x_i\}$ и $Y = \{y_i\}$, где текущая переменная $i = 1, N = 5$

(штук) – соответственно. Хотя и время обычно обозначается символом t (time – время), для более общего случая лучше использовать традиционное обозначение независимого аргумента через «x».

После введения таких символов табл. 1 примет следующий вид:

Таблица 2

Эксплицированные результаты ежемесячных наблюдений

Номер наблюдения, i	Месяцы, мес. x_i	Оставшие-ся суммы, тыс.руб., y_i
1	$x_1 = 1$	$Y_1 = 1$
2	$x_2 = 2$	$y_1 = 1$
3	$x_3 = 3$	$y_1 = 1$
4	$x_4 = 4$	$y_1 = 3$
5	$x_5 = 5$	$y_1 = 4$

Располагая информацией в табл. 2 (впрочем, и в табл. 1) можно визуализировать ее на плоскости в координатах (x, y), но для дополнительного понимания исследуемого процесса это вряд ли прибавит что-то новое.

Другое дело, если бы мы в качестве искомого тренда (закономерности) по пытались бы заменить имеющеюся *эмпирическую* (опытную) дискретную информацию, представленную в табл. 2, в виде некоторой *теоретической* парной зависимостью между переменными «x» и «y», например, вида (2), то поставленная задача была бы решена. Такая замена дискретных данных, выраженных в общем случае рядом вещественных статистических показателей, называется *аппроксимацией*, которую можно осуществить для наглядности и на умозрительном уровне, как это представлено на рис. 1. Для этого попытаемся расположить нашу искомую прямую (ее график) так, чтобы он отстоял от эмпирических точек «y_i» от соответствующих им значений теоретических значений функции для одних и тех же «x_i» возможно ближе ко всем эмпирическим точкам одновременно (графическое решение поставленной задачи).

Понятно, что вариантов графического решения задачи существует бесчисленное множество, особенно с увеличением масштаба (цены делений на каждой координатной оси). Один из них может быть и таким, как на рис. 1.

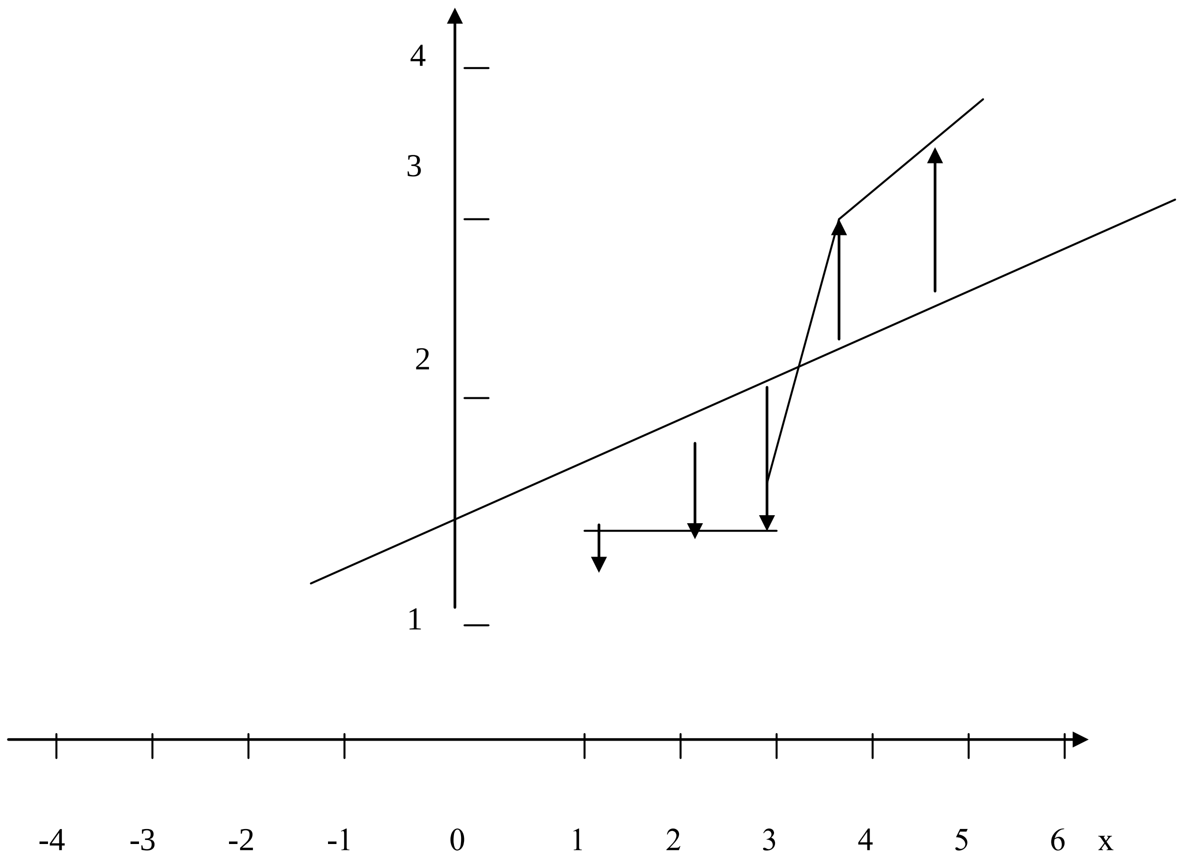

Рис. 1. Исходные данные и аппроксимирующая их прямая

И если наш эмпирический набор данных (см. табл. 1 и 2) имеет область изменения аргумента x [1; 5 мес.], а область определения зависимой от аргумента функции y [1; 4 тыс. руб.], то положение начерченной нами прямой, область определения которой [- ∞ мес.; + ∞ тыс. руб.], как нами было уже отмечено, далеко не единственно возможное.

Стрелками на рис. 1 показано расстояние между эмпирическими и возможными теоретическими значениями искомой аппроксимирующей функции вида (2). Чтобы отличать гипотетические теоретические значения от эмпирических, введем обозначение функции для теоретической прямой:

$$\hat{y}_x = a + b\cdot x. \qquad (6)$$

Вообще говоря, даже в рамках тех сведений, что нам о линейных функциях известно еще со школы, мы даже можем примерно оценить неизвестные коэффициенты линейной функции вида (6) a и b. Действительно, при x = 0 выражение (6) обращается в равенство $\hat{y}_x = a + b\cdot x = a + 0\cdot x = a$. Из графика на рис, 1 видно, что прямая пересекает ось ОУ в точке x = 0, для которой y ≈ 0,9 т.р. Величина коэффициента уравнения (2) «а» нами приблизительно оценена. Однако с даже с приблизительной оценкой величины коэффициента «b» несколько сложнее. Мы уже знаем, что он характеризует скорость изменения функции: чем больший наклон к оси абцисс графика функции, тем его скорость изменения выше, и которую можно измерить как отношение отрезков, которые «отсекает» прямая от начала координат по его осям. Это отношение как отношение катета, противолежащего углу (он равен ≈ 0,9 т.р.) к катету, принадлежащему углу (он равен ≈ 1,6 мес.), называется тангенсом угла φ, образованного нашей прямой и осью абцисс:

$$\mathrm{tg}\ \varphi \approx \frac{0{,}9}{1{,}6} = 0{,}56.$$

Со школы известно, что tg 45^0 = 1 (когда оба катета одинаковы по длине). А поскольку здесь меньше – всего 0,56, это значит, что наклон прямой к оси OX гораздо менее 45^0 , что мы и наблюдаем на рис. 1 (так называемый «грубый еонтроль»). Тогда уравнение искомой теоретической прямой, которая будет заменять нам исходные эмпирические данные табл. 1 и 2, примет вид:

$$\hat{y}_x \approx 0{,}9 + 0{,}56\cdot x. \qquad (7)$$

Как также известно еще со школы, чтобы построить график прямой на плоскости, достаточно наличие всего двух точек. Одна для формулы (7) уже есть при x = 0: y ≈ 0,9. Вторую точку надо взять подальше по оси OX, например, 4: y ≈ 0,9 + 0,56 · 4 = 3,14 (т.р.), что в целом соответствует положению начертанного нами графика прямой с учетом специфики рисования подобных графиков в windows.

Таким образом, параметры прямой можно определить, как и в случае выражения (5) и на умозрительном уровне. Однако возникает вопрос, насколько такие результаты надежны. Ведь мы, аппроксимируя таким вот графическим способом, положение прямой выбрали «на глаз», очень приближенно.

Поэтому для решения проблемы аппроксимации, и не только линейной зависимостью вида (6), разработан специальный метод, названный методом наименьших квадратов (МНК). Точнее, не только любой аппроксимации, которую только что реализовали мы сами без помощи данного метода, но *оптимальной аппроксимации*, такой, чтобы на графике, подобном рис. 1, новая прямая, созданная уже на основе МНК занимала *самое наилучшее положение* из бесконечного числа возможных.

1.2. Сущность метода наименьших квадратов

Сущность МНК можно сформулировать достаточно просто. Метод позволяет получить такие параметры аппроксимирующей линейной функции «a» и «b» в выражении (6), при которых сумма разностей квадратов отклонений дискретных значений функции (y_i) от значений функций, лежащей на прямой ($\hat{y}_{xi}$), то есть $(y_i - \hat{y}_{xi})^2$ при одних и тех же значениях аргумента (x_i), была бы минимальной. В соответствующих символах сказанное можно эксплицировать (формализовать) так:

$$S = \sum_{i=1}^{N} (y_i - \hat{y}_{xi})^2 \rightarrow \min. \qquad (8)$$

Таких сумм в выражении (8) будет N = 5, что можно рассмотреть их графическую интерпретацию на рис. 1. Стрелками на рис. 1 показано разности $(y_i - \hat{y}_{xi})$ так, что если знак такой разницы отрицателен, то стрелки направлены вниз, если положителен, то вверх. Если сложить эти суммы, не возводя в квадрат, то для какого-то случая можем в итоге вычисления (8) получить и нуль, когда значения разностей с плюсами и минусами скомпенсируют друг друга. Однако при возведении в квадрат, как это показано в формуле (8), все разности приобретают только положительные значения, которое (мы хотим), чтобы имело минимальное значение из всех возможных.

Для этого нам каким-то образом необходимо получить в итоге новые, оптимальные значения параметров прямой вида (6) «a» и «b». Для этого перепишем условие (8) с учетом вида уравнения прямой (6):

$$S = \sum_{i=1}^{N} (y_i - a - bx_i)^2 \rightarrow \min. \qquad (9)$$

Понятно, проведя график прямой линии на рис. 1, как уже отмечалось, «на глаз», в относительной близи к заданным нами же эмпирическим точкам, хотя бы приблизительно подсчитаем значение суммы площадей всех воображаемых «квадратиков», которые можно построить на их вертикальных сторонах, обозначенные отрезками со стрелками на рис. 1. Для вычисления площадей этих, мысленно представленных, «квадратиков» необходимо и достаточно возвести длину этих отрезков со стрелками в квадрат.

Так, для квадрата, построенного на приблизительной длине первой «стрелки», измеренной в соответствии с ценой деления оси ординат: ≈ (0,5 x 0,5) = $0{,}5^2$ = 0,25 (мес. · тыс. руб.); для второй стрелки ≈ $1{,}0^2$ = 1,00; для третьей ≈ $1{,}3^2$ = 1,69; для четвертой ≈ $0{,}4^2$ = 0,16 и для пятой ≈ $0{,}8^2$ = 0,64 (тыс. руб.)^2. В итоге из (9) получим значение нами построенной эмпирической суммы:

$$S = \sum_{i=1}^{N} (y_i - \hat{y}_{xi})^2 \approx 0{,}25 + 1{,}00 + 1{,}69 + 0{,}16 + 0{,}64 = 3{,}24. \qquad (10)$$

Понятно, что на минимальное значение из всех возможных сумм значением в 3,24 (мес.·тыс. руб.) мы претендовать не можем, но все же запомним это значение для будущего сравнения с оптимальным значением той же функции (8) в более подробном виде (9).

Из выражения (9) также видим, что сумма S как функция зависит от четырех аргументов-символов: $S = f(y_i, x_i, a, b)$. Однако значения y_i и x_i нам известны в качестве исходных данных (см. табл. 2), тогда как параметры искомой прямой (6) a и b – неизвестны. Формулы для их вычисления каким-то образом надо найти. Для этого необходимо вспомнить ранее пройденный школьный материал в качестве фрагмента математического анализа при поиске экстремума непрерывных функций, в основе которого положены, если кто помнит, теоремы Роля, Лагранжа и Коши.

Тогда содержание такого анализа можно продемонстрировать на следующем наглядном примере.

1.3. Пример нахождение экстремума при реализации МНК

Пример анализа функции на экстремум рассмотрим для наиболее наглядной для такого случая вида функции – параболической. Перед нами в данном случае стоят две задачи: определить аналитическим путем, есть ли у функции экстремум (максимум или минимум), а затем - оценить, что характеризует собой выявленный экстремум - максимум или минимум.

Пусть вид функции будет представлен в выражении (11), где a = 1; b = 3:

$$y = f(x) = a + (x - b)^2 = 1 + (x - 3)^2 . \qquad (11)$$

Для лучшей наглядности визуализируем (отобразим на графике) параболическую функцию (11), для чего зададимся значениями «x», а по этой формуле найдем соответствующие значения «y» и занесем их в рабочую табл.3.

Таблица 3

Рабочая таблица для графика функции $y = 1 + (x - 3)^2$

№	X	Y
1	1,0	5,0
2	2,0	2,0
3	3,0	1,0

4	4,0	2,0
5	5,0	5,0

Из рис. 2 (да и из табл. 3) очевидно, что, во-первых, экстремальное значение функции наблюдается при x = 3 и что значение функции при этом минимальное, то есть y_{min} = 1. Ясно, что при всех остальных значениях x значения функции y будут больше, чем 1.

Данные табл. 3 представим графически на рис. 1.

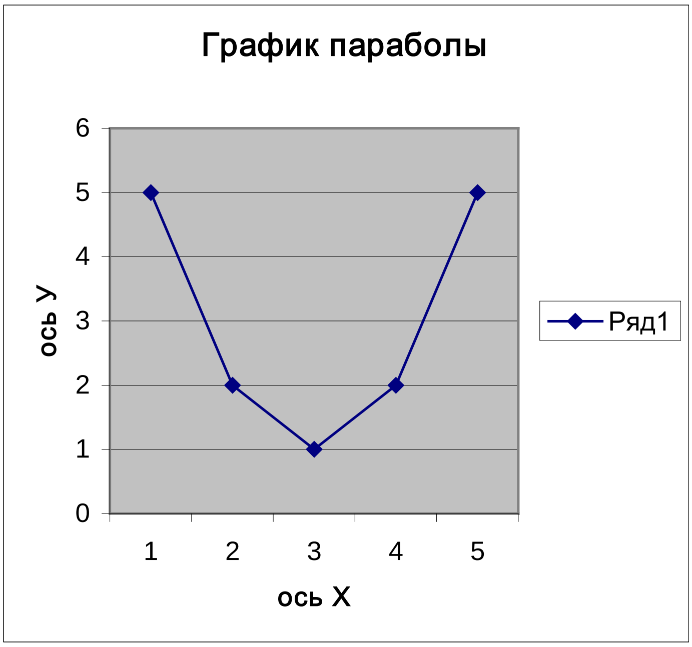

Рис. 2. График функции (11), исследуемой на экстремум

А теперь сделаем то же самое аналитически. Для этого формулу (11) необходимо продифференцировать, иначе говоря, найти первую производную по x по известному правилу дифференцирования степенных функций, что записывают так:

$y_x' = [1 + (x - 3)^2]_x'$ или более наглядно

$$\frac{dy}{dx} = \frac{d}{dx}[1 + (x - 3)^2] = 0 + 2\,(x - 3)^{2-1} = 2\cdot(x - 3)^1 = 2x - 6. \qquad (12)$$

Как мы помним, производная суммы равна сумме производных: производная постоянной величина – здесь «1» равна нулю, - а производная степенной функции равна произведению степени функции на ее основание в степени, уменьшенной на единицу, что и сделано в выражении (12). Однако само полученное выражение $y_x' = 2x - 6 = -6 + 2x$ является уравнением прямой вида $y = a + bx$, где $a = -6$, а $b = 2$. Эта прямая в общем случае не является параллельной оси абцисс (тогда бы $b = 0$), но, исходя из вида дифференцируемой функции (11) расположена под углом с тангенсом наклона 2 (этот угол явно больше 45^o).

Для того, чтобы установить, является ли функция экстремальной, полученное выражение для ее первой производной (12) достаточно *самим приравнять к нулю* (тем самым определяем точку, в которой касательная к дифференцируемой функции будет параллельна оси X. И в этой точке производная функции точно равна нулю (сами приравняли):

$$2x - 6 = 0 \qquad (13)$$

В формуле (13) имеем одно уравнение, линейное по отношению к аргументу x и одно неизвестное. Из (13) находим $x_{экстр} = 6 / 2 = 3$ (что видно и из содержания табл. 3). Подставляя $x_{экстр} = 3$ в выражение для параболы (11), получим:

$$y = f(x) = a + (x - b)^2 = 1 + (3 - 3)^2 = 1 + 0 = 1 = y_{экстр}. \qquad (14)$$

Итак, аналитическим путем мы установили, что экстремум степенной функции вида (11) располагается на плоскости в точке с прямоугольными координатами (x,y) = (3,1) – см. рис. 2.

Далее остается установить, является ли значение y = 1 при x = 3 для функции (11) *максимальным* или *минимальным* по известному «правилу зонтика»: если вторая производная (от первой производной) выражения (13) будет отрицательной, то функция (11) достигает в точке x = 3 свое *максимальное* значение; если же положительным, то функция (11) достигает в той же экстремальной точке *минимальное* значение.

$$\frac{d}{dx}\left(\frac{dy}{dx}\right) = \frac{d}{dx}(2x - 6) = \frac{d}{dx}(2 \cdot x^1) + \frac{d}{dx}(6) = 2 \cdot x^{1-1} + 0 = 2 \cdot 1 = 2 > 0. \qquad (15)$$

Значение второй производной (15) положительно. Следовательно, функция у при значении х = 3 достигает значения 1 еще по выражению (14) своего *минимального* значения.

Таким образом, функция (11) *имеет экстремум* в координатах (х, у) = (3; 1), и этот экстремум означает *минимум* этой исследуемой функции. Задача на поиск экстремума и его характер (максимум или минимум) решена.

Следовательно, располагая только лишь видом функции (11), даже не визуализируя ее ни в табл. 3, ни на рис. 2, можно определить ее экстремальное значение (если оно существует) и установить его характер. Рассмотрение возможного вопроса о единственности экстремума выходит за рамки данного фрагмента.

В данном модельном примере (11) коэффициенты а и b известны и равны 1 и 3 – соответственно, значения «х» задаем сами, значения функции «у» получаем из заданного вида (11). Однако в нашей постановке модельной задачи все прямо *наоборот*: х и у нам известны (см. табл. 1 и 2), известен вид искомой прямой, аппроксимирующей исходную дискретную эмпирическую информацию (у = а + bх), но неизвестны значения коэффициентов «а» и «b».

Решением задач в такой постановке и занимается *регрессионный анализ*. Если переменных всего две, одна из которых подразумевается аргументом (х), а другая функцией (у), то аппарат определения неизвестных коэффициентов носит названия *парной регрессии* или задачей аппроксимации исходных эмпирических данных в виде того или иного вида парной зависимости (здесь – линейной) методом наименьших квадратов (МНК).

Если же аргументов более одного, например, $y = f(x_1, x_2)$, то подобная задача решается в терминах *множественной регрессии* вида $y = a + b_1x_1 + b_2x_2$.

В этой связи рассмотрим механизм парной регрессии, широко используемый при формализации в различных областях знания, в том числе и в социально-экономических исследованиях. Механизм будет рассмотрен на примере линейной парной зависимости не столько потому, что она наиболее проста, наглядна, хорошо интерпретирует социально-экономические исследования, но и служит средством линеаризации всех остальных видов парных зависимостей при применении к ним МНК, как будет показано далее.

2. РЕГРЕССИОННЫЙ АНАЛИЗ НА ОСНОВЕ МНК

2.1. Особенность решения задачи оптимальной аппроксимации

Как нами ранее было отмечено, сущность регрессионного анализа, в отличие, скажем, от общего функционального анализа (это более широкое понятие) состоит в том, что нам известны (здесь – в результате собственных наблюдений) эмпирические точечные (дискретные) значения элементов статистических совокупностей, одну из которых мы принимаем в качестве аргумента, а другую - в качестве функции. Затем мы хотим то и другое связать (аппроксимировать) некоторой непрерывной функциональной зависимостью, например, в виде простой линейной функции вида $y = a + bx$.

Такую связь мы и попытались установить. Пусть и не на научном, а лишь на *умозрительном уровне*, что и проиллюстрировали для тех же наших исходных данных на рис. 1. В конечном итоге мы даже выявили приблизительное графо-аналитическое решение вида (6), определив к тому же еще и его, ранее неизвестные нам, параметры в формуле (7): $\hat{y}_x \approx 0{,}9 + 0{,}56 \cdot x$. И даже оценили значение левой части требования МНК вида – сумму разностей квадратов (8) или (9), насколько это позволил графо-аналитический метод, которое приблизительно равно 3,24 (мес.· тыс. руб.). При этом нами было проигнорировано требование МНК, выраженное в правой его части, а именно требование «→ min» по причине невозможности его реализации на том же умозрительном уровне.

Хотя, как сказать: если бы мы запаслись множеством рисунков типа рис, 1 с различным положением *гипотетической* аппроксимирующей прямой на плоскости с попутным вычислением площадей новых вариантов «квадратиков», то из всего множества рисунков могли бы выбрать такой, сумма вида (9) или (10) на котором бы имела наименьшее (не теоретически минимальное!) эмпирическое значение. А затем графо-аналитическим путем получить приблизительные искомые значения коэффициентов линейного уравнения вида (6). Мало того, что такой опытный (эмпирический) путь получения нужного результата (замены дискретных опытных значений аналитической функцией линейного вида) излишне затратен по времени, он еще и не гарантирует необходимой точности решения. Хотя в качестве первого приближения и он отражает суть предстоящих преобразований дискретной информации в непрерывную (аналитическую, теоретическую).

Следовательно, МНК выступает *теоретической альтернативой* нашему практическому графо-аналитическому способу достижению поставленной цели – определению параметров выбранного вида аппроксимирующей функции (решение задачи спецификации, о чем будет изложено далее). Для регрессионной парной зависимости линейного вида (6) эти параметры называются: «a» - свободный член уравнения парной регрессии (начальное значение аппроксимирующей функции в начале прямоугольных координат); «b» - коэффициент регрессии (скорость изменения регрессионной линейной функции), что хорошо интерпретируется в социально-экономических исследованиях.

2.2. Получение параметров уравнения парной регрессии

С целью нахождения параметров парной регрессии как средства аппроксимации дискретной информации, созданной нами и эксплицированной в виде табл. 2, необходимо реализовать идею МНК, записанной в виде приближенного выражения (10). То есть неизвестные параметры искомой линейной зависимости a и b, следуя только что рассмотренному решению задачи на экстремум, пусть даже со степенной функцией (11), необходимо:

1) продифференцировать выражение (9) для нахождения первой производной сначала по одному неизвестному параметру функции (6) «a»;

2) приравнять полученные результаты дифференцирования к нулю;

3) путем алгебраических преобразований сделать так, чтобы члены полученного выражения распределились в нем так, чтобы члены, содержащие неизвестные параметры, расположились в левой части равенства, а известные – в правой;

4) продифференцировать выражение (9) для нахождения первой производной по другому неизвестному параметру функции (6) «b»;

5) приравнять полученные результаты дифференцирования к нулю;

6) путем алгебраических преобразований сделать так, чтобы члены полученного еще одного выражения распределились в нем так, чтобы члены, содержащие неизвестные параметры, расположились в левой части равенства, а известные – в правой;

Поскольку искомая S является функцией от двух неизвестных: S = f (a, b), будем вычислять не обычные, а *частные производные*. Сначала находим частную производную по неизвестному параметру «a» (для удобства нижние и верхние пределы вычисления сумм, а он все вычисляются от 1 до N = 5, при описании опускаем, опускаем и сами подстрочные индексы переменных «i»):

$$\frac{\partial S}{\partial a} = 2\,\Sigma\,(y - a - bx) \cdot (0 - 1 - 0) = 2\,\Sigma\,(- y + a + bx) = 0. \qquad (16)$$

Все, что расположено в левой части (16), получилось в результате нахождения частного дифференциала, а то, что справа, это *сделали мы сами* для вычисления искомого экстремума. Из алгебры известно, что если произведение двух сомножителей рано 0, то либо они оба, либо один из сомножителей равен 0. Понятно, что 2 нулю не равно (2≠ 0), следовательно нулю равно оставшийся множитель в выражении (16)

$$\Sigma\,(y - a - bx) = 0. \qquad (17)$$

Преобразуем выражение (17) по правилу применения знака Σ следующим образом.

$$\Sigma\,(y - a - bx) = \Sigma y - \Sigma a - \Sigma b = 0; \quad \Sigma y = \Sigma a + b\,\Sigma x = a\,\Sigma + b\,\Sigma x.$$

Поскольку $\sum_{i=1}^{N} = \sum_{i=1}^{N} 1 = (1 + 1 + \ldots + 1) = N$, то выражение (17) примет вид:

$$aN + b\,\Sigma x = \Sigma y. \qquad (18)$$

В выражении (18) члены уравнения, содержащие неизвестные a и b расположились слева от знака равенства, а известный член Σy расположился справа. То, что он известен для данного модельного примера, можно легко убедиться из табл. 2: Σy = 1 + 1 + 1 + 3 + 4 = 10 (тыс. руб.). Если Σx = 1 + 2 + 3 + 4 + 5 = 15, а N = 5, то выражение (18) для конкретного случая приняло бы вид: a5 + b15 = 10, но нас интересует не какой-нибудь частный случай, а именно решение *в общем виде*, так, как это записано в выражении (18). То есть мы хотим решить задачу в общем виде *один раз* с получением формул для вычисления неизвестных параметров a и b при *любых* исходных данных, как это представлено в табл. 1 и 2. Вторая производная по «а» от (18) равна N > 0. Значит найден минимум.

К тому же уравнение (18), в которые неизвестные входят линейно (в первой степени), является линейным относительно неизвестных a и b, имеет бесчисленное множество решений. Если мы получим в результате каких-то операций второе уравнение такого же вида, то два уравнения с двумя неизвестными имеют лишь одно решение.

Для получения еще одного уравнения проделаем такие же операции по отношению к еще одной неизвестной величине – коэффициенту b.

$$\frac{\partial S}{\partial b} = 2\,\Sigma\,(y - a - bx)\,(0 - 0 - x) = 2\,\Sigma\,(-\,yx + ax + bx^2) = 0. \qquad (19)$$

Далее рассуждаем как прежде. Если в выражении (19) $2 \neq 0$, то нулю остальная часть равенства (19):

$\Sigma\,(-\,yx + ax + bx^2) = \ \Sigma yx - \Sigma ax - \Sigma\, bx^2 = 0$, откуда

$$a\,\Sigma x + b\,\Sigma x^2 = \Sigma xy. \qquad (20)$$

Вторая производная по «b» от (20) равна $\Sigma x^2 > 0$. Найден минимум. Полученные линейные уравнения (18) и (20) составляют систему двух уравнений (21) с двумя неизвестными, коэффициентами a и b, а это, в свою очередь, означает, что данная система уравнений имеет *единственное решение,* которую запишем в классическом виде:

$$\begin{cases} a N + b \Sigma x = \Sigma y, \\ a \Sigma x + b \Sigma x^2 = \Sigma xy. \end{cases} \qquad (21)$$

Решение системы уравнений (21) может быть осуществлено как методом подстановки, когда одно неизвестное выражается через другое, или методом Крамера (метод определителей), а также матричным методом. Заметим, однако, что применение первых двух способов оправдано лишь в случаях, когда число неизвестных не превышает трех. Матричный метод – наиболее универсальный, и именно он используется в вычислительных процедурах на ЭВМ средствами пакетов прикладных программ (ППП), что рассмотрим несколько ниже.

Поскольку у нас имеется выбор в методах решения системы уравнений (21), воспользуемся методом определителей Крамера как наиболее наглядным, для чего перепишем систему уравнений в следующем виде.

(для a) (для b) (для правых частей системы уравнений)

$$\begin{vmatrix} N & \Sigma x \\ \Sigma x & \Sigma x^2 \end{vmatrix} = \begin{vmatrix} \Sigma y \\ \Sigma xy \end{vmatrix}. \qquad (22)$$

По левому определителю (22) вычислим главный определитель по известной перекрестной схеме, а также частные определители по известным правилам, когда столбцы при соответствующих неизвестных замещаются правыми частями выражения (22):

$\Delta = N \Sigma x^2 - (\Sigma x)^2$,

$\Delta_a = \Sigma y \Sigma x^2 - \Sigma x \Sigma xy$,

$\Delta_b = N\,\Sigma xy - \Sigma x\,\Sigma y.$

Тогда искомые значения коэффициентов a и b будут следующими:

$$a = \frac{\Delta_a}{\Delta} = \frac{\Sigma y\,\Sigma x^2 - \Sigma x\,\Sigma xy}{N\,\Sigma x^2 - (\Sigma x)^2}, \tag{23}$$

$$b = \frac{\Delta_b}{\Delta} = \frac{N\,\Sigma xy - \Sigma x\,\Sigma y}{N\,\Sigma x^2 - (\Sigma x)^2}, \tag{24}$$

Если коэффициенты регрессии a и b по выражениям (23) и (24) вычислены корректно (правильно), то в этом легко убедиться по выполнению тождества (25), иллюстрирующего тот факт, что если мы подставим среднее значение x, то при верно найденных коэффициентах получим среднее значение y, в чем и проявляется сущность уравнений регрессии, которые они связывают средние значения исследуемых переменных, здесь – y и x:

$$y_{ср} \equiv a + b\,x_{ср}\,. \tag{25}$$

Далее рассмотрим процесс нахождения величин коэффициентов линейной функции вида (6) средствами МНК (9) на модельных исходных данных (см. табл. 1 и 2).

2.3. Процесс нахождения параметров уравнения линейной регрессии

Как следует из содержания предыдущего подраздела, для вычисления искомых параметров a и b необходимо в общем случае произвести вычисления по формулам (23) и (24), подставляя в них соответствующие элементы исходных данных, которые выражены в числе пар данных N (см. таблицу исходных данных – содержание табл. 2), а также суммы: Σx , Σy , Σxy и Σx^2 .при изменении текущей переменной i от нижнего предела = 1 до верхнего предела

N = 5 (шт.). Для этого необходимо исходную базу на основе табл. 2 несколько расширить так, как то показано в табл. 4.

Следовательно, нахождение параметров уравнения парной линейной регрессии вида (6) сопряжено с дополнительной подготовке исходных данных (дополнительно вычислим и сумму Σy^2 , которая понадобиться в дальнейших вычислениях) и вычислений по формулам (23) и (24).

Таблица 4

Эмпирические данные и промежуточные вычисления

i	x_i	y_i	$x_i\ y_i$	x_i^2	y_i^2
1	1	1	1	1	1
2	2	1	2	4	1
3	3	1	3	9	1
4	4	3	12	16	9
5	5	4	20	25	16
N = 5	$\Sigma x_i = 15$	$\Sigma y_i = 10$	$\Sigma x_i\ y_i = 38$	$\Sigma x_i^2 = 55$	$\Sigma y_i^2 = 28$

Опуская для большей наглядности индексы и подставляя данные последней строки табл. 4 в выражения (23) и (24), получим:

$$a = \frac{\Sigma y\ \Sigma x^2 - \Sigma x\ \Sigma xy}{N\ \Sigma x^2 - (\Sigma x)^2} = \frac{10 \bullet 55 - 15 \bullet 38}{5 \bullet 55 - (15)^2} = \frac{550 - 570}{275 - 225} = \frac{-20}{50} = -0{,}4; \qquad (26)$$

$$b = \frac{N\ \Sigma xy - \Sigma x\ \Sigma y}{N\ \Sigma x^2 - (\Sigma x)^2} = \frac{5 \bullet 38 - 15 \bullet 10}{5 \bullet 55 - (15)^2} = \frac{190 - 150}{275 - 225} = \frac{40}{50} = +0{,}8. \qquad (27)$$

Искомые коэффициенты a (тыс. руб.) и b (тыс. руб./мес.) найдены, и символьное выражение (6) с найденными коэффициентами примет явный вид:

$$\hat{y}_x = -0{,}4 + 0{,}8 \cdot x. \qquad (28)$$

Осуществим поверку (верификацию) корректности вычисления коэффициентов уравнения линейной регрессии по выражению (28). Если коэффициенты уравнения a и b нами найдены без ошибок в вычислениях, тождество (25) будет соблюдено,

Как видим, тождество (25) выполняется:

$2 \equiv -0{,}4 + 0{,}8 \cdot 3 = -0{,}4 + 2{,}4 = 2$, где по табл. 4:

$y_{ср} = (1/N) \cdot \sum y_i = (1/5) \cdot 10 = 2$ (тыс. руб.); $x_{ср} = (1/N) \cdot \sum x_i = (1/5) \cdot 15 = 3$ (месяца), что с учетом единиц измерения будет так:

2 (тыс. руб.) $\equiv -0{,}4 + 0{,}8 \cdot 3 = 2$ (тыс. руб.) при размерностях

$$[\text{тыс. руб.}] = [\text{тыс. руб.}] + \left[\frac{\text{тыс. руб.}}{\text{месяц}}\right] \cdot [\text{месяц}].$$

Следовательно, коэффициенты регрессии a – *свободный член уравнения линейной парной регрессии* и b – *коэффициент регрессии* (скорость изменения линейной функции в (тыс. руб. / мес.) найдены верно. Иначе говоря, располагая *исходными дискретными эмпирическими данными*, приведенными в табл. 2 (2 и 3 столбцы), мы, воспользовавшись МНК, в качестве аппроксимирующей *непрерывной функции* выбрали линейную парную зависимость вида (6) и получили в итоге ту же зависимость (28), но уже с конкретными значениями искомых коэффициентов a = - 0,4 и b = 0,8.

Следовательно, по своему физическому смыслу коэффициент «a» (тыс. руб.) характеризует точу на оси прибыли ОУ в начале координат при x = 0 (может быть положительным и отрицательным, а также равным нулю – когда аппроксимирующая прямая проходит через начало координат). Коэффициент «b» (тыс. руб. / мес.) характеризует *скорость изменения функции* «y» относительно аргумента «x» (может быть положительным – прибыль возрастает; может быть отрицательным – прибыль уменьшается; может быть равным нулю – прибыль не изменяется). Для последней ситуации прямая будет параллельной оси ОХ, а исходное уравнение (6) приобретет вид: $y = a + 0 \cdot x = a$.

В геометрической интерпретации, как уже упоминалось, коэффициент b = tg φ , где φ – угол, образованный между осью OX и найденной прямой.

2.4. Графическое отображение оптимальной линейной функции. Прогнозирование

Для того, чтобы отобразить график полученной линейной функции (27) на плоскости, необходимо наличие двух любых точек, принадлежащих этому уравнению. Так, задаваясь значением аргумента x = 0 (мес.), получим по (27) значение y = - 0,4 (тыс. руб.); одна точка с координатами (x;y) = (0; - 0,4) определена. Вторую точку для большего удобства при построении возьмем много правее: пусть x = 5 (мес.), тогда по (27) y = 3,6 (тыс. руб.); вторая точка с координатами (x;y) = (5; 3,6) также определена. Поскольку новая линейная функция определена в области от - ∞ до + ∞ , продолжим немного начертанную прямую в обе стороны.

Отобразим также разности (y_i - $\hat{y}_{xi}$) из формулы для МНК между новыми теоретическими значениями и исходными дискретными значениями для одних и тех же x_i насколько позволяет масштабы, принятые по осям координат (две последние разности вследствие малости геометрических размеров помещать на график не будем). Новые графические построения при прежних исходных данных приведены на рис. 3.

Поставленная задача аппроксимации исходной информации (см. табл. 2) в виде линейного уравнения парной регрессии (28) решена и получила графическую интерпретацию на рис. 3, на котором, в частности, представлено действие принципа МНК, записанного в виде (8).

Обратим также внимание на то, что области определения эмпирических исходных данных в табл. 1 и 2 ограничено наблюдениями от 1-го до 5-го месяцев, а прибыли от 1 до 4 тыс. руб. Но график непрерывной прямой, аппроксимирующей наши дискретные исходные данные, располагается от минус бесконечности до плюс бесконечности. На рис. 3 ее график пересекает ось ординат в точке приблизительно равной величине свободного члена линейного уравнения a = - 0,4 тыс. руб.

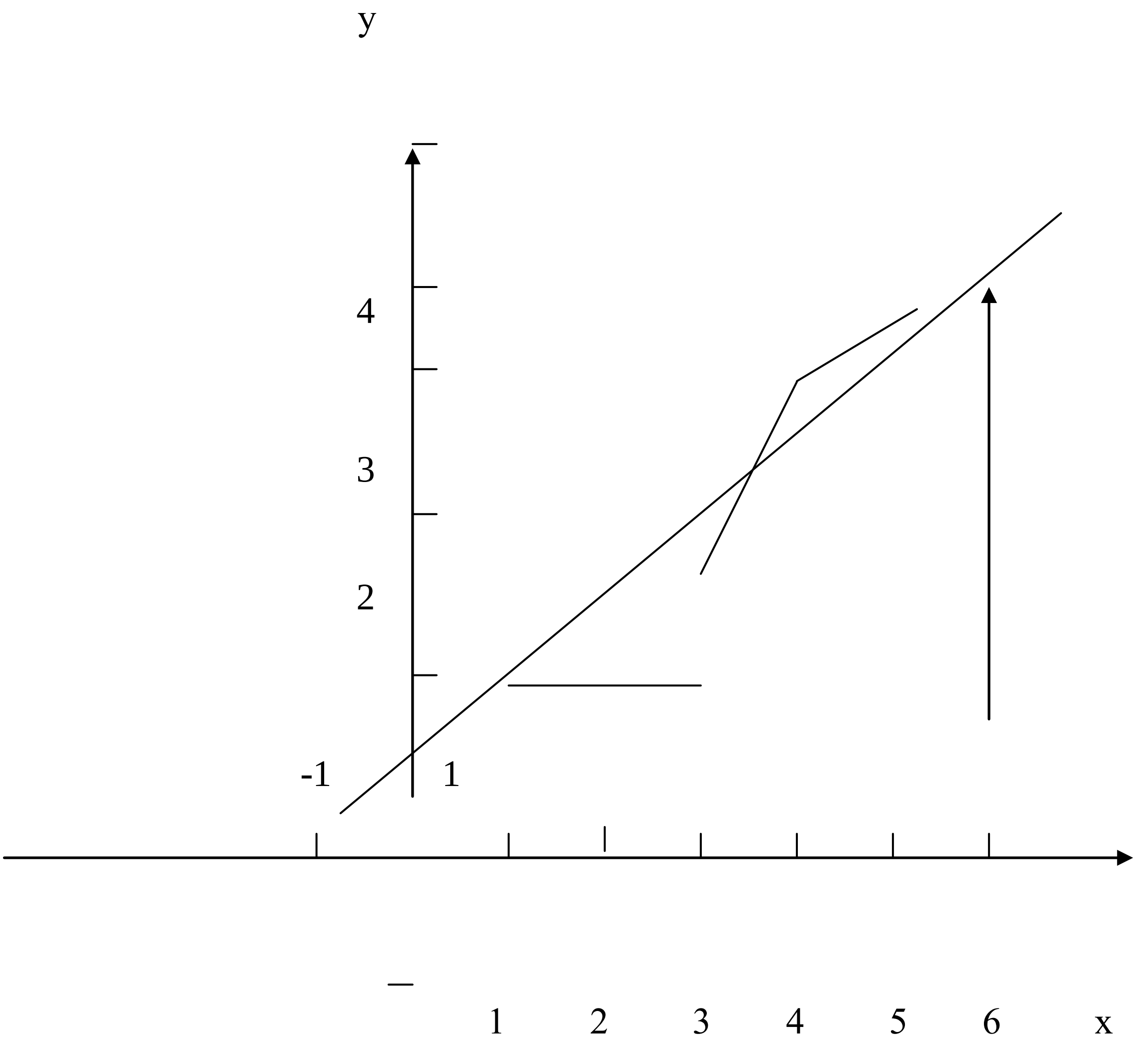

Рис. 3. Уравнение прямой при оптимальных значениях параметров a и b

По аналогии с прежними графо-аналитическими построениями приблизительно оценим значение суммы квадратов разностей исходных (дискретных эмпирических) данных и аппроксимированных непрерывных (теоретических) результатов при дискретных значениях аргумента x_i. Для этого, как и прежде, оценим сумму площадей квадратов, чтобы не перегружать рисунок, мысленно построенных на отображенных на рис. 4 разностях. Вертикальная сторона первого квадрата равна приблизительно 0,7 тыс. руб. Тогда площадь квадрата будет $0,3^2 = 0,49$ (мес.· тыс. руб.); второго $0,0^2 = 0,00$; третьего $0,7^2 = 0,49$; четвертого $0,50^2 = 0,25$; пятого $0,25^2 = 0,06$.

Тогда сумма площадей квадратов из выражения для МНК (8)

$$S = \sum_{i=1}^{N} (y_i - \hat{y}_{xi})^2 \approx 0,49 + 0,00 + 0,49 + 0,25 + 0,06 = 1,29. \qquad (29)$$

Сравнивая новое значение суммы 1,29 (мес.·тыс. руб.) из выражения (28) с прежним выражением (7), где е значение было равно 3,24 (мес.·тыс. руб.) на этапе попытки решения поставленной задачи на умозрительном уровне, констатируем, что, во-первых, значение $1,29 < 3,24$, а во-вторых, значение суммы $S \approx 1,29$ и есть минимально возможное согласно требованию МНК (8) при прочих равных условиях. То есть при аппроксимации данного расположения дискретных точек (см. табл. 2) линейным уравнением (28) сумма (8), приблизительно равная 1,29 мес.·тыс. руб. есть минимально возможная. Все остальные варианты сумм при ином расположении аппроксимирующей прямой будут превышать 1,29 по величине. Иначе говоря, $S = 1,29$ – есть *минимально возможная сумма*, что и отражено в выражении МНК (8): $S \to \min$. Меньше суммы площадей «квадратиков», чем 1,29 при наших исходных данных и аппроксимации их в виде линейной зависимости достичь принципиально невозможно.

Этот факт можно установить лишь теоретически, тогда как разница в суммарных площадях обоих подходов (умозрительного и теоретического) вполне очевидна при сравнении содержания рис. 1 и рис. 3. То есть эти выводы при сравнении графической информации становятся очевидными даже на обыденном уровне.

С позиции формализованного (научного) подхода на основе МНК суммы площадей квадратов (29) и (10) различаются вследствие различий коэффициентов регрессии a и b в уравнениях, аппроксимирующих исходную дискретную информацию (см. табл. 2) в виде линейных уравнений парной регрессии (7): $\hat{y}_x \approx 0{,}9 + 0{,}56 \cdot x$ и (28) вида $\hat{y}_x = -\ 0{,}4 + 0{,}8 \cdot x$. Причем различия коэффициентов регрессии выражено достаточно убедительно: вместо $a = 0{,}9$ должно быть $a = -\ 0{,}4$; вместо $b = 0{,}56$ должно быть $b = 0{,}80$. Уравнение парной линейной регрессии (7) было получено нами для наглядности при оценке разницы между оптимальным (28) и любым другим положением прямой на плоскости.

Таким образом, на данном этапе исследования, располагая видом аппроксимирующей функции (28) $\hat{y}_x = -\ 0{,}4 + 0{,}8 \cdot x$ и содержанием рис. 3, можно осуществить построение некоторых видов прогнозов.

3. ВОПРОСЫ ПРИМЕНЕНИЯ РЕГРЕССИОНОГО АНАЛИЗА

3.1. Получение точечных прогнозов

Анализ поставленной проблемы осуществляется в тех или иных границах, представленных исходными данными. На примере решаемой здесь задачи первичные наблюдения осуществлялись с января по май (см. табл. 1). Тогда период ретроспекции занимает положение по оси OX с 1-го месяца по 5-й: $T^{ретро} = 5$. Прогнозирование является составной частью принятия управленческих решений и состоит в оценке исследуемого социально-экономического явления или процесса с течением времени, называемом периодом упреждения $T^{упр}$. Для определенности возьмем период упреждения, равного одному месяцу: $T^{упр} = 1$ мес. Тогда получим прогноз на шестой месяц – на июнь.

При построении прогнозов предполагается, что период ретроспекции намного превышает период упреждения: $T^{ретро} >> T^{упр}$ или, что то же самое, отношение их ($T^{ретро} / \ T^{ретро}$) достаточно велико (для нашего примера $T^{ретро} = 5$ месяцам – с января по май). Если нам требуется построить точечный прогноз на июнь (6-й месяц в терминах табл. 2), то соотношение обоих периодов будет 5 : 1. Нами же на периоде ретроспекции получено уравнение парной линейной регрессии (28), представленное на рис. 3. Поэтому оценка прогнозного значения на июнь (6-й месяц) может быть осуществлена двумя способами: аналитически и графо-аналитическим.

При аналитическом способе мы в уравнение (28) задаем значение x = 6, тогда

$\hat{y}_x = - 0,4 + 0,8 \cdot 6 = - 0,4 + 4,8 = 4,4$ (тыс. руб.).

Графо-аналитическое решение показано на рис. 3 вертикальной стрелкой, которая начинается у цифры «6» на оси абцисс и пересекается с прямой регрессии. Длина этой стрелки, параллельной оси ординат, и является результатом графо-аналитического точечного прогноза ≈ 4,5 (тыс. руб.) на рис. 3 (несоответствие с результатами аналитического прогноза в данном случае можно объяснить особенностями графических построений в среде windows).

Графо-аналитическое решение при защите самостоятельной работы на данную тему целесообразно приводить как для приблизительного контроля аналитического прогноза по формуле (28), так и для лучшего уяснения решения задач прогнозирования экстраполяционным способом, когда результаты моделирования на периоде ретроспекции распространяются дальше, на величину шага упреждения $T^{упр}$ (здесь шаг равен одному месяцу).

Здесь же отметим то, что большинство анализируемых социально-экономических процессов в выпускных квалификационных работах (ВКР) рассматриваются во времени. Тогда предложенные для последующего анализа исходные дискретные данные (табл. 1 и 2) представляют собой ни что иное, как динамический (временной) ряд. Временные ряды подразделяются на интервальные (когда по оси абцисс, на которой отображается и масштабируется время «t», именованное в нашем изложении как более привычная для студентов переменная «x» представлено интервалами) и моментные (время представлено в виде дискретных значений, как в нашем примере). Причем те и другие могут на той же оси масштабироваться с постоянным (как в нашем примере) или с переменным шагом. Значения переменной «y» называются «уровнями ряда».

В общем случае при проведении анализа динамических рядов исследуется не только тренд (тенденция) Т (в нашем случае нами задана линейная тенденция или тренд в виде формулы 6), но и наличие или отсутствие циклической составляющей S и автокорреляция остатков E:

$$\hat{y}_x = T + S + E. \qquad (30)$$

В нашем случае Т = f (x) = a + bx, тогда как в общем случае вид аппроксимирующей функции (тренд) может быть выражен через решение *задачи спецификации* (подбора вида аппроксимирующей функции так, чтобы относительная ошибка аппроксимации – см. ниже, - была бы наименьшей), содержание которой подробно рассматривается в рамках учебной дисциплины «Эконометрика». В курсе «Статистика» традиционно рассматривается лишь вид и параметры тренда Т, а также некоторые дополнительные статистические характеристики, которые рассмотрим далее.

Таким образом, после того, как искомые коэффициенты уравнения линейной регрессии a и b найдены и точечный прогноз аналитическим и графо-аналитическим методами построен, для завершения поставленной задачи остается определить точность аппроксимации и оценить наличие или отсутствия достоверной статистической связи между изучаемыми явлениями (здесь – между оставшимися в конце месяцев суммами в тыс. руб. и временем, выраженном в месяцах).

Таким образом, процедуры для завершения проводимого анализа и построения аппроксимирующей зависимости также формализуются, и для их реализации достаточно проведения последующих вычислений, а также построения соответствуюших суждений и умозаключений при поверке содержания выдвигаемой нами же «нулевой гипотезы».

3.2. Оценка точности аппроксимации

Различия в представлении эмпирической зависимости по отношению к моделирующей ее аналитической функции оценивается коэффициентом аппроксимации ε, который вычисляется как среднее модуля величин эмпирических и теоретических значений, отнесенных к соответствующим эмпирическим значениям. Относительная ошибка аппроксимации эмпирической зависимости с помощью линейной функции:

$$\varepsilon = \frac{1}{N}\sum\left|\frac{y_i - \hat{y}_{xi}}{y_i}\right| \cdot 100\%. \qquad (30)$$

Здесь в выражении (31) значения y_i мы берм из табл. 2, а значения $\hat{y}_{xi}$ получаем из формулы (28). Схема вычисления суммы в выражении (30) приведена в рабочей табл. 5.

Таблица 5

Рабочая таблица для вычислении суммы выражения (30)

i	y_i	$\hat{y}_{xi}$ по (27)	$\lvert y_i - \hat{y}_i \rvert$	$\lvert y_i - \hat{y}_i / y_i \rvert$
1	1	0,4	0,6	0,60
2	1	1,2	0,2	0,20
3	1	2,0	1,0	1,00
4	3	2,8	0,2	0,06
5	4	3,6	0,4	0,13
				$\sum \lvert y_i - \hat{y}_i / y_i \rvert = 1{,}99$

Тогда по формуле (30) получим:

$$\varepsilon = \frac{1}{5} \cdot 1{,}99 \cdot 100\% = 38{,}6\%,$$

что, конечно, свидетельствует о далеко не малой ошибке (предпочтительной считается ошибка, не превышающая 10%). Обычно аппроксимирующую функцию на этапе спецификации (выбор вида аппроксимирующей зависимости – степенной, показательной, логарифмической и др.) подбирают так, чтобы ошибка аппроксимации не превышала единиц процентов. В данном случае задачу аппроксимации методом наименьших квадратов было бы целесообразно решить еще раз, скажем, для степенной или показательной функции. Однако вычисления по сравнению с линейной функцией намного усложнятся вследствие необходимости предварительной линеаризации и последующей делинеаризации аппроксимирующей функции.

При использовании стандартных компьютерных пакетов прикладных программ (ППП) «Статистика», «Статграфик», «SPSS» и др., как уже упоминалось, от пользователя требуется лишь занесение исходных данных. Затем выбранной компьютерной программой они используются всеми имеющимися парными функциями различного вида, и пользователю предоставляется возможность выбрать ту функцию (не обязательно линейную), относительная ошибка аппроксимации которой является минимальной среди имеющегося встроенного банка функций $\hat{y}_i$..

Набор таких функций, кроме уже рассмотренной линейной, может быть любым. Вот некоторые виды парных зависимостей

1) степенная $\hat{y}_x = ax^b$,

2) гиперболическая $\hat{y}_x = a + \dfrac{b}{x}$,

3) показательная $\hat{y}_x = ab^x$,

4) логарифмическая $\hat{y}_x = a + b\lg x$

5) параболическая $\hat{y}_x = a + bx + cx^2$

Понятно, что рядом с исходными данными на плоском графике какие-то из графиков парных зависимостей будут более приближенными к исходным точкам в тех же координатах, какие-то – менее приближенными к ним. Поэтому подбор вида парных зависимостей (иначе говоря - решение задачи *спецификации*) является одним из основных этапов решения задачи замены дискретных данных теоретической зависимостью, которая, в свою очередь, называется решением задачи *аппроксимации*.

Осуществляя такую аппроксимацию на интуитивном уровне, можно легко убедиться в том, что провести, например, прямую линию вида (6) так, чтобы она максимально близко пролегала вблизи эмпирических точек – далеко не всегда возможно. Поэтому стараются подобрать такие парные зависимости, которые лучше аппроксимируют исходную статистическую информацию в виде табл. 2.

При этом специализированные компьютерные программы, например для выбранной степенной парной зависимости вида $\hat{y}_x = ax^b$, сначала линеаризуют, применяют метод МНК вида (8), а затем делинеаризуют с выводом формул для вычисления коэффициентов регрессии a и b. Сам процесс начальной линеаризации сводится к тому, чтобы придать выбранной парной зависимости линейный вид. Так, для выбраннного нами вида степенной функции при решении задачи спецификации линеаризуется весьма просто путем логарифмирования (по любому основанию) ее обеих частей:

$\lg(\hat{y}_x) = \lg(a) + b \cdot \lg(x)$.

В полученном уравнении роль функции и аргумента выполняют их логарифмы, то есть какие-то числа, и оба неизвестных a и b будут в первой степени, а значит и уравнение – линейным относительно их и т.д. Затем к такому уравнению применяют МНК, а результаты делинеаризируют.

Обычно пакеты программ перебирают все имеющиеся в их базе виды парных зависимостей и рекомендуют исследователю одну из них либо по минимальной среди возможных величине ошибки аппроксимации вида (31), либо по максимальному коэффициенту детерминации (33), который формально является квадратом ранее полученного коэффициента линейной корреляции. Однако по существу (если продолжить аналитические расчеты и выводы) коэффициент детерминации отражает отношение объясненной дисперсии по отношению к аппроксимирующей функции к общей дисперсии, что также рассматривается в учебной дисциплине «Эконометрика».

3.3. Вычисление коэффициента линейной корреляции

Коэффициент линейной корреляции ρ, в отличие от подобного ему по содержанию коэффициента парной ранговой (порядковой) корреляции, может быть вычислен по нескольким, аналогичным по сути, формулам, однако воспользуемся следующей (наиболее простой , на наш взгляд, для последующих вычислений. – Авт.):

$$\rho = \frac{N \sum xy - \sum x \sum y}{\{[N \sum x^2 - (\Sigma x)^2]\,[N \sum y^2 - (\Sigma y)^2]\}^{1/2}}. \qquad (32)$$

Для нашего примера (см. табл. 4):

$$\rho = \frac{5 \cdot 38 - 15 \cdot 10}{([5 \cdot 55 - (15)^2]\,[5 \cdot 28 - (10^2)]\}^{1/2}} = \frac{40}{(2000)^{1/2}} = \frac{40}{44{,}73} = 0{,}894.$$

Отметим, что по степени тесноты связи между двумя переменными по своей абсолютной величине (модулю) корреляционные связи считаются: «отсутствующими» при $|\rho| \approx 0$ «слабыми» при $|\rho| = 0{,}2 - 0{,}3$; «существенными» при $|\rho| = 0{,}5 - 0{,}7$ и «сильными» при $|\rho| \approx 0{,}9$, «функциональными» при $|\rho| = 1$, когда исходные дискретные данные строго лежат на аппроксимирующей их функции.

Область изменения коэффициента линейной корреляции, как и коэффициента парной ранговой корреляции: $-1 \leq \rho \leq +1$.

В данном случае имеем *сильную прямую зависимость* между временем (в месяцах) и величиной прибыли (в тыс. руб.). Кстати, коэффициент детерминации

$$R^2 = \rho^2 = (0{,}894)^2 = 0{,}799 \approx 0{,}80, \qquad (33)$$

или искомое уравнение (28) отражает влияние времени на прибыль с охватом до 80% всех действующих *причин*. То есть фактор (регрессор) «время» охватывает в реальности до 80% всех возможных факторов, действующих на прибыль «у». (Для парной линейной регрессии это далеко не плохо. – Авт.).

Нам остается лишь выяснить степень достоверности вычисленной тесноты связи.

3.4. Формирование и поверка нулевых гипотез

В общем случае для оценки тесноты связи аргумента и функции, значимости полученных коэффициентов и надежности уравнения регрессии исследователь формирует для каждого названного этапа так называемые соответствующие «нулевые гипотезы» и производит их верификацию по соответствующим правилам.

Общее правило формирования нулевых гипотез состоит в следующем. Сначала формулируется утверждение о том, что то, что мы собираемся установить в качестве реально существующего с заданным уровнем значимости, *как бы отсутствует*. Здесь рассмотрим лишь формирование и поверку нулевой гипотезы H_0 относительно тесноты связи аргумента и функции.

Нулевая гипотеза H_0 в данном случае формируется так: «Прибыль «у» и время «х» функционально не связаны: $y \neq f(x)$, или, иными словами, размер прибыли от времени не зависит, если некоторый рабочий параметр $t_{рас} < t_{табл}$».

И наоборот, H_0 нами *опровергается*, то есть аргумент и функция (здесь – время и прибыль) связаны, если $t_{рас} \geq t_{табл}$.

То есть *для опровержения или принятия* данной гипотезы необходимо произвести дополнительные вычисления – рассчитать параметр $t_{рас}$ (как функцию от коэффициента линейной корреляции ρ и числа пар исходных данных N) и сравнить его значение с табличным параметром $t_{таб}$ (как функции от числа степеней свободы db и уровня значимости α).

Сначала определим расчетный параметр Стьюдента:

$$t_{рас} = \frac{|\rho|(N-2)^{1/2}}{(1-\rho^2)^{1/2}}. \qquad (34)$$

Вполне очевидно, что величина $t_{рас}$ всегда больше нуля. При подстановке наших данных в выражение (24) получим:

$$t_{рас} = \frac{|\rho|(N-2)^{1/2}}{(1-\rho^2)^{1/2}} = \frac{0{,}894 \bullet (5-2)^{1/2}}{(1-0{,}894^2)^{1/2}} = \frac{1{,}548}{(0{,}201)^{1/2}} = \frac{1{,}548}{0{,}448} = 3{,}455.$$

Далее производится сравнение расчетного и табличного параметра. При этом, если

$$t_{рас} \geq t_{таб}, \qquad (35)$$

с заданным уровнем ошибки α (в относительных единицах; если умножить ее значения на 100, получим проценты), то нулевая гипотеза H_0 о независимости времени и прибыли *отвергается*, следовательно, связь между переменными x и y *существует* как «отрицание отрицания» и является значимой.

Если же нестрогое неравенство (34) не выполняется, то «нулевая гипотеза» *принимается*: связь прибыли от времени является статистически незначимой.

Для определения табличных значений $t_{таб}$ воспользуемся таблицей Стьюдента, взятой из статистического справочника; фрагмент таблицы для t-параметра Стьюдента приведен в табл. 6. Вход в таблицу осуществляется по числу степеней свободы df, которое вычисляется так:

$$df = N - 2. \qquad (36)$$

В нашем случае согласно (36) число степеней свободы df = 5 – 2 = 3.

Обычно в социально-экономических исследованиях приняты уровни значимости P в 90%, 95% и 99%, что соответствует значениям ошибки в относительных единицах α в 0,10; 0,05 и 0,01 (или 10%, 5% и 1%) – соответственно. Иначе говоря, табличное значение параметра $t_{таб}$ зависит не только от числа степеней свободы, но и от уровня рассматриваемых значимостей: $t_{таб} = t_{таб}$ (db,α) – всего 3 значения – для ошибки в 10%, 5% и 1%, тогда как $t_{рас}$ = $t_{рас}$ (N,ρ) нами получено в одном экземпляре. То есть необходимо сравнить два параметра – одно значение $t_{рас}$,с тремя значениями $t_{таб}$, суть которых как функций от несовпадающих аргументов строго различна, то есть они получены независимо друг от друга. В этом и состоит сущность верификации.

Сначала выбираем из табл. 6 Стьюдента строчку, соответствующей нашим степеням свободы db = 3 (в табл. 6 строка выделена шрифтом):

для α:	0,10	0,05	0,01
$t_{таб}$ (db=3,α) =	2,35	3,18	5,84.

(37)

Понятно, что наше расчетное значение $t_{рас} = 3{,}46$ при сравнении с предыдущим фрагментом расположится где-то между 3,18 и 5,84. Тогда, согласно условию (35), можем прийти к следующему умозаключению: нулевую гипотезу H_0 о независимости времени и прибыли мы отвергаем с риском ошибиться не более, чем в 5-ти случаях из ста (в выписке 36 берется *левое* ближайшее табличное значение из трех). Или: время и прибыль связаны с достоверностью не менее 95-ти процентов. Или: утверждая, что время и прибыль связаны, мы рискуем ошибиться не более, чем в пяти случаях из ста.

Все приведенные умозаключения непротиворечивы и являются по смыслу равноценными.

Поскольку табличных значений всегда три, возможны множество ситуаций относительно взаимного расположения их и $t_{рас}$. Понятно, чем теснее корреляционная связь, тем t расчетное больше. А вот если бы $t_{рас} = 12{,}00$, мы бы пришли к умозаключению следующего вида: нулевая гипотеза о несвязанности времени и прибыли нами отвергается (то есть они – связаны) – с риском ошибиться «не более, чем в одном случае из ста».

И напротив, если, например, $t_{рас}$ = 2,05, то, согласно табличной выписке (37), расчетное значение t-параметра Стьюдента не превосходит ни одного из табличных значений (находится левее таблицы): тогда нулевая гипотеза H_0 о не связанности времени и прибыли (аргумента и функции) *подтверждается*, так как $t_{рас}$ = 2,05 < $t_{табл}$ = 2,35 даже для ошибки в 10% . Конечно, формально мы могли бы утверждать, что и в данном случае нулевая гипотеза может быть опровергнута, но с одной лишь деталью: «с ошибкой более чем в 10-ти случаев из ста» (а сколько это реально – 12%, 15%, 17%? – неясно), что в статистических исследованиях не принято. Надежность менее 90% (или ошибка более 10%) в социально-экономических исследованиях считается неприемлемой (без специальных оговорок).

Итак, в нашем случае нестрогое неравенство (35) выполняется на строке таблицы (выделено шрифтом) для вероятности более, чем 95% (т.е. с ошибкой менее α = 5%), но менее 99% (т.е. с ошибкой более, чем 1%).

Таблица 6

Значение t-критерия Стьюдента при уровне значимости ά

Число степеней свободы df	Уровень значимости ά, отн. ед.		
	0,10	0,05	0,01
1	6,3138	12,706	63,657
2	2,2900	4,3027	9,9248
3	**2,3534**	**3,1825**	**5,8409**
4	2,1318	2,7764	4,6041
5	2,0150	2,5706	4,0321
6	1,9432	2,4469	3,7074
7	1,8946	2,3646	3,4995
8	1,8595	2,3060	3,3554
9	1,8331	2,2622	3,2498
10	1,8125	2,2281	3,1693

Таким образом, можно сделать следующий вывод: «нулевая гипотеза» о несвязанности аргумента и функции может быть опровергнута с вероятностью, не менее 95% (или принята с вероятностью менее 5%). То есть, отвергая H_0, мы можем ошибиться менее, чем в пяти случаев из ста, тогда как принимая ее, мы ошибемся в более, чем 95-ти случаях из 100.

В итоге полученным результатам точечного прогнозирования в 4,4 тыс. руб. на июнь мы в известном смысле доверяем. Задача на данном этапе решена.

4. ОБЩИЕ ВЫВОДЫ И МЕТОДИЧЕСКИЕ РЕКОМЕНДАЦИИ

Вначале мы располагали лишь эмпирическими данными между временем (в месяцах) и размерами прибыли (в тыс. руб.) – см. табл. 1 и 2. В результате применения метода наименьших квадратов для аппроксимирующей парной функции линейного вида получили значения коэффициентов a и b по формулам (23) и (24) - соответственно, убедились в правильности их значений по тождеству (25), построили прогноз на шестой месяц по формуле (28), рассчитали ошибку аппроксимации по (31), оценили степень тесноты связи функции и аргумента по (32) и сделали выводы о приемлемости «нулевой гипотезы» с помощью параметра Стьюдента по выражениям (34) – (37).

Хотя и аргумент и функция связаны достаточно тесно, однако ошибка аппроксимации (31) довольно высока (38,6%). Поэтому надежность полученного точечного прогноза вызывает известные сомнения. Для повышения точности прогноза необходимо попытаться либо аппроксимировать данную нам эмпирическую зависимость каким-либо другим видом парной зависимости (показательной, степенной и др.), либо увеличить число наблюдений (параметр N).

Заметим также, что полученный точечный прогноз дожжен быть дополнен доверительными интервалами σ – средним квадратическим отклонением, тогда конечный результат прогнозирования выглядел бы так: $y_{прог} = y_6 = 4{,}4 \pm \sigma$ тыс. руб. Однако с подобными расчетами, а также с формированием и поверкой ряда других «нулевых гипотез» («все коэффициенты уравнения регрессии неотличимы от нуля» - при оценке надежности полученного уравнения регрессии по Фишеру; «отдельные параметры уравнения регрессии неотличимы от нуля» - при оценке степени значимости коэффициентов по Стьюденту и др.) также подробно изучаются в рамках учебной дисциплины «Эконометрика».

Далее предлагается провести подобные расчеты для вариантов, приведенных в табл. 7 и интерпретировать полученные результаты по следующему алгоритму.

1. Оформить в рабочих тетрадях запись с указанием номера варианта (для заочного обучения – фамилию, инициалы, номер группы, специализацию, номер варианта).

2. Переписать исходные данные для 2-го и 3-го столбцов таблицы вида табл. 2 для своего варианта. Аргументы у всех вариантов одинаковы – от января (код 1) до июля (код 7), значения функции - разные.

3. Представить исходные данные в графическом виде, как это показано на рис. 3 и определить примерный характер аппроксимирующей линейной функции – вид тренда с предварительной оценкой знаков при искомых коэффициентах для *будущей прямой* вида (28) на рис. 3. Все графические работы рекомендуется делать эскизно, «от руки».

4. Рассчитать коэффициенты уравнения линейной регрессии по формулам (23) и (24).

5. Убедиться в том, что знаки и значения коэффициентов найдены верно по выражению (25), то есть подтвердить или опровергнуть справедливость тождество (25)..

6. Представить полученную линейную зависимость на том же графике (см. рис. 3), где ранее отображены эмпирические данные, для чего определить на графике две точки – при $x = 0$ и $x = 6$ или 7 (для более точного отображения

на графике прямой линии - подальше от начала графика) и найти соответствующие значения функции по формуле (28).

7. Определить аналитический точечный прогноз на следующий месяц по общему виду (28), подставив в него значение аргумента $x_8 = 8$.

8. Отобразить значение графо-аналитического прогноза на том же графике, как это *частично показано* на рис. 3.

7. Рассчитать относительную ошибку аппроксимации по выражению (31) так, как это показано на примере табл. 5.

8. Найти степень тесноты связи (коэффициент линейной корреляции между аргументом и функцией) по выражению (32) и сделать выводы о знаке связи и степени тесноты связи («практически отсутствует», «слабая», «существенная» «сильная», «функциональная», когда эмпирические точки точно лежат на аппроксимирующей прямой).

Проверить «нулевую гипотезу» о не связанности х и у, для чего:

9. Рассчитать значение параметра Стьюдента по формуле (34).

10. Определить значение степеней свободы по выражению (35).

11. Выбрать в табл. 6 пороговые табличные значения параметра Стьюдента для вероятностей 90, 95 и 99 процентов (или ошибок 10, 5 и 1 процентов – соответственно) и выписать отдельно, как показано на примере записи (37).

10. Осуществить сравнение рассчитанного и табличных параметров Стьюдента по нестрогому неравенству (35). Если оно выполняется – «нулевая гипотеза» о несвязанности времени и прибыли *отвергается* (связь между аргументом и функцией статистически существует); если же нестрогое неравенство (35) не выполняется, то «нулевая гипотеза» о несвязанности времени и прибыли *принимается* (связь между аргументом и функцией статистически не существует).

11. *Доказательно и кратко* сформулировать вывод о степени принятия или непринятия «нулевой гипотезы» о взаимосвязи функции и аргумента.

12. Сформулировать общий вывод по сделанной работе.

Далее предлагаются варианты самостоятельной работы 3..

5. ВАРИАНТЫ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

Таблица 7

Варианты контрольной работы

Месяц	x_i	Варианты y_i							
		1	2	3	4	5	6	7	8
Январь	1	4	3	4	6	4	5	4	5
Февраль	2	5	4	5	7	7	4	5	6
Март	3	3	6	5	2	3	3	2	3
Апрель	4	2	1	3	4	1	2	3	2
Май	5	1	2	1	2	2	1	2	0
Июнь	6	1	1	2	1	1	0	0	1
Июль	7	1	1	1	1	1	1	1	1

Месяц	x_i	Варианты y_i							
		9	10	11	12	13	14	15	16
Январь	1	1	3	5	7	8	5	2	3
Февраль	2	2	1	7	5	9	8	4	5
Март	3	3	4	4	4	5	4	3	2
Апрель	4	4	3	0	2	4	7	7	1
Май	5	5	2	1	0	3	1	1	0
Июнь	6	6	2	1	1	2	1	1	1
Июль	7	8	2	1	1	2	1	1	0

Месяц	x_i	Варианты y_i							
		17	18	19	20	21	22	23	24
Январь	1	2	8	9	8	3	2	2	5
Февраль	2	3	7	3	4	2	2	2	6
Март	3	4	3	5	7	3	5	5	3
Апрель	4	0	2	3	9	4	4	1	2
Май	5	1	1	2	9	5	5	0	2
Июнь	6	1	0	2	9	6	6	1	1
Июль	7	1	1	2	9	6	6	0	2

Месяц	x_i	Варианты y_i							
		25	26	27	28	29	30	31	32
Январь	1	1	2	2	2	2	1	1	2
Февраль	2	2	1	1	2	2	3	3	3
Март	3	3	3	4	3	3	4	4	4
Апрель	4	4	4	4	3	4	4	5	4
Май	5	5	4	6	4	5	5	6	5
Июнь	6	5	5	7	4	6	6	4	6
Июль	7	5	6	7	5	6	6	7	7

ЛИТЕРАТУРА

1. Громыко Г.Л. Общая теория статистики: Практикум. М.: ИНФРА-М, 1999. 139 с. (Высшее образование).

2. Ниворожкина Л.И., Морозова З.А. Математическая статистика с элементами теории вероятностей в задачах и решениях: Учебное пособие. Москва: ИКЦ «МарТ»; Ростов-н/Д: Издательский центр «МарТ», 2005. 608 с. (Серия «Учебный курс»).

3. Рудакова Р.П., Букин Л.Л., Гаврилов В.И. Статистика. 2-е изд. СПб: Питер, 2007. 288 с.: ил. (Серия «Учебное пособие»).

Учебное издание

Шихалёв Анатолий Михайллович

РЕГРЕССИОННЫЙ АНАЛИЗ. ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Дизайн обложки
М.А. Ахметов

Подписано в печать 15.06.2015.

Бумага офсетная. Печать цифровая.
Формат 60х84 1/16. Гарнитура «Times New Roman». Усл. печ. л. .
Тираж экз. Заказ

Отпечатано с готового оригинал-макета
в типографии Издательства Казанского университета

420008, г. Казань, ул. Профессора Нужина, 1/37
тел. (843) 233-73-59, 233-73-28