7. КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

7.1. МАТРИЦА ДАННЫХ

Многие объекты исследования характеризуются множеством параметров, и по результатам наблюдения за их функционированием формируются многомерные совокупности (матрицы) ЭД

$$X = \begin{vmatrix} x_{11} & x_{12} & . & x_{1m} \\ x_{21} & x_{22} & . & x_{2m} \\ . & . & . & . \\ x_{n1} & x_{n2} & . & x_{nm} \end{vmatrix} \qquad (7.1)$$

Строки такой матрицы соответствуют результатам регистрации всех наблюдаемых параметров объекта в одном эксперименте, а столбцы содержат результаты наблюдений за одним параметром (фактором, вариантой) во всех экспериментах. Обозначим количество параметров через m (m>1), а количество наблюдений – через n.

В матрице элемент x_{ij} соответствует значению j-й варианты в i-м наблюдении. Матрица, вообще говоря, может содержать пустые значения некоторых элементов, например, из-за пропусков в регистрации значений параметров. В многомерном анализе желательно устранить пропущенные значения. Для этого существуют специальные приемы, в частности, вычеркивание соответствующих строк матрицы или занесение средних значений вместо отсутствующих. В дальнейшем будем считать, что матрица не содержит пустых элементов, а параметры объекта характеризуются непрерывными случайными величинами.

Методы обработки матрицы ЭД основаны на следующем предположении: если объект подвергнуть новому обследованию и получить, вообще говоря, другую матрицу данных, то после ее обработки с помощью тех же методов будут получены результаты, близкие к результатам обработки первой матрицы. Данное предположение основано на статистической гипотезе формирования матрицы ЭД. Матрица порождается случайным образом в соответствии с определенной вероятностной закономерностью, а именно: в m-мерном пространстве параметров существует некоторое (пусть и неизвестное) распределение вероятностей, и каждая строка матрицы появляется в соответствии с этим распределением независимо от появления других строк.

Каждый столбец матрицы представляет собой случайную выборку значений одного параметра объекта. Указанное предположение означает, во-первых, что оценки моментов и параметров распределения, вычисленные по

выборке, будут близки к истинным значениям, во-вторых, значения непрерывных функций, построенных по этим оценкам, будут близки к значениям функций, построенным по истинным значениям параметров.

Таким образом, объектом исследования в многомерном анализе является многомерная случайная величина, представленная выборкой конечного объема. К такой выборке применимы все методы и оценки, рассмотренные при обработке одномерных ЭД. Конечно, приведенные суждения не являются доказательством допустимости применения рассматриваемых методов, но вполне подтверждаются практикой.

Параметры, характеризующие объект исследования, имеют разный физический смысл, и матрица данных существенно изменяется, если изменяются шкалы, в которых измеряются те или иные параметры. Матрицу данных еще до проведения анализа целесообразно привести к стандартному виду, т.е. стандартизовать значения вариант (напомним, что среднее значение стандартизованной варианты равно нулю, дисперсия – единице). В тех случаях, когда все варианты измеряются в одной шкале, это преобразование все-таки желательно, ибо оно упрощает последующие преобразования. Стандартизованную матрицу будем обозначать через $\boldsymbol{U}$. Переход от исходной к стандартизованной матрице осуществляется следующим образом.

1. По каждой варианте вычисляются оценки:

- математического ожидания $m^*(x_j) = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$;
- дисперсии $\mu_2(x_j) = \sigma^2(x_j) = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \mu_1(x_i))^2$.

2. Вычисляются элементы стандартизованной матрицы

$$u_{ij} = (x_{ij} - m^*(x_j))/\sigma(x_j), i = 1,\dots,n, j = 1,..,m.$$

Элементы матрицы $\boldsymbol{U}$ являются безразмерными величинами. Именно матрица $\boldsymbol{U}$ будет являться объектом последующей обработки.

7.2. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Величины, характеризующие различные свойства объектов, могут быть независимыми или взаимосвязанными. Различают два вида зависимостей между величинами (факторами): функциональную и статистическую.

При функциональной зависимости двух величин значению одной из них обязательно соответствует одно или несколько точно определенных значений другой величины. Функциональная связь двух факторов возможна лишь при условии, что вторая величина зависит только от первой и не зависит ни от каких других величин. Функциональная связь одной величины с множеством других возможна, если эта величина зависит только от этого множества факторов. В реальных ситуациях существует бесконечно большое количество свойств самого объекта и внешней среды, влияющих друг на друга, поэтому

такого рода связи не существуют, иначе говоря, функциональные связи являются математическими абстракциями. Их применение допустимо тогда, когда соответствующая величина в основном зависит от соответствующих факторов.

При исследовании сложных систем многие параметры следует считать случайными, что исключает проявление однозначного соответствия значений. Воздействие общих факторов, наличие объективных закономерностей в поведении объектов приводят лишь к проявлению статистической зависимости. Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения других (другой), и эти другие величины принимают некоторые значения с определенными вероятностями.

Важным частным случаем статистической зависимости является корреляционная зависимость, характеризующая взаимосвязь значений одних случайных величин со средним значением других, хотя в каждом отдельном случае любая взаимосвязанная величина может принимать различные значения.

Если же у взаимосвязанных величин вариацию имеет только одна переменная, а другая является детерминированной, то такую связь называют не корреляционной, а регрессионной. Например, при анализе скорости обмена с жесткими дисками можно оценивать регрессию этой характеристики на определенные модели, но не следует говорить о корреляции между моделью и скоростью.

При исследовании зависимости между одной величиной и такими характеристиками другой, как, например, моменты старших порядков (а не среднее значение), то эта связь будет называться статистической, а не корреляционной.

Корреляционная связь описывает следующие виды зависимостей:

причинную зависимость между значениями параметров. Примером такой зависимости является взаимосвязь пропускной способности канала передачи данных и соотношения сигнал/шум (на пропускную способность влияют и другие факторы – характер помех, способ кодирования сообщений и др.). Установить однозначную связь между конкретными значениями указанных параметров не удается. Но очевидно, что пропускная способность зависит от соотношения уровней сигнала и помех в канале. Иногда при этом причину и следствие особо не выделяют. В некоторых случаях такая корреляция является бессмысленной, например: если в качестве исходного фактора взять доходы разработчиков антивирусных программ, а за результат – количество вновь появляющихся вирусов, то можно сделать вывод, что разработчики антивирусов "стимулируют" создание вирусов;

"зависимость" между следствиями общей причины. Подобная зависимость характерна, в частности, для скорости и безошибочности набора текста оператором (указанные факторы зависят от квалификации оператора).

Корреляционная зависимость определяется различными параметрами, среди которых наибольшее распространение получили показатели, характеризующие взаимосвязь двух случайных величин (парные показатели): корреляционный момент, коэффициент корреляции.

Оценка *корреляционного момента (коэффициента ковариации*) двух вариант x_j и x_k вычисляется по исходной матрице $\boldsymbol{X}$

$$K *_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - m^*(x_j))(x_{ik} - m^*(x_k)). \quad (7.2)$$

Этот показатель неудобен для практического применения, так как имеет размерность, равную произведению размерностей вариант, и по его величине трудно судить о зависимости параметров.

Коэффициент ковариации r_{jk} нормированных случайных величин называют коэффициентом корреляции, его оценка

$$r *_{jk} = \frac{1}{n}\sum_{i=1}^{n} u_{ij}u_{ik} = \frac{\sum_{i=1}^{n}(x_{ij} - m^*(x_j))(x_{ik} - m^*(x_k))}{nS_jS_k}. \quad (7.3)$$

Значение коэффициента корреляции лежит в пределах от –1 до +1. Если случайные величины X_j и X_k независимы, то коэффициент r_{jk} обязательно равен нулю, обратное утверждение неверно. Коэффициент r_{jk} характеризует значимость линейной связи между параметрами:

- при r_{jk}=1 значения u_{ij} и u_{ik} полностью совпадают, т.е. значения параметров принимают одинаковые значения. Иначе говоря, имеет место функциональная зависимость: зная значение одного параметра, можно однозначно указать значение другого параметра;
- при r_{jk}= – 1 величины u_{ij} и u_{ik} принимают противоположные значения. И в этом случае имеет место функциональная зависимость;
- при r_{jk} = 0 величины u_{ij} и u_{ik} практически не связаны друг с другом линейным соотношением. Это не означает отсутствия каких-то других (например, нелинейных) связей между параметрами;
- при $|r_{jk}| > 0$ и $|r_{jk}| < 1$ однозначной линейной связи величин x_{ij} и x_{ik} нет. И чем меньше абсолютная величина коэффициента корреляции, тем в меньшей степени по значениям одного параметра можно предсказать значение другого.

Используя понятие коэффициента корреляции, матрице ЭД можно поставить в соответствие квадратную матрицу оценок коэффициентов корреляции (корреляционную матрицу)

$$r^* = \begin{vmatrix} r *_{11} & r *_{12} & . & r *_{1m} \\ r *_{21} & r *_{22} & . & r *_{2m} \\ . & . & . & . \\ r *_{n1} & r *_{n2} & . & r *_{nm} \end{vmatrix}. \quad (7.4)$$

К числу характерных свойств корреляционной матрицы относят: симметричность относительно главной диагонали, $r^*_{jk}=r^*_{kj}$, ; единичные значения элементов главной диагонали, $r_{kk}=1$ (r_{kk} соответствует дисперсии стандартизованного параметра X_k), .

Оценка коэффициента корреляции, вычисленная по ограниченной выборке, практически всегда отличается от нуля. Но из этого еще не следует, что коэффициент корреляции генеральной совокупности также отличен от нуля. Требуется оценить значимость выборочной величины коэффициента или, в соответствии с постановкой задач проверки статистических гипотез, проверить гипотезу о равенстве нулю коэффициента корреляции. Если гипотеза H_0 о равенстве нулю коэффициента корреляции будет отвергнута, то выборочный коэффициент значим, а соответствующие величины связаны линейным соотношением. Если гипотеза H_0 будет принята, то оценка коэффициента не значима, и величины линейно не связаны друг с другом (если по физическим соображениям факторы могут быть связаны, то лучше говорить о том, что по имеющимся ЭД эта взаимосвязь не установлена). Проверка гипотезы о значимости оценки коэффициента корреляции требует знания распределения этой случайной величины. Распределение величины r_{ik} изучено только для частного случая, когда случайные величины U_j и U_k распределены по нормальному закону.

В качестве критерия проверки нулевой гипотезы H_0 применяют случайную величину

$$t = |r *_{ik}| \frac{\sqrt{n-2}}{\sqrt{1-r*^2_{ik}}}.$$

Если модуль коэффициента корреляции относительно далек от единицы, то величина t при справедливости нулевой гипотезы распределена по закону Стьюдента с $n–2$ степенями свободы. Конкурирующая гипотеза H_1 соответствует утверждению, что значение r_{ik} не равно нулю (больше или меньше нуля). Поэтому критическая область двусторонняя.

Проверка гипотезы H_0 о равенстве нулю генерального коэффициента парной корреляции двумерной нормально распределенной случайной величины осуществляется в следующей последовательности:

- вычисляется значение статистики t;
- при уровне значимости a для двусторонней области определяется критическая точка распределения Стьюдента $t_{кр}(n–2; a)$, табл;
- сравнивается значение статистики t с критическим значением $t_{кр}(n–2; a)$. Если $t < tкр\ (n–2; a)$, то нет оснований отвергнуть нулевую гипотезу, иначе гипотеза H_0 отвергается (коэффициент корреляции значим).

Когда модуль величины r^*_{ik} близок к единице, распределение r^*_{ik} отличается от распределения Стьюдента, так как значение $|r^*_{ik}|$ ограничено справа единицей. В этом случае применяют преобразование

$$y_{ik}=0,5ln[(1+|r_{ik}|)/(1–|r_{ik}|)].$$

Величина y_{ik} не имеет указанного ограничения, она при $n>10$ распределена приблизительно нормально с центром

$$m^*(r_{ik})=0{,}5\ln[(1+|r_{ik}|)/(1-|r_{ik}|)]+0{,}5|r_{ik}|/(n-1)$$

и дисперсией

$$\mu_2^*(r_{ik})=s^2(r_{ik})=1/(n-3).$$

Если значение центрированной и нормированной величины

$$(y_{ik}-m_1^*(r_{ik}))/s(r_{ik})$$

превышает значение квантили уровня $1-a/2$ нормального распределения стандартизованной величины, то нулевая гипотеза отвергается.

Таким образом, *постановка задачи линейного корреляционного анализа* формулируется в следующем виде.

Имеется матрица наблюдений вида (7.1).

Необходимо определить оценки коэффициентов корреляции для всех или только для заданных пар параметров и оценить их значимость. Незначимые оценки приравниваются к нулю.

Допущения:

- выборка имеет достаточный объем. Понятие достаточного объема зависит от целей анализа, требуемой точности и надежности оценки коэффициентов корреляции, от количества факторов. Минимально допустимым считается объем, когда количество наблюдений не менее чем в 5–6 раз превосходит количество факторов;
- выборки по каждому фактору являются однородными. Это допущение обеспечивает несмещенную оценку средних величин;
- матрица наблюдений не содержит пропусков.

Если необходима проверка значимости оценки коэффициента корреляции, то требуется соблюдение дополнительного условия – распределение вариант должно подчиняться нормальному закону.

Задача анализа решается в несколько этапов:

- проводится стандартизация исходной матрицы;
- вычисляются парные оценки коэффициентов корреляции;
- проверяется значимость оценок коэффициентов корреляции, незначимые оценки приравниваются к нулю. По результатам проверки делается вывод о наличии связей между вариантами (факторами).

Пример 7.1. Результаты наблюдений за характеристиками канала представлены в табл. 7.1.

Таблица 7.1

№ пп	Пропускная способность	Отношение сигнал/шум	Остаточное затухание		
			1020	1800	1400
	X_1	X_2	X_3	X_4	X_5
1	26.37	41.98	17.66	16.05	22.85
2	28.00	43.83	17.15	15.47	23.25
3	27.83	42.83	15.38	17.59	24.55
4	31.67	47.28	18.39	16.92	26.59
5	23.50	38.75	18.32	15.66	26.22
6	21.04	35.12	17.81	17.00	27.52
7	16.94	32.07	21.42	16.77	25.76
8	37.56	54.25	26.42	15.68	23.10
9	18.84	32.70	17.23	15.92	23.41
10	25.77	40.51	30.43	15.29	25.17
11	33.52	49.78	21.71	15.61	25.39
12	28.21	43.84	28.33	15.70	24.56
13	28.76	44.03	30.42	16.87	24.45
14	24.60	39.46	21.66	15.25	23.81
15	24.51	38.78	25.77	16.05	24.48

Необходимо определить наличие линейных корреляционных связей между пропускной способностью и остальными факторами. Предполагается, что выборки по всем вариантам подчиняются нормальному закону. Проверку гипотезы о значимости оценок коэффициентов корреляции произвести с уровнем значимости a, равным 0,1.

Решение. Стандартизация исходной матрицы начинается с вычисления выборочной средней m_1^*, несмещенной оценки дисперсии μ_2^* и среднеквадратического отклонения S по каждой варианте, табл.7.2.

Таблица 7.2

Оценка параметра распределения	Варианта				
	X_1	X_2	X_3	X_4	X_5
m^*	26.47	41.68	21.87	16.12	24.74
S^2	29.10	36.47	26.37	0.52	1.88
S	5.39	6.04	5.13	0.72	1.37

В результате перехода к величинам $u_{ij} = {(x_{ij} - m_j^*)}/{\sigma_j}$ формируется стандартизованная матрица исходных данных, табл. 7.3.

Таблица 7.3

№ пп	Пропускная способность	Отношение сигнал/шум	Остаточное затухание на частоте		
			1020	1800	1400
	U_1	U_2	U_3	U_4	U_5
1	-0.02	-0.05	-0.82	-0.10	-1.38
2	0.28	0.36	-0.92	-0.90	-1.09
3	0.25	0.19	-1.26	2.03	-0.14
4	0.96	0.93	-0.68	1.10	2.35
5	-0.55	-0.49	-0.69	-0.64	1.08
6	-1.01	-1.09	-0.79	1.21	2.03
7	-1.77	-1.59	-0.09	0.90	0.74
8	2.06	2.08	0.89	-0.61	-1.20
9	-1.42	-1.49	-0.90	-0.28	-0.97
10	-0.13	-0.19	1.67	-1.15	0.31
11	1.31	1.34	-0.03	-0.71	0.47
12	0.32	0.36	1.26	-0.58	-0.13
13	0.42	0.39	1.66	1.03	-0.21
14	-0.35	-0.37	-0.04	-1.21	-0.68
15	-0.36	-0.48	0.76	-0.19	-0.19

Оценки коэффициентов корреляции

$$r *_{1k} = \frac{1}{15}\sum_{i=1}^{15} u_{1i}u_{ki}, \qquad (k = 2, 3, 4)$$

представлены в табл. 7.4. В этой же таблице приведены значения статистик критерия $t = |r *_{1k}| \sqrt{n-2} \Big/ \sqrt{1 - r *_{1k}^2}$ Стьюдента для вычисленных оценок коэффициентов корреляции при n = 15.

Таблица 7.4

	X_2	X_3	X_4	X_5
r_{1j}	0.93	0.25	-0.13	-0.22
t	9.12	0.93	0.47	0.81

Критическое значение

$$t_{кр}\ (n–2;\ a) = t_{кр}\ (13;\ 0{,}1) = 1{,}77.$$

Статистика критерия больше критического значения только для r_{12}. Это означает, что только для указанного коэффициента оценка значима (коэффициент корреляции генеральной совокупности не равен нулю), а остальные коэффициенты следует признать равными нулю.

Корреляционная зависимость не обязательно устанавливается только для двух величин, с ее помощью можно анализировать связи между несколькими

вариантами (множественная корреляция). А кроме линейной существуют и другие виды корреляции.

7.3. РЕГРЕССИОННЫЙ АНАЛИЗ

7.3.1. Постановка задачи

Одной из типовых задач обработки многомерных ЭД является определение количественной зависимости показателей качества объекта от значений его параметров и характеристик внешней среды. Примером такой постановки задачи является установление зависимости между временем обработки запросов к базе данных и интенсивностью входного потока. Время обработки зависит от многих факторов, в том числе от размещения искомой информации на внешних носителях, сложности запроса. Следовательно, время обработки конкретного запроса можно считать случайной величиной. Но вместе с тем, при увеличении интенсивности потока запросов следует ожидать возрастания его среднего значения, т.е. считать, что время обработки и интенсивность потока запросов связаны корреляционной зависимостью.

Постановка задачи регрессионного анализа формулируется следующим образом.

Имеется совокупность результатов наблюдений вида (7.1). В этой совокупности один столбец соответствует показателю, для которого необходимо установить функциональную зависимость с параметрами объекта и среды, представленными остальными столбцами. Будем обозначать показатель через y^* и считать, что ему соответствует первый столбец матрицы наблюдений. Остальные $m–1$ ($m > 1$) столбцов соответствуют параметрам (факторам) $x_2, x_3, \ldots, x_m$.

Требуется: установить количественную взаимосвязь между показателем и факторами. В таком случае задача регрессионного анализа понимается как задача выявления такой функциональной зависимости $y^* = f(x_2 , x_3 , \ldots, x_m)$, которая наилучшим образом описывает имеющиеся экспериментальные данные.

Допущения:

- количество наблюдений достаточно для проявления статистических закономерностей относительно факторов и их взаимосвязей;
- обрабатываемые ЭД содержат некоторые ошибки (помехи), обусловленные погрешностями измерений, воздействием неучтенных случайных факторов;
- матрица результатов наблюдений является единственной информацией об изучаемом объекте, имеющейся в распоряжении перед началом исследования.

Функция $f(x_2 , x_3 , \ldots, x_m)$, описывающая зависимость показателя от параметров, называется уравнением (функцией) регрессии. Термин "регрессия" (regression (лат.) – отступление, возврат к чему-либо) связан со спецификой одной из конкретных задач, решенных на стадии становления

метода. Его ввел английский статистик Ф. Гальтон. Он исследовал влияние роста родителей и более отдаленных предков на рост детей. По его модели рост ребенка определяется наполовину родителями, на четверть – дедом с бабкой, на одну восьмую прадедом и прабабкой и т.д. Другими словами, такая модель характеризует движение назад по генеалогическому дереву. Ф. Гальтон назвал это явление регрессией как противоположное движению вперед – прогрессу. В настоящее время термин "регрессия" применяется в более широком плане – для описания любой статистической связи между случайными величинами.

Решение задачи регрессионного анализа целесообразно разбить на несколько этапов:

- предварительная обработка ЭД;
- выбор вида уравнений регрессии;
- вычисление коэффициентов уравнения регрессии;
- проверка адекватности построенной функции результатам наблюдений.

Предварительная обработка включает стандартизацию матрицы ЭД, расчет коэффициентов корреляции, проверку их значимости и исключение из рассмотрения незначимых параметров (эти преобразования были рассмотрены в рамках корреляционного анализа). В результате преобразований будут получены стандартизованная матрица наблюдений $\boldsymbol{U}$ (через y будем обозначать стандартизованную величину y*) и корреляционная матрица **r**.
Стандартизованной матрице $\boldsymbol{U}$ можно сопоставить одну из следующих геометрических интерпретаций:

- в m-мерном пространстве оси соответствуют отдельным параметрам и показателю. Каждая строка матрицы представляет вектор в этом пространстве, а вся матрица – совокупность n векторов в пространстве параметров;
- в n-мерном пространстве оси соответствуют результатам отдельных наблюдений. Каждый столбец матрицы – вектор в пространстве наблюдений. Все вектора в этом пространстве имеют одинаковую длину, равную $\sqrt{n}$. Тогда угол между двумя векторами характеризует взаимосвязь соответствующих величин. И чем меньше угол, тем теснее связь (тем больше коэффициент корреляции).

В корреляционной матрице особую роль играют элементы левого столбца – они характеризуют наличие или отсутствие линейной зависимости между соответствующим параметром u_i (i =2, 3, …, m) и показателем объекта y. Проверка значимости позволяет выявить такие параметры, которые следует исключить из рассмотрения при формировании линейной функциональной зависимости, и тем самым упростить последующую обработку.

7.3.2. Выбор вида уравнения регрессии

Задача определения функциональной зависимости, наилучшим образом описывающей ЭД, связана с преодолением ряда принципиальных трудностей. В общем случае для стандартизованных данных функциональную зависимость показателя от параметров можно представить в виде

$$y = f(u_1, u_2, ..., u_p) + \varepsilon, \qquad (7.5)$$

где f – заранее не известная функция, подлежащая определению;

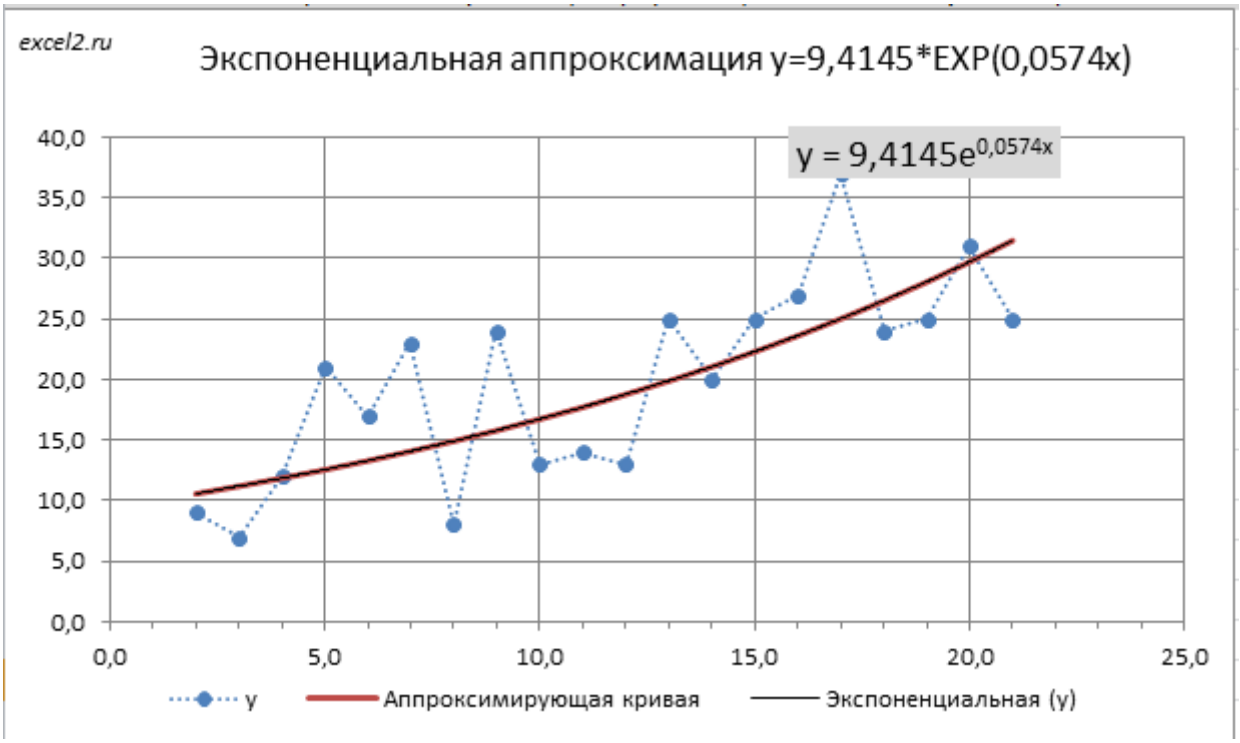
ε - ошибка аппроксимации ЭД.

Указанное уравнение принято называть выборочным уравнением регрессии y на u. Это уравнение характеризует зависимость между вариацией показателя и вариациями факторов. А мера корреляции измеряет долю вариации показателя, которая связана с вариацией факторов. Иначе говоря, корреляцию показателя и факторов нельзя трактовать как связь их уровней, а регрессионный анализ не объясняет роли факторов в создании показателя.

Еще одна особенность касается оценки степени влияния каждого фактора на показатель. Регрессионное уравнение не обеспечивает оценку раздельного влияния каждого фактора на показатель, такая оценка возможна лишь в случае, когда все другие факторы не связаны с изучаемым. Если изучаемый фактор связан с другими, влияющими на показатель, то будет получена смешанная характеристика влияния фактора. Эта характеристика содержит как непосредственное влияние фактора, так и опосредованное влияние, оказанное через связь с другими факторами и их влиянием на показатель.

В регрессионное уравнение не рекомендуется включать факторы, слабо связанные с показателем, но тесно связанные с другими факторами. Не включают в уравнение и факторы, функционально связанные друг с другом (для них коэффициент корреляции равен 1). Включение таких факторов приводит к вырождению системы уравнений для оценок коэффициентов регрессии и к неопределенности решения.

Функция f должна подбираться так, чтобы ошибка ε в некотором смысле была минимальна. Существует бесконечное множество функций, описывающих ЭД абсолютно точно (ε = 0), т.е. таких функций, которые для всех значений параметров $u_{j,2}$, $u_{j,3}$, …, $uj_{,m}$ принимают в точности соответствующие значения показателя y_i , i =1, 2, …, n. Вместе с тем, для всех других значений параметров, отсутствующих в результатах наблюдений, значения показателя могут принимать любые значения. Понятно, что такие функции не соответствуют действительной связи между параметрами и показателем.



В целях выбора функциональной связи заранее выдвигают гипотезу о том, к какому классу может принадлежать функция *f*, а затем подбирают "лучшую" функцию в этом классе. Выбранный класс функций должен обладать некоторой "гладкостью", т.е. "небольшие" изменения значений аргументов должны вызывать "небольшие" изменения значений функции.

Простым, удобным для практического применения и отвечающим указанному условию является класс полиномиальных функций

$$y = a_0 + \sum_{j=2}^{m} a_j u_j + \sum_{j=2}^{m-1} \sum_{k=2}^{m} a_{jk} u_j u_k + \sum_{j=2}^{m} a_{jj} u_j^2 + \ldots + \varepsilon \qquad (7.6)$$

Для такого класса задача выбора функции сводится к задаче выбора значений коэффициентов a_0 , a_j , a_{jk} , ..., a_{jj} , ... Однако универсальность полиномиального представления обеспечивается только при возможности неограниченного увеличения степени полинома, что не всегда допустимо на практике, поэтому приходится применять и другие виды функций.

Частным случаем, широко применяемым на практике, является полином первой степени или уравнение линейной регрессии

$$y = a_0 + \sum_{j=2}^{m} a_j u_j + \varepsilon. \qquad (7.7)$$

Это уравнение в регрессионном анализе следует трактовать как векторное, ибо речь идет о матрице данных

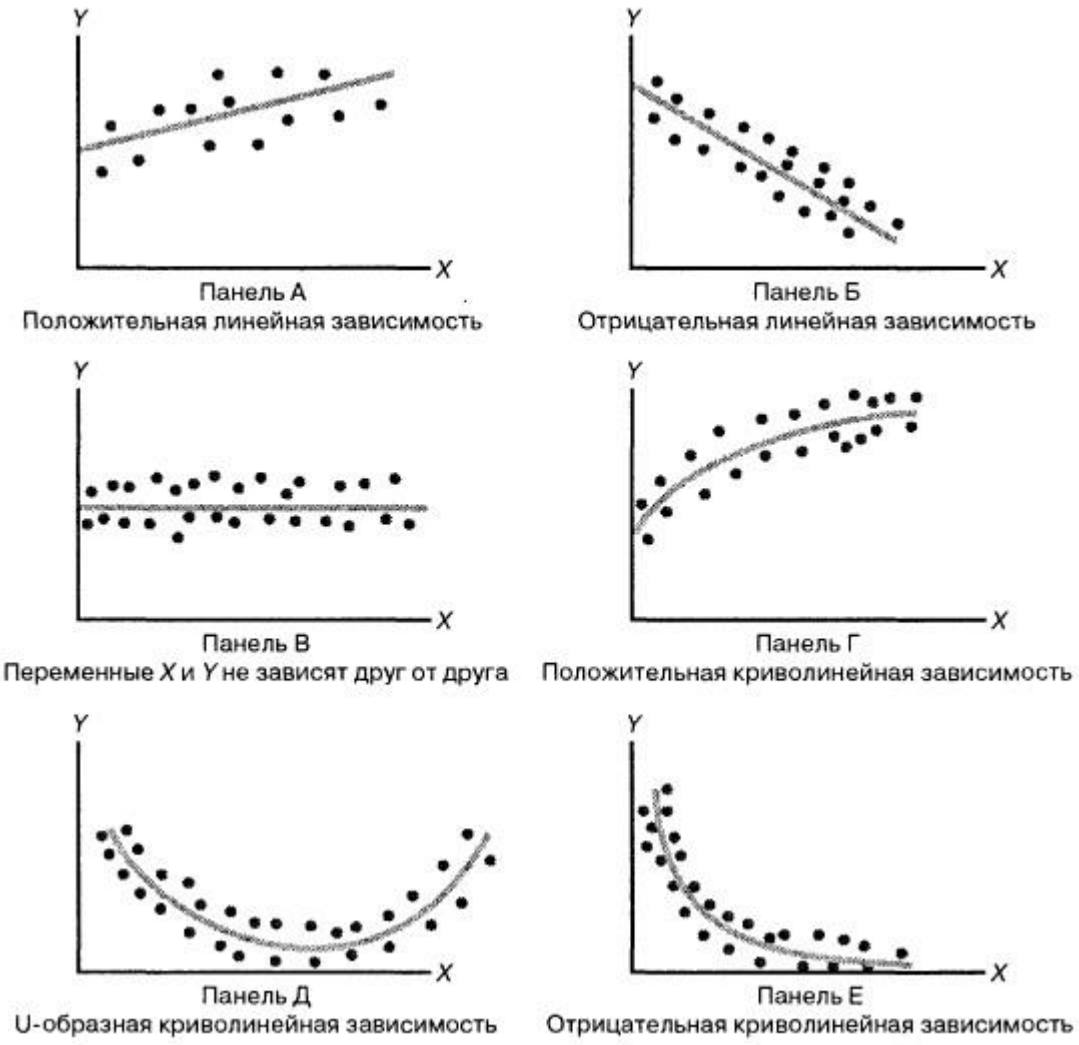
$$y_i = a_0 + \sum_{j=2}^{m} a_j u_{ij} + \varepsilon_i, \; i = 1, 2, \ldots, n. \qquad (7.8)$$

Обычно стремятся обеспечить такое количество наблюдений, которое превышало бы количество оцениваемых коэффициентов модели. Для

линейной регрессии при $n > m$ количество уравнений превышает количество подлежащих определению коэффициентов полинома. Но и в этом случае нельзя подобрать коэффициенты таким образом, чтобы ошибка в каждом скалярном уравнении обращалась в ноль, так как к неизвестным относятся a_j и e_i, их количество $n + m - 1$, т.е. всегда больше количества уравнений n. Аналогичные рассуждения справедливы и для полиномов степени, выше первой.

Для выбора вида функциональной зависимости можно рекомендовать следующий подход:

- в пространстве параметров графически отображают точки со значениями показателя. При большом количестве параметров можно строить точки применительно к каждому из них, получая двумерные распределения значений;
- по расположению точек и на основе анализа сущности взаимосвязи показателя и параметров объекта делают заключение о примерном виде регрессии или ее возможных вариантах;
- после расчета параметров оценивают качество аппроксимации, т.е. оценивают степень близости расчетных и фактических значений;
- если расчетные и фактические значения близки во всей области задания, то задачу регрессионного анализа можно считать решенной. В противном случае можно попытаться выбрать другой вид полинома или другую аналитическую функцию, например периодическую.



7.3.4. Вычисление коэффициентов уравнения регрессии

Систему уравнений (7.8) на основе имеющихся ЭД однозначно решить невозможно, так как количество неизвестных всегда больше количества уравнений. Для преодоления этой проблемы нужны дополнительные допущения. Здравый смысл подсказывает: желательно выбрать коэффициенты полинома так, чтобы обеспечить минимум ошибки аппроксимации ЭД. Могут применяться различные меры для оценки ошибок аппроксимации. В качестве такой меры нашла широкое применение среднеквадратическая ошибка. На ее основе разработан специальный метод оценки коэффициентов уравнений регрессии – метод наименьших квадратов (МНК). Этот метод позволяет получить оценки максимального правдоподобия неизвестных коэффициентов уравнения регрессии при нормальном распределения вариант, но его можно применять и при любом другом распределении факторов.

В основе МНК лежат следующие положения:

- значения величин ошибок и факторов независимы, а значит, и некоррелированы, т.е. предполагается, что механизмы порождения помехи не связаны с механизмом формирования значений факторов;

- математическое ожидание ошибки ε должно быть равно нулю (постоянная составляющая входит в коэффициент a_0), иначе говоря, ошибка является центрированной величиной;
- выборочная оценка дисперсии ошибки $\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2$ должна быть минимальна.

Рассмотрим применение МНК применительно к линейной регрессии стандартизованных величин. Для центрированных величин u_j коэффициент a_0 равен нулю, тогда уравнения линейной регрессии

$$\hat{y}_i = \sum_{j=2}^{m} a_j u_{ij} + \varepsilon, i = \overline{1,n}. \quad (7.9)$$

Здесь введен специальный знак "^", обозначающий значения показателя, рассчитанные по уравнению регрессии, в отличие от значений, полученных по результатам наблюдений.

По МНК определяются такие значения коэффициентов уравнения регрессии, которые обеспечивают безусловный минимум выражению

$$w = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=2}^{m} a_j u_{ij}\right)^2. \quad (7.10)$$

Минимум находится приравниванием нулю всех частных производных выражения (7.10), взятых по неизвестным коэффициентам, и решением системы уравнений

$$\frac{dw}{da_k} = -\frac{2}{n}\sum_{i=1}^{n}\left(y_i - \sum_{j=2}^{m} a_j u_{ij}\right)u_{ik} = 0, k = \overline{2,m}. \quad (7.11)$$

Последовательно проведя преобразования и используя введенные ранее оценки коэффициентов корреляции

$$\frac{1}{n}\sum_{i=1}^{n}\left(y_i u_{ik} - \sum_{j=2}^{m} a_j u_{ij} u_{ik}\right) = 0, k = \overline{2,m};$$

$$\frac{1}{n}\sum_{j=1}^{n} y_i u_{ik} - \sum_{j=2}^{m} a_j \frac{1}{n}\sum_{i=1}^{n} u_{ij} u_{ik} = 0, k = \overline{2,m},$$

получим

$$\rho_{yk} - \sum_{j=2}^{m} a_j \rho_{jk} = 0, k = \overline{2,m}.. \quad (7.12)$$

Итак, получено m–1 линейных уравнений, что позволяет однозначно вычислить значения $a_2, a_3, \ldots, a_m$.

Если же линейная модель неточна или параметры измеряются неточно, то и в этом случае МНК позволяет найти такие значения коэффициентов, при которых линейная модель наилучшим образом описывает реальный объект в смысле выбранного критерия среднеквадратического отклонения.

Когда имеется только один параметр, уравнение линейной регрессии примет вид

$$\hat{y} = a_2 u_2.$$

Коэффициент a_2 находится из уравнения

$$r_{y2} - a_2 r_{22} = 0.$$

Тогда, учитывая, что $r_{2,2} = 1$, искомый коэффициент

$$a_2 = r_{y,2}. \qquad (7.13)$$

Соотношение (7.13) подтверждает ранее высказанное утверждение, что коэффициент корреляции является мерой линейной связи двух стандартизованных параметров.

Подставив найденное значение коэффициента a_2 в выражение для w, с учетом свойств центрированных и нормированных величин, получим минимальное значение этой функции, равное $1 - r^2_{y,2}$. Величину $1 - r^2_{y,2}$ называют остаточной дисперсией случайной величины y относительно случайной величины u_2. Только при $|r_{y,2}| = 1$ остаточная дисперсия равна нулю, и, следовательно, не возникает ошибки при аппроксимации показателя линейной функцией.

Переходя от центрированных и нормированных значений показателя и параметра

$$\frac{x_1 - m(x_1)}{\sigma(x_1)} = r_{y2} \frac{x_2 - m(x_2)}{\sigma(x_2)},$$

можно получить для исходных величин

$$\hat{y} = m(x_1) - r_{y2} m(x_2) \frac{\sigma(x_1)}{\sigma(x_2)} + r_{y2} \frac{\sigma(x_1)}{\sigma(x_2)} x_2. \qquad (7.14)$$

Это уравнение также линейно относительно коэффициента корреляции. Нетрудно заметить, что центрирование и нормирование для линейной регрессии позволяет понизить на единицу размерность системы уравнений, т.е. упростить решение задачи определения коэффициентов, а самим коэффициентам придать ясный смысл.

Применение МНК для нелинейных функций практически ничем не отличается от рассмотренной схемы (только коэффициент a0 в исходном уравнении не равен нулю).

Например, пусть необходимо определить коэффициенты параболической регрессии

$$\hat{y} = a_0 + a_2 u_2 + a_{22} u_2^2.$$

Выборочная дисперсия ошибки

$$w = \frac{1}{n}\sum_{i=1}^{n}(y_i - a_0 - a_2 u_{i2} - a_{22} u_{i2}^2)^2.$$

На ее основе можно получить следующую систему уравнений

$$\begin{cases} \dfrac{dw}{da_0} = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - a_0 - a_2 u_{i2} - a_{22} u_{i2}^2) = 0, \\ \dfrac{dw}{da_1} = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - a_0 - a_2 u_{i2} - a_{22} u_{i2}^2) u_{i2} = 0, \\ \dfrac{dw}{da_2} = -\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(y_i - a_0 - a_1 u_{i2} - a_{22} u_{i2}^2) u_{i2}^2 = 0. \end{cases}$$

После преобразований система уравнений примет вид

$$\begin{cases} m(y) - a_0 - a_2 m(u_2) - a_{22} = 0, \\ r_{y2} - a_0 m(u_2) - a_2 - a_{22}\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} u_{i2}^3 = 0, \\ \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} y_i u_{i2}^2 - a_0 - a_2\dfrac{1}{n}\sum_{i=1}^{n} u_{i2}^3 - a_{22}\dfrac{1}{n}\sum_{i=1}^{n} u_{i2}^4 = 0. \end{cases}$$

Учитывая свойства моментов стандартизованных величин, запишем

$$\begin{cases} a_0 + a_{22} = 0, \\ r_{y2} - a_2 - a_{22}\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} u_{i2}^3 = 0, \\ \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} y_i u_{i2}^2 - a_0 - a_2\dfrac{1}{n}\sum_{i=1}^{n} u_{i2}^3 - a_{22}\dfrac{1}{2}\sum_{i=1}^{n} u_{i2}^4 = 0. \end{cases}$$

С ростом степени уравнения регрессии возрастает и степень моментов распределения параметров, используемых для определения коэффициентов. Так, для определения коэффициентов уравнения регрессии второй степени используются моменты распределения параметров до четвертой степени включительно. Известно, что точность и достоверность оценки моментов по ограниченной выборке ЭД резко снижается с ростом их порядка. Применение в уравнениях регрессии полиномов степени выше второй нецелесообразно.

Качество полученного уравнения регрессии оценивают по степени близости между результатами наблюдений за показателем и предсказанными по уравнению регрессии значениями в заданных точках пространства

параметров. Если результаты близки, то задачу регрессионного анализа можно считать решенной. В противном случае следует изменить уравнение регрессии (выбрать другую степень полинома или вообще другой тип уравнения) и повторить расчеты по оценке параметров.

При наличии нескольких показателей задача регрессионного анализа решается независимо для каждого из них.

Анализируя сущность уравнения регрессии, следует отметить следующие положения. Рассмотренный подход не обеспечивает раздельной (независимой) оценки коэффициентов – изменение значения одного коэффициента влечет изменение значений других. Полученные коэффициенты не следует рассматривать как вклад соответствующего параметра в значение показателя. Уравнение регрессии является всего лишь хорошим аналитическим описанием имеющихся ЭД, а не законом, описывающим взаимосвязи параметров и показателя. Это уравнение применяют для расчета значений показателя в заданном диапазоне изменения параметров. Оно ограниченно пригодно для расчета вне этого диапазона, т.е. его можно применять для решения задач интерполяции и в ограниченной степени для экстраполяции.

Главной причиной неточности прогноза является не столько неопределенность экстраполяции линии регрессии, сколько значительная вариация показателя за счет неучтенных в модели факторов. Ограничением возможности прогнозирования служит условие стабильности неучтенных в модели параметров и характера влияния учтенных факторов модели. Если резко меняется внешняя среда, то составленное уравнение регрессии потеряет свой смысл. Нельзя подставлять в уравнение регрессии такие значения факторов, которые значительно отличаются от представленных в ЭД. Рекомендуется не выходить за пределы одной трети размаха вариации параметра как за максимальное, так и за минимальное значения фактора.

Прогноз, полученный подстановкой в уравнение регрессии ожидаемого значения параметра, является точечным. Вероятность реализации такого прогноза ничтожна мала. Целесообразно определить доверительный интервал прогноза. Для индивидуальных значений показателя интервал должен учитывать ошибки в положении линии регрессии и отклонения индивидуальных значений от этой линии. Средняя ошибка прогноза показателя y для фактора x составит

$$m_{ош}[x_k] = \sqrt{m^2(y_k) + \sigma^2(y)},$$

где $m(y_k) = \sigma(y)\sqrt{1/n + [x_k]}$ – средняя ошибка положения линии регрессии в генеральной совокупности при $x = x_k$;

$\sigma^2(y) = \dfrac{1}{n-2}\sum_{i=1}^{n}[y_i - \mu_1(y)]^2$ – оценка дисперсии отклонения показателя от линии регрессии в генеральной совокупности;

x_k – ожидаемое значение фактора.

Доверительные границы прогноза, например, для уравнения регрессии (7.14), определяются выражением $y[x_k] \pm m_{ош}[x_k]$.

Отрицательная величина свободного члена a_0 в уравнении регрессии для исходных переменных означает, что область существования показателя не включает нулевых значений параметров. Если же $a_0 > 0$, то область существования показателя включает нулевые значения параметров, а сам коэффициент характеризует среднее значение показателя при отсутствии воздействий параметров.

Задача 7.2. Построить уравнение регрессии для пропускной способности канала по выборке, заданной в табл. 7.1.

Решение. Применительно к указанной выборке построение аналитической зависимости в основной своей части выполнено в рамках корреляционного анализа: пропускная способность зависит только от параметра "соотношение сигнал/шум". Остается подставить в выражение (7.14) вычисленные ранее значения параметров. Уравнение для пропускной способности примет вид

$$\hat{y} = 26{,}47 - 0{,}93 \times 41{,}68 \times 5{,}39/6{,}04 + 0{,}93 \times 5{,}39/6{,}03 \times x = -8{,}121 + 0{,}830x.$$

Результаты расчетов представлены в табл. 7.5.

Таблица 7.5

N пп	Пропускная способность канала	Соотношение сигнал/шум	Значение функции	Погрешность
	Y	X	ŷ	ε
1	26.37	41.98	26.72	-0.35
2	28.00	43.83	28.25	-0.25
3	27/83	42.83	27.42	0.41
4	31.67	47.28	31.12	0.55
5	23.50	38.75	24.04	-0.54
6	21.04	35.12	21.03	0.01
7	16.94	32.07	18.49	-1.55
8	37.56	54.25	36.90	0.66
9	18.84	32.70	19.02	-0.18
10	25.77	40.51	25.50	0.27
11	33.52	49.78	33.19	0.33
12	28.21	43.84	28.26	-0.05
13	28.76	44.03	28.42	0.34
14	24.60	39.46	24.63	-0.03
15	24.51	38.78	24.06	0.45

Остаточная дисперсия стандартизованной величины Y относительно 37.56 стандартизованной величины X равна $1 - 0{,}93^2 = 0{,}14$, т.е. является малой величиной. Погрешность аппроксимации и величина остаточной дисперсии показывают высокую точность линейной модели, поэтому задачу регрессионного анализа можно считать решенной.