

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ
РОССИЙСКОЙ ФЕДЕРАЦИИ
САНКТ-ПЕТЕРБУРГСКАЯ АКАДЕМИЯ УПРАВЛЕНИЯ И
ЭКОНОМИКИ
НОВОСИБИРСКИЙ ФИЛИАЛ

Воскобойников Ю.Е., Воскобойникова Т.Н.

РЕШЕНИЕ ЗАДАЧ ЭКОНОМЕТРИКИ В EXCEL

Учебное пособие

	A	B	C	D	E	F
1	Исходные данные		=ПРЕДСКАЗ(A3:\$B\$3:\$B\$12;\$A\$3:\$A\$12)			
2	x_i	y_i	\hat{y}_i			
3	8	5	5,377			
4	11	10	8,426			
5	12	10	9,443			
6	9	7	6,393			
7	8	5	5,377			
8	8	6	5,377			
9	9	6	6,393			
10	9	5	6,393			
11	8	6	5,377			
12	12	8	9,443			
13						
14						
15		b_0	-2,754	=ОТРЕЗОК(B3:B12;A3:A12)		
16		b_1	1,016	=НАКЛОН(B3:B12;A3:A12)		
17		S	1,024	=СТОШУХ(B3:B12;A3:A12)		

$$y = X\beta + \varepsilon$$

$$b = (X^T X)^{-1} X^T y$$

Новосибирск 2006

УДК 330.43(075.8)
ББК 65.вб.я73
В

Печатается по решению учебно-методического совета Новосибирского филиала Академии управления и экономики, г. Санкт-Петербург.

Рецензент:
заведующий кафедрой НГАВТ, д.э.н., профессор
А.С. Овсянников

Учебное пособие содержит основные теоретические положения и расчетные соотношения для решения часто встречающихся в эконометрике задач регрессионного анализа и анализа временных рядов. Основное внимание уделяется реализации этих соотношений в табличном процессоре Excel и учебное пособие можно рассматривать как справочник по численному решению задач эконометрики в Excel. Учебное пособие содержит копии большого числа фрагментов документов Excel, которые позволят студентам не только лучше понять и усвоить учебный материал, но и эффективно использовать Excel при выполнении курсовых работ и дипломной работы.

Кроме решения задач учебное пособие содержит набор лабораторных и контрольных работ по каждой теме, ориентированных на заочную и дистанционную формы обучения.

Учебное пособие рекомендуется студентам экономических специальностей вузов, а также для аспирантов и преподавателей по прикладной экономике и финансам.

© Ю.Е. Воскобойников,
Т.Н. Воскобойникова, 2006

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ	5
ГЛАВА 1. ЭКОНОМЕТРИЧЕСКИЕ МОДЕЛИ И ОСНОВНЫЕ ЗАДАЧИ ЭКОНОМЕТРИКИ	7
1.1 Регрессии	7
1.2. Случайные процессы и временные ряды	11
1.3. Системы одновременных уравнений	13
1.4. Типы переменных эконометрических моделей	14
1.5. Рассматриваемые задачи и вычислительная среда решения этих задач	14
1.6. Точечные оценки и их вычисление в табличном процессоре Excel	17
КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ	21
ГЛАВА 2. ПАРНАЯ РЕГРЕССИЯ	23
2.1 Постановка задачи парной регрессии	23
2.2. Выбор вида функции регрессии	28
2.3. Линейная парная регрессия и вычисление ее коэффициентов	31
2.4. Интервальные оценки функции регрессии и ее параметров	42
2.5. Значимость уравнения регрессии и коэффициент детерминации	48
2.6. Нелинейная парная регрессия	56
2.7. Построение нелинейных регрессий в Excel	64
2.8. Робастные методы оценивания и метод наименьших модулей	73
ЛАБОРАТОРНЫЕ РАБОТЫ	77
КОНТРОЛЬНАЯ РАБОТА	80
КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ ...	81
ГЛАВА 3. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ	83
3.1. Классическая линейная модель множественной регрессии	84
3.2. Оценка коэффициентов линейной модели методом наименьших квадратов	87

3.3. Интервальные оценки для функции регрессии и ее коэффициентов	97
3.4. Значимость множественной регрессии и ее коэффициентов	103
3.5. Построение линейной множественной регрессии в Excel	108
3.6. Нелинейные модели множественной регрессии. Производственная функция Кобба-Дугласа	115
3.7. Мультиколлинеарность модели множественной регрессии	123
3.8. Гетероскедастичность модели и метод, взвешенный наименьших квадратов ...	130
ЛАБОРАТОРНЫЕ РАБОТЫ	141
КОНТРОЛЬНАЯ РАБОТА	144
КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ	146
ГЛАВА 4. ВРЕМЕННЫЕ РЯДЫ	148
4.1. Временные ряды и их числовые характеристики	148
4.2. Выделение трендовой составляющей временного ряда	154
4.3. Выделение периодических составляющих временного ряда	168
4.4. Построение авторегрессионных моделей временного ряда	178
4.5. Временные ряды с коррелированными возмущениями	185
4.6. Выделение тренда временного ряда обобщенным методом наименьших квадратов	195
ЛАБОРАТОРНЫЕ РАБОТЫ	204
КОНТРОЛЬНАЯ РАБОТА	206
КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ ...	207
БИБЛИОГРАФИЧЕСКИЙ СПИСОК	208
ПРИЛОЖЕНИЕ	209

ВВЕДЕНИЕ

В последнее время специалисты, обладающие знаниями и навыками проведения прикладного экономического анализа с использованием современных математических и программных средств, пользуются спросом на рынке труда. Одной из центральных дисциплин в подготовке таких специалистов является дисциплина «Эконометрика». Дословный перевод слова «Эконометрика» означает «экономические измерения», но определение дисциплины «Эконометрика» гораздо шире этого перевода. Ниже приводятся два определения известных ученых, позволяющие получить представления о различном толковании эконометрики.

Эконометрика – это раздел экономики, занимающийся разработкой и применением статистических методов для измерений взаимосвязей между экономическими переменными (С. Фишер).

Эконометрика – это самостоятельная научная дисциплина, объединяющая совокупность теоретических результатов, приемов, методов и моделей, предназначенных для того, чтобы на базе

– экономической теории;

– экономической статистики;

– математико-статистического инструментария

придать конкретное количественное выражение общим качественным закономерностям, обусловленным экономической теорией (С.А. Айвазян).

Из этих определений можно сформулировать основную цель эконометрики: модельное описание конкретных количественных взаимосвязей, обусловленных общими качественными закономерностями, изучаемыми в экономической теории.

Составленное модельное описание называется *эконометрической моделью*. Области применения эконометрических моделей связаны с целями эконометрического планирования, основными из которых являются:

- прогноз экономических и социально-экономических показателей, характеризующих состояние и развитие анализируемой системы;

- имитация различных возможных сценариев социально-экономического развития анализируемой системы.

Заметим, что в качестве анализируемой системы могут выступать страна в целом, регионы, отрасли и корпорации, а также предприятия и фирмы.

Построение эконометрических моделей обуславливает (особенно при большом объеме исходных данных) существенный объем вычислений. На этом этапе многие исследователи сталкиваются с проблемами численной реализации необходимого вычислительного алгоритма той или иной задачи эконометрики и графической интерпретации результатов решения. Этой стороне эконометрики в учебной литературе уделяется крайне мало внимания, что затрудняет использования современных алгоритмов решения эконометрических задач на практике.

Поэтому основной целью данного пособия является *изложение численных методик решения основных задач эконометрики в вычислительной среде табличного процессора Excel XP.*

Для каждой из рассматриваемых задач эконометрики приводится необходимый теоретический материал, математическая запись алгоритма решения (т.е. формулы или расчетные соотношения), а затем даются фрагменты документов Excel XP, реализующих алгоритмы решения задачи.

При этом алгоритм решения может быть реализован путем программирования арифметических или логических выражений в ячейках электронной таблицы или путем обращения к «стандартным» функциям или модулям Excel XP. Поэтому предполагается, что читатель знаком с адресацией ячеек (относительной, абсолютной и смешанной), арифметическими операциями и программированием простейших выражений в ячейках Excel.

Данное учебное пособие, хотя и содержит необходимый теоретический материал, но *не заменяет учебник по эконометрике, а является своеобразным справочником по численному решению задач эконометрике в Excel XP.* Учебное пособие можно также

рассматривать как дополнение к основному учебнику по эконометрике, которое будет полезным при выполнении курсовых и дипломных работ, а также при самостоятельном решении практических задач эконометрики.

Кроме решения задач учебное пособие содержит набор лабораторных и контрольных работ по каждой теме, ориентированных на заочную и дистанционную формы обучения.

Предполагается, что читатель знаком с основными понятиями теории вероятностей и математической статистики. При необходимости он может обратиться к литературе [1-3].

Структура и содержание рассматриваемых задач соответствует требованиям государственного образовательного стандарта высшего профессионального образования для специальностей направления «Экономика и менеджмент», в частности для специальностей 060400 – «Финансы и кредит» и 060500 – «Бухгалтерский учет и аудит».

Глава 1. ЭКОНОМЕТРИЧЕСКИЕ МОДЕЛИ И ОСНОВНЫЕ ЗАДАЧИ ЭКОНОМЕТРИКИ

В общем случае эконометрическая модель – это вероятностно-статистическая модель, описывающая функционирование экономической или социально-экономической системы или объекта.

Важным требованием к эконометрической модели является ее *адекватность объекту-оригиналу*: модель должна с необходимой степенью точности отражать закономерности процесса функционирования реального объекта или системы.

Это требование обуславливает несколько основных типов эконометрических моделей, рассматриваемых в этом разделе.

1.1. Регрессии

В таких моделях присутствуют несколько независимых переменных (в дальнейшем обозначаемых X_1, X_2, \dots, X_k , где k – число переменных), которые влияют на значения зависимой переменной Y . Величины X_1, X_2, \dots, X_k называют *объясняющими пе-*

ременными. Значение Y зависит также от случайной величины ε , в дальнейшем называемую *случайной ошибкой* или *случайным возмущением* модели. Таким образом, получена эконометрическая модель вида

$$Y = f(x_1, x_2, \dots, x_k) + \varepsilon \quad (1.1.1)$$

Так как влияние возмущения ε носит случайный характер, то необходимо построить некоторое аналитическое выражение для функции $f(x_1, x_2, \dots, x_k)$, которую называют *объясненной частью эконометрической модели* (1.1.1), и это выражение не должно содержать возмущение ε .

Наиболее естественным выбором объясненной части случайной величины Y является ее *среднее значение – условное математическое ожидание* $M(Y | x_1, x_2, \dots, x_k)$, полученное при данном (фиксированном) наборе объясняющих переменных x_1, x_2, \dots, x_k . При таком выборе объясненной части эконометрическая модель имеет вид:

$$Y = M(Y | x_1, x_2, \dots, x_k) + \varepsilon. \quad (1.1.2)$$

Уравнение (1.1.2) часто называют *уравнением регрессионной модели*, а выражение

$$f(x_1, x_2, \dots, x_k) = M(Y | x_1, x_2, \dots, x_k) \quad (1.1.3)$$

называют *функцией регрессии*. В дальнейшем под регрессионной моделью будем понимать модель вида (1.1.2). Переменные x_1, x_2, \dots, x_k , входящие в модель (1.1.2) называют *регрессорами модели*.

Частным случаем уравнения (1.1.2) (которое называют уравнением множественной регрессии – присутствуют «много» переменных) является парная регрессионная модель, содержащая только две переменные Y и X и имеющая вид

$$Y = M(Y | x) + \varepsilon$$

Остановимся на одном важном свойстве регрессионной модели. Возьмем условное математическое ожидание от обеих частей (1.1.2):

$$M(Y | x_1, x_2, \dots, x_k) = M(Y | x_1, x_2, \dots, x_k) + M(\varepsilon | x_1, x_2, \dots, x_k).$$

Предполагается, что возмущение ε не зависит от объясняющих переменных и поэтому

$$M(\varepsilon | x_1, x_2, \dots, x_k) = M(\varepsilon).$$

Тогда из предыдущего равенства следует важное условие к возмущению регрессионной модели

$$M(\varepsilon) = 0. \quad (1.1.4)$$

Невыполнение этого условия может быть вызвано неполным учетом объясняющих переменных при определении структуры регрессионной модели.

Определение функции $f(x_1, x_2, \dots, x_k)$ существенно упрощается, если функция допускает параметризацию, т.е. зависит от набора коэффициентов (параметров), которые и необходимо определить. Например, в качестве $f(x_1, x_2, \dots, x_k)$ часто используют линейную функцию вида

$$f(x_1, x_2, \dots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

где $\beta_0, \beta_1, \dots, \beta_k$ - коэффициенты функции регрессии.

В качестве примера рассмотрим следующую ситуацию. Допустим, мы хотим продать автомобиль и решили дать объявление о продаже. Естественно, возникает вопрос: какую цену указать в объявлении. Очевидно, мы будем руководствоваться информацией о ценах, которые выставляют другие продавцы *подобных автомобилей*, а именно автомобилей, обладающих близкими значениями таких факторов, как год выпуска, пробег, мощность двигателя.

Формализуем описанную задачу: необходимо определить цену автомобиля, зависящую от ряда факторов (год выпуска, пробег и т. д.).

Цена автомобиля является *зависимой* величиной, а факторы, от которых она зависит, являются *независимыми*.

Предположим, что в нашем примере продажи автомобиля в качестве объясняющих переменных были приняты: X_1 – срок экс-

плуатации автомобиля (в годах); X_2 - пробег автомобиля (в тыс. км) и получена функция регрессии вида:

$$f(x_1, x_2) = 18000 - 1000x_1 - 5x_2 \quad (1.1.5)$$

Каково практическое применение полученного выражения?

Во-первых, выражение (1.1.5) позволяет выявить от каких факторов и в какой степени зависит рассматриваемая экономическая переменная – цена на автомобиль. Во-вторых, позволяет прогнозировать цену на продаваемый автомобиль, если известны его основные параметры (т. е. значения переменных x_1, x_2). Например, предположим, что $x_1=5, x_2=80$. Подставляя эти значения в (1.1.5), получаем

$$f(5, 80) = 18000 - 1000 \cdot 5 - 5 \cdot 80 = 12600 \text{ (усл. ден. ед.)}.$$

Теперь менеджеру не составляет большого труда определить *ожидаемую цену* вновь поступившего для продажи автомобиля, даже если его год выпуска и пробег не встречался ранее в данном автомобильном салоне или в автомобильном магазине.

К **основным задачам** построения регрессионной модели следует отнести:

- отбор значимых независимых переменных;
- выбор вида функции $f(x_1, x_2, \dots, x_k)$;
- построение оценок b_0, b_1, \dots, b_k для коэффициентов $\beta_0, \beta_1, \dots, \beta_k$;
- построение доверительных интервалов для $\beta_0, \beta_1, \dots, \beta_k$ и функции регрессии;
- проверка значимости вычисленных оценок и построенного уравнения регрессии.

Исходные данные для решения этих задач образуют *пространственную выборку*.

Предположим, что эконометрическая модель включает величины Y, X_1, X_2, \dots, X_k , над которыми выполнены n наблюдений (как правило, над n - объектами). Тогда результаты наблюдений могут быть представлены таблицей

$$\begin{vmatrix} x_{11} & x_{12} \dots x_{1k} & y_1 \\ x_{21} & x_{22} \dots x_{2k} & y_2 \\ \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} \dots x_{nk} & y_n \end{vmatrix} \quad (1.1.5)$$

где x_{ij} означает результат измерения j -ой переменной x_j в i -ом наблюдении (эксперименте). Такой тип данных называется *пространственной выборкой* или данными поперечного среза (cross-section data). Данные не имеют временного параметра, и порядок их следования в таблице (1.1.5) не существенен.

1.2. Случайные процессы и временные ряды

При исследовании поведения экономической системы во времени независимой переменной является временной параметр (час, день, месяц, год), обозначаемый в дальнейшем τ . Тогда зависимая переменная будет сочетать два фактора: а) при фиксированном времени τ является случайной величиной; б) является функцией аргумента τ . Такую величину называют случайной функцией или случайным процессом и ее будем обозначать, как $Y(\tau)$. В общем виде модели случайных процессов описывающих поведение экономических систем можно записать в виде:

$$Y(\tau) = q(\tau) + \xi(\tau), \quad (1.2.1)$$

где $\xi(\tau)$ - случайная составляющая, которая в каждый момент времени имеет нулевое среднее, т.е. $M[\xi(\tau)] \equiv 0$. Детерминированная (регулярная) составляющая $q(\tau)$ допускает в общем случаи следующую запись:

$$q(\tau) = t(\tau) + s(\tau) + p(\tau), \quad (1.2.2)$$

где $t(\tau)$ - тренд, как правило, параметризованная функция (например, параболическая $t(\tau) = \beta_0 + \beta_1 t + \beta_2 t^2$); $s(\tau)$ - сезонная составляющая; $p(\tau)$ - периодическая составляющая. Часто тренд

$t(\tau)$ называют *полиномиальной составляющей*, а $s(\tau)$ и $p(\tau)$ - *тригонометрическими составляющими* случайного процесса.

На практике детерминированная составляющая $q(\tau)$ может включать одно (например, тренд $t(\tau)$) или два слагаемых (например, тренд $t(\tau)$ и сезонную составляющую $s(\tau)$).

Кроме аддитивной модели (1.2.1) возможна мультипликативная модель случайного процесса:

$$Y(\tau) = q(\tau) \cdot \xi(\tau), \quad (1.2.3)$$

Временной выборкой случайного процесса (в зарубежной литературе *time-series data*) называется совокупность наблюдений $\{y(\tau_1), y(\tau_2), \dots, y(\tau_n)\}$ случайной величины $Y(\tau)$ в дискретные моменты времени $\tau_i, i = 1, 2, \dots, n$.

Например, взяты n выпусков некоторого рекламного издания, и они упорядочены по дате выпуска. Из каждого выпуска взята цена автомобиля определенного класса. В этом случае получаем временную выборку, составленную из наблюдений $y(\tau_1), y(\tau_2), \dots, y(\tau_n)$, где τ_i - время выхода i -го рекламного издания.

Временная зависимость данных делает существенным порядок следования наблюдаемых значений $y(\tau_i)$ во временной выборке. Это означает, что перестановка $y(\tau_i)$ во временной выборке может существенно сказаться на характеристиках исследуемой зависимости $Y(\tau)$.

К основным задачам анализа случайного процесса относятся:

- выделение детерминированной составляющей $q(\tau)$;
- выделение тренда $t(\tau)$;
- выделение сезонной составляющей $s(\tau)$;
- выделение периодической составляющей $p(\tau)$;
- построение модели случайной составляющей $\xi(\tau)$;
- прогнозирование развития изучаемого процесса на основе построенной модели временного ряда.

Здесь под выделением понимается не только разделение детерминированной составляющей $q(\tau)$ на присутствующие в ней слагаемые, но и построение соответствующего математического описания для каждого слагаемого $t(\tau), s(\tau), p(\tau)$.

Временным рядом (или *дискретным случайным процессом*) называется совокупность случайных величин $\{Y(\tau_1), Y(\tau_2), \dots, Y(\tau_n)\}$, сформированную из случайной величины $Y(\tau)$ в моменты $\tau = \tau_i, i = 1, 2, \dots, n$. Учитывая что моменты τ_i жестко фиксированы временной ряд можно записать как совокупность $\{Y_1, Y_2, \dots, Y_n\}$, состоящую из n случайных величин $Y_i = Y(\tau_i)$. Задачи анализа временного ряда аналогичны перечисленным выше задачам анализа случайного процесса. Временная выборка временного ряда также состоит из n наблюдений (т.е. уже не случайных величин) $y_i = Y(\tau_i), i = 1, 2, \dots, n$.

1.3. Системы одновременных уравнений

Такие системы могут состоять из тождеств и регрессионных уравнений, каждое из которых, кроме «собственных» объясняющих переменных, может включать в себя объясняемые переменные из других уравнений системы.

Примером может служить модель спроса и предложения. Пусть $Q_D(t)$ – спрос на товар в момент времени t ; $Q_S(t)$ – предложение товара в момент времени t ; $P(t)$ – цена на товар в момент времени t ; $Y(t)$ – доход в момент t . Система имеет вид:

$$Q_S(t) = \alpha_1 + \alpha_2 P(t) + \alpha_3 P(t-1) + \varepsilon(t) \quad (\text{предложение}). \quad (1.3.1)$$

$$Q_D(t) = \beta_1 + \beta_2 P(t) + \beta_3 Y(t) + u(t) \quad (\text{спрос}). \quad (1.3.2)$$

$$Q_S(t) = Q_D(t) \quad (\text{равновесие}). \quad (1.3.3)$$

Цена на товар $P(t)$ и спрос на товар $Q(t) = Q_S(t) = Q_D(t)$ определяются из уравнения модели. Объясняющими переменными являются доход $Y(t)$ и значение цены $P(t-1)$ в предыдущий момент времени $t-1$.

К основным задачам, возникающим при построении таких моделей можно отнести:

- определение вида входящих функций регрессии;
- оценивание коэффициентов регрессионных зависимостей;
- определение решений, удовлетворяющих системе тождеств и регрессионных уравнений.

1.4. Типы переменных эконометрических моделей.

Применимо к рассмотренным моделям можно ввести следующую классификацию переменных:

- *экзогенные переменные* – переменные, задаваемые из вне рассматриваемой системы и в определенном смысле управляемые;
- *эндогенные переменные* – переменные, значения которых формируются в процессе и внутри функционирования анализируемой системы;
- *лаговые эндогенные переменные* – переменные, входящие в уравнения анализируемой системы, но измерены в прошлые моменты, а, следовательно, являются уже известными заданными.
- *предопределенные переменные* – все экзогенные переменные модели и лаговые эндогенные переменные.

Обобщая изложенное, можно сказать, что *эконометрическая модель позволяет объяснить поведение эндогенных переменных в зависимости от значений экзогенных и лаговых эндогенных переменных*.

1.5. Рассматриваемые задачи и вычислительная среда решения этих задач

В данном учебном пособии рассматривается решение основных задач, возникающих при *построении регрессионных моделей и анализе временных рядов*. Для решения задач, связанных с системами одновременных уравнений используются те или иные варианты метода наименьших квадратов (например, косвенный метод наименьших квадратов, двухшаговый метод наименьших квадратов). Численная реализация метода наименьших квадратов

подробно рассматривается в главах 2,3 и поэтому реализация метода наименьших квадратов при решении систем одновременных уравнений в учебном пособии отдельно не рассматривается.

Решение рассматриваемых задач эконометрики сопряжено с выполнением большого числа вычислительных операций и хранения большого объема данных (пространственной или временной выборки). Поэтому для успешного решения этих задач необходима некоторая вычислительная среда, в которой будут выполняться необходимые вычисления. В данном учебном пособии в качестве такой среды используется табличный процессор (проще – электронная таблица) Excel XP.

Возникает вопрос – почему табличный процессор Excel, а не универсальные статистические пакеты, такие как Statgraphics, EViews, Statistica и т.д.?

По нашему мнению использование табличного процессора Excel является более предпочтительным по следующим причинам:

- табличный процессор Excel является доступной русифицированной лицензионной программой, в то время как названные статистические пакеты труднодоступны и в основном являются контрафактными;
- использование табличного процессора Excel подразумевает программирование расчетных выражений, что способствует лучшему усвоению расчетных соотношений и методов эконометрического моделирования.

Табличный процессор Excel предоставляет две возможности для реализации вычислений при построении эконометрических моделей:

- программирование необходимых вычислений в ячейках Excel;
- обращение к соответствующим функциям и модулям Excel.

Первый подход более универсальный (так как позволяет реализовать любой вычислительный алгоритм), но требует определенных затрат времени и знаний основ алгоритмизации вычислений. Второй путь более простой, но ограничен имеющимся набором «стандартных» функций и модулей Excel.

В учебном пособии будут использоваться обе рассмотренные возможности реализации требуемого вычислительного алгоритма.

Поэтому предполагается, что читатель имеет достаточные навыки для реализации вычислений в Excel с использованием:

- программирования арифметических выражений в ячейках электронной таблицы;
- функций Excel (в основном математических и статистических).

Замечание 1.5.1. В тексте пособия при описании той или иной функции в качестве *формальных параметров* используются имена переменных, определенные в тексте пособия. При обращении к функции в качестве *фактических параметров* могут использоваться константы, адреса ячеек, диапазоны адресов и арифметические выражения. Например, описание функции для вычисления среднего арифметического значения (выборочного среднего) имеет вид:

$$\text{CPЗНАЧ}(x_1; x_2; \dots; x_m),$$

где x_1, x_2, \dots, x_m – формальные параметры, число которых не превышает 30 ($m \leq 30$). Для вычисления среднего значения величин, находящихся в ячейках B3, B4, B5, B6, C3, C4, C5, C6, обращение к функции в соответствующей ячейке имеет вид

$$=\text{CPЗНАЧ}(B3:B6; C3:C6),$$

т.е. в качестве фактических параметров используются два диапазона ячеек.

Замечание 1.5.2. Так как в запрограммированной ячейке выводится результат вычислений и не видно самого запрограммированного выражения, то в некоторых случаях рядом с результатом приводится (в другой ячейке) запрограммированное выражение (своеобразный комментарий к выполняемым вычислениям). В случаях, когда не очевидно к какой ячейке относится приводимое выражение, используется стрелка, указывающая на нужную ячейку.

В качестве примера такого комментария на рис. 1.1 показан фрагмент документа Excel, вычисляющего среднее значение чисел, размещенных в ячейках В3:В6, С3:С6. Результат вычислений находится в ячейке В8, а правее показано выражение, запрограммированное в этой ячейке.

	A	B	C	D
1				
2				
3		-0,278	0,227	
4		-0,176	0,100	
5		0,159	0,181	
6		-0,082	-0,172	
7				
8			-0,005	
9				
10		=СРЗНАЧ(В3:В6;С3:С7)		
11				

Рис. 1.1. Фрагмент вычисления среднего значения

1.6. Точечные оценки и их вычисление в табличном процессоре Excel

При построении эконометрических моделей часто используются так называемые *точечные* (или *выборочные*) оценки различных коэффициентов модели. Поэтому кратко остановимся на понятии точечной оценки, ее свойствах и ее вычислении в Excel.

Определение точечной оценки. Пусть над непрерывной случайной величиной X проведены n наблюдений, т.е. получены n значений x_1, x_2, \dots, x_n , которые составляют *выборочную совокупность* объемом n . Обозначим через θ некоторый неизвестный параметр закона распределения величины X (например, математическое ожидание). В качестве статистической оценки $\hat{\theta}_n$ этого параметра примем некоторую функцию от значений x_1, x_2, \dots, x_n , т.е. $\hat{\theta}_n = \varphi(x_1, x_2, \dots, x_n)$. Нижний индекс обозначает объем выборки. Такая оценка, представленная одним числом, называется *точечной*.

Свойства точечных оценок. В отличие от параметра θ оценка $\hat{\theta}_n$ является случайной величиной (как функция случайных величин) и очевидно, что $\hat{\theta}_n$ в общем случае не совпадает с θ . Для того чтобы $\hat{\theta}_n$ была «хорошей» оценкой для θ необходимо, чтобы она была:

- *несмещенной*;
- *эффективной*;
- *состоятельной*.

Оценка $\hat{\theta}_n$ называется *несмещенной*, если $M(\hat{\theta}_n) = \theta$, т.е. среднее значение оценки $\hat{\theta}_n$ равно оцениваемому параметру. В противном случае оценка называется *смещенной*. Видно, что требование несмещенности гарантирует отсутствие систематических ошибок процедуры оценивания.

Возможные значения несмещенной оценки $\hat{\theta}_n$ рассеяны вокруг. Оценка $\hat{\theta}_n$ называется *эффективной*, если среди всех других несмещенных оценок она имеет наименьшую дисперсию, т.е. в меньшей степени отклонена от θ .

Оценка $\hat{\theta}_n$ называется *состоятельной*, если при увеличении объема выборки n дисперсия оценки будет уменьшаться (а точность оценки будет увеличиваться).

Рассмотрим часто используемые в эконометрике точечные оценки числовых характеристик случайной величины X .

Точечные оценки для числовых характеристик случайной величины. Оценкой для математического ожидания $M(X)$ случайной величины является *выборочное среднее*

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^n x_i. \quad (1.6.1)$$

Можно показать, что оценка \bar{x} является несмещенной, эффективной и состоятельной, т.е. удовлетворяет всем требованиям «хорошей» оценки. В дальнейшем операцию усреднения каких-

либо значений будем обозначать горизонтальной чертой над обозначением этих значений. Например, $\bar{x}^2 = \frac{1}{n} \sum_{i=1}^n x_i^2$.

Оценкой для дисперсии $\sigma_X^2 = D(X)$ случайной величины X является **выборочная дисперсия**

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (1.6.2)$$

На практике для вычисления s_X^2 часто используют следующую формулу:

$$s_X^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - (\bar{x})^2 = \bar{x}^2 - (\bar{x})^2. \quad (1.6.3)$$

Оценка s_X^2 является состоятельной, но смещенной. Несмещенная оценка имеет вид:

$$s_X^2 = \frac{n}{n-1} s_X^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2. \quad (1.6.4)$$

При большом объеме выборки n отличие между этими оценками пренебрежимо мало.

Рассмотрим точечную оценку m_{XY} для корреляционного момента μ_{XY} и точечную оценку r_{XY} для коэффициента корреляции ρ_{XY} случайных величин X, Y , определяемых по выборке объемом n . Оценки вычисляются по следующим формулам:

$$m_{XY} = \overline{xy} - \bar{x} \cdot \bar{y}, \quad r_{XY} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_X \cdot s_Y}, \quad (1.6.5)$$

где $\overline{xy} = \frac{1}{n} \cdot \sum_{i=1}^n x_i y_i$.

Вычисление точечных оценок в Excel. Точечные оценки можно вычислить двумя способами:

- программируя в ячейке соответствующее арифметическое выражение;
- используя соответствующие статистические функции Excel.

Рассмотрим на примерах эти два способа.

Пример 1.6.1. На основе наблюдений получена выборка объемом $n = 12$ значений случайной величины X , приведенная на рис. 1.2 в ячейках B2, B3, ..., B13. Вычислить точечные оценки для математического ожидания и дисперсии, используя выражения (1.6.1), (1.6.2) и (1.6.4).

Решение. Первоначально введем в таблицу исходные данные следующим образом: в ячейки A2:A13 занесем порядковые номера выборочных значений, а в ячейки B2:B13 – сами выборочные значения (рис. 1.2). По этим данным построим диаграмму, называемую диаграммой рассеяния (рис. 1.2). Далее, в ячейке B14 запрограммируем формулу (1.6.1), а в ячейках C2:C13 вычислим квадраты разностей $(x_i - \bar{x})^2$. При этом обратите внимание на использование абсолютного адреса \$B\$14 ячейки, где находится значение \bar{x} . Затем в ячейке C14 вычислим несмещенную точечную оценку (1.6.4). Заметим, что математическое ожидание случайной величины (выборочные значения которой занесены в столбе B) равно 0, а дисперсия равна $1/12 = 0.0833$. Видно отличие значений точечных оценок от «точных» значений числовых характеристик случайной величины. ●

Для вычисления точечных оценок для математического ожидания и дисперсии в Excel определены следующие статистические функции:

- = СРЗНАЧ(*диапазон ячеек*) – реализует формулу (1.6.1);
- = ДИСП(*диапазон ячеек*) – реализует формулу (1.6.4);
- = ДИСПР(*диапазон ячеек*) – реализует формулу (1.6.2).

Пример 1.6.2. По выборочным данным примера 1.6.1 вычислить точечные оценки для математического ожидания и дисперсии, используя статистические функции Excel.

Решение. В ячейке G13 запрограммируем функцию СРЗНАЧ, в ячейке G14 функцию ДИСП, а в ячейке G15 функцию ДИСПР (см. рис. 1.2). ●

Для вычисления *выборочного корреляционного момента* m_{XY} используется статистическая функция Excel:

- = КОВАР(*диапазон ячеек X; диапазон ячеек Y*).

Для вычисления *выборочного коэффициента корреляции* r_{XY} используются статистические функции Excel:

=КОРРЕЛ(диапазон ячеек X; диапазон ячеек Y);

=ПИРСОН(диапазон ячеек X; диапазон ячеек Y),

Эти функции дают один и тот же результат.

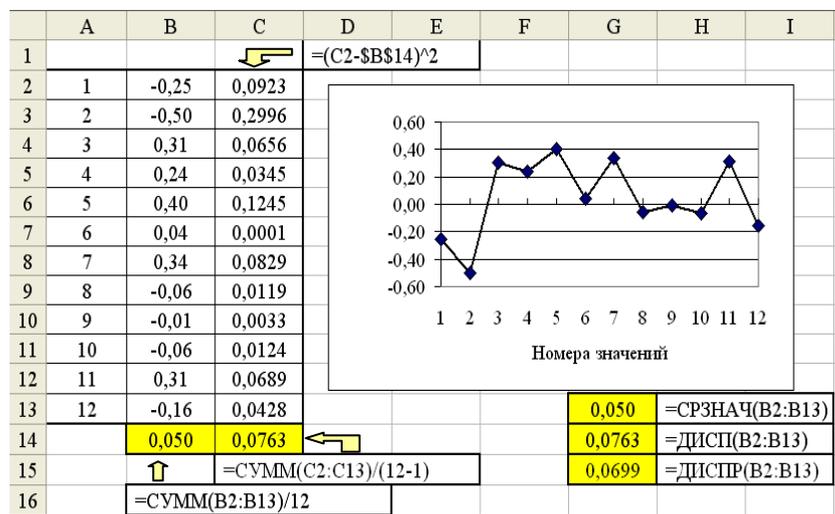


Рис. 1.2. Вычисление точечных оценок в Excel

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Кроме срока эксплуатации и пробега определите третий фактор, влияющий на цену автомобиля, и введите его как третью объясняющую переменную в уравнение регрессии (1.1.5) с соответствующим коэффициентом. Используйте новое уравнение для прогнозирования цены для двух различных наборов значений объясняющих переменных.

2. Составьте линейное уравнение регрессии (вида (1.2.1)), определяющего стоимость вторичного жилья в зависимости от срока эксплуатации здания и удаленности от центра.

3. В таблице 1.1 приведены выборочные значения случайных величин X, Y (объем выборки равен 10).

Таблица 1.1

X	0,039	0,066	-0,076	-0,450	0,211	0,021	-0,312	-0,412	0,146	0,385
Y	4,084	-2,707	2,224	1,770	2,293	1,271	3,395	2,999	1,021	3,171

Используя табличный процессор Excel по этим двум выборкам необходимо вычислить:

- Выборочное среднее (оценка математического ожидания) для каждой из случайных величин;
- Выборочную дисперсию (оценка дисперсии) для каждой из случайных величин.

Сделать вывод о соотношении между собой математических ожиданий и дисперсии случайных величин X, Y .

Рекомендация. Для вычисления характеристик используйте стандартные функции Excel.

4. Используя табличный процессор Excel, по двум выборкам случайных величин X, Y , приведенных в таблице 1.1 (объем выборки равен 10) вычислить:

- Выборочный корреляционный момент m_{XY} (оценка корреляционного момента μ_{XY});
- Выборочный коэффициент корреляции r_{XY} (оценка коэффициента корреляции ρ_{XY}).

Рекомендация. Вычисления точечных оценок m_{XY}, r_{XY} осуществить двумя способами: используя стандартные функции Excel и программируя выражения (1.6.5).

Глава 2. ПАРНАЯ РЕГРЕССИЯ

В этой главе решаются задачи построения регрессионных моделей для случая, когда объясненная часть $f(X)$ модели (1.1.1) является функцией одной независимой переменной X . Рассматриваемые задачи включают установление формы зависимости между переменными, оценку функции регрессии (включая оценку параметров), проверку достоверности построенной функции регрессии и ее параметров, оценку неизвестных значений (прогноз значений) зависимой переменной.

2.1. Постановка задачи парной регрессии

Рассмотрим некоторый экономический объект (процесс, явление, систему) и выделим только две переменные, характеризующие этот объект. Независимая (объясняющая) переменная X оказывает воздействие на значения переменной Y , которая, таким образом, является зависимой переменной.

Далее мы располагаем n парами выборочных наблюдений над величинами X, Y (т. е. имеем пространственную выборку):

$$x_1, x_2, \dots, x_n; \quad (2.1.1)$$

$$y_1, y_2, \dots, y_n.$$

Напомним (см. параграф 1.1), что функция $f(x)$ называется функцией регрессии Y по X , если она описывает изменение условного среднего значения переменной Y в зависимости от значения переменной x :

$$f(x) = M(Y|x). \quad (2.1.2)$$

Таким образом, в качестве объясненной части эконометрической модели (1.1.1) выступает регрессия (2.1.2), а моделью рассматриваемой в этой главе является уравнение регрессионной связи между Y и X вида

$$Y = f(x) + \varepsilon. \quad (2.1.3)$$

Выборка (2.1.1) соответствует модели измерений:

$$y_i = f(x_i) + \varepsilon_i; \quad i = 1, 2, \dots, n. \quad (2.1.4)$$

Присутствие в модели (2.1.3) случайного члена ε , который будем называть возмущением или ошибкой модели, обусловлено следующими причинами:

1. *Ошибки спецификации модели*, обусловленные не включением важных объясняющих переменных, неправильную функциональную спецификацию модели. Математическое ожидание таких ошибок отличается от нуля.

2. *Ошибки измерения*, обусловленные погрешностью сбора и измерения исходных данных. Математическое ожидание таких ошибок может равняться нулю.

3. *Ошибки, связанные со случайностью человеческих реакций*. Обусловлено тем, что поведение и непосредственное участие человека в сборе и подготовке данных может внести определенные погрешности. Математическое ожидание таких ошибок может равняться нулю.

Условия Гаусса-Маркова на парную регрессионную модель. Перечислим ряд предположений относительно рассматриваемой регрессионной модели (2.1.3) и модели измерений (2.1.4), известных как условия Гаусса-Маркова:

P1. Объясняющая переменная X является неслучайной (детерминированной) величиной.

P2. Возмущения ε_i имеют нулевое среднее, т.е.

$$M(\varepsilon_i) = 0, \quad i = 1, 2, \dots, n. \quad (2.1.5)$$

Это условие означает, что случайный член ε может быть отрицательным или положительным, но он не должен иметь систематического смещения. Условие непосредственно вытекает из условия (1.1.4), полученного для общего уравнения регрессионной модели.

P3. Корреляционные моменты случайных величин $\varepsilon_i, \varepsilon_j$ удовлетворяют условию

$$M(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2, & \text{если } i = j; \\ 0, & \text{если } i \neq j. \end{cases} \quad (2.1.6)$$

Первая строка означает *постоянство дисперсии возмущений* ε_i , и это свойство называют *гомоскедастичностью*. Зависимость дис-

персии возмущения ε_i от номера наблюдения i или от величины переменной X называется *гетероскедастичностью*. Характерные диаграммы рассеяния для случаев гомоскедастичности и гетероскедастичности показаны на рис. 2.1 а) и б) соответственно.

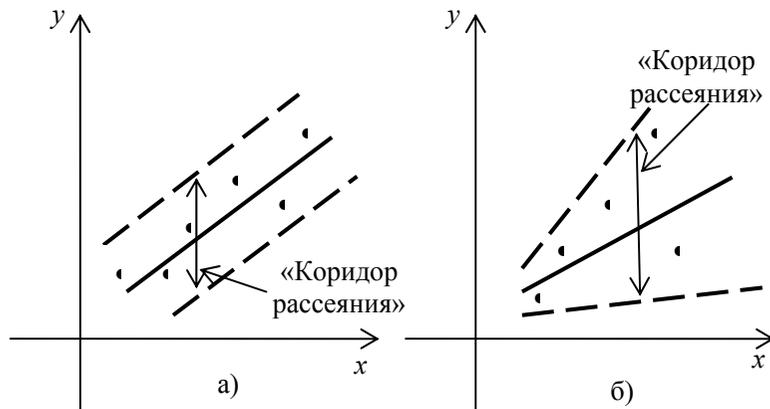


Рис. 2.1. Диаграммы рассеяния

Забегая вперед, заметим, что, если условие гомоскедастичности не выполняется, то вычисленные коэффициенты являются неэффективными оценками, хотя и несмещенными. Вторая строка означает *отсутствие корреляции между двумя значениями ε_i и ε_j при $i \neq j$* .

При этих допущениях и на основе выборочных значений (2.1.1) необходимо построить функцию $f(x) = M(Y|x)$. Однако по выборке ограниченного объема (2.1.1) невозможно точно вычислить условное математическое ожидание $M(Y|x)$, а можно только его оценить. Поэтому по выборке ограниченного объема можно построить только оценку для $f(x)$, обозначаемую в дальнейшем как \hat{f} или $\hat{f}(x)$ и называемую *выборочным уравнением регрессии* (также называемую эмпирическим уравнением регрессии). Кроме этого необходимо проверить соответствие (адекватность) выборочного уравнения регрессии исходным данным и проверить другие статистические гипотезы, характеризующие

«качество» построенного выборочного уравнения регрессии как оценки для $f(x)$.

Замечание 2.1.1. Для сокращения в дальнейшем $f(x)$ будем называть функцией регрессии, а выборочное уравнение регрессии – уравнением регрессии.

Ниже решение задач парного регрессионного анализа будет иллюстрироваться на пространственной выборке следующего примера [5].

Пример 2.1.1. Для определения зависимости между сменной добычей угля на одного рабочего (переменная Y , измеряемая в тоннах) и мощностью угольного пласта (переменная X , измеряемая в метрах) на 10 шахтах были проведены исследования, результаты которых представлены таблицей 2.1. ☉

Таблица 2.1

i	1	2	3	4	5	6	7	8	9	10
x_i	8	11	12	9	8	8	9	9	8	12
y_i	5	10	10	7	5	6	6	5	6	8

Построение выборочного уравнения регрессии содержит два этапа:

- **определение вида функции регрессии $f(x)$** (линейная, полиномиальная и т. д.) и соответственно вида выборочного уравнения регрессии;
- **вычисление коэффициентов уравнения регрессии**, являющихся оценками для коэффициентов функции регрессии.

Заметим, что построение уравнения регрессии подразумевает *наличие между переменными X и Y статистической зависимости*. Как определить степень такой зависимости?

Для этого можно использовать корреляционный момент (часто называемый ковариацией), определяемый выражением

$$\mu_{XY} = M((X - m_X)(Y - m_Y)), \quad (2.1.7)$$

где $M(\cdot)$ - означает оператор математического ожидания. Напомним, что математическое ожидание $m_X = M(X)$ и дисперсия

$\sigma_X^2 = D(X)$ случайной величины X , имеющей плотность распределения $p(x)$, определяются соотношениями:

$$m_X = \int xp(x)dx, \quad \sigma_X^2 = \int (x - m_X)^2 p(x)dx = M(X^2) - (m_X)^2,$$

где интегралы вычисляются по всему интервалу значений случайной величины.

Таким образом, корреляционный момент характеризует среднее значение произведений отклонений X, Y от их математических ожиданий. Если $\mu_{XY} = 0$, то величины X и Y называют некоррелированными. Корреляционный момент есть величина размерная, что затрудняет его использование. Этому недостатка лишен коэффициент корреляции, определяемый по формуле:

$$\rho_{XY} = \frac{\mu_{XY}}{\sigma_X \sigma_Y} \quad (2.1.8)$$

Коэффициент корреляции величина безразмерная и характеризует *тесноту линейной зависимости* между величинами X и Y .

Свойства коэффициента корреляции:

- $-1 \leq \rho_{XY} \leq 1$;
- $\rho_{XY} = 0$, если X и Y некоррелированы;
- если $\rho_{XY} = -1$ или $\rho_{XY} = 1$, то между X и Y существует линейная функциональная (не случайная) связь.

Замечание 2.1.2. Значения ρ_{XY} близкие к нулю означают *отсутствие линейной статистической зависимости между переменными X и Y* . Но при этом вполне возможно наличие *нелинейной статистической зависимости между X и Y* .

Если даны выборочные значения $\{x_i, y_i\}$, $i = 1, \dots, n$, случайных величин X и Y , то оценкой для ρ_{XY} является выборочный коэффициент корреляции r_{XY} , который можно вычислить, используя следующую функцию Excel (формула для вычисления имеет вид (2.3.15)):

$$\text{КОРРЕЛ}(\text{диапазон ячеек } X; \text{ диапазон ячеек } Y). \quad (2.1.9)$$

Например, применение этой функции к данным таблицы 2.1 дало значение $r_{XY} = 0.86$, что означает наличие линейной статистической зависимости между X и Y .

2.2. Выбор вида функции регрессии

Построение оценки для функции $f(x)$ существенно упрощается, если функция $f(x)$ допускает *параметризацию*, т.е. зависит от набора коэффициентов (параметров), которые и необходимо определить. На практике в качестве функции $f(x)$ для парной регрессии используются следующие виды функций:

1. Линейная $- f(x) = \beta_0 + \beta_1 x$. (2.2.1)

2. Полиномиальная k -го порядка $-$

$$f(x) = \beta_0 + \beta_1 x + \dots + \beta_k x^k. \quad (2.2.2)$$

4. Экспоненциальная $- f(x) = \beta_0 \exp(\beta_1 x)$. (2.2.3)

5. Степенная $- f(x) = \beta_0 x^{\beta_1}$. (2.2.4)

6. Показательная $- f(x) = \beta_0 \beta_1^x$. (2.2.5)

7. Логарифмическая $- f(x) = \beta_0 + \beta_1 \ln x$. (2.2.6)

Кроме этих функций на практике находят применение и более сложные функции, такие как:

$$f(x) = \frac{1}{\beta_0 + \beta_1 x}; \quad f(x) = \frac{\beta_0}{1 + \beta_1 e^{-\beta_2 x}}.$$

Возникает вопрос: какой вид функции взять? Для ответа на этот вопрос используют следующие подходы.

1. **Аналитический.** Анализируется априорная информация о содержательной экономической сущности исследуемой зависимости. На основе этого анализа выбирается подходящий вид функции $f(x)$.

Например, для шахт другого угольного района было установлено, что зависимость между производительностью шахтера и толщиной угольного пласта является линейной. Поэтому в качестве функции $f(x)$ для примера 2.1.1 также можно принять линейную функцию $f(x) = \beta_0 + \beta_1 x$.

2. Графический. В декартовой системе координат строят n точек с координатами (x_i, y_i) , определяемыми заданной пространственной выборкой. Построенная диаграмма называется диаграммой рассеяния (или полем корреляции). Затем на основе визуального анализа расположения точек принимают решение о типе функции $f(x)$.

Заметим, что из-за наличия случайной составляющей ε_i , значения y_i имеют определенный разброс и не нужно подбирать $f(x)$, проходящую через все точки (тем самым возмущение ε было бы включено в функцию регрессии $f(x)$). Необходимо, чтобы $f(x)$ в «равной степени близости» проходила около всех точек диаграммы рассеяния.

Пример 2.2.1. По пространственной выборке примера 2.1.1 построить диаграмму рассеяния и определить тип функции $f(x)$.

Строим декартову систему координат и наносим точки с координатами (x_i, y_i) (см. рис. 2.2). Из этого рисунка видно, что с увеличением x_i возрастают значения y_i , и это возрастание носит линейный характер. Поэтому в качестве $f(x)$ можно принять линейную функцию. Для иллюстрации этого вывода на рисунке нанесена функция регрессии $f(x) = -2.75 + 1.016x$, которая «достаточно близко» проходит от точек (x_i, y_i) . ●

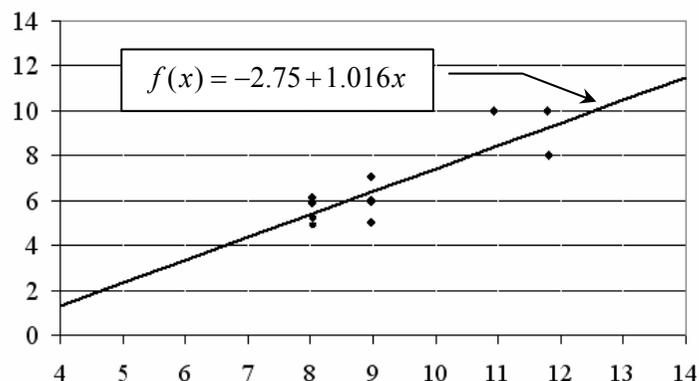


Рис 2.2. Диаграмма рассеяния и линейная регрессия

3. Экспериментальный. Для нескольких наиболее подходящих функций регрессии строятся соответствующие уравнения регрессии (т.е. вычисляются коэффициенты уравнения регрессии). Выбор «наилучшего» уравнения осуществляется путем сравнения некоторых показателей, характеризующих близость уравнения регрессии к заданным значениям y_i . Часто в качестве такого показателя используют следующую сумму квадратов:

$$Q_e = \sum_{i=1}^n (y_i - \hat{f}_i)^2,$$

где \hat{f}_i – значение уравнения регрессии при $x = x_i$. Однако, при таком выборе вида регрессии необходимо помнить о приведенном ниже *принципе минимальной сложности*. В силу своей трудоемкости экспериментальный метод подразумевает применение вычислительной техники и соответствующего программного обеспечения (например, табличного процессора Excel).

Принцип минимальной сложности можно сформулировать следующим образом: при наличии нескольких альтернативных функций $f(x)$ первоначально принимают самую «простую» и, если она не адекватна заданной выборке, то переходят к более сложной функции $f(x)$. При этом в качестве критерия сложности можно принять количество коэффициентов функции $f(x)$.

В примере 2.2.1 в качестве $f(x)$ можно принять линейную функцию $\beta_0 + \beta_1 x$ и параболическую $\beta_0 + \beta_1 x + \beta_2 x^2$, но первоначально следует рассмотреть линейную регрессию $f(x) = \beta_0 + \beta_1 x$.

Дополнением принципа минимальной сложности является следующая рекомендация: *число наблюдений n должно в 6 – 7 раз превышать число вычисляемых коэффициентов функции регрессии при объясняющей переменной X* . Так для расчета коэффициентов параболической регрессии уже потребуется не менее 14 наблюдений (для линейной регрессии – всего 7). При нарушении этой рекомендации вычисленные коэффициенты могут иметь большие дисперсии и оказываются статистически незначимыми.

Таким образом, после определения вида регрессии мы имеем функцию $f(x)$ с неизвестными коэффициентами β_j . Следующим этапом является вычисление оценок для этих коэффициентов. В качестве таких оценок выступают коэффициенты b_j выборочного уравнения регрессии, вид которого однозначно определяется видом функции регрессии. Так для функции (2.2.1) уравнение регрессии имеет вид $\hat{f}(x) = b_0 + b_1x$, для функции (2.2.2) - $\hat{f}(x) = b_0 + b_1x + b_2x^2 + \dots + b_kx^k$ и т.д.

Сначала рассмотрим оценивание коэффициентов линейной функции регрессии.

2.3. Линейная парная регрессия и вычисление ее коэффициентов

Предположим, что регрессия (2.1.3) является линейной функцией относительно объясняющей переменной x , т. е.

$$f(x) = \beta_0 + \beta_1x. \quad (2.3.1)$$

Напомним, что $f(x)$ является условным математическим ожиданием, т. е. вычисляется усреднением по большому ансамблю значений Y при каждом значении величины X . В нашем распоряжении есть только одна выборка, т. е. каждому значению X соответствует одно значение Y . По этой выборке можно построить только «выборочную» регрессию вида

$$\hat{f}(x) = b_0 + b_1x. \quad (2.3.2)$$

Выражение (2.3.2) в дальнейшем будем называть уравнением регрессии и для упрощения записи часто $\hat{f}(x)$ будем обозначать \hat{f} . Коэффициенты b_0, b_1 являются оценками β_0, β_1 и желательно, чтобы они обладали «хорошими» свойствами несмещенности, состоятельности и эффективности, определенные в параграфе 1.6.

Вопрос: «Как же вычислить оценки для коэффициентов β_0, β_1 с такими свойствами».

Очевидно, что, если функция $\hat{f}(x)$ соответствует (2.3.1) и (2.1.3), то $\hat{f}(x)$ должно «достаточно близко» проходить от точек (x_i, y_i) . Мера близости характеризуют некоторым функционалом. В зависимости от вида функционала, определяющего близость $\hat{f}(x)$ к точкам (x_i, y_i) , существует несколько методов вычисления коэффициентов b_0, b_1 . На практике в большинстве случаев используется *метод наименьших квадратов* (МНК), иногда называемый *обыкновенным МНК* или *классическим МНК*.

Метод наименьших квадратов. Согласно этому методу неизвестные коэффициенты b_0, b_1 вычисляются таким образом, чтобы величина функционала

$$F(b_0, b_1) = \sum_{i=1}^n (y_i - \hat{f}_i)^2 = \sum_{i=1}^n (y_i - b_0 + b_1x_i)^2 \quad (2.3.3)$$

была минимальной. Значения \hat{f}_i определяются по формуле (2.3.2) при $x = x_i$, т. е.

$$\hat{f}_i = \hat{f}(x_i) = b_0 + b_1x_i. \quad (2.3.4)$$

Введем величину $e_i = y_i - \hat{f}_i$, характеризующую отклонение выборочного значения y_i от предсказанного \hat{f}_i . Эту величину назовем *невязкой* (или *остатком*) регрессии в i -ой точке. Тогда измеренные значения y_i можно записать выражением

$$y_i = b_0 + b_1x_i + e_i$$

а функционал (2.3.3) можно переписать в виде $F(b_0, b_1) = \sum_{i=1}^n e_i^2$.

Для функционала (2.3.3) необходимыми и достаточными условиями минимума являются условия равенства частных производных нулю, т. е. условия минимума функционала определяются системой из двух следующих уравнений:

$$\begin{cases} \frac{\partial F(b_0, b_1)}{\partial b_0} = 2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) \cdot (-1) = 0 \\ \frac{\partial F(b_0, b_1)}{\partial b_1} = 2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) \cdot (-x_i) = 0 \end{cases} \quad (2.3.5)$$

относительно двух неизвестных b_0, b_1 . Выполнив простые преобразования, получаем *систему нормальных уравнений* для вычисления коэффициентов b_0, b_1 линейной регрессии:

$$\begin{cases} b_0 \cdot n + b_1 \cdot \sum_{i=1}^n x_i = \sum_{i=1}^n y_i \\ b_0 \cdot \sum_{i=1}^n x_i + b_1 \cdot \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i \cdot y_i \end{cases} \quad (2.3.6)$$

Для упрощения записи и дальнейших вычислений введем следующие средние (по выборке) величины:

$$\begin{aligned} \bar{x} &= \frac{1}{n} \sum_{i=1}^n x_i; & \bar{y} &= \frac{1}{n} \sum_{i=1}^n y_i; \\ \overline{xy} &= \frac{1}{n} \sum_{i=1}^n x_i \cdot y_i; & \overline{x^2} &= \frac{1}{n} \sum_{i=1}^n x_i^2. \end{aligned}$$

Тогда систему (2.3.6) можно записать в виде

$$\begin{cases} b_0 + b_1 \cdot \bar{x} = \bar{y} \\ b_0 \cdot \bar{x} + b_1 \cdot \overline{x^2} = \overline{xy} \end{cases} \quad (2.3.7)$$

Решая эту систему уравнений, получаем

$$b_1 = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{\overline{x^2} - (\bar{x})^2} = \frac{m_{XY}}{s_X^2}; \quad (2.3.8)$$

$$b_0 = \bar{y} - b_1 \cdot \bar{x}, \quad (2.3.9)$$

где m_{XY} – выборочное значение корреляционного момента, определенное по формуле:

$$m_{XY} = \overline{xy} - \bar{x} \cdot \bar{y}, \quad (2.3.10)$$

s_X^2 – выборочное значение дисперсии величины X , определяемой по формуле:

$$s_X^2 = \overline{x^2} - (\bar{x})^2. \quad (2.3.11)$$

Коэффициент b_1 называют *коэффициентом регрессии Y по X* , и он показывает, на сколько единиц в среднем меняется переменная Y при изменении X на одну единицу.

Чтобы убедиться в этом, подставим (2.3.9) во второе уравнение системы (2.3.7). Получаем новое уравнение регрессии

$$\bar{y} - \bar{y} = b_1 (\bar{x} - \bar{x}), \quad (2.3.12)$$

которое подтверждает данное выше определение.

Коэффициент регрессии b_1 является размерной величиной, и он также как корреляционный момент m_{XY} характеризует «тесноту связи» между Y и X . Коэффициент b_1 связан с выборочным коэффициентом корреляции формулой

$$r_{XY} = b_1 \cdot \frac{s_X}{s_Y}, \quad (2.3.13)$$

где s_Y – выборочное значение среднеквадратического отклонения s_Y величины Y , определяемое выражением

$$s_Y = \sqrt{\overline{y^2} - (\bar{y})^2}. \quad (2.3.14)$$

Задание. Докажите справедливость формулы (2.3.13), используя формулу (2.3.8).

Непосредственно выборочный коэффициент корреляции r_{XY} (или проще коэффициент корреляции) можно вычислить по формуле

$$r_{XY} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{s_X \cdot s_Y}, \quad (2.3.15)$$

где s_X – определяется выражением

$$s_X = \sqrt{x^2 - (\bar{x})^2}. \quad (2.3.16)$$

Пример 2.3.1. По выборочным данным примера 3.1.1 вычислить коэффициенты b_0, b_1 линейного уравнения регрессии.

Решение. Вычислим эти коэффициенты, используя табличный процессор Excel (версия XP). На рис. 2.3 показан фрагмент документа Excel, в котором: а) размещены выборочные данные таблицы 1; б) запрограммировано вычисление коэффициентов системы (2.3.7); в) запрограммировано вычисление b_0, b_1 по формулам (2.3.9) (2.3.8) соответственно.

	A	B	C	D	E	F	G	H
1		Исходные данные			=B3^2		=B3*C3	
2		x_i	y_i	x_i^2	$x_i \cdot y_i$			
3		8	5	64	40		= (E13-B13*C13)/(D13-B13^2)	
4		11	10	121	110			
5		12	10	144	120	b_1	1,016	
6		9	7	81	63	b_0	-2,75	
7		8	5	64	40			
8		8	6	64	48		=C13-G5*B13	
9		9	6	81	54			
10		9	5	81	45			
11		8	6	64	48			
12		12	8	144	96			
13	Средние значения	9,4	6,8	90,8	66,4			
14								
15		=СРЗНАЧ(B3:B12)			=СРЗНАЧ(E3:E12)			

Рис. 2.3. Вычисление коэффициентов линейной регрессии

Заметим, что для вычисления средних значений используется функция Excel СРЗНАЧ(диапазон ячеек).

В результате выполнения запрограммированных вычислений получаем $b_0 = -2.75$; $b_1 = 1.016$, а само уравнение регрессии имеет вид

$$\hat{f}(x) = -2.75 + 1.016x, \quad (2.3.17)$$

или

$$\hat{f} - 6.8 = 1.016(x - \bar{x}).$$

Прямая линия, соответствующая этим уравнениям, показана на рис. 2.2. ●

Задание. Используя уравнение (2.3.17), определите производительность труда шахтера, если толщина угольного слоя равна: а) 8.5 метров (интерполяция данных); б) 14 метров (экстраполяция данных).

Пример 2.3.2. Используя формулу (2.3.15) и таблицу 2.1, вычислите выборочный коэффициент корреляции. Сделайте выводы о величине взаимосвязи между величинами X и Y .

Решение. Фрагмент документа Excel, вычисляющего величины коэффициента корреляции (формула (2.3.15)), s_X (формула (2.3.16)), s_Y (формула (2.3.14)), приведен на рис. 2.4.

Задание. Используя формулу (2.3.13) и вычисления примера 3.3.2, определите выборочный корреляционный момент m_{XY} .

Свойства оценок $b_0, b_1, \hat{y}(x)$. Напомним, что коэффициенты b_0, b_1 являются оценками для коэффициентов β_0, β_1 линейной регрессии $f(x) = M(Y | x) = \beta_0 + \beta_1 x$. Возникает вопрос: какими свойствами обладают оценки b_0, b_1 ?

При справедливости допущений P1, P2, P3 относительно случайных величин ε_i модели (2.1.4) коэффициенты b_0, b_1 как оценки для β_0, β_1 обладают следующими свойствами:

C1. Коэффициенты b_0, b_1 являются случайными величинами (так как зависят от случайной величины \bar{y});

C2. Коэффициенты b_0, b_1 являются несмещенными оценками т. е.

$$M(b_0) = \beta_0, \quad M(b_1) = \beta_1. \quad (2.3.18)$$

	A	B	C	D	E	F	G
1	Исходные данные						
2	x_i	y_i	x_i^2	y_i^2	$x_i \cdot y_i$		
3	8	5	64	25	40	=КОРЕНЬ(C13-A13^2)	
4	11	10	121	100	110		
5	12	10	144	100	120	s_X	1,562
6	9	7	81	49	63	s_Y	1,833
7	8	5	64	25	40		
8	8	6	64	36	48	r_{XY}	0,866
9	9	6	81	36	54		
10	9	5	81	25	45	=(E13-A13*B13)/(G5*G6)	
11	8	6	64	36	48		
12	12	8	144	64	96		
13	9,4	6,8	90,8	49,6	66,4	<i>Средние значения</i>	

Рис. 2.4. Вычисление выборочного коэффициента корреляции

С3. Уравнение регрессии $\hat{f}(x)$ является несмещенной оценкой для функции регрессии, т.е. $M(\hat{f}(x)) = \beta_0 + \beta_1 x = M(Y|x)$, что доказывает свойство несмещенности оценки $\hat{f}(x)$.

С4. Оценки b_0, b_1 имеют наименьшую дисперсию (т. е. минимально отклоняются от β_0, β_1) в классе всех линейных несмещенных оценок. Это свойство является особенно привлекательным – оно утверждает, что любые другие b_0, b_1 , линейно зависящие от \bar{y} (или от y_i) будут иметь больший разброс, а, следовательно, и меньшую точность.

С5. Величина

$$s^2 = \frac{\sum_{i=1}^n (y_i - \hat{f}_i)^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2}. \quad (2.3.19)$$

Является несмещенной оценкой для дисперсии σ^2 случайной составляющей ε .

Все эти «хорошие» свойства обуславливают широкое применение метода наименьших квадратов для оценивания параметров на протяжении трех последних столетий.

Дисперсии оценок b_0, b_1 . В заключение приведем формулы, определяющие дисперсию оценок b_0, b_1 .

$$D(b_0) = \sigma^2 \frac{\overline{x^2}}{\sum_{i=1}^n (x_i - \bar{x})^2}; \quad D(b_1) = \sigma^2 \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Из этих соотношений можно сделать следующие выводы:

- дисперсии оценок b_0, b_1 прямо пропорциональны дисперсии σ^2 ;
- чем больше дисперсия (разброс значений) объясняющей переменной (т. е. чем шире область ее изменения), тем больше

величина $\sum_{i=1}^n (x_i - \bar{x})^2$ и тем меньше дисперсия оценок;

- при увеличении объема выборки n увеличивается величина $\sum_{i=1}^n (x_i - \bar{x})^2$, а, следовательно, уменьшается дисперсия оценок.

На практике дисперсия σ^2 , как правило, неизвестна. Поэтому вместо σ^2 используют ее оценку s^2 (см. (2.3.19)), и тогда приходим к оценкам дисперсии $D(b_0), D(b_1)$:

$$s_{b_0}^2 = s^2 \cdot \frac{\overline{x^2}}{\sum_{i=1}^n (x_i - \bar{x})^2}; \quad (2.3.20)$$

$$s_{b_1}^2 = s^2 \cdot \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}. \quad (2.3.21)$$

Величины s_{b_0}, s_{b_1} , являющиеся квадратными корнями из (2.3.20), (2.3.21) называют стандартными ошибками коэффициентов регрессии.

Пример 2.3.3. Вычислить оценки $s_{b_0}^2, s_{b_1}^2$ для дисперсий коэффициентов b_0, b_1 , определенных в примере 2.3.1.

Решение. Вычисления проведем, используя табличный процессор. На рис. 2.5 показан фрагмент документа Excel, в котором выполнены вычисления по формулам (2.3.20), (2.3.21). Получаем следующие значения: $s^2 = 1.049, s_{b_0}^2 = 3.904, s_{b_1}^2 = 0.043$. По приведенному фрагменту сделаем следующие замечания:

- значения коэффициентов b_0, b_1 взяты из примера 2.3.1 и ячейки (B1,B2), в которых они находятся, имеют абсолютную адресацию (\$B\$1, \$B\$2) в выражениях, вычисляющих значения регрессии \hat{y}_i ;
- значение \bar{x}^2 (ячейка B19) взято из примера 2.3.1. ●

Функции Excel для вычисления коэффициентов парной линейной регрессии. Приведем некоторые статистические функции Excel, полезные при построении парной линейной регрессии.

Функция ОТРЕЗОК. Вычисляет коэффициент b_0 и обращение имеет вид

ОТРЕЗОК(диапазон_значений_y; диапазон_значений_x).

Функция НАКЛОН. Вычисляет коэффициент b_1 и обращение имеет вид

НАКЛОН(диапазон_значений_y; диапазон_значений_x).

Функция ПРЕДСКАЗ. Вычисляет значение линейной парной регрессии при заданном значении независимой переменной (обозначена через z) и обращение имеет вид

ПРЕДСКАЗ(z ; диапазон_значений_y; диапазон_значений_x).

	A	B	C	D	E	F
1	b_0	-2,75				
2	b_1	1,016				
3	Исходные данные					
4	x_i	y_i	\hat{y}_i	$e_i = \hat{y}_i - y_i$	e_i^2	$(x_i - \bar{x})^2$
5	8	5	5,378	0,378	0,143	1,96
6	11	10	8,426	-1,574	2,477	2,56
7	12	10	9,442	-0,558	0,311	6,76
8	9	7	6,394	-0,606	0,367	0,16
9	8	5	5,378	0,378	0,143	1,96
10	8	6	5,378	-0,622	0,387	1,96
11	9	6	6,394	0,394	0,155	0,16
12	9	5	6,394	1,394	1,943	0,16
13	8	6	5,378	-0,622	0,387	1,96
14	12	8	9,442	1,442	2,079	6,76
15	9,4	6,8			8,393	24,40
16					=СУММ(E5:E14)	
17					=СУММ(F5:F14)	
18	\bar{x}	9,400			=B20*B19/F15	
19	\bar{x}^2	90,800		$s_{b_0}^2$	3,904	
20	s^2	1,049		$s_{b_1}^2$	0,043	
21						
22		=E15/(10-2)				=B20/F15

Рис. 2.5. Вычисление дисперсий оценок b_0, b_1

Функция СТОШУХ. Вычисляет оценку s для среднеквадратического отклонения σ возмущений ε_i и обращение имеет вид (YX – латинские буквы):

СТОШУХ(диапазон_значений_y; диапазон_значений_x).

Пример 2.3.4. По данным таблицы 2.1 вычислить, используя функции Excel величины b_0, b_1, s и найти значения линейной регрессии при $x = x_i$.

Решение. Фрагмент документа Excel, вычисляющего требуемые величины приведен на рис. 2.6. Обратите внимание на использовании абсолютной адресации при вычислении $\hat{\varepsilon}_i$ ●

	A	B	C	D	E	F
1	Исходные данные		=ПРЕДСКАЗ(A3;\$B\$3:\$B\$12;\$A\$3:\$A\$12)			
2	x_i	y_i	\hat{y}_i			
3	8	5	5,377			
4	11	10	8,426			
5	12	10	9,443			
6	9	7	6,393			
7	8	5	5,377			
8	8	6	5,377			
9	9	6	6,393			
10	9	5	6,393			
11	8	6	5,377			
12	12	8	9,443			
13						
14						
15		b_0	-2,754	=ОТРЕЗОК(B3:B12;A3:A12)		
16		b_1	1,016	=НАКЛОН(B3:B12;A3:A12)		
17		s	1,024	=СТОШУХ(B3:B12;A3:A12)		

Рис. 2.6. Использование функций Excel

Замечание 2.3.1. Коэффициент b_1 является размерной величиной, и поэтому вычисляют **коэффициент эластичности** по формуле

$$E = b_1 \cdot \frac{\bar{x}}{\bar{y}}, \quad (2.3.22)$$

который показывает, на сколько процентов (от средней) изменится в среднем величина Y при увеличении переменной X на 1% от своего среднего значения. Для нелинейной парной регрессии коэффициент эластичности определяется выражением:

$$E = f'(x) \cdot \frac{\bar{x}}{\bar{y}} \quad (2.3.23)$$

и зависит от значения переменной x , при котором вычисляется производная $f'(x)$. ♥

Пример 2.3.4. Вычислить коэффициент эластичности для уравнения регрессии (2.3.17).

Решение. Из примера 2.3.2 берем значение $b_1 = 1.016$, а из рис. 3.2 значения $\bar{x} = 9.4$, $\bar{y} = 6.8$. Подставляя эти значения в формулу (2.3.22) получаем $E = 1.016 \cdot 9.4 / 6.8 = 1.40$. ●

2.4. Интервальные оценки функции регрессии и ее параметров

В предыдущем параграфе было говорилось, что при малом объеме выборки дисперсия оценок b_0, b_1 будет большой, т. е. b_0, b_1 могут существенно отклоняться от β_0, β_1 . В этом случае переходят к построению интервальных оценок. Напомним, что *интервальной оценкой* параметра θ называют числовой интервал $(\hat{\theta}_n^{(H)}, \hat{\theta}_n^{(G)})$, в который с заданной вероятностью γ попадает неизвестное значение параметра θ , т. е.

$$P(\hat{\theta}_n^{(H)} < \theta < \hat{\theta}_n^{(G)}) = \gamma.$$

Интервал $(\hat{\theta}_n^{(H)}, \hat{\theta}_n^{(e)})$ называют *доверительным*, а вероятность γ - *доверительной вероятностью* или *надежностью* интервальной оценки.

Необходимым условием для построения интервальных оценок является задание закона распределения возмущения ε . Поэтому введем следующее дополнительное предположение:

P4. Возмущения ε_i подчинялись нормальному распределению $\varepsilon_i \sim N(0, \sigma^2)$.

Интервальные оценки для коэффициентов β_0, β_1 . Если $\varepsilon_i \sim N(0, \sigma^2)$, то оценки b_0, b_1 также будут распределены по нормальному закону, как линейные комбинации нормально распределенных величин y_i , т. е.

$$b_0 \sim N(\beta_0, D(b_0)); \quad b_1 \sim N(\beta_1, D(b_1)). \quad (2.4.1)$$

Отсюда следует, что статистики:

$$T_{b_0} = \frac{b_0 - \beta_0}{s_{b_0}}; \quad T_{b_1} = \frac{b_1 - \beta_1}{s_{b_1}}$$

имеют распределение Стьюдента с $k = n - 2$ степенями свободы. Тогда с вероятностью γ будут выполняться следующие неравенства:

$$b_0 - t(\gamma, n-2) \cdot s_{b_0} \leq \beta_0 \leq b_0 + t(\gamma, n-2) \cdot s_{b_0};$$

$$b_1 - t(\gamma, n-2) \cdot s_{b_1} \leq \beta_1 \leq b_1 + t(\gamma, n-2) \cdot s_{b_1},$$

где $t(\gamma, n-2)$ вычисляется с помощью функции Excel:

$$t(\gamma, n-2) = \text{СТЮДРАСПОБР}(1-\gamma; n-2). \quad (2.4.2)$$

Величины $s_{b_0} = \sqrt{s_{b_0}^2}$, $s_{b_1} = \sqrt{s_{b_1}^2}$, $s_{b_0}^2$, $s_{b_1}^2$ вычисляются по формулам (2.3.20), (2.3.21). Следовательно, интервалы:

$$\left[b_0 - t(\gamma, n-2) \cdot s_{b_0}, b_0 + t(\gamma, n-2) \cdot s_{b_0} \right]; \quad (2.4.3)$$

$$\left[b_1 - t(\gamma, n-2) \cdot s_{b_1}, b_1 + t(\gamma, n-2) \cdot s_{b_1} \right] \quad (2.4.4)$$

являются интервальными оценками для коэффициентов β_0, β_1 с надежностью (доверительной вероятностью), равной γ .

Интервальная оценка для дисперсии σ^2 . Величина s^2 (см. (2.3.17)) использовалась нами как оценка для дисперсии σ^2 . Введем статистику ns^2/σ^2 , которая имеет χ^2 -распределение с $k = n - 2$ степенями свободы. Поэтому интервальная оценка для σ^2 с доверительной вероятностью $\gamma = 1 - \alpha$ имеет вид

$$\left[\frac{ns^2}{\chi_{1-\alpha/2, n-2}^2}, \frac{ns^2}{\chi_{\alpha/2, n-2}^2} \right]. \quad (2.4.5)$$

где $\chi_{\alpha/2, n-2}^2, \chi_{1-\alpha/2, n-2}^2$ - квантили χ^2 -распределения с $k = n - 2$ степенями свободы уровней $\alpha/2, 1 - \alpha/2$ соответственно.

Квантили определяются следующими выражениями:

$$\chi_{\alpha/2, n-2}^2 = \text{ХИ2ОБР}(1-\alpha/2; n-2), \quad (2.4.6)$$

$$\chi_{1-\alpha/2, n-2}^2 = \text{ХИ2ОБР}(\alpha/2; n-2). \quad (2.4.7)$$

Напомним, что квантилем уровня q для случайной величины X с плотностью распределения $p(x)$ называется величина x_q , определяемая уравнением

$$P(X < x_q) = \int_{-\infty}^{x_q} p(x) dx = q,$$

где $P(X < x_q)$ - вероятность случайного события $X < x_q$.

Пример 2.4.1. Построить интервальные оценки для коэффициентов регрессии β_0, β_1 и дисперсии σ^2 с надежностью $\gamma = 0.95$.

Решение. По формуле (2.4.2) определяем $t(0.95, 8) = 2.31$. Тогда $t(0.95, 8) \cdot s_{b_0} = 4.56$; $t(0.95, 8) \cdot s_{b_1} = 0.48$, а сами интервальные оценки для β_0, β_1 определяются интервалами:

$$[-2.75 - 4.56, -2.75 + 4.56] = [-7.31, 1.81];$$

$$[1.016 - 0.48, 1.016 + 0.48] = [0.537, 1.496].$$

Далее по формулам (2.4.6), (2.4.7) находим значения квантилей: $\chi_{0.025,8}^2 = 2.18$; $\chi_{0.975,8}^2 = 17.53$ и получаем интервальную оценку для σ^2 :

$$\left[\frac{10 \cdot 1.049}{17.53}, \frac{10 \cdot 1.049}{2.18} \right] = [0.589, 4.81] \quad \ominus$$

Интервальная оценка для функции регрессии. Построим интервал, в который с вероятностью γ попадает функция регрессии $f(x) = M(Y|x)$. Для этого уравнение регрессии (2.3.12) перепишем в виде (подчеркивая зависимость от x):

$$\underline{f}(x) = \bar{y} + b_1(x - \bar{x}). \quad (2.4.8)$$

Если справедливо предположение **P4** ($\varepsilon_i \sim N(0, \sigma^2)$), то $\underline{f}(x)$ также подчинится нормальному распределению с математическим ожиданием $M(\underline{f}(x))$ и дисперсией $D(\underline{f}(x))$, которые зависят от x . Как было показано ранее, из несмещенности оценок b_0, b_1 следует $M(\underline{f}(x)) = M(Y|x)$, т. е. $\underline{f}(x)$ также является несмещенной оценкой для функции регрессии. Далее можно показать, что

$$D(\underline{f}(x)) = \sigma_{\underline{f}}^2(x) = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right].$$

Заменяя неизвестную дисперсию σ^2 на ее оценку s^2 , получаем оценку для $\sigma_{\underline{f}}^2$, равную

$$s_{\underline{f}}^2(x) = s^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]. \quad (2.4.9)$$

Статистика

$$T_{\underline{f}}(x) = \frac{\underline{f}(x) - M(Y|x)}{s_{\underline{f}}(x)}$$

для каждого фиксированного x имеет распределение Стьюдента с $k = n - 2$ степенями свободы. Поэтому с вероятностью γ будет выполняться неравенство

$$\underline{f}(x) - t(\gamma, n - 2) \cdot s_{\underline{f}}(x) \leq M(Y|x) \leq \underline{f}(x) + t(\gamma, n - 2) \cdot s_{\underline{f}}(x).$$

Следовательно, интервал

$$\left[\underline{f}(x) - t(\gamma, n - 2) \cdot s_{\underline{f}}(x), \underline{f}(x) + t(\gamma, n - 2) \cdot s_{\underline{f}}(x) \right] \quad (2.4.10)$$

будет являться интервальной оценкой для $M(Y|x)$ с надежностью, равной γ .

Так как $s_{\underline{f}}(x)$ зависит от x , то и «ширина» интервала (2.4.7) также зависит от x . Минимальная ширина достигается при $x = \bar{x}$.

Задание. Докажите справедливость этого утверждения.

Пример 2.4.2. Построить интервальную оценку для функции регрессии $M(Y|x)$ с надежностью $\gamma = 0.95$, используя для этого уравнение регрессии $\underline{f}(x)$, построенное в примере 2.3.1.

Решение. Значения граничных точек y_i^H (нижняя), y_i^B (верхняя) интервальной оценки будем вычислять для $x = x_i$, $i = 1, \dots, 10$, приведенных в таблице 2.1 по формуле (2.4.10). Фрагмент документа Excel, осуществляющего вычисление граничных точек и значений $\underline{f}(x_i)$ показан на рис. 2.7. Величины $\sum_{i=1}^{10} (x_i - \bar{x})^2$, s^2 ,

\bar{x} и коэффициенты b_0, b_1 взяты из предыдущих примеров. \ominus

Интервальная оценка для индивидуальных значений зависимой переменной. Построенная интервальная оценка (2.4.10) определяет возможное положение математического ожидания $M(Y|x)$, но не отдельных возможных значений зависимой переменной Y , которые отклоняется от $M(Y|x)$. Такие значения будем называть *индивидуальными значениями* зависимой переменной.

При построении интервальной оценки для индивидуальных значений (обозначим эти значения y^*) зависимой переменной необходимо учитывать еще один источник отклонений – рассеяние вокруг линии регрессии $M(Y|x)$. Дисперсия таких отклонений равна σ^2 . Следовательно, оценку дисперсии $s_{\hat{y}}^2(x)$ необходимо увеличить на s^2 (оценка для σ^2). В результате оценка дисперсии значений y^* равна

	A	B	C	D	E	F	G
1	b_0	-2,75	=B\$1+B\$2*A5				
2	b_1	1,016		=B\$16*(1/10+(A5-B\$17)^2/B\$18)			
3	Исходные данные				=C5-2,31*КОРЕНЬ(D5)		
4	x_i	y_i	\hat{y}_i	s_y^2	y_i^H	y_i^B	
5	8	5	5,378	0,189	4,373	6,383	
6	11	10	8,426	0,215	7,355	9,497	
7	12	10	9,442	0,396	7,989	10,895	
8	9	7	6,394	0,112	5,622	7,166	
9	8	5	5,378	0,189	4,373	6,383	
10	8	6	5,378	0,189	4,373	6,383	
11	9	6	6,394	0,112	5,622	7,166	
12	9	5	6,394	0,112	5,622	7,166	
13	8	6	5,378	0,189	4,373	6,383	
14	12	8	9,442	0,396	7,989	10,895	
15							
16	s^2	1,049					
17	\bar{x}	9,400					
18	$\sum_{i=1}^{10} (x_i - \bar{x})^2$	24,400					

Рис. 2.7. Вычисление интервальной оценки для $M(Y|x)$

$$s_{y^*}^2(x) = s^2 \left[1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right], \quad (2.4.11)$$

а соответствующая интервальная оценка определяется интервалом

$$\left[\hat{f}(x) - t(\gamma, n-2) \cdot s_{y^*}(x), \hat{f}(x) + t(\gamma, n-2) \cdot s_{y^*}(x) \right] \quad (2.4.12)$$

Для построения интервальной оценки для y^* можно использовать фрагмент документа Excel, приведенный на рис. 2.7 с одним изменением в столбце D – выражение, стоящее в скобках надо увеличить на 1 (см. (2.4.11)).

2.5. Значимость уравнения регрессии и коэффициент детерминации

Проверить значимость уравнения регрессии – значит установить, соответствует ли построенное уравнение регрессии экспериментальным данным и достаточно ли включенных в уравнение объясняющих переменных для описания зависимой переменной. Проверка значимости может проводиться по следующим направлениям:

- проверка значимости коэффициентов уравнения регрессии;
- проверка значимости уравнения регрессии;

Проверка статистической значимости коэффициентов регрессии. Напомним, что коэффициенты b_0, b_1 являются случайными величинами, значения которых отклоняются от их математических ожиданий: $M(b_0) = \beta_0, M(b_1) = \beta_1$. Поэтому часто возникают вопросы, подобные данному: при вычисленном значении $b_0 = 0.125$ может ли $\beta_0 = 0$? Коэффициент $b_j, j = 0, 1$ уравнения регрессии является значимым, если соответствующий ему коэффициент β_j отличен от нуля.

Для ответа на вопрос о значимости коэффициентов регрессии используем методы проверки статистических гипотез.

Напомним, что *статистической гипотезой* называется любое предположение о виде или параметре неизвестного закона распределения. Проверяемую гипотезу обычно принимают *нулевой* и обозначают H_0 . Наряду с нулевой гипотезой рассматривают *альтернативную* гипотезу H_1 , являющуюся логическим отрицанием H_0 . Нулевая и альтернативная гипотезы представляют собой две возможности выбора, осуществляемого на основе *проверки статистических гипотез*. Для этого используется некоторая величина K , называемая *статистическим критерием*. Значение критерия зависит от выборочных данных x_1, x_2, \dots, x_n и, будучи случайной величиной, критерий K подчиняется при выполнении гипотезы H_0 некоторому известному закону распределения. В области возможных значений критерия K выделяют подобласть, называемую *критической*. Если вычисленное значение критерия попадает в критическую область, то гипотеза H_0 отвергается и принимается альтернативная H_1 .

Поскольку принятие той или иной гипотезы носит вероятностный характер, то возможны следующие ситуации:

S1. Гипотеза H_0 *верна*, и при проверке она *не отвергается*;

S2. Гипотеза H_0 *верна*, но при проверке она *отвергается*;

S3. Гипотеза H_0 *не верна*, и при проверке она *отвергается* (в пользу альтернативной H_1);

S4. Гипотеза H_0 *не верна*, но при проверке она *принимается*.

Очевидно, что ситуации S1, S3 являются «правильными» ситуациями, S2, S4 – «ошибочными». Ситуация S2 называется *ошибкой I рода*, и вероятность ее появления называется *уровнем значимости* (обозначается α). Обычно $\alpha = 0.025 \div 0.05$. Ситуация S4 называется *ошибкой II рода*, и вероятность ее появления обозначают β .

Для проверки значимости коэффициента b_0 сформулируем следующие *статистические гипотезы*:

$H_0: \beta_0 = 0$ (коэффициент b_0 не значим);

$H_1: \beta_0 \neq 0$ (коэффициент b_0 значим)

и примем уровень значимости (вероятность ошибки первого рода) равным α (обычно $\alpha = 0.05$). В качестве критерия для проверки гипотезы H_0 примем случайную величину

$$T_{b_0} = \frac{b_0}{s_{b_0}}, \quad (2.5.1)$$

которая при справедливости гипотезы H_0 имеет распределение Стьюдента с $k = n - 2$ степенями свободы (s_{b_0} – стандартная ошибка коэффициента b_0 (см. 2.3.20)). Гипотеза H_0 отвергается с уровнем значимости α , если

$$|T_{b_0}| > t(1 - \alpha, n - 2) \quad (2.5.2)$$

где $t(1 - \alpha, n - 2)$ – величина, определяемая выражением (2.4.2). Таким образом, если выполняется неравенство (2.5.2), то говорят, что коэффициент b_0 является *значимым с уровнем значимости α* .

Для проверки значимости коэффициента b_1 сформулируем следующие статистические гипотезы:

$H_0: \beta_1 = 0$ (коэффициент b_1 не значим);

$H_1: \beta_1 \neq 0$ (коэффициент b_1 значим)

и примем уровень значимости α . В качестве критерия для проверки гипотезы H_0 примем случайную величину

$$T_{b_1} = \frac{b_1}{s_{b_1}}, \quad (2.5.3)$$

которая при справедливости гипотезы H_0 имеет распределение Стьюдента с $k = n - 2$ степенями свободы (s_{b_1} – стандартная ошибка коэффициента b_1 (см. 2.3.21)). Гипотеза H_0 отвергается с уровнем значимости α , если

$$|T_{b_1}| > t(1 - \alpha, n - 2). \quad (2.5.4)$$

Таким образом, если выполняется неравенство (2.5.4), то коэффициент b_1 является *значимым с уровнем значимости α* .

Проверка статистической значимости выборочного коэффициента корреляции. Напомним, что выборочный коэффициент корреляции r_{XY} , определяемый формулой (2.3.15), является случайной величиной, значение которой может отклоняться от «теоретического» коэффициента корреляции ρ_{XY} , определяемого выражением (2.1.8).

Для проверки значимости коэффициента r_{XY} сформулируем две гипотезы:

H_0 : $\rho_{XY} = 0$ (коэффициент r_{XY} не значим);

H_1 : $\rho_{XY} \neq 0$ (коэффициент r_{XY} значим)

и примем уровень значимости, равный α . В качестве критерия для проверки H_0 примем случайную величину

$$T_r = \frac{|r_{XY}| \sqrt{n-2}}{\sqrt{1-r_{XY}^2}}, \quad (2.5.5)$$

которая при справедливости гипотезы H_0 имеет распределение Стьюдента с $k = n - 2$ степенями свободы. Следовательно, если выполняется неравенство

$$|T_r| > t(1-\alpha, n-2), \quad (2.5.6)$$

то гипотеза H_0 отвергается с уровнем значимости α .

Пример 2.5.1. Проверить значимость коэффициентов b_0 , b_1 , вычисленных в примере 3.3.1.

Решение. Для проверки значимости коэффициента b_0 вычислим значение критерия (стандартную ошибку s_{b_0} возьмем из

примера 3.3.3): $T_{b_0} = \frac{b_0}{s_{b_0}} = \frac{-2.41}{1.98} = -1.217$. Неравенства

(2.5.2) не выполняется ($|-1.27| < 2.31$) и, следовательно, принимается гипотеза H_0 , т.е. коэффициент b_0 незначим с уровнем значимости $\alpha = 0.05$.

Аналогично проверим значимость коэффициента b_1 . Значение критерия T_{b_1} равно (стандартную ошибку s_{b_1} берем из при-

мера 3.3.3): $T_{b_1} = \frac{b_1}{s_{b_1}} = \frac{1.016}{0.21} = 4.84$. Неравенство (2.5.4) выполняется ($|4.90| > 2.31$) и поэтому делается вывод, что коэффициент b_1 значим с уровнем значимости $\alpha = 0.05$. ●

Пример 2.5.2. Проверить значимость выборочного коэффициента корреляции r_{XY} , вычисленного в примере 2.3.2 (уровень значимости $\alpha = 0.05$).

Решение. Для этого вычисляем значение критерия по формуле (2.5.5):

$$T_r = \frac{0.866 \cdot \sqrt{10-2}}{\sqrt{1-0.866^2}} = 4.90.$$

Неравенство (2.5.6) выполняется (так как $|4.90| > 2.31$), и поэтому нулевая гипотеза отвергается, а принимается альтернативная гипотеза о значимости r_{XY} с уровнем значимости $\alpha = 0.05$. ●

Задание. Проверьте значимость коэффициента корреляции r_{XY} с уровнем значимости $\alpha = 0.025$.

Проверка статистической значимости уравнения регрессии. Отклонение значений \hat{y}_i , вычисленных по уравнению регрессии (2.3.2) при $x = x_i$, $i = 1, \dots, n$ от «заданных» значений y_i может быть вызвано двумя основными причинами:

- наличием случайного слагаемого ε в регрессионной модели;
- принятая функция $f(x)$ не адекватна объясненной части эконометрической модели (неправильно выбран вид функции $f(x)$, например, взята линейная функция вместо параболической, или не учтены другие объясняющие переменные).

Если первая причина приводит к ухудшению точности прогнозирования исследуемого процесса по построенному уравнению регрессии, то вторая причина вносит систематическую ошибку (т. е. $M(\varepsilon) \neq 0$) и делает построенное уравнение регрессии неприемлемым для описания исследуемого экономического процесса.

Как же убедиться в том, что построенное уравнение регрессии «правильно» отражает связь между величинами Y и X ? Другими словами, соответствует ли уравнение регрессии исходным экспериментальным данным, т. е. является уравнение регрессии значимым?

Для проверки значимости рассмотрим две суммы:

- *объясненная (или факторная) сумма квадратов*

$$Q_r = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad (2.5.7)$$

интерпретируемая как мера разброса, «объяснимого» с помощью построенного уравнения регрессии;

- *остаточная сумма квадратов*

$$Q_e = \sum_{i=1}^n (\hat{y}_i - y_i)^2, \quad (2.5.8)$$

являющаяся мерой разброса (разброса точек относительно линии регрессии), не «объясненного» построенным уравнением регрессии.

Замечание 2.5.1. Для уравнения регрессии *со свободным членом* (т.е. присутствует коэффициент b_0), построенного на основе МНК, справедливо следующее равенство

$$Q = Q_r + Q_e,$$

где $Q = \sum_{i=1}^n (y_i - \bar{y})^2$ - *полная сумма квадратов*, интерпретируемая как мера общего рассеивания переменной Y относительно среднего значения \bar{y} . ♥

Уравнение парной регрессии значимо с уровнем значимости α , если выполняется следующее неравенство:

$$F = \frac{Q_r \cdot (n-2)}{Q_e} > F_{1-\alpha; 1; n-2}, \quad (2.5.9)$$

где $F_{\gamma; 1; n-2}$ – значения квантиля уровня γ F -распределения с числами степеней свободы $k_1 = 1$ и $k_2 = n - 2$. Для вычисления квантиля можно использовать следующее выражение

$$F_{1-\alpha; 1; n-2} = \text{FPАСПОБР}(\alpha; 1; n-2). \quad (2.5.10)$$

Критерий F часто называют *критерием Фишера* или *F-критерием*.

Пример 2.5.3. По данным таблицы 2.1 оценить на уровне $\alpha = 0.05$ значимость уравнения регрессии, построенного в примере 2.3.1.

Решение. На рис. 2.8 приведен фрагмент документа Excel, вычисляющего значения Q_e , $Q_r = Q - Q_e$ и критерий F . Получены следующие значения $Q_r = 25.207$, $Q_e = 8.393$, $F = 24.025$. По формуле (2.5.10) вычисляем квантиль $F_{0.95; 1; 8} = 5.32$. Неравенство (2.5.9) выполняется, т. е. $24.04 > 5.32$ и поэтому уравнение регрессии (2.3.16) значимо с уровнем значимости $\alpha = 0.05$. ☉

Одной из *наиболее эффективных оценок адекватности уравнения регрессии* (мерой качества «подгонки» регрессионной модели к «наблюдаемым» значениям y_i) является *коэффициент детерминации* R^2 , определяемый по формуле:

$$R^2 = \frac{Q_r}{Q} = 1 - \frac{Q_e}{Q}. \quad (2.5.11)$$

Величина R^2 показывает, *какая часть (доля) вариации зависимой переменной обусловлена вариацией объясняющей переменной* и изменяется в диапазоне

$$0 \leq R^2 \leq 1 \quad (2.5.12)$$

Чем ближе R^2 к 1, тем лучше регрессия аппроксимирует эмпирические данные. Если $R^2 = 1$, то эмпирические точки (x_i, y_i) лежат на линии регрессии ($Q_e = 0$), и между X и Y существует линейная функциональная зависимость. Если $R^2 = 0$ ($Q_e = Q$), то вариации Y полностью обусловлены воздействием неучтенных в уравнении регрессии переменных, и линия регрессии параллельна оси абсцисс.

	A	B	C	D	E	F
1	b_0	-2,75				
2	b_1	1,016	$=(B5-\$B\$15)^2$		$=(D5-B5)^2$	
3	Исходные данные					
4	x_i	y_i	$(y_i - \bar{y})^2$	\hat{y}_i	$(\hat{y}_i - y_i)^2$	
5	8	5	3,240	5,378	0,143	
6	11	10	10,240	8,426	2,477	
7	12	10	10,240	9,442	0,311	
8	9	7	0,040	6,394	0,367	
9	8	5	3,240	5,378	0,143	
10	8	6	0,640	5,378	0,387	
11	9	6	0,640	6,394	0,155	
12	9	5	3,240	6,394	1,943	
13	8	6	0,640	5,378	0,387	
14	12	8	1,440	9,442	2,079	
15		6,800	33,600		8,393	
16		\bar{y}	Q		Q_e	
17						
18				$=\text{СУММ}(E5:E14)$		
19	$Q_r = Q - Q_e$		25,207			
20	F		24,025			
21				$=C19*(10-2)/E15$		

Рис. 2.8. Вычисление величины F – критерия

Внимание! Коэффициент R^2 имеет смысл рассматривать, если в уравнении регрессии присутствует свободный член (в случае парной линейной регрессии – коэффициент b_0). Только в этом случае справедливо равенство (2.5.7), а, следовательно, и (2.5.10).

В случае парной линейной регрессии имеет место важное тождество

$$R^2 = r_{XY}^2. \quad (2.5.13)$$

Пример 2.5.4. По данным таблицы 2.1 определить коэффициент детерминации для уравнения регрессии, построенного в примере 3.3.1.

Решение. Из примера 3.5.3 возьмем следующие значения: $Q = 33.600$, $Q_e = 8.393$. Получаем $R^2 = 1 - \frac{Q_e}{Q} = 0.750$. Такая величина коэффициента детерминации означает, что вариация зависимой переменной Y – добыча угля на одного рабочего – на 75% объясняется изменением величины X – толщиной угольного пласта. Остальные 25% могут быть объяснены влиянием случайных факторов (т.е. возмущением ϵ). ☹

Проверка значимости уравнения регрессии с использованием коэффициента детерминации. Если известен коэффициент детерминации R^2 , то уравнение парной линейной регрессии значимо с уровнем значимости α , если выполняется условие

$$F_R > F_{1-\alpha;1;n-2}, \quad (2.5.14)$$

где

$$F_R = \frac{R^2 \cdot (n-2)}{(1-R^2)}. \quad (2.5.15)$$

Напомним, что для вычисления квантиля $F_{1-\alpha;1;n-2}$ можно использовать следующее выражение

$$F_{1-\alpha;1;n-2} = \text{FPACПОБР}(\alpha;1;n-2).$$

2.6. Нелинейная парная регрессия

Нелинейность регрессии может быть обусловлена двумя причинами:

- нелинейность по объясняющей переменной;
- нелинейность по коэффициентам регрессии.

Кратко рассмотрим несколько подходов к вычислению коэффициентов парной регрессии в этих случаях.

Нелинейность по объясняющей переменной. Примером такой нелинейности может служить уравнение регрессии вида (гиперболическая регрессия):

$$f(x) = b_0 + b_1 \sqrt{x}$$

В этом случае, вводя новую переменную $Z = X^{1/2}$, приходим к линейной регрессии

$$f(z) = b_0 + b_1 z,$$

коэффициенты b_0, b_1 которой вычисляются на основе метода МНК (см. параграф 2.3). Вычислив коэффициенты, возвращаемся к исходному нелинейному уравнению регрессии.

Пример 2.6.1. Рассмотрим класс регрессионных моделей вида

$$Y = \beta_0 + \beta_1 \ln x + \varepsilon, \quad (2.6.1)$$

которые описывают связь между долей расходов на товары длительного пользования (переменная Y - единицы измерения процентов общей суммы расходов) и доходом американской семьи (переменная X - единица измерения тысяч долларов). Уравнение регрессии для модели (2.6.1) имеет вид

$$\hat{y}(x) = b_0 + b_1 \ln x \quad (2.6.2)$$

Необходимо определить коэффициент этого уравнения по данным, представленным в таблице 2.2.

Таблица 2.2

x_i	1	2	3	4	5	6
y_i	10	13.4	15.4	16.5	18.6	19.1

Для этого введем новую переменную $x' = \ln x$ и приходим к следующей системе уравнений (сравните с (2.3.7)):

$$\begin{cases} b_0 + b_1 \cdot \bar{x}' = \bar{y} \\ b_0 \cdot \bar{x}' + b_1 \cdot (\bar{x}')^2 = \overline{x'y} \end{cases}$$

где $\bar{x}' = \frac{1}{n} \sum_{i=1}^n x'_i = \frac{1}{n} \sum_{i=1}^n \ln x_i$; $(\bar{x}')^2 = \frac{1}{n} \sum_{i=1}^n (x'_i)^2 = \frac{1}{n} \sum_{i=1}^n (\ln x_i)^2$,

$$\overline{x'y} = \frac{1}{n} \sum_{i=1}^n x'_i y_i = \frac{1}{n} \sum_{i=1}^n \ln(x_i) \cdot y_i.$$

Выполнив необходимые вычисления, получаем следующую систему уравнений:

$$\begin{cases} b_0 + \frac{6.5792}{6} b_1 = \frac{93}{6} \\ \frac{6.5792}{6} \cdot b_0 + \frac{9.4099}{6} \cdot b_1 = \frac{113.238}{6} \end{cases}$$

Решая эту систему, находим $b_0 = 9.876$, $b_1 = 5.129$, а само уравнение (2.6.2) принимает

$$\hat{f} = 9.876 + 5.129 \ln x.$$

Значения \hat{f}_i , вычисленные для $x = x_i$, приведены на рис.2.9 (треугольные маркеры).

Видно хорошее согласие построенной регрессии с исходными данными (квадратные маркеры).

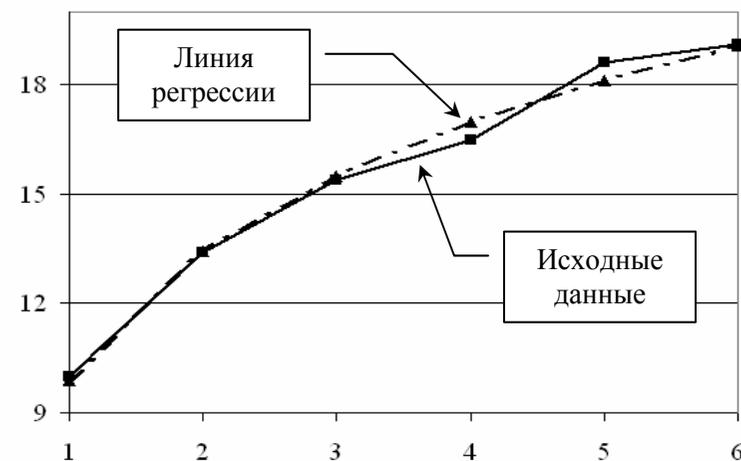


Рис.2.9. Построение нелинейной регрессии

Нелинейность по коэффициентам уравнения регрессии. К такому классу нелинейных регрессий относятся уравнения, в ко-

торых зависимая переменная нелинейным образом зависит от коэффициентов регрессии. Примеры таких нелинейных регрессионных моделей могут служить функции

- степенная $Y = \beta_0 X^{\beta_1} \cdot \varepsilon$; (2.6.3)

- показательная $Y = \beta_0 \beta_1^X \cdot \varepsilon$; (2.6.4)

- экспоненциальная $Y = \beta_0 e^{\beta_1 X} \cdot \varepsilon$. (2.6.5)

Для вычисления коэффициентов нелинейных регрессий возможны два подхода.

Первый подход заключается в применении некоторого (как правило, нелинейного) преобразования, которое приводит к линейной регрессии, но уже относительно новых коэффициентов и (или) новых переменных. Для иллюстрации этого подхода рассмотрим степенную регрессию (2.6.3), широко используемую в эконометрических исследованиях при изучении зависимости спроса от цены. После логарифмирования функции (2.6.3) получаем $\ln Y = \ln \beta_0 + \beta_1 \ln X + \ln \varepsilon$. Введем новые величины

$$Y' = \ln Y, \quad b'_0 = \ln \beta_0, \quad X' = \ln X, \quad \varepsilon' = \ln \varepsilon.$$

Относительно этих величин имеем линейную регрессионную модель

$$Y' = b'_0 + \beta_1 X' + \varepsilon', \quad (2.6.6)$$

которой соответствует уравнение линейной регрессии

$$\hat{y}' = b'_0 + b_1 x'. \quad (2.6.7)$$

Коэффициенты b'_0, b_1 вычисляются на основе МНК по формулам, приведенным в параграфе 2.3. Выполнив обратное преобразование $b_0 = e^{b'_0}$, получаем искомые оценки b_0, b_1 для коэффициентов нелинейной регрессии (2.6.3).

Замечание 2.6.1. Эффективность оценок, получаемых методом наименьших квадратов, основана на допущении о том, что возмущения ε_i не коррелированы между собой и подчиняются нормальному распределению $N(0, \sigma^2)$, т.е. имеет одинаковую

дисперсию σ^2 . К сожалению, выполнение нелинейных преобразований приводит к нарушению этого допущения. Для иллюстрации этого вернемся к преобразованному уравнению регрессии (2.6.7). Коэффициенты этого уравнения будут являться эффективными оценками для β'_0, β_1 , если $\varepsilon' = \ln \varepsilon \sim N(0, \sigma^2)$, т.е. возмущения ε_i исходной модели (2.6.3) должны иметь логарифмически нормальное распределение, что на практике встречается редко. Нарушение свойства гомоскедастичности приводит к тому, что вычисление на основе МНК коэффициенты *будут несмещенными, состоятельными оценками* для соответствующих коэффициентов регрессионной модели, но *они не обладают свойством эффективности*, т.е. возможно вычислить (используя другие алгоритмы) оценки с меньшей дисперсией. ♥

Второй подход используется в случаях, когда не возможно подобрать преобразования для перехода к новой линейной регрессии. Для примера рассмотрим регрессионную модель

$$Y = \beta_0 \cdot X^{\beta_1} + \varepsilon. \quad (2.6.8)$$

Логарифмирование этого уравнения не приводит к линейной регрессионной модели: $\ln Y = \ln(\beta_0 \cdot X^{\beta_1} + \varepsilon)$.

В этих случаях оценки для коэффициентов регрессионной модели вычисляются на основе минимизации функционала некоторого функционала, например, функционала метода наименьших квадратов. Так для модели (2.6.8) уравнение регрессии имеет вид

$$\hat{y} = b_0 x^{b_1}, \quad (2.6.9)$$

а минимизируемый функционал МНК определяется выражением (сравните с (2.3.3)):

$$F(b_0, b_1) = \sum_{i=1}^n (y_i - b_0 x_i^{b_1})^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2. \quad (2.6.10)$$

Существует достаточно большое число алгоритмы минимизации различных функционалов. Некоторые из этих алгоритмов реали-

зованы в табличном процессоре Excel (команда **Поиск решения** пункта меню **Сервис** – подробнее см. параграф 2.7).

Индекс детерминации и значимость нелинейной регрессии. Заметим, что коэффициент корреляции оценивает тесноту связи переменных X, Y только в случае линейной зависимости между этими переменными. В случае нелинейной регрессии абсолютная величина коэффициента корреляции может быть мала, несмотря на наличие нелинейной зависимости между X, Y .

Поэтому в случае нелинейной зависимости между исследуемыми факторами, степень их взаимосвязи характеризуется индексом корреляции I_{xy} , определяемый выражением $I_{xy} = \sqrt{1 - \frac{Q_e}{Q}}$,

где $Q_e = \sum_{i=1}^n (\hat{y}_i - y_i)^2$, $Q = \sum_{i=1}^n (y_i - \bar{y})^2$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, \hat{y}_i - значение

зависимой переменной Y , вычисленное по уравнению нелинейной регрессии при $x = x_i$. Очевидно, что величина этого показателя удовлетворяет неравенству: $0 \leq I_{xy} \leq 1$, причем $I_{xy} = 1$, когда все значения y_i “лежат” на линии регрессии.

Индексом детерминации называется величина

$$R_{xy}^2 = I_{xy}^2 = 1 - \frac{Q_e}{Q}. \quad (2.6.11)$$

Величина индекса детерминации изменяется в пределах

$$0 \leq R_{xy}^2 \leq 1$$

и показывает *какая часть (доля) вариации зависимой переменной Y обусловлена вариацией объясняющей переменной X* , т.е. индекс детерминации имеет тот же смысл, что и коэффициент детерминации R^2 линейной регрессии.

Если уравнение регрессии является линейной функцией, то справедливо тождество: $R_{xy}^2 = R^2$, где R^2 - коэффициент де-

терминации линейной регрессии. Это тождество является теоретическим обоснованием исследования возможности замены нелинейной регрессии линейной функцией. Заметим, что чем больше кривизна линии регрессии, тем величина коэффициента детерминации R^2 меньше индекса детерминации R_{xy}^2 . Близость этих величин означает, что нет необходимости усложнять уравнения регрессии и можно использовать линейную регрессию.

Для проверки нулевой гипотезы H_0 о возможности замены нелинейной регрессии линейной функцией определим следующий критерий

$$T_{нел} = \frac{R_{xy}^2 - R^2}{\delta_{\Delta}}, \quad (2.6.12)$$

где δ_{Δ} - ошибка разности $\Delta = R_{xy}^2 - R^2$, определяемая по формуле

$$\delta_{\Delta} = 2 \cdot \sqrt{\frac{(R_{xy}^2 - R^2) - (R_{xy}^2 - R^2)^2 \cdot (2 - (R_{xy}^2 + R^2))}{n}}. \quad (2.6.13)$$

Нулевая гипотеза H_0 принимается с уровнем значимости α , если выполняется неравенство

$$T_{нел} \leq t(1 - \alpha, n - 2), \quad (2.6.14)$$

где $t(1 - \alpha, n - 2) = \text{СТЮДРАСПОБР}(\alpha; n - 2)$. В противном случае принимается альтернативная гипотеза H_1 о существенном различии между R_{xy}^2 и R^2 и невозможности замены нелинейной регрессии линейной функцией.

Пример 2.6.2. В примере 2.6.1 по данным таблицы 2.2 было построено логарифмическое уравнение парной регрессии

$$\hat{f}(x) = 9.876 + 5.129 \ln(x). \quad (2.6.15)$$

Для этого уравнения индекс корреляции $I_{XY} = 0.99581$. Необходимо проверить возможность замены этого нелинейного уравнения линейным уравнением регрессией вида:

$$\hat{f}(x) = b_0 + b_1 x.$$

Решение. Используя МНК, по данным таблицы 2.2 определяем коэффициенты $b_0 = 9.28$, $b_1 = 1.777$, и получаем уравнение линейной регрессии

$$\hat{f}(x) = 9.28 + 1.777x. \quad (2.6.16)$$

Для этого уравнения коэффициент детерминации $R^2 = (0.97416)^2 = 0.94898$.

Вычислим следующие величины:

$$R_{xy}^2 - R^2 = (0.99581)^2 - (0.97416)^2 = 0.04265;$$

$$R_{xy}^2 + R^2 = (0.99581)^2 + (0.97416)^2 = 1.94063;$$

$$\delta = 2 \cdot \sqrt{\frac{0.04265 - (0.04265)^2 \cdot (2 - 1.94063)}{6}} = 0.16841.$$

Определяем значение критерия $T_{нел} = \frac{0.04265}{0.16841} = 0.25$. Из неравенства (см. (2.6.14)) $0.25 < t(0.95, n - 2) = 2$ следует вывод о возможности замены нелинейного уравнения регрессии (2.6.15) линейным уравнением (2.6.16). К этому выводу можно также прийти из анализа рис.2.9, на котором показан график логарифмической регрессии, близкий к прямой линии.

Используя индекс детерминации R_{xy}^2 можно выполнить проверку значимости построенной нелинейной регрессии. Для этого определим F -критерий

$$F = \frac{R_{xy}^2}{1 - R_{xy}^2} \cdot \frac{n - m - 1}{m}, \quad (2.6.17)$$

где m - число коэффициентов регрессии при переменной X . Тогда построенное уравнение нелинейной регрессии является значимым с уровнем значимости α , если выполняется неравенство

$$F > F_{1-\alpha; m; n-m-1}. \quad (2.6.18)$$

Напомним, что квантиль $F_{1-\alpha; m; n-m-1}$ можно вычислить в Excel с помощью выражения (см. 2.2.9):

$$F_{1-\alpha; m; n-m-1} = \text{FRASPOBR}(\alpha; m; n - m - 1). \quad (2.6.19)$$

Пример 2.6.3. Необходимо определить значимость уравнения регрессии $\hat{y} = 9.876 + 5.129 \cdot \ln x$, построенного в примере 2.6.1.

Решение. Возьмем значение индекса детерминации из примера 3.6.2 $R_{xy}^2 = 0.9916$ и вычислим значение критерия (2.6.17):

$$F = \frac{0.9916}{1 - 0.9916} \cdot (6 - 2) = 474.93.$$

Квантиль $F_{0.95; 1; 4} = 7.70$. Из выполнения первенства (2.6.18): $474.93 > 7.70$ следует вывод о значимости построенной нелинейной регрессии с уровнем значимости $\alpha = 0.05$.

2.7. Построение нелинейных регрессий в Excel

Вычислить коэффициенты нелинейной регрессии в Excel можно одним из следующих способов:

- используя команду *Добавить линию тренда*;
- используя команду *Поиск решения*.

Команда *Добавить линию тренда*. Используется для выделения тренда (медленных изменений) при анализе временных рядов. Однако эту команду можно использовать и для построения уравнения регрессии, рассматривая в качестве времени t независимую переменную x .

Эта команда позволяет построить следующие регрессии:

- линейную $\hat{y} = b_0 + b_1 x$
- полиномиальную $\hat{f} = b_0 + b_1 x + \dots + b_k x^k$ ($k \leq 6$);
- логарифмическую $\hat{y} = b_0 + b_1 \ln x$
- степенную $\hat{y} = b_0 x^{b_1}$;

- экспоненциальную $f = b_0 e^{b_1 x}$.

Для построения одной из перечисленных регрессий необходимо выполнить следующие шаги:

Шаг 1. В выбранном листе Excel ввести по столбцам исходные данные $\{x_i, y_i\}, i = 1, 2, \dots, n$ (см. рис. 2.10).

Шаг 2. По этим данным построить график в декартовой системе координат (см. рис. 2.10).

Шаг 3. Установить курсор на построенном графике, сделать щелчок правой кнопкой и в появившемся контекстном меню выполнить команду *Добавить линию тренда* (см. рис. 2.10).

Шаг 4. В появившемся диалоговом окне (см. рис. 2.11) активизировать закладку «Тип» и выбрать нужное уравнение регрессии.



Рис. 2.10. Построение графика по исходным данным

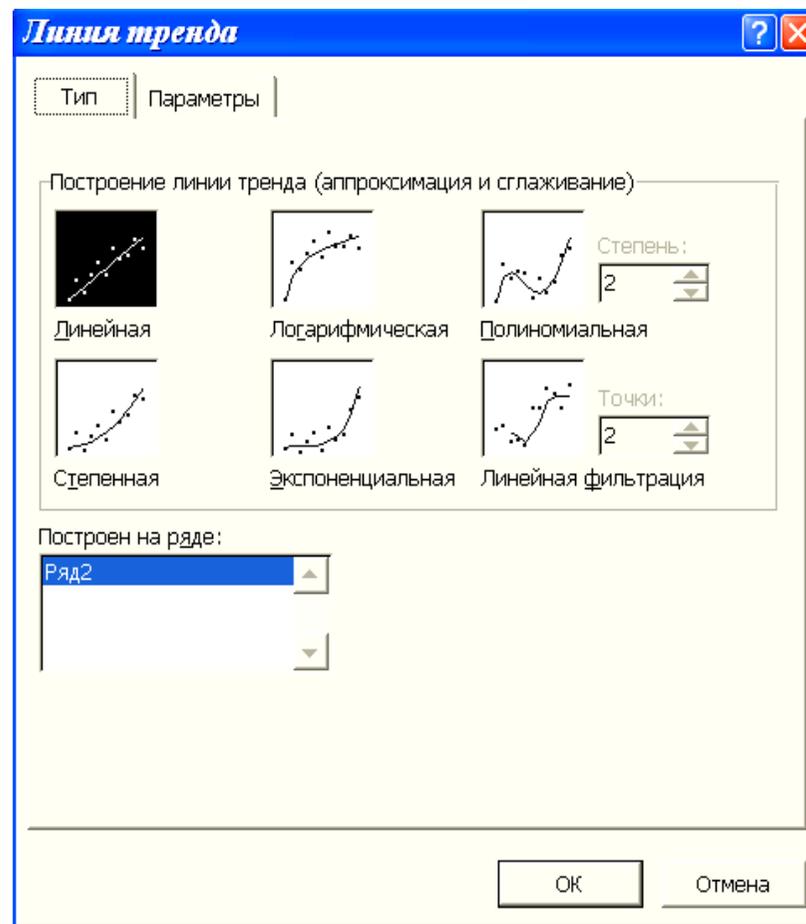


Рис. 2.11. Выбор вида уравнения регрессии

Шаг 5. Активизировать закладку «Параметры» (см. рис. 2.12) и «включить» необходимые для нас опции:

- «Показать уравнение на диаграмме» - на диаграмме будет показано выбранное уравнение регрессии с вычисленным коэффициентом;

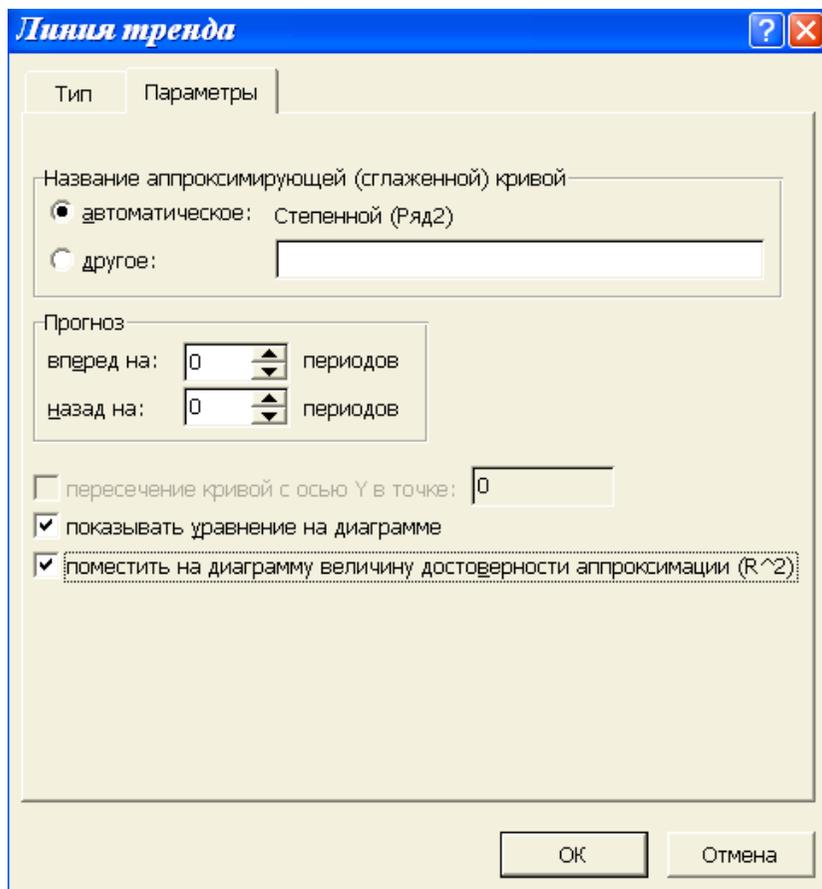


Рис. 2.12. Задание опций вывода информации

- «Поместить на диаграмму величину достоверности аппроксимации (R^2)» - на диаграмме будет показана значение индекса детерминации R_{xy}^2 (см.(2.6.11)), которое можно использовать для проверки значимости построенной регрессии с помощью F - критерия (2.6.18). Если по построенному уравнению регрессии необходимо выполнить прогноз, то нужно указать число периодов прогноза (см. рис. 2.12).

Назначение других опций понятны из своих названий.
Шаг 6. После задания всех перечисленных опций щелкнуть на кнопке «ОК» и на диаграмме появиться формула построенного уравнения регрессии и значение индекса детерминации R_{xy}^2 (выделено на рис. 2.13 затемнением).

Пример 2.7.1. По данным таблицы 2.2 построить уравнения регрессии (предусмотренные командой *Добавить линию тренда*) и по значению индекса детерминации R_{xy}^2 выбрать наилучшее уравнение.

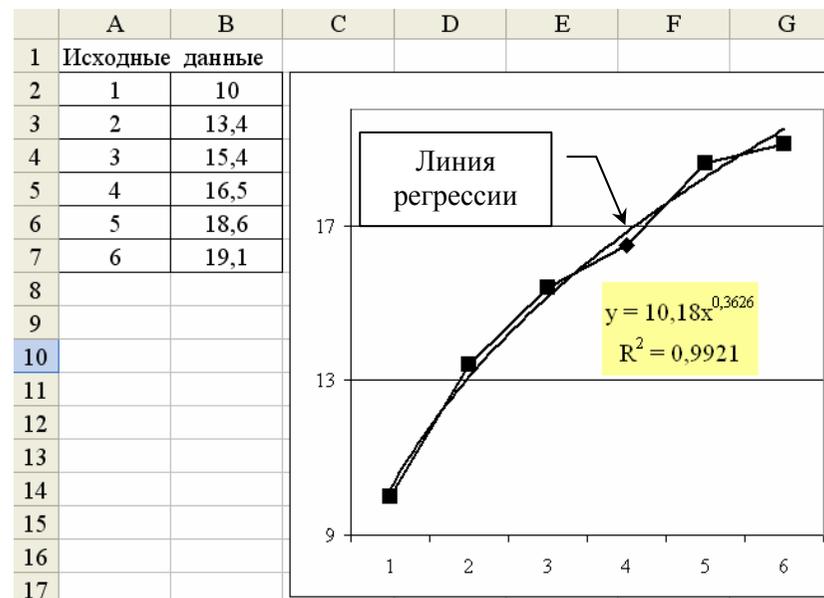


Рис. 2.13. График и уравнение построенной регрессии

Решение. Построение каждого из пяти уравнений осуществляем по описанным выше шагам. Для уравнения $y = b_0 \cdot x^{b_1}$ выполнение шагов иллюстрирует рис. 2.10 ÷ 2.13. В таблицу 2.3

вносим регрессионное уравнение и соответствующее значение R_{xy}^2 . Сравнивая величину индекса детерминации R_{xy}^2 для этих уравнений, в качестве «наилучшего» уравнения выбираем степенную регрессию $\hat{y} = 10.18x^{0.3226}$ (номер 5), для которой индекс детерминации $R_{xy}^2 = 0.9921$.

Таблица 2.3

№	Уравнение	R_{xy}^2	\hat{R}_{xy}^2
1	$\hat{y} = 9.28 + 1.777x$	0.949	0.938
2	$\hat{y} = 9.8759 + 5.1289 \cdot \ln x$	0.9916	0.9895
3	$\hat{y} = 6.93 + 3.5396x - 0.2518x^2$ (полиномиальная, $m = 2$)	0.9896	0.9827
4	$\hat{y} = 5.8333 + 4.9192x - 0.7087x^2 - 0.0435x^3$ (полиномиальная, $m = 3$)	0.9917	0.9792
5	$\hat{y} = 10.18x^{0.3626}$	0.9921	0.9901
6	$\hat{y} = 9.8675 \cdot e^{0.1225x}$	0.9029	0,8786

Замечание 2.6.1. Индекс детерминации R_{xy}^2 характеризует близость построенной регрессии к исходным данным, которые содержат «нежелательную» случайную составляющую ε . Очевидно, что, взяв полином 5-ого порядка, получаем «идеальное» значение $R^2 = 1$, по такое уравнение содержит в себе не только независимую переменную X , но составляющую ε и это снижает точность использования построенного уравнения для прогноза. Поэтому при выборе уравнения регрессии надо учитывать не только величину R_{xy}^2 , но и «сложность» регрессионного уравнения, определяемое качеством коэффициентов уравнения. Такой учет удачно реализован в так называемом *приведенном индексе*

детерминации (для линейной регрессии - *приведенный коэффициент детерминации*):

$$\hat{R}_{xy}^2 = 1 - \frac{(n-1) \cdot Q_e}{(n-m) \cdot Q} = 1 - \frac{n-1}{n-m} \cdot (1 - R_{xy}^2), \quad (2.6.20)$$

где m - количество вычисляемых коэффициентов регрессии. Видно, что при неизменных Q_e, Q увеличение m уменьшает значение \hat{R}_{xy}^2 . Если количество коэффициентов у сравниваемых уравнений регрессии одинаково (например, $m = 2$), то отбор наилучшей регрессии можно осуществлять по величине R_{xy}^2 . Если в уравнениях регрессии меняется число коэффициентов, то такой отбор целесообразно по величине \hat{R}_{xy}^2 .

Для иллюстрации этой рекомендации в таблице 3.3 приведены значения \hat{R}_{xy}^2 . Видно, что по величине \hat{R}_{xy}^2 наилучшей регрессией также является степенная регрессия. Полиномиальная регрессия третьей степени имеет \hat{R}_{xy}^2 значительно меньше коэффициента R_{xy}^2 . ♥

Команда Поиск решения (пункт меню Сервис). Используется для вычисления параметров (коэффициентов) при которых некоторый функционал, зависящий от этих параметров, достигает наименьшего или наибольшего значения. Эта команда позволяет также решать задачи *условной оптимизации*, т.е. когда ищется минимум или максимум функционала с учетом дополнительных ограничений (линейных или нелинейных) на значения искомых параметров. Например, искомый параметр b должен удовлетворять ограничению $0.2 \leq b < 1$. Эта возможность обуславливает существенное преимущество рассматриваемого подхода по сравнению с командой *Добавить линию тренда*. К недостатку следует отнести необходимость программировать «вручную» вычисление индекса детерминации R_{xy}^2 .

Применение команды *Поиск решения* для вычисления коэффициентов нелинейной регрессии на основе метода наименьших квадратов покажем на следующем примере.

Пример 2.7.2. По данным таблицы 2.2 построить уравнения степенной регрессии, используя команду «Поиск решения».

Решение. Первоначально на листе Excel введем исходные данные: значения x_i в ячейках A2 ÷ A7; значения y_i в ячейках B2 ÷ B7. Затем в ячейку B9 введем произвольное значение коэффициента b_0 , а в ячейку B10 – произвольное значение коэффициента b_1 . На рис. 2.14 показан фрагмент документа Excel с введенными данными.

Следующим шагом является вычисление по уравнению регрессии значений $\hat{y}_i = b_0 \cdot x_i^{b_1}$, $i = 1, \dots, 6$. Так для вычисления значения \hat{y}_1 в ячейке C2 программируется выражение `=B$9*A2^$B$10`. Использование абсолютных адресов для ячеек B9, B10 позволяет «размножить» это выражение на ячейки C3 – C7. Далее в ячейках D2 – D7 вычисляется квадрат невязки при соответствующем значении x_i . Так в ячейке D2 вводится выражение `=(C2-B2)^2`, «размножаемое» в ячейках D3 – D7. Значение минимизируемого функционала МНК вычисляется в ячейке D9 (см. рис. 2.14). На этом подготовка необходимой для команды «Поиск решения» информации завершается.

Для выполнения команды «Поиск решения» необходимо обратиться к пункту основного меню **Сервис** и в появившемся меню щелкнуть мышью на команде **Поиск решения**. Затем в появившемся диалоговом окне выполнить следующие действия (см. рис. 2.14):

- в поле ввода *Установить целевую ячейку* ввести адрес ячейки, в которой вычисляется значение минимизируемого функционала (в нашем примере – D9);
- включить опцию *Минимальное значение* (ищутся значения коэффициентов, при которых функционал достигает своего минимального значения);

- в поле ввода *Изменяя значения* ввести адреса ячеек, в которых находятся значения искомых коэффициентов (в нашем примере это ячейки B9, B10);
- щелкнув мышью на кнопке *Добавить* формируем ограничения на значения искомых коэффициентов (в нашем примере это требования неотрицательности искомых коэффициентов).

	A	B	C	D	E	F
1	Исходные данные		\hat{y}_i	$(\hat{y}_i - y_i)^2$		
2	1	10	1	81,000		
3	2	13,4	1,231	148,081		
4	3	15,4	1,390	196,269		
5	4	16,5	1,516	224,529		
6	5	18,6	1,621	288,298		
7	6	19,1	1,712	302,351		
8						
9	b_0	1	$F(b_0, b_1)$	1240,528		
10	b_1	0,3		=СУММ(D2:D7)		

Рис. 2.14. Задание параметров команды *Поиск решения*

После выполнения этих операций щелкнуть на кнопке *Выполнить*. Начинается поиск решения введенной оптимизационной задачи и после некоторого времени на экране появляется новое диалоговое окно *Результаты поиска решения* (см. рис. 2.15). Для сохранения найденных значений коэффициентов в соответствующих ячейках необходимо включить опцию *Сохранить найденное решение* и щелкнуть на кнопке *ОК*.

Из рис. 2.15 видно, что вычисленные значения коэффициентов находятся в ячейках B9, B10 и равны: $b_0 = 10.28299$, $b_1 = 0.354496$. Ячейка D9 содержит значение минимизируемого функционала. Заметим, что найденные значения коэффициентов незначительно отличаются от значений, вычисленных в примере 3.7.1 с помощью команды *Добавить линию тренда*.

В заключении этого параграфа заметим, что использование табличного процессора Excel и двух рассмотренных подходов позволяет построить нелинейную парную регрессии любой «сложности».

2.8. Робастные методы оценивания и метод наименьших модулей

Как уже отмечалось (см. замечание 2.6.1) метод наименьших квадратов вычисляет эффективные оценки в случаях, когда возмущения эконометрической модели ε_i распределены по нормальному закону. Однако возможны случаи нарушения этого предположения. Примером может служить присутствие в выборке аномальных измерений, т.е. измерений, дисперсия которых существенно (на порядок и больше) превосходит дисперсию «нормальных» измерений σ^2 . В этом случае из-за наличия квадратов невязок в функционале метода наименьших квадратов происходит «подтягивание» уравнения регрессии к аномальному измерению и это существенно ухудшает точность построенной эконометрической модели.

Поэтому в последние десятилетия используются методы *робастного оценивания*, которые слабо чувствительны к нарушению априорных предположений о законе распределения или числовых характеристик случайной величины ε_i . Это достигается

заменой суммы квадратов в минимизируемом функционале на другие суммируемые величины. Примером таких методов может служить *метод наименьших модулей* (МНМ).

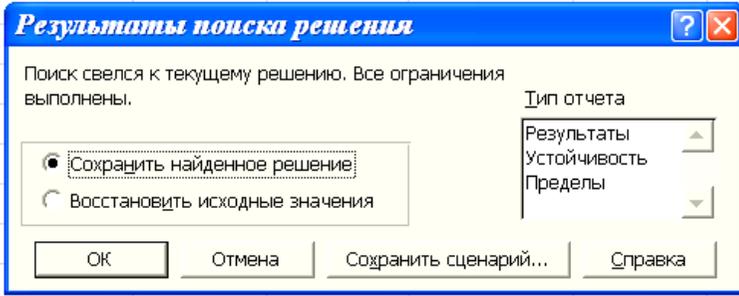
	A	B	C	D	E
1	Исходные данные		\hat{y}_i	$(\hat{y}_i - y_i)^2$	
2	1	10	10,28299	0,080	
3	2	13,4	13,147	0,064	
4	3	15,4	15,179	0,049	
5	4	16,5	16,809	0,096	
6	5	18,6	18,193	0,166	
7	6	19,1	19,408	0,095	
8					
9	b_0	10,28299	$F(b_0, b_1)$	0,549	
10	b_1	0,354496		=СУММ(D2:D7)	
11					
12					
13					
14					
15					
16					
17					
18					

Рис. 2.15. Результаты выполнения команды *Поиск решения*

В этом методе в минимизируемом функционале стоит не сумма квадратов невязок, а *сумма модулей невязок*. Так для линейной парной регрессии функционал МНМ имеет вид

$$F(b_0, b_1) = \sum_{i=1}^n |y_i - \hat{y}_i|, \quad (2.8.1)$$

где $\hat{f}_i = b_0 + b_1 x_i$. Для уравнения регрессии, построенного методом МНМ справедливо примерное равенство количества невязок $e_i = y_i - (b_0 + b_1 x_i)$ с плюсом и минусом.

Пример 2.8.1. В таблице 2.1 примера 2.3.1 измерение номер 4 является аномальным. Новые данные представлены в таблице 2.3 и в ней аномальное измерение выделено цветом.

Необходимо по этим данным построить линейные регрессии, используя методы наименьших квадратов и наименьших модулей.

Таблица 2.3

i	1	2	3	4	5	6	7	8	9	10
x_i	8	11	12	9	8	8	9	9	8	12
y_i	5	10	10	44	5	6	6	5	6	8

Решение. Для нахождения коэффициентов b_0, b_1 МНК воспользуемся фрагментом документа примера 2.3.1, приведенного на рис. 2.4, заменив в нем четвертое измерение. Полученный документ приведен на рис. 2.16. Вычисленные коэффициенты: $b_0 = 6.65, b_1 = 0.410$, а уравнение регрессии имеет вид:

$$\hat{f}(x) = 6.65 + 0.410x.$$

Фрагмент документа для вычисления коэффициентов методом наименьших модулей с использованием команды *Поиск решения* показан на рис 2.17 (более подробно см. пункт 2.7 и пример 2.7.1). Вычисленные коэффициенты $b_0 = -3.14, b_1 = 1.02$, а уравнение регрессии имеет вид:

$$\hat{f}(x) = -3.14 + 1.02x.$$

Сравнивая со значениями коэффициентов $b_0 = -2.75, b_1 = 1.016$, вычисленными по «нормальным» наблюдениям, очевидно, что метод наименьших модулей существенно точнее оценивает ко-

эффициенты линейной регрессии по выборочным данным, содержащим аномальные измерения.

	A	B	C	D	E	F	G	H
1		Исходные данные						
2		x_i	y_i	x_i^2	$x_i \cdot y_i$			
3		8	5	64	40	=(E13-B13*C13)/(D13-B13^2)		
4		11	10	121	110			
5		12	10	144	120	b_1	0,410	
6		9	44	81	396	b_0	6,65	
7		8	5	64	40			
8		8	6	64	48		=C13-G5*B13	
9		9	6	81	54			
10		9	5	81	45			
11		8	6	64	48			
12		12	8	144	96			
13	Средние значения	9,4	10,5	90,8	99,7			

Рис. 2.16. Вычисление коэффициентов регрессии МНК

	A	B	C	D	E	F
1		Исходные данные				=ABS(C3-D3)
2		x_i	y_i	\hat{y}_i	$ e_i $	
3		8	5	5,000011	1,08643E-05	
4		11	10	8,053367	1,946633357	
5		12	10	9,071152	0,928848097	
6		9	44	6,017796	37,98220388	
7		8	5	5,000011	1,08643E-05	
8		8	6	5,000011	0,999989136	
9		9	6	6,017796	0,017796124	
10		9	5	6,017796	1,017796124	
11		8	6	5,000011	0,999989136	
12		12	8	9,071152	1,071151903	
13						
14		b_0	-3,14		$F(b_0, b_1)$	44,964
15		b_1	1,02			
16					=СУММ(E3:E12)	

Рис. 2.17. Вычисление коэффициентов регрессии МНМ

Аналогичный вывод можно сделать из анализа графиков уравнений регрессий, приведенных на рис. 2.18.

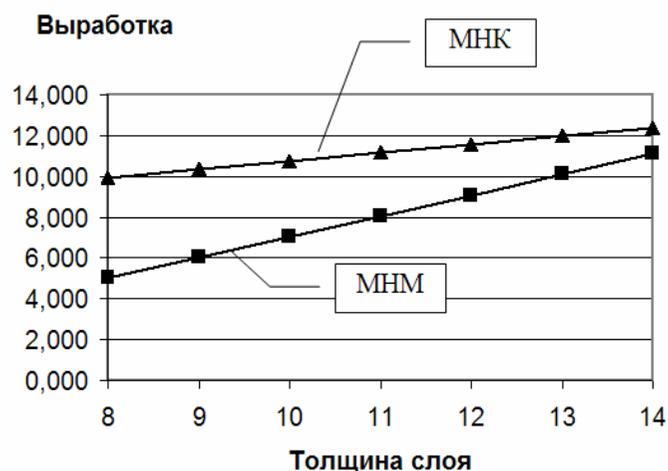


Рис. 2.18. Графики регрессий МНК и МНМ

ЛАБОРАТОРНАЯ РАБОТА № 2.1 «Построение парной линейной регрессии»

Цель работы. Используя табличный процессор Excel, построить линейную парную регрессию, описывающую зависимость удельного веса бракованной продукции от удельного веса рабочих со специальной подготовкой и определить значимость построенного уравнения.

Исходные данные. В таблице Л2.1 приведен удельный вес рабочих со специальной подготовкой в % (объясняющая переменная X) и удельный вес бракованной продукции в % (зависимая переменная Y).

Содержание работы

1. Ввести исходные данные таблицы Л2.1.
2. Построить диаграмму рассеяния.
3. Вычислить коэффициент корреляции r_{XY} (см. (2.3.15)).

4. Вычислить коэффициенты b_0, b_1 выборочного уравнения линейной регрессии.

Таблица Л2.1

№ п/п	1	2	3	4	5	6	7
X	15	25	35	45	55	65	70
Y	18	12	10	8	6	5	3

5. Вычислить по построенному уравнению регрессии значения $\hat{y}_i = b_0 + b_1 x_i$; $e_i = \hat{y}_i - y_i$; $i = 1, 2, \dots, 7$. По значениям y_i, \hat{y}_i, e_i построить диаграмму (по оси абсцисс откладываются значения x_i) и высказать мнение об адекватности построенного уравнения регрессии исходным данным.

6. Проверить значимость построенного уравнения регрессии по критерию Фишера при двух уровнях значимости $\alpha = 0.01, \alpha = 0.05$.

7. Проверить значимость вычисленных коэффициентов b_0, b_1 при двух уровнях значимости $\alpha = 0.01, \alpha = 0.05$.

8. Вычислить коэффициент детерминации R^2 и высказать мнение, насколько хорошо построенная регрессия определяет зависимость Y от X .

Контрольные значения: $r_{XY} = 0.967$, $b_0 = 19.41$, $b_1 = -0.24$, $R^2 = 0.936$, вычисленное значение статистики Фишера $F_{\text{выч}} = 73.46$.

Рекомендации. Вычисление коэффициента корреляции и оценок b_0, b_1 можно осуществить одним из следующих способов:

- запрограммировать в ячейках Excel необходимые вычисления (см. пример 2.3.1);
- использовать соответствующие статистические функции Excel (см. пример 2.3.4).

ЛАБОРАТОРНАЯ РАБОТА № 2.2

«Интервальные оценки для парной линейной регрессии»

Цель работы. Используя табличный процессор Excel, построить интервальные оценки для коэффициентов и доверительные области оценки для коэффициентов и значений линейной регрессии.

Исходные данные. В таблице Л2.1 приведен удельный вес рабочих со специальной подготовкой в % (объясняющая переменная X) и удельный вес бракованной продукции в % (зависимая переменная Y). В лабораторной работе 2.1 были получены следующие оценки:

$$b_0 = 19.41, \quad b_1 = -0.24, \quad r_{XY} = 0.967.$$

Содержание работы.

1. Вычислить оценку s^2 для дисперсии σ^2 (см. (2.3.19)).
2. Вычислить оценки $s_{b_0}^2, s_{b_1}^2$ дисперсий оценок b_0, b_1 (см. (2.3.20), (2.3.21)).
3. Построить интервальную оценку (доверительный интервал) для коэффициента β_0 с надежностью $\gamma = 0.9, \gamma = 0.95$ (см. (2.4.3)).
4. Построить интервальную оценку (доверительный интервал) для коэффициента β_1 с надежностью $\gamma = 0.9, \gamma = 0.95$ (см. (2.4.4)).
5. Вычислить интервальную оценку (т.е. значения y_i^H и y_i^B) для функции регрессии $M(Y|x)$ при $x = x_i, i = 1, 2, \dots, 7$ (см. (2.4.10)) с надежностью $\gamma = 0.95$. Построить диаграмму по значениям $y_i^H, y_i^B, \hat{y}_i, i = 1, 2, \dots, 7$ (по оси X откладываются значения x_i).

Контрольные значения: среднеквадратические ошибки:

$$s_{b_0} = 1.34, \quad s_{b_1} = 0.028.$$

Рекомендации.

1. Для вычисления оценок дисперсий используйте фрагмент программы, приведенный на рис. 2.4 (пример 2.3.3).

2. Для построения доверительных интервалов для β_0, β_1 используйте вычисления примера 2.4.1.

3. Для построения доверительной области для функции регрессии $M(Y|x)$ используйте вычисления примера 2.4.2.

КОНТРОЛЬНАЯ РАБОТА № 2.1

Парная регрессия

Данные, характеризующие прибыль торговой компании «Все для себя» за первые 10 месяцев 2003 года (в тыс. руб.), даны в следующей таблице:

январь	февраль	март	апрель	май
382 + N	402 + N	432 + N	396 + N	454 + N
июнь	июль	август	сентябрь	октябрь
419 + N	460 + N	447 + N	464 + N	498 + N

где N – последняя цифра номера зачетной книжки студента.

Требуется:

1. Построить диаграмму рассеяния.
2. Убедится в наличии тенденции (тренда) в заданных значениях прибыли фирмы и возможности принятия гипотезы о линейном тренде.
3. Построить линейную парную регрессию (регрессию вида $\hat{y}(x) = b_0 + b_1x$). Вычисление коэффициентов b_0, b_1 выполнить методом наименьших квадратов.
4. Нанести график регрессии на диаграмму рассеяния.
5. Вычислить значения статистики F и коэффициента детерминации R^2 . Проверить гипотезу о значимости линейной регрессии.
6. Вычислить выборочный коэффициент корреляции и проверить гипотезу о ненулевом его значении.
7. Вычислить оценку дисперсии случайной составляющей эконометрической модели.

8. Проверить гипотезы о ненулевых значениях коэффициентов β_0, β_1 .

9. Построить доверительные интервалы для коэффициентов β_0, β_1 .

10. Построить доверительные интервалы для дисперсии случайной составляющей эконометрической модели.

11. Построить доверительную область для условного математического ожидания $M(Y|x)$ (диапазон по оси январь – декабрь). Нанести границы этой области на диаграмму рассеяния.

12. С помощью линейной парной регрессии сделать прогноз величины прибыли и нанести эти значения на диаграмму рассеяния. Сопоставить эти значения с границами доверительной области для условного математического ожидания $M(Y|x)$ и сделать вывод о точности прогнозирования с помощью построенной регрессионной модели.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Объясните, чем вызвано появление в модели парной регрессии случайного слагаемого ε ?

2. Каким условиям должны удовлетворять возмущения ε_i ?

3. Почему перед построением модели парной регрессии необходимо вычислять выборочный коэффициент корреляции?

4. Докажите (выполнив все промежуточные преобразования), что из системы (2.3.5) следует система (2.3.6).

5. Зависимость времени разговора покупателя с продавцом спортивного отдела (переменная Y в минутах) от суммы покупки (переменная X в у. е.) определяется следующим уравнением регрессии:

$$\hat{y} = 14.89 + 0.0393 \cdot x.$$

Необходимо вычислить коэффициент эластичности, если $\bar{x} = 130$, $\bar{y} = 20$. Определите экономический смысл вычисленной величины коэффициента эластичности.

6. В таблицу 2.1 внесите следующие изменения: а) значения (x_4, y_4) замените на $x_4 = 10, y_4 = 8$; б) значения (x_1, y_1) замените на $x_1 = 7, y_1 = 4$. Вычислите коэффициенты b_0, b_1 линейной регрессии. Сравните их с коэффициентами b_0, b_1 примера 2.3.1.

7. Какими свойствами обладают оценки b_0, b_1 , вычисленные методом наименьших квадратов при выполнении предположений P1 ÷ P3?

8. По каким показателям можно судить о значимости построенной линейной регрессии в целом?

9. Поясните статистический смысл коэффициента детерминации R^2 .

10. Докажите справедливость диапазона (2.5.4) изменения коэффициента детерминации R^2 .

11. В примере 2.4.1 были построены доверительные интервалы для коэффициентов β_0, β_1 с надежностью $\gamma = 0.95$. Как изменится длина этих интервалов при увеличении γ до значения $\gamma = 0.99$ и уменьшении γ до значения $\gamma = 0.9$.

12. Сформулируйте статистические гипотезы, соответствующие проверке значимости коэффициента b_1 линейной регрессии.

13. Сформулируйте статистические гипотезы, соответствующие проверке значимости коэффициента корреляции r_{XY} .

14. При исследовании корреляционной зависимости между индексом Y нефтяной компании и ценой на нефть X получены следующие значения: $\bar{x} = 16.2$ (усл. ед.), $\bar{y} = 4000$ (усл. ед.), $s_x^2 = 4$, $s_y^2 = 500$, $m_{XY} = 40$.

Необходимо: а) вычислить выборочный коэффициент корреляции r_{XY} ; б) построить уравнение линейной парной регрессии вида $\hat{y}(x) = b_0 + b_1 x$ (т.е. вычислить коэффициенты b_0, b_1); в) определить прогноз индекса Y при цене на нефть 16.5 усл. ед.

Рекомендации: для вычисления r_{XY} использовать формулу (2.3.13); для вычисления b_0, b_1 - формулы (2.3.8), (2.3.9).

Глава 3. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Парная регрессия может дать хороший результат, если изменением других факторов, воздействующих на объект исследования (т.е. на переменную Y) можно пренебречь. Например, при построении модели потребления того или иного товара от дохода исследователь предполагает, что в каждой группе дохода одинаково влияние на потребителя таких факторов, как цена товара, размер семьи, ее состав и т.д. Следовательно, в данном примере построение парной регрессии осуществляется при неизменном уровне других факторов, т.е. мы пренебрегаем влиянием (изменением) этих факторов. В ряде случаев не удастся обеспечить не изменчивость всех прочих условий для оценки влияния одного исследуемого фактора. Тогда следует попытаться выявить влияние других факторов, введя их в эконометрическую модель, т.е. приходим к модели множественной регрессии, определяемой как условное математическое ожидание зависимой величины Y при k фиксированных значениях x_1, x_2, \dots, x_k , объясняющих переменных $X_1, X_2, X_3, \dots, X_k$, т.е.

$$f(x_1, x_2, x_3, \dots, x_k) = M(Y | x_1, x_2, x_3, \dots, x_k)$$

Также к множественной регрессии мы приходим, когда априори известно о влиянии на зависимую переменную Y нескольких объясняющих переменных $X_1, X_2, X_3, \dots, X_k$ (т.е. число объясняющих переменных равно $k > 1$).

Множественная регрессия широко используется в решении проблем спроса, доходности акций, при изучении функции издержек производства и целого ряда других вопросов эконометрики. В настоящее время множественная регрессия – один из наиболее распространенных методов в эконометрике.

Основная цель *множественного регрессионного анализа* – построить регрессионную модель с большим количеством факторов, определив при этом влияние каждого из них в отдельности, а также совокупное их воздействие на зависимую переменную.

3.1. Классическая линейная модель множественной регрессии

В отличие от парной регрессии $f(x) = M(Y|x)$ множественная регрессия определяется как условное математическое ожидание зависимой величины Y при k фиксированных значениях x_1, x_2, \dots, x_k , т.е.

$$f(x_1, x_2, \dots, x_k) = M(Y | x_1, x_2, \dots, x_k) \quad (3.1.1)$$

Линейная множественная регрессия. Часто в качестве функции $f(x_1, x_2, \dots, x_k)$ принимают линейную функцию, и мы приходим к *линейной множественной регрессионной модели* вида

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon, \quad (3.1.2)$$

где $\beta_0, \beta_1, \dots, \beta_k$ – коэффициенты регрессионной модели, ε – случайное слагаемое, называемое возмущением. Обозначим i -ое наблюдение зависимой переменной как y_i , а наблюдаемые значения объясняющих переменных – $x_{i1}, x_{i2}, \dots, x_{ik}$, т.е. в обозначении x_{ij} первый индекс i определяет номер измерения, а второй j – номер переменной. Тогда имеет место следующая модель наблюдений:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n. \quad (3.1.3)$$

Включение в регрессионную модель новых объясняющих переменных усложняет получаемые формулы и вычисления. Это приводит к необходимости использования матричных обозначений и матричных вычислений.

Введем вектор y (другими словами матрицу-столбец), состоящий из n проекций и матрицу X размером $n \times (k+1)$ (состоящую из n строк и $k+1$ столбца):

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}; \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix},$$

а также векторы:

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \text{ – вектор параметров;} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ – случайный вектор возмущений.}$$

В дальнейшем матрицы обозначаются прописными буквами, а векторы – строчными.

Тогда в матричном виде модель наблюдений (3.1.3) примет вид

$$y = X\beta + \varepsilon \quad (3.1.4)$$

Ограничения и условия классической множественной регрессионной модели. По аналогии с парной регрессией приведем ряд условий (известных как условия Гаусса-Маркова), которым должна удовлетворять классическая регрессионная модель (3.1.4).

P1. Матрица X – неслучайная матрица, а ε – случайный вектор.

P2. $M(\varepsilon) = 0_n$, (3.1.5)

где 0_n – вектор, все n проекций которого равны нулю (т.е. нулевой вектор).

P3. $V_\varepsilon = M[\varepsilon\varepsilon^T] = \sigma^2 I$, (3.1.6)

где V_ε – ковариационная матрица размера $n \times n$; I – единичная матрица размера $n \times n$. Напомним, что i, i -ый элемент ковариационной матрицы V_ε определяет дисперсию i -ой проекции вектора ε , а i, j -ый элемент равен корреляционному моменту $\mu_{i,j} =$

$M(\varepsilon_i \cdot \varepsilon_j)$. Если проекции ε_i и ε_j статистически независимы, то $\mu_{i,j} = 0$ и матрица V_ε является диагональной.

P4. Случайный вектор ε подчиняется нормальному распределению $N(0_n, \sigma^2 I)$.

P5. Ранг матрицы X $rank(X)$ удовлетворяет условию

$$rank(X) = k + 1 \leq n. \quad (3.1.7)$$

Поясним некоторые обозначения.

Так как вектор 0_n состоит из n нулевых проекций, то условие (3.1.5) означает, что каждая проекция ε_i имеет нулевое математическое ожидание.

Матрица I_n является диагональной матрицей (т.е. на главной диагонали стоят 1, а остальные элементы равны 0) размером $n \times n$. Тогда условие (3.1.6) можно переписать в виде

$$M(\varepsilon_i \varepsilon_j) = \begin{cases} \sigma^2, & \text{если } i = j; \\ 0, & \text{если } i \neq j. \end{cases} \quad (3.1.8)$$

Напомним, что это условие характеризует свойство гомоскедастичности регрессионной модели.

Ранг матрицы характеризуется количеством линейно независимых строк или столбцов, и он определяет количество линейно независимых решений, которые можно найти из системы уравнений с такой матрицей. В нашем случае число неизвестных коэффициентов линейной регрессии равно $m = k + 1$ и поэтому ранг матрицы X должен быть равен m (условие (3.1.7)). Неравенство $k + 1 \leq n$ требует, чтобы число неизвестных было не больше числа уравнений.

Линейной регрессионной модели (3.1.4) соответствует уравнение множественной линейной регрессии вида

$$\hat{y} = b_0 + b_1 x_1 + \cdots + b_k x_k, \quad (3.1.9)$$

где b_0, b_1, \dots, b_k – коэффициенты регрессии, являющиеся оценками для $\beta_0, \beta_1, \dots, \beta_k$ и которые необходимо вычислить, решая сис-

тему уравнений $y = Xb + e$ по заданным вектору y и матрице наблюдений X .

3.2. Оценка коэффициентов линейной модели методом наименьших квадратов

Обратимся к системе уравнений $y = Xb + e$, которая «содержит информацию» об искомом векторе коэффициентов b . Из-за наличия вектора невязок e эта система, как правило, является *несовместной*, т.е. нельзя найти ни одного вектора b , который бы удовлетворял матричному тождеству $Xb = y$. Поэтому от поиска «точного решения» системы перейдем к вычислению «приближенного решения». Одно из таких приближенных решений можно найти, используя *метод наименьших квадратов*.

Вычисление коэффициентов уравнения регрессии методом наименьших квадратов. Введем функционал (сравните с (2.3.3))

$$F(b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 = (y - Xb)^T (y - Xb) = e^T e, \quad (3.2.1)$$

который характеризует отклонение значений \hat{y}_i , предсказанных линейной регрессией (3.1.9) при $x_1 = x_{i1}, \dots, x_k = x_{ik}$ (вектор Xb) от заданных значений y_i (вектор y). Вектор невязок e имеет n проекций $e_i = y_i - \hat{y}_i$. Согласно методу наименьших квадратов, в качестве решения системы (3.1.10) принимается вектор коэффициентов b , доставляющий минимум функционалу $F(b)$. *Необходимые и достаточные условия минимума* этого функционала определяются матричным тождеством:

$$\frac{\partial F}{\partial b} = 2X^T Xb - 2X^T y = 0, \quad (3.2.2)$$

из которого получаем *систему нормальных уравнений*

$$X^T Xb = X^T y. \quad (3.2.3)$$

Матрица $X^T X$ имеет размер $m \times m$ и следующую структуру:

$$X^T X = \begin{pmatrix} n & \sum x_{i1} & \cdots & \sum x_{ik} \\ \sum x_{i1} & \sum x_{i1}^2 & \cdots & \sum x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ik} & \sum x_{i1}x_{ik} & \cdots & \sum x_{ik}^2 \end{pmatrix},$$

а вектор $X^T y$ имеет m проекций:

$$X^T y = \begin{pmatrix} \sum y_i \\ \sum y_i x_{i1} \\ \vdots \\ \sum y_i x_{ik} \end{pmatrix},$$

где знак \sum подразумевает операцию суммирования $\sum_{i=1}^n$.

В отличие от системы (3.1.10) система нормальных уравнений (3.2.3) всегда имеет решение (т.е. всегда совместна) и для того, чтобы это решение было единственным, необходимо выполнение условия

$$\text{rank}(X^T X) = \text{rank}(X) = k + 1 = m \quad (3.2.4)$$

Это условие гарантирует *существование обратной матрицы* $(X^T X)^{-1}$ и тогда решение метода наименьших квадратов определяется матричным выражением

$$b = (X^T X)^{-1} (X^T y), \quad (3.2.5)$$

которое будет использоваться при вычислении коэффициентов регрессии в Excel.

Заметим, что решение МНК в вычислительной математике называют *псевдорешением системы* $Xb = y$, подчеркивая тем самым приближенный характер решения МНК.

Пример 3.2.1. Данные о сменной добыче угля на одного рабочего (переменная Y – измеряется в тоннах), мощности пласта (переменная X_1 – измеряется в метрах) и уровне механизации работ в шахте (переменная X_2 – измеряется в процентах), характеризующие процесс добычи угля в 10 шахтах приведены в таблице 3.1.

Предполагая, что между переменными Y, X_1, X_2 существует линейная зависимость, необходимо найти аналитическое выражение для этой зависимости, т.е. построить уравнение линейной регрессии.

Таблица 3.1

Номер шахты i	x_{i1}	x_{i2}	y_i
1	8	5	5
2	11	8	10
3	12	8	10
4	9	5	7
5	8	7	5
6	8	8	6
7	9	6	6
8	9	4	5
9	8	5	6
10	12	7	8

Решение. Обозначим

$$Y = \begin{vmatrix} 5 \\ 10 \\ \vdots \\ 8 \end{vmatrix}, \quad X = \begin{vmatrix} 1 & 8 & 5 \\ 1 & 11 & 8 \\ \dots & \dots & \dots \\ 1 & 12 & 7 \end{vmatrix}$$

и вычислим следующие величины:

- матрицу

$$X^T X = \begin{vmatrix} 10 & 94 & 63 \\ 94 & 908 & 603 \\ 63 & 603 & 417 \end{vmatrix}$$

размером 3×3 ;

- вектор

$$X^T y = \begin{vmatrix} 1 & 1 & \dots & 1 \\ 8 & 11 & \dots & 12 \\ 5 & 8 & \dots & 7 \end{vmatrix} \cdot \begin{vmatrix} 5 \\ 10 \\ \dots \\ 8 \end{vmatrix} = \begin{vmatrix} 68 \\ 664 \\ 445 \end{vmatrix},$$

содержащий 3 проекции;

- обратную матрицу

$$A^{-1} = (X^T X)^{-1} = \frac{1}{3738} \cdot \begin{vmatrix} 15027 & -1209 & -522 \\ -1209 & 201 & -108 \\ -522 & -108 & 244 \end{vmatrix}.$$

Тогда в соответствии с выражением (3.2.5) определяем вектор коэффициентов

$$b = A^{-1}(X^T y) = \begin{vmatrix} -3.5393 \\ 0.8539 \\ 0.3670 \end{vmatrix}.$$

Построенное уравнение линейной множественной регрессии имеет вид

$$\hat{y}(x) = -3.54 + 0.854 x_1 + 0.367 x_2.$$

Оно показывает, что при увеличении *только мощности пласта* X_1 (при неизменном X_2) на 1м добыча угля Y увеличивается в среднем на 0.854 т, а при увеличении *только уровня механизации работ* X_2 (при неизменном X_1) – в среднем на 0.367 т. ☉

Стандартизованные коэффициенты регрессии и коэффициенты эластичности. На практике часто бывает необходимым сравнение влияние на зависимую переменную различных объясняющих переменных, когда эти переменные имеют разные единицы измерений. В этом случае используют *стандартизованные*

коэффициенты регрессии b'_j и коэффициенты эластичности E_j , $j = 1, \dots, k$.

Стандартизованный коэффициент регрессии b'_j определяются выражением

$$b'_j = b_j \cdot \frac{S_{x_j}}{S_y}, j = 1, \dots, k, \quad (3.2.6)$$

где

$$s_y = \left(\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 \right)^{1/2}, \quad s_{x_j} = \left(\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right)^{1/2},$$

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^n x_{ij}, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$$

- выборочные средние значения переменной x_j и зависимой переменной y . Стандартизованный коэффициент b'_j показывает, на сколько величин S_y изменяется в среднем зависимая переменная y при увеличении только j -ой объясняющей переменной на S_{x_j} при неизменном среднем уровне других объясняющих переменных.

В отличие от коэффициентов b_j , которые не сравнимы между собой, стандартизованные коэффициенты регрессии можно сравнить между собой и таким образом ранжировать объясняющие переменные по "силе их воздействия" на переменную Y - чем больше значение коэффициента (по модулю), тем больше влияние на Y оказывает переменная, соответствующая этому стандартизованному коэффициенту.

Это обстоятельство позволяет использовать стандартизованные коэффициенты регрессии для исключения из эконометрической модели не значащих или слабо значащих факторов с наименьшими значениями b'_j .

Коэффициент эластичности E_j вычисляется по формуле

$$E_j = b_j \cdot \frac{\bar{x}_j}{\bar{y}}, j = 1, \dots, k. \quad (3.2.7)$$

и показывает, на сколько процентов (от средней) изменится в среднем величина Y при увеличении только X_j на 1%.

Пример 3.2.2. По коэффициентам регрессии необходимо вычислить стандартизованные коэффициенты b'_1, b'_2 и соответствующие коэффициенты эластичности.

Решение. Первоначально вычислим стандартизованные коэффициенты:

$$b'_1 = 0.8539 \cdot \frac{1.56}{1.83} = 0.728; \quad b'_2 = 0.3679 \cdot \frac{1.42}{1.83} = 0.285,$$

а затем коэффициенты эластичности

$$E_1 = 0.8539 \cdot \frac{9.4}{6.8} = 1.180; \quad E_2 = 0.3679 \cdot \frac{6.3}{6.8} = 0.340.$$

Вычисленные показатели указывают, что на сменную добычу угля большее влияние оказывает фактор «мощности пласта», чем фактор «уровень механизации» ●

Свойства оценок метода наименьших квадратов. Напомним, что вектор коэффициентов b , вычисленный из системы нормальных уравнений (3.2.3), является оценкой вектора β . Поэтому рассмотрим некоторые свойства этой оценки при сделанных ранее допущениях $P1 \div P5$ (см. параграф 3.1).

1. **Вектор b является случайным вектором**, так как линейно зависит от случайного вектора y .

2. **Вектор b является несмещенной оценкой**, т.е.

$$M(b) = \beta, \quad (3.2.8)$$

3. **Ковариационная матрица вектора b определяется выражением**

$$V_b = \sigma^2 (X^T X)^{-1}. \quad (3.2.9)$$

Напомним, что ковариационной матрицей случайного вектора b размерности t называется матрица V_b размера $t \times t$, i, j -элемент которой определяется как

$$[V_b]_{i,j} = \mu_{i,j} = M[(b_i - M(b_i))(b_j - M(b_j))], \quad (3.2.10)$$

где запись $[V_b]_{i,j}$ — означает i,j -ый элемент матрицы V_b , а $\mu_{i,j}$ называют корреляционным моментом проекций b_i и b_j вектора b . Диагональные элементы $\mu_{i,i}$ определяют дисперсию i -ой проекции b_i и в дальнейшем обозначаются как $\sigma_{b_i}^2$. В матричном виде корреляционная матрица определяется выражением

$$V_b = M[(b - M(b))(b - M(b))^T]. \quad (3.2.11)$$

Очевидно, что дисперсия $\sigma_{b_i}^2$ i -го коэффициента b_i определяется выражением

$$\sigma_{b_i}^2 = \sigma^2 \left[(X^T X)^{-1} \right]_{i,i}, \quad (3.2.12)$$

а корреляционный момент $\mu_{i,j}$ для проекций b_i и b_j равен

$$\mu_{i,j} = \sigma^2 \left[(X^T X)^{-1} \right]_{i,j}. \quad (3.2.13)$$

4. Вектор b подчиняется нормальному распределению $N(\beta, \sigma^2(X^T X)^{-1})$. Вектор ε в соответствии с допущением **P4** (см. параграф 3.1) распределен нормально $\varepsilon \sim N(0_n, I)$. Известно, что линейная комбинация нормально распределенных величин также имеет нормальный закон распределения. Следовательно, каждая проекция b_i и весь вектор b подчиняются нормальному распределению. Нормальное многомерное распределение характеризуется двумя величинами: математическим ожиданием $M(b)$ и корреляционной матрицей V_b . Учитывая ранее полученные выражения (3.2.8), (3.2.9), приходим к выводу

$$b \sim N\left(\beta, \sigma^2(X^T X)^{-1}\right), \quad (3.2.14)$$

т.е. вектор b подчиняется нормальному распределению с вектором математического ожидания $M(b) = \beta$ и ковариационной матрицей $V_b = \sigma^2(X^T X)^{-1}$.

5. Вектор b является эффективной оценкой в классе линейных несмещенных оценок. Не останавливаясь на доказательстве этого свойства (см. например, [5, стр. 94 – 95]) заметим, что оценки коэффициентов регрессии, найденных методом наименьших квадратов, обладают наименьшей дисперсией среди всех других линейных несмещенных оценок (свойство эффективности оценки). Другими словами, вектор b имеет наименьшее рассеивание (другими словами – разброс) относительно вектора β по сравнению с любым другим вектором несмещенных оценок.

Оценка дисперсии σ^2 . От дисперсии σ^2 случайной составляющей ε зависит дисперсия $\sigma_{b_j}^2$ коэффициентов регрессии (см. (3.2.12)). Однако на практике в большинстве случаев значение σ^2 неизвестно. Поэтому в расчетах вместо σ^2 используется ее оценка

$$s^2 = \frac{1}{(n-m)} \cdot \sum_{i=1}^n e_i^2, \quad (3.2.15)$$

где m – число оцениваемых параметров (для линейной регрессии (3.1.9) $m = k + 1$); $e_i = y_i - \hat{y}_i$ — невязка i -го измерения. Можно показать, что $M(s^2) = \sigma^2$, т.е. величина s^2 является несмещенной оценкой для дисперсии σ^2 .

Вычисление коэффициентов линейной регрессии в Excel. Рассмотрим вычисление вектора b с использованием обратной матрицы $(X^T X)^{-1}$ по формуле $b = (X^T X)^{-1}(X^T y)$ (см. (3.2.5)). Для реализации этой матричной формулы в Excel необходимо выполнить следующие операции: транспонирование; умножение матриц (частный случай – умножение матрицы на вектор); вычисление обратной матрицы. Все эти операции можно реализовать с помощью следующих матричных функций Excel. Для работы с этими функциями можно или а) обратиться к **Мастеру функций**

и выбрать нужную категорию функций, затем указать имя функции и задать соответствующие диапазоны ячеек, или б) ввести с клавиатуры имя функции задать соответствующие диапазоны ячеек.

Транспонирование матрицы осуществляется с помощью функции ТРАНСП (категория функций – Ссылки и массивы). Обращение к функции имеет вид:

$$\text{ТРАНСП}(\text{диапазон ячеек}), \quad (3.2.16)$$

где параметр *диапазон ячеек* задает все элементы транспонируемой матрицы (или вектора).

Умножение матриц осуществляется с помощью функции МУМНОЖ (категория функций – Математические). Обращение к функции имеет вид:

$$\text{МУМНОЖ}(\text{диапазон}_1; \text{диапазон}_2), \quad (3.2.17)$$

где параметр *диапазон_1* задает элементы первой из перемножаемых матриц, а параметр *диапазон_2* – элементы второй матрицы. При этом перемножаемые матрицы должны иметь соответствующие размеры (если первая матрица $n \times k$, вторая - $k \times m$, то результатом будет матрица $n \times m$).

Обращение матрицы (вычисление обратной матрицы) осуществляется с помощью функции МОБР (категория функций – Математические). Обращение к функции имеет вид:

$$\text{МОБР}(\text{диапазон ячеек}), \quad (3.2.18)$$

где параметр *диапазон ячеек* задает все элементы обращаемой матрицы, которая должна быть квадратной и невырожденной.

Замечание 3.2.1. При использовании этих функций необходимо соблюдать следующий порядок действий:

- выделить фрагмент ячеек, в которые будет занесен результат выполнения матричных функций (при этом надо учитывать размеры исходных матриц);
- ввести арифметическое выражение, содержащее обращение к матричным функциям Excel;

- одновременно нажать клавиши [Ctrl], [Shift], [Enter]. Если этого не сделать, то вычислится только один элемент результирующей матрицы или вектора.

Пример 3.2.3. На рис. 3.1 приведены примеры использования матричных функций Excel. ☉

	A	B	C	D	E	F	G
1			Исходные матрица A и вектор x				
2		2	2	6			
3		5	4	8			5
4	матрица A =	7	3	5	вектор x =	6	
5		5	1	3			7
6		6	5	2			
7	Транспонирование и умножение матриц						
8			=ТРАНСП(B2:D6)				
9							
10		2	5	7	5	6	
11	матрица $A^T =$	2	4	3	1	5	
12		6	8	5	3	2	
13							
14		44	66	50			
15	матрица $A^T A =$	66	105	87			
16		50	87	83			
17							
18		64			=МУМНОЖ(B2:D6;B10:F12)		
19	вектор $A \cdot x =$	105					
20		88			=МУМНОЖ(B2:D6;G3:G5)		

Рис. 3.1. Матричные функции Excel

Пример 3.2.4. Используя исходные данные и условия примера 4.2.1, вычислить вектор коэффициентов $b = |b_0, b_1, b_2|^T$ в Excel.

Решение. Сформируем матрицу X (см. § 4.1) и вектор y (см. рис. 3.2). Затем выполним формирование матрицы $X^T X$, векто-

ра $X^T y$ и вычисление вектора $b = |b_0, b_1, b_2|^T$ по формуле (3.2.5). Все эти вычисления показаны на рис. 3.2. ☉

	A	B	C	D	E	F
1		1	8	5		5
2		1	11	8		10
3		1	12	8		10
4		1	9	5		7
5	$X =$	1	8	7	$y =$	5
6		1	8	8		6
7		1	9	6		6
8		1	9	4		5
9		1	8	5		6
10		1	12	7		8
11						
12		10	94	63		68
13	$X^T \cdot X =$	94	908	603	$X^T y =$	664
14		63	603	417		445
15		=МУМНОЖ(ТРАНСП(B1:D10);B1:D10)				
16		=МУМНОЖ(ТРАНСП(B1:D10);F1:F10)				
17		4,0201	-0,323	-0,14		-3,5393
18	$(X^T \cdot X)^{-1} =$	-0,323	0,054	-0,029	$b =$	0,8539
19		-0,14	-0,029	0,0653		0,3670
20						
21		=МОБР(B12:D14)		=МУМНОЖ(B17:D19;F12:F14)		
22						

Рис. 3.2. Вычисление коэффициентов регрессии

3.3. Интервальные оценки для функции регрессии и ее коэффициентов

Напомним, что при малом объеме выборки из-за большой дисперсии оценок b_j отклонение вычисленных оценок b_j от β_j мо-

жет быть весьма существенным. В этом случае переходят к построению интервальных оценок (доверительных интервалов) для β_j . Однако при этом требуется, чтобы вектор возмущения ε подчинялся нормальному распределению $\varepsilon \sim N(0_n, \sigma^2 I)$. Подробно построение интервальных оценок в случае парной регрессии было рассмотрено в параграфе 2.4. Поэтому здесь ограничимся изложением расчетных соотношений.

Интервальные оценки для коэффициентов β_j . С учетом (3.2.12) оценка $s_{b_j}^2$ дисперсии $\sigma_{b_j}^2$ коэффициента регрессии b_j определяется выражением

$$s_{b_j}^2 = s^2 \left[(X^T X)^{-1} \right]_{j,j}, \quad (3.3.1)$$

где s^2 – несмещенная оценка дисперсии σ^2 (см. (3.2.15)); $\left[(X^T X)^{-1} \right]_{j,j}$ – j -ый диагональный элемент матрицы $(X^T X)^{-1}$.

Среднее квадратическое отклонение коэффициента регрессии b_j определяется как

$$s_{b_j} = s \sqrt{\left[(X^T X)^{-1} \right]_{j,j}} \quad (3.3.2)$$

Так как b_j подчиняются нормальному распределению (см. (3.2.14)), то статистика

$$T_{b_j} = \frac{b_j - \beta_j}{s_{b_j}} \quad (3.3.3)$$

имеет распределение Стьюдента с $n - m$ степенями свободы, где m – число оцениваемых коэффициентов регрессии. Следовательно, интервал

$$[b_j - t(\gamma, n - m) \cdot s_{b_j}, b_j + t(\gamma, n - m) \cdot s_{b_j}] \quad (3.3.4)$$

является интервальной оценкой для коэффициента β_j с надежностью равной γ . Другими словами, с вероятностью γ выполняется неравенство

$$b_j - t(\gamma, n - m) \cdot s_{b_j} \leq \beta_j \leq b_j + t(\gamma, n - m) \cdot s_{b_j}, \quad (3.3.5)$$

где $m = k + 1$ - число коэффициентов регрессии.

Напомним, что значение $t(\gamma, n - m)$ можно определить через функцию Excel выражением (см. (2.4.11)):

$$t(\gamma, n - m) = \text{СТЫЮДРАСПОБР}(1 - \gamma; n - m). \quad (3.3.6)$$

Интервальная оценка для дисперсии σ^2 . Строится аналогично парной регрессии по формуле (2.4.5) с соответствующим изменением числа степеней свободы. Поэтому интервальная оценка для σ^2 с доверительной вероятностью $\gamma = 1 - \alpha$ имеет вид

$$\left[\frac{ns^2}{\chi_{1-\alpha/2; n-m}^2}, \frac{ns^2}{\chi_{\alpha/2; n-m}^2} \right], \quad (3.3.7)$$

где $\chi_{\alpha/2; n-m}^2, \chi_{1-\alpha/2; n-m}^2$ - квантили χ^2 - распределения с $n - m$ степенями свободы уровней $\alpha/2, 1 - \alpha/2$ соответственно. Квантили определяются следующими выражениями:

$$\chi_{\alpha/2; n-m}^2 = \text{ХИ2ОБР}(1 - \alpha/2; n - m), \quad (3.3.8)$$

$$\chi_{1-\alpha/2; n-m}^2 = \text{ХИ2ОБР}(\alpha/2; n - m). \quad (3.3.9)$$

Пример 3.3.1. По коэффициентам b_j , вычисленных в примере 3.2.1, построить интервальные оценки с надежностью 95%. Найти интервальную оценку для дисперсии σ^2 .

Решение. Первоначально определим оценку s^2 , если $\sum e_i^2 = 6.329$: $s^2 = \frac{6/329}{10-3} = 0.904$ и $s = \sqrt{0.904} = 0.951$. Затем вычислим среднеквадратические отклонения коэффициентов b_j , используя элементы $(X^T X)_{i,i}^{-1}$ обратной матрицы $(X^T X)^{-1}$, вычисленные в примере 3.2.4:

$$s_{b_0} = 0.951 \cdot \sqrt{4.0201} = 1.907, \quad s_{b_1} = 0.951 \cdot \sqrt{0.054} = 0.221$$

$$s_{b_2} = 0.951 \cdot \sqrt{0.0653} = 0.243.$$

Находим $t(0.95, 10-3) = \text{СТЫЮДРАСПОБР}(0.05; 10-3) = 2.36$ и вычисляем интервальные оценки надежности 95%:

- для коэффициента β_0

$$[-3.54 - 2.36 \cdot 1.907, -3.54 + 2.36 \cdot 1.907] = [-8.04, 0.096];$$

или с вероятностью 0.95 выполняется неравенство

$$-8.04 \leq \beta_0 \leq 0.096;$$

- для коэффициента β_1

$$[0.854 - 2.36 \cdot 0.221, 0.854 + 2.36 \cdot 0.221] = [0.332, 1.376]$$

или с вероятностью 0.95 выполняется неравенство

$$0.332 \leq \beta_1 \leq 1.376;$$

- для коэффициента β_2

$$[0.367 - 2.36 \cdot 0.243, 0.367 + 2.36 \cdot 0.243] = [-0.206, 0.940]$$

или с вероятностью 0.95 выполняется неравенство

$$-0.206 \leq \beta_2 \leq 0.940.$$

Используя выражения (3.3.8), (3.3.9), вычислим следующие квантили: $\chi_{0.025; 7}^2 = 1.69$; $\chi_{0.975; 7}^2 = 16.01$. Тогда по формуле (3.3.7) получаем интервальную оценку для σ^2 с надежностью 95%

$$\left[\frac{10 \cdot 0.904}{16.01}, \frac{10 \cdot 0.904}{1.69} \right] = [0.565, 5.349]$$

или с вероятностью 0.95 выполняется неравенство

$$0.565 \leq \sigma^2 \leq 5.349. \quad \bullet$$

Интервальная оценка для множественной функции регрессии. Так же как и для парной регрессии, интервальная оценка

для условного математического ожидания $M(Y | x)$ (или для функции регрессии) надежности γ имеет вид

$$[\mathcal{F} - t(\gamma, n - m) \cdot s_{\mathcal{F}}(x), \mathcal{F} + t(\gamma, n - m) \cdot s_{\mathcal{F}}(x)] \quad (3.3.10)$$

или с вероятностью γ выполняется неравенство

$$\mathcal{F} - t(\gamma, n - m) \cdot s_{\mathcal{F}}(x) \leq M(Y | x) \leq \mathcal{F} + t(\gamma, n - m) \cdot s_{\mathcal{F}}(x),$$

где $t(\gamma, n - m)$ определяется выражением (3.3.6). Оценка $s_{\mathcal{F}}(x)$ для среднеквадратического отклонения $\sigma_{\mathcal{F}}(x)$ предсказанного значения \hat{y} определяется выражением

$$s_{\mathcal{F}}(x) = s \cdot \sqrt{x^T (X^T X)^{-1} x}, \quad (3.3.11)$$

где $x = |1, x_1, x_2, \dots, x_k|^T$ – вектор, координаты которого определяют значения объясняющих переменных, при которых вычисляется значение регрессии \hat{y} .

В отличие от парной регрессии, где $s_{\mathcal{F}}(x)$ зависит только от одной объясняющей переменной (см. (2.4.9)) для множественной регрессии оценка $s_{\mathcal{F}}(x)$ зависит уже от вектора x , что существенно усложняет геометрическую интерпретацию интервальной оценки.

Интервальная оценка для индивидуальных значений зависимой переменной. Построенная оценка (3.3.10) определяет интервал возможных значений возможного математического ожидания $M(Y | x)$, но не отдельных возможных значений (названных индивидуальными значениями и обозначаемых y^*) переменной Y , которые отклоняются от $M(Y | x)$.

Интервальная оценка для индивидуальных значений y^* надежности γ имеет вид

$$[\mathcal{F} - t(\gamma, n - m) \cdot s_{y^*}(x), \mathcal{F} + t(\gamma, n - m) \cdot s_{y^*}(x)]. \quad (3.3.12)$$

Оценка $s_{y^*}(x)$ для среднеквадратического отклонения $\sigma_{y^*}(x)$ случайной величины Y определяется выражением

$$s_{y^*}(x) = s \sqrt{1 + x^T (X^T X)^{-1} x} \quad (3.3.13)$$

Появление 1 под знаком корня по сравнению с (3.3.11) объясняется учетом дополнительного отклонения значений y^* от своего математического ожидания $M(Y | x)$.

Пример 3.3.2. По данным примера 3.2.1 найти интервальные оценки для среднего значения ($M(Y | x)$) и индивидуального значения y^* сменной добычи угля на одного рабочего для шахт с мощностью пласта 8 м и уровнем механизации работ 6%.

Решение. В примере 3.2.1 было получено уравнение регрессии $\hat{y} = -3.54 + 0.854 x_1 + 0.367 x_2$. По условию задачи необходимо оценить $M(Y | x)$ при $x = |1 \ 8 \ 6|^T$. Такой оценкой является значение регрессии, вычисленное для заданного вектора x

$$\hat{y} = -3.54 + 0.854 \cdot 8 + 0.367 \cdot 6 = 5.49 \text{ (т)}.$$

Для нахождения $s_{\mathcal{F}}(x)$, $s_{y^*}(x)$ вычислим

$$x^T (X^T X)^{-1} x = |1 \ 8 \ 6| \cdot \frac{1}{3738} \cdot \begin{vmatrix} 15027 & -1209 & -522 \\ -1209 & 201 & -108 \\ -522 & -108 & 244 \end{vmatrix} \cdot \begin{vmatrix} 1 \\ 8 \\ 6 \end{vmatrix} = \\ = \frac{699}{3738} = 0.187.$$

Тогда $s_{\mathcal{F}} = 0.951 \cdot \sqrt{0.187} = 0.411$ (т). Величина $t(0.95, 10 - 3) = 2.36$ и интервальная оценка надежности 95% определяется интервалом

$$[5.49 - 2.36 \cdot 0.411, \ 5.49 + 2.36 \cdot 0.411] = [4.52, 6.46]$$

или с вероятностью 0.95 выполняется неравенство

$$4.52 \leq M(Y | x) \leq 6.46 \text{ (т)}.$$

Построим интервальную оценку для индивидуальных значений y^* переменной Y . Вычислим

$$s_{y^*} = 0.951 \sqrt{1 + 0.187} = 1.036 \text{ (т)}$$

и интервальная оценка определяется интервалом

$$[5.49 - 2.36 \cdot 1.036, \ 5.49 + 2.36 \cdot 1.036] = [3.05, 7.93]$$

или с вероятностью 0.95 выполняется неравенство

$$3.05 \leq y^* \leq 7.93 \text{ (т.)}$$

Напомним, что y^* индивидуальные значения переменной Y при векторе $x = | 1 \ 8 \ 6 |$. Видно, что интервал для y^* «шире» интервала для $M(Y|x)$. *Объясните причину этого.* ☹

3.4. Значимость множественной регрессии и ее коэффициентов

Так же как и для парной регрессии выполним проверку значимости коэффициентов построенного уравнения регрессии и значимости самого уравнения регрессии (см. параграф 2.5).

Проверка статистической значимости коэффициентов регрессии. Для проверки значимости коэффициента b_j сформируем статистические гипотезы:

$$H_0: \beta_j = 0 \text{ (коэффициент } b_j \text{ незначим);}$$

$$H_1: \beta_j \neq 0 \text{ (коэффициент } b_j \text{ значим).}$$

В качестве критерия проверки гипотезы примем следующую случайную величину

$$T_{b_j} = \frac{b_j}{s_{b_j}}, \quad (3.4.1)$$

которая при справедливости гипотезы H_0 имеет распределение Стьюдента с $n - m$ степенями свободы. Следовательно, коэффициент b_j значимо отличается от нуля (т.е. принимается гипотеза H_1) на уровне значимости α , если

$$\left| T_{b_j} \right| > t(1 - \alpha, n - m), \quad (3.4.2)$$

где $t(1 - \alpha, n - m)$ определяется выражением (3.3.6), m – число коэффициентов регрессии.

Пример 3.4.1. Проверить значимость коэффициентов b_1, b_2 регрессии

$$\hat{y} = -3.54 + 0.854 x_1 + 0.367 x_2,$$

построенного по данным примера 3.2.1.

Решение. Вычислим значения критериев (оценки s_{b_1}, s_{b_2} возьмем из примера 4.3.1.): $T_{b_1} = \frac{|0.854|}{0.221} = 3.864, T_{b_2} = \frac{|0.367|}{0.243} = 1.510.$

Критическая точка $t(1 - \alpha, n - m) = t(0.95, 7) = 2.36$. Тогда неравенство (3.4.2) выполняется только для критерия T_{b_1} . Следовательно, можно сделать вывод о значимости только одного коэффициента b_1 (т.е. $\beta_1 > 0$), а коэффициент b_2 является незначимым (т.е. принимается гипотеза $H_0: \beta_2 = 0$). ☹

Проверка статистической значимости уравнения множественной регрессии. Уравнение множественной регрессии *значимо*, если гипотеза

$$H_0: \beta_1 = 0, \beta_2 = 0, \dots, \beta_k = 0$$

о равенстве нулю коэффициентов регрессионной модели отвергается. Как и в случае парной регрессии для проверки значимости вновь рассмотрим сумму (см. параграф 2.5):

$$Q = Q_r + Q_e,$$

где Q – полная сумма квадратов; Q_r – сумма квадратов отклонений, обусловленных регрессией; Q_e – остаточная сумма квадратов. В матричных обозначениях эти суммы вычисляются по формулам:

$$Q = y^T y - n(\bar{y})^2 = \sum_{i=1}^n (y_i - \bar{y})^2; \quad (3.4.3)$$

$$Q_e = y^T y - b^T X^T y = \sum_{i=1}^n (y_i - \hat{y}_i)^2; \quad (3.4.4)$$

$$Q_r = b^T X^T y - n(\bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad (3.4.5)$$

где $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Уравнение множественной регрессии значимо с

уровнем значимости α , если F - статистика

$$F = \frac{Q_r \cdot (n-m)}{Q_e \cdot (m-1)} \quad (3.4.6)$$

удовлетворяет условию

$$F > F_{1-\alpha; m-1; n-m}, \quad (3.4.7)$$

где $F_{1-\alpha; m-1; n-m}$ - квантиль распределения Фишера, значение которого определяется выражением

$$F_{1-\alpha; m-1; n-m} = \text{FRACПОБР}(\alpha; m-1; n-m).$$

В качестве эффективных оценок адекватности уравнения регрессии исходным данным в параграфе 2.5 был рассмотрен коэффициент детерминации R^2 . Для множественной регрессии коэффициент детерминации R^2 (или *множественный коэффициент детерминации*) определяется по формуле

$$R^2 = 1 - \frac{Q_e}{Q} = 1 - \frac{(y - Xb)^T (y - Xb)}{(y - \bar{y})^T (y - \bar{y})} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3.4.8)$$

где \bar{y} - вектор размерности n , составленный из средних значений

$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Напомним, что R^2 характеризует долю вариации

зависимой переменной, обусловленной изменением объясняющих переменных x_1, x_2, \dots, x_k . Следовательно, чем ближе R^2 к единице, тем лучше регрессия соответствует исходным данным.

Задание. Определите, в каком случае $R^2 = 1$?

Иногда используют другую формулу

$$R^2 = \frac{Q_r}{Q} = \frac{b^T X^T y - n \bar{y}^2}{y^T y - n \bar{y}^2}. \quad (3.4.9)$$

Если известен коэффициент детерминации R^2 , то статистику F (3.4.6) можно записать в виде

$$F = \frac{R^2 (n-m)}{(1-R^2)(m-1)}. \quad (3.4.10)$$

Замечание 3.4.1. Для выбора наилучшего уравнения регрессии использование только одного коэффициента детерминации R^2 может оказаться недостаточным. Это обусловлено его увеличением при добавлении новых объясняющих переменных, хотя это и не обязательно означает улучшение качества регрессионной модели. «Чрезмерное» увеличение количества объясняющих переменных приводит к «проникновению» в уравнение регрессии случайного слагаемого ε , которое не должно входить в уравнение. Следовательно, необходимо учитывать не только близость значений регрессии к исходным данным (разница $\hat{y}_i - y_i$), но и «сложность» регрессионной модели, которую можно определить количеством объясняющих переменных.

В соответствии со сделанным замечанием предпочтительнее использовать *скорректированный коэффициент детерминации* \hat{R}^2 (с поправкой на число объясняющих переменных), определяемый по формуле

$$\hat{R}^2 = 1 - \frac{n-1}{n-m} \cdot \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad (3.4.11)$$

где m - число коэффициентов регрессии.

Если известен коэффициент R^2 , то скорректированный коэффициент детерминации можно вычислить по формуле:

$$\hat{R}^2 = 1 - \frac{n-1}{n-m} \cdot (1-R^2). \quad (3.4.12)$$

Видно, что в отличие от R^2 (см. 4.4.8) величина \hat{R}^2 может уменьшаться при увеличении количества объясняющих переменных.

Пример 3.4.2. По данным примера 3.2.1 определить множественный коэффициент детерминации и проверить значимость полученного уравнения регрессии

$$\hat{y} = -3.54 + 0.854 x_1 + 0.367 x_2.$$

Решение. Вычислим следующие величины:

$$b^T X^T y = 489.65; \quad y^T y = \sum_{i=1}^{10} y_i^2 = 496; \quad \bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = 6.8.$$

Теперь по формуле (3.4.9) вычисляем

$$R^2 = \frac{489.65 - 10 \cdot 6.8^2}{496 - 10 \cdot 6.8^2} = 0.811$$

Вычисленное значение 0.811 коэффициента R^2 говорит о том, что вариация переменной Y – добыча угля на одного рабочего на 81.1% объясняется изменением мощности угольного пласта (переменная X_1) и уровнем механизации (переменная X_2).

В примере 2.5.4 был вычислен $R^2 = 0.75$ для регрессии, включающей только одну – мощность угольного пласта. Сравнивая 0.811 и 0.75, можно сказать, что добавление второй объясняющей переменной X_2 незначительно увеличило R^2 . Это понятно, так как в примере 3.4.1 была показана незначимость коэффициента b_2 при переменной X_2 .

По формуле (3.4.12) вычислим скорректированный коэффициент детерминации \mathcal{R}^2 для разного количества объясняющих переменных (величина k):

- если $k = 1$, $m = 2$, то $\mathcal{R}^2 = 1 - \frac{9}{8}(1 - 0.75) = 0.720$;
- если $k = 2$, $m = 3$, то $\mathcal{R}^2 = 1 - \frac{9}{7}(1 - 0.811) = 0.757$.

Хотя скорректированный коэффициент детерминации и увеличился при добавлении объясняющей переменной X_2 , но это еще не говорит о значимости коэффициента b_2 (см. пример 3.4.1, где значение статистики $T_{b_2} = 1.51$ не удовлетворяет условию (3.4.2)).

Зная $R^2 = 0.811$, проверим значимость уравнения регрессии по F -

критерию. Вычисленное по формуле (3.4.10) значение критерия F равно

$$F = \frac{0.811(10-3)}{(1-0.811) \cdot 2} = 15.0$$

Квантиль $F_{0.95; 2; 7} = 4.74$. Неравенство (3.4.7) выполняется и с уровнем значимости $\alpha = 0.05$ можно сделать вывод о значимости построенного уравнения регрессии. Следовательно, исследуемая зависимость Y достаточно хорошо описывается включенными в регрессионную модель переменными X_1 и X_2 . ☺

3.5. Построение линейной множественной регрессии в Excel

Табличный процессор Excel содержит модуль *Анализ данных*. Этот модуль позволяет выполнить статистический анализ выборочных данных (построение гистограмм, вычисление числовых характеристик и т.д.). Режим работы *Регрессия* этого модуля осуществляет вычисление коэффициентов множественной регрессии вида (3.1.9), построение доверительные интервалы и проверку значимости уравнения регрессии.

Для вызова режима *Регрессия* модуля *Анализ данных* необходимо:

- обратиться к пункту меню **Сервис**;
- в появившемся меню выполнить команду *Анализ данных*;
- в списке режимов работы модуля *Анализ данных* выбрать режим *Регрессия* и щелкнуть на кнопке **Ок**.

После вызова режима *Регрессия* на экране появляется диалоговое окно (см. рис. 3.3), в котором задаются следующие параметры:

1. *Входной интервал* Y – вводится диапазон адресов ячеек, содержащих значения y_i (ячейки должны составлять один столбец).

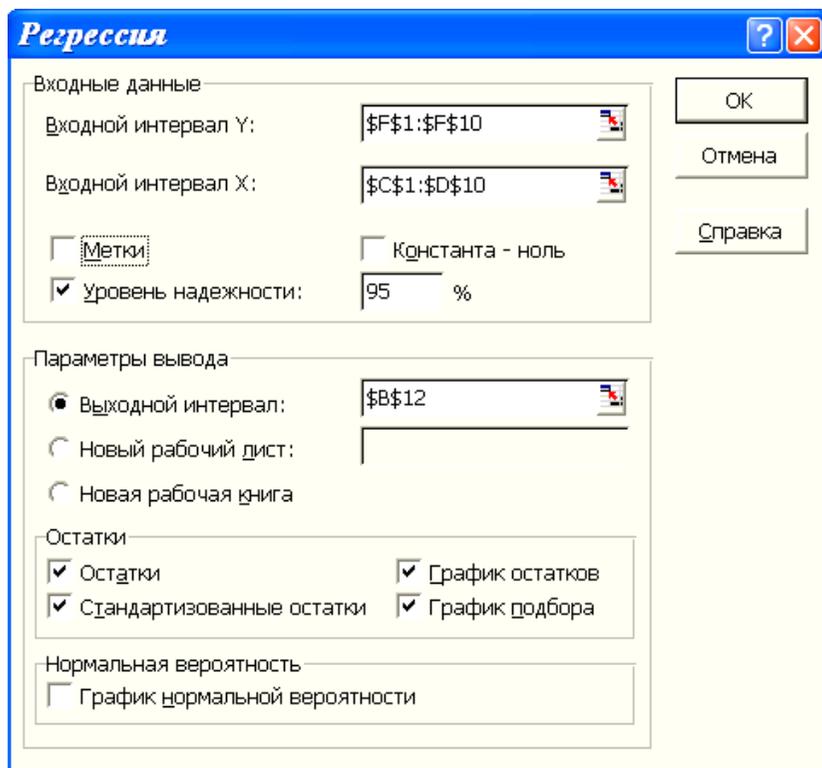


Рис. 3.3. Диалоговое окно режима *Регрессия*

2. *Входной интервал X* – вводится диапазон адресов ячеек, содержащих значения независимых переменных. Значения каждой переменной представляются одним столбцом. Количество переменных не более 16 (т.е. $k \leq 16$).

3. *Метки* – включается если первая строка во входном диапазоне содержит заголовок. В этом случае автоматически будут созданы стандартные названия.

4. *Уровень надежности* – при включении этого параметра задается надежность γ при построении доверительных интервалов.

5. *Константа-ноль* – при включении этого параметра коэффициент $b_0 = 0$.

6. *Выходной интервал* – при включении активизируется поле, в которое необходимо ввести адрес левой верхней ячейки выходного диапазона, который содержит ячейки с результатами вычислений режима *Регрессия*.

7. *Новый рабочий лист* – при включении этого параметра открывается новый лист, в который начиная с ячейки A1 вставляются результаты работы режима *Регрессия*.

8. *Новая рабочая книга* – при включении этого параметра открывается новая книга на первом листе которой начиная с ячейки A1 вставляются результаты работы режима *Регрессия*.

9. *Остатки* – при включении вычисляется столбец, содержащий невязки $y_i - \hat{y}_i, i = 1, \dots, n$.

10. *Стандартизованные остатки* – при включении вычисляется столбец, содержащий стандартизованные остатки.

11. *График остатков* – при включении выводятся точечные графики невязки $y_i - \hat{y}_i, i = 1, \dots, n$, в зависимости от значений переменных $x_j, j = 1, \dots, k$. Количество графиков равно числу k переменных x_j .

12. *График подбора* – при включении выводятся точечные графики предсказанных по построенной регрессии значений \hat{y}_i от значений переменных $x_j, j = 1, \dots, k$. Количество графиков равно числу k переменных x_j .

Пример 3.5.1. По данным таблицы 3.1 используя режим *Регрессия* постройте линейную регрессию.

Решение. Первоначально введем в столбец C десять значений первой переменной, в столбец D – десять значений первой переменной (см. рис. 3.2), а в столбец F – десять значений зависимой переменной.

После этого вызовем режим *Регрессия* и в диалоговом окне зададим необходимые параметры (см. рис. 3.3). Результаты работы приводятся рис. 3.4 – 3.6. Заметим, из-за большой «ширины»

таблиц, в которых выводятся результаты работы режима *Регрессия*, часть результатов помещены в другие ячейки. ●

ВЫВОД ИТОГОВ			
<i>Регрессионная статистика</i>			
Множественный R	0,9009		
R-квадрат	0,8116		
Нормированный R-квадрат	0,7578		
Стандартная ошибка	0,9509		
Наблюдения	10		
<i>Дисперсионный анализ</i>			
	<i>df</i>	<i>SS</i>	<i>MS</i>
Регрессия	2	27,2704	13,635
Остаток	7	6,3296	0,904
Итого	9	33,6000	
		<i>F</i>	<i>Значимость F</i>
		15,0794	0,0029

Рис. 3.4. Результаты работы режима *Регрессия*

Дадим краткую интерпретацию показателям, значения которых вычисляются в режиме *Регрессия*. Первоначально рассмотрим показатели, объединенные названием *Регрессионная статистика* (см. рис. 3.4).

Множественный R - корень квадратный из коэффициента детерминации.

R – квадрат – коэффициент детерминации R^2 .

Нормированный R – квадрат – скорректированный коэффициент детерминации R^2 (см. формулу (3.4.12)).

Стандартная ошибка – оценка s для среднеквадратического отклонения σ .

Наблюдения – число наблюдений n .

Перейдем к показателям, объединенных названием *Дисперсионный анализ* (см. рис. 3.4).

Столбец df - число степеней свободы. Для строки *Регрессия* показатель равен числу независимых переменных $k_r = k = m - 1$; для строки *Остаток* - равен $k_o = n - (k_r + 1) = n - m$; для строки *Итого* – равен $k_r + k_o$.

Столбец SS – сумма квадратов отклонений. Для строки *Регрессия* показатель равен величине Q_r (см. формулу (3.4.5)), т.е.

$$SS_r = Q_r = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2;$$

для строки *Остаток* - равен величине Q_e (см. формулу (3.4.4)), т.е.

$$SS_e = Q_e = \sum_{i=1}^n (\hat{y}_i - y_i)^2;$$

для строки *Итого* – равен $Q = Q_r + Q_e$.

Столбец MS – дисперсии, вычисленные по формуле

$$MS = \frac{SS}{df},$$

т.е. дисперсия на одну степень свободы.

Столбец F – значение F_c , равное F – критерию Фишера, вычисленного по формуле (3.4.6).

Столбец значимость F - значение уровня значимости, соответствующее вычисленной величине F – критерия и равное вероятности $P(F(k_r, k_o) \geq F_c)$, где $F(k_r, k_o)$ - случайная величина, подчиняющаяся распределению Фишера с k_r, k_e степенями сво-

боды. Эту вероятность можно также определить с помощью функции

$$= \text{FRASP}(F_c; k_r; k_e).$$

Если вероятность меньше уровня значимости α (обычно $\alpha = 0.05$), то построенная регрессия является значимой..

Перейдем к следующей группе показателей, объединенных в таблице, показанной на рис. 3.5.

	Коэффициенты	Стандартная ошибка	t-статистика
Y-пересечение	-3,539	1,907	-1,8564
Переменная X 1	0,854	0,221	3,8726
Переменная X 2	0,367	0,243	1,5108
	P-Значение	Нижние 95%	Верхние 95%
	0,1058	-8,0477	0,9690
	0,0061	0,3325	1,3753
	0,1746	-0,2074	0,9415

Рис. 3.5. Продолжение результатов работы режима *Регрессия*

Столбец Коэффициенты – вычисленные значения коэффициентов b_0, b_1, \dots, b_k , расположенных сверху-вниз.

Столбец Стандартная ошибка – значения $s_{b_j}, j = 0, \dots, k$, вычисленные по формуле (3.3.2).

Столбец t-статистика – значения статистик T_{b_j} , вычисленные по формуле (3.4.1).

Столбец P-значение – содержит вероятности случайных событий $P(t(n-m) \geq T_{b_j})$, где $t(n-m)$ – случайная величина, подчиняющаяся распределению Стьюдента с $n-m$ степенями свободы.

Если эта вероятность меньше уровня значимости α , то принимается гипотеза о значимости соответствующего коэффициента регрессии.

Из рис. 3.5 видно, что значимым коэффициентом является только коэффициент b_1 .

Столбцы Нижние 95% и Верхние 95% – соответственно нижние и верхние интервалы для оцениваемых коэффициентов β_j , которые вычисляются по формуле (3.3.4).

Перейдем к следующей группе показателей, объединенных в таблице, показанной на рис. 3.6.

Столбец Наблюдение – содержит номера наблюдений.

Столбец Предсказанное Y – значения \hat{y}_i , вычисленные по построенному уравнению регрессии.

Столбец Остатки – значения невязок $y_i - \hat{y}_i$

ВЫВОД ОСТАТКА			
Наблюдение	Предсказанное Y	Остатки	Стандартные остатки
1	5,127	-0,127	-0,152
2	8,790	1,210	1,443
3	9,644	0,356	0,424
4	5,981	1,019	1,215
5	5,861	-0,861	-1,027
6	6,228	-0,228	-0,272
7	6,348	-0,348	-0,415
8	5,614	-0,614	-0,732
9	5,127	0,873	1,041
10	9,277	-1,277	-1,523

Рис. 3.6. Продолжение результатов работы режима *Регрессия*

В заключении рассмотрения результатов работы режима *Регрессия* приведем график невязок (на рисунке 3.7 невязки на-

званы остатками) $y_i - \hat{y}_i$ при заданных значениях только второй переменной. Наличие чередующихся положительных и отрицательных значений невязок является косвенным признаком *отсутствия систематической ошибки* (неучтенной независимой переменной) в построенном уравнении регрессии.

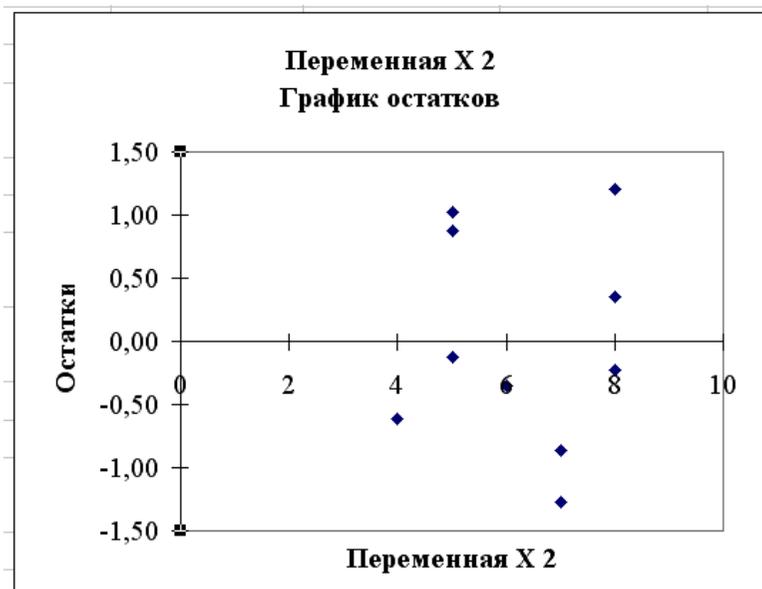


Рис. 3.7. График невязок как функция переменной X_2

3.6. Нелинейные модели множественной регрессии. Производственная функция Кобба-Дугласа

До сих пор мы рассматривали линейные регрессионные модели, в которых переменные имели первую степень. Однако соотношение между социально-экономическими явлениями далеко не всегда можно выразить линейными функциями. Так, например, нелинейными оказываются *производственные функции* (зависимость между объемом произведенной продукции и основными факторами производства), *функции спроса* (зависимость

между спросом на товары или услуги и их ценами или доходом) и другие функции.

Также как и в случае парной нелинейной регрессии (см. § 2.5) можно выделить два вида нелинейности:

- нелинейность по переменным;
- нелинейность по параметрам.

Если **модель нелинейна по переменным**, то введением новых переменных ее можно свести к линейной модели, для оценки параметров которой можно использовать обычный метод наименьших квадратов.

Например, если необходимо оценить коэффициенты *нелинейной регрессионной модели*

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 \sqrt{x_{i2}} + \varepsilon_i, \quad i = 1, \dots, n, \quad (3.6.1)$$

то, вводя новые переменные $Z_1 = X_1^2, Z_2 = \sqrt{X_2}$, получаем *новую линейную модель*

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \varepsilon_i, \quad i = 1, \dots, n, \quad (3.6.2)$$

оценки коэффициентов которой находятся методом наименьших квадратов (см. § 3.2).

К **моделям нелинейным по параметрам** нельзя непосредственно применить линейный МНК. К таким моделям можно отнести следующие модели:

- *степенную модель*

$$y_i = \beta_0 \cdot x_{i1}^{\beta_1} \cdots x_{ik}^{\beta_k} \cdot \varepsilon_i, \quad i = 1, \dots, n; \quad (3.6.3)$$

- *экспоненциальная модель*

$$y_i = e^{\beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik}} \cdot \varepsilon_i, \quad i = 1, \dots, n. \quad (3.6.4)$$

В ряде случаев подбором подходящего преобразования эти модели могут быть приведены к линейной форме. Так для моделей (3.6.3), (3.6.4) таким преобразованием является логарифмирование обеих частей модели. Например, после логарифмирования модель (3.6.3) примет вид:

$$\ln(y_i) = \ln(\beta_0) + \beta_1 \ln(x_{i1}) + \dots + \beta_k \ln(x_{ik}) + \ln(\varepsilon_i), i = 1, \dots, n.$$

Вводя новый параметр $\beta'_0 = \ln(\beta_0)$ и новые переменные $Z_i = \ln(X_i)$, $Y' = \ln(Y)$, приходим к новой линейной модели

$$y'_i = \beta'_0 + \beta_1 z_{i1} + \dots + \beta_k z_{ik} + \ln(\varepsilon_i), i = 1, \dots, n. \quad (3.6.5)$$

Используя МНК, вычисляем оценки b'_0, b_1, \dots, b_k для параметров этой модели. Выполнив обратное преобразование $b_0 = e^{b'_0}$, получаем оценки для коэффициентов нелинейной модели (3.6.3). В качестве примера использования логарифмических преобразований рассмотрим производственную функцию Кобба-Дугласа.

Пример 3.6.1. Производственная функция Кобба-Дугласа имеет вид:

$$Q = A \cdot K^{\beta_1} \cdot L^{\beta_2},$$

где Q – объем производства, K – затраты капитала, затраты труда. Показатели β_1, β_2 являются коэффициентами частной эластичности производства Q соответственно по затратам капитала K и труда L . Это означает, что при увеличении одних только затрат капитала (труда) на 1% объем производства увеличивается на β_1 % (β_2 %).

Учитывая влияние случайных возмущений, получаем нелинейную модель

$$Q = A \cdot K^{\beta_1} \cdot L^{\beta_2} \cdot \varepsilon. \quad (3.6.6)$$

Вычислим оценки для коэффициентов β_1, β_2 по данным табл. 3.3, в которой приведен объем выпуска Q (млн. \$), затраты труда L (чел) и капитала K (млн. \$) в металлургической промышленности. По этим данным необходимо оценить коэффициенты A, β_1, β_2 регрессионной модели (3.6.6).

Решение. Логарифмируя обе части выражение (3.6.6) получаем следующую модель

$$\ln(Q) = \ln(A) + \beta_1 \ln(K) + \beta_2 \ln(L) + \ln(\varepsilon). \quad (3.6.7)$$

Таблица 3.3

Q	657	1200	2427	4257	8095	9849
L	162	245	452	714	1083	1564
K	279	1167	3069	5585	9119	13989

Для удобства дальнейших вычислений переобозначим $Y = \ln(Q)$, $\beta_0 = \ln(A)$, $X_1 = \ln(K)$, $X_2 = \ln(L)$, $\xi = \ln(\varepsilon)$. Тогда имеем линейную регрессионную модель вида

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \xi.$$

Для вычисления оценок b_0, b_1, b_2 используем режим *Регрессия* табличного процессора Excel (см. § 3.5). Результаты работы приведены на рис. 4.8. Вычислены следующие коэффициенты:

$$b_0 = 0.603, b_1 = 1.016, b_2 = 0.127,$$

а само уравнение регрессии примет вид

$$\hat{Y} = 0.603 + 1.016x_1 + 0.127x_2. \quad (3.6.8)$$

Определим $A = e^{b_0} = e^{0.603} = 1.828$, и, возвращаясь к прежним обозначениям запишем выборочное уравнение регрессии для производственной функции Кобба-Дугласа:

$$\hat{Q} = 1.826 \cdot K^{1.016} \cdot L^{0.127}. \quad (3.6.9)$$

На рис. 3.8 приведены также характеристики, вычисляемые в режиме *Регрессия* и определяющие значимость коэффициентов уравнения (3.6.8). Однако ими нельзя воспользоваться по следующей причине.

Напомним, что эффективность оценок, получаемых методом наименьших квадратов, а также проверка значимости коэффициентов регрессии и самого уравнения регрессии основана на допущении о том, что возмущения ε_i не коррелированы между собой и подчиняются нормальному распределению $N(0, \sigma^2)$, т.е. имеет одинаковую дисперсию σ^2 . К сожалению, выполнение

нелинейных преобразований приводит к нарушению этого допущения.

Регрессионная статистика				
Множественный R	0,997			
R-квадрат	0,995			
Нормированный R-квадрат	0,991			
Стандартная ошибка	0,100			
Наблюдения	6,000			
	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение
Y-пересечение	0,603	0,462	1,306	0,283
Переменная X 1	1,016	0,248	4,090	0,026
Переменная X 2	0,127	0,149	0,848	0,459
Дисперсионный анализ				
	F	Значимость F		
	286,146	0,00038		

Рис. 3.8. Вычисление коэффициентов в режиме *Регрессия*

Для иллюстрации этого вернемся к преобразованному уравнению регрессии (3.6.7). Коэффициенты этого уравнения будут являться эффективными оценками, если $\ln(\varepsilon) \sim N(0, \sigma^2)$, т.е. возмущения ε_i должны иметь логарифмически нормальное распределение, что на практике встречается редко.

Минимизация функционала нелинейной множественной регрессии. Рассмотренный режим *Регрессия* применим только для оценивания коэффициентов линейной множественной регрессии.

Возникает вопрос: «Как оценить коэффициенты нелинейной регрессии, которая не приводится к линейной регрессии?». Для этого можно использовать команду *Поиск решения*.

Первоначально формируется функционал метода наименьших квадратов

$$F(b_0, b_1, \dots, b_k) = \sum_{i=1}^n (y_i - \hat{f}_i)^2, \quad (3.6.10)$$

где \hat{f}_i - значение нелинейного уравнения регрессии с коэффициентами b_0, b_1, \dots, b_k при $x = x_i$. Значения этих коэффициентов при которых достигается минимум функционала (3.6.10) принимаются в качестве оценок для $\beta_0, \beta_1, \dots, \beta_k$.

Поиск минимума (т.е. решение задачи безусловной минимизации) можно обратиться к команде *Поиск решения* (пункт меню **Сервис**). При этом возможно задание ограничений на значения искомым коэффициентов (т.е. решается задача условной минимизации). Для иллюстрации этой возможности рассмотрим следующий пример.

Пример 3.6.2. По данным таблицы 3.3 оценить коэффициенты F, β_1, β_2 производственной функции Кобба-Дугласа (рассмотренной в примере 3.6.1) при дополнительном ограничении

$$\beta_1 + \beta_2 = 1 \quad (3.6.11)$$

Решение. Нахождение оценок B, b_1, b_2 для коэффициентов A, β_1, β_2 нелинейной модели будем осуществлять из решения следующей задачи условной минимизации:

$$\min \left[\sum_{i=1}^n \left(Q_i - B \cdot K_i^{b_1} \cdot L_i^{b_2} \right)^2 \right] \quad (3.6.12)$$

при ограничении

$$b_1 + b_2 = 1. \quad (3.6.13)$$

Для решения этой задачи используем команду *Поиск решения*. Первоначально введем в столбцы A, B, C значения

$K_i, L_i, Q_i, i=1, \dots, 6$ (см. рис. 3.9). Затем в ячейках B10, B11, B12 зададим начальные («стартовые») значения искомых коэффициентов: $B=10, b_1=0.5, b_2=0.5$. После этого в соответствующих ячейках столбца D вычислим значения

$$G_i = B \cdot K_i^{b_1} \cdot L_i^{b_2}.$$

в столбце E запрограммируем вычисления значений $(Q_i - G_i)^2$, а в ячейке E10 (выделена цветом) вычислим значения функционала

$$F(B, b_1, b_2) = \sum_{i=1}^6 (Q_i - G_i)^2. \quad (3.6.14)$$

	A	B	C	D	E	F	G
1	Исходные данные						
2	L	K	Q		=B\$10*B3^B\$11*A3^B\$12		
3	162	279	657	425,196	53732,9		
4	245	1167	1200	1069,42	17051		
5	452	3069	2427	2355,58	5100,97		
6	714	5585	4257	3993,84	69253,1		
7	1083	9119	8095	6285,18	3275443		
8	1564	13989	9849	9354,96	244080		
9							
10		2			3664661		
11		0,5					
12		0,5		=СУММ(E3:E8)			
13		1					
14				=B11+B12			

Рис. 3.9. Подготовительные вычисления для решения задачи условной минимизации

После этих подготовительных вычислений обращаемся к команде *Поиск решения* пункт меню **Сервис** и устанавливаем необходимые параметры в уже знакомых диалоговых окнах (см. рис. 3.10).

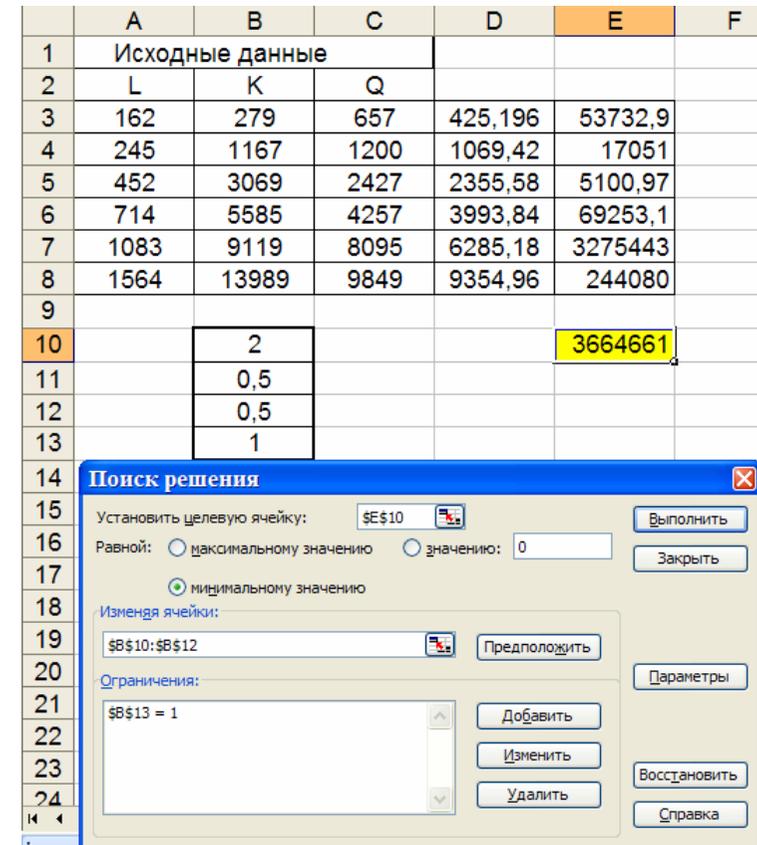


Рис. 3.10. Задание параметров команды Поиск решения

После задания параметров щелкаем на кнопке *Выполнить* и в ячейках B10, B11, B12 выводятся вычисленные значения коэффициентов, а в ячейке E10 – значение функционала (3.6.14) при этих значениях коэффициентов (см. рис. 3.11). Видно, что вычисленные значения коэффициентов $B=3.197, b_1=0.332, b_2=0.668$ удовлетворяют ограничению (3.6.13) и отличаются от коэффициентов, вычисленных в примере 3.6.1 из решения задачи безусловной минимизации.

Задание. Объясните причину отличия коэффициентов нелинейной регрессии, найденных в примере 3.6.1 и 3.6.2.

	A	B	C	D	E
1	Исходные данные				
2	L	K	Q		
3	162	279	657	620,596	1325,28
4	245	1167	1200	1316,29	13522,5
5	452	3069	2427	2732,17	93129
6	714	5585	4257	4523,8	71182
7	1083	9119	8095	7031,87	1130253
8	1564	13989	9849	10361	262103
9					
10		3,197			1571515
11		0,332			
12		0,668			
13		1			

Рис. 3.11. Результаты решения

3.7. Мультиколлинеарность модели множественной регрессии

Серьезной проблемой при построении моделей множественной регрессии на основе метода наименьших квадратов (МНК) является мультиколлинеарность.

Мультиколлинеарность и ее признаки. Одним из условий классической линейной модели является предположение о том, что ранг матрицы X равен числу неизвестных коэффициентов модели, т.е. матрица X – матрица полного ранга. У такой матрицы все столбцы линейно независимы. При нарушении этого условия (т.е. когда один из столбцов матрицы X есть линейная комбинация остальных столбцов) матрица X является вырожденной и, как следствие, вырожденной является матрица $X^T X$. Тогда обратная матрица $(X^T X)^{-1}$ не существует, и в этом случае говорят о *функциональной мультиколлинеарности*. Однако гораздо чаще

приходится сталкиваться с ситуацией, когда матрица X имеет полный ранг (т.е. матрица $(X^T X)^{-1}$ существует), но хотя бы между двумя объясняющими переменными существует тесная корреляционная связь. Такая форма мультиколлинеарности называется *стохастической*, и она будет рассматриваться в дальнейшем.

Мультиколлинеарность модели множественной регрессии – наличие высокой взаимной коррелированности между объясняющими переменными.

Каковы же последствия мультиколлинеарности модели? Приведем самые «неприятные» из них.

- Матрица $X^T X$ хотя и является невырожденной, но величина определителя $|X^T X|$ мала, а, следовательно, элементы обратной матрицы $(X^T X)^{-1}$ становятся очень большими. В результате получаются большие дисперсии $\sigma_{b_j}^2$ (или их оценки $s_{b_j}^2$) коэффициентов b_j .

- Оценки b_j становятся очень чувствительными к незначительному изменению результатов наблюдений и объема выборки. Такая высокая чувствительность характерна для решений плохо обусловленных систем линейных алгебраических уравнений, к которым относится нормальная система уравнений МНК

$$(X^T X)b = X^T y \quad (3.7.1)$$

при наличии мультиколлинеарности в модели.

- Возможно получение неправильных с точки зрения экономической теории значений коэффициентов b_j и даже неверного знака у коэффициентов уравнения регрессии.

- Уменьшаются t -статистики коэффициентов b_j , и оценка их значимости по t -критерию теряет смысл, хотя в целом регрессионная модель может оказаться значимой по F -критерию.

Точных количественных критериев для определения наличия или отсутствия мультиколлинеарности не существует. Тем не ме-

нее, имеются некоторые эвристические подходы по ее выявлению. Кратко остановимся на некоторых из них.

Анализ определителя матрицы парных корреляций. Для оценки мультиколлинеарности факторов (т.е. объясняющих переменных) может использоваться определитель матрицы парных коэффициентов корреляции между факторами. Напомним, что матрицей коэффициентов парных корреляций R_x (или корреляционной матрицей) называют матрицу, i, j элемент которой равен коэффициенту корреляции двух случайных величин X_i, X_j . Очевидно, что диагональные элементы равны 1.

Если факторы не коррелируют между собой, то матрица парных корреляций объясняющих переменных является единичной матрицей размера $k \times k$ и ее определитель равен 1. Если же между объясняющими переменными существует полная линейная зависимость (другой “крайний” случай), то все коэффициенты корреляции равны единице и определитель матрицы R_x равен нулю. Следовательно, чем ближе определитель $\det(R_x)$ матрицы R_x парных корреляций к нулю, тем сильнее мультиколлинеарность факторов и наоборот. Поэтому оценка значимости мультиколлинеарности может быть проведена проверка гипотезы о независимости переменных, т.е.

$$H_0 : \det(R_x) = 1;$$

$$H_1 : \det(R_x) \neq 1.$$

Для проверки этих гипотез определим критерий:

$$K_R = n - 1 - \frac{1}{6}(2k + 5) \lg(\det(R_x)), \quad (3.7.2)$$

где k – число объясняющих переменных, n – количество наблюдений. При справедливости гипотезы H_0 величина K_R подчиняется χ^2 -распределению с $l = \frac{1}{2}n(n-1)$ степенями свободы.

Обозначим через $\chi^2_{1-\alpha, l}$ квантиль χ^2 -распределения с l степе-

нями свободы порядка $1 - \alpha$, вычисляемой с помощью функции Excel $\chi^2_{1-\alpha, l} = \text{ХИ2ОБР}(\alpha; l)$. Если выполняется неравенство

$$K_R > \chi^2_{1-\alpha, l}, \quad (3.7.3)$$

то с уровнем значимости α отвергается гипотеза H_0 и принимается гипотеза о наличии мультиколлинеарности.

Пример 3.7.1. Проверить гипотезу о мультиколлинеарности модели, выборочные данные которой приведены в табл. 3.1 (пример 3.2.1).

Решение. Введем в столбцы А, В значения $x_{i,1}, x_{i,2}$ случайных величин X_1, X_2 (см. рис. 3.12) и вычислим элементы матрицы парных корреляций R_x по формуле:

$$r_{X_l, X_m} = \left[\frac{x_l \cdot x_m - \bar{x}_l \cdot \bar{x}_m}{s_{x_l} \cdot s_{x_m}} \right], \quad l, m = 1, 2.$$

	А	В	С	Д	Е	Ф
1	Исходные данные		Корреляционная матрица			
2	8	5	1	0,488		
3	11	8	0,488	1		
4	12	8				
5	9	5		0,762		
6	8	7				
7	8	8		=МОПРЕД(C2:D3)		
8	9	6		=10-1-1/6*(2*2+5)*LOG10(D5)		
9	9	4	Критерий	9,177		
10	8	5	Граница	61,66		
11	12	7		=ХИ2ОБР(0,05;45)		

Рис. 3.12. Проверка гипотезы о мультиколлинеарности

Для вычисления r_{X_l, X_m} воспользуемся статистической функцией Excel

$$\text{КОРРЕЛ}(\text{диапазон_ячеек_}X_1; \text{диапазон_ячеек_}X_2) \quad (3.7.4)$$

Вычисленные элементы матрицы R_X находятся в ячейках C2:D3 и выделены цветом. Используя математическую функцию Excel:

$$\text{МОПРЕД}(\text{диапазон_ячеек_элементов_матрицы}), \quad (3.7.5)$$

вычисляем $\det(R_X)$ (ячейка D5 – выделена цветом), а затем вычисляем значение 9.177 критерия (3.7.2) при $n=10, k=2$ (ячейка D9). Величина квантиля $\chi^2_{1-\alpha, l}$ равна значению функции ХИ2ОБР(0,05;45)=61.66 (ячейка D10). Неравенство (3.7.3) не выполняется, и поэтому *принимается нулевую гипотезу об отсутствии мультиколлинеарности.* ●

Анализ матрицы парных коэффициентов корреляции объясняющих переменных. По исходным данным $x_{i,j}$, $i=1,2,\dots,n$; $j=1,2,\dots,k$ вычисляют матрицу выборочных парных коэффициентов корреляции и выявляют пары переменных, имеющих высокие коэффициенты корреляции (по модулю больше 0.7 – 0.8). Если такие переменные существуют, то говорят о мультиколлинеарности между ними.

Пример 3.7.2. Выполнить анализ матрицы парных коэффициентов корреляций, вычисленной в примере 3.7.1.

Решение. Матрица парных коэффициентов корреляций имеет вид

$$R_X = \begin{vmatrix} 1.0 & 0.488 \\ 0.488 & 1.0 \end{vmatrix}.$$

Видно, что отсутствуют коэффициенты по модулю превосходящие принятое пороговое значение 0.7. Следовательно, можно говорить об отсутствии мультиколлинеарности. ●

Анализ числа обусловленности матрицы $X^T X$. Числом обусловленности матрицы $X^T X$ называется величина $\text{cond}(X^T X)$, определяемая как (если $(X^T X)^{-1}$ существует):

$$\text{cond}(X^T X) = \|X^T X\| \cdot \|(X^T X)^{-1}\|, \quad (3.7.2)$$

где норма $\|A\|$ матрицы A размера $m \times m$ вычисляемая как

$$\|A\| = \sqrt{\sum_{j=1}^m \sum_{i=1}^m A_{i,j}^2}.$$

Число обусловленности удовлетворяет условию

$$1 \leq \text{cond}(A) < \infty$$

и $\text{cond}(A) = \infty$, если матрица A вырождена. Справедливо следующее неравенство

$$\frac{\|b - \tilde{b}\|}{\|b\|} \leq \text{cond}(X^T X) \frac{\|y - \tilde{y}\|}{\|y\|}, \quad (3.7.5)$$

где вектор \tilde{b} – решение системы $(X^T X)\tilde{b} = X^T \tilde{y}$, вектор b – решение системы $(X^T X)b = X^T y$, а норма вектора $\|y\|$ равна

$$\|y\| = \sqrt{\sum_{i=1}^n y_i^2}.$$

Видно, что чем больше число обусловленной матрицы $X^T X$, тем с большим «коэффициентом усиления» относительная погрешность задания правой части (отношение $\|y - \tilde{y}\|/\|y\|$) передается в относительную погрешность вычисления вектора коэффициентов (отношение $\|b - \tilde{b}\|/\|b\|$).

Значительная величина числа обусловленности матрицы $X^T X$ ($\text{cond}(X^T X) > 10000$) свидетельствует о мультиколлинеарности, а матрицу с таким числом обусловленности называют *плохо обусловленной*.

Пример 3.7.3. Вычислить число обусловленности матрицы $X^T X$ размером 3×3 , сформированной для вычисления коэффициентов уравнения регрессии в примере 3.2.4.

Решение. Введем в столбцы A, B, C элементы матрицы X (см. рис. 3.13) и вычислим матрицы $X^T X, (X^T X)^{-1}$, используя для этого соответствующие матричные функции Excel, описанные в параграфе 3.2 (см. рис. 3.1).

	A	B	C	D	E	F	G
1		1	8	5			
2		1	11	8			
3		1	12	8			
4		1	9	5			
5	X =	1	8	7			
6		1	8	8			
7		1	9	6			
8		1	9	4			
9		1	8	5			
10		1	12	7			
11					=КОРЕНЬ(СУММКВ(B12:D14))		
12		10	94	63			
13	$X^T \cdot X =$	94	908	603		1323,36	
14		63	603	417			
15		=МУМНОЖ(ТРАНСП(B1:D10);B1:D10)					
16							
17		4,02006	-0,3234	-0,1396			
18	$(X^T \cdot X)^{-1} =$	-0,3234	0,05377	-0,0289		4,0519	
19		-0,1396	-0,0289	0,06528			
20					=КОРЕНЬ(СУММКВ(B17:D19))		
21		=МОБР(B12:D14)		=F13*F18			
22							
23			Число обусловленности			5362,132	

Рис. 3.13. Вычисление числа обусловленности

Величина числа обусловленности 5362.132 (ячейка F23) хотя и является достаточно большой, но позволяет сделать вывод об отсутствии мультиколлинеарности.

Методы устранения или уменьшения мультиколлинеарности. Если основная задача моделирования – прогноз будущих значений зависимой переменной, то при достаточно большом коэффициенте детерминации $R^2 > 0.85$ наличие мультиколлинеарности обычно не сказывается на качестве прогноза.

Если же целью исследования является определение степени влияния каждой из объясняющих переменных на зависимую пе-

ренную, то наличие мультиколлинеарности исказит истинные зависимости между переменными.

К сожалению, не существует единого универсального метода устранения мультиколлинеарности. Здесь перечислим только названия наиболее используемых методов (их описание см. [5,6,7]):

- отбор наиболее информативных объясняющих переменных модели;
- преобразование переменных модели;
- вычисление смещенных оценок.

3.8. Гетероскедастичность модели и метод взвешенных наименьших квадратов

Напомним, что существенными условиями классической регрессионной модели являются следующие требования:

1. $M(\varepsilon_i \cdot \varepsilon_j) = 0$, если $i \neq j$.
2. $M(\varepsilon_i \cdot \varepsilon_j) = M(\varepsilon_i^2) = \sigma^2$, если $i = j$.

Первое равенство означает не коррелированность возмущений между собой и это требование для пространственной выборки, как правило, выполняется. Второе требование – равенство дисперсий возмущений ε_i , называемое *условием гомоскедастичности*. На практике в ряде случаев это условие не выполняется и имеет место *гетероскедастичность модели*. Гетероскедастичность «портит хорошие» свойства оценок классической модели и, как правило, ее необходимо «устранить». Поэтому необходимо определить какая ситуация – гомоскедастичность или гетероскедастичность имеет место. Для этого используют специальные статистические тесты, называемые *тестами на гетероскедастичность*. Приведем один из таких тестов.

Тест ранговой корреляции Спирмена. В качестве нулевой гипотезы H_0 гипотезу о гомоскедастичности модели. Идея теста заключается в том, что абсолютные величины невязок e_i являются оценками для σ_i и поэтому в случае гетероскедастичности величины $|e_i|$ и x_i будут коррелированы. Степень коррелиро-

ванности определяется по величине коэффициента ранговой корреляции Спирмена ρ_{xe} .

Для вычисления этого коэффициента необходимо выполнить следующие шаги:

1. Предполагая гомоскедастичность модели, вычислить коэффициенты линейного уравнения регрессии и определить значения невязок e_i .

2. Определить ранг p_{e_i} невязки e_i и ранг p_{x_i} значения x_i . Для вычисления рангов невязок необходимо первоначально упорядочить e_i , например, по возрастанию. Порядковый номер e_i в таком упорядоченном ряду и будет рангом p_{e_i} . Аналогичным образом вычисляются ранги p_{x_i} .

3. Вычислить коэффициента ранговой корреляции по следующей формуле:

$$\rho_{xe} = 1 - \frac{\sum_{i=1}^n d_i}{n^3 - n}, \quad (3.8.1)$$

где $d_i = p_{e_i} - p_{x_i}$ - разность между рангами значений e_i и x_i . Нетрудно заметить, что если ранги $p_{e_i} = p_{x_i}$, то $|\rho_{xe}| = 1$.

После определения ρ_{xe} вычисляется критерий

$$T_\rho = \frac{\rho_{xe} \sqrt{n-2}}{\sqrt{1-\rho_{xe}^2}}. \quad (3.8.2)$$

Если выполняется неравенство

$$|T_\rho| > t(1-\alpha, n-2), \quad (3.8.3)$$

то нулевая гипотеза H_0 отвергается с уровнем значимости α и принимается гипотеза о гетероскедастичности модели. Значение $t(1-\alpha, n-2)$ вычисляется выражением

$$t(1-\alpha, n-2) = \text{СТЮДРАСПОБР}(\alpha; n-2).$$

Заметим, что в приведенном тесте не делается никаких предположений о законе распределений возмущений ε_i и тест следует использовать при $n \geq 10$.

Для вычисления рангов p_{e_i}, p_{x_i} целесообразно использовать функцию РАНГ из категории статистических функций Excel. Обращение к этой функции имеет вид:

РАНГ(число; значения; порядок),

где *число* – значение, для которого определяется ранг;

значения – массив исходных числовых данных (как правило, диапазон ячеек), относительно которого вычисляется ранг заданного числа;

порядок – величина, определяющая способ упорядочивания при вычислении ранга: если *порядок* = 0 или этот параметр опущен в обращении, упорядочивание в порядке убывания; если *порядок* > 0, то упорядочивание в порядке возрастания.

Замечание 3.8.1. Вычисление ранга значений x_i обуславливает наличие в модели только одной независимой переменной. Поэтому **изложенный тест применим только для парной регрессии.**

Пример 3.8.1. Проверить гипотезу о гомоскедастичности данных, представленных в таб. 2.1 и используемые для построения линейной регрессии. Значения коэффициентов уравнения регрессии взять из примера 2.3.1.

Решение. Первоначально введем в ячейки B1, B2 значения коэффициентов b_0, b_1 соответственно (см. рис. 3.14). Затем в ячейках C5 ÷ C14 запрограммируем вычисления значений $\hat{\varepsilon}_i$ регрессионного уравнения $\hat{\varepsilon}(x) = b_0 + b_1x$ при $x = x_i$. После этого запрограммируем вычисления модулей невязок $e_i = y_i - \hat{\varepsilon}_i, i = 1, \dots, 10$. Используя функцию РАНГ, вычисляем ранги p_{x_i}, p_{e_i} (ячейки A17:A26 и ячейки B17:B26 соответственно) и квадраты разностей рангов $d_i^2 = (p_{x_i} - p_{e_i})^2$ (см. рис.

3.14). В ячейке C27 вычисляем $\sum_{i=1}^{10} d_i^2$ (см. рис. 3.15), а в ячейке D28 – значение критерия по формуле (3.8.2), которое равно 0.442. Значение $t(0.95,8) = 2.31$.

	A	B	C	D	E
1	b_0	-2,75			
2	b_1	1,016		=B\$1+B\$2*A5	
3	Исходные данные				
4	x_i	y_i	\hat{y}_i	$ e_i $	
5	8	5	5,378	0,378	
6	11	10	8,426	1,574	
7	12	10	9,442	0,558	
8	9	7	6,394	0,606	
9	8	5	5,378	0,378	
10	8	6	5,378	0,622	
11	9	6	6,394	0,394	
12	9	5	6,394	1,394	
13	8	6	5,378	0,622	
14	12	8	9,442	1,442	
15					
16	ранг x_i	ранг $ e_i $	d_i^2		
17	1	1	0		
18	8	10	4		
19	9	4	25		
20	5	5	0		
21	1	1	0		
22	1	6	25		
23	5	3	4		

Рис. 3.14. Вычисление коэффициента ранговой корреляции

Последним этапом является проверка неравенства (3.8.3). Это неравенство не выполняется, так как $0.442 < 2.31$ и поэтому на уровне значимости $\alpha = 0.05$ принимается нулевая гипотеза о гомоскедастичности модели.

Метод взвешенных наименьших квадратов. Предположим, что неравенство (3.8.3) выполнилось, и была принята гипотеза о гетероскедастичности модели, т.е. каждое возмущение ε_i имеет свою дисперсию σ_i^2 . В этом случае ковариационная матрица вектора возмущения ε остается диагональной, но на главной диагонали стоят не равные элементы и такую матрицу можно записать в виде

$$V_\varepsilon = \text{diag} \{ \sigma_1^2, \sigma_2^2, \dots, \sigma_n^2 \}. \quad (3.8.4)$$

	A	B	C	D	E
24	5	8	9		
25	1	6	25		
26	9	9	0		
27	Сумма	$d_i^2 = 92$			
28			$T_\rho = 0,442$		
29					
30			=1-(6*C27)/(10^3-10)		
31			СТЮДРАСПОБР(0,05;10-2)= 2,31		

Рис. 3.15. Продолжение рис. 3.14

Напомним, что для гомоскедастичной модели ковариационная матрица вектора возмущения ε определяется выражением

$$V_\varepsilon = \text{diag} \{ \sigma^2, \sigma^2, \dots, \sigma^2 \}. \quad (3.8.5)$$

Возникает вопрос: «Какими свойствами будет характеризоваться оценка обыкновенного (классического) МНК вида:

$$b = (X^T X)^{-1} \cdot X^T y, \quad (3.8.6)$$

вычисленная для коэффициентов гетероскедастичной модели?». Дадим ответ на этот вопрос.

1. Оценка (3.8.6) и в этом случае будет *несмещенной и состоятельной*. Это означает, что определение значений зависимой переменной можно осуществлять по уравнению регрессии с коэффициентами (3.8.6).

2. Оценка (3.8.6) *не будет эффективной*, т.е. существуют другие оценки (и соответствующие алгоритмы вычисления коэффициентов уравнения регрессии), *которые имеют меньшую дисперсию*.

3. Изложенный ранее статистический анализ построенной регрессии (точность модели, оценка значимости, построение доверительных интервалов и т.д.) *оказывается неверным для гетероскедастичной модели*.

Возникает традиционный вопрос: «Что делать?». Для ответа на этот вопрос обратимся к функционалу (2.3.3) классического МНК, в котором все квадраты невязок входят с одинаковыми весами, т.к. дисперсия возмущений одинакова. В случае неодинаковых дисперсий квадрат невязки должен «входить» в функционал МНК с весом, обратным величине соответствующей дисперсии. Таким образом, приходим к функционалу

$$F(b) = \sum_{i=1}^n \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} = \sum_{i=1}^n \frac{(e_i)^2}{\sigma_i^2} = (y - Xb)^T \cdot V_\varepsilon^{-1} \cdot (y - Xb), \quad (3.8.7)$$

где матрица V_ε определяется выражением (3.8.4). Можно показать, что вектор b , доставляющий минимум функционалу (3.8.7) является решением следующей матричной системы уравнений:

$$(X^T V_\varepsilon^{-1} X) b = X^T V_\varepsilon^{-1} y \quad (3.8.8)$$

и определяется выражением:

$$b = (X^T V_\varepsilon^{-1} X)^{-1} X^T V_\varepsilon^{-1} y. \quad (3.8.9)$$

Метод вычисления оценок на основе минимизации функционала (3.8.7) получил название *метода взвешенных наименьших квад-*

ратов, а оценка (3.8.9) называется *оценкой метода взвешенных наименьших квадратов*. Эта оценка является *несмещенной, состоятельной и эффективной*. Ковариационная матрица вектора b определяется выражением:

$$V_b = (X^T V_\varepsilon^{-1} X)^{-1}. \quad (3.8.10)$$

Поэтому, используя дисперсию $\sigma_{b_i}^2 = [V_b]_{i,i}$ оценки b_i и соответствующие соотношения параграфов 2.4, 2.5, можно построить доверительные интервалы для коэффициентов β_i , проверить значимость b_i и в случае гетероскедастичности эконометрической модели.

При реализации метода взвешенных наименьших квадратов в Excel будут полезны следующие соотношения:

$$[X^T V_\varepsilon^{-1} X]_{l,k} = \sum_{i=1}^n \frac{[X]_{i,l} \cdot [X]_{i,k}}{\sigma_i^2}, \quad (3.8.11)$$

$$[X^T V_\varepsilon^{-1} y]_l = \sum_{i=1}^n \frac{[X]_{i,l} \cdot [y]_i}{\sigma_i^2}, \quad (3.8.12)$$

где $[X]_{i,l}$ - означает i, l -ый элемент матрицы X . Очевидно, что $X^T V_\varepsilon^{-1} X$ есть матрица размером $m \times m$, $X^T V_\varepsilon^{-1} y$ - вектор из m проекций, m - число коэффициентов регрессионной модели.

Пример 3.8.2. Предположим, что функция регрессии $f(x_1, x_2)$, описывающая зависимость сменной добычи угля от мощности пласта (переменная X_1) и уровня механизации (переменная X_2) имеет вид:

$$f(x_1, x_2) = -3.5 + 0.8 \cdot x_1 + 0.35 \cdot x_2, \quad (3.8.13)$$

что соответствует коэффициентам $\beta_0 = -3.5$; $\beta_1 = 0.8$; $\beta_2 = 0.35$. Пространственная выборка, приведенная в таблице 3.4, соответствует следующей модели измерений:

$$y_i = -3.5 + 0.8 \cdot x_{i1} + 0.35 \cdot x_{i2} + \varepsilon_i, \quad i = 1, 2, \dots, 10, \quad (3.8.14)$$

где x_{i1}, x_{i2} взяты из таблицы 3.1, а ε_i - случайная величина, подчиняющаяся нормальному распределению с нулевым средним и дисперсией

$$\sigma_i^2 = c_\sigma \cdot |f(x_{i1}, x_{i2})|. \quad (3.8.15)$$

Величина c_σ определяет уровень возмущений модели и задавалась равной 0.01, что соответствует относительному уровню возмущений

$$\frac{\max|\varepsilon_i|}{\max|f(x_{i1}, x_{i2})|} = 0.09.$$

Очевидно, что зависимость дисперсий σ_i^2 от значений $f(x_{i1}, x_{i2})$ обуславливает гетероскедастичность модели (3.8.14).

Необходимо вычислить оценки для коэффициентов регрессии, используя метод взвешенных наименьших модулей.

Решение. Исходные данные представлены в таблице 3.5.

Таблица 3.5

	A	B	C	D	E
1	Матрица X			Дисперсия	Вектор y
2	1	8	5	0,22	5,11
3	1	11	8	0,66	8,81
4	1	12	8	0,79	9,71
5	1	9	5	0,3	5,82
6	1	8	7	0,29	4,79
7	1	8	8	0,32	5,74
8	1	9	6	0,34	5,36
9	1	9	4	0,26	5,46
10	1	8	5	0,22	4,57
11	1	12	7	0,73	8,01

Используя матричные функции Excel, формируем матрицы, входящие в систему (3.8.9) и находим вектор коэффициентов b по формуле (3.8.9). Из-за громоздкости этот фрагмент документа

Excel не приводится. Вычисленные коэффициенты: $b_0 = -3.78; b_1 = 0.87; b_2 = 0.27$. Относительная ошибка оценивания $\frac{\max|b - \beta|}{\max|\beta|} = 0.09$. Заметим, что для вектора коэффициентов, вы-

численного с использованием обычного МНК аналогичная ошибка равна 0.132. Сравнивая эти две величины, можно сделать вывод, что использование взвешенного метода наименьших квадратов для гетероскедастичной модели позволяет более точно оценить коэффициенты этой модели. ●

Ранее предлагалось наличие априорной информации о дисперсиях σ_i^2 двух видов:

- заданы значения σ_i^2 ;
- известны аналитические выражения, устанавливающие связь σ_i^2 с объясняющими переменными x_j .

Однако в большинстве случаев такая «подробная» априорная информация отсутствует, однако известно, что существует параметрическая зависимость σ_i^2 от x_j , но параметры этой зависимости не известны.

Для оценки этих параметров используют **двухшаговый метод взвешенных наименьших квадратов**. Кратко опишем этот метод для гетероскедастичной линейной регрессионной модели с двумя объясняющими переменными ($m = 3$):

$$Y = X\beta + \varepsilon,$$

в которой дисперсии σ_i^2 возмущений ε_i зависят от значений x_{ij} как

$$\sigma_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} \quad (3.8.16)$$

где параметры $\alpha_0, \alpha_1, \alpha_2$ - неизвестны.

На первом шаге находим вектор b обычным МНК:

$$b = (X^T X)^{-1} X^T y$$

и вычисляется вектор остатков (невязок)

$$e = y - X^T b.$$

Предполагая, что $D(e_i) = M(e_i^2) = \sigma_i^2$, формируем линейную регрессионную модель, в которой объясненной частью является σ_i^2 , т.е.

$$e_i^2 = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \gamma_i, \quad i = 1, 2, \dots, n, \quad (3.8.17)$$

где случайная величина γ_i имеет нулевое среднее и одинаковую дисперсию. Применяя к этой модели обычный МНК, находим оценки a_0, a_1, a_2 для параметров $\alpha_0, \alpha_1, \alpha_2$.

На втором шаге подставляем оценки a_0, a_1, a_2 в (3.8.16) и вычисляем значения α_i^2 , из которых формируем ковариационную матрицу $\mathcal{K}_\varepsilon = \text{diag}\{\alpha_1^2, \alpha_2^2, \dots, \alpha_n^2\}$ и вычисляем оценку b^* метода взвешенных наименьших квадратов

$$b^* = (X^T \mathcal{K}_\varepsilon^{-1} X)^{-1} X^T \mathcal{K}_\varepsilon^{-1} y. \quad (3.8.18)$$

Метод взвешенных наименьших квадратов для линейной парной регрессии. Рассмотрим следующую модель парной линейной регрессии:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (3.8.19)$$

в которой возмущения ε_i не коррелированы между собой, но имеют разную дисперсию σ_i^2 , т.е. модель (3.8.19) является гетероскедастичной. Можно показать, что в этом случае оценки b_0^* , b_1^* взвешенного МНК для коэффициентов β_0 , β_1 определяются выражениями

$$b_1^* = \frac{\sum_{i=1}^n \frac{1}{\sigma_i^2} \cdot \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2} - \sum_{i=1}^n \frac{x_i}{\sigma_i^2} \cdot \sum_{i=1}^n \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2} \cdot \sum_{i=1}^n \frac{x_i}{\sigma_i^2} - \left(\sum_{i=1}^n \frac{x_i}{\sigma_i^2} \right)^2};$$

$$b_0^* = \frac{\sum_{i=1}^n \frac{y_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} - b_1^* \cdot \frac{\sum_{i=1}^n \frac{x_i}{\sigma_i^2}}{\sum_{i=1}^n \frac{1}{\sigma_i^2}}.$$

Введя обозначения

$$\bar{x} = \sum_{i=1}^n \frac{x_i}{\sigma_i^2}, \quad \bar{y} = \sum_{i=1}^n \frac{y_i}{\sigma_i^2}, \quad \overline{xy} = \sum_{i=1}^n \frac{x_i y_i}{\sigma_i^2},$$

$$\overline{x^2} = \sum_{i=1}^n \frac{x_i^2}{\sigma_i^2}, \quad n_\sigma = \sum_{i=1}^n \frac{1}{\sigma_i^2},$$

приходим к следующим формулам:

$$b_1^* = \frac{n_\sigma \overline{xy} - \bar{x} \cdot \bar{y}}{n_\sigma \overline{x^2} - (\bar{x})^2}; \quad (3.8.20)$$

$$b_0^* = \frac{\bar{y}}{n_\sigma} - b_1^* \frac{\bar{x}}{n_\sigma}, \quad (3.8.21)$$

которые по форме записи совпадают с оценками (2.3.8), (2.3.9) для случая равных дисперсий. Выражения (3.8.20), (3.8.21) легко программируются и вычисление коэффициентов b_0^*, b_1^* не требует обращения матриц.

Если для дисперсий σ_i^2 справедливо выражение

$$\sigma_i^2 = \sigma^2 \cdot x_i^2, \quad (3.8.22)$$

то оценки b_0^*, b_1^* для β_0, β_1 вычисляются по формулам:

$$b_1^* = \frac{\sum_{i=1}^n \frac{y_i}{x_i^2} - \frac{1}{n} \sum_{i=1}^n \frac{1}{x_i} \cdot \sum_{i=1}^n \frac{y_i}{x_i}}{\sum_{i=1}^n \frac{1}{x_i} - \frac{1}{n} \left(\sum_{i=1}^n \frac{1}{x_i} \right)^2}; \quad (3.8.23)$$

$$b_0^* = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{x_i} - \frac{1}{n} b_1^*. \quad (3.8.24)$$

Оценка s^2 для дисперсии σ^2 , входящей в (3.8.22) определяется выражением:

$$s^2 = \frac{1}{n-2} \sum_{i=1}^n \left(\frac{y_i}{x_i} - \frac{b_0^*}{x_i} - b_1^* \right). \quad (3.8.25)$$

ЛАБОРАТОРНАЯ РАБОТА № 3.1

«Построение линейной множественной регрессии»

Цель работы. Используя табличный процессор Excel, построить линейную множественную регрессию, описывающую зависимость себестоимости одной тонны литья (зависимая переменная Y в тыс. руб.) от выработки литья на одного рабочего (объясняющая переменная X_1 в тоннах) и брака литья (объясняющая переменная X_2 в %) и определить значимость построенного уравнения.

Исходные данные. В таблице ЛЗ.1 приведены данные для построения линейной множественной регрессии.

Таблица ЛЗ.1

i	x_{1i}	x_{2i}	y_i
1	14.6	4.2	239
2	13.5	6.7	254
3	21.5	5.5	262
4	17.4	7.7	251
5	44.8	1.2	158
6	111.9	2.2	101
7	20.1	8.4	259
8	28.1	1.4	186
9	22.3	4.2	204
10	25.3	0.9	198
11	56.0	1.3	170

Содержание работы

1. Ввести в лист Excel исходные данные таблицы ЛЗ.1 (см. пример 3.2.1).

2. Используя матричные функции Excel, запрограммируйте вычисление коэффициентов b_0, b_1, b_2 (см. пример 3.2.4).

3. Вычислить стандартизованные коэффициенты регрессии и коэффициенты эластичности (см. пример 3.2.2) и сравнить влияние на зависимую переменную каждой из объясняющих переменных.

4. Проверить значимость построенного уравнения регрессии по критерию Фишера при двух уровнях значимости $\alpha = 0.01$, $\alpha = 0.05$.

5. Вычислить значения коэффициента детерминации R^2 и скорректированного коэффициента детерминации \bar{R}^2 . Высказать мнение, насколько хорошо построенная регрессия определяет зависимость переменной Y от объясняющих переменных X_1, X_2 .

Контрольные результаты:

1. Значения вычисленных коэффициентов

$$b_0 = 213.506, \quad b_1 = -1.170, \quad b_2 = 8.533.$$

Значение критерия Фишера 62.879.

2. Значение коэффициента детерминации $R^2 = 0.940$, скорректированного коэффициента детерминации $\bar{R}^2 = 0.925$.

ЛАБОРАТОРНАЯ РАБОТА № 3.2

«Построение доверительных интервалов для линейной множественной регрессии»

Цель работы. Используя режим *Регрессия* табличного процессора Excel построить доверительные интервалы для коэффициентов $\beta_0, \beta_1, \beta_2$ функции линейной множественной регрессии, описывающей зависимость себестоимости одной тонны литья (зависимая переменная Y в тыс. руб.) от выработки литья на одного рабочего (объясняющая переменная X_1 в тоннах) и брака

литься (объясняющая переменная X_2 в %). Определить значимость коэффициентов b_0, b_1, b_2 .

Исходные данные. В таблице ЛЗ.1 приведены данные для построения линейной множественной регрессии.

Содержание работы

1. Ввести в лист Excel исходные данные таблицы ЛЗ.1 (см. пример 3.2.1).
2. Обратиться к пункту меню *Сервис*, команда *Анализ данных* и включить режим *Регрессия*.
3. В диалоговом окне установить необходимые опции (см. рис. 4.3).
4. Выполнить режим *Регрессия* и проанализировать построенные доверительные интервалы для коэффициентов $\beta_0, \beta_1, \beta_2$.
5. На основе анализа вычисленных t – статистик и P – значения сделать выводы о значимости коэффициентов b_0, b_1, b_2 (см. пример 4.4.1).
6. Построить доверительный интервал для $M(Y|x)$, задав следующие значения объясняющих переменных: $x_1 = 20$; $x_2 = 6$ (см. пример 4.4.2).

Контрольные результаты:

Нижние 95,0%	Верхние 95,0%
185,327	241,685
-1,577	-0,764
4,317	12,749

t -статистика	P -Значение
17,472	1,175E-07
-6,644	1,619E-04
4,667	1,609E-03

КОНТРОЛЬНАЯ РАБОТА № 3.1 Множественная линейная регрессия

По статистическим данным (см. таблицу КЗ.1), описывающим зависимость производительности труда за год в некоторой отрасли производства (переменная Y) от удельного веса рабочих с технической подготовкой (объясняющая переменная X_1) и удельного веса механизированных работ (объясняющая переменная X_2), построить модель множественной линейной регрессии и выполнить статистический анализ построенной модели.

Для вычисления коэффициентов уравнения регрессии

$$\hat{f}(x) = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

и других характеристик множественной регрессии использовать режим *Регрессия* табличного процессора Excel (см. пример 3.6.1).

Требуется:

1. Построить диаграмму рассеяния отдельно по объясняющей переменной X_1 и отдельно по объясняющей переменной X_2 .
2. Используя построенную диаграмму рассеяния, убедиться в наличии линейной зависимости переменной Y от переменной X_1 и от переменной X_2 .
3. Вычислить коэффициенты b_0, b_1, b_2 множественного уравнения регрессии вида

$$\hat{f}(x) = b_0 + b_1 x_1 + b_2 x_2$$

4. Представьте в виде доверительных интервалов для коэффициентов $\beta_0, \beta_1, \beta_2$ значения, приведенные в столбцах *Нижние 95%* и *Верхние 95%* (см. рис. 3.5).
5. Используя вычисленные значения t – статистик (столбец t – статистика рис. 3.5) проверить гипотезы о значимости коэффициентов b_0, b_1, b_2 . Сопоставьте результаты проверки с величинами, приведенными в столбце P – значение (см. рис. 3.5).
Рекомендация: для проверки используйте неравенство (3.4.2).

6. Используя вычисленное значение F – статистики (см. рис. 3.4), проверьте гипотезу о значимости построенного уравнения множественной регрессии. Сопоставьте результат проверки гипотезы с величиной приведенной в ячейке *Значимость F*.

7. Дайте статистическую трактовку вычисленному значению коэффициента детерминации R^2 (см. рис. 3.5).

8. Оформите результаты вычислений отчетом, вставив туда таблицы, сформированные в режиме *Регрессия* (аналогичные тем, что приведены на рис. 3.4, 3.5, 3.6).

Таблица К3.1

№ завода	Удельный вес рабочих с технической подготовкой, %	Удельный вес механизированных работ, %	Производительность труда
1	64 + N	84 + N	4300
2	61 + N	83 + N	4150
3	47 + N	67 + N	3000
4	46 + N	63 + N	3420
5	49 + N	69 + N	3300
6	54 + N	70 + N	3400
7	53 + N	73 + N	3460
8	61 + N	81 + N	4100
9	57 + N	77 + N	3700
10	54 + N	72 + N	3500
11	60 + N	80 + N	4000
12	67 + N	83 + N	4450
13	63 + N	85 + N	4270
14	50 + N	70 + N	3300
15	67 + N	87 + N	4500

где N – последняя цифра в номере зачетной книжки студента.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Чем множественная регрессия отличается от парной?
2. Запишите модель множественной линейной регрессии.
3. Какие условия накладываются на вектор случайных возмущений ε .
4. Запишите функционал метода наименьших квадратов при оценивании коэффициентов множественной линейной регрессии.
5. По статистическим данным (см. таблицу К3.1), описывающим зависимость производительности труда за год в некоторой отрасли производства (переменная Y) от удельного веса рабочих с технической подготовкой (объясняющая переменная X_1) и удельного веса механизированных работ (объясняющая переменная X_2), используя программу Excel, вычислить коэффициенты уравнения регрессии $\hat{f}(x) = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$.

Рекомендация: смотрите пример 3.2.3.

6. Какими свойствами обладают оценки коэффициентов регрессии, вычисленные методом наименьших квадратов?

7. Используя режим *Регрессия* (см. § 3.6), по таблице 3.1 постройте линейную множественную регрессию при предположении $\beta_0 = 0$, т.е. коэффициент уравнения регрессии $b_0 = 0$. Сравните значимость коэффициентов этой регрессии со значимостью коэффициентов примера 3.6.1.

8. Виды нелинейности множественной регрессии?

9. Как преобразовать нелинейную по переменным модель к линейной модели?

10. В чем отличие метода взвешенных наименьших квадратов от обыкновенного (классического) метода наименьших квадратов.

11. Что характеризуют диагональные и недиагональные элементы ковариационной матрицы V_ε вектора возмущений ε ?

12. Запишите функционал метода взвешенных наименьших квадратов. В каком случае этот функционал будет равен нулю?

12. Дана гетероскедастичная модель линейной парной регрессии

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

в которой дисперсии σ_i^2 возмущений ε_i определяются выражением $\sigma_i^2 = \sigma^2 \cdot x_i^2$. Пространственная выборка представлена в следующей таблице ($n=10$)

x_i	2.0	2.4	11.0	8.0	5.6	6.2	4.5	9.8	8.6	3.8
y_i	4.0	5.2	4.5	4.2	4.8	8.0	7.2	12.6	8.5	4.2

По этим данным необходимо:

- построить уравнение регрессии $f(x)$ на основе обычного метода наименьших квадратов;
- построить уравнение регрессии $f^*(x)$ на основе взвешенного метода наименьших квадратов;
- вычислить оценку s^2 для величины σ^2 ;
- на рисунке отобразить исходные данные y_i , значения $f(x_i)$ регрессии, построенной обычным МНК и значения $f^*(x_i)$ регрессии, построенной методом взвешенных наименьших квадратов;
- сравнить графики уравнений регрессий $f(x)$, $f^*(x)$.

Контрольные результаты:

$$b_0 = 3.890, \quad b_1 = 0.392, \quad b_0^* = 3.431, \quad b_1^* = 0.476.$$

ГЛАВА 4. ВРЕМЕННЫЕ РЯДЫ

В этой главе будут рассмотрены методы и алгоритмы решения основных задач анализа временных рядов и их прогнозирования.

4.1 Временные ряды и их числовые характеристики

Напомним (см. пункт 1.2), что временным рядом называется упорядоченная по времени последовательность случайных величин $\{Y(\tau_i)\}, i=1, 2, \dots, n$, где $\tau_i < \tau_{i+1}$. *Временной выборкой* будем называть набор наблюдений $\{y_i\}, i=1, 2, \dots, n$, над случайными величинами $Y(\tau_i)$. Заметим, что иногда в литературе наблюдается подмена понятия «временного ряда» понятием «временной выборки».

Для описания временного ряда достаточно часто используется аддитивная модель вида

$$Y(\tau_i) = q(\tau_i) + \xi(\tau_i), \quad i=1, 2, \dots, n \quad (4.1.1)$$

где детерминированная составляющая $q(\tau_i)$ может включать одну или несколько из следующих компонент: трендовую $t(\tau_i)$, сезонную $s(\tau_i)$ и периодическую $p(\tau_i)$.

Числовые характеристики временного ряда. Из определения временного ряда и модели (4.1.1) следует, что в каждый момент τ_i величина $Y(\tau_i)$ является случайной, подчиняющейся распределению, которое определяется распределением случайной составляющей $\xi(\tau_i)$. Математическое ожидание и дисперсия в момент τ_i определяется выражением:

$$M(Y(\tau_i)) = q(\tau_i); \quad D(Y(\tau_i)) = D(\xi(\tau_i)). \quad (4.1.2)$$

Временной ряд называется стационарным (в широком смысле), если числовые характеристики случайных величин $Y(\tau_i)$ не зависят от времени τ_i , т.е.

$$M(Y(\tau_i)) = q; \quad D(Y(\tau_i)) = \sigma^2. \quad (4.1.3)$$

Для такого временного ряда в качестве оценок величин q , σ^2 используется выборочное среднее \bar{y} и выборочная дисперсия s^2 :

$$\bar{y} = \frac{1}{n} \cdot \sum_{i=1}^n y_i; \quad s^2 = \frac{1}{n} \cdot \sum_{i=1}^n (y_i - \bar{y})^2. \quad (4.1.4)$$

Степень тесноты связи между последовательностями $Y(\tau_1), Y(\tau_2), \dots, Y(\tau_n)$ и $Y(\tau_{1+l}), Y(\tau_{2+l}), \dots, Y(\tau_{n+l})$ (сдвинутых относительно друг друга на l моментов времени, или, как говорят, с лагом l) может быть определена с помощью коэффициента автокорреляции:

$$\rho(l) = \frac{M[(Y(\tau_i) - q)(Y(\tau_{i+l}) - q)]}{\sigma^2}. \quad (4.1.5)$$

В силу стационарности временного ряда $\rho(l)$ зависит только от лага l и для него справедливо следующее равенство

$$\rho(-l) = \rho(l),$$

т.е. достаточно изучать $\rho(l)$ только для положительных лагов l . Если $l=0$, то $\rho(0)=1$. Иногда вместо $\rho(l)$ рассматривают автокорреляционную функцию

$$R(l) = \rho(l) \cdot \sigma^2. \quad (4.1.6)$$

Заметим, что абсолютные значения $\rho(l)$, $R(l)$ убывают с ростом величины l .

Оценкой для $\rho(l)$ является выборочный коэффициент автокорреляции, определяемый по формуле:

$$r(l) = \frac{(n-l) \sum_{i=1}^{n-l} y_i y_{i+l} - (\sum_{i=1}^{n-l} y_i) \cdot (\sum_{i=1}^{n-l} y_{i+l})}{\sqrt{(n-l) \sum_{i=1}^{n-l} y_i^2 - (\sum_{i=1}^{n-l} y_i)^2} \cdot \sqrt{(n-l) \sum_{i=1}^{n-l} y_{i+l}^2 - (\sum_{i=1}^{n-l} y_{i+l})^2}}. \quad (4.1.7)$$

Заметим, что с увеличением l число пар наблюдений y_i, y_{i+l} уменьшается и поэтому число l не должно быть сравнительно большим (рекомендуют $l \leq n/4$). В отличие от $\rho(l)$ для оценки $r(l)$ может быть нарушено свойство монотонного убывания абсолютных значений.

Стационарный временной ряд, у которого математическое ожидание равно 0, а величины $\xi(\tau_i)$ не коррелированы, часто называют “белым шумом”. Очевидно, что для белого шума

$$\rho(l) = \begin{cases} 1, & \text{если } l = 0; \\ 0, & \text{если } l \neq 0. \end{cases}$$

Пример 4.1.1. В столбце А документа Excel (приведенного на рисунке 4.1) представлены 20 значений стационарного временного ряда, являющегося белым шумом. Необходимо вычислить выборочное математическое ожидание, дисперсию и коэффициент автокорреляции $\rho(l)$, $l = 0, 1, 2, 3$.

Решение. Первые две оценки вычисляются по формулам (4.1.4) с использованием стандартных функций Excel (обращение к ним показано на рис. 4.1), а выборочный коэффициент автокорреляции – по формуле (4.1.7), при этом используются предварительно вычисленные суммы: $\sum_{i=1}^{n-l} y_i y_{i+l}$; $\sum_{i=1}^{n-l} y_i$; $\sum_{i=1}^{n-l} y_{i+l}$; $\sum_{i=1}^{n-l} y_i^2$; $\sum_{i=1}^{n-l} y_{i+l}^2$ (см рис. 4.1). Полученные значения оценок приведены в таблице 4.1 (вторая строка). Третья строка таблицы содержит точные значения искомым характеристик. Различие между оценками и точными значениями обусловлены малым объемом выборки. ●

	A	B	C	D	E	F	G
1							
2	27,65		28,53		=СРЗНАЧ(A2:A21)		
3	32,07						
4	33,87		11,994		=ДИСП(A2:A21)		
5	24,83						
6	34,58						
7	24,59		$l=0$	$l=1$	$l=2$	$l=3$	
8	25,00	$\sum y_i$	570,60	541,01	515,38	488,60	
9	25,00	$\sum y_{i+l}$	570,60	542,95	510,89	477,01	
10	26,29	$\sum y_i^2$	16507,4	15631,6	14974,7	14257,4	
11	26,03	$\sum y_{i+l}^2$	16507,4	15742,7	14714,5	13567	
12	33,65	$\sum y_i \cdot y_{i+l}$	16507,4	15417,3	14658,4	13730,5	
13	25,38						
14	28,82	$r(l)$	1	-0,18911	0,14134	0,10467	
15	33,61						
16	27,85						
17	31,46	=(18*E12-E8*E9)/(КОРЕНЬ(18*E10-E8^2)*КОРЕНЬ(18*E11-E9^2))					
18	27,89						
19	26,78						
20	25,63						
21	29,59						

Рис. 4.1. Вычисление числовых характеристик

Таблица 4.1

Характеристики	$M(Y)$	$D(Y)$	$\rho(0)$	$\rho(1)$	$\rho(2)$	$\rho(3)$
Оценка	31.8	8.9	1.0	-0.19	0.14	0.10
Точное значение	30	10	1	0	0	0

Тест стационарности временного ряда. Для стационарности временного ряда достаточно постоянства его числовых характеристик на всем интервале определения временного ряда. Наиболее часто в качестве таких характеристик берут математическое

ожидание и дисперсию. Тогда ответ на вопрос стационарности дискретного временного ряда сводится к проверке следующей пары статистических гипотез:

$$\left. \begin{aligned} H_0 : M(Y(\tau_i)) = const; \\ H_1 : M(Y(\tau_i)) \neq const. \end{aligned} \right\} \begin{array}{l} \text{Постоянство} \\ \text{математического ожидания} \end{array}$$

$$\left. \begin{aligned} H_0 : D(Y(\tau_i)) = const; \\ H_1 : D(Y(\tau_i)) \neq const. \end{aligned} \right\} \begin{array}{l} \text{Постоянство} \\ \text{дисперсии} \end{array}$$

Для проверки этих гипотез используют различные критерии [5,6]. Здесь мы ограничимся только одним (достаточно простым) критерием.

Временной ряд $Y(\tau_i), i=1, 2, \dots, n$, разбивается на две части (не обязательно одинаковые) по количеству содержащихся в них значений $y_i = Y(\tau_i)$. Пусть первая часть (обозначим ее $Y^{(I)}$) содержит n_I наблюдений $Y(\tau_i), i=1, 2, \dots, n_I$, а вторая часть - $Y^{(II)}$ содержит n_{II} наблюдений $Y(\tau_i), i=n_I+1, \dots, n_I+n_{II}$.

Для каждой части временного ряда вычислим (используя формулы (4.1.4)) выборочное среднее \bar{y}_I, \bar{y}_{II} и выборочные дисперсии s_I^2, s_{II}^2 . Далее рассчитаем значение критерия

$$K_S = \frac{|\bar{y}_I - \bar{y}_{II}|}{\sqrt{\frac{s_I^2}{n_I} + \frac{s_{II}^2}{n_{II}}}} \quad (4.1.7)$$

(часто называемого критерием Стьюдента). Если выполняется неравенство

$$K_S > t(1-\alpha, n_I + n_{II} - 2), \quad (4.1.8)$$

то гипотеза о постоянстве математического ожидания отклоняется с уровнем значимости α . Напомним, что значение $t(1-\alpha, n_I + n_{II} - 2)$ вычисляется с использованием следующей функции Excel:

$$t(1-\alpha, n_I + n_{II} - 2) = \text{СТЪЮДРАСПОБР}(\alpha, n_I + n_{II} - 2)$$

Для проверки гипотезы о постоянстве дисперсии определим следующий критерий

$$F_S = \frac{s_I^2}{s_{II}^2}, \quad (4.1.9)$$

где s_I^2, s_{II}^2 - оценки дисперсии, вычисленные по первой (число измерений n_I) и второй (число измерений n_{II}) части временного ряда.

Если **не выполняется неравенство**

$$F_{\frac{\alpha}{2}; n_I - 1; n_{II} - 1} \leq F_S \leq F_{1 - \frac{\alpha}{2}; n_I - 1; n_{II} - 1}, \quad (4.1.10)$$

то гипотеза о постоянстве дисперсии отвергается с уровнем значимости α . Границы критической области вычисляются с помощью следующей функции Excel:

$$F_{\frac{\alpha}{2}; n_I - 1; n_{II} - 1} = \text{FRАСПОБР}(1 - \frac{\alpha}{2}; n_I - 1; n_{II} - 1). \quad (4.1.11)$$

Пример 4.1.2. Осуществить тестирование временного ряда приведенного столбце А на рис. 4.2 на стационарность.

Решение. Разобьем исходный временной ряд на две части по 10 измерений в каждой. Вычислим по каждой из этих частей выборочные оценки (см. рис. 4.2):

$$\bar{y}_I = 30.68, \quad \bar{y}_{II} = 30.14, \quad s_I^2 = 10.19; \quad s_{II}^2 = 8.16.$$

Затем определим значения критериев (4.1.7) и (4.1.9) (см. рис. 4.2): $K_S = 0.40$; $F_S = 1.249$. Проверим выполнение неравенств (4.1.8) и (4.1.10). Неравенство (4.1.8) не выполняется, так как $0.40 < 2.101$, а неравенство (4.1.10) выполняется - $0.248 < 1.249 < 4.026$.

Следовательно, можно сделать вывод о стационарности рассматриваемого временного ряда. ●

	A	B	C	D	E	F	G
1							
2	27,17						
3	28,18						
4	32,21						
5	26,29						
6	33,84		\bar{y}_I	30,68			
7	34,61		s_I^2	10,19			
8	29,33		=СТЪЮДРАСПОБР(0,05;18)				
9	28,14				2,101		
10	34,44		Критерий K			0,40	
11	32,61						
12	27,10		Критерий F			1,249	
13	32,47		=FRАСПОБР(0,975;9;9)				=FRАСПОБР(0,025;9;9)
14	32,58				0,248	4,026	
15	27,75						
16	30,28		\bar{y}_{II}	30,14			
17	30,21		s_{II}^2	8,16			
18	33,97						
19	26,10						
20	33,27						
21	27,64						

Рис. 4.2. Проверка гипотезы о стационарности ряда

4.2 Выделение трендовой составляющей временного ряда.

Трендовая составляющая $t(\tau)$ отражает влияние долговременных факторов и соответствует устойчивой и долговременной тенденции изменения временного ряда. Знание трендовой составляющей позволяет осуществлять долговременное прогнозирование. Поэтому возникает задача выделения тренда, т.е. построение оценки $\hat{t}(\tau)$ для функции $t(\tau)$ (или оценок $\hat{t}(\tau_i)$ для значений $t(\tau_i)$) по заданной временной выборке $\{\tau_i, y_i\}$. При этом предполагается, что остальные составляющие $s(\tau), p(\tau)$ временного

ряда отсутствуют. Для решения этой задачи возможно несколько подходов.

Выделение тренда методами парной регрессии. Временной ряд представляется моделью

$$Y(\tau_i) = t(\tau_i) + \varepsilon(\tau_i), \quad (4.2.1)$$

где случайная составляющая $\varepsilon(\tau)$ в моменты $\tau_i, i=1, 2, \dots, n$, удовлетворяет условиям Гаусса-Маркова, а τ рассматривается как независимая переменная (см. главу 2). Функцию $t(\tau)$ можно оценить, используя методы парной регрессии, изложенные в главе 2. Поэтому здесь рассмотрим только некоторые особенности применения этих методов, обусловленные решаемой задачей.

Одна из особенностей заключается в том, что различный характер тренда (иногда достаточно сложный) обуславливает более широкое использование нелинейных функций. Так наряду с линейной функцией $t(\tau) = \beta_0 + \beta_1\tau$ гораздо чаще используется следующие нелинейные функции:

- Полиномиальная

$$t(\tau) = \beta_0 + \beta_1\tau + \dots + \beta_p\tau^p, \quad (4.2.3)$$

где p - степень полинома (при $p=1$ получаем линейную функцию);

- Экспоненциальная

$$t(\tau) = e^{\beta_0 + \beta_1\tau}, \quad (4.2.4)$$

- Логистическая

$$t(\tau) = \frac{\beta_0}{1 + \beta_1 e^{-\beta_2\tau}}, \quad (4.2.5)$$

Выбор вида функции $t(\tau)$ часто основывается на анализе графического изображения ряда, т.е. на анализе диаграммы рассеяния, построенной по точкам $\{\tau_i, y_i\}$ (см. параграф 2.1).

При применении полиномиальной функции важно правильно определить степень полинома. Для этого можно использовать ме-

тод последовательных разностей, заключающийся в вычислении разностей:

- первого порядка

$$\Delta_i = \Delta_i - \Delta_{i-1}, \quad i = 1, 2, \dots, n-1;$$

- второго порядка

$$\Delta_i^2 = \Delta_i - \Delta_{i-1}, \quad i = 1, 2, \dots, n-2;$$

- k - ого порядка

$$\Delta_i^k = \Delta_i^{k-1} - \Delta_{i-1}^{k-1}, \quad i = 1, 2, \dots, n-k,$$

а также величин

$$d^{(k)} = \frac{1}{n-k} \cdot \frac{\sum_{i=1}^{n-k} (\Delta_i^{(k)})^2}{c_{2k}^k}, \quad (4.2.6)$$

где c_{2k}^k - означает сочетание, определяемое по формуле $c_{2k}^k = \frac{(2k)!}{(k!)^2}$. Величина $d^{(k)}$ первоначально убывает с ростом k , а

затем, начиная с некоторого значения k_0 стабилизируется, оставаясь приблизительно на одном уровне при дальнейшем росте k . Тогда степень полинома определяется по формуле $p = k_0 - 1$.

После выбора вида функции $t(\tau)$ строят уравнение регрессии $\hat{f}(\tau)$, зависящее от коэффициентов b_0, b_1, \dots, b_k , которые являются оценками для коэффициентов $\beta_0, \beta_1, \dots, \beta_k$ функции тренда $t(\tau)$. Так для полиномиального тренда (4.2.1) уравнение регрессии примет вид:

$$\hat{f}(\tau) = b_0 + b_1\tau + \dots + b_p\tau^p \quad (4.2.7)$$

Для вычисления коэффициентов b_0, b_1, \dots, b_k используется метод наименьших квадратов, т.е. коэффициенты находятся из условия минимума функционала

$$F(b_0, b_1, \dots, b_k) = \sum_{i=1}^n (t_i - \hat{f}(\tau_i))^2, \quad (4.2.8)$$

где $\hat{f}(\tau_i)$ - значение уравнения регрессии в точке $\tau = \tau_i$.

Использование нелинейных функций $t(\tau)$ обуславливает следующие виды нелинейности уравнения регрессии: нелинейность по переменной и нелинейность по коэффициентам, рассмотренные в параграфе 2.6. Напомним, что в этих случаях используются два подхода для вычисления коэффициентов регрессии:

а) путем замены переменной или нелинейными преобразованиями осуществляется линеаризация уравнения регрессии, к которому применяется метод наименьших квадратов;

б) непосредственное вычисление коэффициентов из условий минимума функционала (4.2.8).

Первый подход рассмотрен в параграфе 2.6, а второй – в параграфе 2.7 и он является более универсальным, так как позволяет при вычислении коэффициентов учесть априорные ограничения (в общем случае, нелинейные) на искомые коэффициенты (например, $\beta_i \geq 0, i = 0, 1, 2, \dots, k$).

После вычисления коэффициентов b_0, b_1, \dots, b_k , уравнение регрессии принимается в качестве оценки для функции тренда $t(\tau)$ и может быть использовано для дальнейшего анализа временного ряда или его прогнозирования.

Пример 4.2.1. В таблице 4.1 приведены данные, отражающие спрос (в условных единицах) на некоторый товар за восьмилетний период.

Таблица 4.1.

Год	1	2	3	4	5	6	7	8
Спрос	213	171	291	309	317	362	351	361

По этим данным (которые являются временной выборкой) найти оценку $\hat{t}(\tau)$, предполагая, что $t(\tau)$ является квадратичной функцией. Выполнить прогноз временного ряда для десятого года.

Решение. При сделанном предположении оценка $\hat{t}(\tau)$ имеет вид

$$\hat{t}(\tau) = b_0 + b_1\tau + b_2\tau^2 \quad (4.2.9)$$

и это уравнение регрессии нелинейно по переменным. Для перехода к линейному уравнению регрессии введем новые переменные $x_1 = \tau; x_2 = \tau^2$ и получим множественную линейную регрессию:

$$\hat{t}(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$$

вектор коэффициентов $b = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix}$ находим методом наименьших

квадратов, решая уже известную систему нормальных уравнений (см. параграф 3.2):

$$(X^T X) b = X^T y,$$

где X - матрица размером 8×3 , а y - вектор наблюдений. Формирование матрицы X и решение системы показано на рисунке 4.3. Вычисленный вектор коэффициентов (ячейки F16 – F18 выделенные цветом) имеет следующие проекции $b = \begin{pmatrix} 132.3 \\ 55.09 \\ -3.26 \end{pmatrix}$.

Возвращаясь к уравнению (4.2.9), получаем следующую оценку для тренда временного ряда:

$$\hat{t}(\tau) = 132.3 + 55.09 \cdot \tau - 3.26 \cdot \tau^2, \quad (4.2.10)$$

На рисунке 4.4 показана временная выборка $y_i, i = 1, 2, \dots, 8$ (кривая 1, маркированная квадратиками) и график функции $\hat{t}(\tau)$ (кривая 2 – маркированная ромбами). Для выполнения прогноза достаточно в (4.2.10) подставить $\tau = 10$. Получаем значение $\hat{t}(10) = 356.41$. ●

Выделение тренда методами фильтрации. Строится оценка \hat{t}_j для значения функции $t(\tau)$ в точке τ_j как взвешенное

среднее тех исходных значений y_i , которые находятся в некоторой близости от точки τ_j , т.е.

$$\hat{\epsilon}_j = \sum_{l=-L}^L c_l y_{j+l}, \quad j = L+1, L+2, \dots, n-L, \quad (4.2.10)$$

где c_l - весовые множители, удовлетворяющие условию

$$\sum_{l=-L}^L c_l = 1$$

	A	B	C	D	E	F
1	Год	Спрос	Матрица X			\hat{t}
2	1	213	1	1	1	184,13
3	2	171	1	2	4	229,41
4	3	291	1	3	9	268,16
5	4	309	1	4	16	300,38
6	5	317	1	5	25	326,05
7	6	362	1	6	36	345,20
8	7	351	1	7	49	357,80
9	8	361	1	8	64	363,88
10	9					363,41
11	10					356,41
12		8	36	204		2375
13	$X^T \cdot X =$	36	204	1296	$X^T \cdot y =$	11766
14		204	1296	8772		69720
15						
16		1,9464	-0,9107	0,0893		132,3
17	$(X^T \cdot X)^{-1} =$	-0,9107	0,506	-0,0536	b=	55,089
18		0,0893	-0,0536	0,006		-3,2679

Рис. 4.3. Вычисление коэффициентов квадратичного тренда

Видно, что суммируются L значений, находящихся левее точки τ , L значений правее точки τ_j и само значение y_j . Длина интервала суммирования равна $(2L+1)$ точки и этот интервал “скользит” по исходным данным. Наиболее часто используют ме-

тод скользящего среднего, в котором множители задаются выражением:

$$c_l = \frac{1}{2L+1}, \quad l = -L, \dots, 0, \dots, L.$$

Так, если $L=1$, то $c_{-1} = c_0 = c_1 = 1/3$, а сам метод скользящего среднего примет вид:

$$\hat{\epsilon}_j = \frac{1}{3}(y_{j-1} + y_j + y_{j+1}). \quad (4.2.11)$$

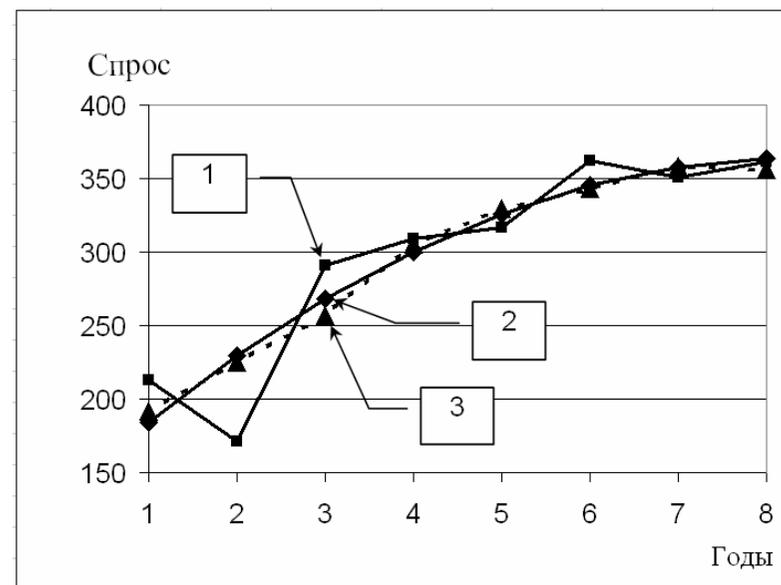


Рис. 4.4. Графики временной выборки и оценок тренда

Очевидно, что чем больше величина L , тем меньше уровень “остаточных” возмущений в оценке \hat{t}_j . Действительно, если $M(\epsilon(\tau_j)) = \sigma^2$, то после алгоритма (4.2.11) дисперсия оценки \hat{t}_j

будет равна $\sigma^2 / (2L + 1) = \sigma^2 / 3$. Однако, следует помнить, что при возрастании L увеличивается систематическая ошибка.

Систематическая ошибка будет мала, если графическое изображение временного ряда напоминает прямую линию. Если же тренд имеет явно нелинейный характер, то фильтр скользящего среднего может привести к значительным искажениям (т.е. к большой систематической ошибке). В таких случаях предпочтительнее использовать метод экспоненциального сглаживания, описанный ниже.

К сожалению, выражение (4.2.10) не определяет «отфильтрованные» значения в первых L точках и последних L точках временного ряда. В этих случаях можно изменить алгоритм (4.2.10), используя под знаком суммирования только известные y_i . На-

пример, в точке τ_1 алгоритм (4.2.11) имеет вид $\hat{t}_1 = \frac{1}{2}(y_1 + y_2)$, а

в точке τ_n $\hat{t}_n = \frac{1}{2}(y_n + y_{n-1})$. Рассмотренные методы фильтрации оценивают значение тренда только в точках $\tau_i, i = 1, 2, \dots, n$, что не позволяет непосредственно использовать этот метод для прогнозирования временного ряда.

Пример 4.2.2. По данным таблицы 4.1 вычислить значение тренда в точках $\tau_j, j = 1, 2, \dots, 8$, используя алгоритм (4.2.11).

Решение. Фрагмент документа Excel, вычисляющий значения \hat{t}_j по формуле (4.2.11) приведен на рисунке 4.5, сами значения нанесены на рис. 4.4 (кривая 3, маркированная треугольниками) Сравнивая эти значения со значением оценки $\hat{t}(\tau)$, построенной в примере 2.1, видим некоторые отличия, которые можно объяснить использованием разных методов для выделения тренда временного ряда. ☺

Выделение тренда методом экспоненциального сглаживания. В отличие от метода скользящего среднего в определении экспоненциальной средней участвуют все наблюдения исходного временного ряда, но с разными весовыми множителями. Алго-

ритм метода экспоненциального сглаживания определяется выражением:

$$\xi_j = (1 - \alpha)\xi_{j-1} + \alpha y_j, \quad j = 1, 2, \dots, n, \quad (4.2.12)$$

где α - коэффициент экспоненциального сглаживания, задаваемый как $0 < \alpha < 1$.

	A	B	C	D	E	F
1	Год	Спрос	\hat{t}			
2	1	213	192	←	=(B2+B3)/2	
3	2	171	225			
4	3	291	257			
5	4	309	305,67	←	=(B4+B5+B6)/3	
6	5	317	329,33			
7	6	362	343,33			
8	7	351	358			
9	8	361	356	←	=(B8+B9)/2	

Рис. 4.5. Выделение тренда методом скользящего среднего

Можно доказать справедливость выражения:

$$\xi_j = \alpha \cdot \sum_{i=0}^{j-1} (1 - \alpha)^i y_{j-i}, \quad (4.2.13)$$

из которого следует, что каждое «старое» измерение y_i входит в оценку $\xi_j (i \leq j)$ с весом $\alpha(1 - \alpha)^i$, т.е. по мере удаления от точки τ_j вес измерения y_i уменьшается. В качестве начального значения \hat{t}_0 может быть принято среднее арифметическое всей временной выборки или только ее части.

Значения \hat{t}_j можно рассматривать как прогнозное значение тренда в момент τ_j и его можно представить как:

$$\hat{\epsilon}_j = \hat{\epsilon}_{j-1} + \alpha(y_j - \hat{\epsilon}_{j-1}), \quad (4.2.14)$$

Из этого выражения видно, что прогноз в момент τ_j состоит из двух слагаемых: прогнозного значения $\hat{\epsilon}_{j-1}$ в предыдущий момент и невязки (ошибки) прогнозирования $y_j - \hat{\epsilon}_{j-1}$, взятой с весом α .

Из выражения (4.2.13) видно, что уменьшая величину α увеличивается степень сглаживания (за счет увеличения числа “значимых” слагаемых). Рекомендуется α определять по формуле

$$\alpha = \frac{2}{n+1}. \quad (4.2.15)$$

Выделение трендовой составляющей в табличном процессоре Excel. Табличный процессор Excel позволяет реализовать все рассмотренные выше методы выделения трендовой составляющей с использованием следующих операций:

- команды *Добавить линию тренда*;
- режима работы *Скользящее среднее* модуля *Анализ данных*;
- режима работы *Экспоненциальное сглаживание* модуля *Анализ данных*.

Кратко рассмотрим эти возможности Excel.

Команду *Добавить линию тренда* следует использовать в случаях когда функция $t(\tau)$ является: линейной, полиномиальной (степени не выше 6), логарифмической, степенной, экспоненциальной, а также для реализации алгоритма скользящего среднего (величина L может задаваться в интервале от 1 до 7).

Выполнение этой команды подробно описано в параграфе 2.7. Поэтому здесь ограничимся только одним примером ее использования.

Пример 4.2.3. Используя команду *Добавить линию тренда* по данным табл. 4.1 найти оценку $\hat{t}(\tau)$, предполагая, что $t(\tau)$ является квадратичной функцией.

Решение. При сделанном предположении оценку $\hat{t}(\tau)$ будем искать в виде

$$\hat{t}(\tau) = b_0 + b_1\tau + b_2\tau^2.$$

Первоначально в документ Excel вводим данные из табл. 4.1 (см. рис. 4.6 – кривая 1). Затем по этим данным строим график и вызываем команду *Добавить линию тренда*. В появившемся диалоговом окне задаем необходимые параметры (подробнее см. параграф 2.7) и щелкаем на кнопке ОК. На экране появится график $\hat{t}(\tau)$ (кривая – 2), уравнение функции $\hat{t}(\tau)$ и значение индекса детерминации (2.6.11), равное 0.849. Заметим, что коэффициент $b_0 = 132.3$, $b_1 = 55.09$, $b_2 = -3.27$ совпадают с коэффициентами, вычисленными в примере 4.2.1. ☉

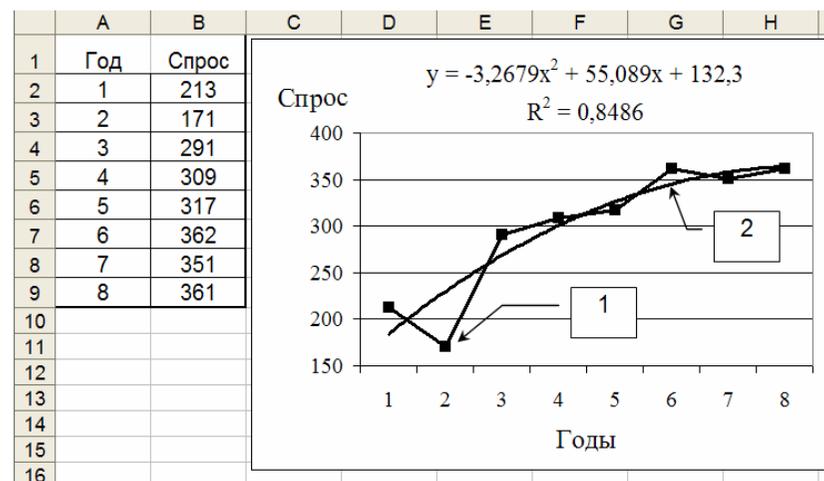


Рис. 4.6. Построение уравнения тренда с помощью команды «Добавить линию тренда»

Режим «Скользящее среднее» реализует алгоритм (4.2.10) с одинаковыми весами $c_l = \frac{1}{2L+1}$. Для вызова режима обратитесь

к пункту главного меню **Сервис**, выполнить команду **Анализ данных**, а затем в появившемся списке режимов работы модуля **Анализ данных** выделить *Скользящее среднее* и щелкнуть на кнопке ОК.

В появившемся диалоговом окне (см. рис. 4.7) задаем следующие параметры:

1. *Входной интервал* – вводится диапазон адресов ячеек, содержащих значения y_i .

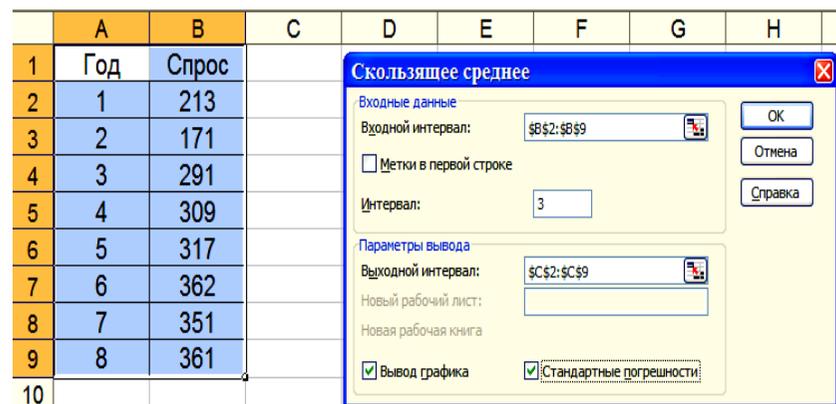


Рис. 4.7. Диалоговое окно режима «Скользящее среднее»

2. *Метки* – включается, если первая строка во входном интервале содержит заголовок. В этом случае автоматически будут созданы стандартные названия.

3. *Интервал* – задается размер окна $(2L + 1)$. По умолчанию размер 3.

4. *Выходной интервал* – содержит адрес верхней ячейки, начиная с которой выводятся вычисленные сглаженные значения.

5. *Вывод графика* – включает вывод графика заданных и сглаженных значений временного ряда.

6. *Стандартные погрешности* – включает вычисление и вывод в виде столбца стандартных погрешностей $\sigma_{\hat{\tau}_j}$, которые вычисляются по формуле

$$\sigma_{\hat{\tau}_j} = \sqrt{\frac{1}{2L+1} \cdot \sum_{i=-L}^L (y_{j-i} - \hat{\tau}_{j-i})^2}.$$

Пример 4.4. По данным табл. 4.1 вычислить значения тренда $\hat{\tau}_j$, используя режим работы *Скользящего среднего* модуля **Анализ данных**.

Решение. Введем в документ Excel исходные данные (см. рис. 4.7), а затем вызовем режим *Скользящего среднего* и зададим необходимые параметры (см. рис. 4.7). На рис. 4.8 показаны графики значений y_j (маркированные ромбами) и $\hat{\tau}_j$ (маркированные квадратами). Здесь же показаны формулы, по которым вычислялись значения $\hat{\tau}_i$ в некоторых ячейках. ●

Заметим, что в режиме *Скользящего среднего* реализована следующая формула:

$$\hat{\tau}_j = \frac{1}{2L+1} \cdot \sum_{i=-2L}^0 y_{j+i}, \quad j = 2L+1, 2L+2, \dots, n.$$

Поэтому в ячейках C2, C3 не определены сглаженные значения. По этой же причине в ячейках D2:D5 документа на рис. 4.8 не определены значения стандартных погрешностей.

Задание. Вычислить значение тренда, задав $(2L+1) = 5$. Сравните с предыдущими результатами.

Режим Экспоненциальное сглаживание реализует алгоритм (4.2.14). Для вызова режима обратиться к пункту главного меню **Сервис**, выполнить команду **Анализ данных**, а затем в появившемся списке режимов работы выделить *Экспоненциальное сглаживание* и щелкнуть на кнопке ОК. В появившемся диалоговом окне (см. рис. 4.9) задать необходимые параметры. Параметры режима *Экспоненциальное сглаживание* совпадают с параметрами режима *Скользящего среднего* за исключением одного

параметра. Вместо параметра *Интервал* необходимо задать *Фактор затухания*, равный величине α в формуле (4.2.12), которая может меняться в интервале (0,1).

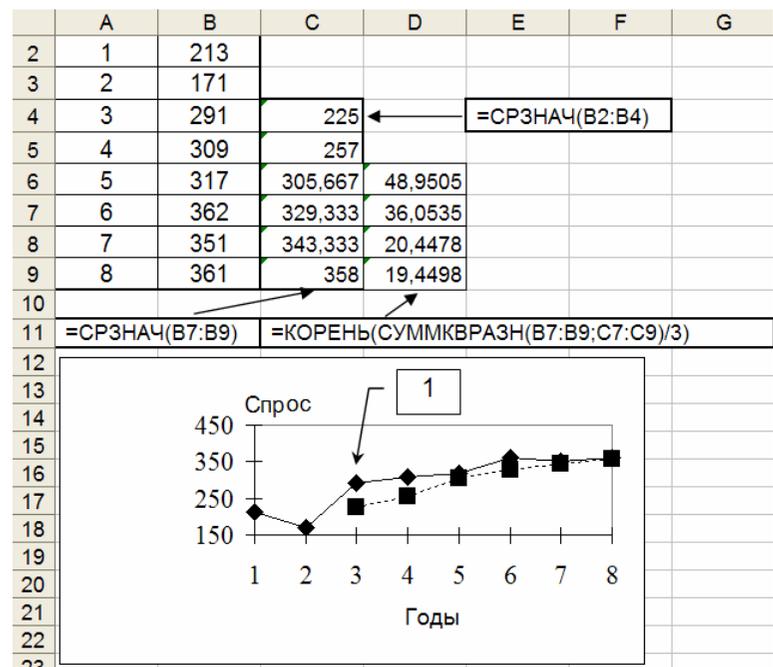


Рис. 4.8. Оценивание тренда в режиме «Скользящее среднее»

Пример 4.5. По данным таблицы 4.1 вычислить значение тренда $\hat{t}_j = \hat{t}(\tau_j)$, используя режим Экспоненциальное сглаживание.

Решение. Введем в документ Excel исходные данные (см. рис. 4.10), вызовем режим Экспоненциальное сглаживание и зададим необходимые параметры, задав $\alpha = 0.2$ (см. рис. 4.9). На рис. 4.10 показаны графики значений y_j (маркированные ромбами) и \hat{t}_j (маркированные квадратами).

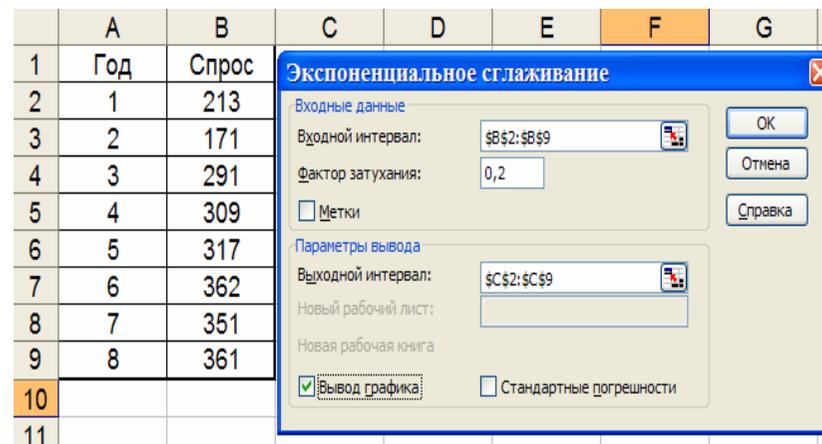


Рис.4.9. Диалоговое окно режима Экспоненциальное сглаживание

Задание. Вычислите значение тренда, задав $\alpha = 0.6$. Сравните полученные результаты с результатами примера 4.5 и объясните отличия.

В заключении этого параграфа можно сделать следующий вывод, основанный на анализе результатов рассмотренных примеров выделения тренда разными методами: *наиболее эффективным является метод, основанный на построении парной регрессии* (см. примеры 4.2.1 и 4.2.2). Этот метод достаточно универсален, позволяет непосредственно решать задачи прогнозирования и лишен недостатка, присущего методам сглаживания при вычислении значений на концах временного интервала.

4.3. Выделение периодических компонент временного ряда

К периодическим (иногда называют тригонометрическим) компонентам относятся: *сезонная составляющая $s(t)$* , отражающая повторяемость экономических процессов в течении не очень длительного периода (года, иногда месяца); *циклическая составляющая $c(t)$* отражающая повторяемость экономических процес-

сов в течении длительных периодов (например, волны экономической активности Кондратьева)

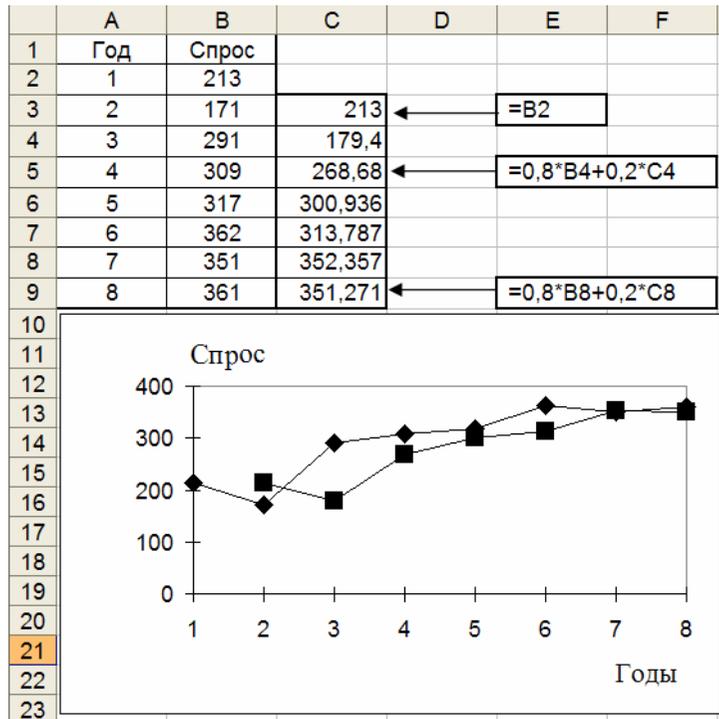


Рис. 4.10. Оценивание тренда в режиме «Экспоненциальное сглаживание»

Эти компоненты являются *периодическими функциями* и их можно объединить общим названием *тригонометрическая составляющая* (обозначим $s(\tau)$) *временного ряда*. Для выделения такой составляющей используют методы *гармонического анализа временного ряда*.

Основы гармонического анализа периодических функций. Гармонический анализ позволяет представить периодическую функцию линейной комбинацией косинусов и синусов.

Предположим, что функция $\varphi(\tau)$ является непрерывной функцией с периодом T . Тогда функцию $\varphi(\tau)$ можно представить рядом Фурье вида

$$\varphi(\tau) = a_0 + \sum_{k=1}^{\infty} [a_k \cos(\frac{2\pi k}{T}\tau) + b_k \sin(\frac{2\pi k}{T}\tau)], \quad 0 \leq \tau \leq T, \quad (4.3.1)$$

где k - номер гармоники, при увеличении которой уменьшается период функций $\cos(\frac{2\pi k}{T}\tau)$, $\sin(\frac{2\pi k}{T}\tau)$. Коэффициенты разложения определяются формулами:

$$a_0 = \frac{1}{T} \int_0^T \varphi(\tau) d\tau;$$

$$a_k = \frac{2}{T} \int_0^T \varphi(\tau) \cos(\frac{2\pi k}{T}\tau) d\tau;$$

$$b_k = \frac{2}{T} \int_0^T \varphi(\tau) \sin(\frac{2\pi k}{T}\tau) d\tau;$$

Аргументы тригонометрических функций \cos , \sin можно трактовать как частоты ω_k , определяемые соответствующим номером гармоники, т.е.

$$\omega_k = \frac{2\pi}{T} k. \quad (4.3.2)$$

Величины $S_k = a_k^2 + b_k^2$ характеризуют «энергетический вклад» k -ой гармоники в функцию $\varphi(\tau)$. Зависимость величины S_k от номера гармоники k (или от частоты ω_k (4.3.2)) характеризует спектральный состав (или спектр) функции $\varphi(\tau)$. Сравнительно большие величины S_k определяют частоты, на которых сосредоточена основная энергия функции $\varphi(\tau)$.

Под аппроксимацией функции $\varphi(\tau)$ рядом Фурье понимают новую функцию $\hat{\varphi}(\tau)$, полученную суммированием первых членов ряда (4.3.1), число которых обозначим K_0 , т.е.

$$\hat{\varphi}(\tau) = a_0 + \sum_{k=1}^{K_0} [a_k \cos(\frac{2\pi k}{T} \tau) + b_k \sin(\frac{2\pi k}{T} \tau)] \quad (4.3.3)$$

Видно, что в функции $\hat{\varphi}(\tau)$ отсутствуют “высокочастотные” гармоники с номерами $k > K_0$, которые присутствовали в исходной функции $\varphi(\tau)$. Такой способ получения функции $\hat{\varphi}(\tau)$ часто называют низкочастотной фильтрацией функции $\varphi(\tau)$.

По аналогии можно построить новую функцию $\hat{\varphi}(\tau)$, содержащую только заданные гармоники, например гармоники с наиболее значимым спектром S_k . Предположим, что такие гармоники имеют номера $k = 3, 8$. Тогда функция $\hat{\varphi}(\tau)$, содержащая только эти гармоники, записывается в виде:

$$\hat{\varphi}(\tau) = a_3 \cos(\frac{2\pi 3}{T} \tau) + b_3 \sin(\frac{2\pi 3}{T} \tau) + a_8 \cos(\frac{2\pi 8}{T} \tau) + b_8 \sin(\frac{2\pi 8}{T} \tau).$$

Такой способ построения функции широко используется для выделения тригонометрической составляющей временного ряда.

Пример 4.3.1. Дана функция

$$\varphi(\tau) = 0.1 + 0.4\tau + 0.5\tau^2 + 3 \sin(\frac{2\pi}{3.2} 5\tau), \quad (4.3.4)$$

определенная на интервале $[0, 3.2]$. График функции показан сплошной линией на рисунке 4.11. Необходимо вычислить спектр S_k этой функции и выделить из функции $\varphi(\tau)$ основную (имеющую наибольшее значение спектра) тригонометрическую составляющую.

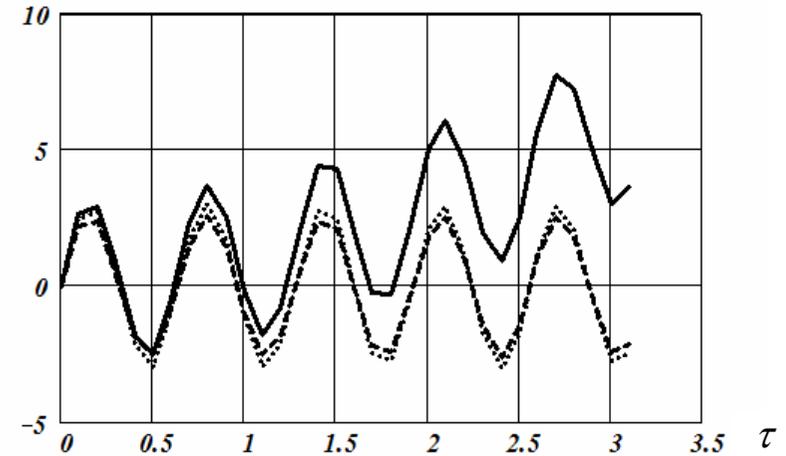


Рис. 4.11. Тригонометрическая составляющая функции $\varphi(\tau)$

Решение. Из аналитического задания $\varphi(\tau)$ (4.3.4) следует, что тригонометрическая составляющая этой функции обусловлена слагаемым $3 \sin(\frac{2\pi}{3.2} \cdot 5\tau)$ и соответствует гармонике с номером 5. Так как функция задана на интервале $[0, 3.2]$, то период этой функции задаем равным $T = 3.2$. Используя приведенные выше формулы, вычисляем коэффициенты $a_0, a_k, b_k, k = 1.., 20$ и определяем спектры $S_0 = a_0^2, S_k = a_k^2 + b_k^2$. Значения S_k приведены на рисунке 4.12 (кривая 1). Большие значения S_0, S_1 обусловлены наличием в функции $\varphi(\tau)$ квадратичного тренда (первые три слагаемых в (4.3.4)), большое значение S_5 обусловлено присутствием в $\varphi(\tau)$ тригонометрической составляющей, для которой вычислены коэффициенты $a_5 = 0.021, b_5 = 2.593$. Построим функцию $\hat{\varphi}(\tau) = 0.021 \cdot \cos(\frac{2\pi}{3.2} 5\tau) + 2.593 \cdot \sin(\frac{2\pi}{3.2} 5\tau)$, которая соответствует этой гармонике. График этой функции приведен на

рисунке 4.12 (штриховая кривая), здесь же приведен график функции $3\sin(\frac{2\pi}{3.2} \cdot 5\tau)$ (точечная кривая), которая входит в исходную функцию $\varphi(\tau)$. Из хорошего совпадения этих графиком можно сделать вывод об эффективности применения методов гармонического анализа для выделения тригонометрических составляющих периодических функций. ●

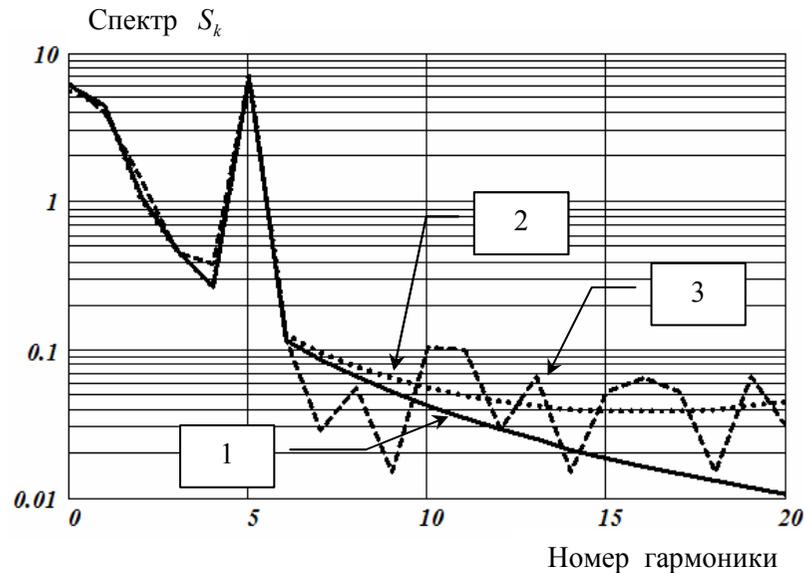


Рис. 4.12. Спектральный состав функции $\varphi(\tau)$

Выделение тригонометрической составляющей временного ряда методами гармонического анализа. Использование рядов Фурье для выделения тригонометрической составляющей временного ряда отличается от рассмотренного выше следующими моментами:

1. Значения временного ряда заданы в дискретные моменты времени τ_i и чаще всего эти моменты представляют собой арифметическую прогрессию с началом Δ_τ , т.е.

$$\tau_i = \tau_{нач} + (i-1) \cdot \Delta_\tau, \quad i = 1, 2, \dots, n, \quad (4.3.5)$$

или $\tau_{i+1} - \tau_i = \Delta_\tau$. Тогда в качестве периода принимается величина $T = \Delta_\tau \cdot n - \tau_{нач}$.

В гармоническом анализе доказывается, что коэффициенты ряда Фурье не зависят от $\tau_{нач}$. В дальнейшем полагается, что τ_i образуют арифметическую прогрессию (4.3.5) при $\tau_{нач} = 0$.

2. Временной ряд $Y(\tau)$ кроме тригонометрической составляющей $c(\tau)$ содержит случайную составляющую $\varepsilon(\tau)$, которую необходимо отделить от тригонометрической составляющей.

Первое отличие обуславливает замену интегралов, определяющих значения a_0, a_k, b_k квадратурными формулами, в которые входят значения $y_i, i = 1, 2, \dots, n$. В качестве примера примем формулу левых прямоугольников и тогда получаем следующие выражения для вычисления интегралов:

$$a_0^* = \frac{\Delta_\tau}{n} \cdot \sum_{i=1}^n y_i; \quad (4.3.6)$$

$$a_k^* = \frac{2\Delta_\tau}{n} \cdot \sum_{i=1}^n y_i \cos\left(\frac{2\pi k}{T} \tau_i\right); \quad (4.3.7)$$

$$b_k^* = \frac{2\Delta_\tau}{n} \sum_{i=1}^n y_i \sin\left(\frac{2\pi k}{T} \tau_i\right). \quad (4.3.8)$$

Символ * означает, что коэффициенты вычислены по дискретным значениям временного ряда, или иначе – по временной выборке.

По аналогии со спектром S_k определим дискретный спектр как $S_k^* = (a_k^*)^2 + (b_k^*)^2$. Дискретность задания значений y_i обуславливает симметричность спектра S_k^* относительно точки $n/2$, т.е. $S_{n/2+j}^* = S_{n/2-j}^*, j = 1, \dots, n/2 - 1$. Поэтому имеет смысл вычислить коэффициенты ряда Фурье для гармоник с номерами $k = 0, 1, 2, \dots, n/2$.

Учет второго отмеченного момента основан на следующем предположении: амплитуда случайной составляющей $\varepsilon(\tau)$ намного меньше амплитуды $c(\tau)$ и спектр $\varepsilon(\tau)$ более менее равномерно «распределен» по гармоникам с различными номерами, т.е. сигнал $\varepsilon(\tau)$ имеет «широкий» спектр, но его вклад в спектр каждой гармоники тригонометрической составляющей сравнительно мал.

Пример 4.3.2. Функция $p(\tau)$ задается формулой

$$p(\tau) = 3 \sin\left(\frac{2\pi}{3.2} 5\tau\right),$$

а значения временного ряда формируются как

$$y_i = 0.1 + 0.4\tau_i + 0.5\tau_i^2 + p(\tau_i) + \delta \cdot \varepsilon(\tau_i),$$

где $\tau_i = (i-1) \cdot 0.1$, $i = 1, 2, \dots, 32$, а $\varepsilon(\tau_i)$ - нормально распределенная величина с нулевым средним и дисперсией $\sigma^2 = 0.11$ (что соответствует относительному уровню 0.10). Необходимо вычислить спектр S_k^{**} временной выборки и выделить тригонометрическую составляющую.

Решение. Первоначально были получены значения y_i при $\delta = 0$ (т.е. отсутствует случайная составляющая $\varepsilon(\tau_i)$) и по формулы (4.3.6) - (4.3.8) вычислены коэффициенты a_0^* , a_k^* , b_k^* и определен спектр S_k^* , значения которых отображены на рисунке 4.12 (кривая 2). Затем были получены значения y_i при $\delta = 1$ (т.е. присутствует случайная составляющая $\varepsilon(\tau_i)$), вычислены коэффициенты a_0^{**} , a_k^{**} , b_k^{**} и определен спектр S_k^{**} , значения которых отображены на рисунке 4.12 (кривая 3). Анализ спектров изображенных на рис. 4.12 позволяет сделать следующие выводы:

- Спектры S_k^* , S_k^{**} , вычисленные по значениям дискретного временного ряда являются симметричными функциями относительно точки $k = n/2 = 16$.

- Во всех трех спектрах присутствует максимум, соответствующий $k = 5$, что говорит о наличии во временном ряду тригонометрической составляющей вида:

$$\varepsilon(\tau) = a_5^{**} \cos\left(\frac{2\pi}{3.2} 5\tau\right) + b_5^{**} \sin\left(\frac{2\pi}{3.2} 5\tau\right),$$

где коэффициенты $a_5^{**} = -0.119$; $b_5^{**} = 2.625$, вычисленные по значениям дискретного временного ряда приближенно равны коэффициентам, вычисленными по непрерывной функции (4.3.4) (см. пример 4.3.1).

Таким образом, используя разложение временного ряда в ряд Фурье, удается достаточно точно выделить тригонометрическую составляющую временного ряда, «отфильтровав» тренд $t(\tau)$ (в нашем примере это полином второй степени $0.1 + 0.4\tau_i + 0.5\tau_i^2$) и случайную составляющую $\varepsilon(\tau)$.

Вычисление коэффициентов ряда Фурье в Excel. В Excel вычислять коэффициенты разложения в ряд Фурье можно двумя способами:

- Программированием в документе Excel формул (4.3.6)-(4.3.8);

- Используя режим *Анализ Фурье* модуля **Анализ данных**.

Первый способ достаточно громоздок, и его можно рекомендовать при сравнительно небольших объемах временной выборки с небольшим числом вычисляемых коэффициентов ряда Фурье.

Второй способ основан на дискретном преобразовании Фурье. Кратко остановимся на этом преобразовании.

Пара дискретных преобразований, определяемая формулами:

$$Z(k) = \sum_{j=0}^{N-1} z(j) e^{-i \frac{2\pi jk}{N}} \quad (\text{прямое ДПФ});$$

$$z(k) = \frac{1}{N} \sum_{j=0}^{N-1} Z(j) e^{i \frac{2\pi jk}{N}} \quad (\text{обратное ДПФ}),$$

где $i = -\sqrt{-1}$ мнимая единица, называется дискретным преобразованием Фурье (ДПФ). Исходная дискретная последователь-

ность $z(j)$ является периодической с периодом N . Последовательность $Z(k)$ (называемая коэффициентами ДПФ) также является периодической с периодом N .

Если коэффициенты $Y(k)$ вычислены по значениям временного ряда $y_j, j = 1, \dots, n$ то связь между коэффициентами ДПФ и коэффициентами разложения в ряд определяется как:

$$a_0^* = \frac{1}{n} Y(0), \quad a_k^* = \frac{2}{n} \operatorname{Re}[Y(k)], \quad b_k^* = \frac{2}{n} \operatorname{Im}[Y(k)], \quad k = 1, \dots, n/2,$$

где $\operatorname{Re}[A]$, $\operatorname{Im}[A]$ - означают вещественную и мнимую части комплексного числа A .

Прямое и обратное ДПФ вычисляются в режиме *Анализ Фурье*. Для вызова этого режима обратиться к пункту **Сервис** главного меню, выполнить команду **Анализ данных** и в появившемся списке режимов выделить *Анализ Фурье* и щелкнуть мышью ОК.

Затем в новом диалоговом окне задать следующие параметры (см. рис. 4.13):

Входной интервал – указывается диапазон ячеек, содержащие вещественные данные, к которым применяется ДПФ.

Метки в первой строке – включается, если первая строка содержит заголовок.

Выходной интервал – вводится адрес левой верхней ячейки выходного диапазона

Инверсия – включается, если необходимо вычислить обратное ДПФ.

Замечание 4.3.1. Используемый для вычисления ДПФ алгоритм (называемый алгоритмом *быстрого преобразования Фурье - БПФ*) требует, чтобы n - число значений временного ряда, должно быть обязательно равным степени 2 (т.е. 8, 16, 32, 64, ...), что является существенным ограничением. Один из путей преодоления этого недостатка – добавление в конец временной выборки нулей до тех пор, пока длина “новой” временной выборки не станет равной степени 2. Однако такой способ, применяемый при цифровой обработке сигналов, далеко не всегда пригоден для обработки данных, характеризующие экономические процессы. По-

этому перед применением режима *Анализ Фурье* необходимо сформировать выборку длиной n , равной степени 2. ♥

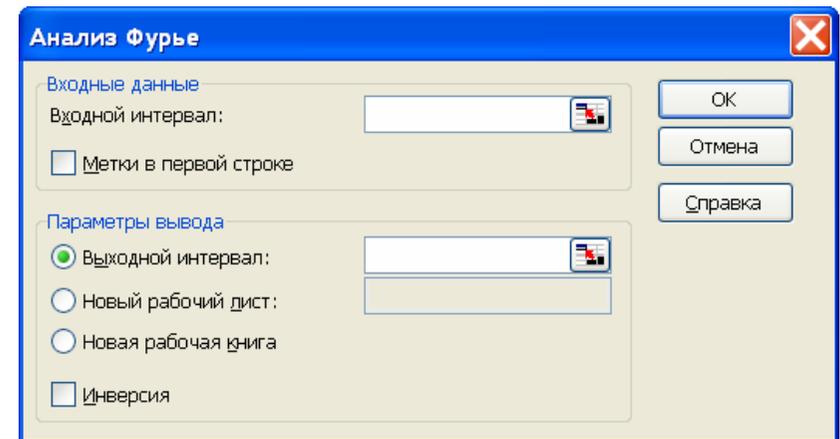


Рис. 4.13. Диалоговое окно режима *Анализ Фурье*

Так как результатом работы режима *Анализ Фурье* будут комплексные числа вида $a_k + ib_k$, то для вычисления вещественной и мнимой части можно использовать следующие функции Excel (категория *Инженерные*): ВЕЩ(), МНИМ().

4.4. Построение авторегрессионных моделей временного ряда

Понятие авторегрессионной модели. Для временного ряда далеко не всегда удастся подобрать адекватную модель (4.1.1), для которой возмущения ε_i будут удовлетворять условиям Гаусса - Маркова (см. параграф 2.1). Ранее, при выделении тренда (см. параграф 4.2) в качестве объясняющей переменной (или регрессора) модели выступало время τ . Однако в эконометрике достаточно широкое распространение получили регрессионные модели, в которых регрессорами выступают *лаговые переменные*,

влияние которых характеризуется некоторым «запаздыванием». Наиболее часто в качестве такой модели используется *авторегрессионная модель*.

Авторегрессионная модель p -го порядка (или модель AR(p)) имеет вид

$$y_i = \beta_0 + \beta_1 y_{i-1} + \beta_2 y_{i-2} + \dots + \beta_p y_{i-p} + \varepsilon_i, i=1, 2, \dots, n, \quad (4.4.1)$$

где $\beta_0, \beta_1, \dots, \beta_p$ - коэффициенты модели, y_{i-j} - лаговые переменные, определяющие зависимость значения y_i временного ряда в момент τ_i от значений в предыдущие моменты времени. Эти лаговые переменные и выступают в роли регрессоров (объясняющих переменных). Возмущения ε_i удовлетворяют условиям Гаусса-Маркова (см. параграф 3.2).

Наиболее часто используются:

- авторегрессионная модель 1-го порядка (или модель AR(1)):

$$y_i = \beta_0 + \beta_1 y_{i-1} + \varepsilon_i, i=1, 2, \dots, n; \quad (4.4.2)$$

- авторегрессионная модель 2-го порядка (или модель AR(2)):

$$y_i = \beta_0 + \beta_1 y_{i-1} + \beta_2 y_{i-2} + \varepsilon_i, i=1, 2, \dots, n. \quad (4.4.3)$$

Оценивание коэффициентов авторегрессионной модели.

Так как лаговые переменные по своей сути являются объясняющими переменными, то модель (4.4.1) можно рассматривать как линейную множественную регрессию, коэффициенты которой можно оценить на основе простого метода наименьших квадратов, записав для модели (4.4.1) следующее уравнение регрессии:

$$\varepsilon_i = b_0 + b_1 y_{i-1} + b_2 y_{i-2} + \dots + b_p y_{i-p}, \quad i=1, 2, \dots, n, \quad (4.4.4)$$

коэффициенты которого являются оценками для $\beta_0, \beta_1, \dots, \beta_p$ соответственно.

Тогда матрица X размером $n \times (p+1)$, входящая в матричную запись системы нормальных уравнений (см. (3.2.3)) имеет вид:

$$X = \begin{pmatrix} 1 & y_0 & y_{-1} & \dots & y_{-p+1} \\ 1 & y_1 & y_0 & \dots & y_{-p+2} \\ 1 & y_2 & y_1 & \dots & y_{-p+3} \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & y_{n-1} & y_{n-2} & \dots & y_{n-p} \end{pmatrix} \quad (4.4.5)$$

Для моделей AR(1), AR(2) матрицы X содержат элементы:

$$X = \begin{pmatrix} 1 & y_0 \\ 1 & y_1 \\ 1 & y_2 \\ \vdots & \vdots \\ 1 & y_{n-1} \end{pmatrix}; \quad X = \begin{pmatrix} 1 & y_0 & y_{-1} \\ 1 & y_1 & y_0 \\ 1 & y_2 & y_1 \\ \vdots & \vdots & \vdots \\ 1 & y_{n-1} & y_{n-2} \end{pmatrix}. \quad (4.4.6)$$

Замечание 4.4.1. «Начальные» значения $y_0, y_{-1}, \dots, y_{-p+1}$, входящие в матрицу X , как правило, неизвестны, и поэтому их необходимо доопределить. Например, $y_0 = \bar{y}$ - среднее значение наблюдаемых значений, $y_{-1} = y_{-2} = \dots = y_{-p+1} = 0$. Этот способ используется ниже в примере 4.4.1. Возможны и другие способы задания этих «начальных значений». ♥

Определив вектор наблюдений y как $y = |y_1, y_2, \dots, y_n|^T$, приходим к системе нормальных уравнений:

$$X^T X b = X^T y,$$

где вектор $b = [b_0, b_1, \dots, b_p]$ содержит искомые оценки для коэффициентов $\beta_0, \beta_1, \dots, \beta_p$ авторегрессионной модели. Предполагая, что обратная матрица $(X^T X)^{-1}$ существует, находим вектор b :

$$b = (X^T X)^{-1} X^T y. \quad (4.4.7)$$

После этого уравнение регрессии (4.4.4) можно использовать для прогнозирования временного ряда, описываемого авторегрессионной моделью.

Замечание 4.4.2. Существенным предположением, при котором вектор b является несмещенной и состоятельной оценкой для вектора коэффициентов β является предположение, что матрица X является не случайной (предположение $P1$ параграфа 3.1). В нашем случае матрица X (4.4.5) является случайной, так как элементами ее являются случайные величины – лаговые переменные y_{i-1}, \dots, y_{i-p} . Нарушение предположения $P1$ приводит к тому, что вектор b уже **не является несмещенной и состоятельной оценкой** для вектора β . Для преодоления этого недостатка используются более сложные методы оценивания, рассмотрение которых выходит за рамки данного учебного пособия (для ознакомления с этими методами см. [1]).

Пример 4.4.1. В таблице 4.2 представлены данные, отражающие динамику курса акций некоторой компании (в условных единицах).

Таблица 4.2

i	1	2	3	4	5	6	7
y_i	971	1166	1044	907	957	727	752
i	8	9	10	11	12	13	14
y_i	1019	972	815	823	1112	1386	1428
i	15	16	17	18	19	20	21
y_i	1364	1241	1145	1351	1325	1226	1189

Необходимо по этим данным построить модель AR(1), AR(2) и определить их значимость при уровне значимости $\alpha = 0.05$, а также построить полиномиальные тренды $\hat{k}(\tau)$ первого, второго порядка и проверить их значимость. Построить прогноз для $i = 22, 23$.

Решение. Первоначально, используя команду “Добавить линию тренда” (см. параграф 4.2), определим коэффициенты уравнения линейного тренда ($p = 1, m = p + 1 = 2$)

$$\hat{k}(\tau) = b_0 + b_1 \tau,$$

квадратичного тренда ($p = 2, m = 3$)

$$\hat{k}(\tau) = b_0 + b_1 \tau + b_2 \tau^2,$$

и вычислим для каждого уравнения тренда значения величин R^2, \hat{R}^2, F по формулам (3.4.8), (3.4.12), (3.4.10) соответственно. Эти значения и соответствующие уравнения трендов приведены в табл. 4.3.

Таблица 4.3

Уравнение тренда	R^2	\hat{R}^2	F	Кван- тиль
$\hat{k}(\tau) = 854.66 + 21.52\tau$ ($p = 1, m = 2$)	0.379	0.346	11.62	4.38
$\hat{k}(\tau) = 943.83 - 1.74\tau + 1.06\tau^2$ ($p = 2, m = 3$)	0.406	0.373	12.99	3.55

Определим квантиль $F_{0.95; m-1; m-n}$ F-распределения по известной формуле:

$$F_{0.95; m-1; m-n} = F_{ПАСПОБР}(0.05; m-1; m-n)$$

при $n = 21$. Значение квантиля, играющего роль границы критической области при проверке гипотезы о значимости построенной регрессии также приведены в табл. 4.3. Из этой таблицы видно, что неравенство (см. параграф 3.4)

$$F > F_{0.95; m-1; m-n} \quad (4.4.8)$$

выполняется как для линейного, так и для квадратичного тренда. Поэтому можно сказать, что построенные модели тренда значимы при уровне значимости 0.05.

На рис. 4.13 нанесены значения временного ряда y_i (отображены маркерами в форме квадратов – кривая 1) и значения квадратичного тренда (кривая 2). Видно, что несмотря на принятие гипотезы о значимости, уравнение тренда не отображает динамику рассматриваемого временного ряда, а представляет “усредненную линию” его поведения. Поэтому перейдем к построению авторегрессионных моделей для этого временного ряда.

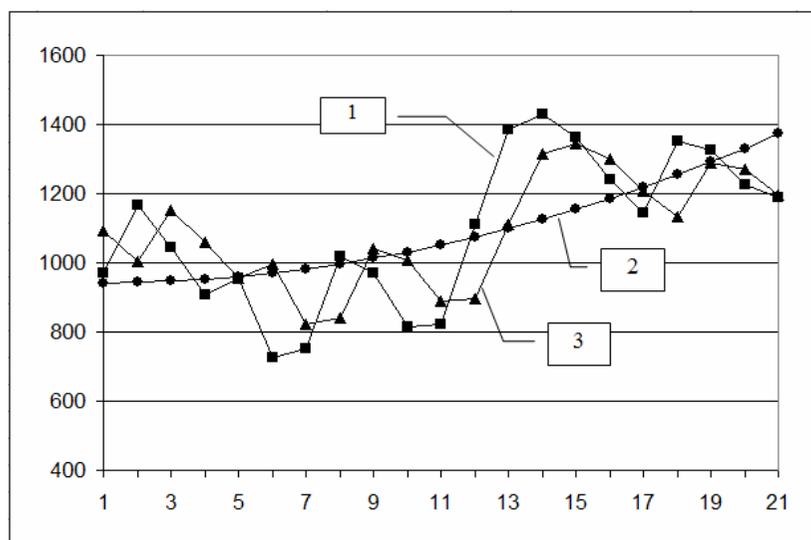


Рис. 4.14. Графики к примеру 4.4.1.

Уравнение регрессии для модели AR(1), определенной выражением (4.4.2) имеет вид:

$$\hat{y}_i = b_0 + b_1 y_{i-1} \quad (4.4.9)$$

Для вычисления коэффициентов b_0, b_1 , которые являются оценками для β_0, β_1 модели (4.4.2) используем обычный метод наименьших квадратов, реализованный в режиме *Регрессия* модуля *Анализ данных* (см. параграф 3.5 и пример 3.5.1). Вычисленные значения: $b_0 = 280.90, b_1 = 0.746$. Значения величин R^2, \hat{R}^2, F приведены в табл. 4.4 (вторая строка).

Уравнение регрессии для модели AR(2), определенной формулой (4.4.3) имеет вид:

$$\hat{y}_i = b_0 + b_1 y_{i-1} + b_2 y_{i-2}. \quad (4.4.10)$$

Используя режим *Регрессия* модуля *Анализ данных*, вычисляем коэффициенты $b_0 = 275.61, b_1 = 0.730, b_2 = 0.022$. В таблице 4.4 приведены значения R^2, \hat{R}^2, F для уравнения (4.4.10).

Таблица 4.4

Модель авторегрессии	R^2	\hat{R}^2	F	Кван- тиль
$\hat{y}_i = 280.90 + 0.746y_{i-1}$	0.553	0.52 8	23.26	4.38
$\hat{y}_i = 275.61 + 0.730y_{i-1} + 0.022y_{i-2}$	0.554	0.52 5	23.22	3.55

Из таблицы 4.4 видно:

- а) уравнения (4.4.9), (4.4.10) имеют существенно большие значения R^2, \hat{R}^2, F по сравнению с квадратичным трендом;
- б) уравнения (4.4.9), (4.4.10) при вычисленных коэффициентах b_0, b_1, b_2 являются значимы при $\alpha = 0.05$ и практически имеют одни и те же характеристики.

Поэтому, исходя из принципа минимальной сложности (см. параграф 2.2), для прогнозирования значений временного ряда будет использоваться уравнение:

$$\hat{y}_i = 280.90 + 0.746y_{i-1}.$$

Для $i = 22$ прогнозное значение равно

$$\hat{y}_{22} = 280.90 + 0.746 \cdot 1189 = 1168.$$

Для $i = 23$ прогнозное значение равно

$$\hat{y}_{23} = 280.90 + 0.746 \hat{y}_{22} = 280.90 + 0.746 \cdot 1168 = 1152.$$

4.5. Временные ряды с коррелированными возмущениями

В этом параграфе будут рассмотрены временные ряды, в которых возмущения ε_i коррелированы. Построены модели для таких коррелированных возмущений и приведены процедуры оценивания параметров этих моделей.

Временные ряды с коррелированными возмущениями. Упорядоченность наблюдений в пространственной выборке временного ряда оказывается существенной в тех случаях, когда присутствуют механизмы влияния результатов предыдущих наблюдений на результаты последующих. Такое влияние имело место при наличии в модели временного ряда лаговых переменных $y_{i-1}, y_{i-2}, \dots, y_{i-p}$. Авторегрессионные модели с такими переменными были рассмотрены в параграфе 4.4. При этом возмущения ε_i временного ряда удовлетворяли условиям Гаусса-Маркова $P2 \div P4$ (см. параграф 3.2), т.е. ε_i подчинялась нормальному распределению с нулевым средним, дисперсией σ^2 и возмущения $\varepsilon_i, \varepsilon_j$ при $i \neq j$ были некоррелированными.

Влияние результатов предыдущих наблюдений на последующие имеет место, даже в случаях, когда отсутствуют лаговые переменные, но возмущения ε_i в следующей регрессионной модели наблюдений временного ряда

$$y_i = q(\tau_i) + \varepsilon_i, i = 1, 2, \dots, n, \quad (4.5.1)$$

оказывается зависимыми случайными величинами, т.е. корреляционный момент $\mu_{i,j} = M(\varepsilon_i \varepsilon_j)$ не равен нулю при $i \neq j$. Оче-

видно, что и коэффициент корреляции $\rho_{\varepsilon_i \varepsilon_j}$ между величинами $\varepsilon_i, \varepsilon_j$ также отличается от нуля.

Регрессионные модели временного ряда, в которых условие $\rho_{\varepsilon_i \varepsilon_j} = 0$, при $i \neq j$ не выполняется, называются моделями с наличием автокорреляции. Возможны два вида автокорреляции: положительная и отрицательная, определяемая знаком коэффициента корреляции между соседними $\varepsilon_i, \varepsilon_{i+1}$.

Положительная автокорреляция проявляется в чередовании временных интервалов, где наблюдаемые значения временного ряда оказываются выше или ниже значений $q(\tau_i)$ объясняемой части регрессионной модели.

Отрицательная автокорреляция характеризуется тем, что наблюдения действуют друг на друга по “принципу маятника” – завышенные значения в предыдущих наблюдениях приводят к занижению последующих значений, т.е. наблюдения y_i слишком часто перескакивают через график объясненной части $q(\tau_i)$.

Возникает вопрос: “К чему приводит наличие автокорреляции временного ряда?”. Использование обычного (классического) метода наименьших квадратов для оценивания коэффициентов объясненной части в этом случае также дает несмещенные и состоятельные оценки, но эти оценки *не являются эффективными* (т.е. существуют оценки с меньшей дисперсией). Более того оценки $s_{b_j}^2$ дисперсий коэффициентов b_j являются смещенными и несостоятельными, что приводит к недостоверным результатам проверки гипотез о значимости вычисленных коэффициентов b_j .

Как же определить наличие автокорреляции в наблюдаемых значениях временного ряда? Достаточно простой критерий, дающий ответ на наличие автокорреляции между соседними наблюдениями дает следующий тест.

Тест Дарбина – Уотсона. Этот тест основан на простой идее: если корреляция между ε_i и ε_{i+1} не равна нулю, то она присутствует и в остатках (невязках) $e_i = y_i - \hat{y}_i$ регрессионной моде-

ли, где $\hat{\epsilon}_i = \hat{\phi}(\tau_i)$ оценка объясненной части временного ряда, построенная обычным методом наименьших квадратов. Определим статистику

$$d = \frac{\sum_{i=2}^n (e_i - e_{i-1})^2}{\sum_{i=2}^n e_i^2}. \quad (4.5.2)$$

Между этой статистикой и выборочным коэффициентом корреляции r имеется связь:

$$d \approx 2(1-r). \quad (4.5.3)$$

В случае отсутствия автокорреляции (т.е. $r \approx 0$) значение статистики близко к двум. Близость статистики к нулю должна означать наличие положительной автокорреляции, к четырем - отрицательной автокорреляции. К сожалению, не определена пороговая точка для статистики d , при принятии или отвержении нулевой гипотезы H_0 : автокорреляция отсутствует. Поэтому весь диапазон значений d делится на ряд интервалов. Если наблюдается значение:

- а) $d_B < d < 4 - d_B$, то гипотеза H_0 принимается;
- б) $d_H < d < d_B$ или $4 - d_B < d < 4 - d_H$, то вопрос о принятии или отвержении гипотезы H_0 остается открытым;
- в) $0 < d < d_H$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза о положительной автокорреляции;
- г) $4 - d_H < d < 4$, то гипотеза H_0 отвергается и принимается альтернативная гипотеза о наличии отрицательной автокорреляции.

Пороговые значения d_H, d_B зависят от числа наблюдений, числа объясняющих переменных в функции $q(\tau)$ и уровня значимости. Эти значения приводятся в специальной таблице (см. например []) и определены для $n \geq 15$. Это ограничение является определенным недостатком этого теста. В таблице 4.4 приведены некоторые значения d_H, d_B для уровня значимости $\alpha = 0.05$. Ис-

пользуя данные таблицы можно экстраполировать d_H, d_B на меньшее число наблюдений.

Таблица 4.4

n	$k=1$		$k=2$		$k=3$		$k=4$	
	d_H	d_B	d_H	d_B	d_H	d_B	d_H	d_B
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72

Пример 4.5.1. Выявить на уровне значимости $\alpha = 0.05$ наличие автокорреляции временного ряда, значения которые приведены в табл. 4.1.

Решение. В качестве оценки $\hat{\phi}(\tau)$ для объясненной части $q(\tau)$ возьмем квадратичный полином (см. пример 4.2.1)

$$\hat{\phi}(\tau) = 132.3 + 55.09\tau - 3.26\tau^2,$$

что соответствует $k = 2$.

Вычислим остатки $e_i, i=1,2,\dots,8$ и статистику d , как показано на фрагменте документа Excel, приведенного на рисунке 4.15. Найденное значение $d = 3.037$. Используя таблицу 4.4, выполним линейную экстраполяцию значений d_H, d_B при $k=2$ для числа наблюдений $n=8$. Получим $d_H \approx 0.74, d_B \approx 1.54$. Видно, что вычисленное значение d находится в пределах от $d_B = 1.54$ до $4 - d_H = 3.26$, что соответствует принятию гипотезы H_0 об отсутствии автокорреляции с уровнем значимости $\alpha = 0.05$. ●

Допустим, что с помощью теста Дарбина-Уотсона (или другого теста) установлено наличие автокорреляции. Кроме этого справедливы следующие предположения о числовых характеристиках возмущений ϵ_i : $M(\epsilon_i) = 0, D(\epsilon_i) = \sigma^2$. При этих предпо-

ложениях возмущения ε_i можно рассматривать как *стационарный дискретный случайный процесс* (другими словами – *стационарный временной ряд*) с коррелированными значениями.

В качестве математических моделей для описания такого процесса используют модели двух классов:

- Авторегрессионные модели;
- Модели скользящего среднего.

Рассмотрим некоторые модели из этих классов.

	H	I	J	K	L	M	N
1	Год	Спрос	\bar{q}	e_i	$e_i - e_{i-1}$	e_i^2	$(e_i - e_{i-1})^2$
2	1	213	184,13	28,87		833,77	
3	2	171	229,41	-58,41	-87,29	3411,81	7618,80
4	3	291	268,16	22,84	81,25	521,63	6601,56
5	4	309	300,38	8,62	-14,21	74,39	202,05
6	5	317	326,05	-9,05	-17,68	81,97	312,53
7	6	362	345,20	16,80	25,86	282,36	668,59
8	7	351	357,80	-6,80	-23,61	46,29	557,30
9	8	361	363,88	-2,88	3,93	8,27	15,43
10				Суммы		5260,48	15976,26
11							
12	Статистика d			3,037			
13						=СУММ(N3:N9)	
14							
15	=N10/M10						

Рис. 4.15. Вычисление статистики Дарбина-Уотсона

Модель авторегрессии первого порядка (модель $AR(1)$).

Эта модель описывает так называемый марковский процесс, состояние которого в каждой следующий момент времени определяется только состоянием в настоящий момент времени и не зависит от того, каким путём процесс достиг этого состояния. Модель $AR(1)$ определяется соотношением

$$\varepsilon_i = \mu\varepsilon_{i-1} + \xi_i, \quad (4.5.4)$$

где μ - коэффициент модели, часто называемый коэффициентом авторегрессии ($|\mu| < 1$), ξ_i - последовательность случайных величин, образующих «белый шум» (см. параграф 4.1) с характеристиками:

$$M(\xi_i) = 0; \quad M(\xi_i \xi_{i \pm j}) = \begin{cases} \sigma^2, & \text{если } j = 0; \\ 0, & \text{если } j \neq 0. \end{cases}$$

Тогда из (4.5.4) следуют выражения для числовых характеристик процесса ε_i :

$$M(\varepsilon_i) = 0, \quad \sigma_{\varepsilon_i}^2 = D(\varepsilon_i) = \frac{\sigma^2}{1 - \mu^2}, \quad \rho(l) = \rho_{\varepsilon_i \varepsilon_{i \pm l}} = \mu^l, \quad (4.5.5)$$

где $\rho(l)$ - коэффициент автокорреляции (4.1.5).

Условие стационарности ряда ε_i имеет вид

$$|\mu| < 1. \quad (4.5.6)$$

Из выражений (4.5.5) видно, что при $|\mu|$ близком к 1 дисперсия случайной величины ε_i будет намного больше дисперсии σ^2 и между соседними значениями имеется сильная корреляция (коэффициент корреляции ρ равен μ).

Для идентификации параметров μ, σ^2 модели $AR(1)$, используются остатки (невязки) $e_i = y_i - \hat{y}_i, i = 1, 2, \dots, n$, где $y_i = \hat{y}(\tau_i)$ - оценка детерминированной составляющей временного ряда (см. параграфы 4.2. – 4.4). Оценки параметров определяются выражениями:

$$\hat{\mu} = \frac{1}{n-1} \sum_{i=1}^{n-1} (e_i - \bar{e})(e_{i+1} - \bar{e}) / s_e^2; \quad (4.5.7)$$

$$\hat{\sigma}^2 = (1 - \hat{\mu}^2) \cdot s_e^2, \quad (4.5.8)$$

где $\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i; \quad s_e^2 = \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2$ - выборочные среднее и дисперсия остатков e_i .

Модель авторегрессии второго порядка (модель AR(2)).
Модель определяется соотношением

$$\varepsilon_i = \mu_1 \varepsilon_{i-1} + \mu_2 \varepsilon_{i-2} + \xi_i, \quad (4.5.9)$$

где ξ_i - последовательность некоррелированных величин с характеристиками

$$M(\xi_i) = 0; \quad M(\xi_i \xi_{i \pm j}) = \begin{cases} \sigma^2, & \text{если } j = 0; \\ 0, & \text{если } j \neq 0. \end{cases} \quad (4.5.10)$$

Математическое ожидание и дисперсия процесса ε_i определяется соотношениями:

$$M(\varepsilon_i) = 0; \quad \sigma_{\varepsilon_i}^2 = D(\varepsilon_i) = \frac{\sigma^2}{\frac{1+\mu_2}{1-\mu_2} \left[(1-\mu_2)^2 - \mu_1^2 \right]}, \quad (4.5.11)$$

а значение коэффициента автокорреляции определяется выражениями:

$$\rho(1) = \frac{\mu_1}{1-\mu_2}; \quad \rho(2) = \mu_2 + \frac{\mu_1^2}{1-\mu_2} \quad (4.5.12)$$

$$\rho(l) = \mu_1 \rho(l-1) + \mu_2 \rho(l-2), \quad l = 3, 4, \dots$$

Условие стационарности процесса ε_i имеют вид:

$$\begin{cases} -1 < \frac{\mu_1}{1-\mu_2} < 1, \\ -1 < \mu_2 + \frac{\mu_1^2}{1-\mu_2} < 1. \end{cases} \Rightarrow \begin{cases} |\mu_1| < 2, \\ \mu_2 < 1 - |\mu_1|. \end{cases}$$

Оценки параметров μ_1, μ_2, σ^2 модели AR(2) находятся по формулам:

$$\hat{\mu}_1 = \frac{r(1)(1-r(2))}{1-r^2(1)}; \quad (4.5.13)$$

$$\hat{\mu}_2 = \frac{r(2) - r^2(1)}{1-r^2(1)} \quad (4.5.14)$$

$$\hat{\sigma}^2 = s_e^2 \cdot \frac{1+\hat{\mu}_2}{1-\hat{\mu}_1} \cdot \left[(1-\hat{\mu}_2)^2 - \hat{\mu}_1^2 \right] \quad (4.5.15)$$

где

$$r(l) = \frac{1}{n-l} \sum_{i=1}^{n-l} (e_i - \bar{e})(e_{i+l} - \bar{e}) / s_e^2, \quad l = 1, 2, \dots, \quad (4.5.16)$$

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i; \quad s_e^2 = \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2.$$

Модель скользящего среднего первого порядка (или модель MA(1)).

Модель определяется соотношением:

$$\varepsilon_i = \xi_i - \theta \xi_{i-1}, \quad (4.5.17)$$

где ξ_i - последовательность некоррелированных случайных величин с характеристиками:

$$M(\xi_i) = 0; \quad M(\xi_i \xi_{i+l}) = \begin{cases} \sigma^2, & \text{если } l = 0; \\ 0, & \text{если } l \neq 0. \end{cases}$$

Коэффициент автокорреляции процесса ε_i равен

$$\rho(l) = \begin{cases} -\frac{\theta}{1+\theta^2}, & \text{если } l = 1; \\ 0, & \text{если } l \geq 2, \end{cases}$$

а дисперсия процесса ε_i

$$\sigma_{\varepsilon_i}^2 = D(\varepsilon_i) = (1+\theta^2) \cdot \sigma^2.$$

Процесс стационарен при любом значении θ .

Оценка параметра θ равна одному из корней θ_1, θ_2 квадратного уравнения

$$\theta^2 + \frac{1}{r(1)}\theta + 1 = 0, \quad (4.5.18)$$

который меньше 1 ($\theta_1 \cdot \theta_2 = 1$), где $r(1)$ вычисляется по формуле (4.5.16) при $l=1$. Такое значение корня обозначим $\hat{\theta}$. Оценка дисперсии вычисляется по формуле

$$\hat{\sigma}^2 = \frac{s_e^2}{1 + \hat{\theta}^2}, \quad (4.5.19)$$

где $s_e^2 = \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2$, $\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i$.

Идентификация параметров моделей в Excel. Из приведенных выше выражений видно, что в оценки параметров моделей входят три величины: $\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i$, $s_e^2 = \frac{1}{n} \sum_{i=1}^n (e_i - \bar{e})^2$,

$\text{cov}(l) = \frac{1}{n-l} \cdot \sum_{i=1}^{n-l} (e_i - \bar{e})(e_{i+l} - \bar{e})$. Значения этих величин удобно вычислять, используя стандартные функции Excel.

Для конкретности предположим, что значения остатков e_i находятся в диапазоне ячеек A3:A22 (т.е. $n = 20$). Тогда

$$\bar{e} = \text{СРЗНАЧ}(A3:A22), \quad (4.5.20)$$

$$s_e^2 = \text{ДИСПР}(A3:A22), \quad (4.5.21)$$

$$\text{cov}(l) = \text{КОВАР}(A3:A22-l; A3+l:A22). \quad (4.5.22)$$

Подставляя вычисленные значения в соответствующие формулы, получаем нужные оценки параметров моделей.

Пример 4.5.1. В ячейках A3:A22 документа Excel, приведенного на рис. 4.16 находятся значения e_i , соответствующие авторегрессионной модели первого порядка (4.5.4) с параметрами $\mu = 0.4$, $\sigma^2 = 0.144$. Эти значения показаны также на графике в документе Excel (см. рис. 4.16). Необходимо по приведенной выборке оценить значения параметров модели и сравнить их с известными параметрами.

Решение. Будем считать, что $e_i = \varepsilon_i$. Тогда, используя выражения (4.5.20) ÷ (4.5.22), вычисляем значения $\bar{e} = 0.061$; $s_e^2 = 0.235$; $\text{cov}(1) = 0.101$, в ячейках D3, D5, D8 соответственно. Затем, подставляя эти значения в формулы (4.5.7), (4.5.8), находим оценки параметров $\hat{\mu} = 0.429$; $\hat{\sigma}^2 = 0.192$. Сравнивая эти значения с «точными» параметрами $\mu = 0.4$, $\sigma^2 = 0.144$, отмечаем не высокую точность оценивания (порядка 20 ÷ 30%). Это объясняется маленьким объемом выборки $n = 20$. ☹

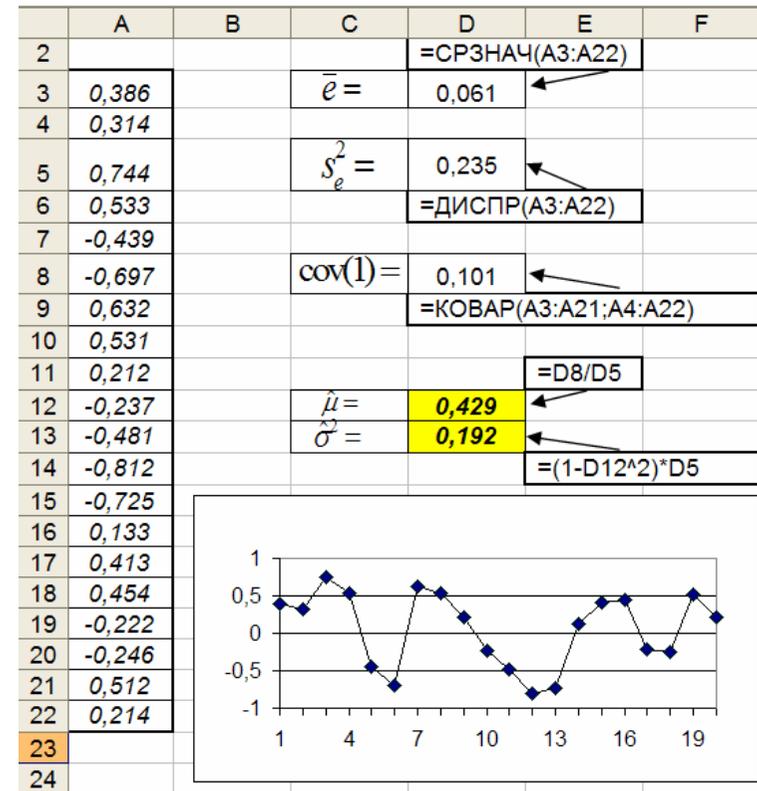


Рис. 4.16. Оценивание параметров модели $AR(1)$

4.6 Выделение тренда временного ряда обобщенным методом наименьших квадратов

При моделировании реальных экономических процессов условия классической линейной модели P2, P3 (см. параграф 3.1) оказываются нарушенными. В частности, могут не выполняться предположения о том, что случайные возмущения ε_i имеют одинаковую дисперсию и не коррелированы между собой. Такая ситуация характерна для временных рядов с коррелированными возмущениями ε_i (см. параграф 4.5). Использование обычного метода наименьших квадратов приводит к неэффективным оценкам коэффициентов функции регрессии, т.е. к оценкам, имеющим не минимальную дисперсию. Возникает вопрос: “Можно ли и в этих случаях построить эффективные оценки?”. Для ответа на этот вопрос первоначально рассмотрим новую модель множественной регрессии.

Обобщенная линейная модель множественной регрессии. Пусть модель наблюдений линейной множественной регрессии имеет вид:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (4.6.1)$$

или в матричном виде

$$y = X\beta + \varepsilon, \quad (4.6.2)$$

где X, y, β, ε - матрица и векторы, определенные в параграфе 3.1. Относительно X, ε сделаем следующие предположения:

P1. Матрица X размером $n \times (k+1)$ является не случайной матрицей.

P2. Вектор ε имеет нулевое среднее, т.е. $M(\varepsilon) = \bar{0}$, где $\bar{0}$ - нулевой вектор.

P3. Ковариационная матрица $V_\varepsilon = M(\varepsilon\varepsilon^T)$ вектора ε не является диагональной (напомним, что не диагональные элементы μ_{ij} характеризуют степень коррелированности случайных величин ε_i и ε_j).

P4. Ранг матрицы X $rank(x) = p + 1 \leq n$.

Множественная регрессия (4.6.1), удовлетворяющая предположениям P1 ÷ P4 получила название *обобщенной линейной модели* множественной регрессии. Сравнивая эту модель с классической (см. параграф 3.1), видим, что она отличается от классической только видом ковариационной матрицы (в классической $V_\varepsilon = \sigma^2 I$, где I - единичная матрица размера $n \times n$).

Оценка $b = (X^T X)^{-1} X^T y$ для обобщенной линейной модели также является несмещенной, состоятельной, но не эффективной. Ковариационная матрица V_b этой оценки равна

$$V_b = (X^T X)^{-1} X^T V_\varepsilon X (X^T X)^{-1}, \quad (4.6.3)$$

в то время как для классической модели

$$V_b = \sigma^2 (X^T X)^{-1}.$$

Для получения эффективной оценки будем использовать другую оценку, получаемую так называемым *обобщенным методом наименьших квадратов*.

Обобщенный метод наименьших квадратов. Можно показать, что при выполнении предположений P1 ÷ P4 обобщенной линейной множественной регрессии линейная оценка

$$b^* = (X^T V_\varepsilon^{-1} X)^{-1} X^T V_\varepsilon^{-1} y. \quad (4.6.4)$$

имеет наименьшую ковариационную матрицу среди всех линейных несмещенных оценок для вектора коэффициентов β . Другими словами, дисперсии оценок $b_j^*, j = 0, 1, \dots, k$ минимальны, т.е. вектор b^* является эффективной оценкой вектора коэффициентов β модели обобщенной линейной множественной регрессии. Ковариационная матрица оценки b^* имеет вид (сравнить с выражением (4.6.3)):

$$V_{b^*} = (X^T V_\varepsilon^{-1} X)^{-1}. \quad (4.6.5)$$

Отметим, что оценка b^* минимизирует функционал

$$F^*(b) = (y - Xb)^T V_\varepsilon^{-1} (y - Xb), \quad (4.6.6)$$

называемый функционалом обобщенного метода наименьших квадратов.

Замечание 4.6.1. Для обобщенной линейной регрессионной модели, в отличие от классической, коэффициент детерминации, вычисленный по формуле (см. (3.4.8)):

$$R^2 = 1 - \frac{(y - XB^*)^T (y - XB^*)}{(y - \bar{y})^T (y - \bar{y})},$$

где \bar{y} - вектор размерности n , составленный из средних значений $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$, не является удовлетворительной мерой качества построенного уравнения регрессии. В общем случае R^2 можно выводить за пределы интервала $[0;1]$. Поэтому в обобщенной модели коэффициент детерминации R^2 может использоваться лишь как приближенная характеристика качества построенного уравнения регрессии.

Применение обобщенного метода наименьших квадратов требует знания элементов ковариационной матрицы V_ε , что в практике эконометрического моделирования встречается очень редко. Поэтому для практической реализации обобщенного МНК необходимо вводить дополнительные условия на структуру матрицы V_ε таким образом, чтобы элементы матрицы V_ε зависели от нескольких параметров. Пример такого задания V_ε рассматривается ниже.

Выделения тренда временного ряда на основе обобщенного метода наименьших квадратов. Вернемся к задаче выделения тренда $t(\tau)$ временного ряда, представленного моделью (см. параграф 4.2):

$$y(\tau_i) = t(\tau_i) + \varepsilon(\tau_i), \quad i = 1, 2, \dots, n. \quad (4.6.7)$$

В качестве уравнения тренда примем полином p -го порядка вида:

$$t(\tau) = \beta_0 + \beta_1 \tau + \beta_2 \tau^2 + \dots + \beta_p \tau^p \dots \quad (4.6.8)$$

Тогда, введя новые переменные

$$x_1 = \tau, \quad x_2 = \tau^2, \quad \dots, \quad x_p = \tau^p,$$

приходим к следующей модели для измеренных значений временного ряда:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (4.6.9)$$

Далее предполагает, что возмущения ε_i имеют нулевое математическое ожидание, но коррелированы между собой, т.е. ковариационная матрица V_ε не является диагональной. Таким образом, приходим к модели обобщенной линейной множественной регрессии. В рамках этой модели несмещенная, состоятельная и эффективная оценка коэффициентов $\beta_0, \beta_1, \dots, \beta_p$ определяется на основе обобщенного МНК и имеет вид (4.6.4). Но для построения такой оценки необходимо задать ковариационную матрицу V_ε . Определим матрицу V_ε для двух следующих моделей возмущений ε_i .

Модель авторегрессии первого порядка. Определяется соотношением (4.5.4). С учетом числовых характеристик возмущений ε_i (см. (4.5.5)) получаем следующую ковариационную матрицу V_ε вектора возмущений ε :

$$V_\varepsilon = \sigma_\varepsilon^2 \begin{pmatrix} 1 & \mu & \mu^2 & \dots & \mu^{n-1} \\ \mu & 1 & \mu & \dots & \mu^{n-2} \\ \mu^2 & \mu & 1 & \dots & \mu^{n-3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu^{n-1} & \mu^{n-2} & \mu^{n-3} & \dots & 1 \end{pmatrix} = \sigma_\varepsilon^2 \cdot \Omega, \quad (4.6.10)$$

где матрица Ω имеет элементы $[\Omega]_{i,j} = \mu^{|i-j|}$

$$\sigma_\varepsilon^2 = \frac{\sigma^2}{1 - \mu^2}. \quad (4.6.11)$$

Можно показать, что обратная матрица Ω^{-1} имеет трехдиагональную структуру:

$$\Omega^{-1} = \frac{1}{1 - \mu^2} \cdot \begin{vmatrix} 1 & -\mu & 0 & \dots & 0 & 0 \\ -\mu & (1 + \mu^2) & -\mu & \dots & 0 & 0 \\ 0 & -\mu & (1 + \mu^2) & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & (1 + \mu^2) & -\mu \\ 0 & 0 & 0 & \dots & -\mu & \mu \end{vmatrix} \quad (4.6.12)$$

и тогда V_ε^{-1} определяется как

$$V_\varepsilon^{-1} = \frac{1}{\sigma_\varepsilon^2} \cdot \Omega^{-1}. \quad (4.6.13)$$

Тогда оценка b^* обобщенного метода наименьших квадратов является решением следующих систем уравнений

$$(X^T \Omega^{-1} X) b = X^T \Omega^{-1} y,$$

или

$$b^* = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} y. \quad (4.6.14)$$

Видимо, что для вычисления вектора b^* достаточно знания параметра μ , а для вычисления ковариационной матрицы V_{b^*} , определяемой соотношением

$$V_{b^*} = \sigma_\varepsilon^2 (X^T \Omega^{-1} X)^{-1},$$

необходимо знание дисперсии σ^2 авторегрессионной модели. В большинстве практических случаев параметры μ, σ^2 априори неизвестны и их приходится оценивать в процессе построения уравнения регрессии, используя следующую итерационную процедуру (обозначенную *IPAR(1)*).

Итерационная процедура *IPAR(1)*. Включает следующие шаги

Шаг 0. По заданной выборке вычисляется оценка простого метода наименьших квадратов

$$\mathcal{B}^{(0)} = (X^T X)^{-1} X^T y$$

и полагаем номер итерации $l = 0$.

Шаг 1. Вычисляется вектор невязки на l -ой итерации:

$$e^{(l)} = y - X \mathcal{B}^{(l)}.$$

Шаг 2. По формулам (4.5.7) находим оценку $\mathcal{K}^{(l)}$ на l -ой итерации и формируем матрицу $\Omega^{(l)}$ согласно (4.6.12).

Шаг 3. Вычисляем вектор

$$\mathcal{B}^{(l+1)} = (X^T (\Omega^{(l)})^{-1} X)^{-1} X^T (\Omega^{(l)})^{-1} y$$

и полагаем $l = l + 1$.

Шаги 1-3 повторяют до тех пор пока различие между $\mathcal{K}^{(l)}$ и $\mathcal{K}^{(l+1)}$ будут малы. Значение $\mathcal{K}^{(l+1)}$, при котором итерационная процедура закончилась, обозначим \mathcal{K} .

Шаг 4. По найденному значению \mathcal{K} вычисляем \mathcal{E}^2 (формула (4.5.8)), матрицу \mathcal{G} (формула (4.6.12)), $\mathcal{E}_\varepsilon^2$ (формула (4.6.11)) и находим вектор

$$\mathcal{E}^* = (X^T \mathcal{G}^{-1} X)^{-1} X^T \mathcal{G}^{-1} y; \quad (4.6.14)$$

и ковариационную матрицу

$$\mathcal{E}_{b^*} = \mathcal{E}_\varepsilon^2 \cdot (X^T \mathcal{G}^{-1} X)^{-1} \quad (4.6.15)$$

Заметим, что часто ограничивается только одной или двумя итерациями описанной процедуры (как это сделано в примере 4.6.1).

Модель скользящего среднего первого порядка. Определяется соотношением (4.5.19). Ковариационная матрица V_ε вектора возмущений ε имеет вид

$$V_\varepsilon = \sigma_\varepsilon^2 \cdot \begin{pmatrix} 1 & -\lambda & 0 & \dots & 0 \\ -\lambda & 1 & -\lambda & \dots & 0 \\ 0 & -\lambda & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & -\lambda \\ 0 & 0 & 0 & -\lambda & 1 \end{pmatrix}, \quad (4.6.16)$$

где $\sigma_\varepsilon^2 = (1+\theta) \cdot \sigma^2$, $\lambda = \frac{\theta}{1+\theta^2}$. Матрица V_ε имеет трехдиагональную структуру, и ее элементы зависят от параметров θ, σ^2 модели скользящего среднего. Если эти параметры априори неизвестны, то для их оценивания можно использовать итерационную процедуру, аналогичную описанной *IPAR(1)*. Ограничимся только записью одной итерации такой процедуры:

Шаг 0. По заданной выборке вычисляется оценка

$$\hat{\beta} = (X^T X)^{-1} X^T y;$$

Шаг 1. Вычисляется вектор невязки $e = y - X\hat{\beta}$.

Шаг 2. По формулам (4.5.18), (4.5.19) вычисляются оценки $\hat{\theta}, \hat{\sigma}^2$ и формируются элементы матрицы \mathcal{V}_ε по формуле

$$\{\mathcal{V}_\varepsilon\}_{i,j} = (1 + \hat{\theta}^2) \cdot \hat{\sigma}^2 \cdot \psi_{i,j}, \quad (4.6.17)$$

$$\text{где } \psi_{i,j} = \begin{cases} 1, & \text{если } i = j; \\ -\frac{\hat{\theta}}{1 + \hat{\theta}^2}, & \text{если } |i - j| = 1; \\ 0, & \text{во всех остальных случаях.} \end{cases}$$

Шаг 3. Подставляя матрицу \mathcal{V}_ε в (4.6.4), получаем оценку

$$\hat{\beta} = (X^T \mathcal{V}_\varepsilon^{-1} X)^{-1} X^T \mathcal{V}_\varepsilon^{-1} y$$

для вектора β коэффициентов полиномиального тренда (4.6.8).

Ковариационная матрица вектора $\hat{\beta}$ определяется выражением:

$$V_{\hat{\beta}} = (X^T \mathcal{V}_\varepsilon^{-1} X)^{-1}. \quad (4.6.18)$$

Пример 4.6.1. В ячейках D3:D17 документа Excel (см. рис.4.17) приведены значения временного ряда $y_i, i=1,2,\dots,15$ (т.е. $n=15$). Предполагается, что возмущения ε_i соответствуют авторегрессионной модели первого порядка (4.5.7). Необходимо выделить трендовую составляющую временного ряда, аппроксимируя ее полиномом второй степени

$$\hat{\mathcal{K}}(\tau) = b_0 + b_1\tau + b_2\tau^2.$$

	A	B	C	D	E	F	G
1	=МУМНОЖ(ТРАНСП(A3:C17);МУМНОЖ(МОБР(ИЗ:W17);A3:C17))						
2							
3	1	1	1,000	2,113	52,310	125,544	345,605
4	1	1,2	1,440	2,908	125,544	347,631	1051,811
5	1	1,4	1,960	3,119	345,605	1051,811	3382,906
6	1	1,6	2,560	3,567			
7	1	1,8	3,240	3,401			
8	1	2	4,000	3,284	0,893	-0,784	0,152
9	1	2,2	4,840	2,706	-0,784	0,737	-0,149
10	1	2,4	5,760	3,144	0,152	-0,149	0,031
11	1	2,6	6,760	2,707			
12	1	2,8	7,840	2,339	=МУМНОЖ(E8:G10;E14:E16)		
13	1	3	9,000	2,348			
14	1	3,2	10,240	2,225	141,800	$\hat{b}^* =$	1,486
15	1	3,4	11,560	2,616	332,162		1,434
16	1	3,6	12,960	2,664	886,888		-0,335
17	1	3,8	14,440	2,241			
18							
19	=МУМНОЖ(ТРАНСП(A3:C17);МУМНОЖ(МОБР(ИЗ:W17);D3:D17))						

Рис. 4.17. Вычисление коэффициентов обобщенным МНК

Решение. Введем новые переменные $x_1 = \tau$; $x_2 = \tau^2$ и сформируем матрицу X размером 15×3 , элементы которой находятся в ячейках $A3:C17$ (см. рис. 4.17). Выполним одну итерацию итерационной процедуры *IPARI* и вычислим оценки параметров авторегрессионного процесса (4.5.4): $\hat{\rho} = 0.403$, $\hat{\sigma}^2 = 0.106$. Этот шаг в документе на рис. 4.17 не отражен из-за недостатка места. Сформируем матрицу V_ε , элементы которой вычисляются по формуле

$$\{V_\varepsilon\}_{i,j} = \frac{\hat{\sigma}^2}{1 - \hat{\rho}^2} \cdot \hat{\rho}^{|i-j|}, \quad i, j = 1, 2, \dots, n.$$

и разместим их в диапазоне $I3:W17$ (на рисунке не показаны). Затем вычисляем: матрицу $X^T V_\varepsilon^{-1} X$ размером 3×3 (ячейки $E3:G5$), обратную матрицу $(X^T V_\varepsilon^{-1} X)^{-1}$ размером 3×3 (ячейки $E8:G10$), вектор $X^T V_\varepsilon^{-1} y$, содержащий три проекции (ячейки $E14:E16$). Далее по формуле

$$\hat{\beta} = (X^T V_\varepsilon^{-1} X)^{-1} X^T V_\varepsilon^{-1} y$$

вычислим вектор B (ячейки $G14:G16$)

$$b = \begin{vmatrix} b_0 \\ b_1 \\ b_2 \end{vmatrix} = \begin{vmatrix} 1.486 \\ 1.434 \\ -0.336 \end{vmatrix}$$

Таким образом, уравнение трендовой составляющей принимает вид:

$$\hat{f}(\tau) = 1.486 + 1.434\tau - 0.336\tau^2$$

На рис. 4.18 показаны исходные значения y_i (кривая 1 отмечена квадратными маркерами) и значения $\hat{f}(\tau_i)$ (кривая 2 - треугольные маркеры).

Заметим, что применение простого метода наименьших квадратов приводит к уравнению:

$$\hat{f}(\tau) = 2.674 + 0.118\tau - 0.010\tau^2,$$

коэффициенты которого существенно отличаются от коэффициентов, найденных обобщенным методом наименьших квадратов.

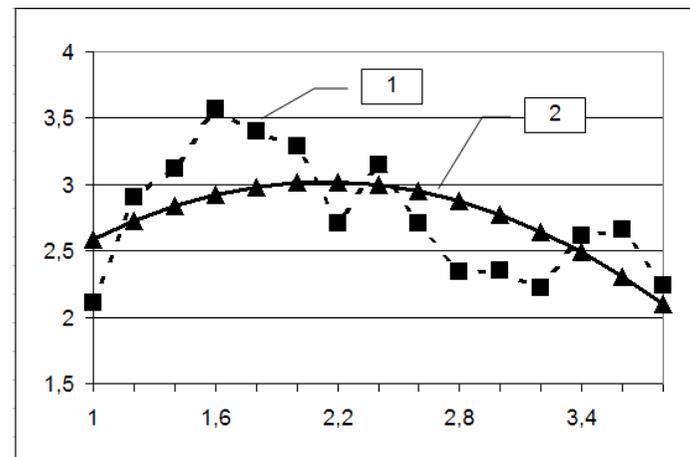


Рис. 4.18. Исходные данные и трендовая составляющая

В заключении этого параграфа отметим, что использование обобщенного метода наименьших квадратов при значительной корреляции возмущений ε_i (например, $\rho_{\varepsilon_i \varepsilon_{i+1}} > 0.4$) позволяет более точно оценить коэффициенты уравнения трендовой составляющей временного ряда.

ЛАБОРАТОРНАЯ РАБОТА № 4.1

Выделение трендовой составляющей временного ряда

Цель работы. Используя табличный процессор Excel и команду *Добавить линию тренда*, определить «наилучшую» модель для трендовой составляющей временного ряда, выборочные значения которого представлены в таблице Л4.1. Переменная Y

определяет темпы роста номинальной заработной платы (за 10 месяцев 1999 года в % к уровню декабря 1998 года). Проверить на наличие корреляции у возмущений временного ряда.

Исходные данные. В таблице Л4.1 приведены значения временного ряда в зависимости от номера месяца.

Таблица Л4.1

Номер месяца	Месяц	y_i
1	Январь	82,9
2	Февраль	87,3
3	Март	99,4
4	Апрель	104,8
5	Май	107,2
6	Июнь	121,6
7	Июль	118,6
8	Август	114,1
9	Сентябрь	123,0
10	Октябрь	127,3

Содержание работы

1. Ввести в лист Excel первый и третий столбец таблицы Л4.1.
2. Построить по данным этих столбцов диаграмму рассеяния (см. пример 4.2.3).
3. Используя команду *Добавить линию тренда* вычислите коэффициенты уравнений $\hat{f}(\tau)$ трендовой составляющей, название которых указаны в таблице Л4.2 и значения коэффициента детерминации R^2 (см. пример 4.2.3). Занесите полученные уравнения и значения коэффициента детерминации в соответствующие ячейки таблицы Л4.2.
4. Для наилучшей модели трендовой составляющей, имеющей максимальное значение R^2 вычислите остатки $e_i = y_i - \hat{f}(\tau_i), i = 1, \dots, 10$.
5. Используя тест Дарбина-Уотсона (см. параграф 4.5) проверить наличие корреляции у остатков e_i

6. Сделать вывод о наличии корреляции у возмущений ε_i временного ряда.

7. По наилучшей модели сделать прогноз темпов роста номинальной заработной платы на ноябрь и декабрь месяцы.

Таблица Л4.2

Название модели тренда	Формула уравнения тренда	R^2
Линейная		
Квадратичная		
Степенная		
Экспоненциальная		

Контрольные результаты:

1. Наилучшей является степенная модель тренда

$$\hat{f}(\tau) = 80.344 \cdot \tau^{0.1935}$$

с коэффициентом детерминации $R^2 = 0.939$.

2. Прогнозируемые значения

$$\hat{f}(11) = 127.78, \quad \hat{f}(12) = 129.95.$$

3. Возмущения временного ряда не коррелированы.

КОНТРОЛЬНАЯ РАБОТА № 4.1 Построение модели временного ряда

В таблице К4.1 приведены данные, отражающие динамику дохода некоторой компании (в млрд. долл. США) за 14 месяцев (i - номер месяца).

Таблица К4.1

i	1	2	3	4	5	6	7
y_i	97	116	104	90	95	72	75
i	8	9	10	11	12	13	14
y_i	101	97	81	82	111	138	142

Требуется:

1. Построить диаграмму значений временного ряда и оценить по ней общую тенденцию изменения значений ряда.

2. Приняв в качестве модели выражение (4.2.1), где $\tau_i = i$ построить оценку $\hat{f}(\tau)$ для тренда исследуемого временного ряда в виде

$$\hat{f}(\tau) = \hat{f}(\tau) = b_0 + b_1 \cdot \tau + b_2 \cdot \tau^2. \quad (Л4.1)$$

Коэффициенты определить методом наименьших квадратов (см. пример 4.4.1)

3. Вычислить коэффициент детерминации и проверить значимость построенной регрессионной модели временного ряда при уровне значимости $\alpha = 0.05$.

4. Приняв в качестве модели модель авторегрессии первого порядка (4.4.2) оценить коэффициенты уравнения

$$\hat{f}_i = b_0 + b_1 \cdot y_{i-1} \quad (Л4.2)$$

Коэффициенты определить методом наименьших квадратов (см. пример 4.4.1)

5. Вычислить коэффициент детерминации и проверить значимость построенной авторегрессионной модели временного ряда при уровне значимости $\alpha = 0.05$.

6. Используя *Мастера диаграмм* табличного процессора Excel представить на одном рисунке три графика: исходных значений y_i , значений функции (Л4.1) при $\tau_i = i$ и значений функции (Л4.2) при $\tau_i = i$. Из анализа рисунка выбрать наилучшую из двух построенных моделей временного ряда.

7. По выбранной наилучшей модели сделать прогноз для $i = 15, 16$.

КОНТРОЛЬНЫЕ ВОПРОСЫ

1. Какие составляющие может включать детерминированная компонента временного ряда?

2. Каким условиям должны удовлетворять числовые характеристики стационарного ряда?

3. Какие методы используются для выделения трендовой составляющей временного ряда?

4. Какой метод используется для выделения тригонометрических составляющих временного ряда?

5. В чем отличие авторегрессионной модели временного ряда от обычной регрессионной модели?

6. Какие модели используются для описания коррелированных возмущений временного ряда?

7. В чем отличие обобщенного метода наименьших квадратов от обыкновенного (классического) МНК?

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. **Тимошенко Е. И.** Теория вероятностей : учеб. пособие / Е. И. Тимошенко, Ю. Е. Воскобойников ; Новосиб. гос. архитектур.-строит. ун-т. – Новосибирск : НГАСУ, 2003 (электр. версия - <http://www.ngasu.nsk.su/prikl/terver.html>).
2. **Воскобойников Ю. Е.** Математическая статистика : учеб. пособие / Ю. Е. Воскобойников, Е. И. Тимошенко ; Новосиб. гос. архитектур.-строит. ун-т. – Новосибирск : НГАСУ, 2000 (электр. версия - <http://www.ngasu.nsk.su/prikl/stat2000.html>).
3. **Гмурман В. Е.** Теория вероятностей и математическая статистика / В. Е. Гмурман. – М.: Высшая школа, 1998.
4. **Воскобойников Ю. Е.** Эконометрика в Excel : учеб. пособие / Ю. Е. Воскобойников. Новосиб. гос. архитектур.-строит. ун-т. – Новосибирск : НГАСУ, 2005.
5. **Кремер Н. Ш.** Эконометрика / Н. Ш. Кремер, Б. А. Путко. – М. : ЮНИТИ, 2002.
6. **Айвазян С. А.** Прикладная статистика и основы эконометрики / Айвазян С. А., В. С. Мхитарян. – М. : ЮНИТИ, 1998.
7. **Минус Я. Р.** Эконометрика. Начальный курс / Я. Р. Минус, Л. К. Катышев, А. А. Пересецкий. – М. : Дело, 2000.
8. **Эконометрика** : под ред. Н. И. Елисейевой. – М. : Финансы и статистика, 2001.
9. **Арженовский С. В.** Эконометрика : учеб. пособие / С. В. Арженовский, О. Н. Федосова. – Ростов н/Д, 2002.

10. **Тихомиров Н. П.** Эконометрика / Н. П. Тихомиров, Е. Ю. Дорохина. – М. : Экзамен, 2003.
11. **Макарова Н. В.** Статистика в EXCEL : учеб. пособие / Н. В. Макарова, В. Я. Трофимец. – М. : Финансы и статистика, 2002.

ПРИЛОЖЕНИЕ

СТАТИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ В ЭКОНОМЕТРИЧЕСКОМ АНАЛИЗЕ

В общем случае процесс моделирования можно определить как замещение *исследуемого объекта (оригинала)* его условным образом, описанием или другим объектом, именуемым *моделью*. Очевидно, что такая модель должна обеспечивать адекватное с оригиналом поведение в рамках некоторых допущений и приемлемых погрешностей. Основной целью моделирования является получение исследователем определенных характеристик, которые адекватны аналогичным характеристикам исследуемого объекта, т.е. характеристикам оригинала.

В *математическом моделировании* в качестве модели выступает совокупность математических соотношений (алгебраические выражения, дифференциальные, интегральные уравнения и т.д.). *Статистическое моделирование* является частным (но весьма распространенным) случаем математического моделирования, когда исследуемые характеристики вычисляются как средние значения по некоторой выборке случайных значений этих характеристик. Такой подход к определению характеристик также называют *методом Монте-Карло*.

Приведем пример использования статистического моделирования для оценки точности вычисления вектора коэффициентов b множественной линейной регрессии. Напомним, что вектор коэффициентов b , вычисляется из системы нормальных уравнений

$$X^T X b = X^T y$$

и является случайным вектором (следствие стохастичности вектора y). Поэтому ошибка оценивания вектора β с использованием

вектора b , определяемая нормой $\Delta(b) = \|b - \beta\| = \sqrt{\sum_{i=1}^k (b_i - \beta_i)^2}$

также является случайной величиной. В качестве неслучайной характеристики этой величины принимают ее математическое ожидание $M[\Delta(b)]$. Для оценивания математического ожидания необходимо вычислить среднее значение по некоторой выборке значений $\Delta(b)$. Для получения такой выборки нужно сгенерировать выборку случайных векторов y , которые отличаются друг от друга реализациями вектора возмущений ε . Затем по каждому вектору y вычисляется свой вектор b и для этого вектора определяется свое значение $\Delta(b)$. По полученной таким образом выборке значений $\Delta(b)$ вычисляется выборочное среднее.

В общем случае статистическое моделирование в регрессионном анализе позволяет оценить характеристики построенных уравнений регрессий, которые невозможно вычислить по аналитическим выражениям.

Статистическое моделирование применимо к задачам регрессионного анализа и включает следующие этапы:

Этап 1. Задание аналитического выражения для объясненной части $f(x)$, которое зависит от коэффициентов $\beta_0, \beta_1, \dots, \beta_k$.

Этап 2. Вычисление вектора \hat{y} , составленного из значений объясненной части при $x = x_i, i = 1, 2, \dots, n$. Очевидно, что вектор \hat{y} есть значения зависимой переменной Y в отсутствии возмущений ε .

Этап 3. Задание объема выборки N_{sam} и генерирование N_{sam} случайных векторов $\varepsilon^{(m)}, m = 1, 2, \dots, N_{sam}$. Проекция $\varepsilon_i^{(m)}$ этих векторов есть псевдослучайные числа с заданным законом распределения и заданными числовыми характеристиками. Так, при выполнении условия гомоскедастичности модели (см. п. 2.1) числа $\varepsilon_i^{(m)}$ подчиняются нормальному распределению с нулевым математическим ожиданием и дисперсией σ^2 и $\varepsilon_i^{(m)}$ не коррелированы между собой и с проекциями вектора \hat{y} .

Для генерации псевдослучайных чисел в табличном процессоре Excel существует несколько возможностей:

– использование функции СЛЧИС() (категория функций *Математические*). При обращении к этой функции параметры не задаются и генерируется одно псевдослучайное число, равномерно распределенное в интервале $[0,1]$;

– использование модуля *Анализ данных*. Для вызова этого модуля обратиться к пункту **Сервис** и выполнить команду *Анализ данных*. Затем в появившемся окне в списке *Инструменты анализа* выбрать *Генерация случайных чисел* и щелкнуть на кнопке ОК. На экране появляется диалоговое окно (рис. П1) в котором задается объем выборки, вид распределения и числовые характеристики генерируемых псевдослучайных чисел.

Параметр *Число переменных* определяется объемом выборки, а параметр *Число случайных чисел* значение n . На рис. П1 показан пример задания параметров для генерирования псевдослучайных чисел подчиняющихся нормальному распределению с нулевым средним и $\sigma = 1$. Величина n равна 14. Заметим, что из закона «трех сигм» следует, что с вероятностью 0.999 генерируемые значения должны находиться внутри интервала $[-3,3]$ (убедитесь в этом). Кроме нормального распределения можно генерировать числа, имеющие равномерное, биномиальное распределение, распределение Пуассона и Бернулли.

Этап 4. Формирование векторов

$$y^{(m)} = \hat{y} + \varepsilon^{(m)}, m = 1, 2, \dots, N_{sam}, \quad (\text{П4})$$

проекции которых трактуются как измеренные (с погрешностью ε) значения зависимой переменной Y .

Этап 5. Вычисление по каждому вектору случайного значения исследуемой характеристики. В нашем примере – вычисление ошибки оценивания

$$\Delta^{(m)} = \| b^{(m)} - \beta \| = \sqrt{\sum_{i=1}^k (b_i^{(m)} - \beta_i)^2}, m = 1, 2, \dots, N_{sam}, \quad (\text{П5})$$

где $b^{(m)}$ – вектор коэффициентов, вычисленный по $y^{(m)}$.

Этап 6. Вычисление по сформированной выборке средних значений или других числовых характеристик. В нашем примере – это среднее значение

$$\bar{\Delta} = \frac{1}{N_{sam}} \cdot \sum_{m=1}^{N_{sam}} \Delta^{(m)}. \quad (\text{П6})$$

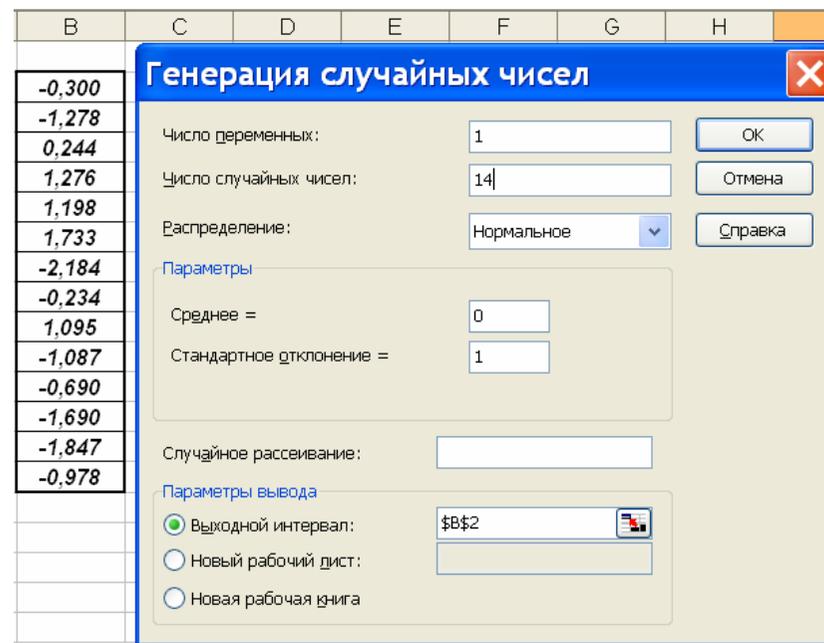


Рис. П1. Генерация псевдослучайных чисел

Проиллюстрируем описанные этапы статистического моделирования на следующем примере.

Пример П.1. Провести исследование точности вычисления коэффициентов β_0, β_1 степенной эконометрической модели

$$Y = \beta_0 \cdot X^{\beta_1} + \varepsilon$$

при заданном уровне случайной составляющей ε .

Решение. Очевидно, что уравнение регрессии имеет вид

$$y(x) = b_0 \cdot x^{b_1}$$

и вычисление коэффициентов будем осуществлять методом наименьших квадратов, т.е. из условия минимума функционала

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

Основные вычисления приведены в документе Excel, показанном на рис. П2. Поясним эти вычисления, следуя описанным выше этапам.

Зададим следующие значения коэффициентов: $\beta_0 = 10$, $\beta_1 = 0.3$ (ячейки B1:B2) и вычислим проекции вектора \hat{y} (этап 2):

$$\hat{y}_i = b_0 \cdot x_i^{b_1}, i = 1, \dots, n = 5$$

для $x_i = 1, 2, 3, 4, 5$ (ячейки B4:B8).

Из-за ограниченной ширины рис. П2 зададим объем выборки равным $N_{sam} = 3$ и, используя модуль *Анализ данных*, сгенерируем три случайных вектора $\varepsilon^{(m)}$, проекции которых распределены по нормальному закону с нулевым средним и $\sigma = 0.5$ (этап 3). Вектор $\varepsilon^{(1)}$ размещен в ячейках C4:C8, $\varepsilon^{(2)}$ – в ячейках D4:D8, $\varepsilon^{(3)}$ – в ячейках E4:E8. Заметим, что относительный уровень возмущений ε равен $\|\varepsilon\|/\|\hat{y}\| \approx 0.02$ или примерно 2 %.

В соответствии с выражением (П4) сформируем векторы $y^{(m)}$, $m = 1, 2, 3$ (этап 4), размещаемые в ячейках: F4:F8 – вектор $y^{(1)}$; G4:G8 – вектор $y^{(2)}$, H4:H8 – вектор $y^{(3)}$.

После этого для каждого вектора $y^{(m)}$ вычислим вектор коэффициентов $b^{(m)}$. Для этого используется команда *Поиск решения* (см. параграф 2.7), а сами вычисления программируются в Excel точно также как в примере 2.7 и поэтому здесь не поясняются. Найденные вектора $b^{(m)}$ размещаются в ячейках: B10:B11, D10:D11, G10:G11. Эти ячейки на рис. П2 выделены серым фоном. Затем по формуле (П5) вычисляются ошибки оценивания $\Delta^{(m)}$, размещенные в ячейках C20:C22 (этап 5).

	A	B	C	D	E	F	G	H	I
1	β_0	10				=B4+C4			
2	β_1	0.350							
3	x_i	\hat{y}_i	$\varepsilon^{(1)}$	$\varepsilon^{(2)}$	$\varepsilon^{(3)}$	$y^{(1)}$	$y^{(2)}$	$y^{(3)}$	
4	1	10,000	0,206	0,374	0,079	10,206	10,374	10,079	
5	2	12,746	-0,090	-0,313	-0,086	12,655	12,433	12,660	
6	3	14,689	-0,054	-0,227	0,406	14,635	14,462	15,095	
7	4	16,245	-0,052	-0,083	-0,285	16,193	16,162	15,960	
8	5	17,565	-0,161	-0,188	-0,005	17,403	17,376	17,560	
9	=B\$10*A13*\$B\$11		=(F4-B13)^2						
10		10,121		10,118			10,096		
11		0,337		0,333			0,343		
12	x_i	y_i	$(y_i^{(1)} - \hat{y}_i)^2$	x_i	y_i	$(y_i^{(2)} - \hat{y}_i)^2$	x_i	y_i	$(y_i^{(3)} - \hat{y}_i)^2$
13	1	10,121	0,007	1	10,118	0,066	1	10,096	0,000
14	2	12,781	0,016	2	12,745	0,097	2	12,801	0,020
15	3	14,650	0,000	3	14,587	0,016	3	14,708	0,150
16	4	16,139	0,003	4	16,054	0,012	4	16,231	0,074
17	5	17,398	0,000	5	17,292	0,007	5	17,521	0,002
18	=СУММ(C13:C17)		0,026			0,198			0,245
19									
20		Δ_1	0,122			=((B1-B10)^2+(B2-B11)^2)^(1/2)			
21		Δ_2	0,119			=((B1-D10)^2+(B2-D11)^2)^(1/2)			
22		Δ_3	0,096			=((B1-G10)^2+(B2-G11)^2)^(1/2)			
23	Средняя ошибка		0,112			=СРЗНАЧ(C20:C22)			

Рис. П2. Вычисление средней ошибки оценивания

По выборочным данным $\Delta^{(m)}$ и формуле (П6) вычисляется среднее значение $\bar{\Delta} = 0.112$ – ячейка C23 (этап 6). Заметим, что относительная ошибка оценивания коэффициентов β_0, β_1 равна $\bar{\Delta}/\|\beta\| \approx 0.011$ или примерно 1 %. Эта величина говорит о высокой точности вычисленных коэффициентов b_0, b_1 . Изменяя значения σ , можно определить влияние уровня возмущений ε на точность оценивания коэффициентов β_0, β_1 . ☉

Задание. Выполните самостоятельно такое исследование, задавая значения $\sigma = 1$ и $\sigma = 3$.