МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

НОВОСИБИРСКИЙ ГОСУДАРСТВЕННЫЙ АРИТЕКТУРНО-СТРОИТЕЛЬНЫЙ УНИВЕРСИТЕТ

Воскобойников Ю.Е.

ЭКОНОМЕТРИКА в EXCEL

Часть 1

(парный и множественный регрессионный анализ)

Учебное пособие

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

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

Новосибирск 2005

УДК 330.43(075.8)
ББК 65.вб.я73
В

Печатается по решению учебно-методического совета Новосибирского государственного архитектурно-строительного университета.

Рецензенты:
Заведующий кафедрой экономики труда и хозяйственной деятельности НГАВТ, д.э.н., профессор
А.С. Овсянников
Заведующий кафедрой экономики и инвестиций НГАСУ кандидат экономических наук, профессор
Т.А. Ивашенцева

Учебное пособие содержит основные теоретические положения по следующим разделам эконометрики: эконометрические модели и эконометрическое моделирование, парный и множественный регрессионный анализ. Приводятся необходимые расчетные соотношения. Большое внимание уделяется реализации этих соотношений в табличном процессоре Excel. Учебное пособие содержит большое количество примеров и копий 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.

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

ОГЛАВЛЕНИЕ

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, занимающейся 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.*

В настоящем пособии даются основные понятия, модели и методы эконометрики. Учебное пособие состоит из двух частей. Часть 1 содержит основные понятия и методы эконометрики для построения регрессионных моделей по пространственной выборке. Часть 2 включает методы построения моделей временных рядов и моделей, описываемых системой одновременных уравнений.

Излагаемый материал делится на основной и дополнительный. Основной материал (глава 2, 3) содержит решение задач эконометрики в «классических» постановках. Дополнительный материал (глава 4) затрагивает решение задач в постановках, отличных от «классических» (такие постановки чаще всего и встречаются на практике).

В качестве вычислительного инструментария используется табличный процессор Excel XP. В пособии приводятся фрагменты документов программы Excel, решающие ту или иную задачу. Это будет существенной помощью читателям пособия при выполнении заданий и решении реальных практических задач.

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 В1. 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:

$$\text{СРЗНАЧ}(x_1; x_2; ...; x_m),$$

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

=СРЗНАЧ(B3:B6;C3:C6) ,

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

Замечание B2. 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.

Изложение материала сопровождается большим числом примеров, а также заданий, решение которых служит закреплению теоретических положений изучаемой дисциплины. Эту же цель преследуют контрольные вопросы и учебные задания, приводимые в конце каждой главы пособия.

Содержание пособия полностью соответствует требованиям государственного образовательного стандарта высшего профессионального образования для специальностей направления «Экономика и менеджмент», в частности для специальностей: 060800-«Экономика и управление на предприятии (в строительстве)».

Глава 1. Основные аспекты эконометрического моделирования

В этой главе вводятся понятия эконометрической модели, эконометрического моделирования и рассматриваются основные этапы эконометрического моделирования.

1.1. Эконометрическая модель и эконометрическое моделирование

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*. Зависимые переменные в эконометрике называют *объясняемыми*, а независимые – *объясняющими*.

Цены, указанные в объявлениях, определяются не только значениями соответствующих независимых переменных, но и зависят от *случайных обстоятельств* – таких, например, как характер продавца, возможные сроки продажи, потребность продавца в конкретной сумме и т. д. Однако, в отличие от объясняющих переменных, которые «постоянно» формируют цену автомобиля, влияние отдельного случайного обстоятельства может присутствовать или отсутствовать в наблюдаемой цене конкретного автомобиля.

Таким образом, нами получена следующая *эконометрическая модель*

$$Y = f(X_1, X_2, ..., X_k) + \varepsilon \qquad (1.1.1)$$

в которой Y – наблюдаемое в рекламных объявлениях значение цены автомобиля, X_1, X_2, …, X_k – набор независимых переменных, т. е. факторов, влияющих на цену автомобиля, ε - *случайная величина*, отражающая влияние случайных обстоятельств на цену автомобиля и в дальнейшем называемую *случайным возмущением* или *случайной ошибкой* модели. Запись $f(X_1, X_2, ..., X_k)$ означает некоторую функцию от k аргументов и эту функцию называют *объясненной частью эконометрической модели*.

В дальнейшем сама переменная обозначается прописной (большой) буквой, а ее конкретные значения строчной (маленькой) буквой.

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

В нашем случае таковым экономическим объектом является рынок автомобилей.

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

Определив эконометрическую модель процесса формирования цены автомобиля, мы еще не получили ответа на вопрос: какую же цену назначить при продаже нашего автомобиля?

Обратимся к модели (1.1.1). Так как влияние возмущения ε носит случайный характер, то нам необходимо построить некоторое аналитическое выражение для объясненной части эконометрической модели $f(X_1, X_2, \ldots, X_k)$ и это выражение не должно содержать возмущения ε.

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

$$Y = M(Y \mid x_1, x_2, ..., x_k) + \varepsilon. \qquad (1.1.2)$$

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

$$f(x_1, x_2, ..., x_k) = M(Y \mid x_1, x_2, ..., x_k) \qquad (1.1.3)$$

называют *функцией регрессии*.

С математической точки зрения регрессионная модель (1.1.2) оказывается более простым объектом, чем эконометрическая модель общего вида (1.1.1). Остановимся на одном важном свойстве регрессионной модели. Возьмем условное математическое ожидание от обеих частей (1.1.2):

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

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

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

Тогда из предыдущего равенства следует важное условие к возмущению регрессионной модели

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

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

Предположим, что в нашем примере продажи автомобиля в качестве объясняющих переменных были приняты: X_1 – срок эксплуатации автомобиля (в годах); X_2 - пробег автомобиля (в тыс. км) и получена функция регрессии вида:

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

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.

Этапы, связанные с построением эконометрической модели, оцениванием параметров модели, проверкой ее работоспособности и составляет содержание эконометрического моделирования.

1.2. Типы данных и типы эконометрических моделей

Выборка наблюдений зависимой переменной Y и объясняющих X_j (j = 1,2,…, k) является отправной точкой любого эконометрического исследования. В курсе эконометрики рассматриваются следующие типы выборочных (часто называемых экспериментальных) данных.

Пространственная выборка (или пространственные данные). Предположим, что эконометрическая модель включает величины Y, X_1, X_2, …,X_k, над которыми выполнены n наблюдений (как правило, над n-объектами). Результаты наблюдений могут быть представлены таблицей

$$\begin{vmatrix} x_{11} & x_{12} \dots & x_{1k} & y_1 \\ x_{21} & x_{22} \dots & x_{2k} & y_2 \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} \dots & x_{nk} & y_n \end{vmatrix} \qquad (1.2.1)$$

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 (1.2.1) не существенен.

Например, данные, полученные из рекламных изданий о цене (Y) и параметрах (X_1, X_2, …,X_k) n автомобилей.

Временная выборка (или временной ряд). Временным рядом (в зарубежной литературе – time-series data) называется выборка наблюдений, в которых важен порядок следования наблюдаемых значений. Чаще всего такая упорядоченность обусловлена тем, что экспериментальные данные представляют собой серию наблюдений одной и той же случайной величины в последовательные моменты времени.

Например, взяты n выпусков некоторого рекламного изделия, и они упорядочены по дате выпуска. Из каждого выпуска была взята цена подобного автомобиля. В этом случае мы имеем временной ряд, составленный из наблюдаемых значений $y(t_1)$, $y(t_2)$, …, $y(t_n)$, где t_i – время выхода i-го выпуска рекламного издания при этом $t_i < t_{i+1}$.

Рассмотренные типы данных в определенной степени обуславливают *типы следующих эконометрических моделей*.

Регрессионные модели с одним уравнением. В таких моделях зависимая переменная Y представляется в виде функции

$$Y = f(X_1, X_2, \dots, X_k, \beta_1, \beta_2, \dots, \beta_m) + \varepsilon = f(X, \beta) + \varepsilon,$$

где X_1, X_2, …, X_k – *независимые (объясняющие) переменные*, β_1, β_2, …,β_m – *коэффициенты (параметры)* модели. В зависимости от вида функции $f(X, \beta)$ модели делятся на линейные и нелинейные. Так уравнение регрессии (1.1.5) соответствовало эконометрической модели вида

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon, \qquad (1.2.1)$$

которая линейна как по параметрам β_0 ,β_1, β_2, так и по объясняющим переменным X_1, X_2. Такие эконометрические модели называют *линейными моделями.*

Модели временных рядов. К этому классу моделей относятся модели:

- *тренда*

$$Y(t) = T(t) + \varepsilon_t, \qquad (1.2.2)$$

где t – время, $T(t)$ – временной тренд заданного параметрического вида (например, $T(t) = \beta_1 + \beta_2 t$), ε_t – случайная составляющая;

- *сезонности*

$$Y(t) = S(t) + \varepsilon_t, \qquad (1.2.3)$$

где $S(t)$ – периодическая функция.

Кроме этих двух простых моделей используются модели временных рядов, в которых присутствуют несколько слагаемых (аддитивные модели) или несколько сомножителей (мультикативные модели).

Системы одновременных уравнений. 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:

$$Q_S(t) = \alpha_1 + \alpha_2 P(t) + \alpha_3 P(t-1) + \varepsilon(t) \quad \text{(предложение)}. \quad (1.2.4)$$

$$Q_D(t) = \beta_1 + \beta_2 P(t) + \beta_3 Y(t) + u(t) \quad \text{(спрос)}. \quad (1.2.5)$$

$$Q_S(t) = Q_D(t) \quad \text{(равновесие)}. \quad (1.2.6)$$

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 $t-1$.

Применимо к рассмотренным моделям можно ввести следующую классификацию переменных:

- *экзогенные переменные* – переменные, задаваемые из вне рассматриваемой системы и в определенном смысле управляемы;
- *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.*

1.3. Основные этапы эконометрического моделирования

Весь процесс эконометрического моделирования можно разбить на шесть основных этапов.

Этап 1 (постановочный) – определение конечных целей моделирования, набора участвующих в модели факторов и показателей.

Этап 2 (априорный) – предмодельный анализ экономической сущности изучаемого явления, формирование и формализация априорной (известной до начала моделирования) информации и исходных допущений.

Заметим, что не всякая экономико-математическая модель исследуемого экономического объекта может считаться эконометрической. Модель становится эконометрической только в том случае, если будет отражать этот объект на основе характеризующих именно его эмпирических (статистических) данных.

Этап 3 (параметризация) – выбор общего вида модели, состава и форм, входящих в нее связей между переменными. Основная задача, решаемая на этом этапе, - выбор вида функции $f(X_1, X_2, \ldots, X_m)$ в эконометрической модели (1.1.1), в частности, возможность использования линейной модели.

Этап 4 (информационный) – сбор необходимой статистической информации, т. е. регистрация значений участвующих в модели факторов и переменных.

Этап 5 (идентификация модели) – статистический анализ модели и, в первую очередь, статистическое оценивание неизвестных параметров модели.

Этап 6 (верификация модели) – сопоставление реальных и модельных данных, проверка адекватности модели, оценка точности модельных данных.

Содержание этапов 1,2,3 часто объединяют одним название *спецификацией модели.* Вопросы, связанные с этими этапами, в пособии рассматриваются кратко (в основном, это выбор объясняющих переменных эконометрической модели). Основное внимание в учебном пособии будет уделено этапам 5, 6.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Может ли выходная переменная быть одновременно и входной? Если *ДА*, то в каких случаях?
2. Выберите правильный вариант ответа на вопрос:

Эконометрическая модель – это модель:

а) гипотетического экономического объекта;

б) конкретного существующего объекта, построенная по гипотетическим данным;

в) конкретного существующего объекта, построенная по статистическим данным.

3. Выберите правильный вариант ответа на вопрос:
Предопределенные переменные включают:
а) все экзогенные и эндогенные переменные;
б) только экзогенные переменные;
в) все экзогенные переменные и лаговые эндогенные переменные;
г) лаговые экзогенные и эндогенные переменные.

4. 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.

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

Глава 2. Парный регрессионный анализ

В этой главе решаются задачи регрессионного анализа 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):

$$x_1, x_2, \ldots, x_n; \qquad y_1, y_2, \ldots, y_n. \tag{2.1.1}$$

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:

$$f(x) = M(Y|\, x). \tag{2.1.2}$$

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

$$Y = f(x) + \varepsilon. \tag{2.1.3}$$

Выборка (2.1.1) соответствует модели измерений:

$$y_i = f(x_i) + \varepsilon_i; \quad i = 1, 2, \ldots, n. \tag{2.1.4}$$

Присутствие в модели (2.1.3) случайного члена ε, который будем называть возмущение 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:

Р1. Объясняющая переменная X является неслучайной (детерминированной) величиной.

Р2. Возмущения ε_i имеют нулевое среднее, т.е.

$$M(\varepsilon_i) = 0,\ i = 1, 2,\ldots,n. \qquad (2.1.5)$$

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.

Р3. Корреляционные моменты случайных величин $\varepsilon_i, \varepsilon_j$ удовлетворяют условию

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

Первая строка означает *постоянство дисперсии возмущений* ε_i, и это свойство называют *гомоскедастичностью*. Зависимость дисперсии возмущения ε_i от номера наблюдения i или от величины переменной X называется *гетероскедакстичностью.* Характерные диаграммы рассеяния для случаев гомоскедастичности и гетероскедакстичности показаны на рис. 2.1 а) и б) соответственно.

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 $M(Y \mid x)$, а можно только его оценить. Поэтому по выборке ограниченного объема можно построить только оценку для $f(x)$, обозначаемую в дальнейшем как $\hat{y}$ или $\hat{y}(x)$ 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 $f(x)$.

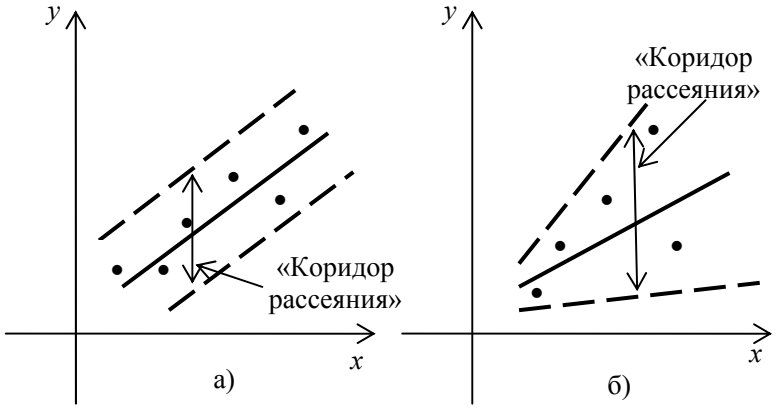

Рис. 2.1. Диаграммы рассеяния

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

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

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:

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

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. Этого недостатка лишен коэффициент корреляции, определяемый по формуле:

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

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

Свойства коэффициента корреляции:

- $-1 \le \rho_{XY} \le 1$;
- $\rho_{XY} = 0$, если *X* и *Y* некоррелированны;
- 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 $\{x_i, y_i\}$, $i = 1,...,n$, 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)):

КОРРЕЛ(*диапазон ячеек X; диапазон ячеек У*). (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*.

3.2. 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 –
$$f(x) = \beta_0 + \beta_1 x + \ldots + \beta_k x^k. \qquad (2.2.2)$$
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 тип 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 $f(x) = –2.75 + 1.016\,x$, которая «достаточно близко» проходит от точек (x_i, y_i). ☻

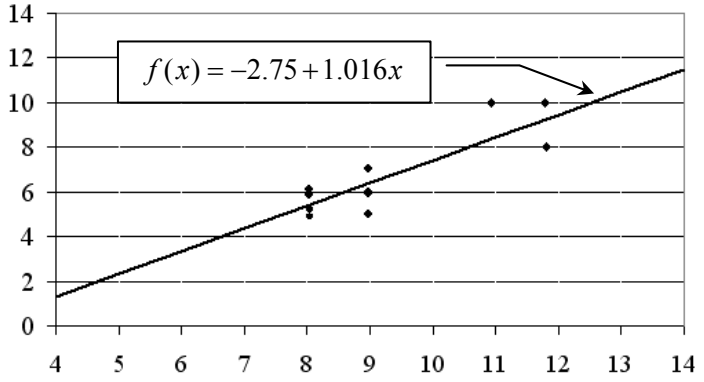

Рис 2.2. Диаграмма рассеяния и линейная регрессия

3. **Экспериментальный.** Для нескольких наиболее подходящих функций регрессий 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:

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

где $\hat{y}_i$ – 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 рассчитываемых параметров 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 $\hat{y}(x) = b_0 + b_1 x$, для функции (2.2.2) - $\hat{y}(x) = b_0 + b_1 x + b_2 x^2 + ... + b_k x^k$ и т.д.

Сначала рассмотрим оценивание коэффициентов функции регрессии для линейной 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.

$$f(x) = \beta_0 + \beta_1 x. \qquad (2.3.1)$$

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

$$\hat{y}(x) = b_0 + b_1 x. \qquad (2.3.2)$$

Выражение (2.3.2) в дальнейшем будем называть уравнением регрессии и для упрощения записи часто $\hat{y}(x)$ будем обозначать $\hat{y}$. Коэффициенты b_0, b_1 являются оценками β_0, β_1 и желательно, чтобы они обладали «хорошими» свойствами. Определим эти свойства.

Обозначим через θ некоторый неизвестный параметр (коэффициент), а через $\hat{\theta}_n$ оценку (называемую точечной оценкой) этого параметра, вычисленную по выборке $\{x_1, x_2, ..., x_n\}$ объемом n, т.е. $\hat{\theta}_n = \varphi(x_1, x_2, \ldots, x_n)$. 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* вероятность того, что оценка $\hat{\theta}_n$ будет отли-

чаться от θ на сколь угодно малую величину ε будет равна нулю, т.е. $\lim\limits_{n\to\infty} P(\left|\hat{\theta}_n - \theta\right| > \varepsilon) = 0$, где запись $P(A)$ означает вероятность события A.

Вопрос: «Как же вычислить «хорошие оценки» для β_0, β_1 ?».

Очевидно, что, если функция $\hat{y}(x)$ соответствует (2.3.1) и (2.1.3), то $\hat{y}(x)$ должно «достаточно близко» проходить от точек (x_i, y_i). Меру близости характеризуют некоторым функционалом. В зависимости от вида функционала, определяющего близость $\hat{y}(x)$ 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).

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

$$F(b_0,b_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - b_0 + b_1 x_i)^2 \qquad (2.3.3)$$

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

$$\hat{y}_i = \hat{y}(x_i) = b_0 + b_1 x_i. \qquad (2.3.4)$$

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

$$y_i = b_0 + b_1 x_i + e_i$$

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

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:

$$\begin{cases} \dfrac{\partial F(b_0,b_1)}{\partial b_0} = 2\sum\limits_{i=1}^{n}(y_i - b_0 - b_1 x_i)\cdot(-1) = 0 \\ \dfrac{\partial F(b_0,b_1)}{\partial b_1} = 2\sum\limits_{i=1}^{n}(y_i - b_0 - b_1 x_i)\cdot(-x_i) = 0 \end{cases} \qquad (2.3.5)$$

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

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

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

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

$$\overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i; \qquad \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2.$$

Тогда систему (2.3.6) можно записать в виде

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

Решая эту систему уравнений, получаем

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

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

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

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

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

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

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

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

которое подтверждает данное выше определение.

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

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

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

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

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

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

где s_X – определяется выражением

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

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

$$\hat{y}(x) = -2.75 + 1.016\,x, \qquad (2.3.17)$$

или

$$\hat{y} - 6.8 = 1.016\,(x - \overline{x}).$$

Прямая линия, соответствующая этим уравнениям, показана на рис. 3.1. ☻

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 3.3.2. 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.

	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***	***Средние значения***	

Рис. 2.4. Вычисление выборочного коэффициента корреляции

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 $f(x) = M(Y \mid x) = \beta_0 + \beta_1 x$. Возникает вопрос: *какими свойствами обладают оценки b_0, b_1?*

При справедливости допущений Р1, Р2, Р3 относительно случайных величин ε_i модели (2.1.4) коэффициенты b_0, b_1 как оценки для β_0, β_1 обладают следующими свойствами:

С1. Коэффициенты b_0, b_1 являются случайными величинами (так как зависят от случайной величины $\overline{y}$);

С2. Коэффициенты b_0, b_1 являются несмещенными оценками т. е.

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

С3. Уравнение регрессии $\hat{y}(x)$ является несмещенной оценкой для функции регрессии $f(x) = M(Y \mid x) = \beta_0 + \beta_1 x$. Действительно, вычислим математическое ожидание функции $\hat{y}(x)$, определяемой (2.3.2):

$$M\left[\hat{y}(x)\right] = M\left[b_0\right] + M\left[b_1\right] \cdot x.$$

С учетом несмещенности оценок b_0, b_1 (см. (2.3.18)) получаем:

$$M\left[\hat{y}(x)\right] = \beta_0 + \beta_1 x = M\left[Y \mid x\right],$$

что доказывает свойство несмещенности оценки $\hat{y}(x)$.

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 $\overline{y}$ (или от y_i) будут иметь больший разброс, а, следовательно, и меньшую точность.

С5. Величина

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

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

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

В заключение приведены формулы, определяющие *дисперсию оценок* b_0, b_1.

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

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* $D(b_0)$, $D(b_1)$:

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

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

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 $\overline{x^2}$ (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		=$B$1+$B$2*A5		
3	Исходные данные					
4	x_i	y_i	$\hat{y}_i$	$e_i=\hat{y}_i-y_i$	e_i^2	$(x_i-\overline{x})^2$
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	$\overline{x}$	9,400			=B20*B19/F15	
19	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 $\hat{y}_i$ ☻

	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

Интерпретация уравнения парной регрессии. После построения уравнения регрессии, возникает вопрос: «Какую информацию несет полученное уравнение?». Другими словами возникает вопрос об интерпретации уравнения регрессии, а точнее – его коэффициентов. Дадим необходимую интерпретацию.

1. Коэффициент b_0 дает прогнозируемое значение переменной Y при $X = 0$. Это значение в зависимости от конкретной ситуации может не иметь экономического смысла. Например, прогнозируемое значение добычи угля, вычисленное по уравнению (2.3.17) при $x = 1$ равно отрицательной величине – 1.634, которая не имеет экономической трактовки.

2. Коэффициент b_1 показывает изменение на b_1 единиц прогнозируемого значения переменной Y при изменении переменной X на одну единицу.

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

$$E = b_1 \cdot \frac{\overline{x}}{\overline{y}}, \qquad (2.3.22)$$

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:

$$E = f'(x) \cdot \frac{\overline{x}}{\overline{y}} \qquad (2.3.23)$$

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).

Из примера 2.3.2 берем значение $b_1 = 1.016$, а из рис. 3.2 значения $\overline{x} = 9.4$, $\overline{y} = 6.8$. 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 $\left(\hat{\theta}_n^{(н)}, \hat{\theta}_n^{(в)}\right)$, в который с заданной вероятностью γ попадает неизвестное значение параметра θ, т. е.

$$P\left(\hat{\theta}_n^{(н)} < \theta < \hat{\theta}_n^{(в)}\right) = \gamma.$$

Интервал $\left(\hat{\theta}_n^{(н)}, \hat{\theta}_n^{(в)}\right)$ 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:

Р4. 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.

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

Отсюда следует, что статистики:

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

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

$$b_0 - t(\gamma, n-2)\cdot s_{b_0} \le \beta_0 \le b_0 + t(\gamma, n-2)\cdot s_{b_0};$$
$$b_1 - t(\gamma, n-2)\cdot s_{b_1} \le \beta_1 \le b_1 + t(\gamma, n-2)\cdot s_{b_1},$$

где $t(\gamma, n-2)$ вычисляется с помощью функции Excel (см. (2.4.11)):

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

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:

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

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

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

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

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

Квантили определяются следующими выражениями:

$$\chi^2_{\alpha/2,n-2} = \text{ХИ2ОБР}(1-\alpha/2;\, n-2), \qquad (2.4.6)$$

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

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: $\chi^2_{0.025,8} = 2.18$; $\chi^2_{0.975,8} = 17.53$ и получаем интервальную оценку для σ^2:

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

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

$$\hat{y}(x) = \bar{y} + b_1(x - \bar{x}). \qquad (2.4.8)$$

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

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

Заменяя неизвестную дисперсию σ^2 на ее оценку s^2, получаем оценку для $\sigma^2_{\hat{y}}$, равную

$$s^2_{\hat{y}}(x) = s^2 \left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \right]. \qquad (2.4.9)$$

Статистика

$$T_{\hat{y}}(x) = \frac{\hat{y}(x) - M(Y \mid x)}{s_{\hat{y}}(x)}$$

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

$$\hat{y}(x) - t(\gamma, n-2)\cdot s_{\hat{y}}(x) \le M(Y \mid x) \le \hat{y}(x) + t(\gamma, n-2)\cdot s_{\hat{y}}(x).$$

Следовательно, интервал

$$\left[\hat{y}(x) - t(\gamma, n-2)\cdot s_{\hat{y}}(x),\ \hat{y}(x) + t(\gamma, n-2)\cdot s_{\hat{y}}(x) \right] \qquad (2.4.10)$$

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

Так как $s_{\hat{y}}(x)$ 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 $M(Y \mid x)$ с надежностью $\gamma = 0.95$, используя для этого уравнение регрессии $\hat{y}(x)$, построенное в примере 2.3.1.

Решение. Значения граничных точек y_i^H (нижняя), y_i^B (верхняя) интервальной оценки будем вычислять для $x = x_i$, $i = 1,\ldots,10$, приведенных в таблице 2.1 по формуле (2.4.10). Фрагмент документа Excel, осуществляющего вычисление граничных точек и значений $\hat{y}(x_i)$ 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. ☻

	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	$\hat{y}_i$	$s^2_{\hat{y}}$	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 $M(Y \mid x)$

Интервальная оценка для индивидуальных значений зависимой переменной. Построенная интервальная оценка (2.4.10) определяет возможное положение математического ожидания $M(Y \mid x)$, но не отдельных возможных значений зависимой переменной Y, которые отклоняется от $M(Y \mid x)$. 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 $M(Y \mid x)$. Дисперсия таких отклонений равна σ^2. Следовательно, оценку дисперсии $s^2_{\hat{y}}(x)$ необходимо увеличить на s^2 (оценка для σ^2). В результате оценка дисперсии значений y^* равна

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

а соответствующая интервальная оценка определяется интервалом

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

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 $x_1, x_2, \ldots, x_n$ 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:

С1. Гипотеза H_0 *верна*, и при проверке она *не отвергается*;

С2. Гипотеза H_0 *верна*, но при проверке она *отвергается*;

С3. Гипотеза H_0 *не верна*, и при проверке она *отвергается* (в пользу альтернативной H_1);

С4. Гипотеза H_0 *не верна*, но при проверке она *принимается*.

Очевидно, что ситуации С1, С3 являются «правильными» ситуациями, С2, С4 – «ошибочными». Ситуация С2 называется *ошибкой I рода*, и вероятность ее появления называется *уровнем значимости* (обозначается α). Обычно $\alpha = 0.025 \div 0.05$. Ситуация С4 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

$$T_{b_0} = b_0 \Big/ s_{b_0}, \qquad (2.5.1)$$

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

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

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

$$T_{b_1} = b_1 \Big/ s_{b_1}, \qquad (2.5.3)$$

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

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

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

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

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

$$\left| T_r \right| > t(1-\alpha, n-2), \qquad (2.5.6)$$

го 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.

Для проверки значимости коэффициента b_0 вычислим значение критерия (стандартную ошибку s_{b_0} возьмем из примера 3.3.3): $T_{b_0} = b_0 / s_{b_0} = -2.41/1.98 = -1.217$. 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 примера 3.3.3): $T'_{b_1} = b_1 / s_{b_1} = 1.016/0.21 = 4.84$. 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 $\alpha = 0.05$).

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 $x = x_i$, $i = 1,\ldots,n$ 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?

Проверка значимости производится на основе дисперсионного анализа. Введем три суммы:

- *полная сумма квадратов* (в переводной литературе обозначается TSS)

$$Q = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

интерпретируемая как мера общего рассеивания переменной Y относительно среднего значения $\bar{y}$;

- *объясненная (или факторная) сумма квадратов* (в переводной литературе – RSS)

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

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

- *остаточная сумма квадратов* (в переводной литературе – ESS)

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

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

Можно показать, что для уравнения регрессии ***со свободным членом*** (т.е. присутствует коэффициент b_0), построенного на основе МНК, справедливо следующее равенство

$$Q = Q_r + Q_e \qquad (2.5.7)$$

Далее, случайные величины

$$s_r^2 = \frac{Q_r}{m-1} \quad и \quad s^2 = \frac{Q_e}{n-m}, \qquad (2.5.8)$$

где m – количество коэффициентов регрессии (для линейной парной регрессии $m = 2$), имеют χ^2-распределение соответственно с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы, а их отношение s_r / s подчиняется F-распределению с теми же степенями свободы.

Таким образом, уравнение парной регрессии значимо с уровнем значимости α, если выполняется условие

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

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

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

Пример 2.5.3. По данным таблицы 3.1 оценить на уровне α = 0.05 значимость уравнения регрессии, построенного в примере 3.3.1.

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 α = 0.05. ☻

	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				=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

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

$$R^2 = \frac{Q_r}{Q} = 1 - \frac{Q_e}{Q}. \qquad (2.5.10)$$

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

$$0 \le R^2 \le 1 \qquad (2.5.11)$$

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.

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

$$R^2 = r_{XY}^2. \qquad (2.5.12)$$

Пример 3.5.4. По данным таблицы 3.1 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

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

где

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

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

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

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):

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

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

$$\hat{y}(z) = b_0 + b_1 z\ ,$$

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

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

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

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

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)):

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

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

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

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

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

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

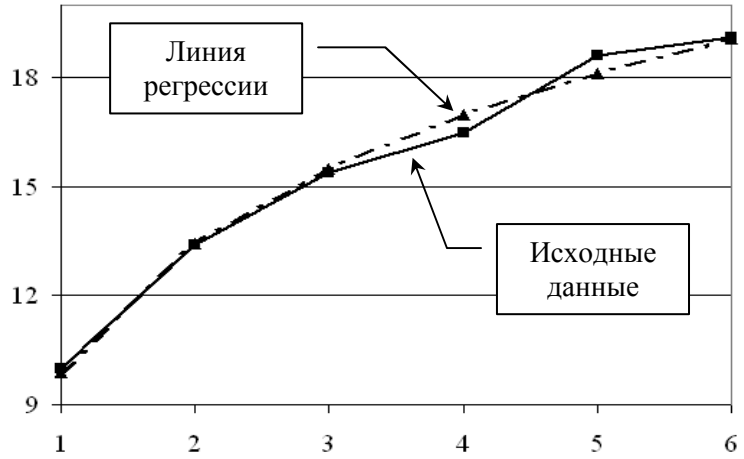

Рис.2.9. Построение нелинейной регрессии

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

$$Y' = \ln Y',\ b_0' = \ln b_0,\ \ X' = \ln X,\ \ \varepsilon' = \ln \varepsilon .$$

Относительно этих величин имеем линейную регрессионную модель

$$Y' = \beta_0' + \beta_1 X' + \varepsilon', \qquad (2.6.6)$$

которой соответствует уравнение линейной регрессии

$$\hat{y}' = b_0' + b_1 x' . \qquad (2.6.7)$$

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).

Напомним, что эффективность оценок, получаемых методом наименьших квадратов, основана на допущении о том, что возмущения ε_i не коррелированны 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

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

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

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

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

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

Существует достаточно большое число алгоритмы минимизации различных функционалов. Некоторые из этих алгоритмов реализованы 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 $Q_e = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$, $Q = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2$, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, $\hat{y}_i$ - значение зависимой переменной Y, вычисленное по уравнению нелинейной регрессии при $x = x_i$. Очевидно, что величина этого показателя удовлетворяет неравенству: $0 \le I_{xy} \le 1$, причем $I_{xy} = 1$, когда все значения y_i "лежат" на линии регрессии.

Индексом детерминации называется величина

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

Величина индекса детерминации изменяется в пределах

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

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 детерминации 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

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

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

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

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

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

где $t(1-\alpha, n-2) = \text{СТЬЮДРАСПОБР}(\alpha; n-2)$. 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

$$\hat{y}(x) = 9.876 + 5.129 \ln(x) \qquad (2.6.15)$$

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

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

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

$$\hat{y}(x) = 9.28 + 1.777 x. \qquad (2.6.16)$$

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^2_{xy} можно выполнить проверку значимости построенной нелинейной регрессии. Для этого определим F - критерий

$$F = \frac{R^2_{xy}}{1 - R^2_{xy}} \cdot \frac{n - m - 1}{m}, \qquad (2.6.17)$$

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

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

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

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

Пример 2.6.3. Определим 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 $R^2_{xy} = 0.9916$ 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 ядов. 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 $\hat{y} = b_0 + b_1 x$
- полиноминальную $\hat{y} = b_0 + b_1 x + \ldots + b_k x^k$ ($k \le 6$);
- логарифмическую $\hat{y} = b_0 + b_1 \ln x$
- степенную $\hat{y} = b_0 x^{b_1}$;
- экспоненциальную $\hat{y} = b_0 e^{b_1 x}$.

Для построения одной из перечисленных регрессий необходимо выполнить следующие шаги:

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

Шаг 2. По этим данным построить график в декартовый 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.

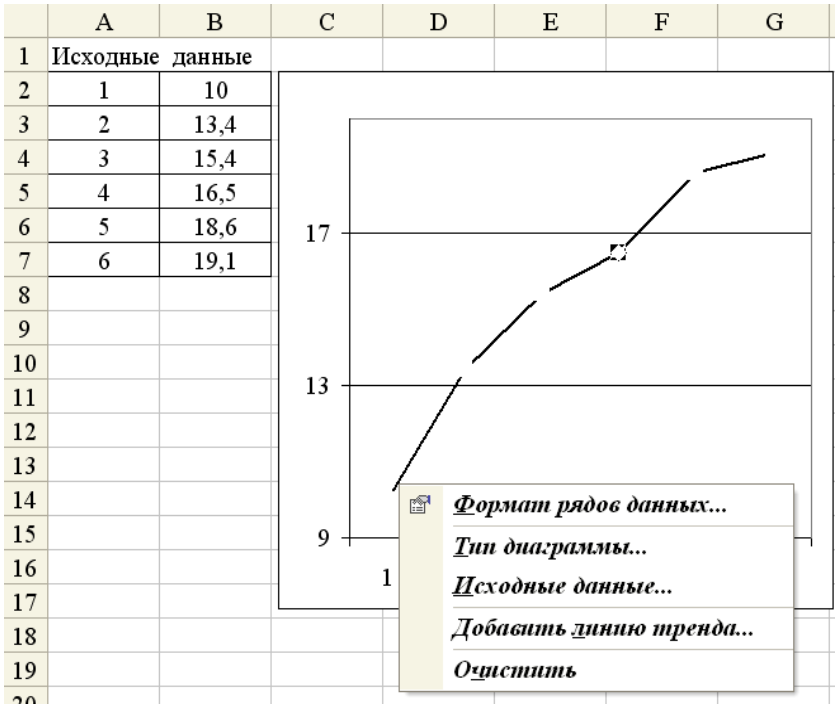

Рис. 2.10. Построение графика по исходным данным

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 коэффициентами.
- *«Поместить на диаграмму величину достоверности аппроксимации (R^2)»* - на диаграмме будет показана значение индекса детерминации R^2_{xy} (см.(2.6.11)), которое можно использовать для проверки значимости построенной регрессии с помощью F - теста (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.

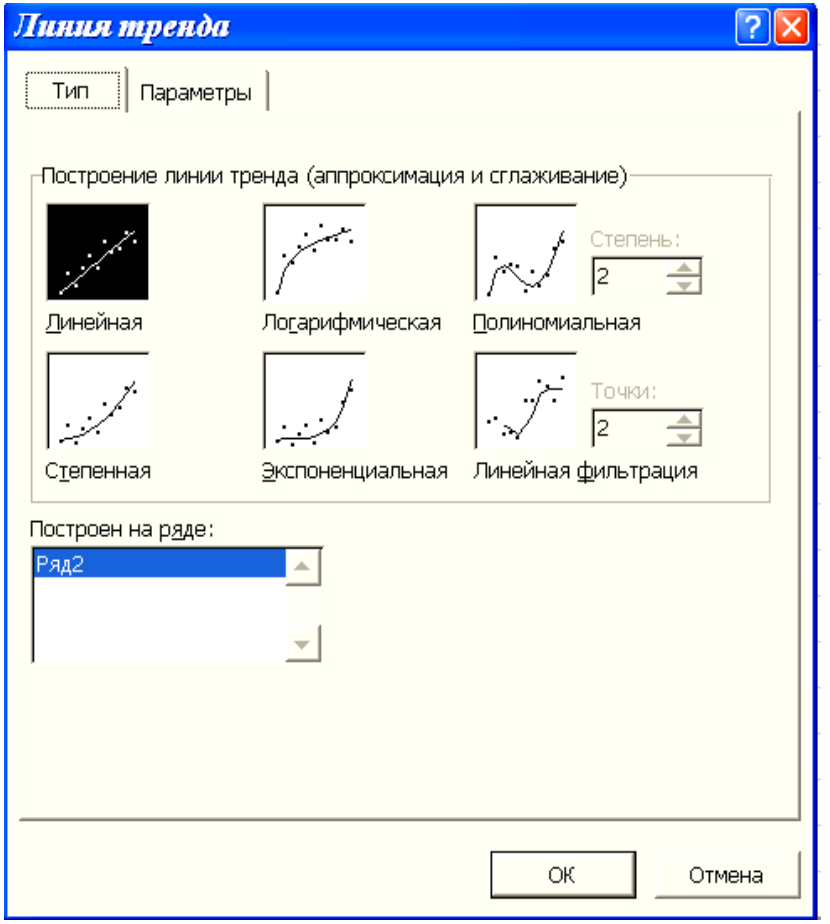

Рис. 2.11. Выбор вида уравнения регрессии

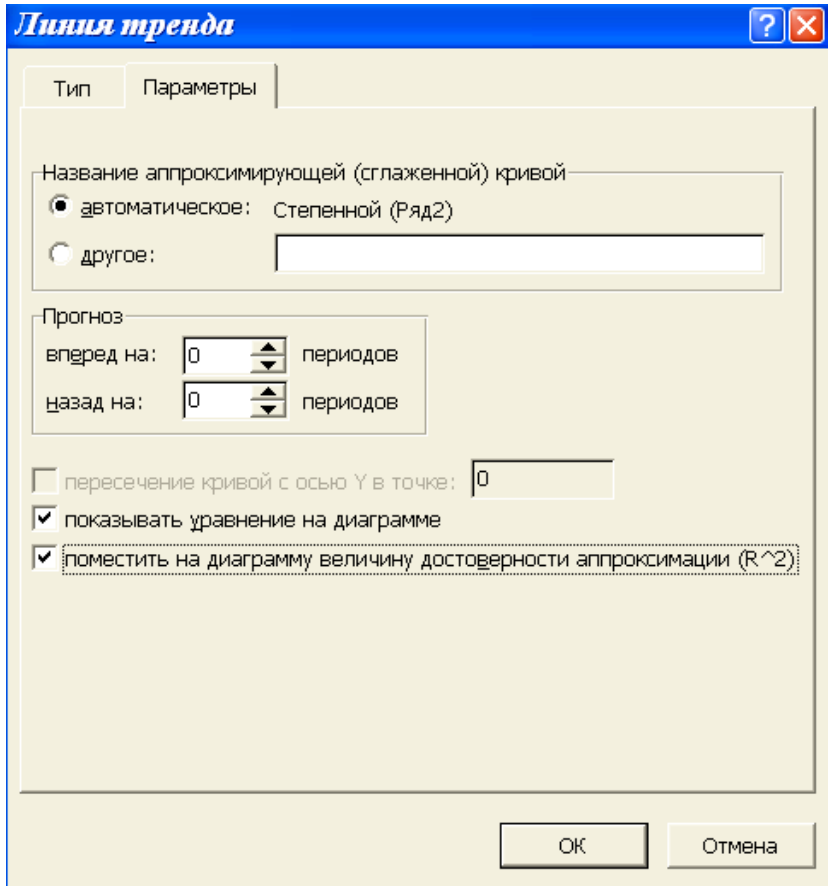

Рис. 2.12. Задание опций вывода информации

Шаг 6. После задания всех перечисленных опций щелкнуть на кнопке «ОК» и на диаграмме появиться формула построенного уравнения регрессии и значение индекса детерминации R^2_{xy} (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^2_{xy} выбрать наилучшее уравнение.

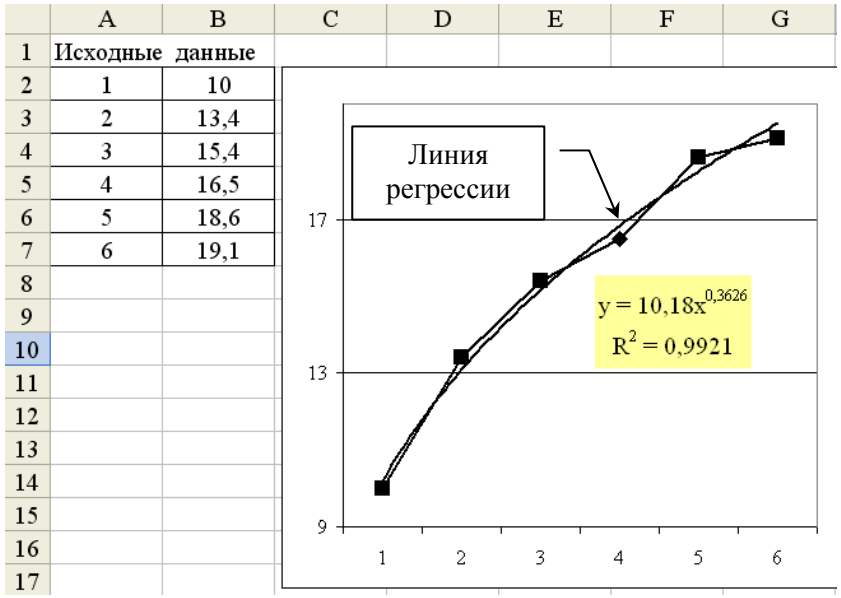

Рис. 2.13. График и уравнение построенной регрессии

Решение. Построение каждого из пяти уравнений осуществляем по описанным выше шагам. Для уравнения $\hat{y} = b_0 \cdot x^{b_1}$ выполнение шагов иллюстрирует рис. 2.10÷2.13. В таблицу 2.3 заносим регрессионное уравнение и соответствующее значение R^2_{xy}. Сравнивая величину индекса детерминации R^2_{xy} для этих уравнений, в качестве «наилучшего» уравнения выбираем степенную регрессию $\hat{y} = 10.18x^{0.3226}$ (номер 5), для которой индекс детерминации $R^2_{xy} = 0.9921$.

Таблица 2.3

№	Уравнение	R^2_{xy}	$\hat{R}^2_{xy}$
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	$\hat{y} = 6.93 + 3.5396x - 0.2518x^2$ (полиноминальная, $m = 2$)	0.9896	0.9827
4	$\hat{y} = 5.8333 + 4.9192x - 0.7087x^2 - -0.0435x^3$ (полиноминальная, $m = 3$)	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^2_{xy} 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^2_{xy}, 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):*

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

где m - количество коэффициентов регрессии, Видно, что при неизменных Q_e, Q увеличение m уменьшает значение $\hat{R}^2_{xy}$. Ес-

ли количество коэффициентов у сравниваемых уравнений регрессии одинаково (например, $m = 2$), то отбор наилучшей регрессии можно осуществлять по величине R^2_{xy}. Если в уравнениях регрессии меняется число коэффициентов, то отбор целесообразно по величине $\hat{R}^2_{xy}$.

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

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 $0.2 \le b < 1$. 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^2_{xy}.

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* коэффици-

ента b_1. На рис. 2.14 показан фрагмент документа Excel с введенными данными.

	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)		
11						
12						
13						
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						

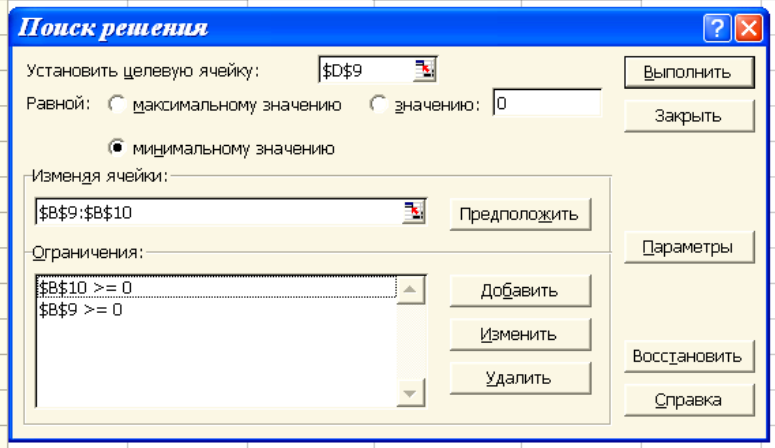

Рис. 2.14. Задание параметров команды ***Поиск решения***

Следующим шагом является вычисление по уравнению регрессии значений $\hat{y}_i = b_0 \cdot x_i^{b_1}$, $i = 1,...,6$. 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).

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 найденной 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».

	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					

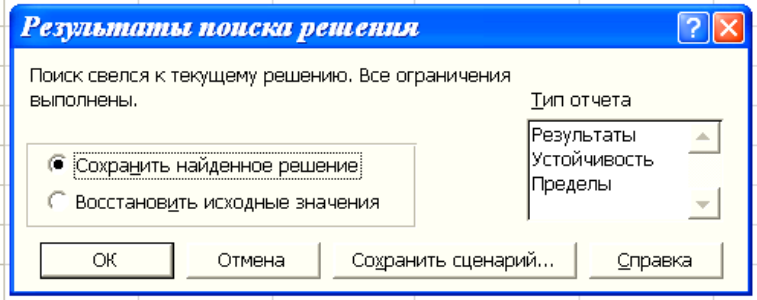

Рис. 2.15. Результаты выполнения команды *Поиск решения*

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

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*).

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

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.
5. Вычислить по построенному уравнению регрессии значения $\hat{y}_i = b_0 + b_1 x_i$; $e_i = \hat{y}_i - y_i$; $i = 1, 2, \ldots, 7$. 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. 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 $s^2_{b_0}$, $s^2_{b_1}$ 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 $x = x_i$, $i = 1, 2,\ldots,7$ (см. (2.4.10)) с надежностью $\gamma = 0.95$. Построить диаграмму по значениям y_i^H, y_i^B, $\hat{y}_i$, $i = 1, 2,\ldots,7$ (по оси X откладываются значения x_i).

Контрольные значения: среднеквадратические ошибки: $s_{b_0} = 1.34$, $s_{b_1} = 0.028$.

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 $\hat{y}(x) = b_0 + b_1 x$). 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 наличии линейного тренда.

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

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

4. Зависимость времени разговора покупателя с продавцом спортивного отдела (переменная *У* 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 $\overline{x} = 130$, $\overline{y} = 20$. Определите экономический смысл вычисленной величины коэффициента эластичности.

5. 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.

6. Какими свойствами обладают оценки b_0, b_1, вычисленные методом наименьших квадратов при выполнении предположений Р1 ÷ Р3?

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

8. Поясните статистический смысл коэффициента детерминации R^2.

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

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

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

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

Глава 3. Множественный регрессионный анализ

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 удается 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 $x_1, x_2, ..., x_k$, объясняющих переменных $X_1, X_2, X_3, \ldots, X_k$, т.е.

$$f(x_1, x_2, x_3, \ldots, x_k) = M(Y \mid x_1, x_2, x_3, \ldots, x_k)$$

Также к множественной регрессии мы приходим, когда априори известно о влиянии на зависимую переменную Y нескольких объясняющих переменных $X_1, X_2, X_3, \ldots, X_k$ (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, $x_2, \ldots, x_k$, т. е.

$$f(x_1, x_2, \ldots, x_k) = M(Y|x_1, x_2, \ldots, x_k) \qquad (3.1.1)$$

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

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

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

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

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):

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

а также векторы:

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

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

Тогда в матричном виде модель наблюдений (3.1.3) примет вид

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

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

Р1. Матрица X – неслучайная матрица, а ε – случайный вектор.

Р2. $M(\varepsilon) = 0_n$, (3.1.5)

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

Р3. $V_\varepsilon = M\left[\varepsilon \varepsilon^T\right] = \sigma^2 I$, (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.

Р4. Случайный вектор ε подчиняется нормальному распределению $N(0_n, \sigma^2 I)$.

Р5. Ранг матрицы X $rank(X)$ удовлетворяет условию

$$rank(X) = k + 1 < n. \qquad (3.1.7)$$

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

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

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 становится очевидным условие (3.1.7). Неравенство $k + 1 < n$ требует, чтобы число неизвестных было меньше числа уравнений.

Регрессионной линейной модели (3.1.4) соответствует *уравнение множественной линейной регрессии вида*

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

где $b_0, b_1, \ldots, b_k$ – коэффициенты регрессии, являющиеся оценками для $\beta_0, \beta_1, \ldots, \beta_k$ и которые необходимо вычислить, решая систему 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 $x_1 = x_{i1}, \ldots, x_k = x_{ik}$ (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:

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

а вектор $X^T y$ имеет m проекций:

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

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

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

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

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

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 уровнем 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

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

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

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

содержащий 3 проекции;

- обратную матрицу

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

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, $j = 1,...,k$.

Стандартизованный коэффициент регрессии b'_j определяются выражением

$$b'_j = b_j \cdot \frac{s_{x_j}}{s_y}, j = 1,\ldots,k,$$

где

$$s_y = \left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2\right)^{1/2} = \sqrt{\overline{y}^2 - (y)^2};$$

$$s_{x_j} = \left(\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2\right)^{1/2} = \sqrt{\overline{x}_j{}^2 - (\overline{x}_j)^2};$$

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

$$E_j = b_j \cdot \frac{\overline{x}_j}{\overline{y}}, \ j = 1,...,k.$$

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:

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

а затем коэффициенты эластичности

$$E_1 = 0.8539 \cdot \frac{9.4}{6.8} = 1.180; \ E_2 = 0.3670 \cdot \frac{6.3}{6.8} = 0.340.$$

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 4.1).

1. ***Вектор b является случайным вектором***. Действительно, с учетом (3.1.4) вектор b можно представить в виде двух слагаемых:

$$b = \left(X^T X\right)^{-1} X^T \left(X\beta + \varepsilon\right) = \left(X^T X\right)^{-1} X^T X\beta + \\ + \left(X^T X\right)^{-1} X^T \varepsilon = \beta + \left(X^T X\right)^{-1} X^T \varepsilon = \beta + \xi. \qquad (3.2.6)$$

Если первое слагаемое – вектор «точных» коэффициентов β – является неслучайным, то второе слагаемое – вектор $\xi = \left(X^T X\right)^{-1} X^T \varepsilon$ есть результат линейного преобразования случайного вектора ε и поэтому также является *случайным вектором.* Таким образом, вектор ξ вызывает отклонение вектора b от вектора β и поэтому вектор ξ можно назвать *вектором ошибки оценивания вектора* β.

2. ***Вектор b является несмещенной оценкой.*** С учетом (3.2.6) вычислим математическое ожидание вектора *b*:

$$M(b) = M(\beta) + \left(X^T X\right)^{-1} X^T M(\varepsilon). \qquad (3.2.7)$$

Для неслучайного вектора β справедливо равенство $M(\beta) = \beta$. Тогда из допущения (3.1.5) $M(\varepsilon) = 0_n$ непосредственно следует

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

так как произведение нулевого вектора на матрицу равно нулевому вектору (второе слагаемое в (3.2.7)).

Таким образом, показано, что вектор b есть *несмещенная оценка* вектора β.

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

$$V_b = V_\xi = \sigma^2 \left(X^T X\right)^{-1}. \qquad (3.2.9)$$

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

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

где запись $\left[V_b\right]_{i,j}$ – означает *i,j* – ый элемент матрицы V_b, а $\mu_{i,j}$ называют корреляционным моментом проекций b_i и b_j вектора *b*. Напомним, что диагональные элементы $\mu_{i,i}$ определяют дисперсию *i*-ой проекции b_i и в дальнейшем обозначаются как $\sigma^2_{b_i}$.

В матричном виде корреляционная матрица определяется выражением

$$V_b = M\left[(b - M(b))(b - M(b))^T\right].$$

Учитывая (3.2.6), (3.2.8), можно записать

$$V_b = M\left[\left(\left(X^T X\right)^{-1} X^T \varepsilon\right)\cdot\left(\left(X^T X\right)^{-1} X^T \varepsilon\right)^T\right] = \qquad (3.2.11)$$
$$= \left(X^T X\right)^{-1} X^T M(\varepsilon\varepsilon^T) X \left(X^T X\right)^{-1}$$

Матрица $V_\varepsilon = M(\varepsilon\varepsilon^T)$ есть ковариационная матрица вектора ε и согласно (3.1.6) равна $\sigma^2 I$, где I – единичная матрица. Подставляя это выражение в (3.2.11), получаем

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

Из (3.2.6), (3.2.8) следует $b - M(b) = \xi$. Следовательно

$$V_\xi = M\left[\xi\xi^T\right] = M\left[(b - M(b))(b - M(b))^T\right] = V_b = \sigma^2 \left(X^T X\right)^{-1}.$$

Последние два матричные выражения доказывают справедливость (3.2.9).

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

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

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

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

4. ***Вектор b подчиняется нормальному распределению*** $N\left(\beta, \sigma^2 \left(X^T X\right)^{-1}\right)$**.** Вернемся к представлению (3.2.6). Вектор ε в соответствии с допущением **P4** (см. параграф 4.1) 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

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

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

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 $\sigma^2_{b_j}$ 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

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

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 (см. [5, стр. 96 – 97]), 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 $\left(X^T X\right)^{-1}$ по формуле $b = \left(X^T X\right)^{-1}\left(X^T y\right)$ (см. (3.2.5)). Для реализации этой матричной формулы в 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:

ТРАНСП (*диапазон ячеек*), (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:

МУМНОЖ(*диапазон_1;диапазон_2*), (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:

МОБР (*диапазон ячеек*), (3.2.18)

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

Замечание 4.2.1. 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. 4.1 приведены примеры использования матричных функций Excel. ☻

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

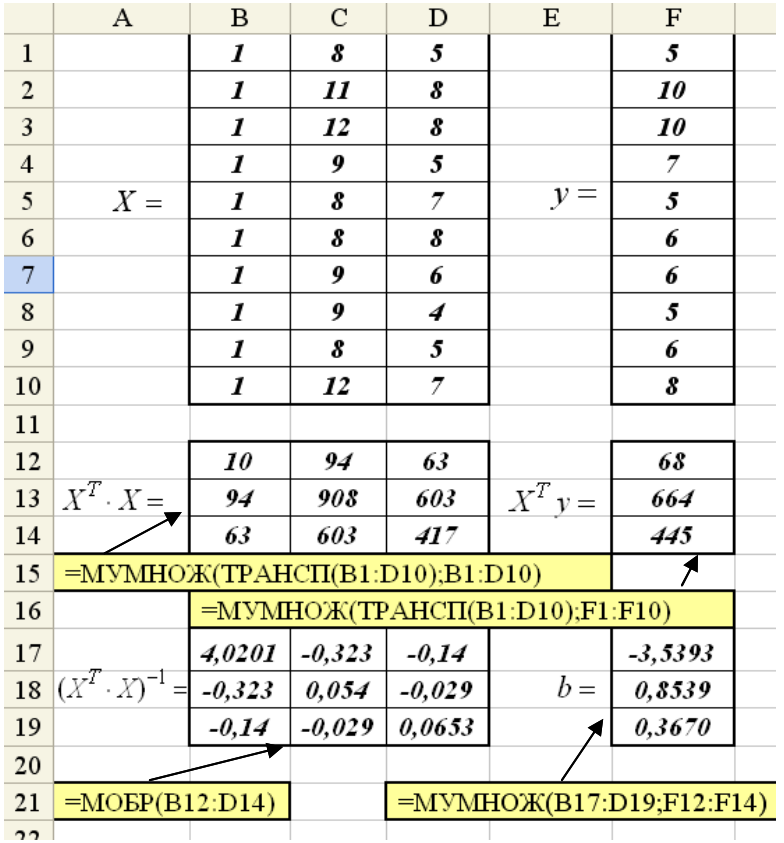

Рис. 3.2. Вычисление коэффициентов регрессии

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 может 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

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

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

Среднее квадратическое отклонение коэффициента регрессии b_j определяется как

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

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

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

имеет распределение Стьюдента с $n - m$ степенями свободы. Следовательно, интервал

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

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

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

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

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

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

Интервальная оценка для дисперсии σ^2. Строится аналогично парной регрессии по формуле (2.4.3) с соответствующим изменением числа степеней свободы. Поэтому интервальная оценка для σ^2 с доверительной вероятностью $\gamma = 1 - \alpha$ имеет вид

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

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

$$\chi^2_{\alpha/2,n-2} = \text{ХИ2ОБР}(1-\alpha/2; n-2), \qquad (3.3.8)$$

$$\chi^2_{1-\alpha/2,n-2} = \text{ХИ2ОБР}(\alpha/2; n-2). \qquad (3.3.9)$$

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 $(X^T X)^{-1}_{i,i}$ обратной матрицы $(X^T X)^{-1}$, вычисленные в примере 4.2.3:

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

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

Находим $t(0.95,10–3)$=СТЬЮДРАСПОБР($0.05;10-3$)=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

$$-8.04 \le \beta_0 \le 0.096;$$

- 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

$$0.332 \le \beta_1 \le 1.376;$$

- 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

$$-0.206 \le \beta_2 \le 0.940.$$

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

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

или с вероятностью 0.95 выполняется неравенство

$$0.565 \le \sigma^2 \le 5.349. \quad ☻$$

Интервальная оценка для множественной функции регрессии. Так же как и для парной регрессии, интервальная оценка для условного математического ожидания $M(Y \mid x)$ (или для функции регрессии) надежности γ имеет вид

$$[\hat{y} - t(\gamma, n-m) \cdot s_{\hat{y}}(x),\ \hat{y} + t(\gamma, n-m) \cdot s_{\hat{y}}(x)] \qquad (3.3.10)$$

или с вероятностью γ выполняется неравенство

$$\hat{y} - t(\gamma, n-m) \cdot s_{\hat{y}}(x) \le M(Y \mid x) \le \hat{y} + t(\gamma, n-m) \cdot s_{\hat{y}}(x),$$

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

$$s_{\hat{y}}(x) = s \cdot \sqrt{x^T \left(X^T X\right)^{-1} x}\,, \qquad (3.3.11)$$

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

В отличие от парной регрессии, где $s_{\hat{y}}(x)$ зависит только от одной объясняющей переменной (см. (2.4.6)) для множественной регрессии оценка $s_{\hat{y}}(x)$ зависит уже от вектора x, что существенно усложняет геометрическую интерпретацию интервальной оценки.

Интервальная оценка для индивидуальных значений зависимой переменной. Построенная оценка (3.3.8) определяет интервал возможных значений возможного математического ожидания $M(Y \mid x)$, но не отдельных возможных значений (названных индивидуальными значениями и обозначаемых y^*) переменной Y, которые отклоняются от $M(Y \mid x)$.

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

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

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

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

Появление 1 под знаком корня по сравнению с (3.3.9) объясняется учетом дополнительного отклонения значений y^* от своего математического ожидания $M(Y \mid x)$.

Пример 3.3.2. По данным примера 3.2.1 найти интервальные оценки для среднего значения ($M(Y \mid x)$) 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 $M(Y \mid x)$ при $x = |\ 1 \quad 8 \quad 6\ |^T$. 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 $s_{\hat{y}}(x)$, $s_{y^*}(x)$ вычислим

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

$$= \frac{699}{3738} = 0.187.$$

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

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

или с вероятностью 0.95 выполняется неравенство

$$4.52 \le M(Y \mid x) \le 6.46 \text{ (т)}.$$

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

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

или с вероятностью 0.95 выполняется неравенство

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

Напомним, что y^* индивидуальные значения переменной Y при векторе $x = |\ 1\ 8 \quad 6\ |$. Видно, что интервал для y^* «шире» интервала для $M(Y \mid x)$. *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 сформулируем статистические гипотезы:

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

H_1: $\beta_j \ne 0$ (коэффициент b_j значим).

В качестве критерия проверки гипотезы примем следующую случайную величину

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

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

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

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 4.2.1.

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

$$H_0: \beta_1 = 0, \beta_2 = 0, \ldots, \beta_k = 0$$

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:

$$Q = y^T y - n(\bar{y})^2 ; \qquad (3.4.3)$$

$$Q_e = y^T y - b^T X^T y ; \qquad (3.4.4)$$

$$Q_r = b^T X^T y - n(\bar{y})^2 , \qquad (3.4.5)$$

где $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$. Уравнение множественной регрессии значимо с уровнем значимости α, если статистика

$$F = \frac{Q_r \cdot (n-m)}{Q_e \cdot (m-1)} \qquad (3.4.6)$$

удовлетворяет условию

$$F > F_{1-\alpha;\, m-1;\, n-m} , \qquad (3.4.7)$$

где $F_{1-\alpha;\, m-1;\, n-m}$ – квантиль распределения Фишера, значение которого определяется выражением

$$F_{1-\alpha;\, m-1;\, n-m} = \text{FРАСПОБР}(\alpha; m-1; n-m).$$

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

$$R^2 = 1 - \frac{(y - Xb)^T (y - Xb)}{(y - \bar{y})^T (y - \bar{y})} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} , \qquad (3.4.8)$$

где $\bar{y}$ – вектор размерности n, составленный из средних значений $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$. Напомним, *что R^2 характеризует долю вариации зависимой переменной, обусловленной изменением объясняющих переменных x_1, x_2,…,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

$$R^2 = \frac{Q_r}{Q} = \frac{b^T X^T y - n\bar{y}^2}{y^T y - n\bar{y}^2} . \qquad (3.4.9)$$

Если известен коэффициент детерминации R^2, то статистику F (3.4.6) можно записать в виде

$$F = \frac{R^2 (n-m)}{(1-R^2)(m-1)}. \qquad (3.4.10)$$

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

$$\hat{R}^2 = 1 - \frac{n-1}{n-m} \cdot \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \overline{y})^2}, \qquad (3.4.11)$$

где m – число коэффициентов регрессии.

Если известен коэффициент R^2, то скорректированный коэффициент детерминации можно вычислить по формуле:

$$\hat{R}^2 = 1 - \frac{n-1}{n-m} \cdot (1-R^2). \qquad (3.4.12)$$

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:

$$b^T X^T y = \begin{vmatrix} -3.54 & 0.854 & 0.367 \end{vmatrix} \cdot \begin{vmatrix} 68 \\ 664 \\ 445 \end{vmatrix} = 489.65;$$

$$y^T y = \sum_{i=1}^{10} y_i^2 = 496; \quad \overline{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = 6.8.$$

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 $\hat{R}^2$ для разного количества объясняющих переменных (величина k):

- если $k = 1, \quad m = 2$, то $\hat{R}^2 = 1 - \frac{9}{8}(1 - 0.75) = 0.720$;
- если $k = 2, \quad m = 3$, то $\hat{R}^2 = 1 - \frac{9}{7}(1 - 0.811) = 0.757$.

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 Ok.

После вызова режима ***Регрессия*** на экране появляется диалоговое окно (см. рис. 3.3), в котором задаются следующие параметры:

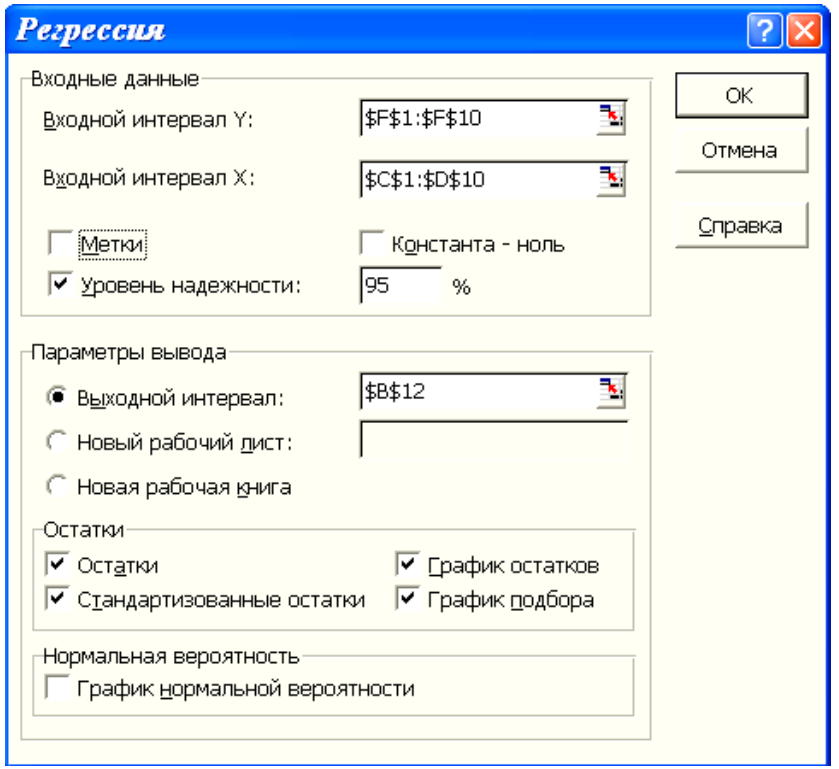

Рис. 3.3. Диалоговое окно режима *Регрессия*

1. Входной интервал Y – вводится диапазон адресов ячеек, содержащих значения y_i (ячейки должны составлять один столбец).

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. $k \le 16$).

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 $y_i - \hat{y}_i, i = 1,...,n$.

10. Стандартизованные остатки – при включении вычисляется столбец, содержащий стандартизованные остатки.

11. График остатков – при включении выводятся точечные графики невязки $y_i - \hat{y}_i, i = 1,...,n$, в зависимости от значений переменных $x_j, j = 1,...,k$. 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 $x_j, j = 1,...,k$. Количество графиков равно числу k переменных x_j.

Пример 3.6.1. 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			
Регрессионная статистика			
Множественный R	0,9009		
R-квадрат	0,8116		
Нормированный R-квадрат	0,7578		
Стандартная ошибка	0,9509		
Наблюдения	10		
Дисперсионный анализ			
	df	*SS*	*MS*
Регрессия	2	27,2704	13,635
Остаток	7	6,3296	0,904
Итого	9	33,6000	
		F	*Значимость F*
		15,0794	0,0029

Рис. 3.4. Результаты работы режима *Регрессия*

Дадим краткую интерпретацию показателям, значения которых вычисляются в режиме *Регрессия.* Первоначально рассмот-

рим 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 $\hat{R}^2$ (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.

$$SS_r = Q_r = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2 ;$$

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 свободы. Эту вероятность можно также определить с помощью функции

$$= \text{FРАСП}(F_c; k_r; k_e).$$

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 $b_0, b_1, ..., b_k$, расположенных сверху-вниз.

Столбец Стандартная ошибка – значения $s_{b_j}, j = 0,...,k$, 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 У – значения $\hat{y}_i$, вычисленные по построенному уравнению регрессии.

Столбец Остатки – значения невязок $y_i - \hat{y}_i$

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное У*	*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 названы 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.

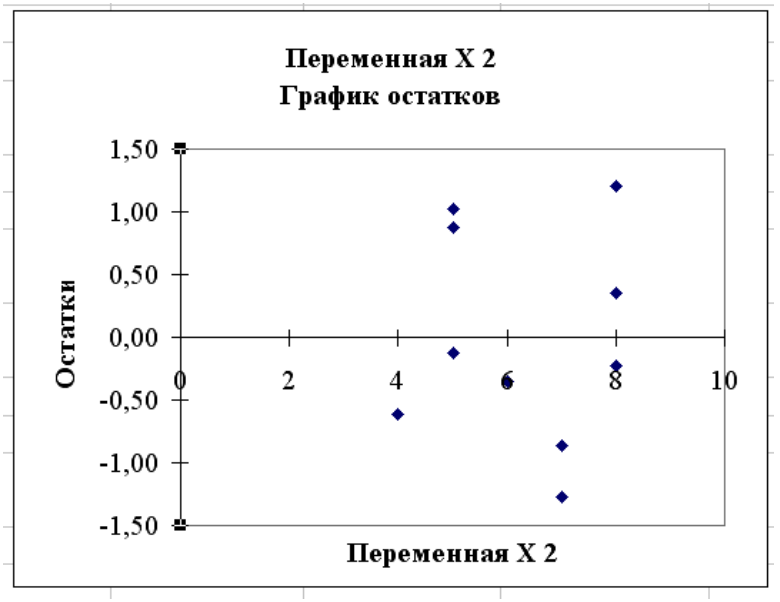

Рис. 3.7. График невязок как функция переменной X_2

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* (за-

висимость 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*

$$y_i = \beta_0 + \beta_1 x_{i1}^2 + \beta_2 \sqrt{x_{i2}} + \varepsilon_i,\ i = 1, ..., n, \qquad (3.6.1)$$

то, вводя новые переменные $Z_1 = X_1^2, Z_2 = \sqrt{X_2}$, получаем *новую линейную модель*

$$y_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \varepsilon_i,\ i = 1, ..., n, \qquad (3.6.2)$$

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*

$$y_i = \beta_0 \cdot x_{i1}^{\beta_1} \cdots x_{ik}^{\beta_k} \cdot \varepsilon_i, i = 1, ..., n; \qquad (3.6.3)$$

- *экспоненциальная модель*

$$y_i = e^{\beta_0 + \beta_1 \cdot x_{i1} + \cdots + \beta_k \cdot x_{ik}} \cdot \varepsilon_i, i = 1, ..., n. \qquad (3.6.4)$$

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 логарифми-

рование обеих частей модели. Например, после логарифмирования модель (3.6.3) примет вид:

$$\ln(y_i) = \ln(\beta_0) + \beta_1 \ln(x_{i1}) + \cdots + \beta_k \ln(x_{ik}) + \ln(\varepsilon_i), i = 1, ..., n.$$

Вводя новый параметр $\beta_0' = \ln(\beta_0)$ и новые переменные $Z_i = \ln(X_i),\ Y' = \ln(Y)$, приходим к новой линейной модели

$$y_i' = \beta_0' + \beta_1 z_{i1} + \cdots + \beta_k z_{ik} + \ln(\varepsilon_i), i = 1, ..., n. \qquad (3.6.5)$$

Используя МНК, вычисляем оценки $b_0', b_1, ..., b_k$ 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

$$Q = A \cdot K^{\beta_1} \cdot L^{\beta_2} \cdot \varepsilon. \qquad (3.6.6)$$

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).

Таблица 3.3

Q	657	1200	2427	4257	8095	9849
L	162	245	452	714	1083	1564
K	279	1167	3069	5585	9119	13989

Решение. Логарифмируя обе части выражение (3.6.6) получаем следующую модель

$$\ln(Q) = \ln(A) + \beta_1 \ln(K) + \beta_2 \ln(L) + \ln(\varepsilon). \qquad (3.6.7)$$

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

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon .$$

Для вычисления оценок b_0, b_1, b_2 используем режим *Регрессия* табличного процессора Excel (см. § 4.5). 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

$$\hat{y} = 0.603 + 1.016 x_1 + 0.127 x_2 . \qquad (3.6.8)$$

Определим $A = e^{b_0} = e^{0.603} = 1.828$, и, возвращаясь к прежним обозначениям запишем выборочное уравнение регрессии для производственной функции Кобба-Дугласа:

$$\hat{Q} = 1.826 \cdot K^{1.016} \cdot L^{0.127} . \quad ☻$$

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 коррелированны 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.

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 Л3.1 приведены данные для построения линейной множественной регрессии.

Таблица Л3.1

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 Л3.1 (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 $\hat{R}^2$. Высказать мнение, насколько хорошо построенная регрессия определяет зависимость переменной Y от объясняющих переменных X_1, X_2.

Контрольные результаты:

1. Значения вычисленных коэффициентов

$$b_0 = 213.506,\ b_1 = -1.170,\ b_2 = 8.533.$$

Значение критерия Фишера 62.879.

2. Значение коэффициента детерминации $R^2 = 0.940$, скорректированного коэффициента детерминации $\hat{R}^2 = 0.925$.

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 литья (объясняющая переменная X_2 в %). Определить значимость коэффициентов b_0, b_1, b_2.

Исходные данные. В таблице Л3.1 приведены данные для построения линейной множественной регрессии.

Содержание работы

1. Ввести в лист Excel исходные данные таблицы Л3.1 (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 $M(Y \mid x)$, 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 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), построить модель множественной линейной регрессии и выполнить статистический анализ построенной модели.

Для вычисления коэффициентов уравнения регрессии

$$\hat{y}(x) = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$$

и других характеристик множественной регрессии использовать режим *Регрессия* табличного процессора Excel (см. пример 3.6.1).

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.

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

$$\hat{y}(x) = b_0 + b_1 x_1 + b_2 x_2$$

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 величинам, 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).

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 коэффициентов уравнения регрессии $\hat{y}(x) = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2$.

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?

Глава 4. Практические аспекты множественного регрессионного анализа

В предыдущих главах была изучена классическая линейная модель регрессии (как парной, так и множественной), приведены оценки ее коэффициентов и выполнен статистический анализ построенных уравнений регрессии. Однако на практике возникает ряд проблем, связанных с невыполнением допущений классической модели и другими обстоятельствами, не рассмотренными ранее.

В этой главе будут рассмотрены аспекты, связанные с мультиколлинеарностью регрессионной модели, отбором объясняющих переменных и нарушением условия гомоскедастичности. Вопросы, обусловленные коррелированностью возмущений ε_i, будут рассмотрены позже при анализе временных рядов.

4.1. 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^TX. Тогда обратная матрица $\left(X^TX\right)^{-1}$ 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 $\left(X^TX\right)^{-1}$ 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^TX хотя и является невырожденной, но величина определителя $|X^TX|$ мала, а, следовательно, элементы обратной матрицы $\left(X^TX\right)^{-1}$ становятся очень большими. В результате получаются большие дисперсии $\sigma^2_{b_j}$ (или их оценки $s^2_{b_j}$) 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

$$\left(X^TX\right)b = X^Ty \qquad (4.1.1)$$

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 менее, имеются некоторые эвристические подходы по ее выявлению. Кратко остановимся на некоторых из них.

- ***Анализ матрицы парных коэффициентов корреляции объясняющих переменных.*** По исходным данным $X_{i,j}$, $i=1,2,\ldots,n$; $j=1,2,\ldots,k$ 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.
- ***Анализ собственных чисел матрицы X^TX.*** Собственным числом λ матрицы X^TX называется скалярная величина, входящая в тождество

$$\left(X^TX\right)z = \lambda z,$$

где z – вектор, называемый собственным. Матрица X^TX размером $m\times m$ имеет m собственных чисел. Значительное отклонение максимального собственного числа λ_{max} от минимального λ_{min} ($\lambda_{max}/\lambda_{min} > 1000$) свидетельствует о наличии мультиколлинеарности.

• ***Анализ числа обусловленности матрицы X^TX.*** Числом обусловленности матрицы X^TX называется величина $cond(X^TX)$, определяемая как (если $\left(X^TX\right)^{-1}$ существует):

$$cond(X^TX) = \left\|X^TX\right\| \cdot \left\|\left(X^TX\right)^{-1}\right\|, \qquad (4.1.2)$$

где норма $\|A\|$ матрицы A размера $m \times m$ определяется как

$$\|A\| = \sqrt{\sum_{j=1}^{m}\sum_{i=1}^{m} A_{i,j}^2}\ .$$

Число обусловленности удовлетворяет условию

$$1 \le cond(A) < \infty$$

и $cond(A) = \infty$, если матрица A вырождена. Справедливо следующее неравенство

$$\frac{\left\|b-\tilde{b}\right\|}{\|b\|} \le cond(X^TX)\frac{\|y-\tilde{y}\|}{\|y\|}, \qquad (4.1.3)$$

где вектор $\tilde{b}$ – решение системы $(X^TX)\tilde{b} = X^T\tilde{y}$, вектор b – решение системы $(X^TX)b = X^Ty$, а норма вектора $\|y\|$ равна $\|y\| = \sqrt{\sum_{i=1}^{n} y_i^2}$. Видно, что чем больше число обусловленной матрицы X^TX, тем с бóльшим «коэффициентом усиления» относительная погрешность задания правой части (отношение $\|y-\tilde{y}\|/\|y\|$) передается в относительную погрешность вычисления вектора коэффициентов (отношение $\left\|b-\tilde{b}\right\|/\|b\|$). Значительная величина числа обусловленности матрицы X^TX ($cond\ (X^TX) > 1000$) свидетельствует о мультиколлинеарности, а матрицу с таким числом обусловленности *называют плохо обусловленной*.

• ***Анализ множественного коэффициента детерминации.*** Находят множественный коэффициент детерминации между одной из объясняющих переменных и некоторой группой из них. Наличие высокого значения (больше 0.6) говорит о наличии мультиколлинеарности.

• ***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 корреляционны между собой, то матрица парных корреляций объясняющих переменных являются единичной матрицей размера $\kappa \times k$ 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.

$$H_0 : \det(R_x) = 1;$$
$$H_1 : \det(R_x) \ne 1.$$

Для проверки этих гипотез определим критерии:

$$K_R = n - 1 - \frac{1}{6}(2k+5)\lg(\det(R_x)),$$

где k – число объясняющих переменных, n – количество наблюдений. При справедливости гипотезы H_0 величина K_R подчиня-

ется 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 степенями свободы порядка $1-\alpha$, вычисляемой с помощью функции Excel $\chi^2_{1-\alpha,l} = \text{ХИ2ОБР}(\alpha; l)$. Если выполняется неравенство

$$K_R > \chi^2_{1-\alpha,l},$$

то с уровнем значимости α отвергается гипотеза H_0 и принимается гипотеза о наличии мультиколлинеарности.

Методы устранения или уменьшения мультиколлинеарсти. 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 переменную, то наличие мультиколлинеарности исказит истинные зависимости между переменными.

К сожалению, не существует единого универсального метода устранения мультиколлинеарности. Рассмотрим основные методы.

- ***Исключение переменной (или переменных) из модели.*** Состоит в том, что из двух объясняющих переменных, имеющих высокий коэффициент парной корреляции (больше 0.8), *одну переменную исключают из модели*. При этом, какую переменную оставить, решают на основании экономических соображений. Если ни одной из переменных нельзя отдать предпочтение, то оставляют ту переменную, которая имеет бóльший коэффициент корреляции с зависимой переменной.
- ***Изменение спецификации модели***. Состоит в *добавлении объясняющих переменных*, не учтенных в первоначальной модели, но существенно влияющих на зависимую переменную.
- ***Преобразование переменных модели***. Состоит в *переходе от исходных переменных* $X_1, X_2,\ldots,X_k$, связанных тесной корреляционной связью, к *новым переменным*, которые между собой слабо коррелированны или вообще не коррелированны.

Например, пусть в регрессии

$$\hat{y} = b_0 + b_1x_1 + b_2x_2$$

переменные x_1 и x_2 коррелированны. В этой ситуации можно построить регрессию относительно новых величин

$$\frac{\hat{y}}{x_2} = b_0 + b_1\frac{x_1}{x_2}.$$

Вполне вероятно, что в этой модели проблема мультиколлинеарности будет отсутствовать.

- ***Вычисление смещенных оценок***. Устойчивое (к возмущениям правой части) решение системы

$$(X^TX)b = X^Ty \qquad (4.1.4)$$

с плохо обусловленной матрицей X^TX можно получить, применяя метод регуляризации А. Н. Тихонова. Суть этого метода заключается в том, что вместо системы (4.1.4) решают «регуляризированную» систему

$$(X^TX + \alpha I)b_\alpha = X^Ty, \qquad (4.1.5)$$

где $\alpha > 0$ – параметр регуляризации, I – единичная (m×m)-матрица. Подбором соответствующей величины α, можно получить решение b_α, имеющее меньшую ошибку по сравнению с решением системы (4.1.4). Однако выбор параметра регуляризации α является на практике весьма трудным вопросом и требует априорной информации о точности задания вектора y (в частности, задания дисперсии σ^2).

Пример 4.1.1. Пусть по данным бюджетного обследования двадцати случайно выбранных семей изучалась зависимость накопления (Y) от дохода (X_1), расходов на питание (X_2) и стоимости имущества (X_3). По выборке бала вычислена матрица парных коэффициентов корреляции, представленная в таблице 4.1.1.

Определите, какие переменные целесообразно включить в линейную множественную регрессию?

Решение. Прежде всего, определим наличие переменных с высоким коэффициентом корреляции (> 0.8). К таким переменным относятся X_1 и X_2 ($r_{X_1 X_2} = 0.93$). Поэтому исключаем X_2 и оставляем переменную X_1, имеющую более высокий коэффициент корреляции с переменной Y: $r_{Y X_1} = 0.85$. Тогда уравнение регрессии имеет вид

$$\hat{y} = b_0 + b_1 x_1 + b_3 x_3. \quad \blacksquare$$

Таблица 4.1.1

	Y	X_1	X_2	X_3
Y	1			
X_1	0.85	1		
X_2	0.81	0.93	1	
X_3	–0.65	–0.38	–0.28	1

4.2. Отбор объясняющих переменных регрессионной модели

Вопросы выбора значимых (информативных) объясняющих переменных играет важную роль при построении множественной регрессионной модели, другими словами, при определении *размерности модели.*

При не учете информативных объясняющих переменных модель будет не полной (заниженная размерность модели), и содержать систематическую ошибку, т.е. M(ε) ≠ 0, а сами коэффициенты b_j являются смещенными оценками. Смещенной окажется оценка s^2 для дисперсии σ^2, а, следовательно, стандартные ошибки и многие статистические тесты, в которых используется s^2, становятся некорректными.

Введение «лишних» (слабо информативных) объясняющих переменных может привести к появлению мультиколлинеарности, а также уменьшится точность построенной модели, обусловленная «переходом» возмущения ε в коэффициенты модели и оценки b_j становятся неэффективными (завышенная размерность модели). Таким образом, возникает *проблема отбора наиболее значимых объясняющих переменных регрессионной модели, или иначе, определение оптимальной размерности модели*.

Наиболее эффективным подходом к решению этой проблемы является использование *пошаговых процедур отбора наиболее информативных объясняющих переменных*. Возможны два варианта пошаговых процедур:

- *процедура добавления объясняющих переменных* – начиная с некоторой «минимальной» регрессионной модели идет *добавление наиболее значимой* на данном шаге объясняющей переменной (такая процедура подробно описывается ниже);
- *процедура удаления объясняющих переменных* – начиная с некоторой «максимальной» регрессионной модели идет *удаление наименее значимой* на данном шаге объясняющей переменной.

Какой из этих процедур отдать предпочтение? Каждая из процедур имеет свои плюсы и минусы, но более предпочтительной представляется процедура добавления объясняющих переменных. Приведем некоторые доводы в пользу этой процедуры.

Во-первых, начиная с «минимальной» модели (предельный случай: одна объясняющая и одна зависимая переменные) для вычисления коэффициентов уравнения регрессии b_j решается система нормальных уравнений небольшой размерности (в предельном случае – 2×2) и небольшим числом обусловленности (несколько десятков или меньше). Это позволяет достаточно точно оценить коэффициенты регрессии и сделать корректные выводы об их значимости (с помощью *t*-критерия). По мере добавления объясняющих переменных размерность системы увеличивается и увеличивается число обусловленности (достигая тысяч и даже сотен тысяч). Это приводит к известным уже трудностям (см. § 4.1).

Во-вторых, принцип добавления переменных хорошо согласуется с *принципом минимальной сложности*, используемый в различных областях науки и техники. Суть его заключается в том, что при разработке конструкции (системы, устройства и т.д.) первоначально берутся простейшие и переход к более сложным конструкциям осуществляется только тогда, когда предыдущие конструкции не удовлетворяют заданным критериям. В нашем случае конструкция – это регрессионная модель, ее сложность определяется количеством объясняющих переменных, а критерием может служить скорректированный коэффициент детерминации.

Процедура добавления объясняющих переменных. *На первом шаге* этой процедуры рассматривается лишь одна объясняющая и переменная, имеющая с зависимой переменной наибольший коэффициент детерминации R^2. Напомним, что для получаемых на этом шаге парных регрессий R^2 равняется квадрату соответствующего коэффициента корреляции.

На втором шаге в регрессионную модель включается новая объясняющая переменная, которая вместе с первоначально отобранной образуют пару объясняющих переменных, имеющую с Y наиболее высокий (по сравнению с другими возможными парами) скорректированный коэффициент детерминации.

На третьем шаге в уравнение регрессии вводится еще одна объясняющая переменная, которая вместе с двумя первоначально отобранными образуют тройку переменных, имеющую с Y наибольший (по сравнению с другими возможными тройками) скорректированный коэффициент детерминации и т.д.

Процедура добавления новых переменных продолжается до тех пор, пока будет увеличиваться соответствующий скорректированный коэффициент детерминации.

Пример 4.2.1. По данным $n = 20$ сельскохозяйственных районов области исследуется зависимость переменной Y – урожайность зерновых культур (в ц/га) от ряда переменных – факторов сельскохозяйственного производства:

X_1 – число тракторов (приведенной мощности на 100 га);

X_2 – число зерновых комбайнов (на 100 га);

X_3 – число орудий поверхностной обработки почвы (на 100 га);

X_4 – количество удобрений, расходуемых на 1 га (т/га);

X_5 – количество химических средств защиты растений, расходуемых на 1 га (ц/га).

Матрица парных корреляций приведена в таблице 4.2.

Таблица 4.2

Переменные	Y	X_1	X_2	X_3	X_4	X_5
Y	1	0.43	0.37	0.40	0.58*	0.33
X_1	0.43	1	0.85*	0.98*	0.11	0.34
X_2	0.37	0.85*	1	0.88*	0.03	0.46*
X_3	0.40	0.98*	0.88*	1	0.03	0.28
X_4	0.58*	0.11	0.03	0.03	1	0.57*
X_5	0.33	0.34	0.46*	0.28	0.57*	1

Знаком * отмечены корреляции, значимые по t-критерию (2.5.5) на уровне значимости 0.05 ($\alpha = 0.05$).

Используя пошаговую процедуру добавления объясняющих переменных, составить уравнение регрессии.

Первоначально проанализируем матрицу парных корреляций для выявления пар, имеющих сильную статистическую зависимость (коэффициент корреляции > 0.8), т.е. мультиколлинеарность (см. §5.1). К таким парам можно отнести:

- X_1 и X_2 (коэффициент корреляции $r_{12} = 0.85$);
- X_1 и X_3 (коэффициент корреляции $r_{13} = 0.98$);
- X_2 и X_3 (коэффициент корреляции $r_{23} = 0.88$).

Следовательно, для устранения мультиколлинеарности часть из этих переменных не следует включать в уравнение регрессии при добавлении переменных.

Пошаговую процедуру отбора представим следующими шагами (при этом приводятся только результаты выполнения необходимых вычислений):

Шаг 1. Из объясняющих переменных $X_1 \div X_5$ выбирается переменная X_4, имеющая с зависимой переменной Y наибольший коэффициент детерминации $R^2_{Y,4} = 0.58^2 = 0.336$ (напомним, что для парной регрессии R^2 равен квадрату соответствующего коэффициента корреляции). Скорректированный коэффициент детерминации $\hat{R}^2_{Y,4} = 1 - \frac{19}{18}(1 - 0.336) = 0.299$. Таким образом, получена парная регрессия $\hat{y}\,(x) = b_0 + b_4 x_4$.

Шаг 2. Для всевозможных пар (X_4, X_j), j = 1,2,3,5 строятся уравнения регрессии $\hat{y}(x) = b_0 + b_4 x_4 + b_j x_j$, j = 1,2,3,5, и для каждого уравнения вычисляют коэффициент детерминации. Сравнение этих коэффициентов показывает, что максимальное значение имеет пара (X_4, X_3), для которой $R^2_{Y,43} = 0.483$, а скорректированный коэффициент детерминации $\hat{R}^2_{Y,43} = 1 - \frac{19}{17}(1 - 0.483) = 0.422$. Видно, что значение $\hat{R}^2$ увеличилось и результатом второго этапа становится уравнение регрессии $\hat{y}(x) = b_0 + b_3 x_3 + b_4 x_4$.

Шаг 3. Для всевозможных троек (X_3, X_4, X_j), j = 1,2,5 строятся уравнения $\hat{y}\,(x) = b_0 + b_3 x_3 + b_4 x_4 + b_j x_j$, j = 1,2,5, и для каждого уравнения вычисляется коэффициент детерминации. Наиболее информативной оказалась тройка (X_4, X_3, X_5), имеющая максимальный коэффициент детерминации $R^2_{Y,435} = 0.513$. Скорректированный коэффициент детерминации $\hat{R}^2_{Y,435} = 0.422$.

Замечание 4.2.1. Так как на втором шаге переменная X_3 была включена в уравнение регрессии, то на третьем шаге можно было не рассматривать тройки (X_4, X_3, X_1) и (X_4, X_3, X_2), так как X_2 и X_1 статистически связаны с X_3 и введение их в уравнение вызвало бы мультиколлинеарность получаемой регрессии.

Вернемся к коэффициенту $\hat{R}^2_{Y,435} = 0.422$ и сравним его с $\hat{R}^2_{Y,43} = 0.422$, вычисленном на втором шаге. Так как скорректированный коэффициент детерминации на третьем шаге не увеличился, то в уравнении регрессии *достаточно ограничиться лишь двумя отобранными ранее переменными* X_4 и X_3. После вычисления коэффициентов b_0, b_3, b_4, получаем

$$\hat{y} = 7.29 + 3.48x_3 + 3.48x_4 .$$
$$(11.0) \quad (26.8) \quad (2.25)$$

В скобках приведены вычисленные значения t-статистик T_{b_j}, $j = 0, 3, 4$ (см. (3.4.1)), которые все больше критической точки $t(0.95,\ 20 - 3) = 2.11$. Следовательно, *коэффициенты* b_0, b_3, b_4 *являются значимыми* на уровне значимости $\alpha = 0.05$.

Замечание 4.2.2. Если в уравнение регрессии включить все пять переменных, то оно примет вид

$$\hat{y} = 3.515 - 0.006x_1 + 15.542x_2 + 0.110x_3 + 4.475x_4 - 2.932x_5$$
$$(0.65) \quad (0.01) \quad (0.72) \quad (0.13) \quad (2.91) \quad (0.95)$$

Сравнивая значения t-статистик, приведенные в скобках, с критической величиной $t(0.95,\ 14) = 2.14$, видим, что значимым (на уровне $\alpha = 0.05$) является только коэффициент b_4. В то же время уравнение регрессии значимо (на уровне 0.05):

$$F = 3.00 > F_{0.95;\ 5;\ 14} = 2.96.$$

Такая ситуация (большинство коэффициентов b_j не значимы, а уравнение регрессии в целом значимо) является следствием мультиколлинеарности модели, включающей все объясняющие переменные.

Замечание 4.2.3. Для определения «момента» прекращения добавления объясняющих переменных в регрессионную модель рассматривается рост величины скорректированного коэффициента детерминации $\hat{R}^2$, не требующего априорного задания величины дисперсии σ^2. В задачах линейной параметрической идентификации динамических систем, построения среднеквадра-

тических приближений для определения «оптимальной размерности» также используются критерии, не требующие задания σ^2. Кратко остановимся на двух из этих критериев, обозначив через m - количество искомых параметров, т.е. «размерность регрессии».

Оценивание оптимальной размерности на основе метода перекрестной значимости. В качестве m принимается значение m_{GCV}, доставляющее минимум функционалу

$$GCV(m) = \frac{\frac{1}{n}\sum_{i=1}^{n} e_i^2}{\left[1 - \frac{m}{n}\right]^2}$$

Оценивание оптимальной размерности на основе информационного критерия. В качестве m принимается значение, доставляющее минимум функционалу

$$AIC(m) = n \cdot \ln\left(\frac{1}{n}\sum_{i=1}^{n} e_i^2\right) + 2m\,.$$

Эти критерии можно рекомендовать как полезное дополнение к критерию $\hat{R}^2$ при определении количества объясняющих переменных в регрессионной модели.

4.3. Фиктивные переменные в линейных регрессионных моделях

До сих пор рассматривались регрессионные модели, в которых в качестве объясняющих переменных (регрессоров) выступали количественные переменные (толщина угольного пласта, уровень механизации и т.д.), которые принимали значение из некоторого непрерывного интервала. Однако на практике достаточно часто возникает необходимость исследования влияния *качественных признаков*, имеющих два или несколько уровней (градаций). Например, пол (мужской, женский), фактор сезонности (зима, весна, лето, осень) и т.д.

Фиктивные переменные. Качественные признаки могут существенно влиять на структуру линейных связей между переменными и приводить к скачкообразному изменению параметров регрессионной модели. В этом случае говорят о *регрессионных моделях с переменной структурой*.

Например, надо изучить зависимость размера заработной платы Y не только от количественных факторов X_1, X_2,…,X_k, но и от качественного признака Z_1 (например, от пола работника). В принципе можно получить оценки регрессионной модели

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + \varepsilon_i \qquad (4.3.1)$$

для каждого уровня качественного признака (т.е. одно уравнение регрессии для мужчин, второе – для женщин), а затем изучать различия между ними.

Но есть и другой подход, позволяющий оценивать влияние количественных переменных и уровней качественных признаков с помощью одного уравнения регрессии благодаря введению так называемых *фиктивных* (или *структурных*) *переменных*. В качестве фиктивных переменных обычно используются бинарные (булевы) переменные, которые принимают всего два значения: «0» или «1» (например, значение переменной $Z_1 = 0$ для работников-женщин и $Z_1 = 1$ – для мужчин). В этом случае первоначальная модель (4.3.1) изменится и примет вид:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots \beta_k x_{ik} + \alpha_1 z_{i1} + \varepsilon_i, \qquad (4.3.2)$$

где

$$z_{i1} = \begin{cases} 1, & \text{если } i \text{ – ый работник мужчина;} \\ 0, & \text{если } i \text{ – ый работник женщина.} \end{cases}$$

Таким образом, работая с моделью (4.3.2), видно, что заработная плата у мужчин на $\alpha_1 \cdot 1 = \alpha_1$ выше, чем у женщин. Проверяя гипотезу H_0: $\alpha_1 = 0$, можно установит влияние фактора «пол» на размер заработной платы.

Пример 4.3.1. Необходимо построить линейную регрессионную модель для исследования зависимости между результатами

письменных вступительных экзаменов (объясняющая переменная X – число решенных задач), полом студента (качественный признак «пол» - мужчина, женщина) и успешно сданными курсовыми работами на первом курсе (зависимая переменная Y – число курсовых работ). Исходные данные представлены в таблице 4.3.

Таблица 4.3

№ студента	x_i	y_i	Пол	№ студента	x_i	y_i	Пол
1	10	6	М	7	6	3	Ж
2	6	4	Ж	8	7	4	М
3	8	4	М	9	9	7	М
4	8	5	Ж	10	6	3	Ж
5	6	4	Ж	11	5	2	М
6	7	7	М	12	7	3	Ж

Решение. Первоначально рассмотрим регрессионную модель, не учитывающую признака «пол», т.е.

$$Y = \beta_0 + \beta_1 X + \varepsilon \qquad (4.3.3)$$

Методами парной регрессии вычислим оценки $b_0 = -1.437$, $b_1 = 0.815$ и уравнение регрессии примет вид:

$$\hat{y} = -1.437 + 0.815\,x \qquad (4.3.4)$$

Коэффициент детерминации $R^2_{Y,X} = 0.530$, и уравнение значимо по F-критерию на уровне 0.05:

$$F = 9.46 > F_{0.95;\,1;\,10} = 4.96.$$

Введем фиктивную переменную Z со значениями:

$$Z = \begin{cases} 1, & \text{если студент мужского пола;} \\ 0, & \text{если студент женского пола,} \end{cases}$$

и рассмотрим регрессионную модель вида

$$Y = \beta_0 + \beta_1 X + aZ + \varepsilon \qquad (4.3.5)$$

Для вычисления коэффициентов соответствующего уравнения регрессии

$$\hat{y} = b_0 + b_1 x + a\,z \qquad (4.3.6)$$

сформируем матрицу X (размером 12×3) и вектор коэффициентов (размерности 3):

$$X = \begin{vmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 10 & 6 & 8 & 8 & 6 & 7 & 6 & 7 & 9 & 6 & 5 & 7 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & 1 & 0 \end{vmatrix}^T ; \quad b = \begin{vmatrix} b_0 \\ b_1 \\ a \end{vmatrix}.$$

Решая систему нормальных уравнений

$$\left(X^T X\right) b = X^T y\,,$$

получаем $b = \left|-1.165,\ 0.743,\ 0.466\right|^T$, а уравнение регрессии (4.3.6) примет вид

$$\hat{y} = -\,1.165 + 0.743x + 0.466z \qquad (4.3.7)$$

Полученное уравнение регрессии значимо по F-критерию на уровне 0.05:

$$F = 5.48 > F_{0.95;\,2;\,9} = 4.26.$$

Коэффициент детерминации $R^2_{Y,XZ} = 0.549$ несколько выше, чем $R^2_{Y,X} = 0.530$ у парной регрессии.

Таким образом, из уравнения (4.3.7) следует, что при том же числе решенных задач на вступительных экзаменах (переменная X) юноши сдают успешно на $0.466 \approx 0.5$ курсовых работ больше. Это различие видно из рис. 4.1, на котором построены графики регрессии (4.3.7) при $Z = 0$ и $Z = 1$.

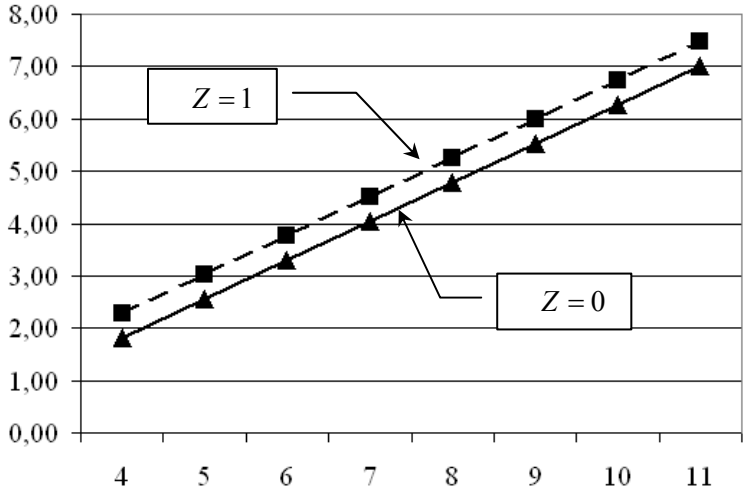

Рис. 4.1. Уравнение регрессии с фиктивными переменными

Для того, чтобы убедиться в справедливости сделанного вывода проверим по *t*-критерию значимость коэффициента *a*:

$$T_a = 1.15 < t\,(0.95, 9) = 2.26.$$

Следовательно, на уровне значимости $\alpha = 0.05$ можно принять гипотезу H_0: $a = 0$ в модели (4.3.5), т.е. влияние фактора «пол» на сдачу курсовых работ оказалось не существенным. К этому выводу приходим, сравнивая скорректированный коэффициент детерминации: $\hat{R}^2_{Y,X} = 0.483$ (для уравнения (4.3.4)) и $\hat{R}^2_{Y,XZ} = 0.449$ (для уравнения (4.3.7)). Видно, что ввод фиктивной переменной Z уменьшил значение скорректированного коэффициента детерминации. Следовательно, фиктивную переменную Z вводить в модель нецелесообразно.

Введение нескольких фиктивных переменных. Если рассматриваемый качественный признак имеет несколько уровней (градаций), то в принципе можно в модель ввести одну фиктивную переменную – дискретную переменную, принимающую такое же количество значений. Однако из-за трудности содержательной интерпретации вычисляемых коэффициентов регрессии обычно вводят несколько бинарных переменных.

Пример 4.3.2. Предположим, что необходимо построить регрессионную модель для исследования зависимости между весом новорожденного и семейным положением матери, а также рожала ли она раньше.

Очевидно, возможны следующие возможные случаи:

1. Замужняя мать, первые роды.
2. Одинокая мать, первые роды.
3. Замужняя мать, не первые роды.
4. Одинокая мать, не первые роды.

В принципе можно ввести одну фиктивную переменную Z, принимающую значения от 0 до 3 (четыре уровня). Однако из-за трудностей последующей интерпретации введем две фиктивных бинарных переменных Z_1, Z_2 со следующими значениями:

$$Z_1 = \begin{cases} 1, & \text{если мать одинока;} \\ 0, & \text{во всех остальных случаях;} \end{cases}$$

$$Z_2 = \begin{cases} 1, & \text{если мать рожала в прошлом;} \\ 0, & \text{во всех остальных случаях.} \end{cases}$$

Тогда выше перечисленным случаям соответствуют следующие значения фиктивных переменных:1)$Z_1 = 0$; $Z_2 = 0$; 2)$Z_1 = 1$; $Z_2 = 0$; 3)$Z_1 = 0$; $Z_2 = 1$; 4)$Z_1 = 1$; $Z_2 = 1$; Использование двух фиктивных переменных упрощает интерпретацию построенной модели. Так коэффициент a_1 при Z_1 представляет разность веса новорожденного, если мать одинока (ожидаем отрицательный знак у коэффициента), а коэффициент a_2 при Z_2 будет определять дополнительный вес новорожденного, если ребенок не является первенцем (ожидаем положительный знак у коэффициента).

4.4. Частная корреляция

Выше для оценки тесноты связи между одной зависимой и одной независимой переменной был введен выборочный коэф-

фициент корреляции. В случае нескольких переменных высокое значение коэффициента корреляции может означать высокую степень линейной зависимости, но может означать и то, что третья переменная, оказывает значительной влияние на две первых переменных и, что именно она служит основной причиной их высокой корреляции. Поэтому необходимо найти «чистую» корреляцию между двумя переменными, исключив влияние других факторов. Это осуществляется путем вычисления коэффициента *частной корреляции.*

Выборочным частным коэффициентом корреляции (или просто ***частным коэффициентом корреляции***) между переменными X_i и X_j при фиксированных значениях остальных $(k-2)$ переменных называется выражение

$$r_{X_iX_j(X_1,X_2,\ldots,X_k)} = \frac{-q_{ij}}{\sqrt{q_{ii}q_{jj}}}, \qquad (4.4.1)$$

где q_{ij} – алгебраическое дополнение элемента r_{ij} матрицы выборочных парных коэффициентов корреляции (размера $k \times k$)

$$R_X = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1k} \\ r_{21} & 1 & \cdots & r_{2k} \\ \cdots & \cdots & \cdots & \cdots \\ r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix},$$

а r_{ij} – выборочный коэффициент корреляции между переменными X_i, X_j. В скобках записи $r_{X_iX_j(X_1,X_2,\ldots,X_k)}$ указываются имена переменных, влияние которых исключаются при вычислении частной корреляции. Количество переменных в скобках определяет порядок коэффициент частный корреляции.

Аналогично можно определить коэффициент частный корреляции между зависимой переменной Y и независимыми переменными $X_1,\ldots,X_k$, рассматривая Y как дополнительную переменную X_{k+1}. Так коэффициенты частной корреляции (первого

порядка) для уравнения регрессии с двумя независимыми переменными рассчитываются по формулам:

$$r_{YX_1(X_2)} = \frac{r_{YX_1} - r_{YX_2} \cdot r_{X_1X_2}}{\sqrt{(1-r_{YX_2}^2)\cdot(1-r_{X_1X_2}^2)}}, \qquad (4.4.2)$$

$$r_{YX_2(X_1)} = \frac{r_{YX_2} - r_{YX_1} \cdot r_{X_1X_2}}{\sqrt{(1-r_{YX_1}^2)\cdot(1-r_{X_1X_2}^2)}}, \qquad (4.4.3)$$

$$r_{X_1X_2(Y)} = \frac{r_{X_1X_2} - r_{YX_1} \cdot r_{YX_2}}{\sqrt{(1-r_{YX_1}^2)\cdot(1-r_{YX_2}^2)}}. \qquad (4.4.4)$$

Частный коэффициент корреляции $r_{X_iX_j(X_1,X_2,\ldots,X_k)}$, как и парный коэффициент $r_{X_iX_j}$, может принимать значения от –1 до +1.

Для проверки значимости частного коэффициента корреляции вычисляют статистику

$$T_r' = \frac{r_{YX_i(X_1,X_2,\ldots,X_k)}}{\sqrt{1-(r_{YX_i(X_1,X_2,\ldots,X_k)})^2}} \cdot \sqrt{n-h-2}, \qquad (4.4.5)$$

где h - число исключенных переменных. Для частных коэффициентов корреляции (4.4.2) ÷ (4.4.4) $h=1$.

Частный коэффициент корреляции $r_{YX_i(X_1,X_2,\ldots,X_k)}$ является значимым с уровнем значимости α, если выполняется неравенство

$$|T_r'| > t(1-\alpha, n-h-2) \qquad (4.4.6)$$

Необходимо заметить, что в эконометрике частные коэффициенты корреляции обычно не имеют самостоятельного значения. В основном их используют на стадии формирования модели при отборе значимых переменных. Так, строя многофакторную модель, на первом шаге определяют уравнение регрессии с полным набором факторов, и рассчитывается матрица частных ко-

эффициентов корреляции. На втором шаге из модели исключается фактор с наименьшей и незначимой по критерию (4.4.6) величиной частного коэффициента корреляции $r_{YX_i(X_1,X_2,\dots,X_k)}$. Исключив этот фактор из модели, строится новое уравнение регрессии. Такая процедура исключения продолжается до тех пор, пока не окажется, что все частные коэффициенты корреляции значимы.

Пример 4.4.1. Для построения зависимости между стоимостью грузовой автомобильной перевозки (переменная Y - тыс. руб.), весом груза (X_1 - тонны) и расстоянием (X_3 – тыс. км.) по выборке объемом $n = 20$ были вычислены следующие коэффициенты корреляции:

$$r_{YX_1} = 0.665\,(3.68), \quad r_{YX_2} = 0.6345\,(3.60), \quad r_{X_1X_2} = 0.125\,(2.80).$$

В скобках указаны значения T – статистик. Необходимо вычислить частные коэффициенты корреляции и определить наиболее значимые факторы.

Решение. По формулам (4.4.2) ÷ (4.4.4) вычислим частные коэффициенты корреляции, а по формуле (4.4.5) – значения соответствующих T – статистик (приведены в круглых скобках).

$$r_{YX_1(X_2)} = \frac{0.655 - 0.635 \cdot 0.122}{\sqrt{(1-(0.635)^2)\cdot(1-(0.122)^2)}} = 0.751 \quad (T'_r = 4.69),$$

$$r_{YX_1(X_2)} = \frac{0.635 - 0.655 \cdot 0.122}{\sqrt{(1-(0.655)^2)\cdot(1-(0.122)^2)}} = 0.738 \quad (T'_r = 4.51),$$

$$r_{X_1X_2(Y)} = \frac{0.125 - 0.655 \cdot 0.635}{\sqrt{(1-(0.655)^2)\cdot(1-(0.635)^2)}} = -0.499 \quad (T'_r = -2.37).$$

Вычислим $t(1-\alpha, n-h-2))$ при $\alpha = 0.05, n = 20, h = 1$, используя известную формулу (3.3.6):

$$t(1-\alpha, n-h-2) = \text{СТЬЮДРАСПОБР}(\alpha; n-h-2).$$

Получаем $t(0.95, 20-3) = 2.11$. Анализируя результаты вычислений, можно сделать следующие выводы:

- все вычисленные частные коэффициенты корреляции являются значимыми на уровне 0.05, так как соответствующие T – статистики удовлетворяют неравенству (4.4.6);
- наиболее сильной является взаимосвязь между стоимостью перевозки и весом груза;
- частные коэффициенты корреляции между Y и X_1, Y и X_2 свидетельствуют о более сильных взаимосвязях независимых переменных с зависимой, чем это показывают значения парных коэффициентов корреляции. ☻

4.5. 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 гетероскедакстично-

сти величины $|e_i|$ и x_i будут коррелированны. Степень коррелированности 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:

$$\rho_{xe} = 1 - \frac{6\sum_{i=1}^{n} d_i}{n^3 - n}, \qquad (4.5.1)$$

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

$$T_\rho = \frac{\rho_{xe}\sqrt{n-2}}{\sqrt{1-\rho_{xe}^2}}. \qquad (4.5.2)$$

Если выполняется неравенство

$$|T_\rho| > t(1-\alpha, n-2), \qquad (4.5.3)$$

то нулевая гипотеза H_0 отвергается с уровнем значимости α и принимается гипотеза о гетероскедакстичности модели. Значение $t(1-\alpha, n-2)$ вычисляется выражением

$$t(1-\alpha, n-2) = \text{СТЬЮДРАСПОБР}(\alpha; n-2).$$

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.

Пример 4.5.1. Проверить гипотезу о гомоскедастичности данных, представленных в таб. 3.1 и используемые для построения линейной регрессии. Значения коэффициентов уравнения регрессии взять из примера 3.3.1.

Решение. Первоначально введем в ячейки B1, B2 значения коэффициентов b_0, b_1 соответственно (см. рис. 4.2). Затем в ячейках C5 ÷ C14 запрограммируем вычисления значений $\hat{y}_i$ регрессионного уравнения $\hat{y}(x) = b_0 + b_1 x$ при $x = x_i$. После этого программируем вычисления модулей невязок $e_i = y_i - \hat{y}_i, i = 1,...,10$. 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. 4.2). В ячейке C27 вычис-

ляем $\sum_{i=1}^{10} d_i^2$ (см. рис. 4.3) , а в ячейке D28 – значение критерия по формуле (4.5.2), 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	$\hat{y}_i$	$\lvert e_i \rvert$	
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** $\lvert e_i \rvert$	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. 4.2. Вычисление коэффициента ранговой корреляции

Последним этапом является проверка неравенства (4.5.3). 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 (4.5.3) 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

$$V_\varepsilon = diag\left\{ \sigma_1^2, \sigma_2^2, ..., \sigma_n^2 \right\}. \qquad (4.5.4)$$

	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	СТЬЮДРАСПОБР(0,05;10-2)=	***2,31***			

Рис. 4.3. Продолжение рис. 4.2

Напомним, что для гомоскедастичной модели ковариационная матрица вектора возмущения ε определяется выражением

$$V_\varepsilon = diag\left\{ \sigma^2, \sigma^2, ..., \sigma^2 \right\}. \qquad (4.5.5)$$

Возникает вопрос: «Какими свойствами будет характеризоваться оценка классического МНК вида:

$$b = (X^T X)^{-1} \cdot X^T y \ , \qquad (4.5.6)$$

построенная для коэффициентов гетероскедастичной модели?». Дадим ответ на этот вопрос.

1. Оценка (4.5.6) и в этом случае будет *несщенной и состоятельной*. Это означает, что определение значений зависимой переменной можно осуществлять по уравнению регрессии с коэффициентами (4.5.6).

2. Оценка (4.5.6) *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 (3.2.1) 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

$$F(b)=\sum_{i=1}^{n}\frac{(y_i-\hat{y}_i)^2}{\sigma_i^2}=\sum_{i=1}^{n}\frac{(e_i)^2}{\sigma_i^2}=(y-Xb)^T\cdot V_\varepsilon^{-1}\cdot(y-Xb),\ (4.5.7)$$

где матрица V_ε определяется выражением (4.5.4). Можно показать, что вектор b, доставляющий минимум функционалу (4.5.7) является решением следующей матричной системы уравнений:

$$(X^TV_\varepsilon^{-1}X)b=X^TV_\varepsilon^{-1}y \qquad (4.5.8)$$

и определяется выражением:

$$b=(X^TV_\varepsilon^{-1}X)^{-1}X^TV_\varepsilon^{-1}y\,. \qquad (4.5.9)$$

Метод вычисление оценок на основе минимизации функционала (4.5.7) получил название *метода взвешенных наименьших квадратов,* а оценка (4.5.9) называется *оценкой метода взвешенных наименьших квадратов.* Эта оценка является *несмещенной, состоятельной и эффективной.* Ковариационная матрица вектора b определяется выражением:

$$V_b=(X^TV_\varepsilon^{-1}X)^{-1}. \qquad (4.5.10)$$

Поэтому, используя дисперсию $\sigma_{b_i}^2=[V_b]_{i,i}$ оценки b_i и соответствующие соотношения параграфов 4.3, 4.4, можно построить доверительные интервалы для коэффициентов β_i, проверить значимость b_i и в случае гетероскедастичности эконометрической модели.

КОНТРОЛЬНЫЕ ВОПРОСЫ И ЗАДАНИЯ

1. Поясните принцип «минимальной сложности» при отборе переменных модели множественной регрессии.

2. Построить регрессионную модель и вычислить оценки для ее коэффициентов для исследования влияния на вес куриных окорочков (переменная *Y*) возраста куры (переменная *X*) и страны происхождения куры (Америка, Канада) по данным таблицы 4.3.

Таблица 4.3

№ измерения	x_i	y_i	Страна	№ измерения	x_i	y_i	Страна
1	28	13.3	К	7	28	13.2	А
2	20	8.9	А	8	26	11.8	А
3	32	15.1	К	9	21	11.5	А
4	22	10.4	А	10	27	14.2	К
5	29	13.1	К	11	29	15.4	А
6	27	12.4	А	12	31	15.1	К

Проверить значимость построенной регрессии и коэффициентов регрессии.

Рекомендация. Введите в модель фиктивную переменную Z, равную

$$Z = \begin{cases} 1, & \text{если куры из Канады;} \\ 0, & \text{если куры из Америки.} \end{cases}$$

3. Имеются следующие данные о потреблении некоторого продукта Y (усл. ед.) в зависимости от уровня урбанизации (доли городского населения) X_1, относительного образовательного уровня X_2 и относительного заработка X_3 для девяти географических районов:

i **(*номер района*)**	x_{i1}	x_{i2}	x_{i3}	y_i
1	42.2	11.2	31.9	167.1
2	48.6	10.6	13.2	174.4
3	42.6	10.6	28.7	160.8
4	39.0	10.4	26.1	162.0
5	34.7	9.3	30.1	140.8
6	44.5	10.8	8.5	174.6
7	39.1	10.7	24.3	163.7
8	40.1	10.0	18.6	174.5
9	45.9	12.0	20.4	185.7

Средние значения $\bar{x}_1 = 41.85; \bar{x}_2 = 10.62; \bar{x}_3 = 24.42; \bar{y} = 167.07$.

Стандартные отклонения $s_{x_1} = 4.176$; $s_{x_2} = 0.7463$; $s_{x_3} = 7.928$;

$s_y = 12.645$. Корреляционная матрица:

	X_1	X_2	X_3	Y
X_1	1	0.684	–0.616	0.802
X_2	0.684	1	–0.173	0.770
X_3	–0.616	–0.173	1	–0.629
Y	0.802	0.770	–0.629	1

Используя пошаговую процедуру отбора наиболее информативных объясняющих переменных, определить подходящую регрессионную модель, исключив при этом мультиколлинеарность. Оценить значимость коэффициентов регрессии полученной модели по *t*-критерию.

Рекомендация. Для отбора наиболее значимых объясняющих переменных используйте процедуру добавления объясняющих переменных (параграф 4.2) и смотрите пример 4.2.1.

4. Какая идея положена в основу теста на гетероскедастичность модели?

5. В чем отличие между коэффициентом корреляции и частным коэффициентом корреляции?

6. С целью исследования влияния факторов X_1 – среднемесячного количества профилактических наладок автоматической линии и X_2 – среднемесячного числа обрывов нити на показатель Y – среднемесячную характеристику качества ткани (в баллах) по данным 37 предприятий легкой промышленности были вычислены парные коэффициенты корреляции: $r_{YX_1} = 0.105$, $r_{YX_2} = 0.024$ и $r_{X_1X_2} = 0.996$. Определить частные коэффициенты корреляции $r_{YX_1(X_2)}$ и $r_{YX_2(X_1)}$ и оценить их значимость на уровне 0.05.

7. Чем отличается взвешенный метод наименьших квадратов от классического МНК ?

ИНТЕРНЕТ – РЕСУРСЫ

1. **Воскобойников Ю.Е., Тимошенко Е.И**. Теория вероятностей. – Новосибирск: Новосибирский государственный архитектурно-строительный университет. 2003 (электронная версия книги размещена по адресу http://www.ngasu.nsk.su/prikl.html).
2. **Воскобойников Ю.Е., Тимошенко Е.И.** Математическая статистика. Новосибирск: Новосибирский государственный архитектурно-строительный университет. 2000 (электронная версия книги размещена по адресу http://www.ngasu.nsk.su/prikl/stat2000.html).

3. Персональная страница (в прямоугольной рамке URL - адрес страницы)

Ресурсы по статистике и эконометрике

Составители: Сергей Моргулис -Якушев и Петр Савельев.

Оглавление:
1. Введение
2. Учебные пособия по статистике и эконометрике
3. Статьи и научные доклады по эконометрике
4. Программное обеспечение (статистические пакеты)
5. Источники данных

http://dist-economics.eu.spb.ru/HTML/predmet/econometrics.htm

4. Персональная страница (в прямоугольной рамке URL - адрес страницы)

Арженовский С.В., Молчанов И.Н. Статистические методы прогнозирования: Учебное пособие. - Ростов-на-Дону: РГЭУ «РИНХ», 2001. - 74 с.

Аннотация или *Полный текст* (1,35 Мбайт) (формат PDF Acrobat Reader)

NEW!!! Арженовский С.В., Федосова О.Н. Эконометрика: Учебное пособие. - Ростов-на-Дону: РГЭУ «РИНХ», 2002. - 102 с.

В двух частях: *Часть 1* (714273 байт) и *Часть 2* (897631 байт)

Статистический пакет прикладных программ Microstat

http://molchanov.narod.ru/

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. **Тимошенко Е.И., Воскобойников Ю.Е.** Теория вероятностей: Учебное пособие. – Новосибирск: НГАСУ, 2003.
2. **Воскобойников Ю.Е., Тимошенко Е.И.** Математическая статистика: Учебное пособие. – Новосибирск: НГАСУ, 2000.
3. **Гмурман В.Е.** Теория вероятностей и математическая статистика. – М.: Высшая школа, 1998.
4. **Калинина В.Н., Панкин В.Ф.** Математическая статистика. – М.: Высшая школа. 1994.
5. **Кремер Н.Ш., Путко Б.А.** Эконометрика. – М.: ЮНИТИ, 2002.
6. **Айвазян С.А., Мхитарян В.С.** Прикладная статистика и основы эконометрики. – М. ЮНИТИ, 1998.
7. **Минус Я.Р., Катышев Л.К., Пересецкий А.А.** Эконометрика. Начальный курс. – М.: Дело, 2000.
8. **Эконометрика** // Под ред. Н.И. Елисеевой. – М.: Финансы и статистика, 2001.
9. **Доугерти К.** Введение в эконометрику. – М.: ИНФРА-М, 1999.
10. **Арженовский С.В., Федосова О.Н.** Эконометрика. Учебное пособие. – Ростов-на-Дону, 2002.
11. **Тихомиров Н.П., Дорохина Е.Ю.** Эконометрика. – М.: Экзамен, 2003.
12. **Макарова Н.В., Трофимец В.Я.** Статистика в EXCEL. Учебное пособие. – М.: Финансы и статистика, 2002.

ПРИЛОЖЕНИЕ

ТОЧЕЧНЫЕ ОЦЕНКИ И ИХ ВЫЧИСЛЕНИЕ В ТАБЛИЧНОМ ПРОЦЕССОРЕ EXCEL

Определение точечной оценки. Пусть над непрерывной случайной величиной X проведены n наблюдений, т. е. получены n значений $x_1, x_2,\ldots,x_n$, 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 $x_1, x_2,\ldots,x_n$, т. е. $\hat{\theta}_n = \varphi(x_1, x_2,\ldots,x_n)$. 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 вероятность того, что оценка $\hat{\theta}_n$ будет отличаться от θ на сколь угодно малую величину ε будет равна нулю, т. е.

$$\lim_{n\to\infty} P(\left|\theta_n - \theta\right| > \varepsilon) = 0$$

Рассмотрим точечные оценки для числовых характеристик случайной величины X.

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \qquad (П1)$$

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, $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$.

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

$$s_X^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (П2)$$

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

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

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

$$\hat{s}_X^2 = \frac{n}{n-1} s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2. \qquad (П3)$$

При больших n отличие между этими оценками пренебрежительно 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 формулам

$$m_{XY} = \overline{xy} - \bar{x} \cdot \bar{y}, \qquad r_{XY} = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{s_X \cdot s_Y}$$

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

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 П1. На основе наблюдений получена выборка объемом n = 12 значений случайной величины X, приведенная на рис. 2.9 в ячейках B2, B3,…, B13. Вычислить точечные оценки для математического ожидания и дисперсии, используя выражения (П1), (П2) и П3).

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 (см. рис. П1). По этим данным построим диаграмму, называемую диаграммой рассеяния (см. рис. П1). Далее, в ячейке B14 запрограммируем формулу (П1), а в ячейках C2 ÷ C13 вычислим квадраты разностей $(x_i - \bar{x})^2$. При этом обратите внимание на использование абсолютного адреса $B$14 для ячейки, где находится значение $\bar{x}$. Затем в ячейке C14 вычислим точечную оценку (П3). Заметим, что математическое ожидание случайной величины равно 0, а дисперсия равна 1/12 = 0.0833. ☻

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

=СРЗНАЧ(*диапазон ячеек*) – реализует формулу (П1);
=ДИСП(*диапазон ячеек*) – реализует формулу (П3);
=ДИСПР(*диапазон ячеек*) – реализует формулу (П2).

	A	B	C	D	E	F	G	H	I
1				=(C2-$B$14)^2					
2	1	-0,25	0,0923						
3	2	-0,50	0,2996						
4	3	0,31	0,0656						
5	4	0,24	0,0345						
6	5	0,40	0,1245						
7	6	0,04	0,0001						
8	7	0,34	0,0829						
9	8	-0,06	0,0119						
10	9	-0,01	0,0033						
11	10	-0,06	0,0124						
12	11	0,31	0,0689						
13	12	-0,16	0,0428				0,050	=СРЗНАЧ(B2:B13)	
14		0,050	0,0763				0,0763	=ДИСП(B2:B13)	
15			=СУММ(C2:C13)/(12-1)				0,0699	=ДИСПР(B2:B13)	
16		=СУММ(B2:B13)/12							

Рис. П1. Вычисление точечных оценок в Excel

Пример П2.. По выборочным данным примера П1 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. П1). ☻

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

=КОВАР(*диапазон ячеек X; диапазон ячеек У*).

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

=КОРРЕЛ(*диапазон ячеек X; диапазон ячеек У*) ;
=ПИРСОН(*диапазон ячеек X; диапазон ячеек У*),

Эти функции дают один и тот же результат.