ОСНОВЫ РЕГРЕССИОННОГО АНАЛИЗА

ПОНЯТИЕ КОРРЕЛЯЦИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА

Для решения задач экономического анализа и прогнозирования очень часто используются статистические, отчетные или наблюдаемые данные. При этом полагают, эти данные являются значениями случайной величины.

Случайной величиной называется переменная величина, которая в зависимости от случая принимает различные значения с некоторой вероятностью. Закон распределения случайной величины показывает частоту ее тех или иных значений в общей их совокупности.

При исследовании взаимосвязей между экономическими показателями на основе статистических данных часто между ними наблюдается стохастическая зависимость. Она проявляется в том, что изменение закона распределения одной случайной величины происходит под влиянием изменения другой.

Взаимосвязь между величинами может быть полной (функциональной) и неполной (искаженной другими факторами).

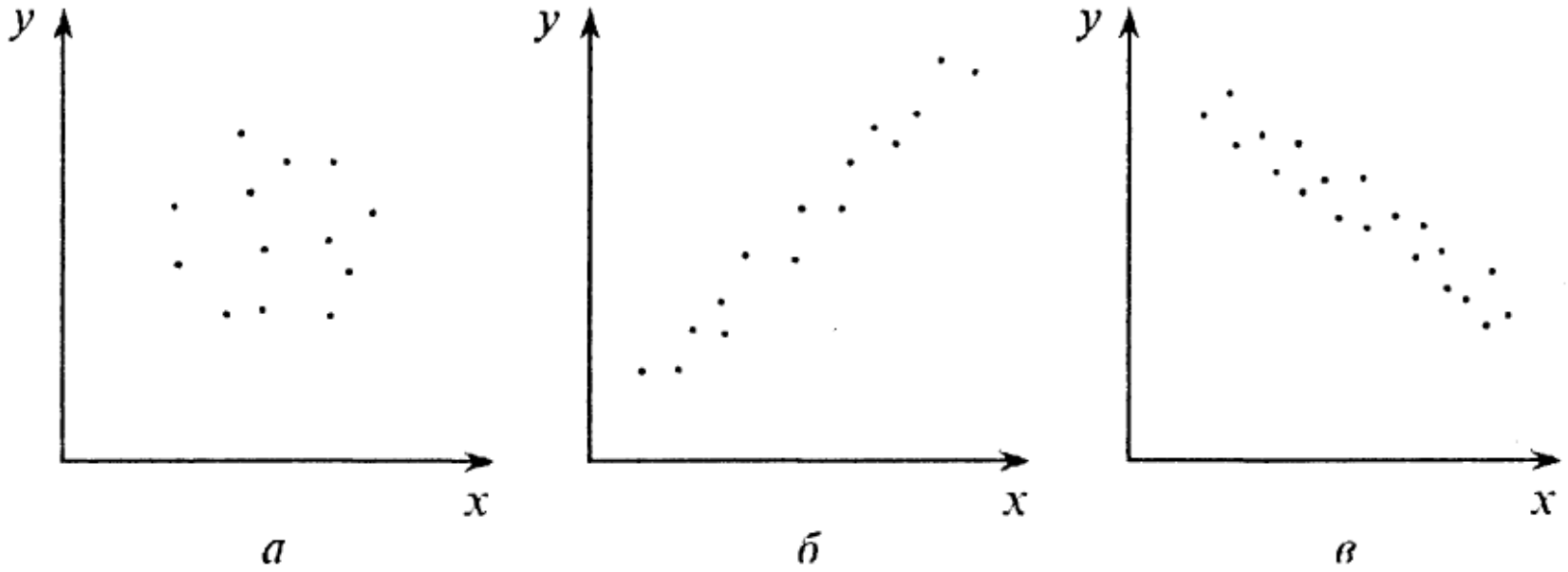
Пример функциональной зависимости — выпуск продукции и ее потребление в условиях дефицита.

Неполная зависимость наблюдается, например, между стажем рабочих и их производительностью труда.

Обычно рабочие с большим стажем трудятся лучше молодых, но под влиянием дополнительных факторов — образование, здоровье и т.д. эта зависимость может быть искажена.
Раздел математической статистики, посвященный изучению взаимосвязей между случайными величинами, называется корреляционным анализом {от лат. соггеїаііо — соотношение, соответствие).

Основная задача корреляционного анализа — это установление характера и тесноты связи между результативными (зависимыми) и факторными (независимыми) показателями (признаками) в данном явлении или процессе. Корреляционную связь можно обнаружить только при массовом сопоставлении фактов.

Характер связи между показателями определяется по корреляционному полю. Если y — зависимый признак, а x— независимый, то, отметив каждый случай x(i) с координатами x_i и y_i получим корреляционное поле. По расположению точек можно судить о характере связи



Примеры корреляционных полей:
А — переменные л: и у не коррелируют; б — наблюдается сильная положительная корреляция; в — наблюдается слабая отрицательная корреляция

Теснота связи определяется с помощью коэффициента корреляции, который рассчитывается специальным образом и лежит в интервалах от минус единицы до плюс единицы.

Если значение коэффициента корреляции лежит в интервале от 1 до 0,9 по модулю, то отмечается очень сильная корреляционная зависимость.

В случае, если значение коэффициента корреляции лежит в интервале от 0,9 до 0,6, то говорят, что имеет место слабая корреляционная зависимость.

Наконец, если значение коэффициента корреляции находится в интервале от -0,6 до 0,6, то говорят об очень слабой корреляционной зависимости или полном ее отсутствии.

Таким образом, корреляционный анализ применяется для нахождения характера и тесноты связи между случайными величинами

Регрессионный анализ своей целью имеет вывод, определение (идентификацию) уравнения регрессии, включая статистическую оценку его параметров. Уравнение регрессии позволяет найти значение зависимой переменной, если величина независимой или независимых переменных известна.

Практически, речь идет о том, чтобы, анализируя множество точек на графике (т.е. множество статистических данных), найти линию, по возможности точно отражающую заключенную в этом множестве закономерность (тренд, тенденцию), — линию регрессии.

По числу факторов различают одно-, двух- и многофакторные уравнения регрессии.

По характеру связи однофакторные уравнения регрессии подразделяются:

а) на линейные:

$$y = a + bx \ ,$$

где х — экзогенная (независимая) переменная, у — эндогенная (зависимая, результативная) переменная, а, Ь — параметры;

Б) СТЕПЕННЫЕ:

$$y = a \cdot x^b \ ,$$

в) показательные:

$$y = a \cdot b^x \ ,$$

г) прочие.

Определение параметров линейного однофакторного уравнения регрессии

Пусть у нас имеются данные о доходах (x) и спросе на некоторый товар y>) за РЯД лет (n):

Год i	Доход x	Спрос y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
...	...	...
n	x_n	y_n

Предположим, что между x и y существует линейная взаимосвязь, т.е.

$$y = a + bx ,$$

Для того, чтобы найти уравнение регрессии, прежде всего нужно исследовать тесноту связи между случайными величинами x; и y, т.е. корреляционную зависимость.

Пусть

$$x_1,\ x_2,\ \ldots,\ x_n$$

— совокупность значений независимого, факторного признака;

$$y_1,\ y_2,\ \ldots,\ y_n$$

-совокупность соответствующих значений зависимого, результативного признака;

n — количество наблюдений.

Для нахождения уравнения регрессии вычисляются следующие величины:

1. СРЕДНИЕ ЗНАЧЕНИЯ

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$$

— ДЛЯ ЭКЗОГЕННОЙ ПЕРЕМЕННОЙ

y = --для эндогенной переменной.

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

2. Отклонения от средних величин

$$\Delta x_i = x_i - \bar{x}\,, \qquad \Delta y_i = y_i - \bar{y}\,.$$

3. ВЕЛИЧИНЫ ДИСПЕРСИИ И СРЕДНЕГО КВАДРАТИЧНОГО ОТКЛОНЕНИЯ

$$D_x = \frac{\sum_{i=1}^{n} \Delta x_i^2}{n-1}, \qquad D_y = \frac{\sum_{i=1}^{n} \Delta y_i^2}{n-1};$$

$$\sigma_x = \sqrt{D_x}, \qquad \sigma_y = \sqrt{D_y}.$$

Величины дисперсии и среднего квадратичного отклонения характеризуют разброс наблюдаемых значений вокруг среднего значения. Чем больше дисперсия, тем больше разброс.

4. Вычисление корреляционного момента (коэффициента ковариации):

$$K_{x,y} = \frac{\Delta x_1 \cdot \Delta y_1 + \Delta x_2 \cdot \Delta y_2 + \ldots + \Delta x_n \cdot \Delta y_n}{n-1} = \frac{\sum_{i=1}^{n} \Delta x_i \cdot \Delta y_i}{n-1}.$$

Корреляционный момент отражает характер взаимосвязи между x и y.

Если Кху > 0, то взаимосвязь прямая. Если Кху < 0, то взаимосвязь обратная.

5. Коэффициент корреляции вычисляется по формуле

$$R_{x,y} = \frac{K_{x,y}}{\sigma_x \sigma_y}.$$

Доказано, что коэффициент корреляции находится в интервале от минус единицы до плюс единицы (-1 < Кху < 1). Коэффициент корреляции в квадрате (Кху) называется коэффициентом детерминации.

Если Кху > |0,8|, то вычисления продолжаются.
6. Вычисления параметров регрессионного уравнения.
Коэффициент b находится по формуле

$$b = \frac{K_{x,y}}{D_x}.$$

После чего можно легко найти параметр а:

$$a = \bar{y} - b\bar{x}.$$

Пример. Пусть у нас имеются статистические данные о доходах (x) и спросе (y). Необходимо найти корреляционную зависимость между ними и определить параметры уравнения регрессии.

Год *i*	Доход *x*	Спрос *y*
1	10	6
2	12	8
3	14	8
4	16	10,3
5	18	10,5
6	20	13

Коэффициенты а и Ь находятся методом наименьших квадратов, основная идея которого состоит в том, что за меру суммарной погрешности принимается сумма квадратов разностей (остатков) между фактическими значениями результативного признака y_i и его расчетными значениями полученными при помощи уравнения регрессии

$$y_{i\text{p}} = a + bx_i \,.$$

При этом величины остатков находятся по формуле $u_i = y_i - y_{i\text{p}}\,,$

где y_i — фактическое значение y; y_{ip} — расчетное значение y.

Предположим, что между нашими величинами существует линейная зависимость. Тогда расчеты лучше всего выполнить в Excel, используя статистические функции:
СРЗНАЧ — для вычисления средних значений; ДИСП — для нахождения дисперсии;
СТАНДОТКЛОН — для определения среднего квадратичного отклонения;
КОРЕЛЛ — для вычисления коэффициента корреляции.

Корреляционный момент можно вычислить, найдя отклонения от средних значений

ему

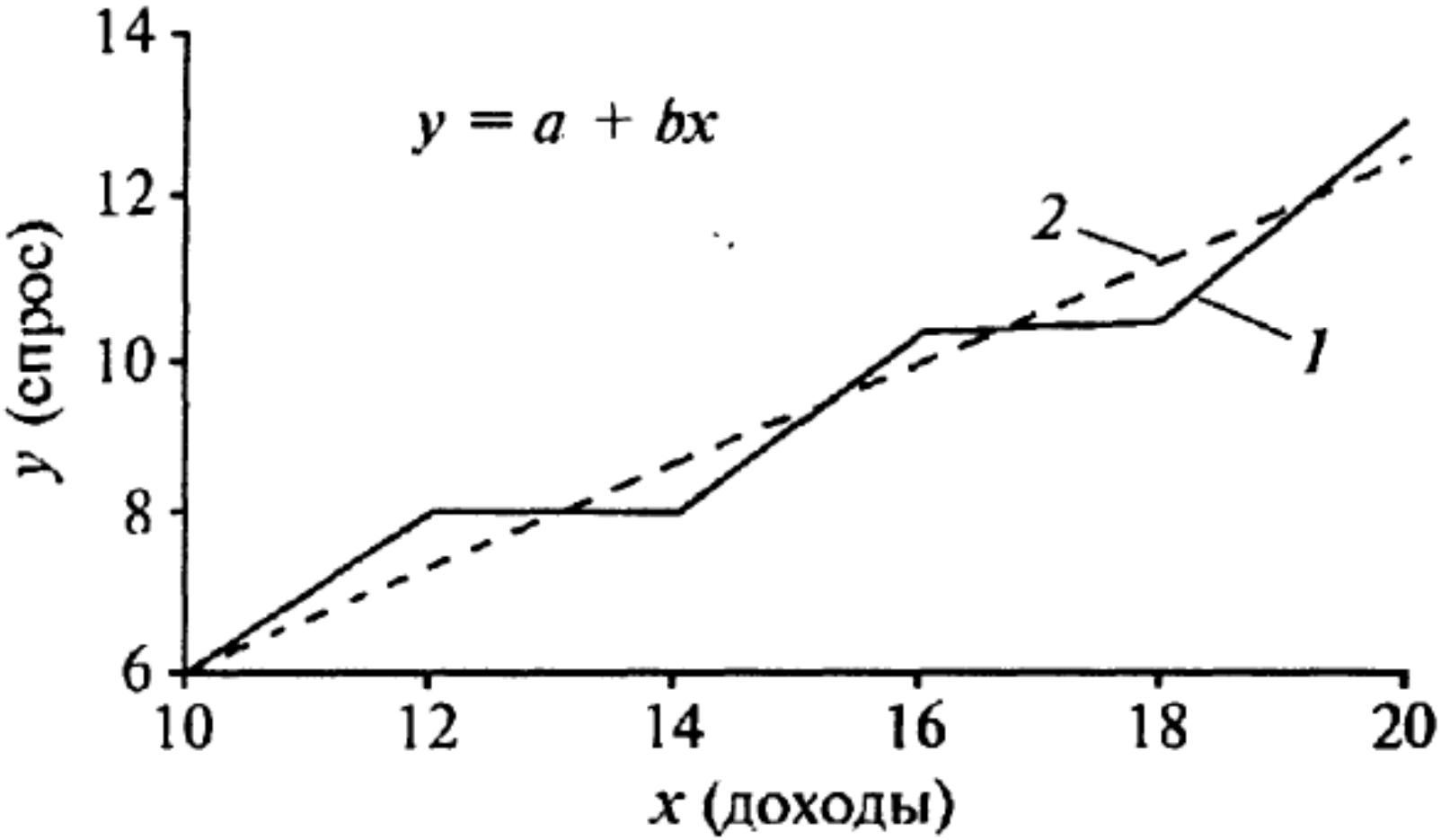
Параметры линейного однофакторного уравнения регрессии

Показатели	x	y
Среднее значение	15	9,3
Дисперсия	14	6,08
Среднее квадратичное отклонение	3,7417	2,4658
Корреляционный момент	8,96	
Коэффициент корреляции	0,9712	
Параметры	$b = 0{,}64$	$a = -0{,}3$

В итоге наше уравнение будет иметь вид

У = -0,3 + 0,64Х .

Используя это уравнение, можно найти расчетные значения у и построить график



Ломаная линия на графике отражает фактические значения у, а прямая линия построена с помощью уравнения регрессии и отражает тенденцию изменения спроса в зависимости от дохода.

Однако встает вопрос, насколько значимы параметры а и Ь

Какова величина погрешности?

Оценка величины погрешности линейного однофакторного уравнения

1. ОБОЗНАЧИМ РАЗНОСТЬ МЕЖДУ ФАКТИЧЕСКИМ ЗНАЧЕНИЕМ РЕЗУЛЬТАТИВНОГО ПРИЗНАКА И ЕГО РАСЧЕТНЫМ ЗНАЧЕНИЕМ КАК u_i

$$u_i = y_i - y_{i\,\text{р}},$$

где y_i — фактическое значение y; y_{ip} — расчетное значение y;

u_i— РАЗНОСТЬ МЕЖДУ НИМИ.

2. В качестве меры суммарной погрешности выбрана величина

$$S = \frac{\sum_{n-2}^{n} u_i^2}{n-2}.$$

Для нашего примера S= 0,432.

Поскольку $\overline{u}$
(среднее значение остатков) равно нулю, то суммарная погрешность равна остаточной дисперсии.

3. ОСТАТОЧНАЯ ДИСПЕРСИЯ НАХОДИТСЯ ПО ФОРМУЛЕ

$$D_u = \frac{\sum (u_i - \bar{u})^2}{n-2} = \frac{\sum u_i^2}{n-2} = S .$$

Для нашего примера $D_u = 0{,}432.$

Можно показать, что $D_u = \left(1 - R^2_{x,y}\right) \cdot D_y .$

Если

$R^2_{x,y} = 1$, то $D_u = 0$;

$R^2_{x,y} = 0$, то $D_u = D_y .$

Таким образом, $0 \le D_u \le D_y$.

Легко заметить, что если

$$R_{x,\,y} = 0{,}9, \text{ то } D_u = (1 - 0{,}81) \cdot D_y = 0{,}91 \cdot D_y .$$

Это соотношение показывает, что в экономических приложениях допустимая суммарная погрешность может составить не более 20% от дисперсии результативного признака D_y.

4. Стандартная ошибка уравнения находится по формуле

$$\sigma_u = \sqrt{D_u},$$

где $D_{и}$ — остаточная дисперсия. В нашем случае $\sigma_{и} = 0{,}6572$.
5. Относительная погрешность уравнения регрессии вычисляется как

$$\vartheta = \frac{\sigma_u}{\bar{y}} \cdot 100\%,$$

где $\sigma_{и}$ — стандартная ошибка; $\bar{y}$— среднее значение результативного признака.
В нашем случае ϑ = 7,07%.

Если величина ϑ мала и отсутствует автокорреляция остатков, то прогнозные качества оцененного регрессионного уравнения высоки.

6. Стандартная ошибка коэффициента Ь вычисляется по формуле

$$S_b = \frac{\sigma_u}{\sqrt{nD_x}}.$$

В нашем случае она равна S_b = 0,07171.

Для вычисления стандартной ошибки коэффициента а используется формула

$$S_a = \sigma_u \sqrt{\frac{D_x + \bar{x}^2}{n \cdot D_x}}.$$

В нашем примере S_A = 1,108.

Стандартные ошибки коэффициентов используются для оценивания параметров уравнения регрессии.
Коэффициенты считаются значимыми, если

$$\frac{S_a}{|a|} < 0{,}5; \qquad \frac{S_b}{|b|} < 0{,}5.$$

В нашем примере

$$\frac{S_a}{|a|} = \frac{1{,}108}{|0{,}3|} = 3{,}69, \qquad \frac{S_b}{|b|} = \frac{0{,}07171}{0{,}64} = 0{,}112.$$

Коэффициент a не значим, так как указанное отношение больше 0,5, а относительная погрешность уравнения регрессии слишком высока — 26,7%.

Стандартные ошибки коэффициентов используются также для оценки статистической значимости коэффициентов при помощи t-критерия Стьюдента. Значения t-критерия Стьюдента содержатся в справочниках по математической статистике.

В таблице приводятся его некоторые значения.

Степени свободы	Уровень доверия (c)	
(n–2)	0,90	0,95
1	6,31	12,71
2	2,92	4,30
3	2,35	3,18
4	2,13	2,78
5	2,02	2,57

Далее находятся максимальные и минимальные значения параметров (b^-, b^+) по формулам:

$$b^- = b - t_{\text{СТ}} \cdot S_b,$$

$$b^+ = b + t_{\text{СТ}} \cdot S_b.$$

Для нашего примера находим

$$b^- = 0{,}64 - 2{,}78 \cdot 0{,}07171 = 0{,}44,$$

$$b^+ = 0{,}64 + 2{,}78 \cdot 0{,}07171 = 0{,}839.$$

Если интервал (b^-, b^+) ДОСТАТОЧНО МАЛ И НЕ СОДЕРЖИТ НОЛЬ, ТО КОЭФФИЦИЕНТ Ь ЯВЛЯЕТСЯ СТАТИСТИЧЕСКИ ЗНАЧИМЫМ НА С-ПРОЦЕНТНОМ ДОВЕРИТЕЛЬНОМ УРОВНЕ.

Аналогично находятся максимальные и минимальные значения параметр а.

Для нашего примера $$a^{-} = -0{,}3 - 2{,}78 \cdot 1{,}108 = -3{,}38,$$

$$a^{+} = -0{,}3 + 2{,}78 \cdot 1{,}108 = 2{,}78.$$

Коэффициент а не является статистически значимым, так как интервал (АА+) велик и содержит ноль.

Вывод: полученные результаты не являются значимыми и не могут быть использованы для прогнозных расчетов.

Ситуацию можно поправить следующими способами:
а) увеличить число п;
Б) УВЕЛИЧИТЬ КОЛИЧЕСТВО ФАКТОРОВ;
в) изменить форму уравнения.

Проблема автокорреляции остатков.

Критерий Дарбина—Уотсона

Часто для нахождения уравнений регрессии используются динамические ряды, т.е. последовательность экономических показателей за ряд лет (кварталов, месяцев), следующих друг за другом.
В этом случае имеется некоторая зависимость последующего значения показателя от его предыдущего значения, которое называется автокорреляцией. В некоторых случаях зависимость такого рода является весьма сильной и влияет на точность коэффициента регрессии.
Пусть уравнение регрессии построено и имеет вид:

$$y_t = a + bx_t + u_t, \quad t = 1, 2, \ldots, n,$$

где u_i — погрешность уравнения регрессии в год i.
Явление автокорреляции остатков состоит в том, что в любой год i остаток u не является случайной величиной, а зависит от величины остатка предыдущего года
В результате при использовании уравнения регрессии могут быть большие ошибки.

Для определения наличия или отсутствия автокорреляции применяется критерий Дарбина—Уотсона:

$$DW = \frac{\sum_{t=2}^{n} (u_t - u_{t-1})^2}{\sum_{t=1}^{n} u_t^2}.$$

Возможные значения критерия DW находятся в интервале от О до 4. Если автокорреляция остатков отсутствует, то DW $\cong$ 2.

Построение уравнения степенной регрессии

Уравнение степенной агрессии имеет вид: $$y = a \cdot x^b,$$

где а, Ь — параметры, которые определяются по данным таблицы наблюдений. Таблица наблюдений составлена и имеет вид

x	x_1	x_2	...	x_n
y	y_1	y_2	...	y_n

Прологарифмируем исходное уравнение и в результате получим

$$\ln y = \ln a + b \cdot \ln x.$$

Обозначим ln(y) через y', lп (a) как a', ln(x): как x'

В результате подстановки получим $y' = a' + bx'$.

Данное уравнение есть не что иное, как уравнение линейной регрессии, параметры которого мы умеем находить.

Для этого прологарифмируем исходные данные:

$\ln x$	$\ln x_1$	$\ln x_2$	...	$\ln x_n$
$\ln y$	$\ln y_1$	$\ln y_2$	...	$\ln y_n$

Далее необходимо выполнить известные нам вычислительные процедуры по нахождению коэффициентов а и Ь, используя прологарифмированные исходные данные. В результате получим значения коэффициентов Ь и а. Параметр а можно найти по формуле $a = e^{a'}$.

В этих же целях можно воспользоваться функцией EXP в Excel.

Двухфакториые и многофакторные уравнения регрессии

Линейное двухфакторное уравнение регрессии имеет вид

$$y = a + b_1 x_1 + b_2 x_2,$$

где а, Ь1, b2 — параметры; х1, х**2** — экзогенные переменные; у — эндогенная переменная.

Идентификацию этого уравнения лучше всего производить с использованием функции Excel ЛИНЕЙН.

Степенное двухфакторное уравнение регрессии имеет вид

$$y = a x_1^{\alpha} \cdot x_2^{\beta},$$

где a, α, β — параметры; x1, x**2** — экзогенные переменные; y — эндогенная переменная.

Для нахождения параметров этого уравнения его необходимо прологарифмировать. В результате получим

$$\ln y = \ln a + \alpha \ln x_1 + \beta \ln x_2.$$

Идентификацию этого уравнения также лучше всего производить с использованием функции Excel ЛИНЕЙН.

Следует помнить, что мы получим не параметр a, а его логарифм, который следует преобразовать в натуральное число.

Линейное многофакторное уравнения регрессии имеет вид

$$y = a + b_1 x_1 + \ldots + b_n x_n ,$$

где a, Ь2, b_n— параметры; x1, x_n— экзогенные переменные; y — эндогенная переменная.

Идентификацию этого уравнения также лучше всего производить с использованием функции Excel ЛИНЕЙН