Федеральное государственное образовательное бюджетное учреждение высшего образования
«Финансовый университет при Правительстве Российской Федерации»
(Финуниверситет)
Смоленский филиал Финуниверситета

Кафедра Математики и информатики

ЭКОНОМЕТРИКА

Методические указания
по выполнению контрольной работы
для самостоятельной работы студентов третьего курса
направления 38.03.01 «Экономика»

(программа подготовки бакалавров)
Очная форма обучения

СМОЛЕНСК 2016

Федеральное государственное образовательное бюджетное учреждение
высшего образования
**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ
ПРИ ПРАВИТЕЛЬСТВЕ РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)
СМОЛЕНСКИЙ ФИЛИАЛ**

Кафедра «Математика и информатика»

УТВЕРЖДАЮ
Директор филиала

В.Д.Голичев
« 23 » сентября 2016 г.

Гусарова О.М..Прохоренков П.А.

Эконометрика

Методические указания
по выполнению практических работ и контрольных заданий
для самостоятельной работы студентов второго курса
направления 38.03.01 «Экономика»

(программа подготовки бакалавров)
Очная форма обучения

Рекомендовано Ученым советом Смоленского филиала
протокол № _36_ от __ «21» сентября___ 2016 г.

Одобрено кафедрой «Математика и информатика»
протокол № 15 от «13» сентября 2016 г.

СМОЛЕНСК 2016

УДК 330.43(073)

ББК 65в631

Г 97

Рецензент: С.В.Земляк, заведующий кафедрой «Экономика и менеджмент».

Эконометрика. Методические указания по выполнению контрольной работы для самостоятельной работы студентов третьего курса направления 38.03.01 «Экономика» (программа подготовки бакалавров), очная форма обучения. – Смоленск: Смоленский филиал Финуниверситета, кафедра «Математика и информатика», 2016. – 61 с.

Учебное электронное издание

Гусарова Ольга Михайловна
Прохоренков Павел Александрович

Методы оптимальных решений

СОДЕРЖАНИЕ

1. РАСЧЕТ СТАТИСТИЧЕСКИХ ХАРАКТЕРИСТИК СЛУЧАЙНЫХ ВЕЛИЧИН

Задача 1.

Бросают 2 игральные кости. Составить распределение вероятностей случайной величины X – разность выпавших очков (Всегда положительная, либо равна нулю). Найти математическое ожидание и дисперсию двумя способами. Построить закон распределения p=f(X).

Красная кость	**Зеленая кость**					
	1	**2**	**3**	**4**	**5**	**6**
1	0	1	2	3	4	5
2	1	0	1	2	3	4
3	2	1	0	1	2	3
4	3	2	1	0	1	2
5	4	3	2	1	0	1
6	5	4	3	2	1	0

Всего 36 исходов

X	**0**	**1**	**2**	**3**	**4**	**5**
p	6/36	10/36	8/36	6/36	4/36	2/36

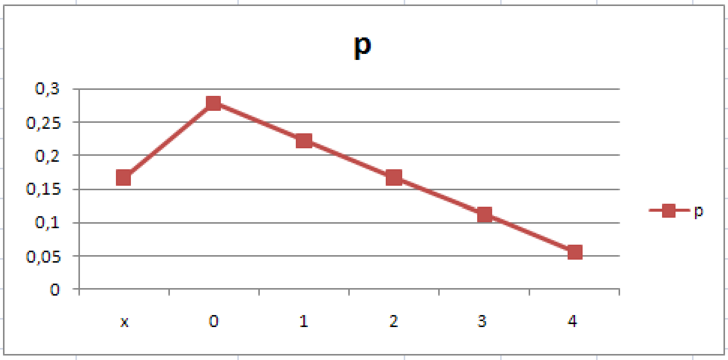
Формулы расчета математического ожидания и дисперсии случайной величины

$$M(X) = \sum_{i=1}^{n} x_i p_i$$

$$\sigma_X^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \, p_i \qquad \sigma_X^2 = M(X^2) - \mu^2$$

Результаты расчетов и график распределения вероятностей имеет следующий вид:

A	B	C	D	E	F	G
x	p	x*p	(x-m(x))кв	D*p	хкв	M(хкв)
0	0,166667	0	3,780864198	0,630144033	0	0
1	0,277778	0,277778	0,891975309	0,247770919	1	0,277778
2	0,222222	0,444444	0,00308642	0,000685871	4	0,888889
3	0,166667	0,5	1,114197531	0,185699588	9	1,5
4	0,111111	0,444444	4,225308642	0,469478738	16	1,777778
5	0,055556	0,277778	9,336419753	0,518689986	25	1,388889
	1	1,944444		2,052469136		5,833333
m(x)	1,944444					
D(x)	2,052469					
м(хкв)	5,833333					
дисп(2сп)	2,052469					



Задача 2.

В результате статистических исследований получены значения X и Y:

№	X	Y
1	1	3
2	2	7
3	3	15
4	4	10
5	5	17
6	3	10
7	4	15

X – уровень производства, Y – затрату на производство.

Определить степень связи между признаками (ковариацию и корреляцию).

Решение.

$\operatorname{cov}(x,y) = \overline{y \cdot x} - \overline{y} \cdot \overline{x}$ – ковариация признаков x и y

$\sigma_x^2 = \overline{x^2} - \overline{x}^2$ – дисперсия признака x

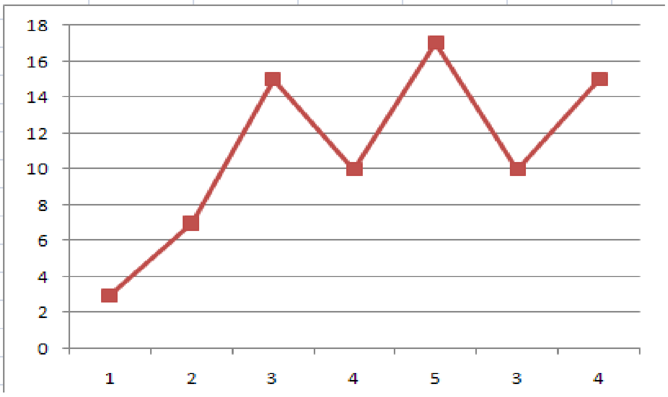
$$\overline{x} = \frac{1}{n}\sum x,\quad \overline{y} = \frac{1}{n}\sum y,\quad \overline{y \cdot x} = \frac{1}{n}\sum y \cdot x,\quad \overline{x^2} = \frac{1}{n}\sum x^2$$

Линейный коэффициент корреляции r_{xy} можно рассчитать по формуле:

$$r_{xy} = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y}.$$

Рпасчеты выполненные в Excel приведены в таблице и на рисунке.

A	B	C	D	E	F
	x	y	y*x	xкв	yкв
1	1	3	3	1	9
2	2	7	14	4	49
3	3	15	45	9	225
4	4	10	40	16	100
5	5	17	85	25	289
6	3	10	30	9	100
7	4	15	60	16	225
Среднее	3,142857	11	39,57143	11,42857	142,4286
Сигма кв				1,55102	21,42857
Сигма				1,2454	4,6291
cov(x,y)	5				
К-нтКорр	0,867291				



2 МОДЕЛЬ ПАРНОЙ РЕГРЕССИИ

2.1 Теоретическое введение

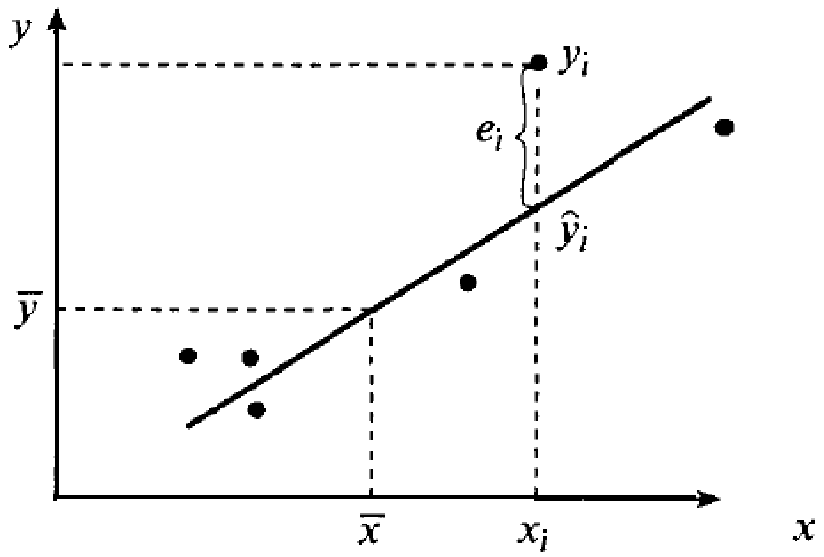
Задача моделирования: на основе выборочного наблюдения оценивается выборочное уравнение регрессии (или линия регрессии):

$$\hat{y} = a + bx, \qquad (2.1)$$

где коэффициенты (a и b) - оценки параметров (α и β).

Метод наименьших квадратов

$$e_i = y_i - \hat{y}_i.$$



Неизвестные значения (a и b) определяются **методом наименьших квадратов** (МНК).

Решение системы:

$$\begin{cases} b = \dfrac{cov(x,y)}{var(x,y)} = \dfrac{\overline{xy} - \overline{x}\cdot\overline{y}}{\overline{x^2} - (\overline{x})^2}, \\ a = \overline{y} - b\overline{x}. \end{cases} \qquad (2.3)$$

Линия регрессии (расчетное значение зависимой переменной):

$$\hat{y} = a + bx, \text{ или } \hat{y} - \overline{y} = b(x - \overline{x}).$$

Коэффициент $\boldsymbol{b}$ есть **угловой коэффициент регрессии**, он показывает, на сколько единиц в среднем изменяется переменная $\boldsymbol{y}$ при увеличении независимой переменной $\boldsymbol{x}$ на единицу.

Постоянная $\boldsymbol{a}$ дает прогнозируемое значение зависимой переменной при $\boldsymbol{x=0}$.

Можно показать, что

$$\boldsymbol{b = \frac{cov(x,y)}{var(x,y)} = r\sqrt{\frac{var(y)}{var(x)}} = r\frac{S_y}{S_x}},$$

где $\boldsymbol{r}$ – коэффициент корреляции между $\boldsymbol{x}$ и $\boldsymbol{y}$,

$\boldsymbol{S_x}$, $\boldsymbol{S_y}$ - их стандартные отклонения. Таким образом, если коэффициент корреляции уже рассчитан, можно найти коэффициенты ($\boldsymbol{a}$, $\boldsymbol{b}$) парной регрессии.

Запишем выборочные дисперсии величин $\boldsymbol{y}, \boldsymbol{\hat{y}}, \boldsymbol{e}$:

$\boldsymbol{var(y) = \frac{1}{n}\sum(y_i - \bar{y})^2}$ – дисперсия наблюдаемых значений $\boldsymbol{y}$;

$\boldsymbol{var(\hat{y}) = \frac{1}{n}\sum(\hat{y}_i - \bar{y})^2}$ – дисперсия расчетных значений $\boldsymbol{y}$;

$\boldsymbol{var(e) = \frac{1}{n}\sum(e_i - \bar{e})^2 = \frac{1}{n}\sum e_i^2}$ – дисперсия остатков.

Разброс значений зависимой переменной характеризуется выборочной дисперсией $\boldsymbol{var(y)}$, которую можно представить как:

$$\boldsymbol{var(y) = var(\hat{y}) + var(e)}. \qquad (2.2)$$

Таким образом, дисперсия $\boldsymbol{var(y)}$ разложена на 2 части:

- $\boldsymbol{var(\hat{y})}$ – часть объясненная регрессионным уравнением,
- $\boldsymbol{var(e)}$ – необъясненная часть.

Коэффициентом детерминации $\boldsymbol{R^2}$ называется отношение:

$$\boldsymbol{R^2 = \frac{var(\hat{y})}{var(y)} = 1 - \frac{var(e)}{var(y)}, \qquad 0 \le R^2 \le 1}.$$

$\boldsymbol{R^2}$ характеризует долю вариации зависимой переменной, объясненную с помощью уравнения регрессии.

Отношение $\boldsymbol{\frac{var(e)}{var(y)}}$ представляет собой долю необъясненной регрессии.

Если $\boldsymbol{R^2 = 1}$, то регрессия точно описывает выборку:

$$\boldsymbol{var(y) = var(\hat{y}),\ var(e) = 0,\ \ y_i = \hat{y}_i,\ \ i = 1, \dots, n},$$

т.е. все точки наблюдения лежат на регрессионной прямой.

Если $\boldsymbol{R^2 = 0}$, то регрессия ничего не дает для описания выборки.

Чем ближе к единице $\boldsymbol{R^2}$, тем $\boldsymbol{\hat{y}}$ более точно аппроксимирует $\boldsymbol{y}$.

Замечание. Вычисление $\boldsymbol{R^2}$ корректно, если константа $\boldsymbol{a}$ включена в уравнение регрессии.

F-тест на качество оценивания

Для определения статистической значимости коэффициента детерминации $\boldsymbol{R^2}$ проверяется гипотеза: $\boldsymbol{H_0: F = 0}$ для $\boldsymbol{F -}$ статистики:

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

Величина $\boldsymbol{F}$ имеет распределение Фишера с $\boldsymbol{v_1 = 1, v_2 = n - 2}$.

Проверку значимости $\boldsymbol{R^2}$ можно выполнить двумя способами.

1. Критическое значение $\boldsymbol{F}_{\text{крит}}$ при заданных $\boldsymbol{\alpha, v_1, v_2}$ определяется по таблице $\boldsymbol{F}$- распределения Фишера или в Excel с помощью функции

$$\boldsymbol{F}_{\text{крит}} = \boldsymbol{F}\text{РАСПОБР}(\boldsymbol{\alpha}, \boldsymbol{v_1}, \boldsymbol{v_2}).$$

Из сравнения наблюдаемого значения $\boldsymbol{F}$ с критическим, получаем:

- Если $\boldsymbol{F} < F_{\text{крит}}$, то $\boldsymbol{H_0}$ принимается, т.е. $\boldsymbol{R^2}$ незначим.
- Если $\boldsymbol{F} > F_{\text{крит}}$, то $\boldsymbol{H_0}$ отвергается, т.е. $\boldsymbol{R^2}$ значим.

2. Наблюдаемому (расчетному) значению критерия $\boldsymbol{F}$ соответствует определенная *значимость* $\boldsymbol{F}$**,** которую можно вычислить в Excel с помощью функции

   Значимость $\boldsymbol{F} = \boldsymbol{FРАСП(F}, \boldsymbol{v_1}, \boldsymbol{v_2})$.

Из сравнения *значимости* $\boldsymbol{F}$ с заданным стандартным уровнем значимости, получаем:

- Если *значимость* $\boldsymbol{F}$ больше стандартного уровня, то $\boldsymbol{R^2}$ незначим.
- Если *значимость* $\boldsymbol{F}$ меньше стандартного уровня, то $\boldsymbol{R^2}$ значим.

Чаще всего $\boldsymbol{F}$- тест используется для оценки того, значимо ли объяснение, даваемое уравнением, в целом.

Средняя ошибка аппроксимации

Оценку качества построенной модели дает коэффициент детерминации, а также средняя ошибка аппроксимации.

Средняя ошибка аппроксимации – среднее отклонение расчетных значений зависимой переменной от фактических значений:

$$\boldsymbol{A} = \frac{\boldsymbol{1}}{\boldsymbol{n}} \sum \left| \frac{\boldsymbol{y} - \widehat{\boldsymbol{y}}}{\boldsymbol{y}} \right| \cdot \boldsymbol{100}\%.$$

Допустимый предел значений $\boldsymbol{A}$ не более 8 – 10%.

2.2 Пример решения задач

Пример 1.

Данные личного дохода $\boldsymbol{x}$ и расходов на питание $\boldsymbol{y}$ за соответствующие годы приведены в таблице.

Год	**2000**	**2001**	**2002**	**2003**	**2004**
$\boldsymbol{x}$	2	6	10	14	18
$\boldsymbol{y}$	1	2	4	11	12

А) построить регрессионную зависимость расходов на питание $\boldsymbol{y}$ и личного дохода $\boldsymbol{x}$,

Б) построить регрессионную зависимость расходов на питание $\boldsymbol{y}$ и времени $\boldsymbol{t}$,

В) оценить качество моделей.

Пошаговое решение задачи.

Пусть истинная модель описывается выражением:

$$\boldsymbol{Y = \alpha + \beta X + \varepsilon.}$$

1. По выборочным наблюдениям определяем оценки ($\boldsymbol{a}$, $\boldsymbol{b}$), расчеты записываем в таблицу:

2. Находим

$$\boldsymbol{cov(x,y) = \overline{xy} - \overline{x} \cdot \overline{y} = 84,8 - 60 = 24,8}$$

$$\boldsymbol{var(x) = \overline{x^2} - (\overline{x})^2 = 132 - 100 = 32,}$$

$$\boldsymbol{b = \frac{cov(x,y)}{var(x)} = \frac{24,8}{32} = 0,775}$$

$$\boldsymbol{a = \overline{y} - b\overline{x} = 6 - 0,775 \cdot 10 = -1,75}$$

Год	x	y	x^2	$x\,y$	$\hat{y}$	$(y-\bar{y})^2$	$(\hat{y}-\bar{y})^2$	$(y-\hat{y})^2$
2000	2	1	4	2	-0,2	25	38,44	1,44
2001	6	2	36	12	2,9	16	9,61	0,81
2002	10	4	100	40	6	4	0	4
2003	14	11	196	154	9,1	25	9,61	3,61
2004	18	12	324	216	12,2	36	38,44	0,04
Сумма	50	30	660	424	30	106	96,1	9,9
Среднее	10	6	132	84,8	6	21,2	19,22	1,98
	$\bar{x}$	$\bar{y}$	$\overline{x^2}$	$\overline{x\,y}$	$\bar{\hat{y}}$	***var(y)***	$var(\hat{y})$	***var(e)***

3. Записываем уравнение регрессии

$$\hat{y} = a + bx = -1,75 + 0,775 \cdot x,$$

4.Делаем выводы по коэффициенту b.

Коэффициент b показывает, что при увеличении дохода на 1 усл.ед. расходы на питание увеличиваются в среднем на 0,775 усл.ед.

5.Проверяем условие (2.2):

$$var(y) = var(\hat{y}) + var(e).$$

6. Находим коэффициент детерминации:

$$R^2 = \frac{var(\hat{y})}{var(y)} = \frac{19,22}{21,2} = 0,907,$$

7. Делаем выводы по величине коэффициента детерминации.

90,7% вариации зависимой переменной (расходы на питание) объясняется регрессией.

8. Значимость коэффициента R^2 проверяем по F- тесту:

$$F = \frac{R^2(n-2)}{1-R^2} = 29,2.$$

Выполняем проверку значимости R^2 двумя способами.

1. При $\alpha = 0{,}05$ $v_1 = 1$ и $v_2 = 3$ по таблице или с помощью функции **FРАСПОБР**(α, v_1 , v_2) находим $F_{кр}$ =10,13

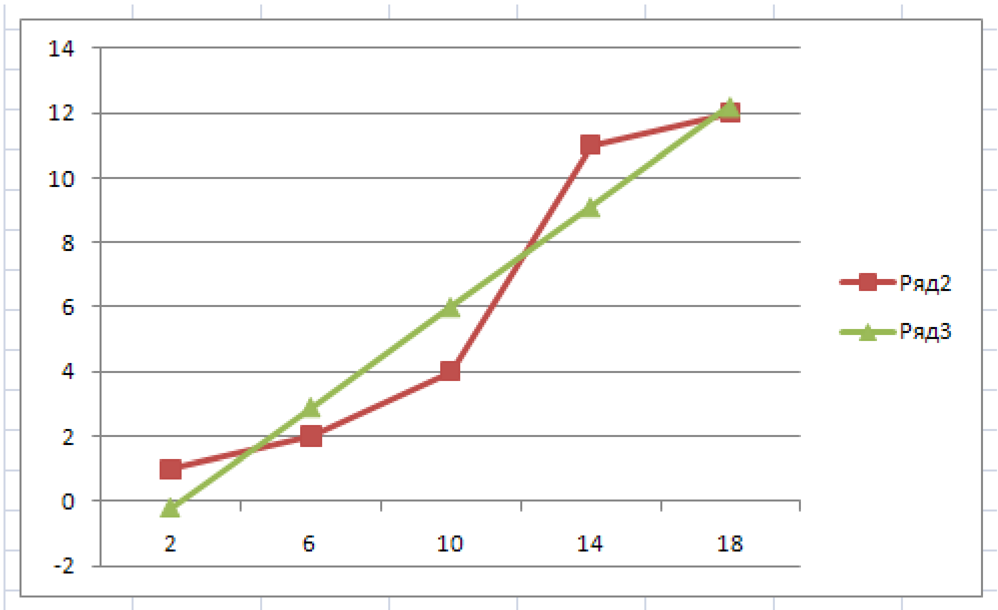
поскольку F = 29,2 > $F_{кр}$ =10,13 , то $\boldsymbol{R^2}$ значим при 5% уровне.

2. Наблюдаемому расчетному значению критерия F=29,2 соответствует значимость F=0,0124, которую можно определить с помощью функции Excel

Значимость F=**FРАСП**(F, v_1 , v_2) , где $v_1 = 1$ и $v_2 = 3$

Поскольку значимость F=0,0124< 0,05 , то $\boldsymbol{R^2}$ значим при 5% уровне.

Построить график разброса точек и модели



3 РАСЧЕТ И ОЦЕНКА ПАРАМЕТРОВ УРАВНЕНИЯ ПАРНОЙ РЕГРЕССИИ

3.1 Построение уравнения парной регрессии

Рассмотрим методику анализа уравнения парной регрессии на примере. Имеются выборочные данные о стоимости квартир и их общей площади в г. Смоленске, декабрь 2006г. (***n=15***):

Y	13,8	13.8	14	22,5	24	28	32	20,9	22	21,5	32	35	24	37,9	27,5
X	33	40	36	60	55	80	95	70	48	53	95	75	63	112	70

X – общая площадь квартиры, $м^2$;

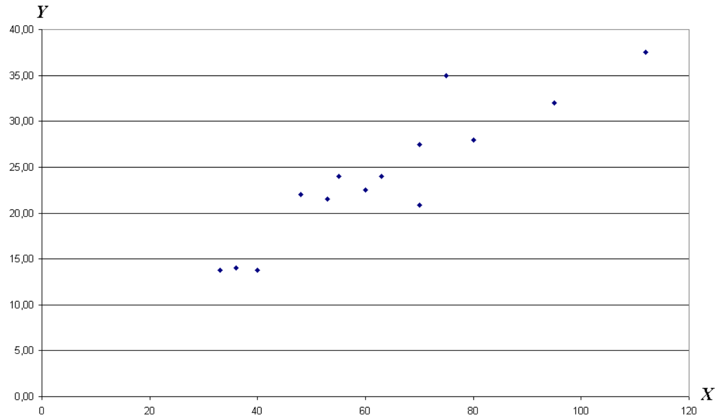
Y – рыночная стоимость квартиры, тыс. у.е.

Требуется:

1. Построить график ***Y=f(X)***, по которому подобрать модель уравнения регрессии.
2. Рассчитать параметры уравнения регрессии методом наименьших квадратов.
3. Оценить качество полученного уравнения регрессии с помощью средней ошибки аппроксимации.
4. Найти коэффициент эластичности.
5. Оценить силу связи между переменными с помощью коэффициентов корреляции и детерминации.
6. Оценить значимость коэффициентов корреляции и регрессии по критерию Стьюдента (***t*** – критерий) при уровне значимости ***α=0,05***.
7. Определить значение результативного признака, если его значение составит ***1,2*** от его среднего значения.

Решение

1. Построим график зависимости ***Y=f(X)*** в прямоугольной системе координат:



Характер расположения точек на графике показывает, что связь между переменными может выражаться линейным уравнением регрессии: $\boldsymbol{y = b_0 + b_1 x}$.

2. Параметры уравнения регрессии находим методом наименьших квадратов. Для этого строим вспомогательную таблицу регрессионного анализа (для этого удобно использовать пакет Excel):

	X	Y	X^2	Y^2	XY
	33	13,80	1089	190,44	455,4
	40	13,80	1600	190,44	552
	36	14	1296	196	504
	60	22,5	3600	506,25	1350
	55	24	3025	576	1320
	80	28	6400	784	2240
	95	32	9025	1024	3040
	70	20,9	4900	436,81	1463
	48	22	2304	484	1056
	53	21,5	2809	462,25	1139,5
	95	32	9025	1024	3040
	75	35	5625	1225	2625
	63	24	3969	576	1512
	112	37,9	12544	1436,41	4244,8
	70	27,5	4900	756,25	1925
Сумма	985	368,90	72111	9867,85	26466,7
Среднее значение	65,66667	24,59	4807,4	657,86	1764,447

Заполнив эти столбцы таблицы, можно вычислить параметры уравнения регрессии:

$$b_1 = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2} = \frac{1764,447 - 65,667 \cdot 24,59}{4807,4 - 65,667^2} = 0,302 .$$

$$b_0 = \overline{Y} - b_1 \cdot \overline{X} = 24,59 - 0,302 \cdot 65,667 = 4,759 .$$

Таким образом, уравнение регрессии имеет вид:

$$\hat{y} = 4,759 + 0,302 \cdot x ,$$

где $\hat{y}$ - оценка значения y, полученная с помощью уравнения регрессии.

Коэффициент регрессии b_1 показывает, что при увеличении общей площади квартиры на 1 м^2 стоимость квартиры в среднем увеличивается на 0,302 тыс у.е., или на 302 у.е.

3.2 Оценка параметров уравнеия регрессии

Воспользовавшись полученным уравнением регрессии, продолжим заполнять вспомогательную таблицу:

6-ой столбец – найдем значения $\hat{y}$ соответствующие, согласно уравнения регрессии, значениям x указанным в 1-ом столбце вспомогательной таблицы

7-ой столбец – найдем значения $y - \hat{y}$ для каждого значения y.

8-ой столбец – найдем значения $(y - \hat{y})^2$.

9-ый столбец – найдем ошибку аппроксимации для каждого значения y:

$$A = \left| \frac{y - \hat{y}}{y} \right| \cdot 100 \ \%.$$

Полностью вспомогательная таблица приведена ниже:

	x	y	x^2	y^2	xy	$\hat{y}$	$y-\hat{y}$	$(y-\hat{y})^2$	A
	33	13,80	1089	190,44	455,4	14,725	-0,92	0,86	6,666667
	40	13,80	1600	190,44	552	16,839	-3,04	9,24	22,02899
	36	14	1296	196	504	15,631	-1,631	2,660161	11,65
	60	22,5	3600	506,25	1350	22,879	-0,379	0,143641	1,684444
	55	24	3025	576	1320	21,369	2,631	6,922161	10,9625
	80	28	6400	784	2240	28,919	-0,919	0,844561	3,282143
	95	32	9025	1024	3040	33,449	-1,449	2,099601	4,528125
	70	20,9	4900	436,81	1463	25,899	-4,999	24,99	23,91866
	48	22	2304	484	1056	19,255	2,745	7,535025	12,47727
	53	21,5	2809	462,25	1139,5	20,765	0,735	0,540225	3,418605
	95	32	9025	1024	3040	33,449	-1,449	2,099601	4,528125
	75	35	5625	1225	2625	27,409	7,591	57,62328	21,68857
	63	24	3969	576	1512	23,785	0,215	0,046225	0,895833
	112	37,9	12544	1436,41	4244,8	38,583	-0,683	0,466489	1,802111
	70	27,5	4900	756,25	1925	25,899	1,601	2,563201	5,821818
Сумма	985	368,90	72111	9867,85	26466,7	368,86	0,05	118,63	135,35
Среднее значение	65,66667	24,59	4807,4	657,86	1764,447	24,59		7,91	9,023591

1. Качество уравнения регрессии принято оценивать с помощью средней ошибки аппроксимации. Ее находят следующим образом:

$$\bar{A} = \frac{\sum A_i}{n} = \frac{135,35}{15} \approx 9,02\%$$ - приведена в таблице в последней строке последнего столбца.

Следовательно, фактические значения стоимости квартир отличаются от стоимостей рассчитанных по уравнению регрессии в среднем на 9,03%.

Качество уравнения регрессии считается хорошим, если ошибка аппроксимации не превышает 8 – 10%. Следовательно, полученное уравнение регрессии можно оценить как вполне хорошее.

2. Оценим коэффициент эластичности. При линейной форме связи средний коэффициент эластичности находят по формуле:

$$Э = b_1 \frac{\bar{X}}{\bar{Y}},$$

где $\bar{X}$ и $\bar{Y}$ - средние значения признаков $\boldsymbol{X}$ и $\boldsymbol{Y}$ соответственно.

$$Э = b_1 \frac{\bar{X}}{\bar{Y}} = 0,302 \cdot \frac{65,667}{24,59} = 0,806.$$

Коэффициент эластичности показывает, что при увеличении общей площади квартиры на 1% ее стоимость в среднем возрастает на 0,806%.

3. Определим коэффициент корреляции, который определяет силу связи между признаками при их линейной зависимости:

Для расчета коэффициента корреляции воспользуемся формулой:

$r = \frac{cov(X, Y)}{\sigma_x \sigma_y} = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sigma_x \sigma_y}$. Для ее использования определим σ_x и σ_x.

Расчетные соотношения для σ_x и σ_x:

$$\sigma_x = \sqrt{\overline{X^2} - \overline{X}^2} = \sqrt{4807,4 - 65,667^2} = 22,254;$$

$$\sigma_y = \sqrt{\overline{Y^2} - \overline{Y}^2} = \sqrt{657,86 - 24,59^2} = 7,293;$$

Все цифры для вычисления σ_x и σ_x берутся из вспомогательной таблицы. Тогда

$$r = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\sigma_x \sigma_y} = \frac{1764,447 - 65,667 \cdot 24,59}{22,254 \cdot 7,293} = 0,922.$$

Такая связь, по шкале Шедока, оценивается как тесная (т.е. является близкой к линейной функциональной).

4. Найдем коэффициент детерминации:

$$d = r^2 = 0,922^2 = 0,85.$$

Коэффициент детерминации показывает какая часть (в %) изменения результативного признака обусловлена изменением факторного признака. В примере ***d=0,85*** обозначает, что 85% различий в стоимости квартир объясняется различием в их общей площади, а 15% - другими неучтенными факторами (местонахождение, этаж, благоустройство, близость к транспорту и т.д.).

5. Оценка значимости найденных коэффициентов

а) Оценка значимости коэффициента корреляции.

Коэффициент корреляции (и все другие оценки) вычислен по выборочным данным. Поэтому необходимо оценить его значимость. Для этого выдвинем нулевую гипотезу

H_0: r=0 (такая гипотеза означает предположение о том, что коэффициент корреляции незначим, т.е. изменение факторного признака ***X*** не приводит к изменению результативного признака ***Y***). Альтернативная гипотеза

H_1: r≠0 (т.е. принятие альтернативной гипотезы означает, что между признаками существует связь).

Для проверки правильности нулевой гипотезы применяется ***t***-критерий Стьюдента.

Определим расчетное значение ***t***- критерия:

$$t_{расч} = |r|\sqrt{\frac{n-2}{1-r^2}} = 0{,}922\sqrt{\frac{15-2}{1-0{,}922^2}} = 8{,}586\ .$$

Находим значение $t_{крит}$ по таблицам ***t***- распределения Стьюдента при уровне значимости ***α=0,05***и числе степеней свободы ***ν=n-2=15-2=13*** для двухсторонней критической области. Получаем $t_{крит}=2{,}16$.

Т.к. $t_{расч}> t_{крит}$, (***8,586>2,16***), то нулевая гипотеза отвергается и принимается альтернативная гипотеза ***H_1*** – общая площадь квартир оказывает статистически существенное влияние на их стоимость.

б) Оценка значимости коэффициента регрессии ***b_1***.

Оценка значимости коэффициента регрессии ***b_1*** также проводится по ***t***- критерию Стьюдента.

Выдвигаем нулевую гипотезу ***H_0: b_1=0*** (такая гипотеза означает предположение о том, что коэффициент регрессии незначим. Альтернативная гипотеза ***H_1: b_1≠0***.

Определим расчетное значение ***t***- критерия:

$t_{расч}=\frac{b_1}{m_{b_1}}$. Величину m_{b_1} находим по формуле:

$$m_{b_1}=\sqrt{\frac{\sum(y-\hat{y})^2}{(n-2)\cdot\sum(x-\overline{X})^2}}=\sqrt{\frac{\sum(y-\hat{y})^2}{(n-2)\cdot\sigma_x^2\cdot n}}=\sqrt{\frac{118,63}{(15-2)\cdot 22,254^2\cdot 15}}=0,035$$

Тогда $t_{расч}=\frac{b_1}{m_{b_1}}=\frac{0,302}{0,035}\approx 8,63$.

Находим значение $\boldsymbol{t_{крит}}$ по таблицам ***t***- распределения Стьюдента при уровне значимости $\boldsymbol{\alpha=0,05}$и числе степеней свободы $\boldsymbol{\nu=n-2=15-2=13}$ для двухсторонней критической области. Получаем как и в предыдущем случае $\boldsymbol{t_{крит}=2,16}$.

Т.к. $\boldsymbol{t_{расч}}>$ $\boldsymbol{t_{крит}}$, (8,63>2,16), то нулевая гипотеза отвергается и принимается альтернативная гипотеза $\boldsymbol{H_1}$ –коэффициент регрессии значим, подтверждается вывод о том, что общая площадь квартир оказывает статистически существенное влияние на их стоимость.

6. Статистическая надежность уравнения регрессии проверяется с помощью ***F***- критерия Фишера.

Для линейного уравнения регрессии расчетное значение F- критерия находится по формуле:

$$F_{расч}=\frac{r^2}{1-r^2}\cdot(n-2)=\frac{0,85}{1-0,85}\cdot(15-2)=73,67.$$

Находим значение $\boldsymbol{F_{крит}}$ по таблицам ***F***- распределения Фишера при уровне значимости $\boldsymbol{\alpha=0,05}$и числе степеней свободы $\boldsymbol{k_1=k=1}$ и $\boldsymbol{k_2=n-k-1=15-1-1=13}$, где ***k*** – число параметров при переменной ***X***. Получаем $\boldsymbol{F_{крит}=F_{0,05;1;13}=4,67}$.

Т.к. $\boldsymbol{F_{расч}}>\boldsymbol{F_{крит}}$, то уравнение регрессии статистически значимое или надежное.

7. Прогнозное значение результативного признака определяется при помощи подстановки в уравнение регрессии возможного значения факторного признака ***X***:

$$x_{расч} = 1,2 \cdot \overline{X} = 1,2 \cdot 65,667 = 78,8.$$

$$\hat{y} = 4,759 + 0,302 \cdot 78,8 = 28,56.$$

Следовательно, возможная стоимость квартиры общей площадью 78,8 м^2 составит 28,56 тыс. у.е.

4 ИНТЕРВАЛЬНАЯ ОЦЕНКА ПАРАМЕТРОВ РЕГРЕССИОННОЙ МОДЕЛИ

4.1 Интервальная оценка коэффициентов регрессии

Задача

Регрессионная модель зависимости расходов на питание $\boldsymbol{y}$ от личного дохода $\boldsymbol{x}$ рассчитанная ранее имеет вид:

$$\hat{y} = -1,75 + 0,775x.$$

Известны следующие данные: ***n=5, var(x)=32,*** $\overline{x^2} = 132$***, var(e)=1,98.***

Найти стандартные ошибки коэффициентов регрессии, оценить значимость коэффициента регрессии $\boldsymbol{b}$ при 5%-ном уровне значимости, построить доверительный интервал для коэффициента $\boldsymbol{\beta}$ при том же уровне значимости.

Решение.

1. Расчет стандартных ошибок коэффициентов регрессии $(\boldsymbol{a}, \boldsymbol{b})$.

   Находим несмещенную оценку дисперсии $\boldsymbol{\sigma^2}$ (ее также называют остаточной дисперсией):

$$S^2 = \frac{n}{n-2} var(e) = \frac{5}{5-2} \cdot 1,98 = 3,3$$

Тогда

$$S_a^2 = \frac{\overline{x^2} \cdot S^2}{n \cdot var(x)} = \frac{132 \cdot 3,3}{5 \cdot 32} = 2,7225$$

$$S_b^2 = \frac{S^2}{n \cdot var(x)} = \frac{3,3}{5 \cdot 32} = 0,020625$$

Тогда стандартные ошибки коэффициентов регрессии равны:

$$S_a = \sqrt{2,7225} = 1,65$$

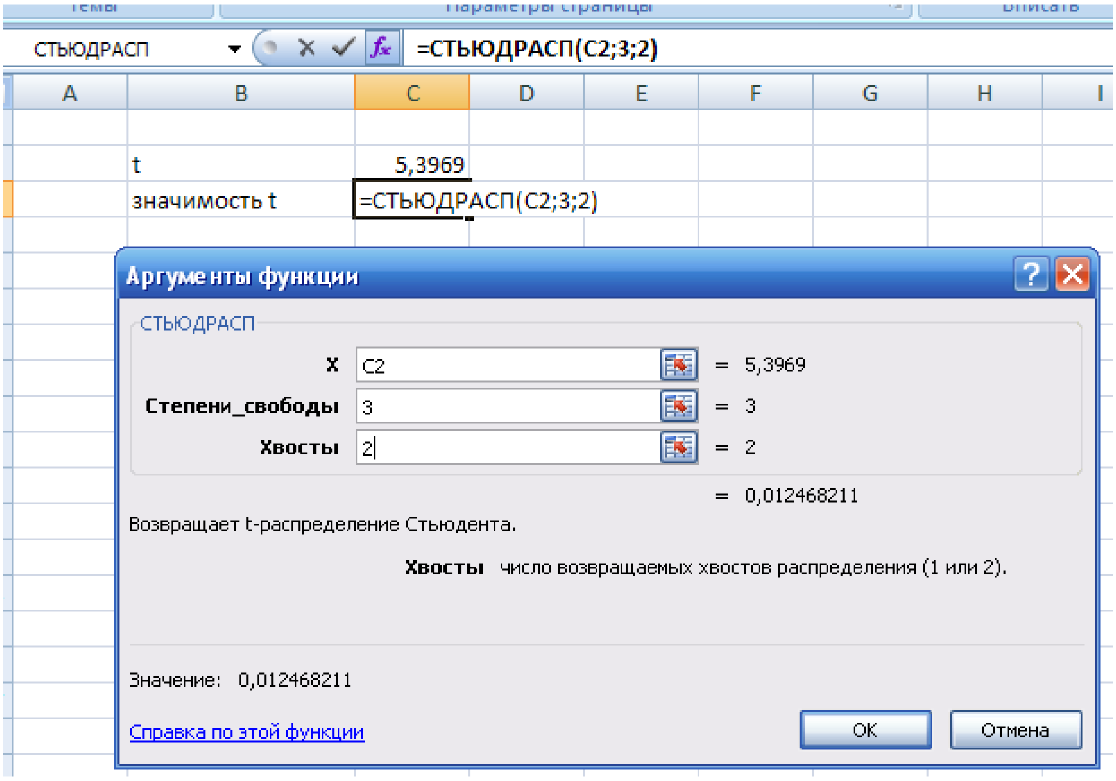
$$S_b = \sqrt{0,020625} = 0,1436.$$

2. Найдем наблюдаемое значение критерия t:

$$t = \frac{b}{S_b} = \frac{0,775}{0,1436} = 5,3969.$$

Для вычисления *значимость* t воспользуемся Excel

Значимость t =СТЬЮДРАСП(*t; v; 2)*.



Находим *значимость* t =*0,01247*.

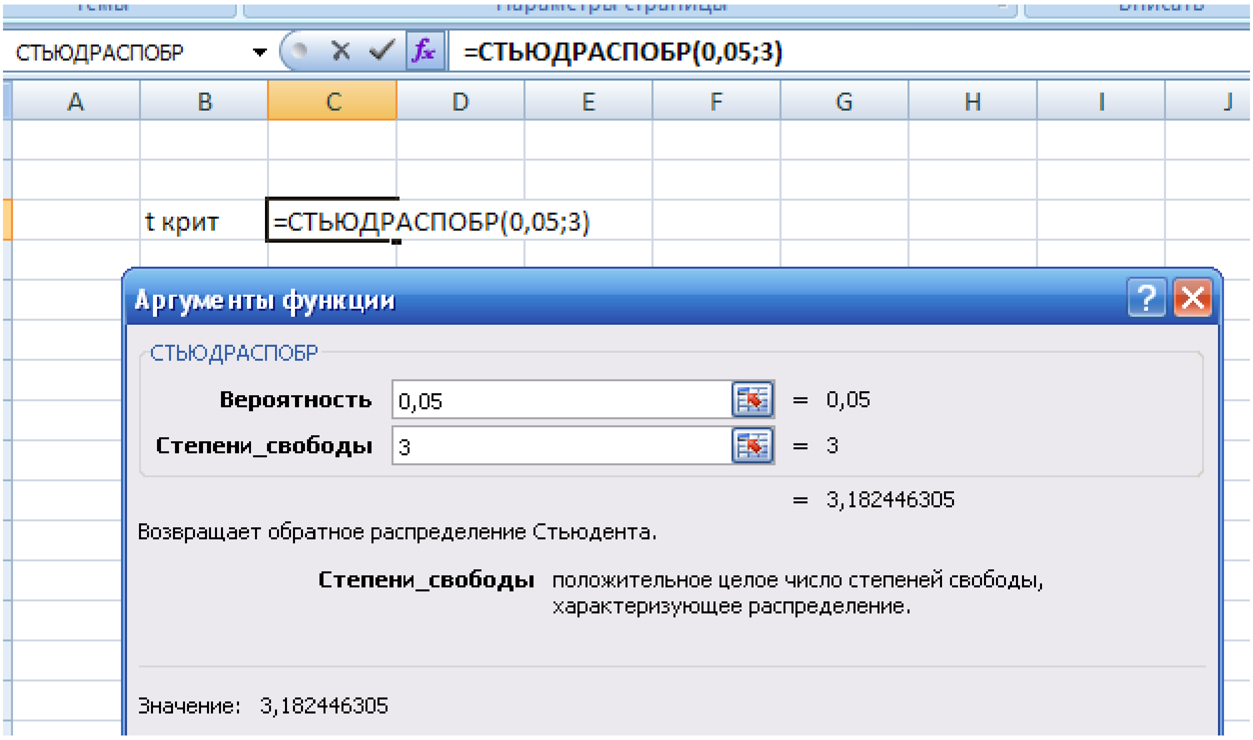
Вывод: т.к *значимость* t меньше заданного стандартного уровня значимости (0,05) , то коэффициент b значим.

3. Найдем доверительный интервал для β.

Для этого требуется найти $t_{крит}$. Воспользуемся таблицами распределения Стьюдента или или функцией Excel

$$t_{крит} = \text{СТЬЮДРАСПОБР}(\alpha,\ v)$$

Получаем:



$$t_{крит} = 3,1824.$$

Доверительный интервал для β имеет вид:

$$b - t_{крит}S_b < \ \beta < b + t_{крит}S_b$$

Подставляя значения, получаем:

$$0,775 - 3,18 \cdot 0,1436 < \ \beta < 0,775 + 3,18 \cdot 0,1436$$

Подсчитывая, получаем доверительный интервал для β:

$$0,318 < \ \beta < 1,23.$$

4.2 Интервальная оценка прогноза эндогенной переменной

Задача.

Известны данные об объеме продаж фирмы и о затратах на рекламу. Оценить предполагаемый объем продаж при затратах на рекламу 5,5 усл .ед.

Найти стандартную ошибку предсказания и 99%-ный доверительный интервал для полученной оценки.

Исходные данные (усл.ед.):

X – затраты на рекламу	5	8	6	5	3	9	12	4	3	10
Y – объем продаж	72	76	78	70	68	80	82	65	62	90

Предсказание можно сделать, воспользовавшись Excel:

$$\mathbf{y_p} = \text{ПРЕДСКАЗ}\left(\boldsymbol{x_p}, \text{массив}\boldsymbol{x}, \text{массив}\boldsymbol{y}\right).$$

Для вычисления стандартной ошибки предсказания

$$\boldsymbol{S_p} = \boldsymbol{S}\sqrt{\mathbf{1} + \frac{\mathbf{1}}{\boldsymbol{n}} + \frac{\left(\boldsymbol{x_p} - \overline{\boldsymbol{x}}\right)^{\mathbf{2}}}{\boldsymbol{nvar}(\boldsymbol{x})}}$$

требуется найти $\overline{\boldsymbol{x}}$**,** $\boldsymbol{var}(\boldsymbol{x})$, $\boldsymbol{S}$.

Для вычислений воспользуемся функциями Excel:

$$\overline{\boldsymbol{x}} = \text{СРЗНАЧ}(\text{массив}\boldsymbol{x}),$$

$$\boldsymbol{var}(\boldsymbol{x}) = \text{ДИСП}(\text{массив}\boldsymbol{x}),$$

$$\boldsymbol{S} = \text{СТОШ}\boldsymbol{yx}(\text{массив } \boldsymbol{y}\text{ , массив } \boldsymbol{x}).$$

Чтобы найти доверительный интервал для действительного значения $\boldsymbol{y_p}$ по выражению

$$\widehat{\mathbf{y_p}} - \mathbf{t}_{\text{крит}}\mathbf{S_p} < \mathbf{y_p} < \widehat{\mathbf{y_p}} + \mathbf{t}_{\text{крит}}\mathbf{S_p},$$

необходимо найти $\mathbf{t}_{\text{крит}}$. Найдем его с помощью функции Excel:

$$\mathbf{t}_{\text{крит}} = \text{СТЬДРАСПОБР}(\boldsymbol{\alpha}, \boldsymbol{v}),$$

где $\boldsymbol{v} = \boldsymbol{n} - \mathbf{2}$, $\boldsymbol{\alpha} = \mathbf{1} - \mathbf{0{,}99} = \mathbf{0{,}01}$.

$\boldsymbol{n} = \mathbf{10}$ (число измерений в заданной таблице).

A	B	C
	x	y
	5	72
	8	76
	6	78
	5	70
	3	68
	9	80
	12	82
	4	65
	3	62
	10	90
среднее	6,5	
var(x)	9,611111	
S	4,240324	
tкрит	3,355387	
yпредск	71,86647	
Sстанд	4,523465	

Доверительный интервал для действительного значения $\boldsymbol{y_p}$:

$$\mathbf{71,866 - 3,355 \cdot 4,52 < \ y_p < 71,866 + 3,355 \cdot 4,52.}$$

$$\mathbf{56,7 < \ y_p <} \ 87,\mathbf{03.}$$

5 НЕЛИНЕЙНЫЕ РЕГРЕССИИ

5.1 Методы лианеризации нелинейных уравнений

Различают два класса нелинейных регрессий:

1. Регрессии, нелинейные относительно включенных в анализ объясняющих переменных, но линейные по оцениваемым параметрам, например:
   - полиномы различных степеней –

$$\hat{y}_x = a + bx + cx^2,$$

$$\hat{y}_x = a + bx + cx^2 + dx^3;$$

   - равносторонняя гипербола –

$$\hat{y}_x = a + \frac{b}{x};$$

   - полулогарифмическая функция –

$$\hat{y}_x = a + b \cdot lnx.$$

2. Регрессии, нелинейные по оцениваемым параметрам, например
   - степенная –

$$\hat{y}_x = ax^b;$$

   - показательная –

$$\hat{y}_x = ab^x;$$

   - экспоненциальная –

$$\hat{y}_x = e^{a+bx}$$

Регрессии, нелинейные по включенным переменным приводятся к линейному виду простой заменой переменных, дальнейшая оценка параметров проводится с помощью МНК.

Например, парабола второй степени $\hat{y}_x = a + bx + cx^2$ приводится к линейному виду с помощью замены $x = x_1,\ x^2 = x_2$. В результате получаем двухфакторное уравнение $\hat{y}_x = a + b \cdot x_1 + c \cdot x_2$

Анализ параметров проводится по МНК для многофакторной модели. Парабола второй степени применяется в случаях, когда для определенного интервала значений фактора x меняется характер связи рассматриваемых признаков, т.е. прямая связь меняется на обратную или обратная связь меняется на прямую.

Также к линейному виду заменой переменной приводятся равносторонняя гипербола, полулогарифмическая функция и многие другие функции.

Регрессии, нелинейные по оцениваемым параметрам, делятся на два вида:

1. нелинейные модели, внутренне линейные – приводятся к линейному виду с помощью соответствующих преобразований, например, логарифмированием;
2. нелинейные модели, внутренне нелинейные – к линейному виду не приводятся.

К внутренне линейным моделям относятся, например, степенная функция, показательная функция, экспоненциальная функция, логистическая функция

$$\hat{y}_x = \frac{a}{1 + b \cdot e^{-cx}},$$

обратная функция

$$\hat{y}_x = \frac{1}{a + bx}.$$

К внутренне нелинейным моделям относятся, например, следующие:

$$\hat{y}_x = a \cdot \left(1 - \frac{1}{1 - x^b}\right),$$

$$\hat{y}_x = a + b \cdot x^c \ .$$

5.2 Кэффициент эластичности нелинейных моделей

При анализе регрессий широко используется коэффициент эластичности. Коэффициент эластичности показывает на сколько процентов измениться в среднем результат, если фактор изменится на 1%. Коэффициент эластичности рассчитывается, в общем виде, по формуле:

$$Э = f'(x) \cdot \frac{x}{y}.$$

Так как для большинства функций (за исключением степенной) коэффициент эластичности не является постоянной величиной, а зависит от соответствующего значения фактора x, то обычно рассчитывается средний коэффициент эластичности:

$$\overline{Э} = f'(x) \cdot \frac{\bar{x}}{\bar{y}}.$$

Формулы для расчета средних коэффициентов эластичности для наиболее часто используемых уравнений регрессии – в таблице.

Вид функции y	Первая производная y'	Средний коэффициент эластичности $\overline{Э}$
1	2	3
$y = a + bx + \varepsilon$	b	$\frac{b \cdot \bar{x}}{a + b \cdot \bar{x}}$
$y = a + bx + cx^2 + \varepsilon$	$b + 2cx$	$\frac{(b + 2c\bar{x}) \cdot \bar{x}}{a + b \cdot \bar{x} + c \cdot \bar{x}^2}$

$y = a + \frac{b}{x} + \varepsilon$	$\frac{-b}{x^2}$	$\frac{-b}{a \cdot \overline{x} + b}$
$y = ax^b\varepsilon$	abx^{b-1}	b
$y = ab^x\varepsilon$	$alnb \cdot b^x$	$\overline{x} \cdot lnb$
$y = a + b \cdot lnx + \varepsilon$	$\frac{b}{x}$	$\frac{b}{a + b \cdot ln\overline{x}}$
$y = \frac{a}{1 + b \cdot e^{-cx+\varepsilon}}$	$\frac{abc \cdot e^{-cx}}{(1 + b \cdot e^{-cx})^2}$	$\frac{bc \cdot \overline{x}}{b + e^{c\overline{x}}}$
$y = \frac{1}{a + bx + \varepsilon}$	$\frac{-b}{(a + b \cdot x)^2}$	$\frac{-b \cdot \overline{x}}{a + b \cdot \overline{x}}$

Возможны случаи, когда расчет коэффициента эластичности не имеет смысла. Это происходит тогда, когда для рассматриваемых признаков бессмысленно определение изменения в процентах.

Эластичность показывает на сколько процентов изменяется функция $y = f(x)$ при изменении независимой переменной на 1%.

Для степенной функции $y = ax^b$ эластичность постоянна и равна b.

Пусть, например, при исследовании получена зависимость расходов на питание от дохода семьи в виде

$$\hat{y} = 0,305 \cdot x^{1,183}.$$

Это означает, что эластичность расходов на продукты питания по доходу составляет $1,183$, т.е. увеличение дохода семьи на 1% приведет к увеличению расходов на питание на $1,183\%$. Коэффициент $0,305$ не имеет экономического смысла.

Для экспоненциального уравнения временного тренда $y = ae^{rt}$ эластичность также постоянна и равна r.

Пусть, например, при исследовании получена зависимость расходов на питание от дохода семьи в виде уравнения временного тренда:

$$\hat{y} = 0,543 \cdot e^{0,667t}.$$

Это означает, что расходы на продукты питания в течение выборочного периода росли с темпом ***66,7%*** в год. Постоянный множитель ***0,543*** показывает, что в момент начала отсчета ***t=0*** общие расходы на питание составили ***0,543*** у.е.

Эластичность линейной функции $y = a + bx$ не является постоянной величиной, а зависит от x:

$$Э = \frac{bx}{y} = \frac{bx}{a + bx}.$$

Поэтому для линейной функции обычно вычисляется средний показатель эластичности:

$$Э = \frac{b\bar{x}}{\bar{y}},$$

где $\bar{x}, \bar{y}$ – средние значения переменных x и y в выборке.

Если, например, при исследовании получена зависимость расходов на питание от дохода семьи в виде линейного уравнения вида:

$$\hat{y} = -1,75 + 0,775x,$$

причем $\bar{x} = 10$, $\bar{y} = 6$, средний показатель эластичности равен $\frac{0,775 \cdot 10}{6} = 1.29$ и показывает, что с увеличением дохода на ***1%*** расходы на питание возрастут в среднем на ***1,29***%.

5.3 Показатели тесноты связи

Уравнение нелинейной регрессии, так же как и в случае линейной зависимости, дополняется показателем тесноты связи. В данном случае это индекс корреляции:

$$\boldsymbol{\rho_{xy}} = \sqrt{\mathbf{1} - \frac{\boldsymbol{vare}}{\boldsymbol{vary}}},$$

где $\boldsymbol{vare} = \frac{1}{n}\sum(\boldsymbol{y} - \widehat{\boldsymbol{y}})^2$ – эту величину часто называют остаточной дисперсией,

$\boldsymbol{vary} = \frac{1}{n}\sum(\boldsymbol{y} - \overline{\boldsymbol{y}})^2$ – эту величину часто называют дисперсией результативного признака $\boldsymbol{y}$.

$$\mathbf{0} \le \boldsymbol{\rho_{xy}} \le \mathbf{1}.$$

Чем ближе значение индекса корреляции к единице, тем теснее связь рассматриваемых признаков, тем более надежно уравнение регрессии.

Квадрат индекса корреляции называется индексом детерминации и характеризует долю дисперсии результативного признака $\boldsymbol{y}$, которая объясняется регрессией:

$$\boldsymbol{\rho_{xy}^2} = \mathbf{1} - \frac{\boldsymbol{vare}}{\boldsymbol{vary}} = \frac{\boldsymbol{var\widehat{y}}}{\boldsymbol{vary}},$$

где $\boldsymbol{var\widehat{y}} = \frac{1}{n}\sum(\widehat{\boldsymbol{y}} - \overline{\boldsymbol{y}})^2$

Т.е. индекс детерминации имеет тот же смысл, что коэффициент детерминации в линейной регрессии.

Индекс детерминации $\boldsymbol{\rho_{xy}^2}$ можно сравнивать с коэффициентом детерминации $\boldsymbol{R^2}$ для обоснования возможности применения линейной функции. Чем больше кривизна линии регрессии, тем величина $\boldsymbol{R^2}$ меньше $\boldsymbol{\rho^2_{xy}}$. А близость этих показателей указывает на то, что нет необходимости усложнять вид уравнения регрессии и можно использовать линейную функцию.

Индекс детерминации используется для проверки существенности в целом уравнения регрессии по **F** – критерию Фишера:

$$\boldsymbol{F}_{\text{набл}} = \frac{\boldsymbol{\rho^2_{xy}}}{\boldsymbol{1-\rho^2_{xy}}} \cdot \frac{\boldsymbol{n-m-1}}{\boldsymbol{m}}$$

где **n** – число наблюдений, **m** – число параметров при переменной **x**.

При рассмотрении парной линейной регрессии для нахождения $\boldsymbol{F}_{\text{набл}}$ использовалось выражение

$$\boldsymbol{F = \frac{R^2(n-2)}{1-R^2}},$$

т.к. в парной линейной регрессии **m =1.**

О качестве нелинейного уравнения регрессии можно также судить по средней ошибке аппроксимации.

5.4 Примеры построения нелинейных регрессийй путем замены переменной

Имеются данные о зависимости между ежегодным потреблением сахара $\boldsymbol{y}$ и годовым доходом $\boldsymbol{x}$ (усл. ед.) десяти семей:

x	1	2	3	4	5	6	7	8	9	10
y	2	7	9	12	10	12	11	12	13	14

Рассмотрим различные варианты уравнения регрессии.

1. Линейной уравнение регрессии вида

$$y = \alpha + \beta x + \varepsilon.$$

Для оценки параметров линейного уравнения регрессии воспользуемся Excel, функцией:

$$b = \text{ЛИНЕЙН}(\text{массив } y; \text{массив } x),$$

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

	x	y	yоценка	(y-ycp)кв	(Yоц-ycp)кв
	1	2	6,018182	64	15,85487603
	2	7	6,90303	9	9,591221304
	3	9	7,787879	1	4,893480257
	4	12	8,672727	4	1,761652893
	5	10	9,557576	0	0,19573921
	6	12	10,44242	4	0,19573921
	7	11	11,32727	1	1,761652893
	8	12	12,21212	4	4,893480257
	9	13	13,09697	9	9,591221304
	10	12	13,98182	4	15,85487603
Среднее	5,5	10		10	6,459393939
ЛинРегр	b	0,884848		var(y)	var(yоц)
	a	5,133333			
Rквадр	к-нт детерм	0,645939			
S	ст.ош. регр.	2,103748			

Оценка параметров уравнения по выборочным данным приводит к уравнению регрессии вида:

$$\hat{y} = 5,13 + 0,88x.$$

Чтобы можно было сравнивать разные уравнения регрессии, рассчитаем коэффициент детерминации R^2 и стандартную ошибку регрессии **S.** (S^2 служит мерой разброса зависимой переменной вокруг линии регрессии).

$$R^2 = \frac{var(\hat{y})}{var(y)} = 0,65$$

Для нахождения $\boldsymbol{S}$ воспользуемся Excel, функцией

$$\boldsymbol{S} = \text{СТОШ}\boldsymbol{yx}(\text{массив } \boldsymbol{y}, \text{массив } \boldsymbol{x}).$$

$$\boldsymbol{S = 2,10}.$$

2. Нелинейной уравнение регрессии вида

$$\boldsymbol{y = \alpha + \beta\sqrt{x} + \varepsilon}.$$

Произведем замену переменной $\boldsymbol{z = \sqrt{x}}$. Уравнение примет вид:

$$\boldsymbol{y = \alpha + \beta z + \varepsilon},$$

т.е. станет линейным. Применим метод МНК для оценки тех же параметров:

получили уравнение линейной регрессии

$$\boldsymbol{\hat{y} = 0,774 + 4,106z},$$

$$\boldsymbol{R^2 = \frac{var(\hat{y})}{var(y)} = 0,762},$$

$$\boldsymbol{S = 1,725}.$$

A	B	C	D	E	F	G
	x	y	z=кор(x)	yоценки	(y-yср)кв	(yоценки-yср)кв
	1	2	1	4,879976282	64	26,21464288
	2	7	1,414214	6,580919453	9	11,69011179
	3	9	1,732051	7,886099053	1	4,468577213
	4	12	2	8,986416355	4	1,027351806
	5	10	2,236068	9,955815357	0	0,001952283
	6	12	2,44949	10,83221905	4	0,692588542
	7	11	2,645751	11,63815542	1	2,683553165
	8	12	2,828427	12,3883027	4	5,703989773
	9	13	3	13,09285643	9	9,56576088
	10	12	3,162278	13,75923991	4	14,13188473
Среднее	5,5	10	2,246828	10	10	7,618041306
ЛинРегр	b	4,10644			var(y)	var(yоц)
	a	0,773536				
Rквадр	к-нт детерм	0,761804				
S	ст.ош. регр.	1,725528				

Замечание: при вычислениях в качестве столбца $\boldsymbol{x}$ использовался столбец $\boldsymbol{z = \sqrt{x}}$.

3. В качестве уравнения регрессии рассмотрим нелинейной уравнение

$$\boldsymbol{y = \alpha + \frac{\beta}{x} + \varepsilon}.$$

Воспользуемся подстановкой $\boldsymbol{z = \frac{1}{x}}$. Тогда уравнение примет линейный вид $\boldsymbol{y = \alpha + \beta z + \varepsilon}$. Воспользовавшись МНК, оценим те же параметры:

получили уравнение линейной регрессии

$$\boldsymbol{\hat{y} = 13{,}42 - 11{,}67z},$$

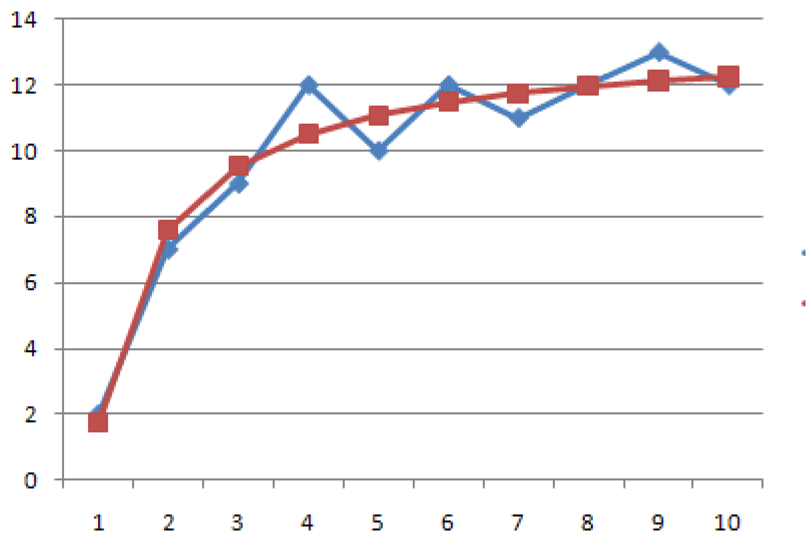
$$\boldsymbol{R^2 = \frac{var(\hat{y})}{var(y)} = 0{,}942},$$

$$\boldsymbol{S = 0{,}85}.$$

A	B	C	D	E	F	G
	x	y	z=1/(x)	yоценки	(y-yср)кв	(yоценки-yср)кв
	1	2	1	1,748572061	64	68,08606303
	2	7	0,5	7,583242471	9	5,840716955
	3	9	0,333333	9,528132607	1	0,222658836
	4	12	0,25	10,50057768	4	0,250578009
	5	10	0,2	11,08404472	0	1,175152947
	6	12	0,166667	11,47302274	4	2,169796004
	7	11	0,142857	11,75086419	1	3,065525419
	8	12	0,125	11,95924528	4	3,838642059
	9	13	0,111111	12,12131946	9	4,499996234
	10	12	0,1	12,2509788	4	5,066905551
Среднее	5,5	10	0,292897	10	10	9,421603504
ЛинРегр	b	-11,6693			var(y)	var(yоц)
	a	13,41791				
Rквадр	к-нт детерм	0,94216				
S	ст.ош. регр.	0,850291				

Качество оценивания 3-его варианта выше, чем у первых двух.

Графики $\boldsymbol{y(x)}$ и $\hat{\boldsymbol{y}}(\boldsymbol{x})$ для третьего уранения имеют вид



5.5 Примеры построения нелинейных регрессийй путем логарифмирования исходного уравнения

Нелинейность по параметру часто удается устранить путем логарифмирования исходного уравнения. Например, Следующие нелинейные уравнения, после логарифмирования, сводятся к линейным:

- Степенная функция $\boldsymbol{y = \alpha x^{\beta} \varepsilon}$. После логарифмирования $\boldsymbol{lny = ln\alpha + \beta \cdot lnx + ln\varepsilon}$.
- Экспоненциальная функция $\boldsymbol{y = \alpha e^{\beta x} \varepsilon}$. После логарифмирования $\boldsymbol{lny = ln\alpha + \beta \cdot x + ln\varepsilon}$.

Но уравнение $\boldsymbol{y = \alpha x^{\beta} + \varepsilon}$, где случайный член $\boldsymbol{\varepsilon}$ является аддитивным такими преобразованиями привести к линейному виде невозможно. В этом случае используются специальные *итерационные методы оценивания нелинейной регрессии*.

В экономике применяются функции вида:

- $\boldsymbol{y = \alpha x^{\beta} \varepsilon}$ при моделировании кривых спроса;

- $y = \alpha e^{\beta x}\varepsilon$ при моделировании временных трендов, при этом в качестве переменной x используется время t, а под β понимается постоянный темп прироста $\mathbf{r}$, т.е. уравнение принимает вид $y = \alpha e^{rt}\varepsilon$.

Пример

Задана зависимость расходов на питание y в зависимости от доходов семьи x. Определить параметры степенной функции и временного тренда, описывающего данную зависимость.

X (доход)	9	6	10	14	18
Y (расходы)	1	2	4	11	12

1. Определим параметры степенной функции $y = \alpha x^{\beta}\varepsilon$, которая является моделью представленных данных. Для этого прологарифмируем функцию, т.е. представим ее в виде линейной функции:

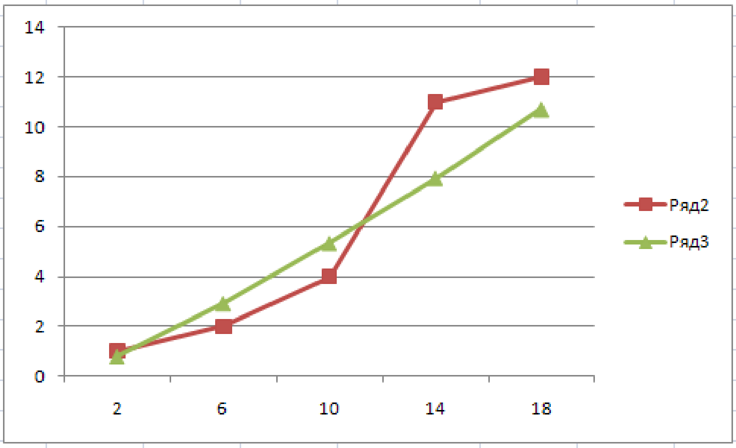
$$ln y = ln\alpha + \beta \cdot ln x + ln\varepsilon.$$

Т.о., для дальнейших расчетов, нужно найти lnx и lny. Если обозначить $lny = u,\ lnx = z,$ то новое уравнение

$$u = ln\alpha + \beta \cdot z + ln\varepsilon$$

будет линейным относительно $\mathbf{u}$ и $\mathbf{z}$, т.е. для построения уравнения регрессии можно пользоваться МНК. Расчетная таблица:

A	B	C	D	E	F	G	H
		x	y	z=lnx	u=lny	uоц=lnyоц	yоц
		2	1	0,693147	0	-0,22922048	0,795153
		6	2	1,791759	0,693147	1,070982113	2,918244
		10	4	2,302585	1,386294	1,67554189	5,341689
		14	11	2,639057	2,397895	2,073755235	7,954639
		18	12	2,890372	2,484907	2,371184706	10,71007
Среднее				2,063384	1,392449		
К-ты лин. регр.							
b	1,183495						
a	-1,04956						
e	2,718282						
е в степ.а	0,350093						



Оценив регрессию (с помощью МНК) между **u** и **z,** получаем

$$\hat{u} = -1,05 + 1,183z,$$

или, подставляя значения **u** и **z,** получаем:

$$ln\hat{y} = -1,05 + 1,183lnx.$$

Выполнив обратные преобразования, получим уравнение для $\hat{y}$:

$$\hat{y} = e^{-1,05} \cdot x^{1,183} = 0,305 \cdot x^{1,183},$$

т.е. параметры степенной функции $y = \alpha x^{\beta}\varepsilon$, которая является моделью представленных данных, определены.

2. Определим параметры временного тренда, описывающего представленные данные. Для этого введем в таблицу параметр время t, считая, что все отсчеты получены последовательно, через равные промежутки времени. Будем пользоваться уравнением временного тренда:

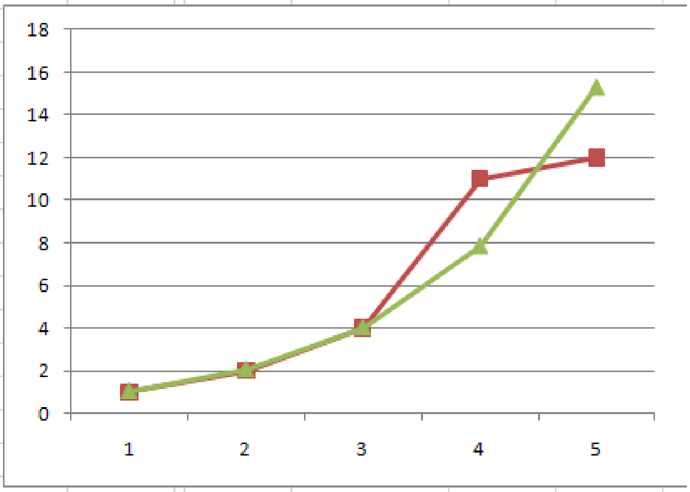
$$y = \alpha e^{rt}\varepsilon.$$

После логарифмирования данное уравнение приводится линейному виду:
$$ln y = ln\alpha + r \cdot t + ln\varepsilon.$$

Обозначив $\boldsymbol{lny = u}$, получим: $\quad \boldsymbol{u = ln\alpha + r \cdot t + ln\varepsilon,}$

т.е. уравнение, связывающее $\boldsymbol{u}$ и $\boldsymbol{t}$ линейное, можно воспользоваться МНК.

A	B	C	D	F	G	H
	t	x	y	u=lny	uоц=lnyоц	yоц
	1	2	1	0	0,057536414	1,059224
	2	6	2	0,693147	0,724992554	2,064716
	3	10	4	1,386294	1,392448693	4,024693
	4	14	11	2,397895	2,059904832	7,845223
	5	18	12	2,484907	2,727360971	15,29248
Среднее	3			1,392449		
К-ты лин. регр.						
b	0,667456					
a	-0,60992					
e	2,718282					



Оценив регрессию (с помощью МНК) между $\mathbf{u}$ и $\boldsymbol{t,}$ получаем

$$\hat{u} = -0,61 + 0,667t,$$

или, подставляя значение $\mathbf{u,}$ получаем:

$$ln\hat{y} = -0,61 + 0,667t.$$

Выполнив обратные преобразования, получим уравнение для $\hat{y}$:

$$\hat{y} = e^{-0,61} \cdot e^{0,667t} = 0,543 \cdot e^{0,667t},$$

т.е. параметры уравнения временного тренда $\boldsymbol{y = \alpha e^{rt}\varepsilon}$, которое является моделью представленных данных, определены.

5.6 Пример выбора модели регрессии

По семи районам области получена следующая статистика:

Район	Расходы на покупку продтоваров в общих расходах % (y)	Среднедневная зарплата одного работающего руб. (x)
1	68,8	45,1
2	61,2	59,0
3	59,9	57,2
4	56,7	61,8
5	55,0	58,8
6	54,3	47,2
7	49,3	55,2

1.Построим модели различного вида:

А) линейную регрессионную

Б) степенную $\boldsymbol{y = ax^b}$

В) показательную $\boldsymbol{y = ab^x}$

Г) гиперболу $\boldsymbol{y = a + b \cdot \frac{1}{x}}$

При составлении моделей Б) – Г) предварительно проведем линеаризацию.

.

Формулы для расчета

$$b = \frac{\overline{XY} - \overline{X} \cdot \overline{Y}}{\overline{X^2} - (\overline{X})^2}.$$

$$a = \overline{Y} - b \cdot \overline{X}$$

$\boldsymbol{var(y) = \frac{1}{n}\sum(y_i - \overline{y})^2 -}$ дисперсия наблюдаемых значений $\boldsymbol{y}$;

$\boldsymbol{var(\hat{y}) = \frac{1}{n}\sum(\hat{y}_i - \overline{y})^2}$ – дисперсия расчетных значений $\boldsymbol{y}$;

$$\boldsymbol{var(e)} = \frac{1}{n}\sum(e_i - \bar{e})^2 = \frac{1}{n}\sum e_i^2$$ – дисперсия остатков.

$$r = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\sqrt{varx \cdot vary}}$$

$$R^2 = \frac{var(\hat{y})}{var(y)} = 1 - \frac{var(e)}{var(y)} = r^2$$

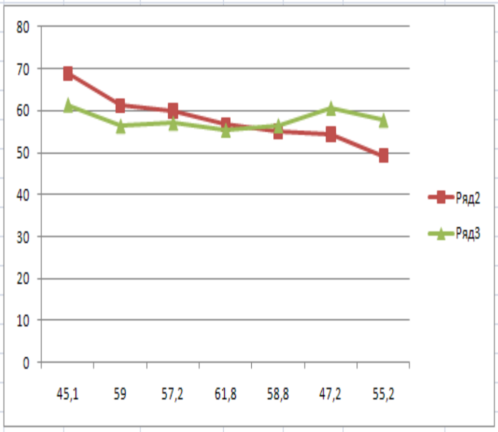
$$F = \frac{R^2(n-2)}{1-R^2}.$$

Величина F имеет распределение Фишера с $v_1 = 1, v_2 = n - 2.$

$$A = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right| \cdot 100\%$$

А) Результаты расчета для линейной модели

A	B	C	D	E	F	G	H	I	J	K	L	M	N
Район	x	y	xквадр	xy	varx	var y	yоценки	var yоценки	e	var e	A	к-нт корреляции	К-нт детермин
1	45,1	68,8	2034,01	3102,88	96,04	119,1216	61,34764	11,98490029	7,452364241	55,53773	10,83%	r	Rквадрат
2	59	61,2	3481	3610,8	16,81	10,98449	56,43736	2,097731922	4,762640616	22,68275	7,78%	-0,353257293	0,124790715
3	57,2	59,9	3271,84	3426,28	5,29	4,057347	57,07322	0,660142883	2,826777489	7,990671	4,72%		
4	61,8	56,7	3819,24	3504,06	47,61	1,405918	55,44824	5,94128595	1,251761037	1,566906	2,21%		
5	58,8	55	3457,44	3234	15,21	8,327347	56,50801	1,898066778	-1,508010842	2,274097	2,74%		
6	47,2	54,3	2227,84	2562,96	59,29	12,85735	60,6058	7,398841503	-6,305795443	39,76306	11,61%		
7	55,2	49,3	3047,04	2721,36	0,09	73,71449	57,77974	0,011231164	-8,479737098	71,90594	17,20%		
среднее	54,9	57,88571	3048,344	3166,049	34,33429	32,92408	57,88571	4,284600069	-8,12049E-15	28,81731	8,16%		
			Линейная модель										
b	-0,35326		Проверка выполнения условий дисперсионного анализа										
a	77,27954		33,10191										
Проверка		Значимость R квадрат по Фишеру											
-0,34593													
		F набл	0,712919										
		Значимость F		0,831912									
		Уравнение незначимо на уровне 5%											



Ошибка аппроксимации 8,16%, коэффициент корреляции -0,353257.

Б) Степенная модель.

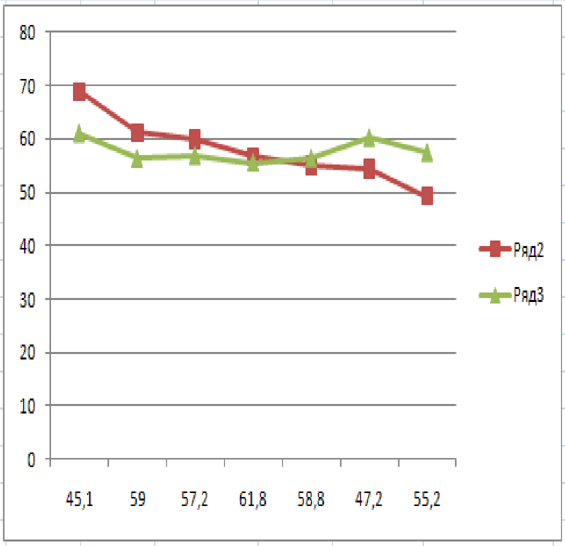
Логарифмируем левую и правую части степенной функции

$$lny = lna + blnx$$

Обозначим $lny = Y,\ lna = A,\ \ lnx = X,$ получаем уравнение $Y = A + bX,$ где коэффициенты **A** и **b** вычисляются методом наименьших квадратов.

Результаты расчета

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
Район	x	y	X=lnx	Y=lny	XY	Хквадрат	Yоценка	varX	varY	e	vare	varYоценка	A Ошибка аппрок	Yреальное
1	45,1	68,8	3,808882	4,231204	16,11616	14,50758	4,110534	0,036328	0,031524	0,120669379	0,014561	0,00323529	2,85%	60,9792941
2	59	61,2	4,077537	4,114147	16,77559	16,62631	4,030361	0,006093	0,003659	0,08378601	0,00702	0,00054259	2,04%	56,2812352
3	57,2	59,9	4,046554	4,092677	16,56124	16,3746	4,039607	0,002216	0,001523	0,053069089	0,002816	0,00019733	1,30%	56,8040381
4	61,8	56,7	4,123903	4,037774	16,65139	17,00658	4,016524	0,015481	0,000252	0,02124974	0,000452	0,00137866	0,53%	55,507851
5	58,8	55	4,074142	4,007333	16,32644	16,59863	4,031375	0,005574	0,002146	-0,02404132	0,000578	0,00049641	0,60%	56,3382953
6	47,2	54,3	3,854394	3,994524	15,39647	14,85635	4,096953	0,021051	0,003496	-0,102428367	0,010492	0,0018747	2,56%	60,1566861
7	55,2	49,3	4,010963	3,897924	15,63443	16,08782	4,050229	0,000132	0,024252	-0,152304531	0,023197	1,1738E-05	3,91%	57,4105803
Среднее			3,999482	4,053655	16,20882	16,00827		0,012411	0,00955	-2,53765E-16	0,008445	0,00110525	1,97%	
		Лианеризованная степенная модель				Выполнение условия дисперсионного анализа					0,00955	выполняется		
b	-0,29842													
a	5,247197		К-нт корреляции r			-0,34019								
			К-нт детерминации Rквадрат			0,115729								
			Значимость по Фишеру											
е в степ а	190,0328		будет незначимо так как R квадрат мал											



Для расчета реального значения y используется уравнение:

$$y = e^{5,247197} x^{-0,29842}.$$

Ошибка аппроксимации 1,97%, коэффициент корреляции -0,34019.

В) Показательная функция

Логарифмируем показательную функцию показательную $y = ab^x$, получаем

$$lny = lna + xlnb$$

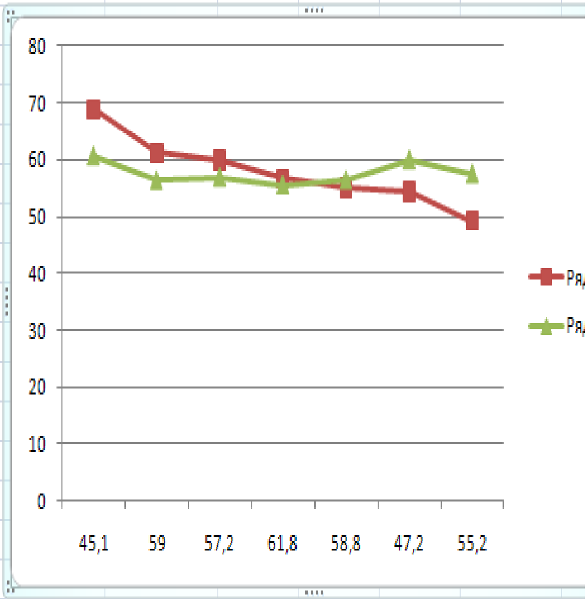
Обозначим $lny = Y,\ lna = A,\ \ lnb = B,$ получаем уравнение

$Y = A + Bx,$ где коэффициенты **A** и **B** вычисляются методом наименьших квадратов.

Результаты расчетов приведены в таблице

Район	x	y	lny=Y	xY	xквадрат	Yоценка	var Y	var Yоценка	e	var e	A	var x	y реальное
1	45,1	68,8	4,231204	190,8273	2034,01	4,106005	0,031524	0,002740514	0,125199089	0,015675	2,96%	96,04	60,7037003
2	59	61,2	4,114147	242,7347	3481	4,031753	0,003659	0,000479676	0,082393952	0,006789	2,00%	16,81	56,3596365
3	57,2	59,9	4,092677	234,1011	3271,84	4,041369	0,001523	0,000150951	0,051307976	0,002633	1,25%	5,29	56,9041645
4	61,8	56,7	4,037774	249,5344	3819,24	4,016796	0,000252	0,001358558	0,020978094	0,00044	0,52%	47,61	55,5229316
5	58,8	55	4,007333	235,6312	3457,44	4,032822	0,002146	0,000434019	-0,025488418	0,00065	0,64%	15,21	56,4198814
6	47,2	54,3	3,994524	188,5415	2227,84	4,094787	0,003496	0,001691848	-0,100262589	0,010053	2,51%	59,29	60,0265411
7	55,2	49,3	3,897924	215,1654	3047,04	4,052052	0,024252	2,56816E-06	-0,154128105	0,023755	3,95%	0,09	57,5153683
Среднее	54,9		4,053655	222,3622	3048,344	4,053655	0,00955	0,000979733	-5,07531E-16	0,008571	1,98%	34,33429	

		Линеаризация показательной модели				
			Проверка условий дисперсионного анализа			
B	-0,00534		0,00955	выполняется		
A	4,346921		К-нт корреляции		-0,32029	
			К-нт детерминации		0,102587	
			Ошибка аппроксимации		1,98%	
a	77,24028					
b	0,994672					



Для расчета у реального используется уравнение:

$$y = 77,24028 \cdot 0,994672^x$$

Учитываем, что $lna = A = 4,346921$, откуда $a = e^{4,34} = 77,24$

$lnb = B = -0,00534$, откуда $b = e^{-0,005} = 0,994672.$

Ошибка аппроксимации чуть больше, чем в предыдущей модели 1,98%, но коэффициент корреляции минимальный из всех моделей -0,32029.

Г) Гиперболическая функция.

Гиперболу $\boldsymbol{y = a + b \cdot \frac{1}{x}}$ линеаризуем с помощью замены $\boldsymbol{\frac{1}{x} = z}$.

получаем уравнение $\boldsymbol{y = a + bz}$, где коэффициенты $\boldsymbol{a}$ и $\boldsymbol{b}$ вычисляются методом наименьших квадратов.

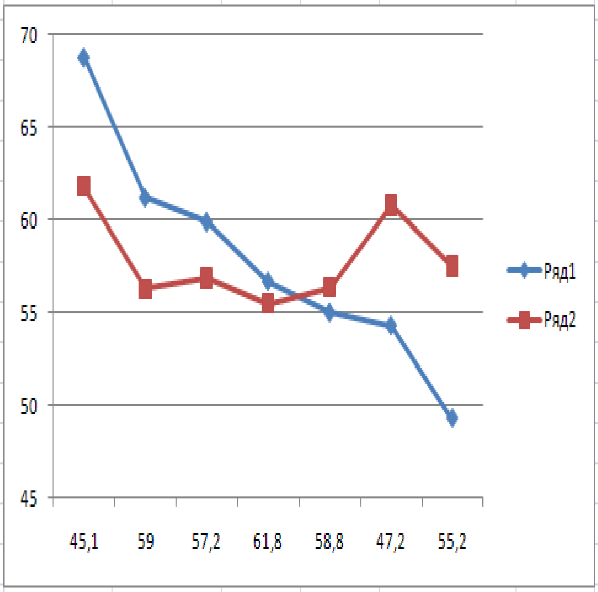
Для расчета у реального используется уравнение:

$$\boldsymbol{y = 38,43534 + 1054,67 \cdot \frac{1}{x}}$$

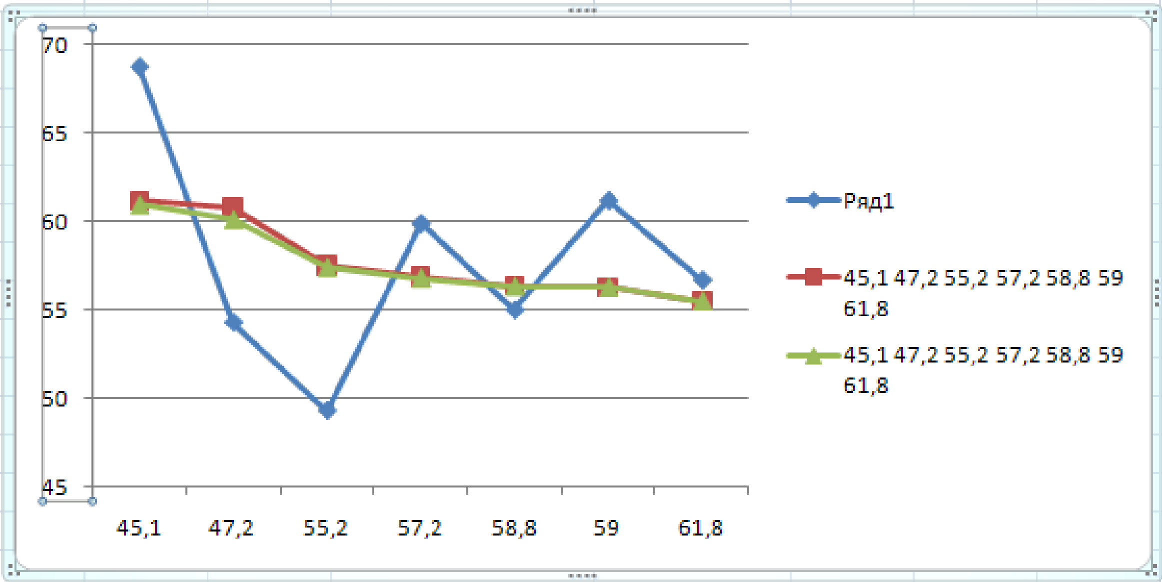
Ошибка аппроксимации 8,06%, но коэффициент корреляции максимальный из всех моделей 0,392246.

Район	x	y	z=1/x	zy	zквадрат	y оценка	var y	var yоценка	e	var e	A	varz	y реальн
1	45,1	68,8	0,022173	1,525499	0,000492	61,82048	119,1216	15,48236499	6,979522064	48,71373	10,14%	4,55405E-06	61,82048
2	59	61,2	0,016949	1,037288	0,000287	56,3111	10,98449	2,47941885	4,888902763	23,90137	7,99%		56,3111
3	57,2	59,9	0,017483	1,047203	0,000306	56,87362	4,057347	1,024332484	3,026378834	9,158969	5,05%		56,87362
4	61,8	56,7	0,016181	0,917476	0,000262	55,50119	1,405918	5,685945409	1,198807765	1,43714	2,11%		55,50119
5	58,8	55	0,017007	0,935374	0,000289	56,3719	8,327347	2,291636286	-1,371899143	1,882107	2,49%		56,3719
6	47,2	54,3	0,021186	1,150424	0,000449	60,78004	12,85735	8,377105924	-6,480037337	41,99088	11,93%		60,78004
7	55,2	49,3	0,018116	0,893116	0,000328	57,54167	73,71449	0,118363067	-8,241674946	67,92521	16,72%		57,54167
Среднее		57,88571	0,018442	1,07234	0,000345	57,88571	32,92408	5,065595287	-7,10543E-15	27,85849	8,06%		

b	1054,67	Линеаризация гиперболы		
a	38,43534	Проверка условий дисперсионного анализа	32,92408163	выполняется
		К-нт корреляции r	0,392246	
	-0,34593	К-нт детерминации R квадрат	0,153857	
		Fнаблюд	0,909166	
	1054,67	Fкритич	10,12796	незначим
		Уров. Знач. 0,05		



Район	x	y	y реал гипербола	y реал степен
1	45,1	68,8	61,2	60,98
6	47,2	54,3	60,78	60,16
7	55,2	49,3	57,54	57,41
3	57,2	59,9	56,87	56,8
5	58,8	55	56,37	56,34
2	59	61,2	56,3	56,28
4	61,8	56,7	55,5	55,5
Среднее				



Графики гиперболической и степенной зависимостей при фильтрации x по возрастанию

	К-нт корреляции	К-нт детерминации	Ошибка аппроксимации
Линейная функция	-0,353	0,124	8,16
Степенная функция	-0,34	0,115	1.97
Показательная функция	-0,32	0,102	1,98
Гиперболическая	0,392	0,153	8,06

В данном примере трудно выделить абсолютно лучшую модель по всем трем критериям.

6 ИСПОЛЬЗОВАНИЕ EXCEL ДЛЯ ПОСТРОЕНИЯ МОДЕЛИ ПАРНОЙ РЕГРЕССИИ

Задача 1

В качестве показателей, характеризующих результаты финансово-хозяйственной деятельности предприятия, могут быть выбраны:

- Объем выпуска продукции (Y, млн.руб.)
- Объем капиталовложений (X, млн.руб.)

Требуется:

1. Для характеристики зависимости объема выпуска продукции от объема капиталовложений построить следующие модели:
   - Линейную;
   - Экспоненциальную;
   - Гиперболическую.
2. Результаты расчетов отразить на графике.
3. Выбрать лучшую по качеству модель. Ответ обосновать.

Y	38	42	46	50	54	62	68
X	25	31	33	35	47	54	61

Решение

I. Для построения модели линейной парной регрессии y=a+b*x воспользуемся встроенной статистической функцией **ЛИНЕЙН.**

Для этого необходимо:

1) Создайте новый лист Excel (щелчок правой кнопкой на рабочем столе=>Создать=>Лист Microsoft Excel)
2) Введите исходные данные в соответствии с вариантом

	A	B
1	Y	X
2	38	25
3	42	31
4	46	33
5	50	35
6	54	47
7	62	54
8	68	61
9		

Рисунок 1- Ввод исходных данных

3) Выделите область пустых ячеек 5×2 (5 строк, 2 столбца)
4) Выберите в главном меню **Вставка =>Функция**
5) В меню **Категория** выберите **Статистические** в окне выбора функций – функцию **ЛИНЕЙН,** нажмите **ОК** (Рисунок 2)

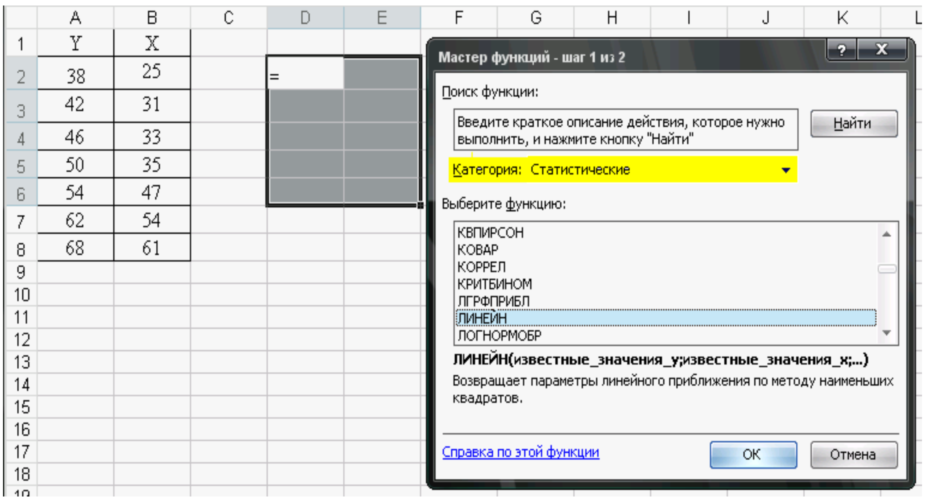


Рисунок 2- Окно выбора линейной функции

6) Заполните аргументы функции
   - Значения Y- диапазон, содержащий значения результативного признака (нажать кнопку в правой части соответствующей ячейки и выделить нужный диапазон A2:A8)
   - Значения X – диапазон, содержащий данные факторов независимого признака (аналогично B2:B8)
   - Константа=1
   - Статистика=1 (см.рисунок 3)

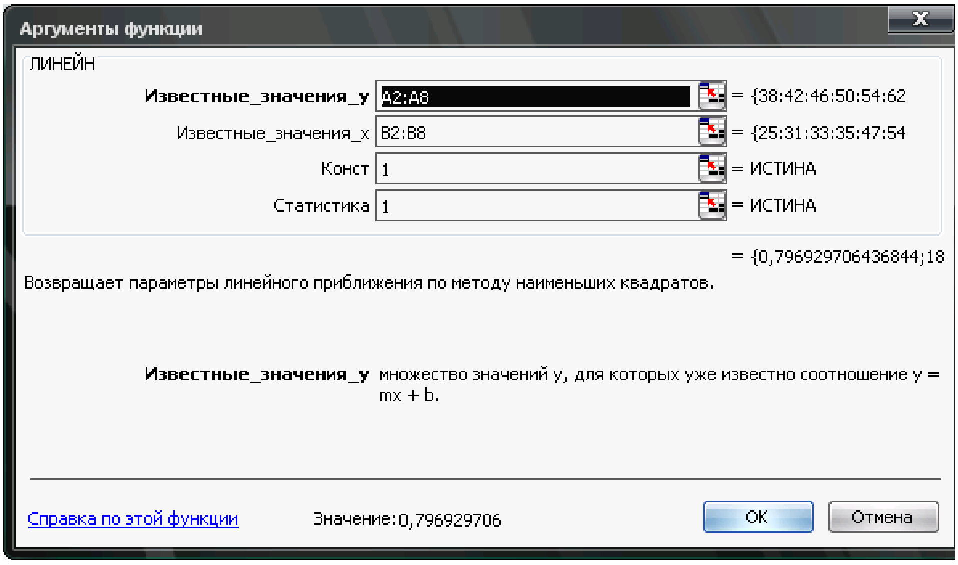
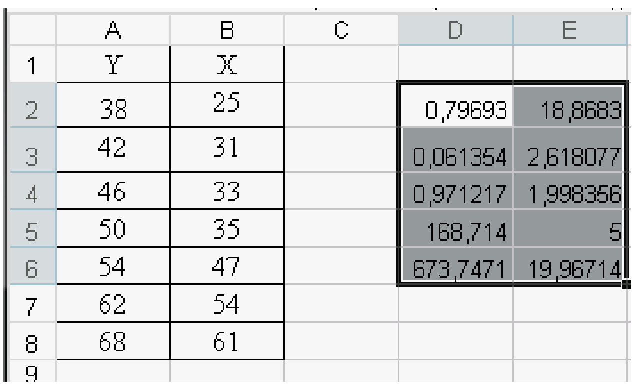


Рисунок 3- Окно аргументов функции “Линейн”

7) В левой верхней ячейке появится первый элемент итоговой таблицы. Чтобы раскрыть всю таблицу нажмите F2, а затем комбинацию клавиш <CTRL>+<SHIFT>+<ENTER>.

Дополнительная регрессионная статистика выводится в следующем порядке

Значение коэффициента b	Значение коэффициента a
Среднеквадратическое отклонение b	Среднеквадратическое отклонение a
Коэффициент детерминации R^2	Среднеквадратическое отклонение y
F-статистика	Число степеней свободы
Регрессионная сумма квадратов	Остаточная сумма квадратов



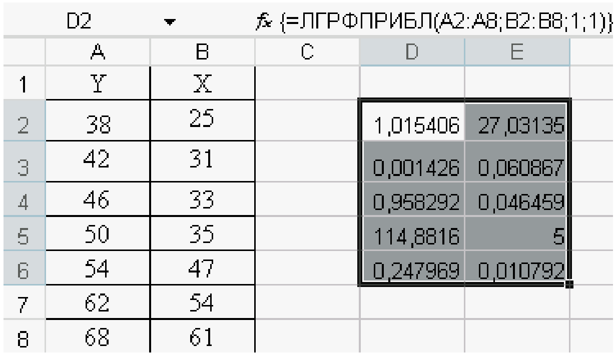
	A	B	C	D	E
1	Y	X			
2	38	25		0,79693	18,8683
3	42	31		0,061354	2,618077
4	46	33		0,971217	1,998356
5	50	35		168,714	5
6	54	47		673,7471	19,96714
7	62	54			
8	68	61			
9					

Рисунок 4- Результаты построения линейной модели

В соответствии с полученными результатами можно записать уравнение линейной регрессии:

$$y=18{,}8683+0{,}79693*x$$

II. Для вычисления параметров экспоненциальной кривой $y=\alpha*\beta^{x}$ воспользуйтесь встроенной статистической функцией **ЛГРФПРИБЛ.** Порядок вычислений аналогичен применению предыдущей функции. Результаты показаны на рисунке 5.


D2 fx {=ЛГРФПРИБЛ(A2:A8;B2:B8;1;1)}

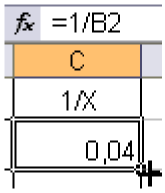
	A	B	C	D	E
1	Y	X			
2	38	25		1,015406	27,03135
3	42	31		0,001426	0,060867
4	46	33		0,958292	0,046459
5	50	35		114,8816	5
6	54	47		0,247969	0,010792
7	62	54			
8	68	61			

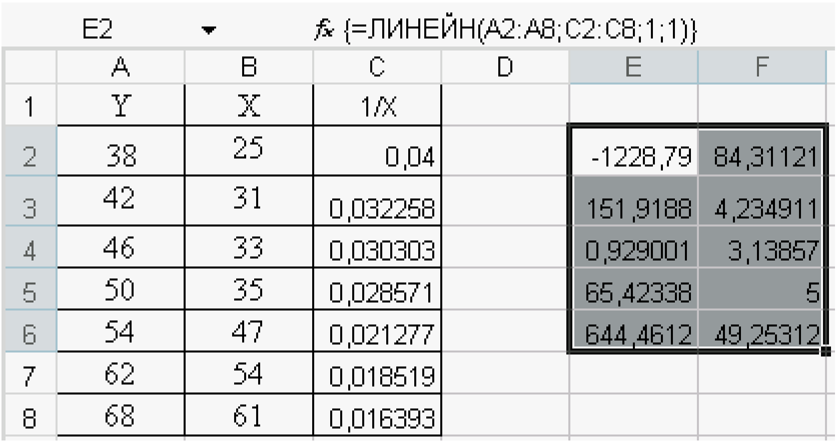
Рисунок 5- Результаты построения экспоненциальной модели

В соответствии с полученными результатами можно записать уравнение:

$$y=27{,}03135*e^{1{,}0154\,x}$$

Для построения гиперболической функции воспользуйтесь функцией ЛИНЕЙН, взяв в качестве факторного признака величину 1/X. Для этого в столбце C в ячейке C1 записать название столбца “1/X”, а в ячейку C2 ввести формулу =1/B2, после чего, выделив ячейку C2, навести курсор мыши в правый нижний угол ячейки (появится крестик) нажать левую кнопку и, не отпуская, растянуть выделенную ячейку мышкой до ячейки C8 включительно. Затем воспользоваться уже известной функцией Линейн, взяв в качестве параметра Известные_значения_x диапазон C2:C8. Результаты представлены на рисунке 6.





E2 f_x {=ЛИНЕЙН(A2:A8;C2:C8;1;1)}

	A	B	C	D	E	F
1	Y	X	1/X			
2	38	25	0,04		-1228,79	84,31121
3	42	31	0,032258		151,9188	4,234911
4	46	33	0,030303		0,929001	3,13857
5	50	35	0,028571		65,42338	5
6	54	47	0,021277		644,4612	49,25312
7	62	54	0,018519			
8	68	61	0,016393			

Рисунок 6- Результат построения гиперболической функции

Запишем уравнение гиперболической функции:

$$y=84,31121-1228,79/x$$

III. Отразите полученные результаты расчетов на графике.

Для этого выполните следующую последовательность действий:

1. Выберите в меню **Вставка->Диаграмма, Тип-Точечная.** (рисунок 7)

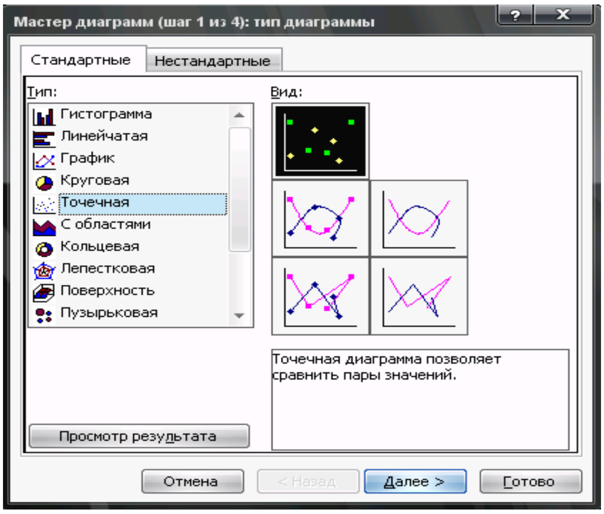


Рисунок 7 - Выбор типа диаграммы

2. Введите диапазон входных данных (рисунок 8)

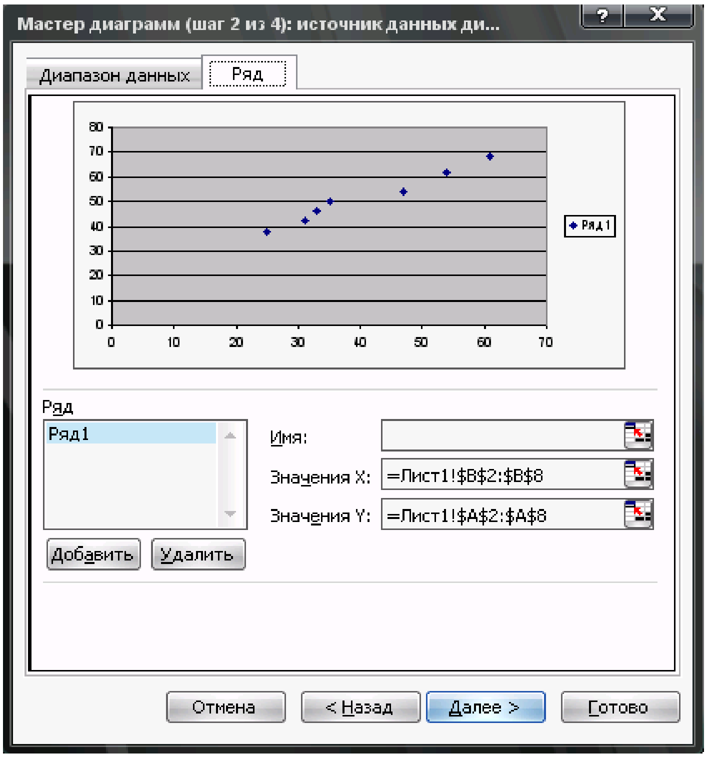


Рисунок 8- Выбор источника данных для диаграммы

3. Оформите вид полученного графика (подписи осей, название графика, и т.д.)
4. Выберите место отображения графика (в имеющемся).
5. Добавьте линию тренда:

- выделите полученную диаграмму и выберите на панели управления **Диаграмма->Добавить линию тренда ;**
- выбираем **Тип-Линейная;**
- в закладке **Параметры** ставим галочки:

∨ **Показывать уравнение на диаграмме**

∨ **Поместить на диаграмме величину достоверности аппроксимации (R^2).**

Полученная диаграмма представлена на рисунке 9.

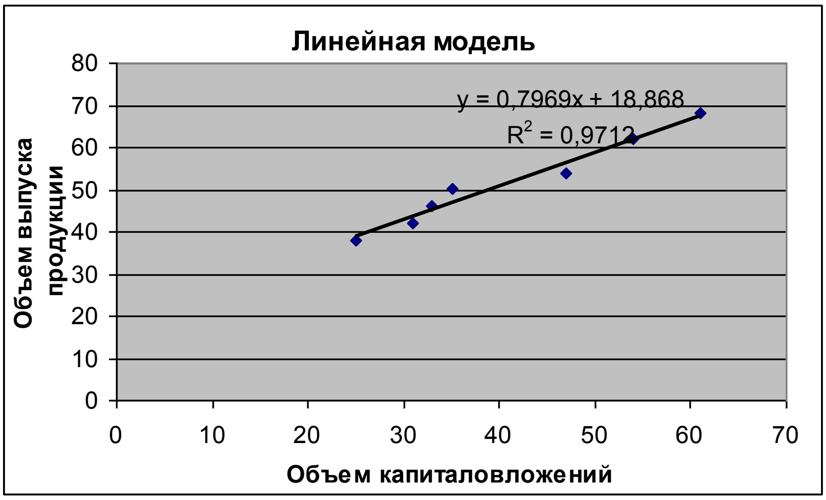


Рисунок 9- Графическое представление линейной модели

V. Осуществляя аналогичные действия, но, указав в **Параметрах линии тренда Тип-Экспоненциальная**, получим графическое представление экспоненциальной модели (рисунок 10).

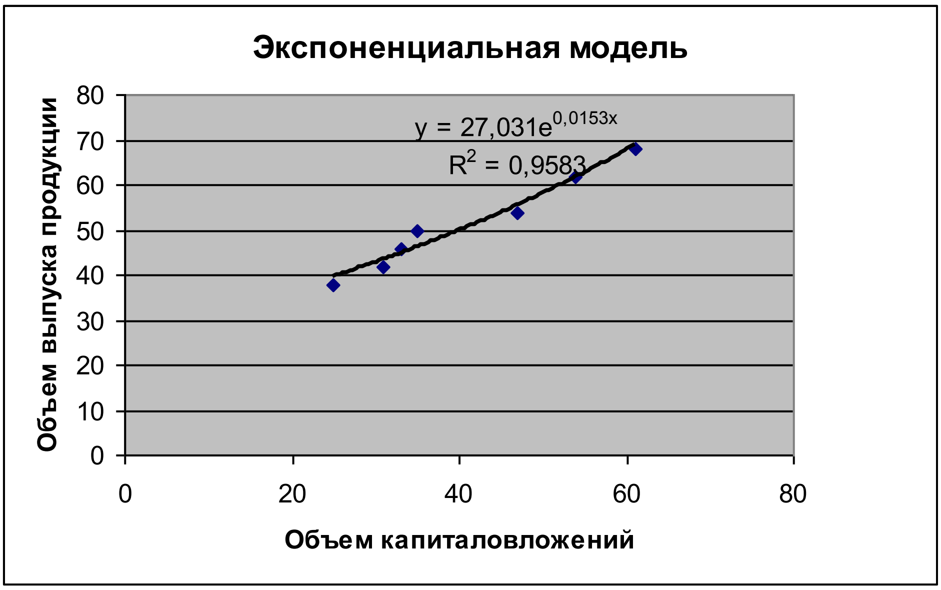


Рисунок 10- Графическое представление экспоненциальной модели

VI. Для гиперболической модели порядок действий аналогичен построению линейной модели, но входной диапазон для X-столбца необходимо указать со значениями 1/X (C2:C8).

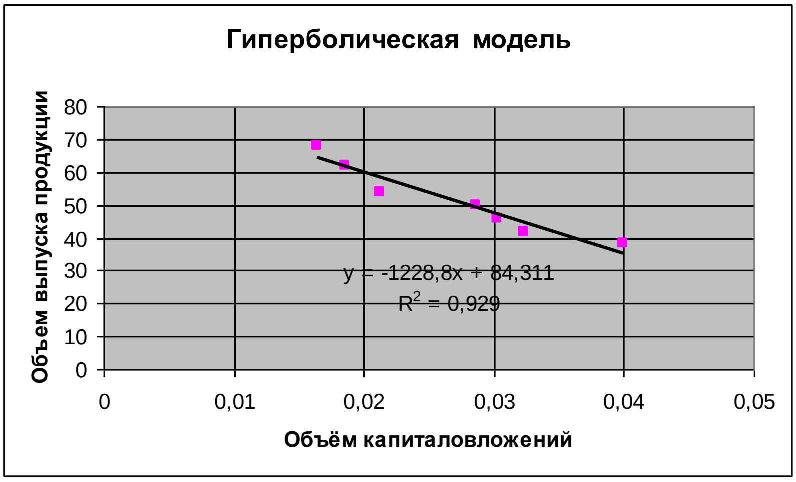


Рисунок 11- Графическое представление гиперболической модели

VII. Сравним по качеству полученные модели. Для этого результаты вычислений представим в сводной таблице.

Таблица 1 - Сводная таблица результатов вычислений

Модель / Показатель	Линейная	Экспоненциальная	Гиперболическая
Коэффициент детерминации R^2	0,971217	0,958292	0,929001
Значение F-критерия Фишера	168,714	114,8816	65,42338

Вывод: Лучшей по качеству моделью признается модель, обладающая большим значением коэффициента детерминации. Коэффициент детерминации R^2 показывает долю вариации результативного признака под воздействием факторных признаков, введенных в модель регрессии.

Для линейной модели 97,12 % вариации объема выпуска продукции определяется вариацией объема капиталовложений, что является максимальным значением коэффициента детерминации из всех 3-х моделей. Следовательно, в качестве лучшей модели следует выбрать линейную модель.

Проверка статистической значимости уравнения регрессии осуществляется по F-критерию Фишера, для этого находим $F_{Табл.}$= 5,987378 с помощью функции FРАСПОБР (для a = 0,05, k1 = 1, k2 = 6). $F_{Табл.} < F_{Расч.}$

для каждой из моделей, следовательно, полученные уравнения регрессий следует признать адекватными, а модели значимыми, при этом лучшей моделью по значению коэффициента детерминации признается модель линейной регрессии.

Задания для самостоятельной работы к задаче 1

Вариант	Признаки	Наблюдения						
1	Y	36	38	46	44	48	42	40
	X	60	68	64	72	78	74	70
2	Y	26	28	36	34	38	44	42
	X	40	39	43	46	50	53	57
3	Y	176	170	156	172	162	160	166
	X	150	154	146	134	132	126	133
4	Y	60	68	74	76	84	86	92
	X	50	54	60	62	70	66	74
5	Y	32	40	44	38	50	56	50
	X	60	68	80	76	74	87	96
6	Y	36	38	46	44	48	42	40
	X	70	78	74	82	88	84	80
7	Y	152	148	146	134	130	136	134
	X	86	94	100	96	93	104	122
8	Y	64	56	52	48	50	46	38
	X	64	68	82	76	84	96	100
9	Y	50	54	60	62	70	74	81
	X	60	68	74	82	88	94	100
10	Y	40	44	48	52	56	64	70
	X	22	28	30	32	44	51	58

7 МОДЕЛЬ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

7.1 Методика построения модели множестьвенной регрессии

Рассмотрим методику на примере.

Изучается 20 предприятий региона: y – выработка продукции на одного работника (тыс. руб.), $x_1 -$ ввод в действие новых основных фондов (% от стоимости фондов на конец года), $x_2 -$ процент рабочих высокой квалификации к общей численности рабочих.

Номер предприятия	y	x_1	x_2	Номер предприятия	y	x_1	x_2
1.	7,0	3,9	10,0	11.	9,0	6,0	21,0
2.	7,0	3,9	14,0	12.	11,0	6,4	22,0
3.	7,0	3,7	15,0	13.	9,0	6,8	22,0
4.	7,0	4,0	16,0	14.	11,0	7,2	25,0
5.	7,0	3,8	17,0	15.	12,0	8,0	28,0
6.	7,0	4,8	19,0	16.	12,0	8,2	29,0
7.	8,0	5,4	19,0	17.	12,0	8,1	30,0
8.	8,0	4,4	20,0	18.	12,0	8,5	31,0
9.	8,0	5,3	20,0	19.	14,0	9,6	32,0
10.	10,0	6,8	20,0	20.	14,0	9,0	36,0

Задание:

1. Построить линейную модель множественной регрессии. Записать стандартизированное уравнение множественной регрессии. На основе стандартизированных коэффициентов регрессии и средних коэффициентов эластичности провести ранжирование факторов по степени их влияния на результат.

2. Найти коэффициенты парной, частной и множественной корреляции. Провести их анализ.
3. Найти скорректированный коэффициент множественной детерминации. Сравнить его с нескорректированным (общим) коэффициентом детерминации.
4. С помощью F – критерия Фишера оценить статистическую надежность уравнения регрессии и коэффициента детерминации $\boldsymbol{R^2}$.
5. С помощью t – критерия оценить статистическую значимость коэффициентов чистой регрессии.
6. С помощью частных F – критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора $\boldsymbol{x_1}$ после $\boldsymbol{x_2}$ и фактора $\boldsymbol{x_2}$ после $\boldsymbol{x_1}$.
7. Составить уравнение линейной парной регрессии, оставив лишь один значащий фактор. Найти ошибку аппроксимации для парной регрессии. Сравнить ее с ошибкой аппроксимации для множественной регрессии. Построить графики $\boldsymbol{y}$ и $\boldsymbol{\hat{y}}$ для парной регрессии на одних осях.

Решение

Находим вариации (дисперсии) признаков:

$$\boldsymbol{var\, y = \frac{1}{n}\sum(y-\overline{y})^2 = \overline{y^2} - \overline{y}^2 = 97,9 - 9,6^2 = 5,74}$$

$$\boldsymbol{var\, x_1 = \frac{1}{n}\sum(x_1-\overline{x_1})^2 = \overline{x_1{}^2} - \overline{x_1}^2 = 41,887 - 6,19^2 = 5,74}$$

$$\boldsymbol{var\, x_1 = \frac{1}{n}\sum(x_1-\overline{x_1})^2 = \overline{x_1{}^2} - \overline{x_1}^2 = 41,887 - 6,19^2 = 3,571}$$

$$var\ x_2 = \frac{1}{n}\sum(x_2 - \overline{x_2})^2 = \overline{x_2{}^2} - \overline{x_2}^2 = 541{,}4 - 22{,}3^2$$
$$= 44{,}11.$$

1. Найдем коэффициенты уравнения чистой регрессии по готовым формулам:

$$b_1 = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{1 - r_{x1x2}^2} \cdot \sqrt{\frac{vary}{varx_1}};$$

$$b_2 = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{1 - r_{x1x2}^2} \cdot \sqrt{\frac{vary}{varx_2}}$$

$$a = \overline{y} - b_1 \cdot \overline{x_1} - b_2 \cdot \overline{x_2}.$$

Для расчета коэффициентов требуется предварительно рассчитать парные коэффициенты корреляции:

$$r_{yx1} = \frac{cov(y, x_1)}{\sqrt{vary \cdot varx_1}} = \frac{\overline{y \cdot x_1} - \overline{y} \cdot \overline{x_1}}{\sqrt{vary \cdot varx_1}};$$

$$r_{yx2} = \frac{cov(y, x_2)}{\sqrt{vary \cdot varx_2}} = \frac{\overline{y \cdot x_2} - \overline{y} \cdot \overline{x_2}}{\sqrt{vary \cdot varx_2}};$$

$$r_{x1x2} = \frac{\overline{x_1 \cdot x_2} - \overline{x_1} \cdot \overline{x_2}}{\sqrt{varx_1 \cdot varx_2}}.$$

Получаем:

$$r_{yx1} = \frac{63{,}815 - 9{,}6 \cdot 6{,}19}{\sqrt{5{,}74 \cdot 3{,}5709}} = 0{,}97;$$

$$r_{yx2} = \frac{229,05 - 9,6 \cdot 22,3}{\sqrt{5,74 \cdot 44,11}} = 0,941;$$

$$r_{x1x2} = \frac{149,87 - 6,19 \cdot 22,3}{\sqrt{6,19 \cdot 22,3}} = 0,943.$$

Теперь можно посчитать коэффициенты уравнения чистой регрессии:

$$b_1 = \frac{0,97 - 0,941 \cdot 0.943}{1 - 0,943 \cdot 0.943} \cdot \sqrt{\frac{5,74}{3,57}} = 0,946;$$

$$b_2 = \frac{0,941 - 0,97 \cdot 0,943}{1 - 0,943 \cdot 0.943} \cdot \sqrt{\frac{5.74}{44,11}} = 0,0856;$$

$$a = 9,6 - 0,946 \cdot 6,19 - 0,0856 \cdot 22,3 = 1,835.$$

Получено следующее уравнение множественной регрессии:

$$\hat{y} = 1,835 + 0,946 \cdot x_1 + 0,0856 \cdot x_2.$$

Уравнение регрессии показывает, что при увеличении ввода в действие основных фондов на 1% (при неизменном уровне удельного веса рабочих высокой квалификации) выработка продукции на одного рабочего увеличится в среднем на 0,946 тыс. руб., а при увеличении удельного веса рабочих высокой квалификации в общей численности рабочих на 1% (при неизменном уровне ввода в действие новых основных фондов) выработка продукции на одного рабочего увеличивается в среднем на 0,086 тыс. руб.

Найдем $\boldsymbol{vare}$ (остаточную дисперсию) и среднюю ошибку аппроксимации:

$$var\ e = \frac{1}{n}\sum(y - \hat{y})^2 = \overline{e^2} - \bar{e}^2 = 0,305,$$

$$A = \frac{1}{n}\sum\left|\frac{y - \hat{y}}{y}\right| \cdot 100\% = 4,77\%.$$

Качество модели, исходя из средней ошибки аппроксимации, признается хорошим, т.к. средняя ошибка аппроксимации меньше 10%.

7.2 Стандартизованное уравнение регерессии

Стандартизированное уравнение регрессии имеет вид:

$$\hat{t}_y = \gamma_1 \cdot t_{x1} + \gamma_2 \cdot t_{x2}.$$

Стандартизированным уравнением регрессии называется уравнение, в котором все переменные (x и y) заданы как центрированные и нормированные. Центрированной называется переменная, у которой математическое ожидание (для выборки – среднее значение) равно нулю. Нормированной называется переменная, у которой среднее квадратическое отклонение (для выборки – корень из вариации) равно единице. В связи с этим, переменные вводятся следующим образом:

$$t_y = \frac{y - \bar{y}}{\sqrt{vary}}; \quad t_{x_i} = \frac{x_i - \bar{x_i}}{\sqrt{varx_i}}.$$

Т.к. все переменные заданы как центрированные и нормированные, стандартизированные коэффициенты регрессии γ_i можно сравнивать между собой. Это позволяет ранжировать факторы по силе воздействия на результат: чем больше величина коэффициента γ_i, тем сильнее влияние данного признака. В этом основное достоинство стандартизированного уравнения регрессии. К стандартизированному уравнению применим МНК, коэффициенты γ_i рассчитываются по следующим формулам:

$$\gamma_1 = b_1 \cdot \sqrt{\frac{varx_1}{vary}}; \qquad \gamma_2 = b_2 \cdot \sqrt{\frac{varx_2}{vary}}.$$

Подставив, получаем:

$$\boldsymbol{\gamma_1 = 0,946 \cdot \sqrt{\frac{3,57}{5,74}} = 0,746};$$

$$\boldsymbol{\gamma_2 = 0,0856 \cdot \sqrt{\frac{44,11}{5,74}} = 0,237}.$$

Стандартизированное уравнение имеет вид:

$$\boldsymbol{\hat{t}_y = 0,745 \cdot t_{x1} + 0,237 \cdot t_{x2}}$$

Т.к. стандартизированные коэффициенты регрессии можно сравнивать между собой, то можно сделать вывод, что ввод в действие новых основных фондов оказывает большее влияние на выработку продукции (более, чем в три раза), чем удельный вес рабочих высокой квалификации по отношению к общему числу рабочих.

Сравнивать влияние факторов на результат можно также при помощи средних коэффициентов эластичности, которые находятся для каждой переменной $\boldsymbol{x}$:

$$\overline{\text{Э}_1} = \boldsymbol{b_1} \cdot \frac{\overline{\boldsymbol{x_1}}}{\overline{\boldsymbol{y}}}; \qquad \overline{\text{Э}_2} = \boldsymbol{b_2} \cdot \frac{\overline{\boldsymbol{x_2}}}{\overline{\boldsymbol{y}}}$$

Находим:

$$\overline{\text{Э}_1} = \boldsymbol{0,946 \cdot \frac{6,19}{9,6} = 0,61};$$

$$\overline{\text{Э}_2} = \boldsymbol{0,0856 \cdot \frac{22.3}{9.6} = 0,2}.$$

Вывод: увеличение только основных фондов на 1% от их среднего значения увеличивает в среднем выработку продукции на 0,61%; увеличение только удельного веса рабочих высокой квалификации на 1% от их среднего значения увеличивает в среднем выработку продукции на 0,2%. Т.е. подтверждается большее влияние на результат фактора $\boldsymbol{x_1}$.

7.3 Коэффициенты корреляции

Коэффициенты парной корреляции найдены в п.1:

$$\boldsymbol{r_{yx1}} = 0{,}97;\quad \boldsymbol{r_{yx2} = 0,941};\quad \boldsymbol{r_{x1x2} = 0,943}.$$

Величина этих коэффициентов указывает на сильную связь каждого фактора с результатом, а также на высокую связь между факторами. Силу связи, после нахождения коэффициента корреляции, принято оценивать по шкале Шедока:

- если $|r| < 0{,}2$ – связи нет;
- если $0{,}2 \le |r| \le 0{,}5$ - связь слабая;
- если $0{,}5 \le |r| \le 0{,}75$ - связь средняя;
- если $0{,}75 \le |r| < 0{,}95$ - связь тесная;
- если $0{,}95 \le |r| < 1$ - связь очень тесная.

При $\boldsymbol{r_{x1x2} = 0,943}$ можно считать, что между факторами связь близкая к линейной, в этом случае рекомендуется один из факторов исключить из модели.

Частные коэффициенты корреляции характеризуют тесноту связи между результатом и соответствующим фактором при элиминировании (устранении влияния) других факторов, включенных в регрессионную модель.

При двух факторах частные коэффициенты корреляции рассчитываются следующим образом:

$$\boldsymbol{r_{yx1\cdot x2} = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{\sqrt{\left(1 - r_{yx2}^2\right) \cdot \left(1 - r_{x1x2}^2\right)}}};$$

$$r_{yx2\cdot x1} = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{\sqrt{\left(1 - r_{yx1}^2\right) \cdot \left(1 - r_{x1x2}^2\right)}}$$

Подставляя значения, получаем:

$$r_{yx1\cdot x2} = \frac{0,97 - 0,941 \cdot 0,943}{\sqrt{(1 - 0,941^2) \cdot (1 - 0,943^2)}} = 0,734,$$

$$r_{yx2\cdot x1} = \frac{0,941 - 0,97 \cdot 0,943}{\sqrt{(1 - 0,97^2) \cdot (1 - 0,943^2)}} = 0,325.$$

При сравнении коэффициентов парной и частной корреляции можно сделать вывод: из-за высокой факторной зависимости (т.е. тесной связи между факторами x_1 и x_2) коэффициенты парной корреляции дают завышенные оценки тесноты связи. Именно по этой причине при наличии тесной связи между факторами рекомендуется исключать из уравнения тот фактор, у которого теснота парной зависимости меньше, чем теснота межфакторной связи.

Коэффициент множественной корреляции можно определить тремя способами.

1 способ – через матрицы парных коэффициентов корреляции:

$$R_{yx1x2} = \sqrt{1 - \frac{\Delta r}{\Delta r_{11}}},$$

где $\Delta r -$ определитель матрицы парных коэффициентов корреляции

$$\Delta r = \begin{vmatrix} 1 & r_{yx1} & r_{yx2} \\ r_{yx1} & 1 & r_{x1x2} \\ r_{yx2} & r_{x1x2} & 1 \end{vmatrix}$$

$\Delta r_{11} -$ определитель матрицы межфакторной корреляции

$$\Delta \boldsymbol{r}_{11} = \begin{vmatrix} \mathbf{1} & \boldsymbol{r}_{x1x2} \\ \boldsymbol{r}_{x1x2} & \mathbf{1} \end{vmatrix}.$$

Подставляя данные, находим:

$$\Delta \boldsymbol{r} = \begin{vmatrix} \mathbf{1} & \mathbf{0,97} & \mathbf{0,941} \\ \mathbf{0,97} & \mathbf{1} & \mathbf{0,943} \\ \mathbf{0,941} & \mathbf{0,943} & \mathbf{1} \end{vmatrix} = \mathbf{0,0058}$$

$$\Delta \boldsymbol{r}_{11} = \begin{vmatrix} \mathbf{1} & \mathbf{0,943} \\ \mathbf{0,943} & \mathbf{1} \end{vmatrix} = \mathbf{0,1108}.$$

Коэффициент множественной корреляции равен:

$$\boldsymbol{R}_{yx1x2} = \sqrt{\mathbf{1} - \frac{\mathbf{0,0058}}{\mathbf{0,1108}}} = \mathbf{0,973}.$$

2 способ – расчет по формуле

$$\boldsymbol{R}_{yx1x2} = \sqrt{\mathbf{1} - \frac{\boldsymbol{vare}}{\boldsymbol{vary}}} = \sqrt{\mathbf{1} - \frac{\mathbf{0,305}}{\mathbf{5,74}}} = \mathbf{0,973}.$$

3 способ – расчет по формуле с использованием коэффициентов стандартизированного уравнения регрессии

$$\boldsymbol{R}_{yx1x2} = \sqrt{\sum \boldsymbol{\gamma}_{\boldsymbol{i}} \cdot \boldsymbol{r}_{\boldsymbol{yxi}}}$$

Получаем

$$\boldsymbol{R}_{yx1x2} = \sqrt{\mathbf{0,746 \cdot 0,97 + 0,237 \cdot 0,941}} = \mathbf{0,973}.$$

Коэффициент множественной регрессии указывает на очень тесную связь всего набора факторов с результатом.

7.4 Расчет коэффициента множественной детерминации

Нескорректированный коэффициент множественной детерминации $\boldsymbol{R^2_{yx1x2}}$ оценивает долю дисперсии результата, которая объясняется регрессионным уравнением в общей доле дисперсии результата.

$$\boldsymbol{R^2_{yx1x2} = 0,973^2 = 0,947}.$$

Здесь эта доля составляет ***94,7%*** и указывает на то, что результат и факторы имеют тесную связь.

Скорректированный коэффициент множественной регрессии дает оценку тесноты связи, которая не зависит от числа факторов в модели, поэтому позволяет сравнивать разные модели с различным числом факторов.

Скорректированный коэффициент множественной регрессии рассчитывается с учетом числа степеней свободы вариации (дисперсии) $\boldsymbol{y}$ и вариации (дисперсии остатка) ошибки $\boldsymbol{e}$:

$$\boldsymbol{\widehat{R^2} = 1 - \left(1 - R^2\right) \cdot \frac{n-1}{n-k-1}}.$$

Подставляя, получаем:

$$\boldsymbol{\widehat{R^2} = 1 - (1 - 0,947) \cdot \frac{20-1}{20-2-1} = 0,941}.$$

Оба коэффициента детерминации (скорректированный и нескорректированный) указывают на высокую зависимость результата $\boldsymbol{y}$ от факторов $\boldsymbol{x_1}$ и $\boldsymbol{x_2}$.

Оценку надежности уравнения регрессии в целом и коэффициента множественной корреляции (показателя тесноты связи) проводят с использованием $\boldsymbol{F}$ - критерия Фишера.

$$\boldsymbol{F = \frac{R^2}{1 - R^2} \cdot \frac{n-k-1}{k}}.$$

Рассчитаем фактическое значение критерия Фишера (наблюдаемое значение):

$$F_{\text{набл}} = \frac{0,973^2}{1-0,973^2} \cdot \frac{20-2-1}{2} = 151,65.$$

Найдем $F_{\text{крит}}$, воспользовавшись функцией

$$F_{\text{крит}} = F\text{РАСПОБР}(\alpha, v_1, v_2),$$

где, для уравнения множественной регрессии, имеем:

$$v_1 = k = 2, \quad v_2 = n - k - 1 = 17.$$

Задав уровень значимости $\alpha = 0,05$, получаем $F_{\text{крит}} = 3,59$.

$$F_{\text{набл}} > F_{\text{крит}}.$$

Следовательно, полученное значение коэффициента множественной корреляции не случайно, оно сформировалось под влиянием существенных факторов. Т.о. подтверждается статистическая значимость всего уравнения регрессии в целом и, также, статистическая значимость коэффициента множественной корреляции.

7.5 Оценка значимости коэффицентов модели

Оценим статистическую значимость параметров чистой регрессии с помощью t – критерия Стьюдента. Для этого рассчитаем стандартные ошибки коэффициентов регрессии. Для множественной регрессии формулы для расчета стандартных ошибок коэффициентов регрессии b_1 и b_2 имеют вид:

$$S_{b1} = \sqrt{\frac{vary \cdot (1 - R^2_{yx1x2})}{varx_1 \cdot \left(1 - r^2_{x1x2}\right) \cdot (n-k-1)}};$$

$$S_{b2} = \sqrt{\frac{vary \cdot (1 - R^2_{yx1x2})}{varx_2 \cdot \left(1 - r^2_{x1x2}\right) \cdot (n - k - 1)}}.$$

Подставив значения, получим:

$$S_{b1} = \sqrt{\frac{5,74 \cdot (1 - 0,973^2)}{3,571 \cdot (1 - 0,943^2) \cdot (20 - 2 - 1)}} = 0,2125$$

$$S_{b2} = \sqrt{\frac{5,74 \cdot (1 - 0,973^2)}{44,11 \cdot (1 - 0,943^2) \cdot (20 - 2 - 1)}} = 0,0605$$

Находим фактическое (наблюдаемое) значение t – критерия для каждого коэффициента регрессии. Используем следующие формулы:

$$t_{b1} = \frac{b_1}{S_{b1}}; \qquad t_{b2} = \frac{b_2}{S_{b2}} \ .$$

Получаем:

$$t_{b1} = \frac{0,946}{0,2125} = 4,45; \qquad t_{b2} = \frac{0,0856}{0,0605} = 1,41.$$

Для нахождения критического значения критерия пользуемся функцией

$$t_{\text{крит}} = \text{СТЬЮДРАСПОБР}(\alpha,\ v)$$

где $v = n - k - 1 = 17,\ \alpha = 0,05.$

$$t_{\text{крит}} = 2,11.$$

Т.о. признается статистическая значимость коэффициента b_1, т.к. $t_{b1} > t_{\text{крит}}$. Коэффициент b_2 имеет случайную природу формирования, он незначим, т.к. $t_{b2} < t_{\text{крит}}$.

Найдем доверительные интервалы для коэффициентов чистой регрессии. Используем формулы:

$$\boldsymbol{b_1 - S_{b1} \cdot t_{\text{крит}} \leq \beta_1 \leq b_1 + S_{b1} \cdot t_{\text{крит}}};$$

$$\boldsymbol{b_2 - S_{b2} \cdot t_{\text{крит}} \leq \beta_2 \leq b_2 + S_{b2} \cdot t_{\text{крит}}}.$$

Подставив, получаем:

$$\boldsymbol{0,497 \leq \beta_1 \leq 1,394};$$

$$\boldsymbol{-0,042 \leq \beta_2 \leq 0,2132}.$$

7.6 Оценка целесообразности включения факторов в модель регрессии

С помощью частных $\boldsymbol{F}$ – критериев Фишера оценим целесообразность включения в уравнение множественной регрессии фактора $\boldsymbol{x_1}$ после того, как в уравнение уже включен фактор $\boldsymbol{x_2}$**;** и целесообразность включения фактора $\boldsymbol{x_2}$, после того, как в уравнение уже включен фактор $\boldsymbol{x_1}$**.**

Теоретическая справка. Частный $\boldsymbol{F}$ – критерий Фишера построен на сравнении прироста вариации $\boldsymbol{\hat{y}}$ (факторной дисперсии), который вызван включением в модель дополнительного фактора и вариации ошибки (остаточной дисперсии) на одну степень свободы. В общем виде для фактора $\boldsymbol{x_i}$ частный $\boldsymbol{F}$ – критерий Фишера определяется как:

$$\boldsymbol{F_{x_i} = \frac{R^2_{yx_1x_2\ldots x_k} - R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k}}{1 - R^2_{yx_1x_2\ldots x_k}} \cdot \frac{n-k-1}{1}},$$

где $\boldsymbol{R^2_{yx_1x_2\ldots x_k}}$- коэффициент множественной детерминации для модели с полным набором факторов, $\boldsymbol{R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_k}}$-

коэффициент множественной детерминации для модели, в которую не включен фактор $\boldsymbol{x_i}$, $\boldsymbol{k}-$ число параметров в модели без свободного члена.

Фактическое (наблюдаемое) значение критерия Фишера сравнивается с критическим

$$\boldsymbol{F}_{\text{крит}} = \boldsymbol{F}\text{РАСПОБР}(\boldsymbol{\alpha}, \boldsymbol{v_1}, \boldsymbol{v_2}),$$

где, в данном случае, имеем:

$$\boldsymbol{v_1 = 1}, \quad \boldsymbol{v_2 = n - k - 1 = 17}.$$

Если фактическое значение $\boldsymbol{F_{x_i} > F}_{\text{крит}}$, то дополнительное включение в модель фактора $\boldsymbol{x_i}$ оправдано.

Проведем расчеты. Рассмотрим включение в уравнение множественной регрессии фактора $\boldsymbol{x_1}$ после того, как в уравнение уже включен фактор $\boldsymbol{x_2}$, т.е. $\boldsymbol{x_i = x_1}$. Формула для частного критерия Фишера, в этом случае, имеет вид (учтем, что ***k=2***):

$$\boldsymbol{F_{x_1} = \frac{R^2_{yx_1x_2} - R^2_{yx_2}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-k-1}{1}}.$$

Найдем предварительно $\boldsymbol{R^2_{yx_2}}-$ коэффициент детерминации двухфакторной регрессионной линейной модели вида $\boldsymbol{\hat{y} = a_2 + b_2^* x_2}$. Известно, что для модели такого вида, выполняется соотношение:

$$\boldsymbol{R^2_{yx_2} = r^2_{yx_2} = 0,941^2 = 0,885}$$

Подставив, получаем:

$$\boldsymbol{F_{x_1} = \frac{0,973^2 - 0,885}{1 - 0,973^2} \cdot \frac{20-2-1}{1} = 19,801}.$$

Рассмотрим включение в уравнение множественной регрессии фактора $\boldsymbol{x_2}$ после того, как в уравнение уже включен фактор $\boldsymbol{x_1}$, т.е. $\boldsymbol{x_i = x_2}$. Формула для частного критерия Фишера, в этом случае, имеет вид (учтем, что $\boldsymbol{k=2}$):

$$\boldsymbol{F_{x_2} = \frac{R^2_{yx_1x_2} - R^2_{yx_1}}{1 - R^2_{yx_1x_2}} \cdot \frac{n-k-1}{1}}.$$

Найдем предварительно $\boldsymbol{R^2_{yx_1}}$ – коэффициент детерминации двухфакторной регрессионной линейной модели вида $\boldsymbol{\hat{y} = a_2 + b_2^* x_2}$. Известно, что для модели такого вида, выполняется соотношение:

$$\boldsymbol{R^2_{yx_1} = r^2_{yx_1} = 0,97^2 = 0,9406}.$$

Подставив, получаем:

$$\boldsymbol{F_{x_2} = \frac{0,973^2 - 0,941}{1 - 0,973^2} \cdot \frac{20-2-1}{1} = 2,004}.$$

Найдем критическое значение критерия Фишера

$$\boldsymbol{F}_{\text{крит}} = \boldsymbol{F}\text{РАСПОБР}(\boldsymbol{\alpha}, \boldsymbol{v_1}, \boldsymbol{v_2}),$$

где, в данном случае, имеем:

$$\boldsymbol{\alpha = 0,05}; \quad \boldsymbol{v_1 = 1}; \quad \boldsymbol{v_2 = n - k - 1 = 17}.$$

$$\boldsymbol{F}_{\text{крит}} = \boldsymbol{4,45}.$$

Вывод: $\boldsymbol{F_{x_1}} > \boldsymbol{F}_{\text{крит}}$ $(\boldsymbol{19,801} > 4{,}45)$, т.е., если включить в модель фактор $\boldsymbol{x_2}$, а затем фактор $\boldsymbol{x_1}$, то прирост дисперсии за счет включения фактора $\boldsymbol{x_1}$ является существенным. Фактор $\boldsymbol{x_1}$ должен присутствовать в уравнении, даже, если включать его после $\boldsymbol{x_2}$.

$\boldsymbol{F_{x_2} < F_{\text{крит}}}$ $(\boldsymbol{2,004} < 4{,}45)$**,** т.е., если включить в модель фактор $\boldsymbol{x_1}$**,** а затем фактор $\boldsymbol{x_2}$, то прирост дисперсии за счет включения фактора $\boldsymbol{x_2}$ является несущественным. Фактор $\boldsymbol{x_2}$ включать в уравнение после $\boldsymbol{x_2}$ не следует.

2. Общий вывод: множественная модель с факторами $\boldsymbol{x_1}$ и $\boldsymbol{x_2}$**,** и с нескорректированным коэффициентом множественной детерминации $\boldsymbol{R^2_{yx1x2} = 0,947}$ содержит неинформативный фактор $\boldsymbol{x_2}$ **.** Если исключить из уравнения $\boldsymbol{x_2}$ **,** то можно ограничиться уравнением парной регрессии:

$$\boldsymbol{\hat{y} = a + bx_1}.$$

Найдем параметры этого уравнения:

$$\boldsymbol{b = \frac{\overline{y \cdot x_1} - \overline{y} \cdot \overline{x_1}}{varx_1} = \frac{63,815 - 6,19 \cdot 9,6}{3,571} = 1,23},$$

$$\boldsymbol{a = \overline{y} - b \cdot \overline{x_1} = 9,6 - 1,23 \cdot 6,19 = 1,99}.$$

Т.о. получили:

$$\boldsymbol{\hat{y} = 1,99 + 1,23 \cdot x_1},$$

коэффициент детерминации:

$$\boldsymbol{R^2_{yx_1} = r^2_{yx_1} = 0,97^2 = 0,941}.$$

8 ОПРЕДЕЛЕНИЕ ПАРАМЕТРОВ МНОЖЕСТВЕННОЙ РЕГРЕССИИ В EXCEL

8.1 Методика построения модели множественной регрессии в Excel

Задача 1.

Аналитический отдел предприятия осуществляет анализ зависимости прибыли предприятия (Y, млн.руб.) от величины оборотного капитала (X1, млн.руб.) и стоимости основных средств (X2 , млн.руб).

Y	148	152	144	132	30	124	132	124	116	118
X1	63	71	67	75	81	91	93	85	95	97
X2	35	45	49	33	55	61	55	61	65	67

Требуется:

1. Вычислить статистические характеристики для каждого показателя.
2. Построить матрицу коэффициентов парной корреляции и осуществить её анализ. Обосновать выбор факторных признаков.
3. Построить однофакторную линейную регрессию, рассчитать её параметры. Проанализировать значение коэффициента регрессии.
4. Вычислить точечный прогноз факторного признака, построив по нему линейный тренд.
5. Определить прогнозные значения результативного признака.
6. Результаты расчетов отразить на графике.

Решение

I. Сводную таблицу основных статистических характеристик для одного или нескольких массивов данных можно получить с помощью инструмента анализа данных **Описательная статистика**. Для этого выполните следующие шаги:

1. Введите исходные данные в MS Excel.
2. В главном меню выберите **Сервис->Анализ данных->Описательная статистика** (если в закладке Сервис пункт Анализ данных отсутствует, то выберите Сервис-Надстройки и поставьте галочку напротив Пакет анализа) (рисунок 1)

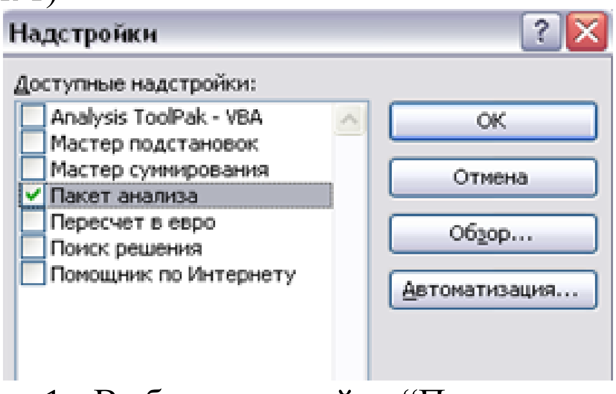


Рисунок 1- Выбор надстройки “Пакет анализа”

Заполните диалоговое окно ввода параметров:

- Входной интервал – диапазон, содержащий анализируемые данные (3 столбца в нашем случае).
- Группирование – по столбцам.
- Метки - флажок, который указывает, содержит ли первая строка названия столбцов или нет (в нашем случае - содержит).
- Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона.
- Итоговая статистика – поставить галочку.

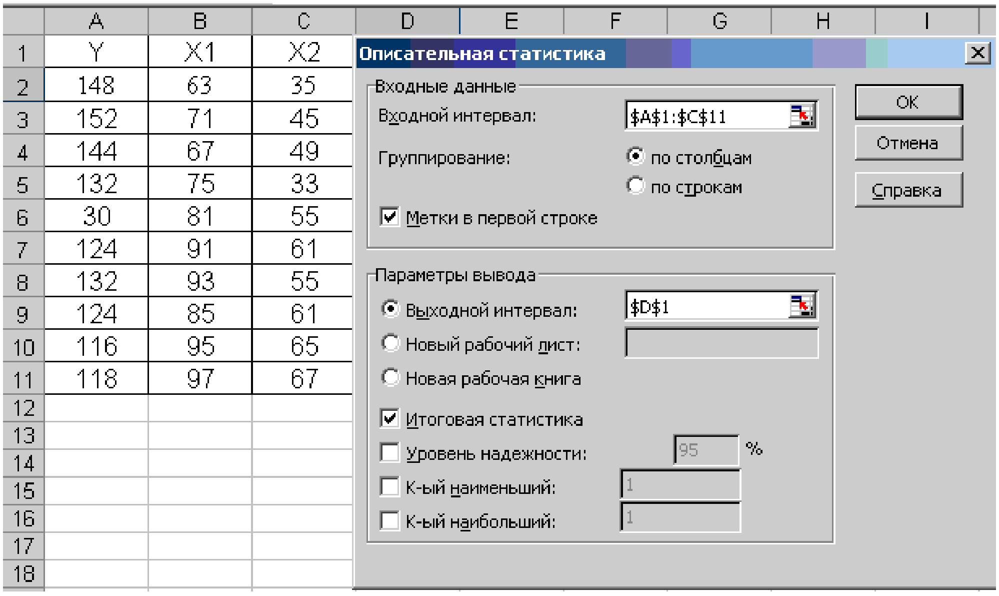


Рисунок 2- Параметры инструмента «Описательная статистика»

Результаты вычислений представлены на рисунке 3.

E	F	G	H	I	J
Y		X1		X2	
Среднее	122	Среднее	81,8	Среднее	52,6
Стандартная ошибка	10,94633	Стандартная ошибка	3,889587	Стандартная ошибка	3,768878
Медиана	128	Медиана	83	Медиана	55
Мода	132	Мода	#Н/Д	Мода	55
Стандартное отклонение	34,61535	Стандартное отклонение	12,29995	Стандартное отклонение	11,91824
Дисперсия выборки	1198,222	Дисперсия выборки	151,2889	Дисперсия выборки	142,0444
Эксцесс	6,782477	Эксцесс	-1,52261	Эксцесс	-0,83778
Асимметричность	-2,41677	Асимметричность	-0,25623	Асимметричность	-0,60393
Интервал	122	Интервал	34	Интервал	34
Минимум	30	Минимум	63	Минимум	33
Максимум	152	Максимум	97	Максимум	67
Сумма	1220	Сумма	818	Сумма	526
Счет	10	Счет	10	Счет	10

Рисунок 3 - Описательная статистика

II. Матрицу парных коэффициентов корреляции можно рассчитать, используя инструмент Анализа данных **Корреляция**. Для этого:

1. В главном меню выбрать **Сервис->Анализ данных->Корреляция**
2. Заполнить диалоговое окно ввода параметров, в качестве входного интервала следует указать весь диапазон представленных данных (рисунок 4).

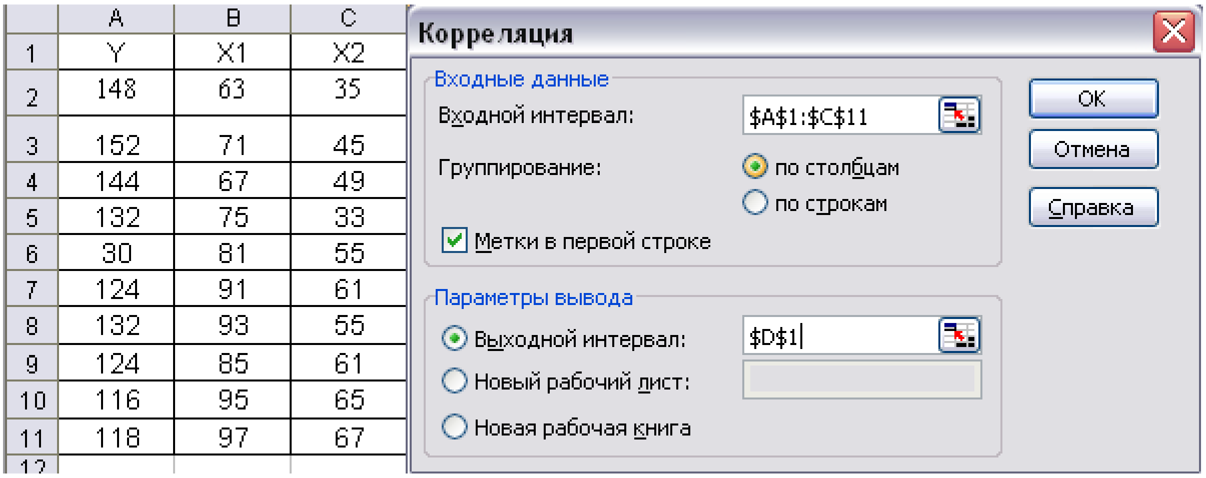


Рисунок 4 - Параметры инструмента «Корреляция»

3. Результаты построения матрицы коэффициентов парной корреляции представлены на рисунке 5.

	Y	X1	X2
Y	1		
X1	-0,29646	1	
X2	-0,33827	0,842236542	1

Рисунок 5 - Матрица коэффициентов парной корреляции

Коэффициент парной корреляции между прибылью предприятия и величиной оборотного капитала $r_{y\,x_1} = -0.29646$ имеет отрицательную величину, следовательно, между этими признаками имеет место обратная связь, т.е. при увеличении величины оборотного капитала прибыль предприятия уменьшается. Значение коэффициента невелико по абсолютной величине, следовательно, между прибылью предприятия и величиной оборотного капитала имеет место слабая связь. Эта тенденция имеет негативный характер и в качестве рекомендаций предприятию следует предложить осуществление детального анализа всех направлений использования оборотного капитала.

Коэффициент парной корреляции между прибылью предприятия и стоимостью основных средств $r_{y\,x_2} = -0.33827$ также имеет отрицательную величину, следовательно, между этими признаками имеет место обратная связь, что также свидетельствует о необходимости анализа всех направлений финансово-хозяйственной политики предприятия.

Из матрицы парных коэффициентов корреляции видно, что между величиной оборотного капитала и стоимостью основных средств (факторами X1 и X2) существует эффект мультиколлинеарности (коэффициент корреляции между ними $r_{x_1 x_2} = 0.8422$ >0,8), следовательно, оба фактора одновременно в модель регрессии включать нельзя, один из факторов необходимо из рассмотрения исключить. Из рассмотрения исключается тот фактор, который оказывает меньшее влияние на прибыль предприятия (результативный признак).

В данной задаче из рассмотрения необходимо исключить X1, так он оказывает меньшее влияние на результативный признак (коэффициент парной корреляции между X1 и Y меньше по модулю, чем между X2 и Y) , т.е.

$$\left| r_{yx_1} = -0.29646 \right| < \left| r_{yx_2} = -0.33827 \right|.$$

Итак, по результатам анализа матрицы парных коэффициентов корреляции в качестве факторного признака для построения однофакторной регрессии должен быть выбран фактор X2 (стоимость основных средств).

III. С помощью инструмента анализа данных **Регрессия,** помимо результатов регрессионной статистики, дисперсионного анализа и доверительных интервалов, можно получить остатки и графики подбора линии регрессии, остатков и нормальной вероятности. Порядок действий следующий:

1. Для построения модели парной регрессии в главном меню выберите **Сервис->Анализ данных->Регрессия**
2. Заполните диалоговое окно ввода данных и параметров вывода (рисунок 6)
   - Входной интервал Y – диапазон, содержащий данные результативного признака.
   - Входной интервал X – диапазон, содержащий данные факторов независимого признака (так как модель однофакторная, то построим её на основе фактора X2)
   - Метки – флажок, который указывает, содержит ли первая строка названия столбцов или нет.

- Константа-ноль – флажок, указывающий на наличие или отсутствие свободного члена в уравнении (метку напротив параметра не ставить).
- Выходной интервал – достаточно указать левую верхнюю ячейку диапазона с выходными данными регрессионного анализа.

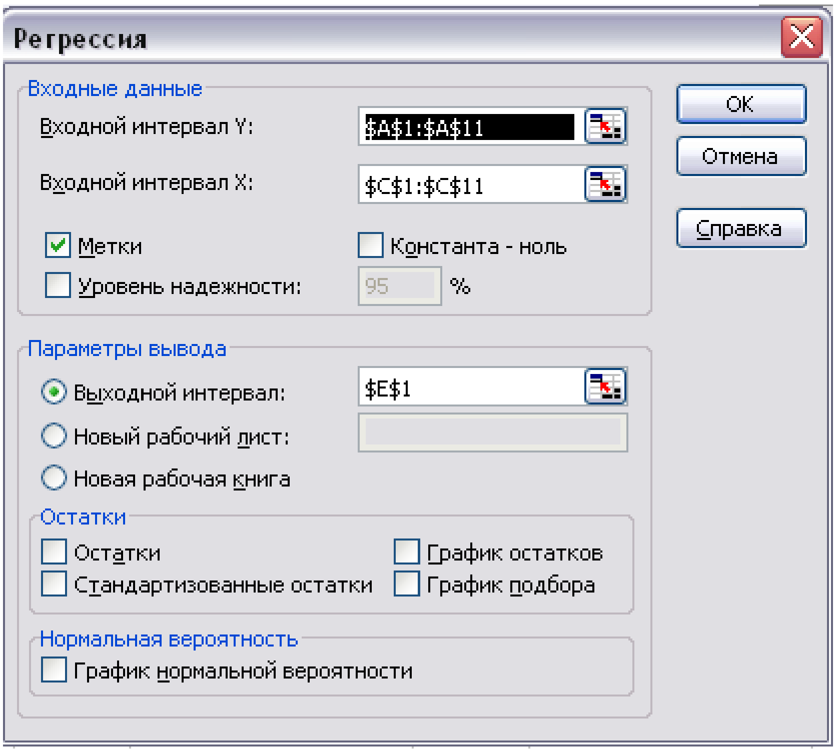


Рисунок 6 - Параметры инструмента «Регрессия»

Результаты регрессионного анализа представлены на рисунке 7.

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,33827							
R-квадрат	0,11443							
Нормированный R-квад	0,00373							
Стандартная ошибка	34,5507							
Наблюдения	10							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	1	1233,992491	1233,99	1,033710174	0,33905			
Остаток	8	9550,007509	1193,75					
Итого	9	10784						
	Коэффициен	*Стандартная ошибка*	*татисти*	*P-Значение*	*ижние 95%*	*ерхние 95%*	*ижние 95,0%*	*рхние 95,0*
Y-пересечение	173,678	51,98975382	3,34063	0,010221637	53,7898	293,567	53,7898	293,567
X2	-0,98248	0,966325593	-1,01672	0,339047923	-3,21083	1,24587	-3,21083	1,24587

Рисунок 7 - Результаты регрессионного анализа

На основании этого можно записать уравнение линейной парной регрессии:

$$y = 173.678 - 0.98248 * x$$

При увеличении стоимости основных средств (X2) на 1 млн.руб. прибыль предприятия (у) уменьшается на 0,98248 млн.рублей, что свидетельствует о нерациональном использовании финансовых ресурсов предприятия и необходимости пересмотра направлений финансово-хозяйственной политики предприятия.

IV. Для определения точечного прогноза факторного признака (X2), необходимо воспользоваться мастером диаграмм, в качестве параметра взяв диапазон для X2 – C2:C11 (рисунок 8).

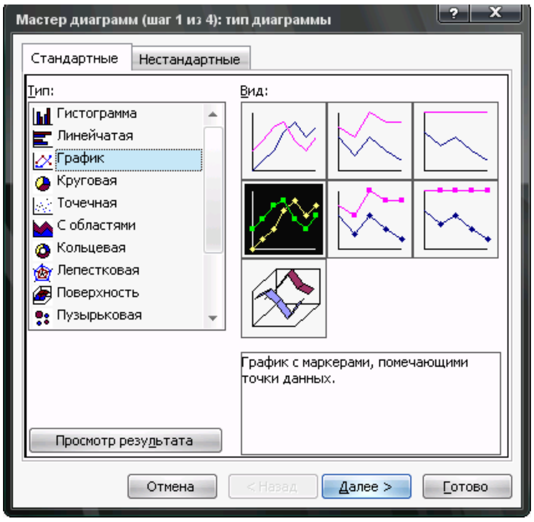


Рисунок 8 - Выбор типа диаграммы

На полученном графике необходимо добавить линию тренда.

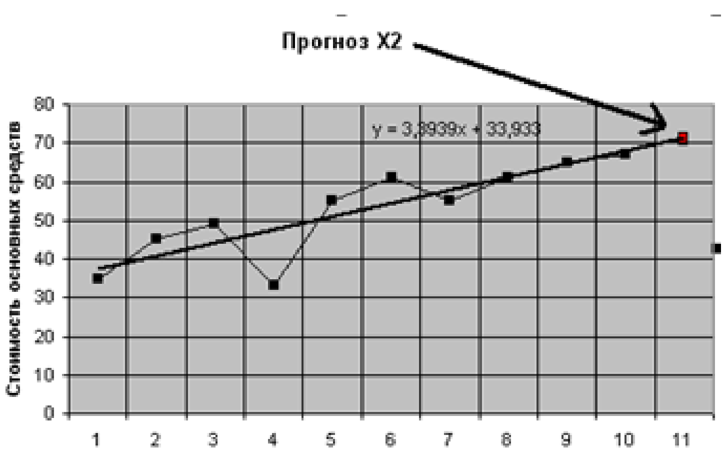


Рисунок 9 - Построение линейного тренда по факторному признаку X2

Исходя из полученного уравнения тренда, можно рассчитать прогнозное значение фактора X2 (для следующего, 11-го, момента времени):

X2 =3,3939*11+33,933=71,2659 (млн.руб.)

Прогнозное значение стоимости основных средств (X2) может быть отображено на рис.20).

V. Для получения прогнозного значения прибыли предприятия (результативного признака) необходимо в уравнение однофакторной регрессии подставить полученное значение прогноза стоимости основных средств (факторного признака):

y = 173.678 – 0.98248 * 71.2659 = 103.66068 (млн.руб.)

VI. Для отображения прогнозного значения результативного признака нужно воспользоваться мастером диаграмм (рисунок 10)

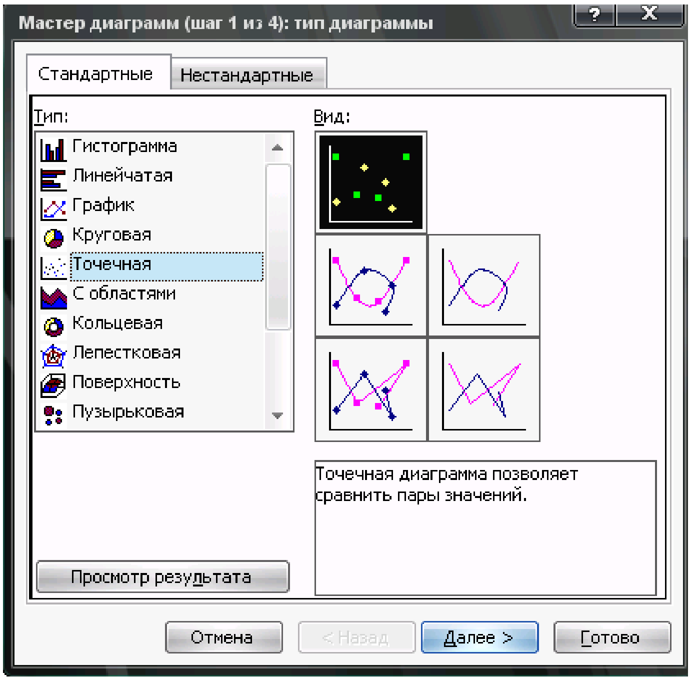


Рисунок 10 - Выбор типа диаграммы

В качестве входных диапазонов необходимо взять данные для факторного признака X2 - (C2:C11), для результативного признака Y - (A2:A11), на полученном графике отразить линию тренда (рисунок 11).

Для отображения на графике прогноза прибыли предприятия (Y) необходимо:

- Выделить полученный график
- Щелкнуть на нем правой кнопкой мыши и выбрать **Исходные данные**
- Добавить новый ряд, и ввести прогнозные значения X и Y, рассчитанные ранее.

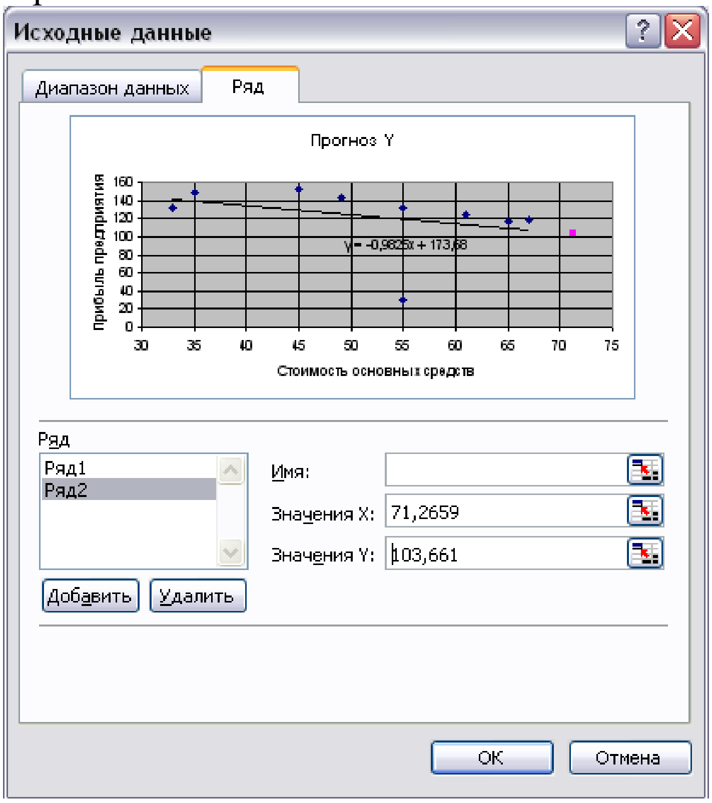


Рисунок 11- Добавление на график прогнозного значения прибыли предприятия

Окончательный вид графика будет следующий (рисунок 12).

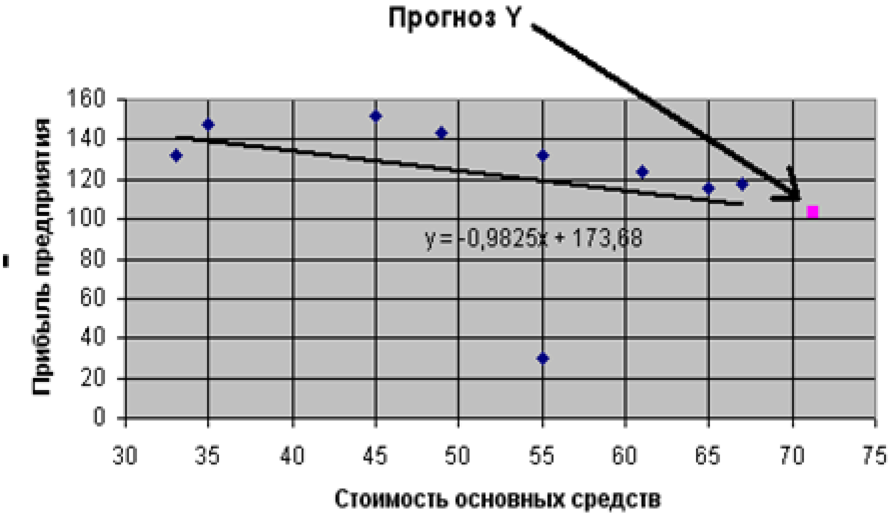


Рисунок 12 - Построение прогноза прибыли предприятия (Y)

8.2 Задачи для самостоятельного решения

Построить модель множественной регрессии по заданным в таблице данным.

Вариант	Признак	Наблюдения									
1	Y	36	28	66	74	80	84	82	98	112	96
	X1	32	40	44	28	50	56	50	56	60	62
	X2	51	54	60	68	70	74	84	82	86	84
2	Y	32	40	44	28	50	56	51	57	60	63
	X1	50	54	60	62	70	54	84	82	86	84
	X2	176	170	156	172	162	160	166	156	152	138
3	Y	50	54	60	62	70	54	84	82	86	84
	X1	176	170	156	172	162	160	166	156	152	138
	X2	86	94	100	96	134	114	122	118	130	108
4	Y	176	170	156	172	162	160	166	156	152	138
	X1	86	94	100	96	134	114	122	118	130	108
	X2	56	48	52	58	66	62	48	66	70	68
5	Y	86	94	100	96	134	114	122	118	130	108
	X1	56	48	52	58	66	62	48	66	70	68
	X2	64	68	82	76	84	96	100	104	108	102
6	Y	56	48	52	58	66	62	48	66	70	68
	X1	64	68	82	76	84	96	100	104	108	102
	X2	64	56	52	48	50	43	38	54	44	40
7	Y	40	44	38	52	50	64	70	68	78	90
	X1	60	68	80	76	64	96	100	104	106	68
	X2	22	30	20	32	44	34	52	56	66	68
8	Y	60	68	80	76	44	96	100	104	106	98
	X1	22	30	20	32	44	34	52	56	66	68
	X2	150	154	146	134	132	126	134	126	118	120
9	Y	22	30	20	32	44	34	52	56	66	68
	X1	150	154	146	134	132	126	134	128	118	120
	X2	60	68	64	72	78	88	90	82	92	94
10	Y	150	154	146	134	132	126	134	126	118	120
	X1	60	68	64	72	78	88	90	82	92	94
	X2	30	40	44	28	50	56	50	56	60	62

8.3 Мультиколениарность факторов регрессии

Задача

По десяти кредитным учреждениям получены данные, характеризующие зависимость объема прибыли (Y, млн.руб.) от величины доходов по кредитам (X1, млн.руб.), доходов по депозитам (X2, млн.руб.) и размера внутрибанковских расходов (X3, млн.руб.).

Y	32	40	44	28	50	56	50	56	60	62
X1	22	30	20	32	44	34	52	56	66	68
X2	56	48	52	58	66	62	48	66	70	68
X3	64	68	82	76	84	96	100	104	108	102

Осуществить выбор факторных признаков для построения многофакторной регрессионной модели.

1. Рассчитать параметры регрессионной модели. Оценить ее качество.
2. Для характеристики модели определить:
    - средние коэффициенты эластичности;
    - бета-коэффициенты,
    - дельта-коэффициенты.
3. Оценить с помощью t-критерия Стьюдента статистическую значимость коэффициентов уравнения множественной регрессии.
4. Построить регрессионную модель со статистически значимыми факторами. Оценить ее качество.
5. Определить точечный и интервальный прогноз результативного показателя.

Решение

I. Выбор факторных признаков для построения модели осуществляется с помощью матрицы коэффициентов парной корреляции. Для её построения необходимо:

- выбрать **Сервис->Анализ данных->Корреляция**
- заполнить необходимые поля диалогового меню (рисунок 13)

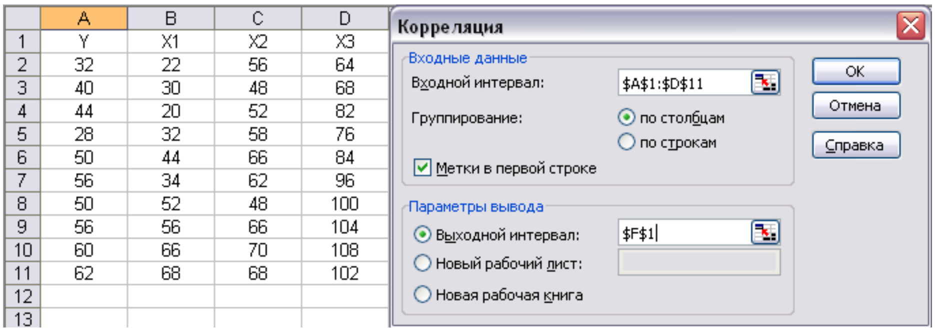


Рисунок 13 - Ввод параметров инструмента «Корреляция»

Результаты представлены на рисунке 14.

	Y	X1	X2	X3
Y	1			
X1	0,783838	1		
X2	0,599604	0,642876	1	
X3	0,883761	0,843608	0,57207	1

Рисунок 14 - Таблица коэффициентов парных корреляций

Для выявления явления мультиколлинеарности необходимо проанализировать коэффициенты парной корреляции между факторными признаками. Если имеют место коэффициенты, значение которых по модулю больше 0,8, то, следовательно, мультиколлинеарность присутствует, и это явление необходимо устранять. Если же значения коэффициентов парной корреляции между факторными признаками, взятые по модулю, меньше величины 0,8, то явление мультиколлинеарности отсутствует, и, следовательно, все факторные признаки можно включать в модель множественной регрессии.

Так как $r_{x_1 x_3} = 0.843608 > 0.8$, т.е. между факторными признаками X1 и X3 существует явление мультиколлинеарности, то для построения модели выбираем тот факторный признак, который оказывает большее влияние на результативный признак (фактор, для которого коэффициент парной корреляции с результативным признаком, взятый по модулю, является большим).

$$\left| r_{yx_1} = 0.783838 \right| < \left| r_{yx_3} = 0.883761 \right|$$

Следовательно, фактор X3 оказывает большее влияние на результативный признак (Y) и этот фактор рекомендуется в модели оставить. Фактор X1 оказывает меньшее влияние на результативный признак (Y) и этот фактор рекомендуется из модели исключить.

Таким образом, для построения модели множественной регрессии выбираются два факторных признака - X2 (величина доходов по депозитам) и X3 (величина внутрибанковских расходов).

II. Расчет параметров регрессионной модели можно осуществить с помощью инструмента анализа данных **Регрессия** (см. задача 2), отличие заключается в том, что в качестве диапазона значений фактора X необходимо указать диапазон значений факторов X2 и X3 (рисунок 15).

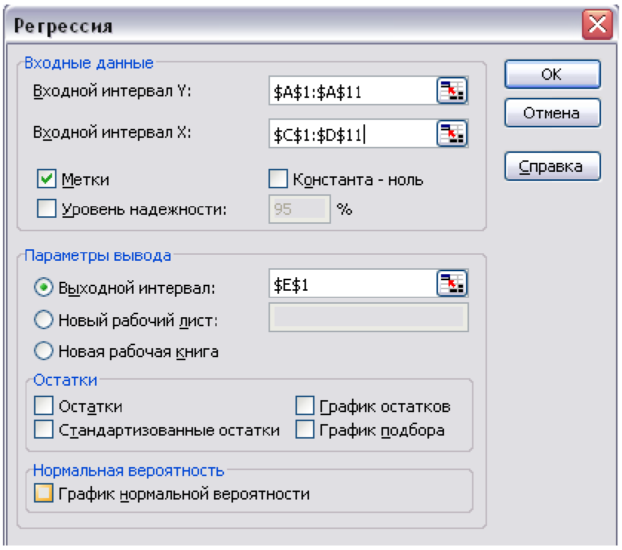


Рисунок 15 - Ввод параметров регрессии

Результаты построение множественной регрессии представлены на рисунке 16.

ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множеств	0,891166							
R-квадрат	0,794176							
Нормиров	0,73537							
Стандартн	5,968678							
Наблюден	10							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	2	962,2242	481,1121	13,50486	0,003956			
Остаток	7	249,3758	35,62511					
Итого	9	1211,6						
	Коэффициен	*дартная о*	*статисти*	*Р-Значени*	*ижние 95%*	*ерхние 95%*	*жние 95,0*	*рхние 95,0%*
Y-пересеч	-16,2872	14,93646	-1,09043	0,311629	-51,6064	19,0319	-51,6064	19,0319
X2	0,197247	0,295027	0,668573	0,525194	-0,50038	0,894874	-0,50038	0,894874
X3	0,592429	0,154086	3,844787	0,006335	0,228073	0,956786	0,228073	0,956786

Рисунок 16 - Вывод итогов регрессии

На основании полученных данных можно записать уравнение множественной регрессии

$$Y=-16,2872 + 0,197247*X2 + 0,592429*X3$$

Оценим качество построенной модели множественной регрессии по следующим направлениям:

- Коэффициент детерминации R^2 = 0.794176 достаточно близок к 1, следовательно, качество модели можно признать высоким.
- Критерий Фишера F = 13,50486 > Fтабл = 4,74 , следовательно, уравнение регрессии признается статистически значимым и может быть использовано для анализа и прогнозирования экономических процессов.

Для вычисления Fтабл необходимо определить:

- степень свободы числителя m=2 (число факторных признаков);

- степень свободы знаменателя n-m-1=10-2-1=7;

- уровень значимости α =0,05.

III. Оценим качество построенной модели множественной регрессии с помощью коэффициентов эластичности, β - и Δ - коэффициентов.

Коэффициент эластичности определяется:

$$Э_i = b_i * \frac{\overline{X_i}}{\overline{Y}} , \tag{1}$$

где $\overline{X_i}$ - среднее значение соответствующего факторного признака,

$\overline{Y}$ - среднее значение результативного признака.

b_i – коэффициенты регрессии соответствующих факторных признаков.

ß-коэффициент определяется по следующей формуле:

$$\beta_i = b_i * \frac{\sigma_{xi}}{\sigma_y} , \tag{2}$$

где σ_{xi} - среднеквадратическое отклонение (СКО) соответствующего факторного признака (рассчитывается как корень квадратный из дисперсии признака),

σ_Y - СКО результативного признака.

Δ-коэффициент определяется по следующей формуле:

$$\Delta i = \beta_i * \frac{r_{yxi}}{R^2} , \tag{3}$$

где r_{yxi} - коэффициент парной корреляции результативного и соответствующего факторного признаков,

R^2 - коэффициент детерминации.

На рисунке 17 представлены формулы расчетов описанных выше коэффициентов

	A	B	C	D
1		Y	X2	X3
2		32	56	64
3		40	48	68
4		44	52	82
5		28	58	76
6		50	66	84
7		56	62	96
8		50	48	100
9		56	66	104
10		60	70	108
11		62	68	102
12				
13	ср.знач	=СРЗНАЧ(B2:B11)	=СРЗНАЧ(C2:C11)	=СРЗНАЧ(D2:D11)
14	эласт.		=C17*C13/B13	=D17*D13/C13
15	дисп	=ДИСП(B2:B11)	=ДИСП(C2:C11)	=ДИСП(D2:D11)
16	ско	=КОРЕНЬ(B15)	=КОРЕНЬ(C15)	=КОРЕНЬ(D15)
17	bi		0,197246886359573	0,59242946329357
18	Bi		=C17*C16/$B$16	=D17*D16/$B$16
19	ryxi		0,599603698767859	0,883760976029946
20	deltai		=C18*C19/B21	=D18*D19/B21
21	R^2	0,794176478516679		

Рисунок 17 - Формулы расчетов коэффициентов

Результаты вычислений представлены в таблице 2.

Таблица 2 - Результаты расчета бета-, дельта- и коэффициентов эластичности

Y	X2	X3
32	56	64
40	48	68
44	52	82
28	58	76
50	66	84
56	62	96
50	48	100
56	66	104
60	70	108
62	68	102

Ср.знач	47,8	59,4	88,4
Эласт.		0,245114	0,881663
Дисп	134,6222	67,6	247,8222
СКО	11,60268	8,221922	15,74237
bi		0,197247	0,592429
β_i		0,139774	0,803801
r_{yx_i}		0,599604	0,883761
Δ_i		0,105529	0,894471

Выводы:

- Частный коэффициент эластичности показывает, на сколько процентов изменится среднее значение результативного признака, если среднее значение конкретного факторного признака изменится на 1 %, т.е., при увеличении на 1% величины доходов по депозитным операциям (X2) прибыль банка увеличится на 0,245 % (Э2 = 0,245), при увеличении на 1% размера внутрибанковских расходов (X3) объём прибыли увеличится на 0,88% (Э3 =0,881).
- β-коэффициент показывает, на какую величину изменится СКО результативного признака, если СКО конкретного факторного признака изменится на 1 единицу, т.е. при увеличении на 1 единицу СКО доходов по депозитам (X2), СКО объёма прибыли увеличится на 0,14 (β_2 =0,139774); при увеличении на 1 единицу СКО внутрибанковских расходов СКО прибыли организации увеличится на 0,804 единицы (β_3 = 0,803801).
- Δ-коэффициент показывает удельный вес влияния конкретного факторного признака в совместном влиянии всех факторных признаков на результативный показатель, т.е. удельный вес влияния внутрибанковских расходов (X3) на объём прибыли (результативный

признак) составляет 89,4% (Δ_3 = 0,8944), а удельное влияние доходов по депозитам (X2) на прибыль составляет 10,5 % (Δ_2 = 0,1055).

IV. Для оценки статистической значимости факторных признаков модели множественной регрессии используется t-критерий Стьюдента.

С помощью функции СТЬЮДРАСПОБР(0,05;7) определим табличное значение t табл = 2,364624.

Сравним расчетные значения t-статистики, взятые по модулю, с табличным значением этого критерия (расчетные значения берутся из столбца t-статистика таблицы 3 регрессионного анализа).

Таблица 3 - Результаты регрессионного анализа

	Коэффициенты	Стандартная ошибка	t-статистика	P-Значение	Нижние 95%	Верхние 95%	Нижние 95,0%	Верхние 95,0%
Y-пересечение	-16,2872	14,93646	-1,090434022	0,311629	-51,6064	19,0319	-51,6064	19,0319
X2	0,197247	0,295027	0,668573305	0,525194	-0,50038	0,894874	-0,50038	0,894874
X3	0,592429	0,154086	3,844787137	0,006335	0,228073	0,956786	0,228073	0,956786

- t_{x2} = 0,668573 < $t_{таб}$=2,364624, следовательно, фактор X_2 признается статистически не значимым. Такой фактор из модели рекомендуется исключить.
- t_{x3} = 3,844787 > $t_{таб}$=2,364624, следовательно, фактор X_3 признается статистически значимым и информативным. Такой фактор рекомендуется в модели регрессии оставить.

IV. Построим регрессионную модель со статистически значимыми факторами. Для конкретного примера статистически значимым фактором является только фактор X3 (величина внутрибанковских расходов). Подробное построение регрессионных моделей рассмотрено ранее. Осуществим следующие установки в окне **Регрессия** (рисунок 18).

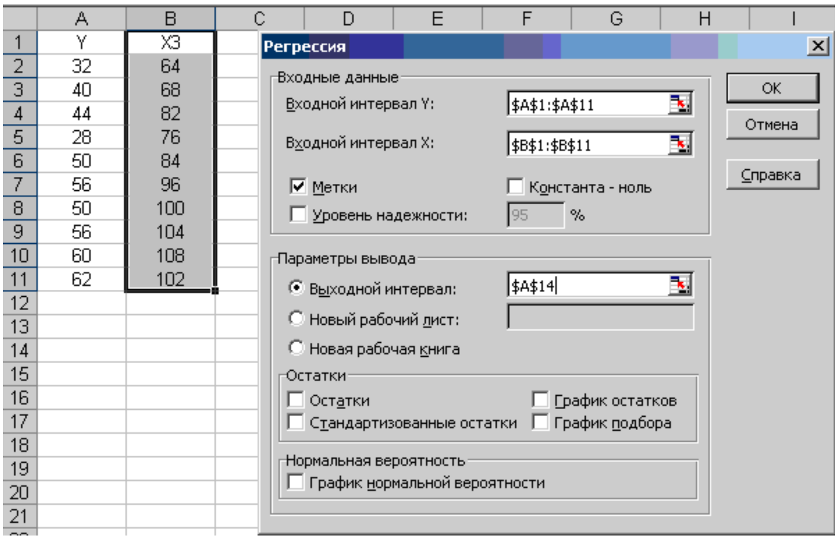


Рисунок 16 - Диалоговое окно Регрессия

Получим следующие результаты (рисунок 30)

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,883761
R-квадрат	0,781033
Нормированный R-квадрат	0,753663
Стандартная ошибка	5,758688
Наблюдения	10

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Знач. F*
Регрессия	1	946,3001	946,3001	28,53526	0,000693
Остаток	8	265,2999	33,16248		
Итого	9	1211,6			

	Коэфф.	*Стандарт. ошибка*	*t-статист.*	*P-Знач.*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-9,78049	10,93189	-0,89467	0,397079	-34,9895	15,4285	-34,9895	15,4285
X3	0,651363	0,121936	5,341841	0,000693	0,370178	0,932548	0,370178	0,932548

Рисунок 17 - Вывод итогов регрессии

Запишем уравнение зависимости прибыли организации от величины внутрибанковских расходов (X3):

Y = 0,651363*X3 – 9,78049

Качество этой модели может быть оценено по коэффициенту детерминации R^2=0,781, следовательно, размер прибыли кредитных организаций на 78,1 % зависит от величины внутрибанковских расходов.

При сравнении качества регрессии y = f (X3) с качеством регрессии y = f (X2, X3) , имеющей R^2=0,794, можно утверждать, что улучшение качества модели не произошло.

Значение F-критерия Фишера составляет 28,53 > Fтабл (1,8)=5,32 , следовательно, построенное уравнение регрессии признается статистически значимым и может быть использовано для анализа и прогнозирования процессов.

V. Построение точечного прогноза прибыли кредитного учреждения (результативного показателя) может быть осуществлено по уравнению множественной регрессии, построенной в пункте 4 задачи, или по уравнению регрессии, содержащего только статистически значимые факторы (пункт 5 задачи).

Воспользуемся уравнением множественной регрессии, так как качество этой модели признано лучшим:

Y=0,197247*X2+0,592429*X3-16,2872

Для построения точечного прогноза результативного признака необходимо рассчитать точечные прогнозы факторных признаков (величины доходов организации по депозитам и величины внутрибанковских расходов). Для этого построим графики X2(t), X3(t) и тренд по каждому из факторов (рисунок 18,19).

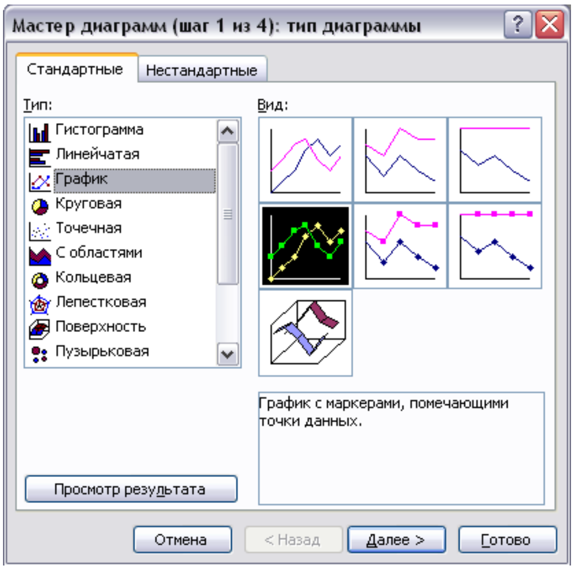


Рисунок 18 - Выбор типа диаграммы

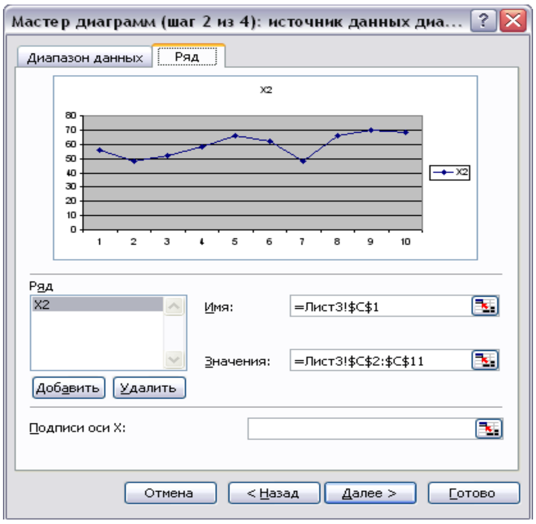


Рисунок 19- Выбор источника данных

На полученной диаграмме необходимо добавить линию тренда:

Диаграмма->Добавить линию тренда.

В настройках тренда в закладке **Параметры** указать (рисунок 20):

- **Прогноз вперед на 1 единицу**
- **Показать уравнение на диаграмме**
- **Поместить на диаграмму величину достоверности аппроксимации.**

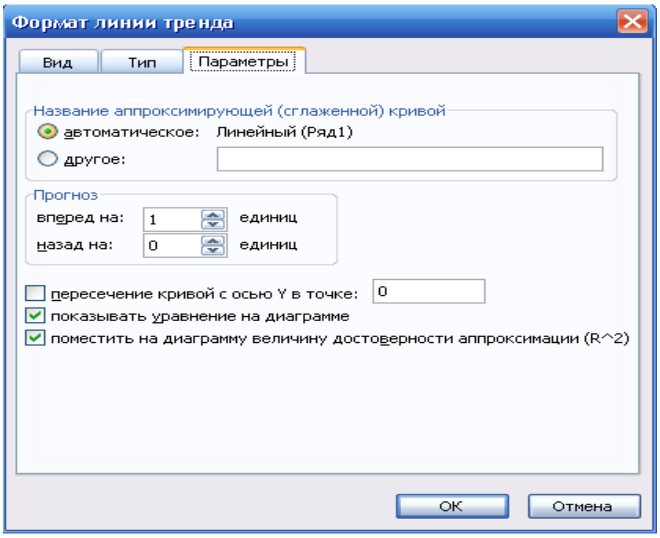


Рисунок 20 - Параметры линии тренда

Результат построения представлен на рисунке 21.

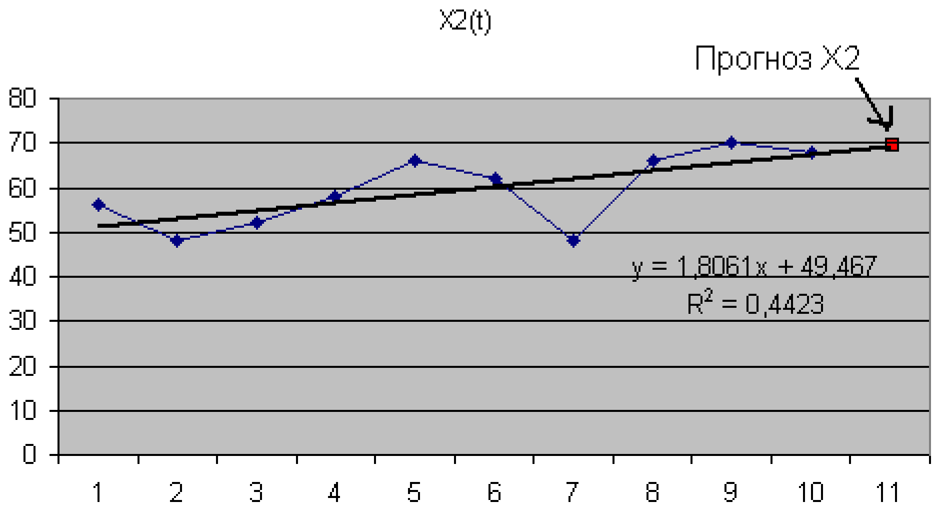


Рисунок 21 - Построение прогноза величины доходов по депозитам (X2)

В полученное уравнение тренда

$$X2 = 1{,}8061*x + 49{,}467 \quad ,$$

в котором в качестве факторного признака выступает «время», необходимо подставить следующий момент времени. Так как временной ряд факторного признака X2 представлен 10 наблюдениями, то следующий момент времени будет представлен числом 11.

Получим:

$X2_{Прогн.}=1,8061*11+49,467 = 69,3341$ (млн.руб.)

Осуществляя аналогичные установки для фактора X3, построим прогноз по величине внутрибанковских расходов (рисунок 22) .

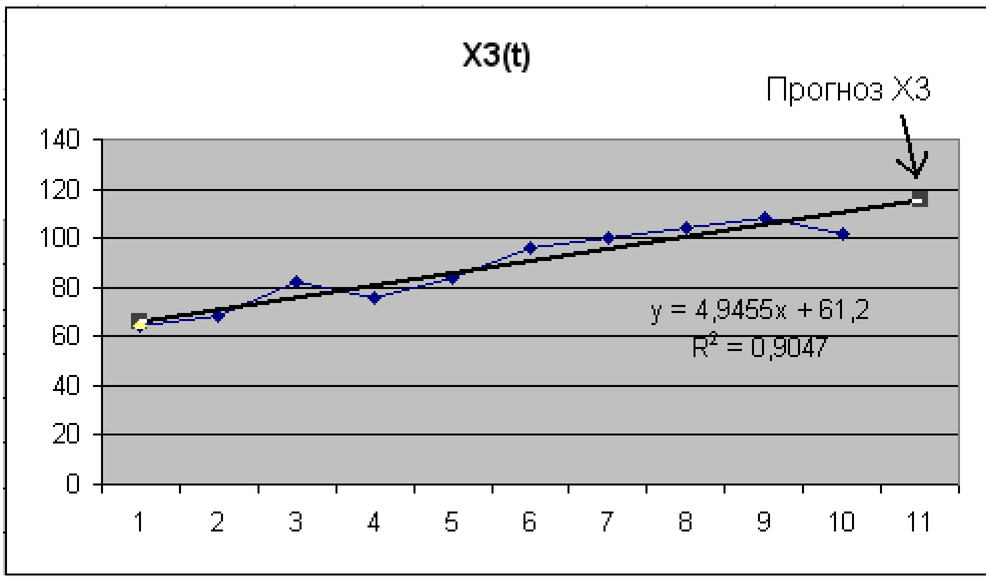


Рисунок 22 - Построение прогноза величины внутрибанковских расходов (X3)

Определим прогнозное значение внутрибанковских расходов из построенного уравнения тренда:

$X3_{Прогн.}=4,9455 *11+61,2=115,6005$ (млн.руб.)

Рассчитанные значения прогнозов по факторам X2 и X3 подставим в уравнение множественной регрессии:

$Y=0,197247*X2 + 0,592429*X3 - 16,2872$

Получим:

$Y_{Прогн.} = 0,197247*X2_{Прогн.} + 0,592429*X3_{прогн.} - 16,2872$

$Y_{Прогн.}=0,197247*69,3341+0,592429*115,6005-16,2872=65,873832$ (млн.руб.)

Определим интервальный прогноз результирующего показателя, для этого рассчитаем ширину доверительного интервала по формуле:

$$U\,(k) = SKp\,\sqrt{1 + 1/n + \frac{(Y_{прогн} - \overline{Y})^2}{\sum (Y - \overline{Y})^2}}\,, \tag{4}$$

где $S = \sqrt{\frac{\sum (Y - \widehat{Y}_x)^2}{n - m - 1}}$, = 5,968678 (стандартная ошибка из таблицы регрессионной статистики, рисунок 27),

$Y_{Прогн.}$ – рассчитанное выше значение точечного прогноза результативного признака,

Кр= $t_{таб}$= 2,364624 табличный коэффициент Стьюдента, можно определить с помощью функции СТЬЮДРАСПОБР(0,05;7)

$\overline{Y}$ - среднее значение результативного признака (прибыли кредитной организации).

Подставляя эти значения в выше записанную формулу, получим:

U(k)= 5,968678*2,364624*√(1+0,1+326,6634/1211,6)= 16,51731

Таким образом, прогнозное значение прибыли кредитных организаций $Y_{прогн}$= 65,873832 , будет находиться между верхней границей, равной

65,873832 + 16,51731 = 82,39113827 (млн.руб.)

и нижней границей, равной

65,873832 – 16,51731= 49,3565254 (млн.руб.)

Вывод: Прогнозное значение прибыли исследуемых кредитных организаций, рассчитанное по уравнению множественной регрессии, будет находиться в интервале от 49,36 мл.руб. до 82,39 млн.руб.

Данное уравнение регрессии признано статистически значимым по критерию Фишера и обладает достаточно высоким качеством, следовательно, результаты расчетов можно признать надежными и достоверными.

8.4 Задания для самостоятельной работы

Строка, соответствующая Вашему номеру варианта, определяет значения признака Y, последующие три строки соответствуют признакам XI, X2, X3.

Вариант	Наблюдения								
1	36	28	66	74	80	84	82	98	112	96
2	40	44	28	52	50	64	70	68	78	90
3	32	40	44	28	50	56	50	56	60	62
4	60	68	80	76	44	96	100	104	106	98
5	50	54	60	62	70	54	84	82	86	84
6	22	30	20	32	44	34	52	56	66	68
7	176	170	156	172	162	160	166	156	152	138
8	150	154	146	134	132	126	134	126	88	120
9	86	94	100	96	134	114	122	118	130	108
10	60	68	64	72	78	88	90	82	92	94
11	56	48	52	58	66	62	48	66	70	68
12	30	40	44	28	50	56	50	56	60	62
13	64	68	82	76	84	96	100	104	108	102

9 ВРЕМЕННЫЕ РЯДЫ

9.1 Тестирование временного ряда на стационареость

Эконометрическую модель можно построить, используя два типа исходных данных:

- данные, характеризующие совокупность различных объектов в определенный момент (период) времени;
- данные, характеризующие один объект за ряд последовательных моментов (периодов) времени.

Модели, построенные по данным первого типа, называются ***пространственными моделями***. Модели, построенные по данным второго типа, называются ***моделями временных рядов.***

Определение:
Временной ряд (динамический ряд, ряд динамики) – это совокупность значений какого-либо показателя за несколько последовательных моментов (периодов) времени.

Ряд динамики может содержать следующие составляющие (***компоненты***):

1) основную тенденцию развития (***тренд***);
2) ***циклические*** (в том числе и сезонные) колебания;
3) ***случайные*** колебания.

Основная задача эконометрического исследования временного ряда: выявление и количественное выражение его компонент (тенденции, периодичности, случайной компоненты) в целях их использования для прогнозирования будущих значений ряда.

Для исследования временых рядов используются математические модели. Чаще всего это модели следующих видов:

1. Аддитивная $\mathrm{Y}_{\mathrm{t}} = T_t + S_t + E_t$
2. Мультипликативная $\mathrm{Y}_{\mathrm{t}} = T_t \times S_t \times E_t$
3. Смешанная $\mathrm{Y}_{\mathrm{t}} = T_t \times S_t + E_t$

Таким образом временной ряд представляет собой случайный процесс, имеющий в общем случае три составляющие. Случайные процессы делятся на стационарные и нестационарные. Для стационарного процесса должны выполняться следующие условия:

$$M(X(t)) = \text{const}; \quad D_x(t) = \text{const}; \quad \rho(t_1, t_2) = \rho(t_2 - t_1).$$

Для исследования временного ряда используется автокорреляционная функция (АКФ). АКФ измеряется с помощью линейного коэффициента корреляции между уровнями исходного временного ряда и уровнями ряда, сдвинутыми на несколько шагов назад во времени:

$$r_\tau = \frac{\sum\limits_{t=\tau+1}^{n}(y_t-\bar{y}_{1\tau})\cdot(y_{t-\tau}-\bar{y}_{2\tau})}{\sqrt{\sum\limits_{t=\tau+1}^{n}(y_t-\bar{y}_{1\tau})^2\cdot\sum\limits_{t=\tau+1}^{n}(y_{t-\tau}-\bar{y}_{2\tau})^2}}$$

$$\bar{y}_{1\tau} = \frac{\sum\limits_{t=\tau+1}^{n} y_t}{n-\tau} \qquad \bar{y}_{2\tau} = \frac{\sum\limits_{t=\tau+1}^{n} y_{t-\tau}}{n-\tau}$$

τ – величина сдвига во времени, или лаг

Например, лаг τ=1 означает, что ряд сдвинут на один период (момент) назад и т.д. С увеличением лага число пар значений, по которым рассчитывается коэффициент автокорреляции, уменьшается.

$$\tau=1 \Rightarrow \quad r_1 = \frac{\sum\limits_{t=2}^{n}(y_t-\bar{y}_1)\cdot(y_{t-1}-\bar{y}_2)}{\sqrt{\sum\limits_{t=2}^{n}(y_t-\bar{y}_1)^2\cdot\sum\limits_{t=2}^{n}(y_{t-1}-\bar{y}_2)^2}}$$

$$\tau=2 \Rightarrow \quad r_2 = \frac{\sum\limits_{t=3}^{n}(y_t-\bar{y}_3)\cdot(y_{t-2}-\bar{y}_4)}{\sqrt{\sum\limits_{t=3}^{n}(y_t-\bar{y}_3)^2\cdot\sum\limits_{t=3}^{n}(y_{t-2}-\bar{y}_4)^2}}$$

Пример 1. **Тестирование временного ряда на стационарность**

Стационарный временной ряд, у которого математическое ожидание равно 0, а величины $e(\tau_i\ )$ не коррелированны, часто называют “*белым шумом*”. Очевидно, что для белого шума

$$r(\tau)=1, \text{ если } \tau=0$$
$$r(\tau)=0, \text{ если } \tau\neq 0$$

В столбце y(0) документа Excel (приведенного на рисунке) представлены 20 значений временного ряда. Необходимо вычислить выборочное математическое ожидание, дисперсию и коэффициент автокорреляции $r\ (\tau)$, $\tau = 0,1,2,3$.

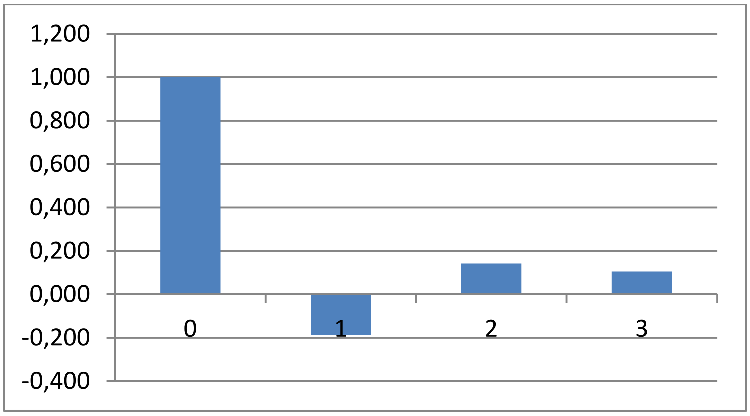
Путем копирования и последующего сдвига построим таблицу значений y для трех лагов y(1), y(2), y(3).

y(0)	Y(1)	Y(2)	Y(3)
27,65			
32,07	27,65		
33,87	32,07	27,65	
24,83	33,87	32,07	27,65
34,58	24,83	33,87	32,07
24,59	34,58	24,83	33,87
25	24,59	34,58	24,83
25	25	24,59	34,58
26,29	25	25	24,59
26,03	26,29	25	25
33,65	26,03	26,29	25
25,38	33,65	26,03	26,29
28,82	25,38	33,65	26,03
33,61	28,82	25,38	33,65
27,85	33,61	28,82	25,38
31,46	27,85	33,61	28,82
27,89	31,46	27,85	33,61
26,78	27,89	31,46	27,85
25,63	26,78	27,89	31,46
29,59	25,63	26,78	27,89
	29,59	25,63	26,78
		29,59	25,63
			29,59

Воспользуемся функцией Excel **КОРРЕЛ**(диапазон1, диапазон 2) расчитаем коэффициенты корреляции для различных лагов. Кроме того вычислим математическое ожидание и дисперсию.

τ=0	τ=1	τ=2	τ=3	M(Y)	D(y)
1,000	-0,189	0,141	0,105	28,529	11,995

Коррелограмма временного ряда имеет вид



Проверка теста на стационарность сводится к проверке статистических гипотез:

$$H_0 : M(y) = const \quad H_1 : M(y) \neq const$$

$$H_0 : D(y) = const \quad H_1 : D(y) \neq const$$

Временной ряд разбивается на две части (не обязательно одинаковые) по количеству содержащихся в них значений. Пусть первая часть (обозначим ее $Y(m)$) содержит m наблюдений, а вторая часть - $Y(k)$ содержит k наблюдений.

Для каждой части временного ряда вычислим выборочное среднее и выборочные дисперсии. Далее рассчитаем значение критерия

$$K_s = \frac{|\bar{y}_m - \bar{y}_k|}{\sqrt{\frac{S_m^2}{m} + \frac{S_k^2}{k}}}$$

Пусть m=10 и k=10. Тогда $\bar{y}_m = 27{,}991 \quad \bar{y}_k = 29{,}066$

$$S_m^2 = 15{,}66 \qquad S_k^2 = 9{,}02$$

$$K_s = 0{,}435$$

Проверяем значимость расчитанного критерия по статистике Стьюдента для α=0,05.

$$t_{кр} = \text{СТЬЮДРАСПОБР}(\alpha\ ,\ m+k-2) = 2{,}01$$

Поскольку $K_s < t_{кр}$, то принимается гипотеза H_0 о постоянстве математического ожидания.

Для проверки гипотезы о постоянстве дисперсии определим следующий критерий:

$$F_s = \frac{S_m^2}{s_k^2} = 1{,}249$$

Проверяем значимость расчитанного критерия по статистике Фишера для α=0,05.

$F_{кр1} =$ FРАСПОБР(0,025,m-1,k-1) = 4,026

$F_{кр2} =$ FРАСПОБР(0,975,m-1,k-1) = 0,248

Проверяем выполнение неравенства: $F_{кр1} < F_s < F_{кр2}$

Поскольку неравенство выполняется можносделать вывод о постоянстве дисперсии временного ряда. Следовательно можно сделать общий вывод о стационарности временного ряда.

Метод Фостера-Стюарта. Является более универсальным и дает более надежные результаты. Каждому уровню ряда y_i, начиная со второго, ставится в соответствие два значения p_i q_i по следующим правилам:

$p_i = 1$, если уровень y_i меньше всех предыдущих уровней, т. е.

$$y_i < y_1, y_2, ..., y_{i-1} ,$$

и $p_i = 0$ в противном случае;

$q_i = 1$, если уровень y_i больше всех предыдущих уровней, т. е.

$$y_i > y_1, y_2 , ..., y_{i+1} ,$$

и $q_i = 0$ в противном случае.

Вычисляется статистика

$$t_p = \frac{\sum_{i=2}^{n}(p_i - q_i)}{2\sum_{i=2}^{n}\frac{1}{i}} .$$

Гипотеза об отсутствии тенденции отвергается, если выполняется условие $t_p > t_1(\alpha, n\text{-}1)$,

где $t_{1-\alpha,\ n-1})$ – табличное значение t-критерия Стьюдента при уровне значимости α и числе степеней свободы $n - 1$.

9.2 Выделение тренда в случае нестационарного временного ряда

Процесс выделения **тренда** (выравнивание ряда) включает в себя три этапа:

1) выбор типа кривой, соответствующей характеру изменения ряда;

2) определение числовых значений параметров кривой;

3) оценка качества подобранной модели тренда.

На практике используется несколько приемов, позволяющих подобрать соответствующую (адекватную) действительности форму кривой.

Первый прием - визуальный - является наиболее простым. Форма кривой выбирается на основе графического изображения временного ряда, формы его корреляционного поля. Корреляционным полем временного ряда $Y(t)$ называется множество точек $(t, Y(t))$ на плоскости (t, OY). Однако при этом есть риск субъективного и произвольного выбора. Результат выбора зависит от масштаба графического изображения. Но при относительно простых конфигурациях корреляционного поля и незначительных отклонениях (помехах) от тенденции развития визуальный подход дает вполне приемлемые результаты. Для уменьшения субъективизма следует привлекать по возможности всю информацию об исследуемом процессе, в том числе о внутренней структуре и движущих силах процесса.

Часто используются другие критерии отбора. Обычно в качестве критерия принимают сумму квадратов отклонений фактических значений от расчетных, полученных выравниванием. Выбирается кривая, которой соответствует минимальное значение критерия. Однако минимум суммы квадратов еще не означает, что тенденция развития временного ряда описана наилучшим образом. Это объясняется тем, что через любой набор из N точек на плоскости всегда можно провести многочлен степени $N - 1$. Поэтому, выбрав достаточно сложную кривую, можно сделать нулевой сумму квадратов отклонений расчетных значений от заданных. Однако в этом случае вряд ли будет отражена тенденция дальнейшего развития процесса.

Предположим, что для заданного временного ряда $Y(t_i)$ путем выравнивания получен выровненный ряд $\overline{Y}(t_i)$.

Вычитая из данных выровненные значения получаем остатки, случайную составляющую **тренда**

$$e(t_i) = Y(t_i) - \overline{Y}(t_i)$$

Обычно считается, что выравнивание удовлетворительное, если остатки $e(t_i)$ образуют стационарный процесс с нулевым математическим ожиданием.

Если $e(t_i)$ коррелируют между собой, то в модели присутствует автокорреляция остатков. Значительная корреляция остатков сигнализирует о том, что либо кривая подобрана неудачно.

Модель считается адекватной, если остатки:

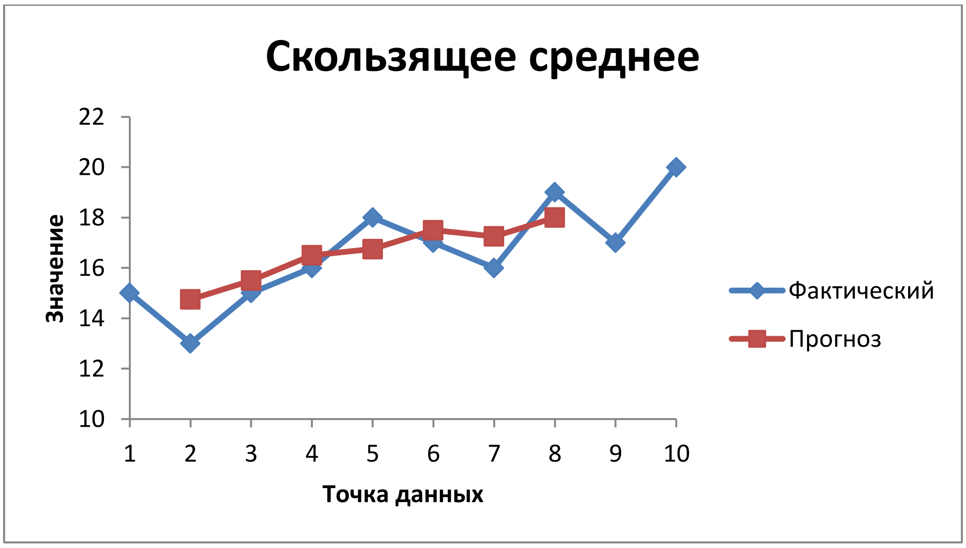
1. являются случайными;
2. распределены по нормальному закону;
3. имеют равное нулю среднее значение $e = 0$;
4. независимы между собой.

Пример выравнивания по методу скользящей средней

Рассчитаем скользящую среднюю об урожайности зерновых культур (ц/га) за 10 лет. Исходные данные и расчетные показатели приведены в таблице.

		скол. средняя		
	Y	3-х летняя	4ч летняя	центрир.ск.ср.
1996	15			
1997	13	14,33	14,75	
1998	15	14,67	15,5	15,13
1999	16	16,33	16,5	16,00
2000	18	17,00	16,75	16,63
2001	17	17,00	17,5	17,13
2002	16	17,33	17,25	17,38
2003	19	17,33	18	17,63
2004	17	18,67		
2005	20			

График исходной и сглаженной зависимости будет иметь вид:



Период скольжения может быть четным и нечетным. Для получения выровненного ряда воспользлзуйтесь средствами Excel. Для этого выбираем **Данные - анализ данных - скользящее среднее**.

В форме определяем входной и выходной интервал данных, интервал сглаживания и условие вывода графика.

9.3 Аналитическое выравнивание временного ряда с использованием аддитивной модели

Общий вид аддитивной модели $\boldsymbol{Y=T+S+e}$

Основные этапы анализа:

1. Выравнивание исходного ряда методом скользящей средней
2. Расчет значений сезонной компоненты S
3. Устранение сезонной компоненты из исходных уровней ряда $(Y\text{-}S)$ и получение выровненных данных $(T+e)$
4. Аналитическое выравнивание уровней $(T+e)$ и расчет значений T с использованием полученного уравнения тренда
5. Расчет по модели полученных значений $(T+S)$
6. Расчет абсолютных ошибок

Рассмотрим методику анализа на примере.

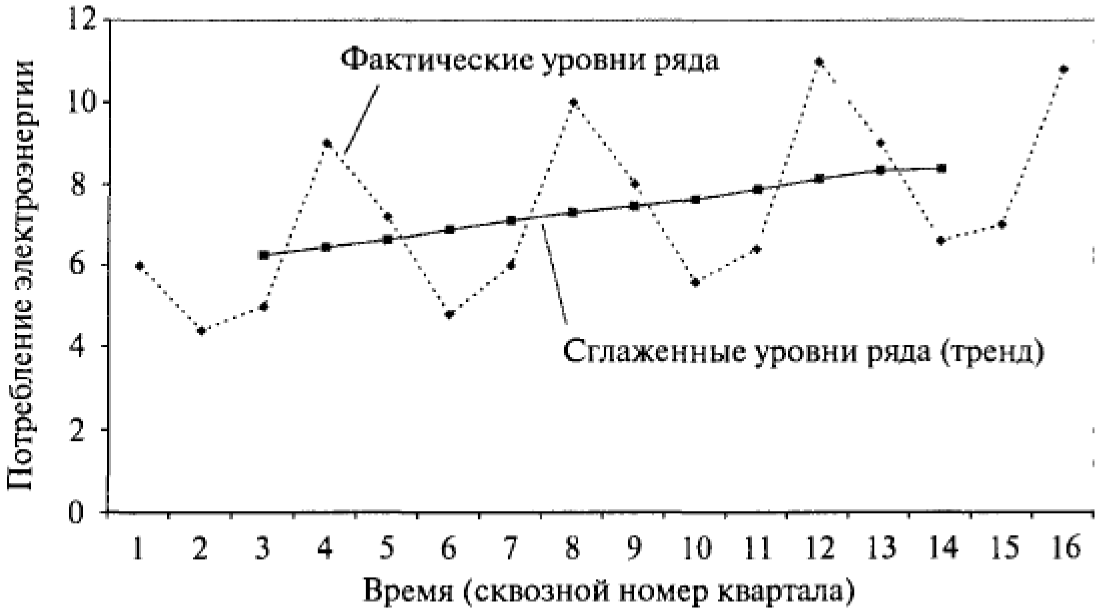
Имеется поквартальные данные о объеме потребления электроэнергии y в некотором районе за четыре года (усл.ед.).

Квартал	**Год**			
	1	**2**	**3**	**4**
I	6,0	7,2	8,0	9,0
II	4,4	4,8	5,6	6,6
III	5,0	6,0	6,4	7,0
IV	9,0	10,0	11,0	10,8

Будем использовать аддитивную модель временного ряда. Проведем выравнивание исходного ряда методом скользящей средней за четыре квартала с процедурой центрирования.

Сквозной номер квартала	Потребление электроэнергии	Скользящая средняя за четыре квартала	Центрированная скользящая средняя	Оценка сезонной вариации
1	6,0	-	-	-
2	4,4	6,10	-	-
3	5,0	6,40	6,250	-1,250
4	9,0	6,50	6,450	2,550
5	7,2	6,75	6,625	0,575
6	4,8	7,00	6,875	-2,075
7	6,0	7,20	7,100	-1,100
8	10,0	7,40	7,300	2,700
9	8,0	7,50	7,450	0,550
10	5,6	7,75	7,625	-2,025
11	6,4	8,00	7,875	-1,475
12	11,0	8,25	8,125	2,875
13	9,0	8,40	8,325	0,675
14	6,6	8,35	8,375	-1,775
15	7,0	-	-	-
16	10,8	-	-	-

Построим графики фактических уровней ряда и центрированных. Оценки сезонной вариации определяются как разность между фактическими уровнями ряда и центрированными скользящими средними.



Расчет сезонной компоненты выполним в следующей таблице, в которой оценки сезонной вариации записываются под соответствующим номером квартала в году.

Показатели	Год	Номер квартала в году				
		I	II	III	IV	
	1	-	-	-1,250	2,550	
	2	0,575	-2,075	-1,100	2,700	
	3	0,550	-2,025	-1,475	2,875	
	4	0,675	-1,775	-	-	
Итого		1,800	-5,875	-3,825	8,125	**Сумма**
Среднее		0,600	-1,958	-1,275	2,708	0,075
Скорректированное S_i		0,581	-1,977	-1,294	2,690	0

В строке **Среднее** рассчитаны средние сезонные вариации по годам за каждый квартал и их сумма равна 0,075. В аддитивной модели предполагается, что сумма всех сезонных компонентпо всем кварталам должна быть равна нулю.

В строке Скорректированное S_i рассчитаны значения сезонных компонет как разность между средней сезонной вариацией и корректирующим коэффициентом 0,075/4 , при этом

$$\sum S_i = 0.$$

Расчет трендовой компоненты и ошибок выполнен в следующей таблице.

t	Y	S	Y-S=T+e	T	Ошибка e
1	6,0	0,581	5,419	5,893	-0,474
2	4,4	-1,977	6,337	6,088	0,256
3	5,0	-1,294	6,294	6,268	0,025
4	9,0	2,690	6,310	6,455	-0,145
5	7,2	0,581	6,619	6,642	-0,023
6	4,8	-1,977	6,777	6,829	-0,052
7	6,0	-1,294	7,294	7,016	0,277
8	10,0	2,690	7,310	7,204	0,106
9	8,0	0,581	7,419	7,391	0,027
10	5,6	-1,977	7,577	7,578	-0,001
11	6,4	-1,294	7,694	7,765	-0,001
12	11,0	2,690	8,310	7,952	0,357
13	9,0	0,581	8,419	8,139	0,279
14	6,6	-1,977	8,577	8,326	0,250
15	7,0	-1,294	8,294	8,514	-0,220
16	10,8	2,690	8,110	8,701	-0,591

В столбце ***Y-S=T+e*** исключается влияние сезонной компоненты. Проводя аналитическое выравнивание ряда ***T+e*** с помощью линейного тренда, получим следующее уравнение линии тренда:

$$T = 5{,}706 + 0{,}187t.$$

Расчет ошибок в аддитивной модели осуществляется по формуле:

$$e = Y-(T+S)$$

Дисперсию фактического ряда и ошибки вычисляем с помощью функции **ДИСП.**
Получаем var(y)=4,196; var(e)=0,0684
Для оценки качества модели используем выражение:

$$1-\frac{var(e)}{var(y)}=1-\frac{0{,}0684}{4{,}196}=0{,}984$$

т.е. аддитивная модель объясняет 98,4% общей вариации уровней исходного временного ряда.

9.4 Аналитическое выравнивание временного ряда с использованием мультипликтивной модели

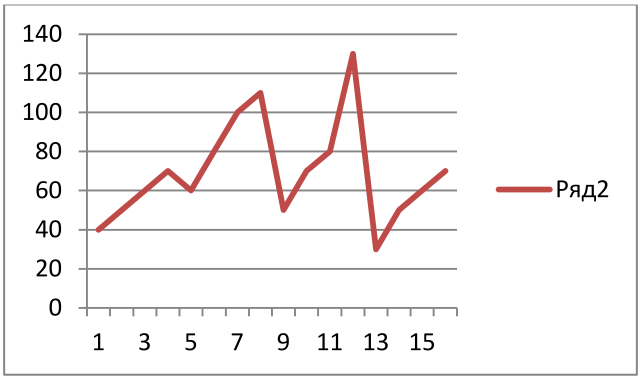
Общий вид аддитивной модели $Y = TSe$
Основные этапы анализа:

1. Выравнивание исходного ряда методом скользящей средней
2. Расчет значений сезонной компоненты *S*
3. Устранение сезонной компоненты из исходных уровней ряда *(Y/S)* и получение выровненных данных *(Te)*
4. Аналитическое выравнивание уровней *(Te)* и расчет значений *T* с использованием полученного уравнения тренда
5. Расчет по модели полученных значений *(TS)*
6. Расчет абсолютных ошибок

Рассмотрим методику анализа на примере.

Квартал	Год			
	1	**2**	**3**	**4**
I	40	60	50	30
II	50	80	70	50
III	60	100	80	60
IV	70	110	130	70

Построим график для определения характера зависимости.



По графику ряда можно установить наличие приблизительно линейного тренда и сезонных колебаний (период равен 4) возрастающей амплитуды, поэтому используется мультипликативная модель. Определим ее компоненты, сезонную и трендовую.

Для исключения влияния сезонной компоненты проведем выравнивание исходного ряда методом скользящей средней за четыре квартала и процедуру центрирования.. Результаты расчетов представлены в таблице.

Сквозной номер квартала	**Размер дивидентов y_i**	**Скользящая средняя за четыре квартала**	**Центрированная скользящая средняя**	**Оценка сезонной вариации**
1	40	-	-	-
2	60	45	-	-
3	50	47,2	46,25	1,081
4	30	52,8	50	0,600
5	50	57,5	55	0,909
6	80	62,5	60	1,333
7	70	65	63,75	1,098
8	50	70	67,5	0,741
9	60	72,5	71,25	0,842
10	100	75	73,75	1,356
11	80	77,5	76,25	1,049
12	60	80	78,75	0,762
13	70	92,5	86,25	0,811
14	110	95	93,75	1,173
15	130	-	-	-
16	70	-	-	-

Оценки сезонной вариации для мультипликативной модели определяются как частное от деления фактических уровней ряда y_i на центрированные скользящие средние.

Расчет сезонной компоненты выполним в следующей расчетной таблице, в которой оценки сезонной вариации записываются под соответствующим номером квартала в году.

Показатели	**Год**	**Номер квартала в году**				
		I	**II**	**III**	**IV**	
	1	-	-	1,081	0,600	
	2	0,909	1,333	1,098	0,741	
	3	0,8842	1,356	1,049	0,762	
	4	0,811	1,173	-	-	
Итого		2,562	3,862	3,228	2,103	**Сумма**
Среднее		0,854	1,287	1,076	0,707	3,918
Скорректированное S_i		0,872	1,314	1,099	0,715	4

В строке **Среднее** рассчитаны средние сезонной вариации по годам за каждый квартал и их сумма, равная 3,918.

В мультипликативной модели предполагается, что сумма всех сезонных компонент по всем кварталам должна быть равна четырем - числу сезонов в году (условие взаимопогашаемости сезонных воздействий).

В строке **Скорректированное** S_i рассчитаны значения сезонных компонент S_i как произведение соответствующей средней сезонной вариации на коректирующий коэффициент 4/3,918=1,021, при этом $\sum S_i = 4.$

Расчет трендовой компоненты и ошибки выполним в следующей таблице.

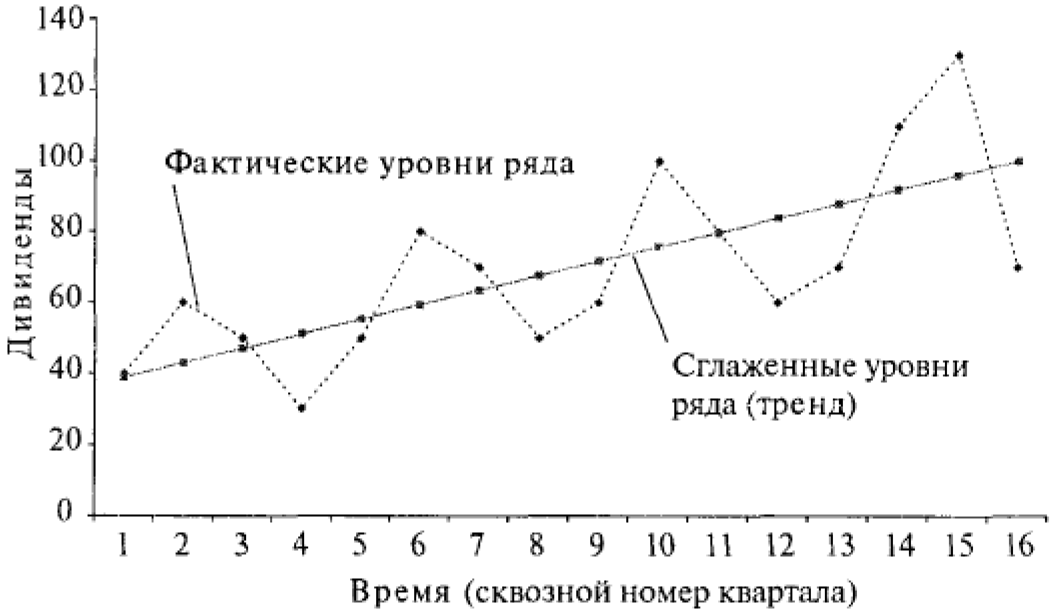
t	**Y**	**S**	**Y/S=Te**	**T**	**e=y/(TS)**	**e=Y-TS**
1	40	0,872	45,871	38,97	1,18	6,02
2	60	1,314	45,662	43,046	1,06	3,44
3	50	1,099	45,496	47,122	0,96	-1,79
4	30	0,715	41,958	51,197	0,82	-6,60
5	50	0,872	57,335	55,273	1,04	1,80
6	80	1,314	60,883	59,349	1,02	2,01
7	70	1,099	63,694	63,425	1,00	0,29
8	50	0,715	69,930	67,501	1,03	1,73
9	60	0,872	68,807	71,577	0,96	-2,41
10	100	1,314	76,103	75,652	1,00	0,59
11	80	1,099	72,793	79,728	0,91	-7,62
12	60	0,715	83,916	83,804	1,00	0,08
13	70	0,872	80,725	87,880	0,91	-6,63
14	110	1,314	83,713	91,956	0,91	-10,83
15	130	1,099	118,289	96,032	1,23	24,46
16	70	0,715	97,902	100,108	0,98	-1,58

Разделив каждый уровень исходного ряда на соответствующее значение сезонной компоненты (столбец Y/S = Te) , исключим влияние сезонной компоненты и в результате получим только тенденцию и случайную компоненту.

Проведя аналитическое выравнивание ряда (*Te*) с помощью линейного тренда, получим следующее уравнение линии тренда:

$$T = 34{,}89 + 4{,}087t$$

Уровни ряда *T* для каждого t =1,2,....,16 указаны в вышеприведенной таблице. График исходного ряда имеет вид.



Расчет ошибки в мультипликативной модели осуществляется по формуле:

$$e = Y/(TS)$$

Чтобы сравнить мультипликативную модель с другими моделями временого ряда, ошибки в мультипликативной модели определяются как

$$e = Y - TS$$

Дисперсии фактического ряда и ошибки, рассчитанные в Excel с помощью функции ДИСПР, составляют: var(y) = 643,36 var(e) = 59,18.

Для оценки качества построенной модели можно по аналогии с аддитивной моделью использовать выражение:

$$1 - \frac{var(e)}{var(y)} = 1 - \frac{58{,}18}{643{,}36} = 0{,}91$$

т.е. мультипликативная модель объясняет 91% общей вариации уровней исходного временного ряда.

10 ДИНАМИЧЕСКИЕ МОДЕЛИ