Простейшая обработка данных. Линейная регрессия. Коэффициент корреляции. Его значимость

Цель: научиться находить коэффициент корреляции и определять его значимость; находить коэффициенты регрессии и строить уравнение регрессии.

Основные сведения

Парная регрессия – это уравнение связи двух переменных y и x:

$$y=f(x),$$

где y – зависимая переменная (результат, отклик);

x – независимая, объясняющая переменная (фактор).

Различают *линейные и нелинейные* регрессии.

Линейная регрессия: $y=a+bx$.

Построение уравнения регрессии сводится к оценке ее параметров. Для оценки параметров регрессий, линейных по параметрам, используют *метод наименьших квадратов* (МНК). МНК позволяет получить такие оценки параметров, при которых сумма квадратов отклонений фактических значений результативного признака y от теоретических y_x минимальна.

Для линейных и нелинейных уравнений, приводимых к линейным, решается следующая система относительно a и b:

$$\begin{cases} na + b\sum x = \sum y, \\ a\sum x + b\sum x^2 = \sum xy. \end{cases}$$

Можно воспользоваться готовыми формулами, которые вытекают из этой системы:

$$b = \frac{\overline{yx} - \bar{y}\cdot\bar{x}}{\overline{x^2} - \bar{x}^2} = \frac{\operatorname{cov}(x,y)}{\operatorname{var}(x)} = \frac{\operatorname{cov}(x,y)}{\sigma_x^2}.$$

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

где

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}, \qquad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{y_1 + y_2 + \ldots + y_n}{n},$$

$$\operatorname{var}(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \overline{x^2} - \bar{x}^2 = \sigma_x^2, \qquad \operatorname{var}(y) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 = \overline{y^2} - \bar{y}^2 = \sigma_y^2,$$

$$\operatorname{cov}(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \overline{xy} - \bar{x}\cdot\bar{y},$$

$$\sigma_y = \sqrt{\frac{\sum_{k=1}^{n}(y_k - \bar{y})^2}{n}} = \sqrt{\operatorname{var}(y)}, \qquad \sigma_x = \sqrt{\frac{\sum_{k=1}^{n}(x_k - \bar{x})^2}{n}} = \sqrt{\operatorname{var}(x)}.$$

Коэффициент регрессии b показывает абсолютную силу связи между вариацией x и вариацией y.

Тесноту связи изучаемых явлений оценивает линейный коэффициент парной корреляции r_{xy} для линейной регрессии ($-1 \le r_{xy} \le 1$):

$$r_{xy} = \frac{\overline{yx} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}} = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\,\mathrm{var}(y)}} = в \frac{\sigma_x}{\sigma_y}.$$

Теснота линейной связи между переменными может быть оценена на основании шкалы Чеддока:

Теснота связи	Значение коэффициента корреляции при наличии:	
	Прямой связи	Обратной связи
Слабая	0,1–0,3	(–0,3)–(–0,1)
Умеренная	0,3–0,5	(–0,5)–(–0,3)
Заметная	0,5–0,7	(–0,7)–(–0,5)
Высокая	0,7–0,9	(–0,9)–(–0,7)
Весьма высокая	0,9–1	(–1)–(–0,9)

Положительное значение коэффициента корреляции говорит о положительной связи между *x* и *y*, когда с ростом одной из переменных другая тоже растет. Отрицательное значение коэффициента корреляции означает, с ростом одной из переменных другая убывает, с убыванием одной из переменной другая растет.

Оценку статистической значимости коэффициента корреляции проводят с помощью *t*-критерия Стьюдента. Выдвигают гипотезу H_0 о статистически незначимом отличии коэффициента от нуля. Оценка значимости коэффициента корреляции с помощью *t*-критерия Стьюдента проводится путем сопоставления его значения с величиной случайной ошибки:

$$t_r = r/m_r.$$

Стандартная (случайная) ошибка коэффициента корреляции определяется по формуле:

$$ò_r = \sqrt{\frac{1-r^2}{n-2}};$$

Сравнивая фактическое и табличное (критическое) значения *t*-статистики – $t_{табл}$ и $t_{факт}$ – принимает или отвергаем гипотезу H_0.

Если $t_{табл} < t_{факт}$, то гипотеза H_0 отклоняется, коэффициент корреляции не случайно отличается от. Если $t_{табл} > t_{факт}$, то гипотеза H_0 не отклоняется и признается случайная природа формирования коэффициента корреляции.

Порядок выполнения работы.

По заданной выборке исследовать зависимость результата y от фактора x. Для этого

1. Создать таблицу данных.
2. Найти средние значения $\bar{x}, \bar{y}$, выборочные дисперсии S_x^2, S_y^2, исправленные средние квадратические отклонения $\bar{S}_x, \bar{S}_y$.
3. Найти коэффициент корреляции и проверить его значимость.
4. Найти коэффициенты линейного уравнения регрессии.
5. Построить график прямой регрессии.

Пример выполнения лабораторной работы.

В табл. 1.1 приведены данные об объеме производства y (тыс.ед.) в зависимости от численности занятых x (тыс.чел.) некоторой фирмы.

Таблица 1.1.

Исходные данные

x	11	13	15	18	20	22	24	25	27
y	15	17	21	20	28	33	34	32	29

1. В диапазоне B3:C11 подготовим исходные данные.
2. Вводим следующие формулы:

Ячейка	Формула	Примечание
D3	=B3*C3	Копируем в диапазон D3:D11
E3	=B3*B3	Копируем в диапазон E3:E11
F3	=C3*C3	Копируем в диапазон F3:F11
B12	=СРЗНАЧ(B3:B11)	Копируем в диапазон B12:F12
A17	=E12-B12*B12	Выборочная средняя фактора
B17	=F12-C12*C12	Выборочная средняя результата
A20	=СТАНДОТКЛОН(B3:B11)	Исправленное среднее квадратическое отклонение фактора
B20	=СТАНДОТКЛОН(C3:C11)	Исправленное среднее квадратическое отклонение результата

Получим следующие результаты (см. рис. 1.1).

	A	B	C	D	E	F
1	**Простейшая обработка данных**					
2		x	y	xy	x^2	y^2
3	1	11	25	275	121	625
4	2	13	27	351	169	729
5	3	15	31	465	225	961
6	4	18	30	540	324	900
7	5	20	38	760	400	1444
8	6	22	43	946	484	1849
9	7	24	44	1056	576	1936
10	8	25	42	1050	625	1764
11	9	27	49	1323	729	2401
12	среднее значение	19,44	36,56	751,78	405,89	1401,00
13						
14						
15	**Выборочные средние**					
16	S_x^2	S_y^2				
17	27,80	64,69				
18	**Исправленные средние квадратические отклонения**					
19	$S_{хиспр}$	$S_{уиспр}$				
20	5,59	8,53				
21						

Рис. 1.1. Результаты простейшей обработки данных

3. Для определения коэффициента корреляции воспользуемся формулой $r_{xy} = \dfrac{\overline{yx} - \bar{x} \cdot \bar{y}}{\sqrt{(\overline{x^2} - \bar{x}^2)(\overline{y^2} - \bar{y}^2)}}$. Для этого в ячейку **E16** вводим формулу **=(D12-B12*C12)/КОРЕНЬ(A17*B17).**

Из расчетов следует, что коэффициент корреляции *r*=0,97. Это свидетельствует о том, что связь между объемом выпуска продукции и численностью занятых весьма высокая и положительная.

4. Для проверки значимости коэффициента корреляции введем вспомогательные данные:

Ячейки

К16 9 число предприятий;

К17 0,05 уровень значимости.

5. Далее вводим следующие формулы:

H19	=КОРЕНЬ((1-E16*E16)/(K16-2))	Стандартная ошибка
H20	=E16/H19	*t*-статистика
H21	=СТЬЮДРАСПОБР(K17;K16-2)	Критическое значение *t*-статистики

H22	=ЕСЛИ(ABS(H20)>H21;"Значим";"Незначим")	Вывод

Таким образом, получим данные, представленные на рис. 1.2.

	C	D	E	F	G	H	I	J	K
13									
14									
15		**Коэффициент корреляции**					**Вспомогательные данные**		
16		r	0,97				n		9
17							уровень значимости		0,05
18		**Проверка значимости коэффициента корреляции**							
19		стандаратная ошибка				0,10			
20		t-статистика				9,91			
21		Критическое значение t-статистики				2,36			
22		Вывод				Значим			
23									

Рис. 1.2. Анализ значимости коэффициента корреляции

6. Для определения коэффициентов уравнения линейной регрессии на основе формул

$$в = \frac{\overline{yx} - \bar{y} \cdot \bar{x}}{\overline{x^2} - \bar{x}^2};\ a = \bar{y} - в\bar{x},$$

следует в ячейки **I3**, **I4** ввести соответственно следующие формулы:

=(D12-B12*C12)/A17;

=C12-I3*B12.

Уравнение регрессии y=7,9+1,47x.

Значение коэффициента b=1,47 говорит о том, что при увеличении численности занятых на 1 тыс.чел. объем продукции увеличится на 1,74 тыс.ед.

Результаты расчетов приведены на рис.1.3.

	A	B	C	D	E	F	G	H	I	J	K
1	**Простейшая обработка данных**										
2		x	y	xy	x^2	y^2		**Коэффициенты регрессии**			
3	1	11	25	275	121	625		в	1,47		
4	2	13	27	351	169	729		a	7,90		
5	3	15	31	465	225	961					
6	4	18	30	540	324	900					
7	5	20	38	760	400	1444					
8	6	22	43	946	484	1849					
9	7	24	44	1056	576	1936					
10	8	25	42	1050	625	1764					
11	9	27	49	1323	729	2401					
12	среднее значение	19,44	36,56	751,78	405,89	1401,00					
13											
14											
15	**Выборочные средние**			**Коэффициент корреляции**					**Вспомогательные данные**		
16	S_x^2	S_y^2		r	0,97				n		9
17	27,80	64,69							уровень значимости		0,05
18	**Исправленные средние квадратические отклонения**			**Проверка значимости коэффициента корреляции**							
19	$S_{хиспр}$	$S_{уиспр}$		стандаратная ошибка				0,10			
20	5,59	8,53		t-статистика				9,91			
21				Критическое значение t-статистики				2,36			
22				Вывод				Значим			

Рис. 1.3. Результаты расчетов

7. Для построения графика выделим диапазон B3:C11. Вызовем **Мастер диаграмм**. Чтобы ось отражала фактические данные, выберем тип диаграммы **Точечная.** После чего нажмем кнопку **Готово**. На построенной диаграмме выделим график функции, щелкнув по нему левой кнопкой мыши. Выделение обозначается светлыми маркерами на функции. Нажав правую кнопку мыши, выведем контекстно-зависимое меню, в котором выберем опцию **Добавить линию тренда**. В окне **Линия тренда** по вкладке **Тип** выберем тип функции **Линейная**, а во вкладке **Параметры** – установим флажок **показывать уравнение на диаграмме**. В результате на диаграмме появиться вид теоретической кривой – тренда и ее уравнение (рис.1.4).

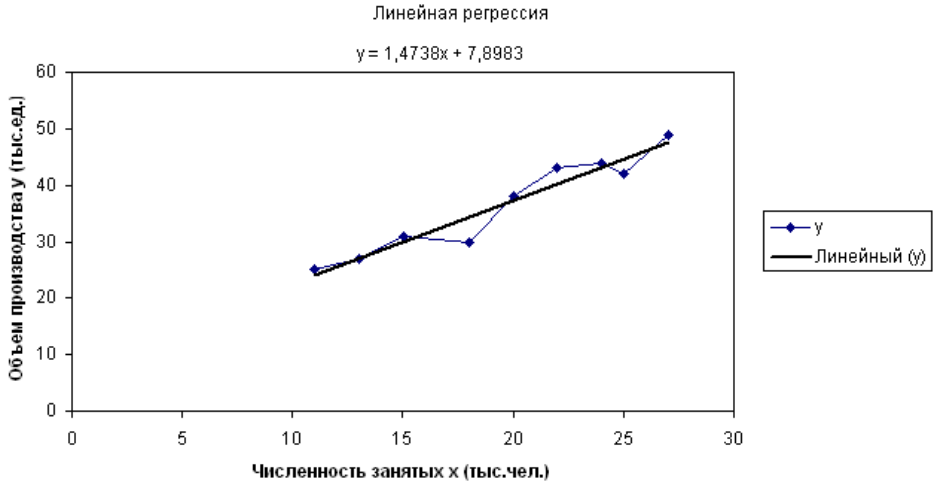

Рис. 1.4. Графики фактических данных и построенной регрессии

8. Вычисление параметров регрессии с помощью статистических функций Excel:

КОРРЕЛ(массив1;массив2) вычисляет коэффициент корреляции между двумя переменными; значения первой из них приведены в диапазоне массив1, значения второй – в диапазоне массив2;

НАКЛОН(известные_значения_y;известные_значения_x) служит для определения коэффициента *b*;

ОТРЕЗОК(известные_значения_y;известные_значения_x) служит для определения коэффициента *a*.

Вводим формулы:

C27	=КОРРЕЛ(B3:B11;C3:C11)	Коэффициент корреляции
C28	=НАКЛОН(C3:C11;B3:B11)	Коэффициент *b*
C29	=ОТРЕЗОК(C3:C11;B3:B11)	Коэффициент *a*

Встроенная статистическая функция **ЛИНЕЙН** определяет параметры линейной регрессии. Порядок вычислений следующий:

1) выделите область пустых ячеек 5x2 (5 строк, 2 столбца) с целью вывода результатов регрессионной статистики (A27:B3);

2) в главном меню выберите **Вставка/Функция**;

3) в строке **Категория** (рис.1.5) выберите **Статистические**, в окне **Функция – ЛИНЕЙН**. Щелкните **ОК**.

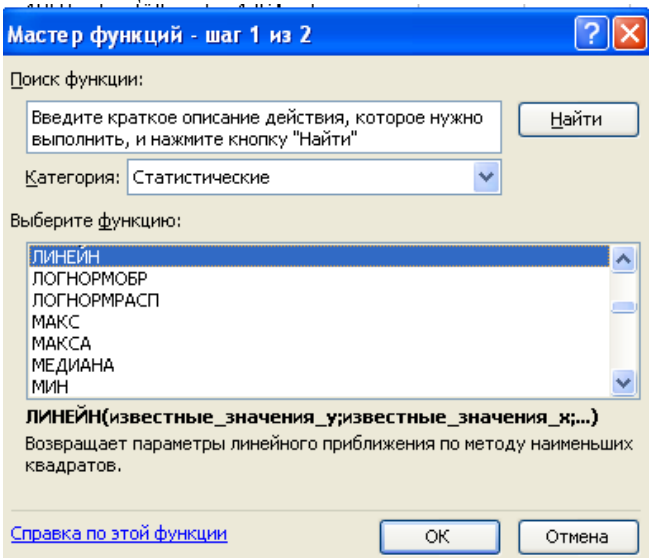

Рис. 1.5. Диалоговое окно «Мастер функций»

4) Заполните аргументы функции (рис.1.6.):

Известные_значения_y – диапазон, содержащий данные результативного признака;

Известные_значения_x – диапазон, содержащий данные факторов независимого признака;

Константа – логическое значение, которое указывает на наличие или на отсутствие свободного члена в уравнении; если *Константа* = 1, то свободный член рассчитывается обычным образом, если *Константа* = 0, то свободный член равен 0.

Статистика – логическое значение, которое указывает выводить дополнительную информацию по регрессионному анализу или нет. Если *Статистика* = 1, то дополнительная информация выводится, если *Статистика* = 0, то выводится только оценки параметров уравнения. Далее **ОК**.

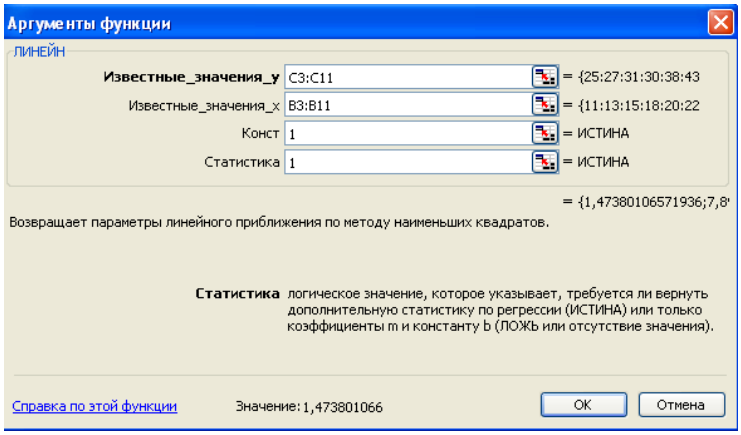

Рис.1.6. Диалоговое окно ввода аргументов функции **ЛИНЕЙН**

5) В левой верхней ячейке выделенной области появится первый элемент итоговой таблицы. Чтобы раскрыть всю таблицу, нажмите на

клавишу **F2,** а затем – на комбинацию клавиш **CTRL**+**SHIFT**+**ENTER**. Дополнительная регрессионная статистика будет выводиться в порядке, указанном в следующей схеме:

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.

Результаты регрессионного анализа представлены на рис.1.7.

	A	B	C	D	E	F
25						
26	**Линейн**					
27	1,4738011	7,89831261	0,97	коэффициент корреляции		
28	0,1486756	2,99532023	1,47	b		
29	0,9335011	2,35181208	7,90	a		
30	98,264891	7				
31	543,50508	38,7171403				

Рис. 1.7.Результаты регрессионного анализа