Простейшая обработка данных. Линейная регрессия. Коэффициент корреляции.

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: $y=a+bx$,

$$b=\frac{\overline{yx}-\overline{y}\cdot\overline{x}}{\overline{x^2}-\overline{x}^2}$$

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

где

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

$$\overline{y}=\frac{1}{n}\sum_{i=1}^{n}y_i=\frac{y_1+y_2+\ldots+y_n}{n},$$

$$\overline{xy}=\frac{1}{n}\sum_{i=1}^{n}x_iy_i=\frac{x_1y_1+x_2y_2+\ldots+x_ny_n}{n},$$

$$\overline{x^2}=\frac{1}{n}\sum_{i=1}^{n}x_i^2=\frac{x_1^2+x_2^2+\ldots+x_n^2}{n}.$$

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 ($-1\le r_{xy}\le 1$):

$$r_{xy}=\frac{\overline{yx}-\overline{x}\cdot\overline{y}}{\sqrt{(\overline{x^2}-\overline{x}^2)(\overline{y^2}-\overline{y}^2)}}.$$

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.

Порядок выполнения работы.

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

1. Создать таблицу данных.
2. Построить поле корреляции.
3. Найти коэффициенты линейного уравнения регрессии.
4. Найти коэффициенты корреляции и детерминации.
5. Построить график прямой регрессии.

Пример выполнения лабораторной работы.

В табл. 1. приведены данные об объеме производства *y* (тыс.ед.) в зависимости от численности занятых *x* (тыс.чел.) некоторой фирмы.

Таблица 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

Получим следующие результаты (см. рис. 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

Рис. 1. Результаты простейшей обработки данных

3. Для построения поля корреляции выделим диапазон B3:C11. Вызовем **Мастер диаграмм**. Чтобы ось отражала фактические данные,

выберем тип диаграммы **Точечная.** После чего нажмем кнопку **Готово**.

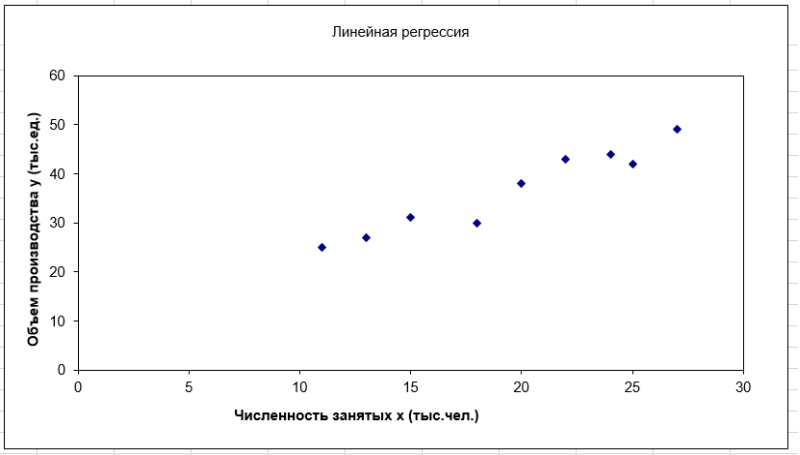

Рис. 2. Поле корреляции

4. Для определения коэффициентов уравнения линейной регрессии на основе формул

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

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

=(D12-B12*C12)/(E12-B12^2);

=C12-I3*B12.

Уравнение регрессии y=7,9+1,47x.

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

Результаты расчетов приведены на рис.3.

	A	B	C	D	E	F	G	H	I	J
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		а	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				

Рис. 3. Результаты расчетов

5. Для определения коэффициента корреляции воспользуемся формулой $r_{xy} = \frac{\overline{yx} - \overline{x} \cdot \overline{y}}{\sqrt{(\overline{x^2} - \overline{x}^2)(\overline{y^2} - \overline{y}^2)}}$. Для этого в ячейку **E16** вводим формулу

=(D12-B12*C12)/КОРЕНЬ((E12-B12^2)*(F12-C12^2)).

Из расчетов следует, что коэффициент корреляции r=0,97.

Таким образом, получим данные, представленные на рис. 4.

	C	D	E	F
13				
14				
15		Коэффициент корреляции		
16		r	0,97	
17				

Рис. 4. Коэффициент корреляции

Вывод: По шкале Чеддока - связь между объемом выпуска продукции и численностью занятых прямая и весьма высокая.

Коэффициент детерминации равен $R^2=r^2=0,97^2=0,94$.

Вывод: Уравнением регрессии объясняется 94% дисперсии результативного признака, а на долю случайных факторов приходится 6%.

6. Для построения графика линейной регрессии 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.5).

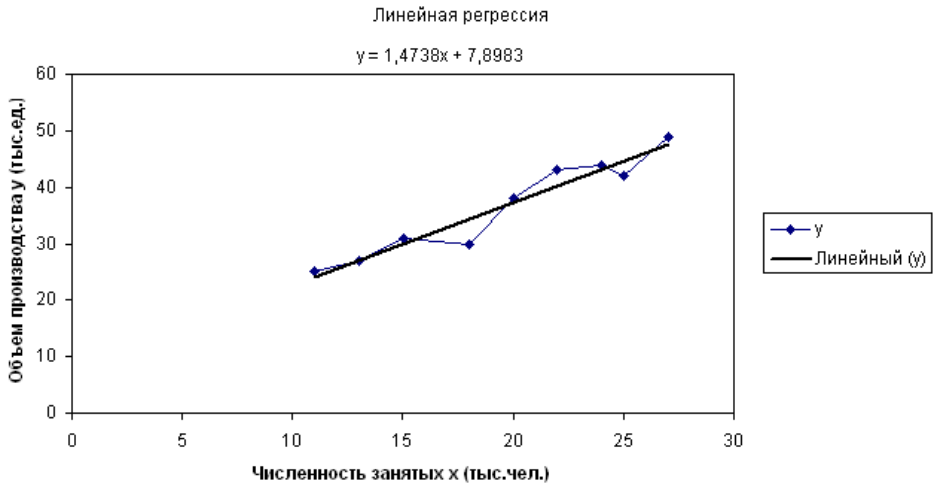

Рис. 5. Графики фактических данных и построенной регрессии

7. 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.6) выберите **Статистические**, в окне **Функция** – **ЛИНЕЙН**. Щелкните **ОК**.

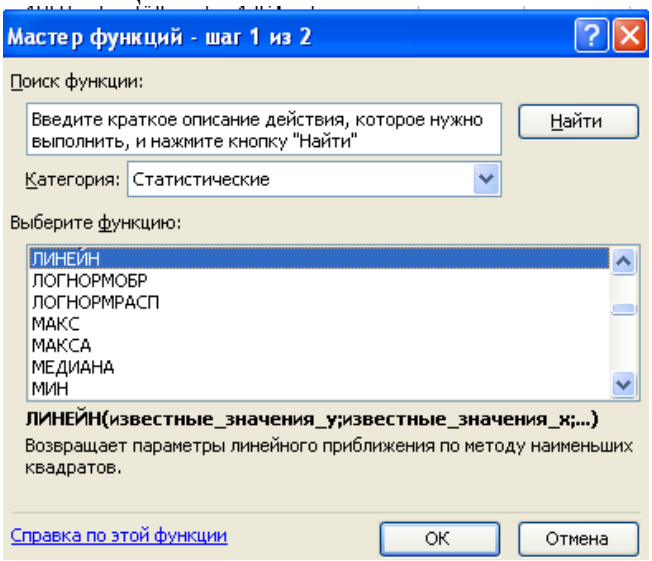

Рис. 6. Диалоговое окно «Мастер функций»

4) Заполните аргументы функции (рис.7):

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**.

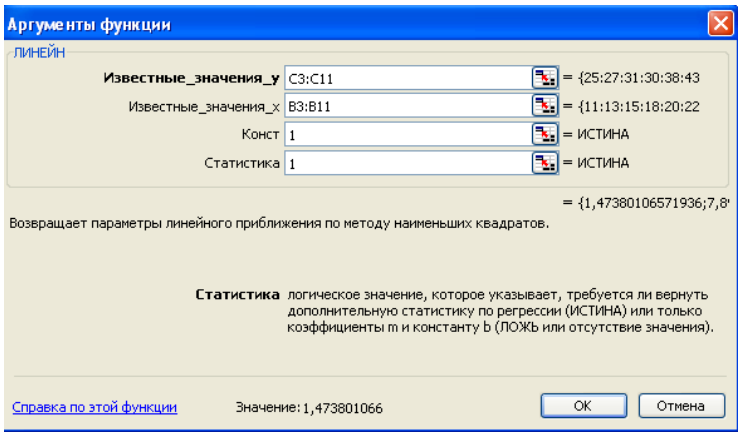

Рис.7. Диалоговое окно ввода аргументов функции **ЛИНЕЙН**

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.8.

	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. 8. Результаты регрессионного анализа

Варианты заданий

Вариант 1

Исследование зависимости между среднемесячными доходами (x) на семью и расходами (y) на покупку кондитерских изделий представлены в таблице:

Семья	1	2	3	4	5	6	7
Доход семьи, тыс. руб.	48	38	54	42	34	46	34
Расходы на кондитерские изделия, руб.	750	680	780	710	640	730	660

Вариант 2

По семи регионам приводятся следующие данные:

№ региона	1	2	3	4	5	6	7

Среднедушевой прожиточный минимум в день одного трудоспособного, у.е., x	78	82	87	79	89	106	67
Среднедневная заработная плата, у.е., y	133	148	134	154	162	195	139

Вариант 3. Взаимосвязь между ценой спроса (x) и ценой предложения (y) наиболее ликвидных на внебиржевом рынке акций характеризуется следующими данными (см. табл.):

Ценная бумага	БМП	ГУМ	ЕЭС	ЗИЛ	КаОк	Лукойл	ТНК
Цена спроса	34,1	33,6	30,3	13,5	13,9	26,5	18,1
Цена предложения	60,6	40,7	33,8	22,1	30,0	34,5	20,9

Вариант 4

В таблице приведены данные о темпе прироста внутреннего национального продукта (y , %) и промышленного производства (x , %) семи развитых стран мира за 1992 г.

Страна	Дания	США	Германия	Франция	Италия	Канада	Австралия
Промышленное производство, (%)	4,3	4,6	2,0	3,1	3,0	3,4	2,6
Темп прироста, (%)	3,5	3,1	2,2	2,7	2,7	3,1	1,8

Вариант 5

По семи регионам приводятся следующие данные:

№ региона	1	2	3	4	5	6	7
Среднедушевой прожиточный минимум в день одного трудоспособного, у.е.., x	81	77	85	79	93	100	72
Среднедневная заработная плата, у.е., y	124	131	146	139	143	159	135

Задача 6

По семи регионам приводятся следующие данные:

№ региона	1	2	3	4	5	6	7
Среднедушевой прожиточный минимум в день одного трудоспособного, у.е.., x	74	81	90	79	89	87	77

Среднедневная заработная плата, у.е., у	122	134	136	125	120	127	125

Вариант 7

Взаимосвязь между производительностью труда (y) и энерговооруженностью труда (x) (в расчете на одного работника) для семи предприятий характеризуется следующими данными:

Предприятие	1	2	3	4	5	6	7
Энерговооруженность труда, кВт	2,8	2,2	3,0	3,5	3,2	3,7	4,0
Производительность труда, тыс. руб.	6,7	6,9	7,2	7,3	8,4	8,8	9,1

Вариант 8

С целью анализа взаимного влияния зарплаты и текучести рабочей силы на семи однотипных фирмах с одинаковым числом работников проведены измерения уровня месячной зарплаты (x) и числа уволившихся за год рабочих (y):

Фирма	1	2	3	4	5	6	7
Уровень месячной зарплаты, $	100	150	200	250	300	350	400
Кол-во уволившихся за год, чел.	60	35	20	20	15	10	4

Вариант 9

Провели исследование, сколько сберегает население (y) и сколько оно зарабатывает за год (x). Были получены следующие данные для случайно отобранных семи человек:

Граждане	1	2	3	4	5	6	7
Доход, тыс. руб	15	6	9	3	20	11	14
Сбережения, руб.	2000	200	500	100	2500	1800	1500

Вариант 10

В таблице приведены статистические данные, описывающие зависимость спроса на товар (y) от его цены(x):

№	1	2	3	4	5	6	7
Цена товара, руб.	99	82	77	69	52	44	31
Спрос на товар, шт.	100	115	210	270	323	478	544

Контрольные вопросы

1. Сущность и задачи корреляционного анализа.
2. Меры связи в метрических и неметрических шкалах.
3. Парная корреляция.
4. Корреляционное поле.
5. Статистическая значимость корреляционной связи.
6. Получение регрессионных моделей в решении задач оптимизации функционирования социально-экономических систем разного уровня.
7. Проверка обоснованности регрессионной модели.
8. Адекватность модели.

9. Коэффициент детерминации.
10. Компьютерные методы поиска решения задач