ЛАБОРАТОРНАЯ РАБОТА № 3
ПОСТРОЕНИЕ И ПРОВЕРКА АДЕКВАТНОСТИ МОДЕЛИ ЛИНЕЙНОЙ РЕГРЕССИИ

Цель: Построить уравнение линейной регрессии, освоить инструменты MS EXCEL, применяющиеся при построении линейных регрессионных моделей.

Задачи:

1. Определить зависимость y от x;
2. построить корреляционные поля и график уравнения линейной регрессии y на x, то есть $\hat{y} = a + b \cdot x$;
3. сделать вывод о качестве модели и рассчитать прогнозное значение y при прогнозном значении x, составляющем 107% от среднего уровня.
4. ответить на контрольные вопросы.

Основные сведения

Можно указать два варианта рассмотрения взаимосвязей между двумя переменными X и Y. В первом случае обе переменные считаются равноценными в том смысле, что они не подразделяются на первичную и вторичную (независимую и зависимую) переменные. Основным в этом случае является вопрос о наличии и силе взаимосвязи между этими переменными. Например, между ценой товара и объемом спроса на него, между урожаем картофеля и урожаем зерна. При исследовании силы линейной зависимости между такими переменными обращаются к корреляционному анализу, основной мерой которого является коэффициент корреляции. Вполне вероятно, что связь в этом случае вообще не носит направленного характера. Например, урожайность картофеля и зерновых обычно изменяется в одном и том же направлении, однако очевидно, что ни одна из этих переменных не является определяющей.

Другой вариант рассмотрения взаимосвязей выделяет одну из величин как независимую (объясняющую), а другую как зависимую (объясняемую). В этом случае изменение первой из них может служить причиной для изменения

другой. Например, рост дохода ведет к увеличению потребления; рост цены — к снижению спроса; снижение процентной ставки увеличивает инвестиции. Однако такая зависимость не является однозначной в том смысле, что каждому конкретному значению объясняющей переменной (набору объясняющих переменных) может соответствовать не одно, а множество значений из некоторой области. Другими словами, каждому конкретному значению объясняющей переменной соответствует некоторое вероятностное распределение зависимой переменной (рассматриваемой как случайная величина). Поэтому анализируют, как объясняющая переменная влияет на зависимую переменную «в среднем». Зависимость такого типа, выражаемая соотношением

$$M(Y|x) = f(x) \qquad (1)$$

называется функцией регрессии Y на X. При этом X называется независимой (объясняющей) переменной (регрессором), Y – зависимой (объясняемой) переменной. При рассмотрении зависимости двух СВ говорят о парной регрессии.

В настоящее время под регрессией понимается функциональная зависимость между объясняющими переменными и условным математическим ожиданием (средним значением) зависимой переменной, которая строится с целью предсказания (прогнозирования) этого среднего значения при фиксированных значениях первых.

Для отражения того факта, что реальные значения зависимой переменной не всегда совпадают с ее условными математическими ожиданиями и могут быть различными при одном и том же значении объясняющей переменной (наборе объясняющих переменных), фактическая зависимость должна быть дополнена некоторым слагаемым ε, которое, по существу, является случайной величиной. Из этого следует, что связи между зависимой и объясняющей(ими) переменными выражаются соотношением

$$Y = M(Y|x) + \varepsilon. \qquad (2)$$

Порядок выполнения лабораторной работы

Для выполнения данной работы студенту необходимо выбрать свой вариант из таблицы 1.2. Номер варианта определяется согласно номеру студента по списку (уточняется у преподавателя). Рассмотрим пример. По 15 предприятиям отрасли (таблица 1.1) известны: x – объем произведенной продукции (тыс. ед.) и y – затраты на выпуск этой продукции (тыс. ден. ед.).

Таблица 1.1 – Исходные данные

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x	2,7	1,5	8,2	4,5	3,3	5,8	3,0	7,1	1,2	10,4	4,9	5,2	11,5	9,4	6,5
y	110	70	310	120	75	170	100	180	30	440	190	150	390	310	230

1. В Excel составим вспомогательную таблицу (рисунок 1)

	A	B	C	D	E	F
1		x	y	xy	x^2	y^2
2	1	2,7	110	297	7,29	12100
3	2	1,5	70	105	2,25	4900
4	3	8,2	310	2542	67,24	96100
5	4	4,5	120	540	20,25	14400
6	5	3,3	75	247,5	10,89	5625

Рисунок 1 Вспомогательная таблица

Для вычисления выборочных средних используем формулу

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}.$$

Выборочные средние для $\bar{x} = 5{,}68$; $\bar{y} = 191{,}67$ показывают, что средний объем произведенной продукции по 15 предприятиям отрасли составляет 5,68 тыс. ед., а средние затраты на выпуск этой продукции – 191,67 тыс. ден. ед.

Для вычисления выборочной ковариации между x и y используем формулу $Cov(x, y) = \overline{xy} - \bar{y} \cdot \bar{x}$ и получим 345,5.

Выборочную дисперсию для x найдем по $Var(x) = \overline{x^2} - \bar{x}^2$ и получим 9,37 (рисунок 1.1). Аналогично определяем $Var(y) = 13838{,}89$.

Выборочный коэффициент корреляции рассчитывается по формуле

$$r_{xy} = \frac{Cov(x,y)}{\sqrt{Var(x)}\sqrt{Var(y)}}.$$

Коэффициент корреляции $r_{xy} = 0{,}96$ очень высокий, что указывает на прямую и весьма сильную связь между x и y, т. е. с ростом объема произведенной продукции (х) увеличиваются затраты на выпуск этой продукции (у). Проверку необходимо осуществлять с помощью статистических функций **КОВАРИАЦИЯ.Г**, **ДИСП.Г**, **КОРРЕЛ**.

Выборочный коэффициент линейной регрессии

$$b = \frac{Cov(x,y)}{Var(x)} = 36{,}87;$$

параметр $a = \bar{y} - b \cdot \bar{x} = -17{,}78$. Значит, уравнение парной линейной регрессии имеет вид $\boldsymbol{\hat{y} = -17{,}78 + 36{,}87 \cdot x}$.

Коэффициент b показывает, что при увеличении объема произведенной продукции (х) на 1 тыс. ед. затраты на выпуск этой продукции (у) в среднем увеличатся на 36,87 тыс. ден. ед. 2 Подставляя в найденное уравнение регрессии фактические значения x, определим теоретические (расчетные) значения $\hat{y}$.

В результате получим следующие данные (рисунок 2)

	A	B	C	D	E	F	G
1		x	y	xy	x^2	y^2	$\hat{y}$
2	1	2,7	110	297	7,29	12100	81,78
3	2	1,5	70	105	2,25	4900	37,53
4	3	8,2	310	2542	67,24	96100	284,59
5	4	4,5	120	540	20,25	14400	148,15
6	5	3,3	75	247,5	10,89	5625	103,91
7	6	5,8	170	986	33,64	28900	196,09
8	7	3	100	300	9	10000	92,84
9	8	7,1	180	1278	50,41	32400	244,03
10	9	1,2	30	36	1,44	900	26,47
11	10	10,4	440	4576	108,16	193600	365,71
12	11	4,9	190	931	24,01	36100	162,90
13	12	5,2	150	780	27,04	22500	173,97
14	13	11,5	390	4485	132,25	152100	406,28
15	14	9,4	310	2914	88,36	96100	328,84
16	15	6,5	230	1495	42,25	52900	221,90
17	**сумма**	85,2	2875	21512,5	624,48	758625	2875
18	**среднее**	5,68	191,67	1434,17	41,632	50575	191,667
19	**n**	15					
20							
21			Проверка				
22	Cov(x,y)	345,5	345,5				
23	Var(x)	9,37	9,37				
24	Var(y)	13838,9	13838,9				
25	r_{xy}	0,96	0,96				
26	b	36,87					
27	a	-17,78					
28							

Рисунок 2 Итоговая таблица

2. С помощью Мастера диаграмм строим поле корреляции (выделяя столбцы со значениями x и y) и уравнение линейной регрессии (выделяя столбцы со значениями x и $\hat{y}$). Выбираем тип диаграммы – Точечная и, следуя рекомендациям Мастера диаграмм, вводим нужные параметры (название, подписи к осям, легенду и т. п.). В результате получим рисунок 3.

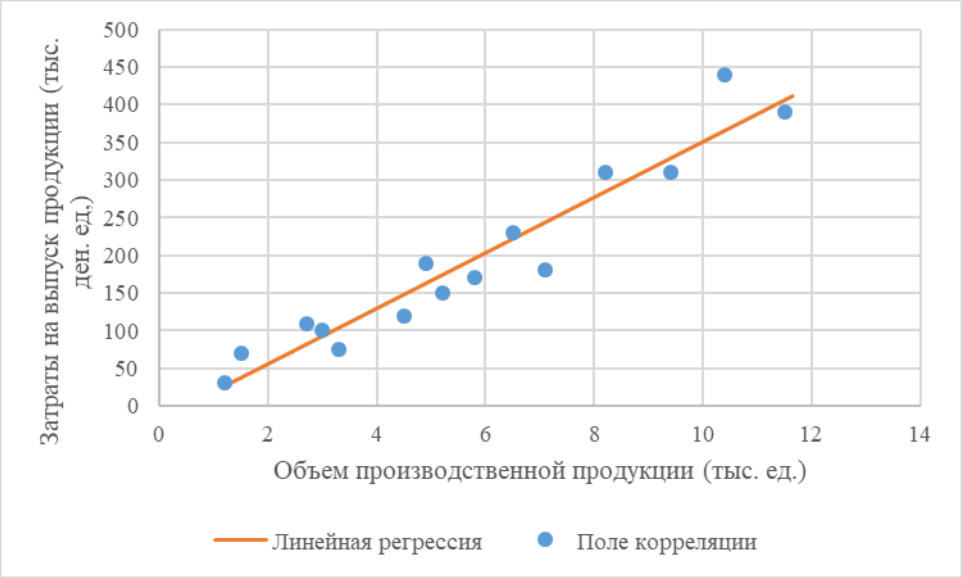

Рисунок 3 График зависимости объема произведенной продукции от теоретических и фактических затрат на выпуск этой продукции.

3. Для оценки качества построенной модели регрессии вычислим

- коэффициент детерминации $R^2 = r_{xy}^2 = 0{,}92$ который показывает, что изменение затрат на выпуск продукции (y) на 92% объясняется изменением объема произведенной продукции (x), а 8% приходится на долю неучтенных в модели факторов, что указывает на хорошее качество построенной регрессионной модели;

- среднюю ошибку аппроксимации. Для этого, продолжая вспомогательную таблицу вычислим $\varepsilon = y - \hat{y}$ и $\left|\frac{y-\hat{y}}{y}\right|$ (рисунок 4).

	A	B	C	D	E	F	G	H	I
1		x	y	xy	x^2	y^2	$\hat{y}$	$y-\hat{y}$	$\left\|\frac{y-\hat{y}}{y}\right\|$
2	1	2,7	110	297	7,29	12100	81,78	28,22	0,26
3	2	1,5	70	105	2,25	4900	37,53	32,47	0,46
4	3	8,2	310	2542	67,24	96100	284,59	25,41	0,08
5	4	4,5	120	540	20,25	14400	148,15	-28,15	0,23

Рисунок 4 Вспомогательная таблица (продолжение)

Среднее значение по элементам последнего столбца, умноженного на 100%, дает $\bar{A} = 18{,}2\%$. Следовательно, в среднем теоретические значения $\hat{y}$ отклоняются от фактических y на 18,2%.

4. С помощью F-критерия Фишера оценим значимость уравнения регрессии в целом:

$$F_{fact} = \frac{R^2}{1 - R^2}(n - 2) = 150{,}74.$$

На уровне значимости $\alpha = 0{,}05$ $F_{tabl} = 4{,}67$. Табличное значение F-критерия находится с помощью встроенной функции **F.ОБР.ПХ.** В качестве аргументов вводим уровень значимости, число степеней свободы 1=1, число степеней свободы 2=$(n - 2)$.

Так как $F_{fact} > F_{tabl}$, то уравнение регрессии значимо.

5. С помощью t-критерия Стьюдента оценим значимость параметров a, b уравнения регрессии. Для этого рассчитаем стандартные ошибки коэффициентов a, b по формулам

$$m_a = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2}{n - 2} \cdot \frac{\sum x_i^2}{n \cdot \sum(x_i - \bar{x})^2}};$$

$$m_b = \sqrt{\frac{\sum(y_i - \hat{y}_i)^2 / (n - 2)}{\sum(x_i - \bar{x})^2}}.$$

И вычислим фактические значения t-критерия

$$t_a = \frac{a}{m_a}; t_b = \frac{b}{m_b}.$$

На уровне значимости $\alpha = 0{,}05$ $t_{tabl} = 2{,}16$. Табличное значение t-критерия находится с помощью встроенной функции СТЬЮДЕНТ.ОБР.2Х. В качестве аргументов вводим уровень значимости, число степеней свободы =$(n - 2)$.

Так как $|t_a| < t_{tabl}$ и $|t_b| > t_{tabl}$ то коэффициент a регрессии статистически незначим (и им можно пренебречь), а коэффициент b статистически значим.

Параметры линейной регрессии $y = a + bx$ в Excel можно определить гораздо проще.

С помощью инструмента анализа данных Регрессия можно получить результаты регрессионной статистики, дисперсионного анализа,

доверительные интервалы, остатки, графики подбора линий регрессии, графики остатков и нормальной вероятности. Порядок действий следующий:

1. Необходимо проверить доступ к Пакету анализа. Для этого в главном меню нужно выбрать Параметры Excel →Надстройки, установить курсор на Пакет анализа и нажать кнопку Перейти. В появившемся окне необходимо установить флажок Пакет анализа.

2. Выбрать в главном меню Данные / Анализ данных / Регрессия и заполнить диалоговое окно: Входной интервал *Y* – диапазон, содержащий данные результативного признака *Y*; Входной интервал *X* – диапазон, содержащий данные объясняющего признака *X*; Метки – флажок, который указывает, содержит ли первая строка названия столбцов или нет; Константа-ноль – флажок, указывающий на наличие или отсутствие свободного члена в уравнении; Выходной интервал – достаточно указать левую верхнюю ячейку будущего диапазона; Новый рабочий лист – можно задать произвольное имя нового листа, на который будут выведены результаты.

3. Для получения информации об остатках, графиков остатков, подбора и нормальной вероятности нужно установить соответствующие флажки в диалоговом окне.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,96
R-квадрат	0,92
Нормированный R-квадрат	0,91
Стандартная ошибка	35,61
Наблюдения	15

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	191102,48	191102,48	150,74	0,00
Остаток	13	16480,86	1267,76		
Итого	14	207583,33			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-17,78	19,38	-0,92	0,38	-59,65	24,08	-59,65	24,08
Переменная X 1	36,87	3,00	12,28	0,00	30,39	43,36	30,39	43,36

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*
1	81,78	28,22
2	37,53	32,47

Рисунок 4 Результаты применения инструмента Регрессия

В результате получится таблица (рисунок 4). В ней желтым цветом подсвечены те значения, которые получались в ходе разбора предложенного примера.

Задания по вариантам

Таблица 1.2 Варианты исходных данных

Вариант 1		Вариант 2		Вариант 3		Вариант 4		Вариант 5	
x	y	x	y	x	y	x	y	x	y
27,7	172	0,6	67,9	92,7	342,7	52,7	252,7	18,7	208,7
26,5	132	0,6	58,6	96,5	296,5	51,5	281,5	17,5	167,5
33,2	372	0,9	92,5	98,2	348,2	68,2	168,2	24,2	174,2
29,5	182	0,1	14,8	94,5	344,5	54,5	254,5	20,5	170,5
28,3	137	0,3	28,4	93,3	193,3	53,3	253,3	23,3	193,3
30,8	232	0,9	90,8	95,8	345,8	55,8	305,8	21,8	171,8
28	170	0,9	88,2	93	343	53	253	19	169
32,1	242	0,4	48,8	97,1	347,1	57,1	257,1	23,1	173,1
26,2	120	0,2	21,2	91,2	371,2	51,2	251,2	17,2	167,2
35,4	502	0,2	16,5	100,4	350,4	90,4	290,4	26,4	176,4
32,9	252	0,9	93,5	104,9	354,9	54,9	254,9	29,9	179,9
33,2	235	0,2	24,2	95,2	345,2	55,2	255,2	21,2	171,2
36,5	452	0,7	66,7	101,5	351,5	81,5	381,5	27,5	177,5
34,4	372	0,2	18	99,4	399,4	59,4	259,4	25,4	205,4
36,5	292	0,1	12	96,5	346,5	56,5	256,5	22,5	172,5
Вариант 6		Вариант 7		Вариант 8		Вариант 9		Вариант 10	
x	y	x	y	x	y	x	y	x	y
21,7	241,7	11,7	73,7	52,7	352,7	31,7	151,7	0,6	70,4
13,5	213,5	3,5	65,5	46,5	246,5	23,5	173,5	0,7	70,3
18,2	218,2	10,2	68,2	48,2	348,2	29,2	149,2	0,8	75,7
16,5	216,5	6,5	68,5	44,5	344,5	26,5	146,5	0,4	38,9
15,3	215,3	5,3	67,3	43,3	343,3	25,3	145,3	0,6	58,8
17,8	227,8	7,8	69,8	45,8	345,8	27,8	147,8	0,5	47,9
15	215	9	71	83	383	25	145	0,1	11,8
19,1	219,1	9,1	71,1	47,1	347,1	29,1	149,1	0,7	89,1
13,2	213,2	3,2	68,2	41,2	341,2	23,2	143,2	0,9	87,3
28,4	228,4	18,4	80,4	70,4	380,4	38,4	158,4	0,6	55,1
16,9	216,9	8,9	76,9	54,9	354,9	26,9	166,9	0,3	29,5
17,2	217,2	7,2	69,2	45,2	245,2	32,2	152,2	0,4	38,7
29,5	279,5	19,5	81,5	51,5	351,5	39,5	199,5	0,9	90,5
21,4	221,4	11,4	89,4	49,4	349,4	31,4	201,4	0,1	11,8
18,5	218,5	8,5	70,5	46,5	346,5	28,5	208,5	0,6	62,3

Контрольные вопросы

1. Что такое функция регрессии?
2. Назовите этапы регрессионного анализа.
3. Что понимается под спецификацией модели, и как она проявляется?
4. Как применяются F- и t-статистики в регрессионном анализе?
5. В чем состоит различие между теоретическим и эмпирическим уравнением регрессии?
6. В чем суть метода наименьших квадратов?
7. В чем суть коэффициента детерминации R^2?