МОДУЛЬ 3. ЭЛЕМЕНТЫ КОРРЕЛЯЦИОННОГО, ОДНОФАКТОРНОГО РЕГРЕССИОННОГО И ДИСПЕРСИОННОГО АНАЛИЗА

Практическое занятие №7.
НАХОЖДЕНИЕ МОМЕНТА И КОЭФФИЦИЕНТА КОРРЕЛЯЦИИ

Цель занятия. Уметь вычислять момент и коэффициент корреляции

Учебные вопросы:

1.Однофакторная линейная регрессия.

2. Статистическая оценка точности и надежности уравнения регрессии

Методика решения заданий такая же, как и для практического занятия №9.

Практическое занятие №8.
ПОСТРОЕНИЕ И СТАТИСТИЧЕСКИЙ АНАЛИЗ НАДЕЖНОСТИ И ДОСТОВЕРНОСТИ УРАВНЕНИЯ ОДНОФАКТОРНОЙ РЕГРЕССИИ

Цель занятия. Уметь строить уравнение однофакторной нелинейной регрессии в MS Excel. Получить навыки статистического анализа надежности и достоверности уравнения однофакторной регрессии

Учебные вопросы:

1.Однофакторная линейная регрессия.

2. Статистическая оценка точности и надежности уравнения регрессии

Методика решения заданий такая же, как и для практического занятия №9.

Практическое занятие №9.
НАХОЖДЕНИЕ УРАВНЕНИЙ ОДНОФАКТОРНОЙ НЕЛИНЕЙНОЙ РЕГРЕССИИ

Цель занятия. Уметь строить уравнение однофакторной нелинейной регрессии в MS Excel.

Учебные вопросы:

1.Основные виды уравнений нелинейной регрессии.

2. Остаточная дисперсия и индекс детерминации.

В MS Excel имеется несколько способов построения уравнения регрессии:

1) автоматический анализ тренда на основе диаграммы экспериментальных данных;
2) использование функций Excel из категории «Статистические»;
3) вычисление коэффициентов уравнения регрессии в ячейках Excel;
4) использование надстройки **Анализ данных. Регрессия.**

Задание 1

Построить в MS Excel 5 видов уравнений регрессии с помощью автоматического анализа тренда для экспериментальных данных признаков *X* и *Y*. Экспериментальные данные приведены в табл. 3.1.

Таблица 3.1

X	97	73	79	99	86	91	85	77	89	95	72	115
Y	161	131	135	147	139	151	135	132	161	159	120	160

На основе полученных значений индекса (коэффициента для линейной зависимости) детерминации выбрать линию регрессии, наиболее согласованную с экспериментальными данными.

Решение.

1. Построить диаграмму данных в виде графика или точечной диаграммы.
2. Активизировать диаграмму и выполнить команду **Макет. Линия тренда** (другой способ – навести курсор на линию графика или точку диаграммы; появится сообщение **Ряд "Y". Точка "1"**; через контексное меню выбрать команду **Добавить линию тренда**).
3. В окне «Формат линии тренда. Параметры линии тренда» выбрать вид линии тренда, поставить флажки в пунктах «Показывать уравнение на диаграмме», «Поместить на диаграмму величину достоверности аппроксимации (*R^2*)».

В MS Excel можно выбрать тренд из пяти видов аппроксимирующих линий (табл. 3.2).

Таблица 3.2

Название линии регрессии	Вид уравнения аппроксимирующей линии
Экспоненциальная	$Y = ae^{bX}$
Линейная	$Y = bX + a$
Логарифмическая	$Y = b\ln(X) + a$
Полиномиальная	$Y = b_1X^6 + b_2X^5 + b_3X^4 + b_4X^3 + b_5X^2 + b_6X + a$, *степень полинома можно выбрать от 2 до 6*
Степенная	$Y = bX^a$

Результаты решения задания представлены на рис. 3.1.

Анализ графиков линий тренда, значений индекса детерминации показывает, что наиболее подходит к экспериментальным данным полиномиальная регрессия. Для данной функции индекс детерминации R^2 равен 0,92 (хорошее качество уравнения регрессии). Уравнение регресии представляется в виде

$$y = -0{,}000001x^6 + 0{,}0008x^5 - 0{,}1974x + 25{,}116x^3 - 1781x^2 + 66773x - 1000000.$$

Параметры лучшего из пяти уравнений регрессии выделены на рисунке полужирным текстом.

Визуально все другие 4 вида линий тренда сливаются, т. е. между ними практически нет разницы. Об этом же свидетельствуют и примерно одинаковые индексы детерминации для этих трендов (диапазон значений – 0,684–0,727).

Следует отметить, что в MS Excel при построении линий тренда для всех функций рассчитывается линейный коэффициент корреляции, а не индекс корреляции.

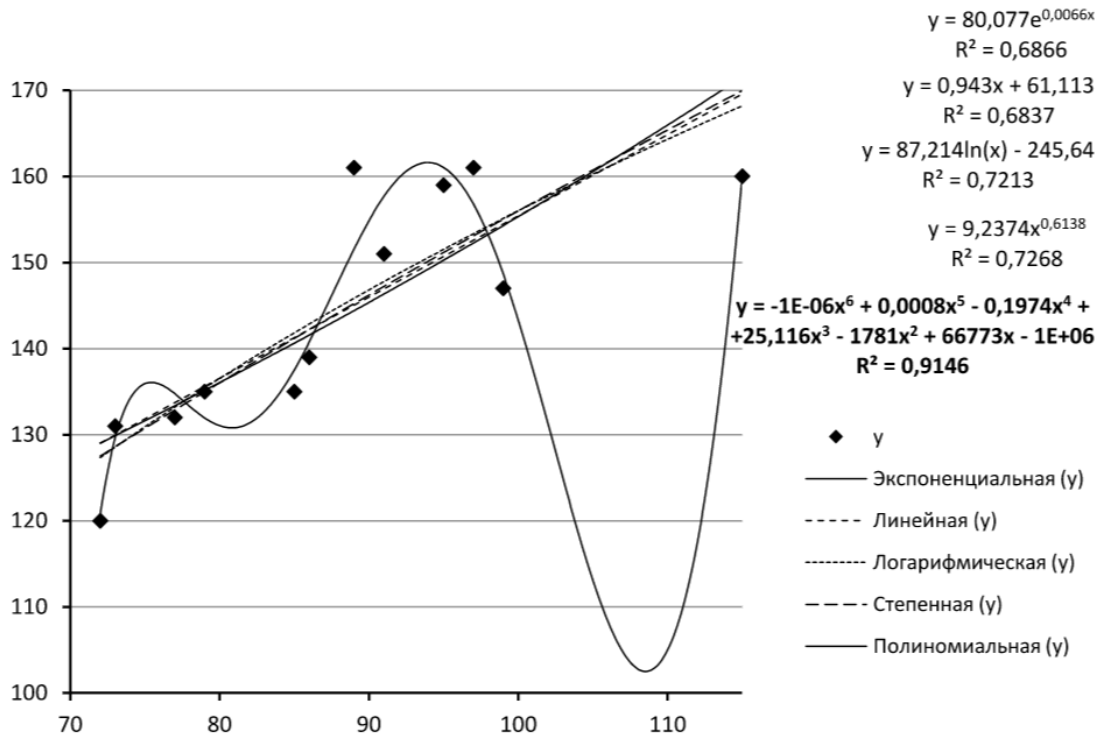


Рис. 3.1. Виды линий тренда

Задание 2

Построить уравнение регрессии для экспериментальных данных задания 1, используя функции Excel.

Решение.

Для получения коэффициентов уравнения линейной регрессии $Y = bX + a$ используем функции:

a =ОТРЕЗОК($Y;X$) = 61,113, b =НАКЛОН($Y;X$) = 0,943;

=ЛИНЕЙН(известные_значения_*y*, [известные_значения_*x*], [константа], [статистика]).

Если *Константа* равна 1, то константа a вычисляется обычным образом; если *Константа* = 0, то a = 0; если *Статистика* = 1, то выводится дополнительная регрессионная статистика (массив из 10 значений); если *Статистика* = 1, то выводится только коэффициент b.

Дополнительная регрессионная статистика:

Значение коэффициена (*b*)	Значение коэффициена *a*
Среднеквадратическое отклонение *b (Sb)*	Среднеквадратическое отклонение *a (Sa)*
Коэффициент детерминации R^2 *(R^2)*	Среднеквадратическое отклонение *Y (E)*

F-статистика *(F)*	Число степеней свободы *(n – k)*
Регрессионная сумма квадратов *(SS1)*	Остаточная сумма квадратов *(SS2)*

Функция ЛИНЕЙН работает как функция массива. Поэтому вначале выделяется массив ячеек 2×10. Затем вызывается функция ЛИНЕЙН, нажимается комбинация клавиш Ctrl + Shift + Enter.

Результаты работы функции ЛИНЕЙН для рассматриваемого примера:

b	0,943	61,113	*a*
Sb	0,203	18,043	*Sa*
R^2	0,684	8,327	*E*
F	21,615	10	*n – k*
SS1	1498,82	693,43	*SS2*

Для корректности задания значений по оси *x* значения упорядочены по возрастанию. Анализ данных, приведенных в ячейках выше, показывает, что значения коэффициентов *b* и *a*, полученные с помощью функции ЛИНЕЙН, совпадает со значениями, полученными с помощью функций ОТРЕЗОК и НАКЛОН. Уравнение регрессии имеет вид

$$\hat{y} = 0{,}943 \cdot x + 61{,}113.$$

Рассчитаем $\hat{y}$ для каждого значения *x* (табл. 3.3) и построим линию регрессии (рис. 3.2).

Таблица 3.3

X	*Y*	*Y^*
72	120	129,006
73	131	129,949
77	132	133,721
79	135	135,607
85	135	141,264
86	139	142,207
89	161	145,036
91	151	146,922
95	159	150,694
97	161	152,579
99	147	154,465
115	160	169,552

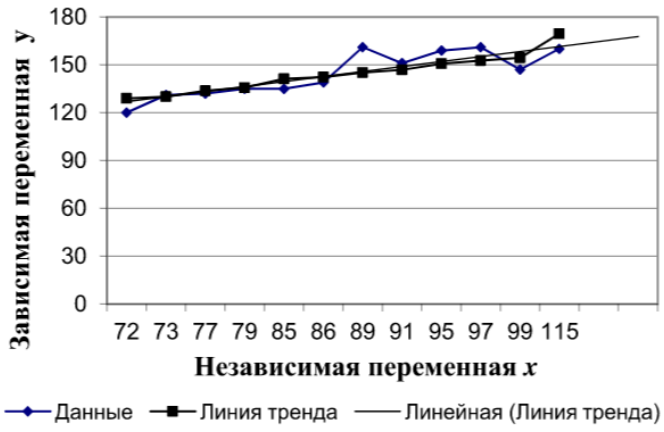


Рис. 3.2. Линия регрессии и прогноз

Задание 3

С помощью MS Excel провести регрессионный анализ данных:

1. Рассчитать простое уравнение линейной регрессии.
2. Проверить адекватность уравнения регрессии (модели) исходным данным с уровнем значимости 0,05.
3. Проверить достоверность коэффициентов модели.

4. Провести анализ остатков.

5. Дать прогноз на 2 единицы вперед от $x = 115$.

Решение.

Результаты расчетов по пункту 1 в ячейках Excel приведены в табл. 3.4.

Таблица 3.4

	A	B	C	D	E	F
1	*X*	*Y*	*Y^*	*(Y^ -MY)^2*	*Y-Y^*	*(Y-Y^)^2*
2	72	120	129,006	232,374	–9,006	81,111
3	73	131	129,949	204,515	1,051	1,104
4	77	132	133,721	110,864	–1,721	2,961
5	79	135	135,607	74,707	–0,607	0,368
6	85	135	141,264	8,915	–6,264	39,241
7	86	139	142,207	4,173	–3,207	10,286
8	89	161	145,036	0,618	15,964	254,851
9	91	151	146,922	7,139	4,078	16,632
10	95	159	150,694	41,519	8,306	68,998
11	97	161	152,579	69,379	8,421	70,907
12	99	147	154,465	104,351	–7,465	55,730
13	115	160	169,552	640,195	–9,552	91,242
14	*X* среднее	88,167	Сумма	1498,748	–0,002	693,430
15	*Y среднее*	144,250	*b*	0,943	*a*	61,115
16	*k*	2	*n*	12		
17	$F_{рас}$	21,61352	$F_{кр}$	4,965	*P*	0,001

При расчетах использовались формулы Excel (табл. 3.5).

Таблица 3.5

Ячейка Excel	Формула Excel
B14	=СРЗНАЧ(A2:A13)
B15	=СРЗНАЧ(B2:B13)
D15	Первоначально 0
F15	Первоначально 0
C2	=$D$15*A2+$F$15
D2	=(C2-$B$15)^2
E2	=B2-C2
F2	=(B2-C2)^2
B16	2
B17	=(D14*(D16-B16))/(F14*(B16-1))
D16	=СЧЁТ(A2:A13)
D17	=F.ОБР.ПХ(B17;B16-1;D16-B16;0)
F17	=F.РАСП.ПХ(B17;B16-1;D16-B16)

Для расчета коэффициентов *b* и *a* уравнения линейной регрессии использовалось средство «Поиск решения». Его установки: **оптимизировать целевую функцию** $F$14; **до минимума**; **изменяя ячейки переменных** $D$15;$F$15; **найти решение**. Полученные значения *b* и *a* совпадают с результатами пункта 1.

По пункту 2 проверка адекватности уравнения регрессии исходным данным осуществляется по критерию Фишера. Так как на уровне значимости $\alpha = 0{,}05$ $F_{рас} > F_{кр}$, то уравнение считается адекватным исходным данным. Рассчитанное на основе функции =F.РАСП.ПХ(B17;B16-1;D16-B16) значение вероятности равно 0,001.

По пункту 3 проверка значимости коэффициентов уравнения регрессии осуществляется по критерию Стьюдента: $t_b = b / Sb = 4{,}65$, $t_a = a / Sa = 3{,}39$. Рассчитаем значение функции Excel: =СТЬЮДЕНТ.РАСП.2Х(4,65;10)=0,007 и значение функции Excel: =СТЬЮД.РАСП.2Х (3,39;10)=0,001. Критическое значение $t_{кр}$ при уровне значимости $\alpha = 0{,}05$ =СТЬЮДЕНТ.ОБР.2Х(0,05;10) равно 2,23, т. е. коэффициенты линейной модели значимы.

Полученный согласно пункту 4 график остатков приведен на рис. 3.3.

Одно из требований к остаткам регрессии – они должны иметь нормальное распределение с нулевым средним. Не проводя тесты на нормальность, сопоставим значения трех параметров с соответствующими значениями нормального распределения.

Среднее значение остатков: =СРЗНАЧ($Y - Y$^) =0 (норма 0); асимметрия: =СКОС($Y - Y$^) =0,655 (норма 0); эксцесс: =ЭКСЦЕСС($Y – Y$^) = 0,279 (норма 0). Таким образом, остатки регрессии не соответствуют предъявляемым к ним требованиям.

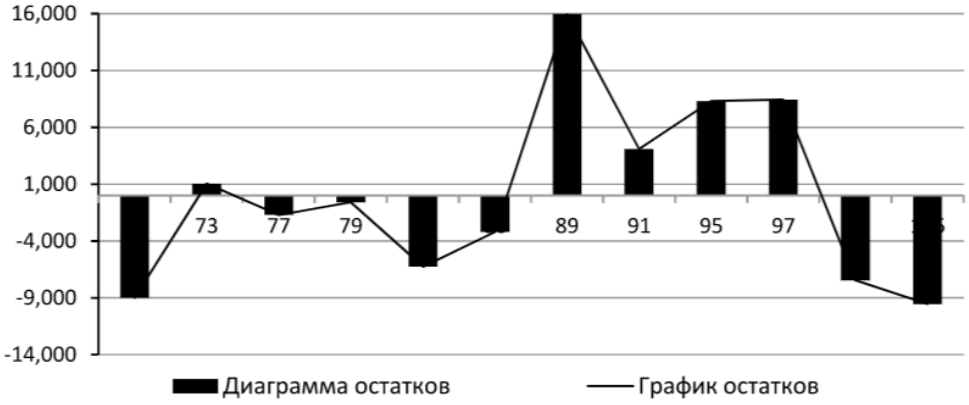


Рис. 3.3. Диаграмма и график остатков

По пункту 5 прогнозирование значения результативного признака на 10 единиц вперед производится по формуле

$$Y_{\text{прогноз}} = a + b \cdot X_{\text{прогноз}} = 61{,}113 + 0{,}943 \cdot (115 + 2) = 171{,}444.$$

Прогноз показан на рис. 3.4.

Задание 4

С помощью надстройки MS Excel **Анализ данных. Регрессия** построить уравнение регрессии для данных пункта 1, сравнить результаты с полученными в предыдущих заданиях.

Решение.

Задаем входные интервалы Y и X, метки, верхнюю левую ячейку выходного интервала; ставим флаги в окошках «График подбора» и «График остатков».

Результаты регрессионной статистики в табл. 3.6 по стандартной ошибке и значению R-квадрата совпадают со значениями аналогичных показателей,

полученных с помощью функции ЛИНЕЙН. Результаты дисперсионного анализа (табл. 3.7) совпадают со значениями аналогичных показателей для *F*, значимости *F*, регрессионной суммы и суммы остатков, полученных с помощью функции ЛИНЕЙН.

Таблица 3.6

Регрессионная статистика	
Множественный *R*	0,827
R-квадрат	0,684
Нормированный *R*-квадрат	0,652
Стандартная ошибка	8,327
Наблюдения	12

В табл. 3.8 находятся данные о коэффициентах уравнения регрессии ($Y_{пересечение}$ – это *a*, коэффициент при *X* – *b*), *t* – статистика и значимость коэффициентов. Их значения совпадают с теми, которые получены ранее (t_b = 4,65, t_a = 3,39, значимость 0,001 и 0,007).

Таблица 3.7

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1,000	1498,820	1498,820	21,615	0,001
Остаток	10,000	693,430	69,343		
Итого	11,000	2192,250			

Таблица 3.8

	Коэффициенты	*Станд. ошибка*	*t-статистика*	*P-Значение*	*Нижние 95 %*	*Верхние 95 %*	*Нижние 95 %*	*Верхние 95 %*
Y-перес.	61,113	18,043	3,387	0,007	20,911	101,316	20,911	101,316
X	0,943	0,203	4,649	0,001	0,491	1,395	0,491	1,395

На рис. 3.4 помещен скорректированный (по минимальным и максимальным значениям по осям, типу диаграммы) график подбора линии регрессии.

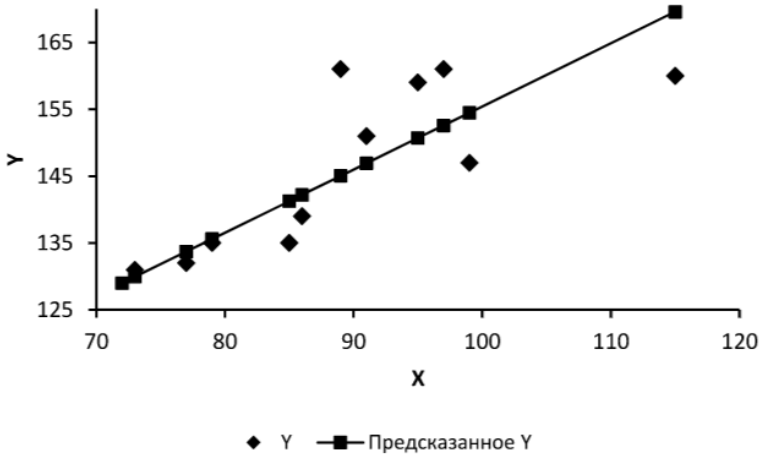
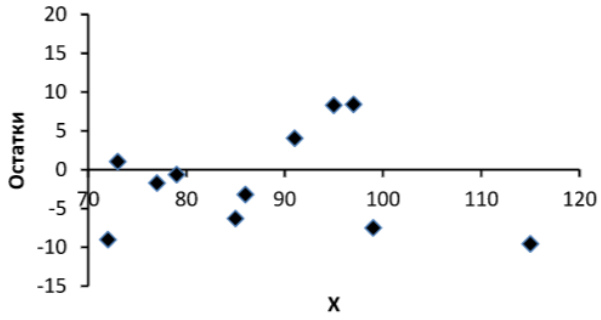


Рис. 3.4. График подбора линии регрессии



На рис. 3.5 помещен график остатков, полученный с помощью надстройки MS Excel «Анализ данных». При коррекции графика учтены максимальное и минимальное значения по осям.

Рис. 3.5. График остатков линии регрессии