N 3

3. ЛАБОРАТОРНАЯ РАБОТА №3. РЕГРЕССИОННЫЙ АНАЛИЗ В MS EXCEL

Цель работы – освоение методики проведения регрессионного анализа средствами MS Excel.

3.1 Теоретические сведения

Популярное офисное приложение MS Excel содержит мощные встроенные инструменты статистической обработки данных.

Инструмент анализа «Регрессия» применяется для подбора графика для набора наблюдений с помощью метода наименьших квадратов. Инструмент «Регрессия» использует функцию ЛИНЕЙН.

Для построения линейной регрессионной модели необходимо подготовить список из n строк и m столбцов, содержащий экспериментальные данные (столбец, содержащий выходную величину y должен быть либо первым, либо последним в списке) и обратиться к меню **Сервис/Анализ данных/Регрессия**.

Внешний вид окна «Анализ данных» MS Excel представлен на рисунке 3.1.

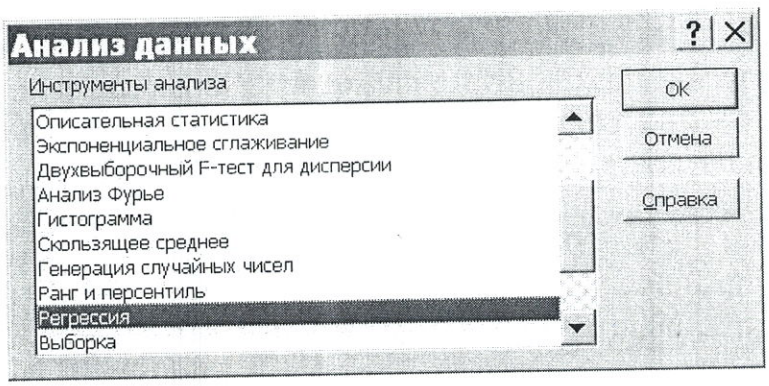

Рисунок 3.1 – Внешний вид окна «Анализ данных» MS Excel

Если пункт **«Анализ данных»** в меню **«Сервис»** отсутствует, то необходимо обратиться к пункту **«Надстройки»** того же меню и установить флажок **«Пакет анализа»**.

Значения параметров, установленных в диалоговом окне **«Регрессия»**, представлены на рисунке 3.2.

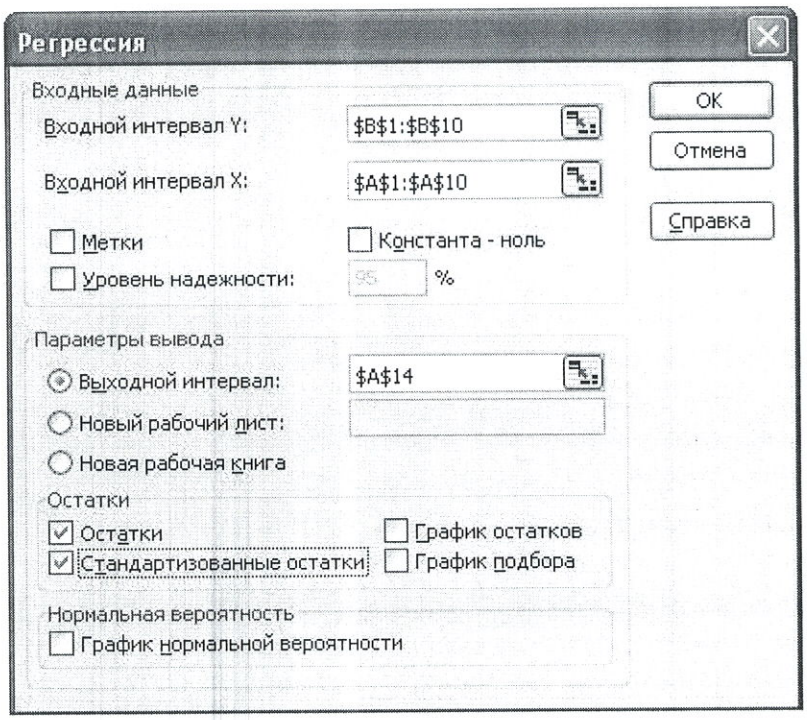

Рисунок 3.2 – Диалоговое окно «Регрессия» MS Excel

В диалоговом окне задаются следующие параметры:

Входной интервал Y – диапазон ячеек, содержащий данные результативного признака;

Входной интервал X – диапазон ячеек, содержащий данные факторного признака;

Метки – флажок, который указывает, содержит ли первая строка названия столбцов или нет;

Константа-ноль – данный флажок необходимо установить, чтобы линия регрессии прошла через начало координат;

Уровень надежности – этот флажок необходимо использовать, если требуется уровень надежности отличный от 95%, принятый по умолчанию;

Выходной интервал – верхняя левая ячейка интервала, в который будут помещаться результаты вычислений (можно разместить результаты на новом рабочем листе или в новой рабочей книге).

Если необходимо получить дополнительную информацию и графики остатков, установите соответствующие флажки в диалоговом окне. Щелкните по кнопке ОК;

MS Excel автоматически сгенерирует результаты по регрессионной статистике, представленные в виде 3 таблиц.

Результаты регрессионной статистики будут представлены в виде таблицы, изображенной на рисунке 3.1.

Регрессионная статистика	
Множественный R	0,969525973
R-квадрат	0,939980612
Нормированный R-квадрат	0,935363736
Стандартная ошибка	14,22893673
Наблюдения	15

Рисунок 3.1 – Результаты регрессионного анализа в MS Excel

Регрессионная статистика, включает в свой состав:

- Множественный *R* – коэффициент множественной корреляции;
- *R*- квадрат – множественный коэффициент детерминации;
- Нормированный *R*-квадрат – скорректированный коэффициент детерминации;
- Стандартная ошибка – стандартная ошибка регрессии;
- Наблюдения – количество наблюдений.

Результаты дисперсионного анализа будут представлены в виде таблицы, показанной на рисунке 3.2.

Дисперсионный анализ					
	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	41220,72106	41220,72106	203,5966782	2,55346E-09
Остаток	13	2632,014326	202,4626405		
Итого	14	43852,73538			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	4,746	7,003	0,678	0,510	-10,384	19,876	-10,384	19,876
Переменная X 1	9,595	0,672	14,269	0,000	8,142	11,048	8,142	11,048

Рисунок 3.2 – Результаты дисперсионного анализа в MS Excel

В группе результатов дисперсионного анализа использован ряд общепринятых сокращений:

df – степени свободы (degree of freedom);

SS – сумма квадратов отклонений (Sum of squares);

MS – средний квадрат отклонения (Mean square);

F – отношение дисперсий;

Значимость ***F*** – критическое значение квантиля распределения Фишера, на котором отвергается нулевая гипотеза отсутствия влияния фактора.

Третья группа результатов включает в свой состав значения коэффициентов регрессии, а также статистики, на основании которых проверяется значимость влияния фактора для каждого коэффициента, включенного в модель:

Коэффициенты – значения коэффициентов;

Стандартная ошибка – стандартная ошибка коэффициентов;

t-статистика – значение статистики критерия;

P–значение – уровень значимости отклонения гипотезы равенства коэффициентов нулю;

Нижние 95% - нижняя граница доверительного интервала, в котором находится значение коэффициента генеральной совокупности;

Верхние 95% - верхняя граница доверительного интервала, в котором находится значение коэффициента генеральной совокупности.

При необходимости есть возможность вывести таблицу стандартных и простых остатков, где для каждого значения ряда выводится предсказанное значение, с которым сопоставляется остаток, представляющий разность между прогнозным и реальным значением ряда.

Кроме вывода табличной информации, есть возможность просмотреть графики остатков, что позволяет визуально проконтролировать качество подбора модели и отсутствие закономерности в остатках.

Для приведенного на рисунке 3.2 примера уравнение регрессии будет выглядеть следующим образом:

$$\tilde{y} = 4.746 + 9.595 \cdot x \qquad (3.1)$$

Направление связи между переменными определяется на основании знаков (отрицательный или положительный) у коэффициента регрессии (коэффициента b). В нашем случае знак коэффициента регрессии положительный, следовательно, связь также является положительной.

Выводимые результаты позволяют проверить значимость коэффициентов регрессии: a и b. Сравнивая попарно значения столбцов ***Коэффициенты*** и ***Стандартная ошибка*** в таблице (рисунок 3.2), видим, что абсолютное значение коэффициента b больше чем его стандартная ошибка. К тому же этот коэффициент является значимым, о чем можно судить по значениям показателя ***P-значение*** в таблице, которые меньше заданного уровня значимости α=0,05.

Обратная картина наблюдается для коэффициента a, который в приведенном примере является незначимым, о чем дополнительно свидетельствуют границы доверительного интервала от -10,384 до 19,876.

В доверительный интервал попадает ноль, а значит значения коэффициента *a* может быть принято равным нулю.

В таблице, представленной на рисунке 3.3, показаны результаты вывода остатков. При помощи этой части отчета можно увидеть отклонения каждой точки от построенной линии регрессии. Наибольшее абсолютное значение остатка в нашем случае - 32,680, наименьшее - 1,405.

Вывод остатка			
Наблюдение	*Предсказанное Y*	*Остатки*	*Стандартные остатки*
1	14,34109109	2,46894016	0,180065478
2	33,53122802	-6,286822857	-0,458512433
3	43,12629649	-4,678139819	-0,341187484
4	81,50657035	14,56087	1,0619577
5	71,91150189	4,064625487	0,296442474
6	91,10163882	12,19267418	0,889239737
7	148,6720496	-9,493250037	-0,692364534
8	139,0769811	14,86694901	1,084280745
9	119,8868442	-32,68030161	-2,383449471
10	196,6473919	-11,22156612	-0,818414596
11	158,2671181	23,01022885	1,678188851
12	110,2917758	1,840441652	0,134227638
13	23,93615956	4,164485469	0,303725492
14	52,72136495	-11,40440323	-0,831749326
15	62,31643342	-1,404731148	-0,10245027

Рисунок 3.3 – Результаты вывода остатков в MS Excel

Простым и наглядным способом проверки удовлетворительности регрессионной модели является графическое представление отклонений, которое MS Excel представляет в виде графика остатков, представленном на рисунке 3.4.

Если регрессионная модель близка к реальной зависимости, то отклонения будут носить случайный характер и их сумма будет близка к нулю. В рассмотренном примере:

$$\sum_{i=1}^{n}(\tilde{y}_i - y_i) = 0.000001$$

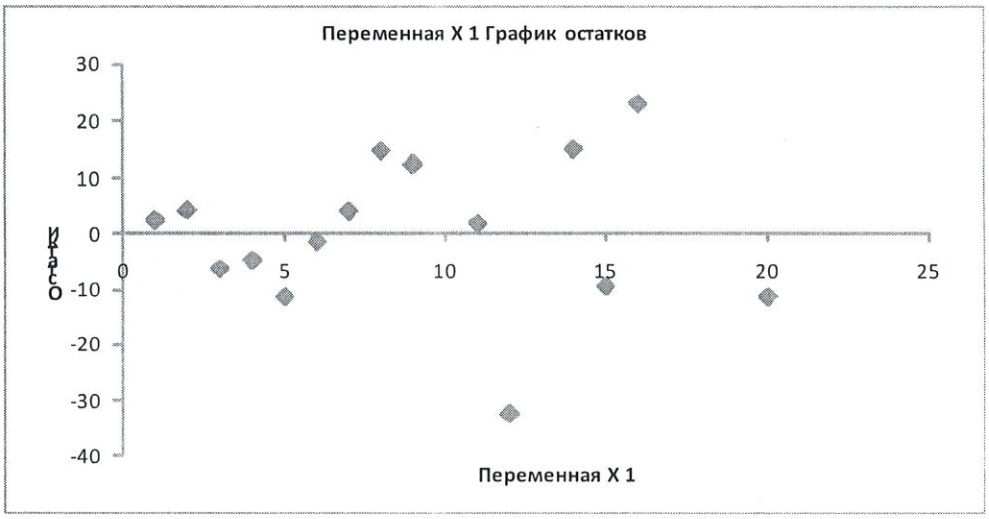

Рисунок 3.4 – Результаты графического представления остатков в MS Excel

Обычно мерой ошибки регрессионной модели служит среднее квадратическое отклонение:

$$\sigma_\varepsilon = \left[\left(\sum_{i=1}^{n} \varepsilon_i^{\,2}\right) \Big/ (n-2)\right]^{1/2} = \left\{\left[\sum_{i=1}^{n} (\tilde{y}_i - y_i)^2\right] \Big/ (n-2)\right\}^{1/2} \qquad (3.2)$$

Для нормально распределенных процессов приблизительно 67% точек находится в пределах одного отклонения σ_ε от линии регрессии и 95% - в пределах $2\sigma_\varepsilon$ (на рисунке 3.5 эти интервалы выделены разным цветом).

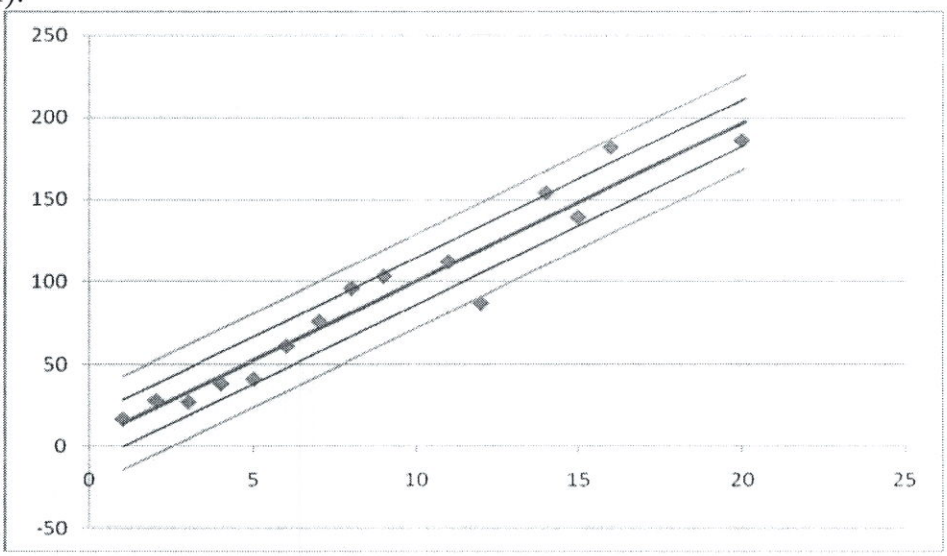

Рисунок 3.5 –Графическое представление результатов регрессионного анализа

3.2. Методические указания

Используя исходные данные **Лабораторной работы №2** произвести регрессионный анализ данных средствами MS Excel.

1. Создать лист с именем «Регрессионный анализ».
2. Произвести математическую формулировку задачи в виде таблицы исходных данных. согласно номеру своего варианта.
3. Запустить пакет регрессионного анализа через меню «Сервис - Пакет анализа – Регрессия».
4. В окне «Регрессия» задать интервалы, содержащие Y и X. Результаты обработки расположить на листе «Регрессионный анализ».
5. Проанализировать значения *t*-статистики для каждого коэффициента модели.
6. Построить график остатков уравнения регрессии
7. Построить доверительные интервалы для линии регрессии
8. Провести анализ полученных результатов моделирования

3.3. Требования к отчету

Отчет по лабораторной работе должен содержать:

1. Цель работы
2. Результаты проведенных вычислений
3. Выводы

3.4. Контрольные вопросы

1. Что понимается под регрессией в теории вероятностей и математической статистике?
2. Какие функции используются для построения уравнения парной регрессии в MS Excel ?
3. Какие задачи решаются при построении уравнения регрессии?
4. Что означает уровень значимости при проверке статистических гипотез?
5. Как вычисляется коэффициент детерминации и что он характеризует ?
6. Как проверяется значимость уравнения регрессии?
7. Как проверяется значимость коэффициентов уравнения регрессии?
8. По какой формуле вычисляется коэффициент парной корреляции r_{xy} ?
9. Как строится доверительный интервал для линейного коэффициента парной корреляции?