

# Множественный корреляционный анализ

**Выполнила:**  
студент(ка) группы 1к-Пот.1 -МГЭ  
Кондрашова Анна Николаевна

**Проверил:**  
д. т. н., профессор Ядыкин Евгений  
Александрович

**Понятие корреляции** появилось в середине XIX века в работах английских статистиков *Ф. Гальтона* и *К. Пирсона*. Этот термин произошел от латинского "correlatio" - соотношение, взаимосвязь. **Понятие регрессии** (латинское "regressio" - движение назад) также введено *Ф. Гальтоном*, который, изучая связь между ростом родителей и их детей, обнаружил явление "регрессии к среднему" - рост детей очень высоких родителей имел тенденцию быть ближе к средней величине.

Теория и методы корреляционного анализа используются для выявления связи между случайными переменными и оценки ее тесноты. Основной задачей регрессионного анализа является установление формы и изучение зависимости между переменными.

## Статистическая зависимость

- Изменение одной из величин влечет изменение распределения другой.

## Корреляционная зависимость

- Статистическая зависимость, при которой изменение одной из величин влечет изменение среднего значения другой

Функция  $\hat{y} = f(x_1, x_2, \dots, x_p)$ ,

**описывающая зависимость показателя от параметров, называется уравнением (функцией) регрессии.**

Уравнение регрессии показывает ожидаемое значение зависимой переменной при определенных значениях независимых переменных .

В зависимости от количества включенных в модель факторов  $X$  модели делятся на однофакторные (парная модель регрессии) и многофакторные (модель множественной регрессии).

В зависимости от вида функции  $f(X_1, X_2, \dots, X_k)$  модели делятся на линейные и нелинейные.

Модель множественной линейной регрессии имеет вид:

$$y_{i(1)} = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \dots + \alpha_k x_{ik} + \varepsilon_i$$

- количество наблюдений.

Коэффициент регрессии  $\alpha_j$  показывает, на какую величину в среднем изменится результирующий признак, если переменную  $x_j$  увеличить на единицу измерения, т. е.  $\alpha_j$  является нормативным коэффициентом.

Коэффициент может быть отрицательным. Это означает, что область существования показателя не включает нулевых значений параметров. Если же  $\alpha_0 > 0$ , то область существования показателя включает нулевые значения параметров, а сам коэффициент характеризует среднее значение показателя при отсутствии воздействий параметров.

Анализ уравнения (1) и методика определения параметров становятся более наглядными, а расчетные процедуры существенно упрощаются, если воспользоваться матричной формой записи:

$$Y = Xa + \varepsilon \quad (2)$$

Где  $Y$  – вектор зависимой переменной размерности  $n \times 1$ , представляющий собой  $n$  наблюдений значений .

- матрица  $n$  наблюдений независимых переменных , размерность матрицы равна  $n \times (k+1)$  . Дополнительный фактор , состоящий из единиц, вводится для вычисления свободного члена. В качестве исходных данных могут быть временные ряды или пространственная выборка.

$k$ - количество факторов, включенных в модель.

$a$  — подлежащий оцениванию вектор неизвестных параметров размерности  $(k+1) \times 1$ ;

$\varepsilon$  — вектор случайных отклонений (возмущений) размерности  $n \times 1$ .  $\varepsilon$  отражает тот факт, что изменение  $y$  будет неточно описываться изменением объясняющих переменных  $x$ , так как существуют и другие факторы, неучтенные в данной модели.

$k$  - количество факторов, включенных в модель.

$a$  — подлежащий оцениванию вектор неизвестных параметров размерности  $(k+1) \times 1$ ;

$\varepsilon$  — вектор случайных отклонений (возмущений) размерности  $n \times 1$ . отражает тот факт, что изменение будет неточно описываться изменением объясняющих переменных, так как существуют и другие факторы, неучтенные в данной модели.



Таким образом,

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk} \end{pmatrix}, \quad a = \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_k \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Уравнение (2) содержит значения неизвестных параметров  $a_0, a_1, a_2, \dots, a_k$

Эти величины оцениваются на основе выборочных наблюдений, поэтому полученные расчетные показатели не являются истинными, а представляют собой лишь их статистические оценки. Модель линейной регрессии, в которой вместо истинных значений параметров подставлены их оценки (а именно такие регрессии и применяются на практике), имеет вид

$$Y = XA + e = \hat{Y} + e,$$

где  $A$  — вектор оценок параметров;  $e$  — вектор «оцененных» отклонений регрессии, остатки регрессии  $e = Y - XA$ ; — оценка значений  $Y$ , равная  $XA$ .

Построение уравнения регрессии осуществляется, как правило, **методом наименьших квадратов (МНК)**, суть которого состоит в минимизации суммы квадратов отклонений фактических значений результатного признака от его расчетных значений, т.е.:

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 \rightarrow \min$$

Формулу для вычисления параметров регрессионного уравнения по методу наименьших квадратов приведем без вывода

$$A = (X'X)^{-1} X'Y$$

Для того чтобы регрессионный анализ, основанный на обычном методе наименьших квадратов, давал наилучшие из всех возможных результаты, должны выполняться следующие условия, известные как условия Гаусса – Маркова.

Первое условие. Математическое ожидание случайной составляющей в любом наблюдении должно быть равно нулю.

$$M(\varepsilon_i) = 0 \quad \text{для всех } i = 1, 2, \dots, n$$

Второе условие означает, что дисперсия случайной составляющей должна быть постоянна для всех наблюдений. Эта постоянная дисперсия обычно обозначается  $\sigma^2$ , или часто в более краткой форме  $\sigma_\varepsilon^2$ . Данное условие записывается следующим образом:

$$D(\varepsilon_i) = D(\varepsilon_j) = \sigma_\varepsilon^2 \quad \text{для любых наблюдений } i \text{ и } j$$

Выполнимость данного условия называется **гомоскедастичностью** (постоянством дисперсии отклонений). невыполнимость данной предпосылки называется **гетероскедастичностью**, (непостоянством дисперсии отклонений).

Третье условие предполагает отсутствие систематической связи между значениями случайной составляющей в любых двух наблюдениях. В силу того, что  $M(\varepsilon_i) = M(\varepsilon_j) = 0$ , данное условие можно записать следующим образом:

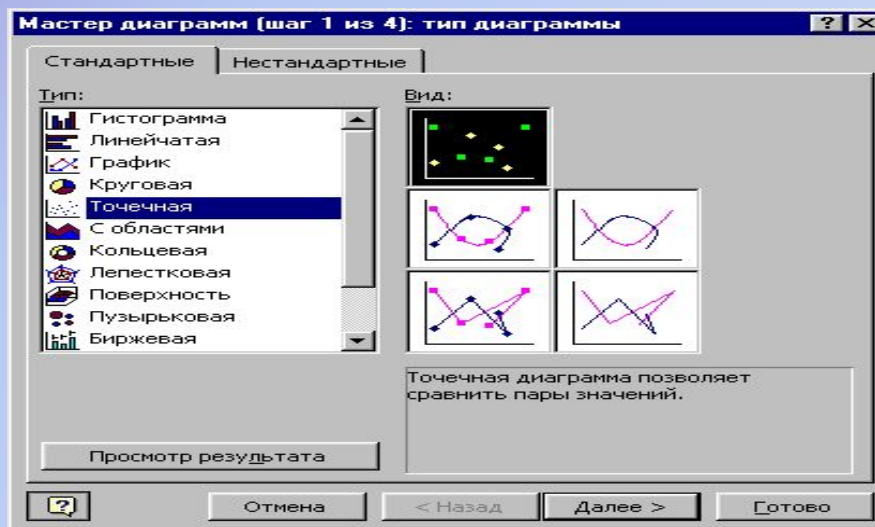
$$M(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j)$$

Возмущения  $\varepsilon_i$  и  $\varepsilon_j$  не коррелированы (условие независимости случайных составляющих в различных наблюдениях). Это условие означает, что отклонения регрессии (а значит, и сама зависимая переменная) не коррелируют.

Четвертое условие состоит в том, что в модели (1) возмущение  $\varepsilon_i$  (или зависимая переменная  $y_i$ ) есть величина случайная, а объясняющая  $x_i$  переменная - величина неслучайная. Если это условие выполнено, то теоретическая ковариация между независимой переменной и случайным членом равна нулю.

# КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ В MS EXCEL

1. Создайте файл исходных данных в MS Excel (например, таблица 2)
2. Построение корреляционного поля
3. Для построения корреляционного поля в командной строке выбираем меню **Вставка/ Диаграмма**. В появившемся диалоговом окне выберите тип диаграммы: **Точечная**; вид: **Точечная диаграмма**, позволяющая сравнить пары значений (Рис. 5).



Нажимаем кнопку **Далее**>. В появившемся диалоговом окне (Рис. 6) указываем диапазон значений, в нашем примере = Лист1!A2:В26 и указываем расположение данных: **в столбцах**.

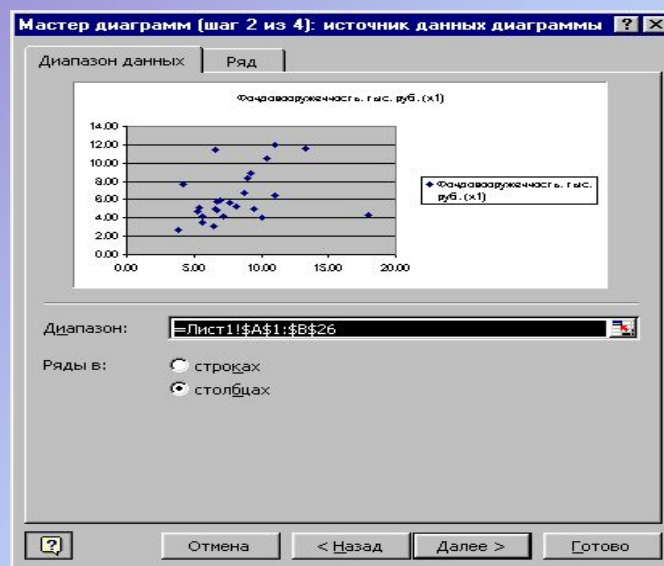


Рисунок 6– Вид окна при выборе диапазона и рядов

Нажимаем кнопку **Далее>**. В следующем диалоговом окне (рис. 7) указываем название диаграммы, наименование осей. Нажимаем кнопку **Далее>**, и **Готово**.

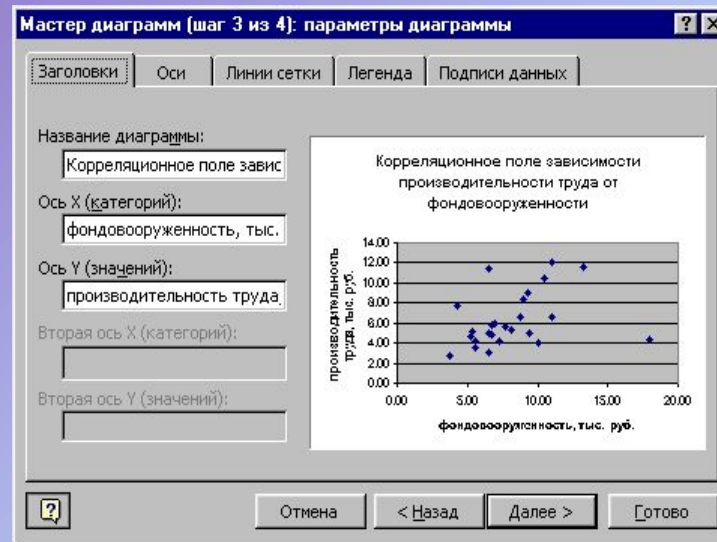


Рисунок 7 – Вид окна, шаг 3.

Таким образом, получаем корреляционное поле зависимости  $y$  от  $x$ . Далее добавим на графике линию тренда, для чего выполним следующие действия:



В области диаграммы щелкнуть левой кнопкой мыши по любой точке графика, затем щелкнуть правой кнопкой мыши по этой же точке. Появляется контекстное меню (рис. 8).

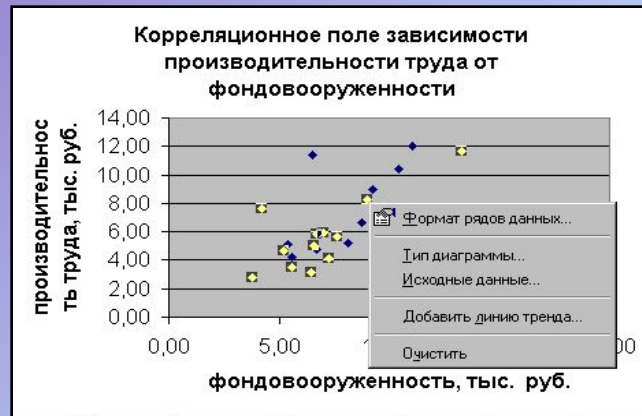


Рисунок 8 – Вид окна, шаг 4

В контекстном меню выбираем команду **Добавить линию тренда**.

В появившемся диалоговом окне выбираем тип графика (в нашем примере линейная) и параметры уравнения, как показано на рисунке 9.

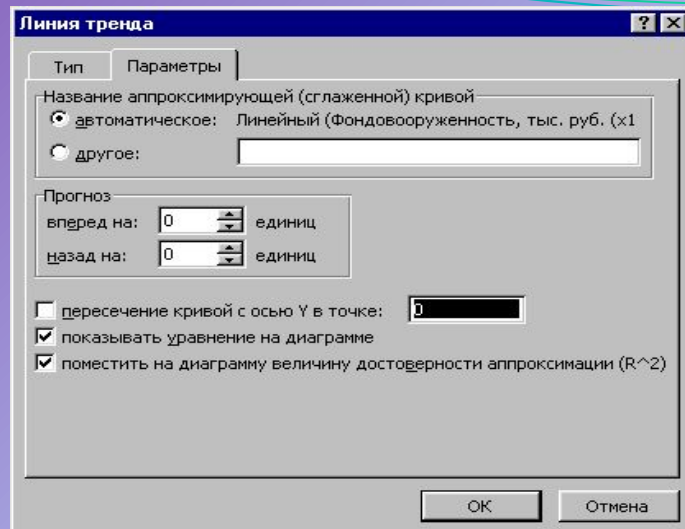


Рисунок 9 – Установка параметров линии тренда

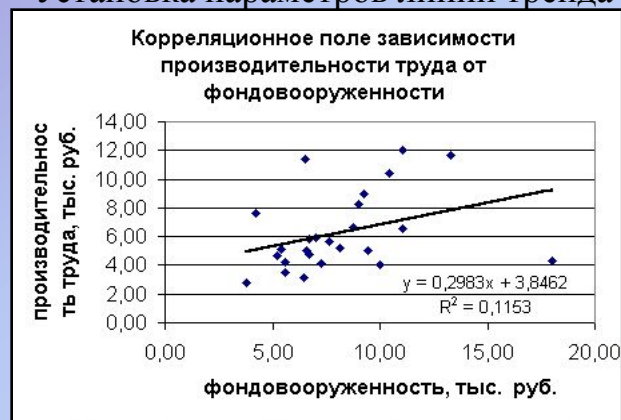


Рисунок 10– Корреляционное поле зависимости производительности труда от фондовооруженности

Аналогично строим корреляционное поле зависимости производительности труда от коэффициента сменности оборудования. (рисунок 11).

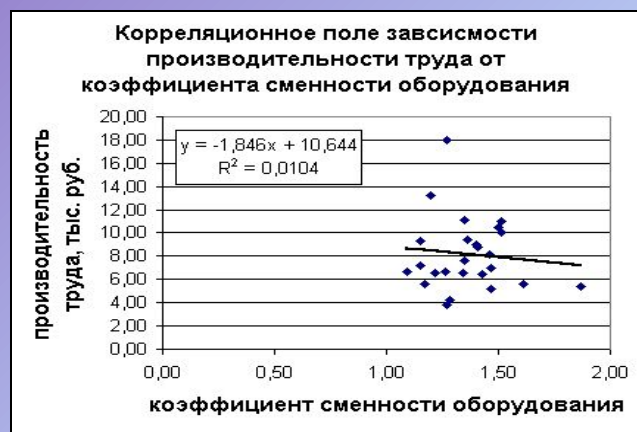


Рисунок 11 – Корреляционное поле зависимости производительности труда от коэффициента сменности оборудования

Построение корреляционной матрицы.

Для построения корреляционной матрицы в меню **Сервис** выбираем **Анализ данных**.

С помощью инструмента анализа данных **Регрессия**, помимо результатов регрессионной статистики, дисперсионного анализа и доверительных интервалов, можно получить остатки и графики подбора линии регрессии, остатков и нормальной вероятности. Для этого необходимо проверить доступ к пакету анализа. В главном меню последовательно выберите **Сервис/ Надстройки**. Установите флажок **Пакет анализа** (Рисунок 12)

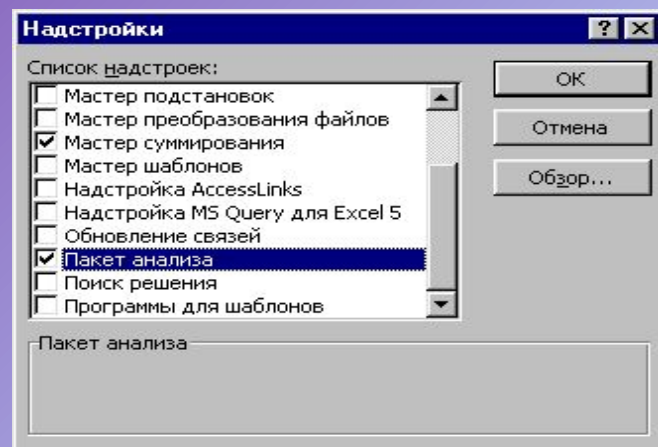
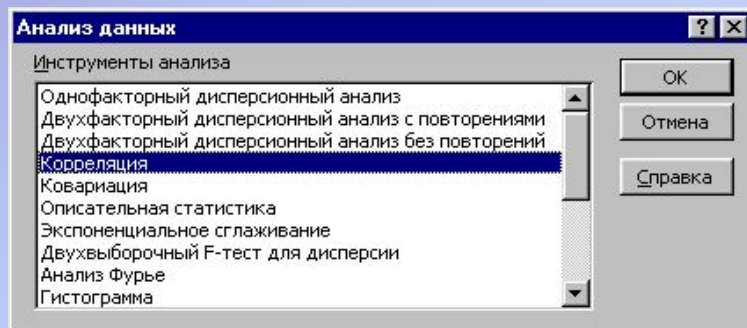


Рисунок 12 – Подключение надстройки **Пакет анализа**

В диалоговом окне **Анализ данных** выбираем **Корреляция** (Рисунок 13).



После нажатия ОК в появившемся диалоговом окне указываем входной интервал (в нашем примере A2:D26), группирование (в нашем случае по столбцам) и параметры вывода, как показано на рисунке 14.

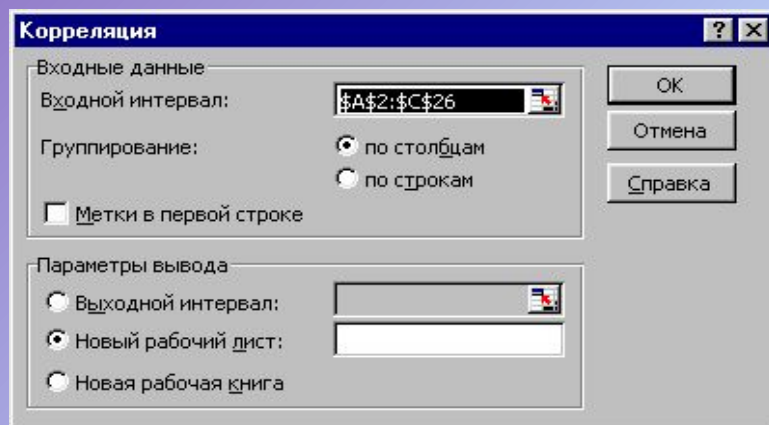


Рисунок 14 – Диалоговое окно **Корреляция**

Результат расчетов представлен в таблице 4.

Таблица 4 – Корреляционная матрица

	<i>Столбец 1</i>	<i>Столбец 2</i>	<i>Столбец 3</i>
<i>Столбец 1</i>	1		
<i>Столбец 2</i>	0,3395753	1	
<i>Столбец 3</i>	-0,102020	-0,161494	1