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 $\hat{y} = f(x1,x2,\ldots,xp)$,

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

Уравнение регрессии показывает ожидаемое значение зависимой переменной при определенных значениях зависимых 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 $f(X1, X2, \ldots Xk)$ модели делятся на линейные и нелинейные.

Модель множественной линейной регрессии имеет вид:

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_k x_{ik} + \varepsilon_i \quad (1)$$

- количество наблюдений.

Коэффициент регрессии α_j показывает, на какую величину в среднем изменится результативный признак , если переменную x_j увеличить на единицу измерения, т. е. α_j является нормативным коэффициентом.

Коэффициент может быть отрицательным. Это означает, что область существования показателя не включает нулевых значений параметров. Если же $a_0>0$, 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:

$Y=Xa+\varepsilon$ (2)

Где – 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 $(k+1) \times 1$;

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

k - количество факторов, включенных в модель.

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

ε — 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 пара метров $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_k$

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 под ставлены их оценки (а именно такие регрессии и применяются на практике), имеет вид $Y = XA + e = Y + e,$

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.:

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

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.

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

Второе условие означает, что дисперсия случайной составляющей должна быть постоянна для всех наблюдений**.** Эта постоянная дисперсия обычно обозначается σ_{ε}^2, или часто в более крат кой форме $\sigma^2(\varepsilon)$, а 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:

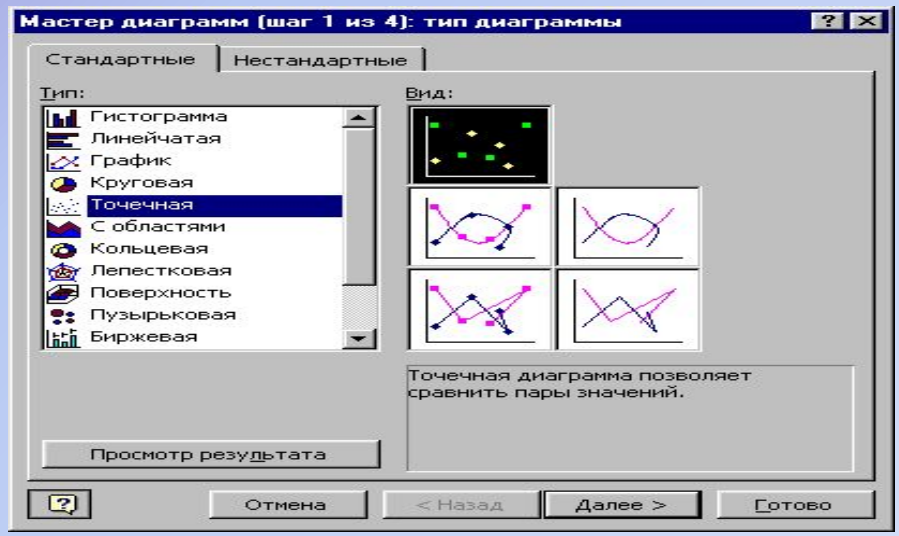
$$M(\varepsilon_i, \varepsilon_j) = 0 \quad (i \neq j)$$

Возмущения ε_i и ε_j не коррелированны (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 ε_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:B26 и указываем расположение данных: **в столбцах**.

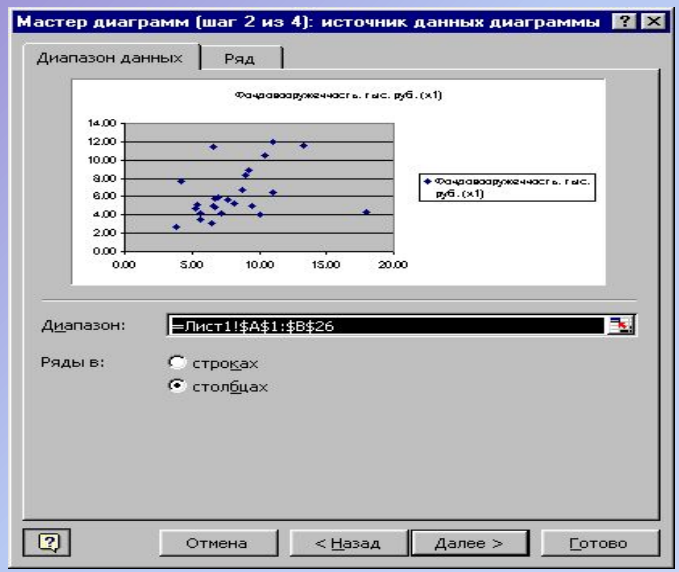


Рисунок 6– Вид окна при выборе диапазона и рядов

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

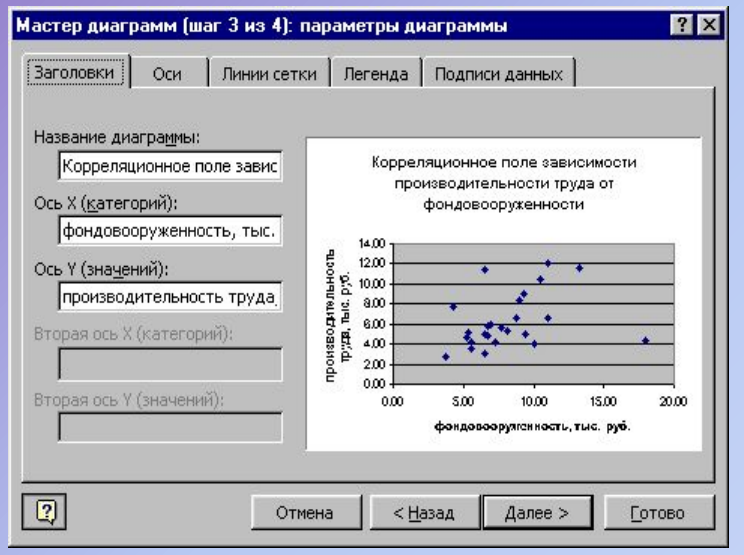


Рисунок 7 – Вид окна, шаг 3.

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).

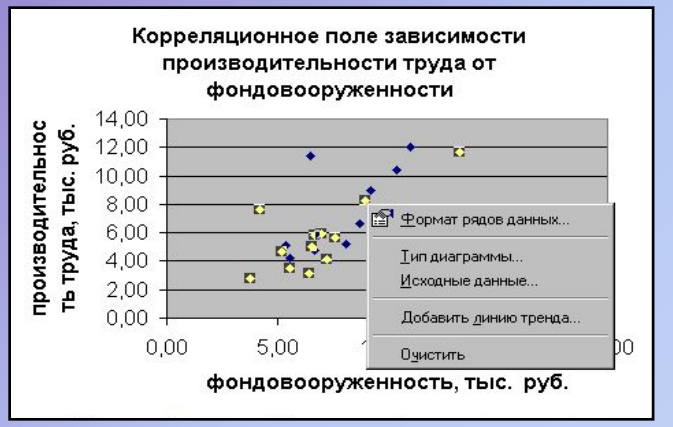


Рисунок 8 – Вид окна, шаг 4

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.

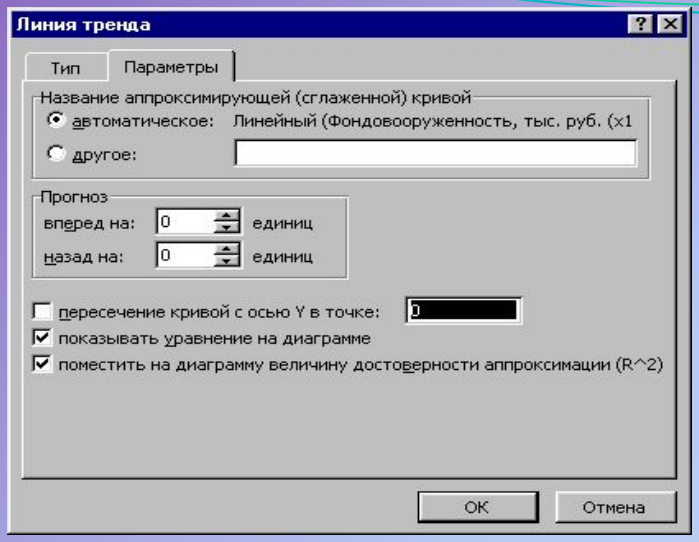


Рисунок 9 – Установка параметров линии тренда

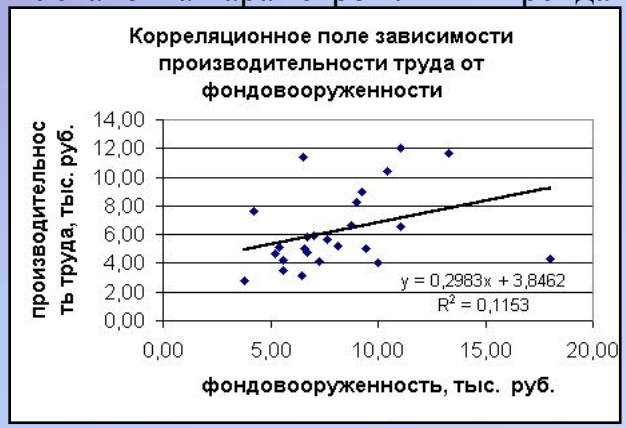


Рисунок 10– Корреляционное поле зависимости производительности труда от фондовооруженности

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

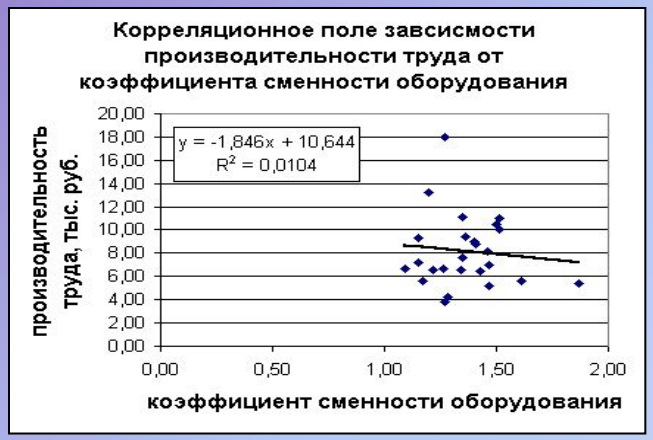


Рисунок 11 – Корреляционное поле зависимости производительности труда
от коэффициента сменности оборудования

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)

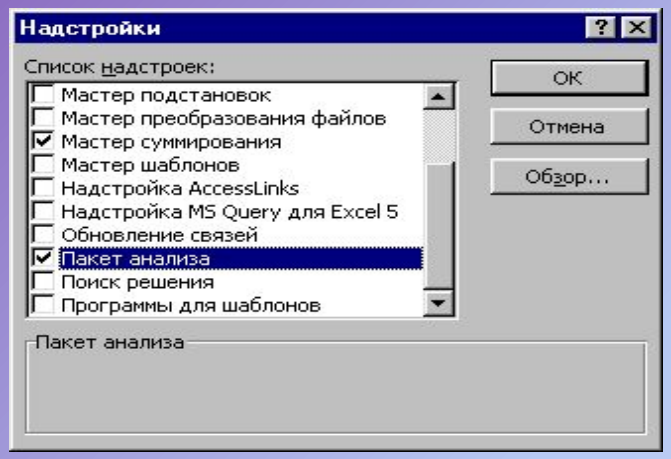
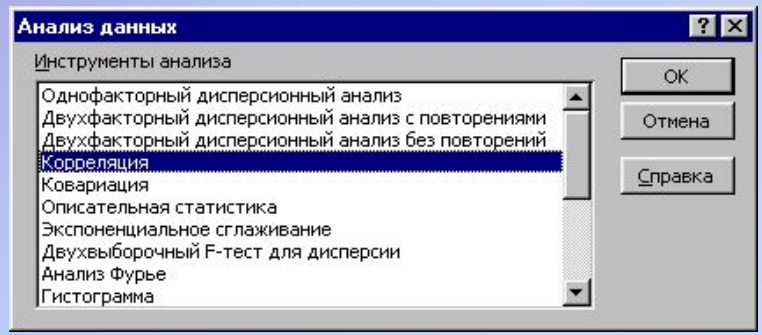


Рисунок 12 – Подключение надстройки **Пакет анализа**

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.

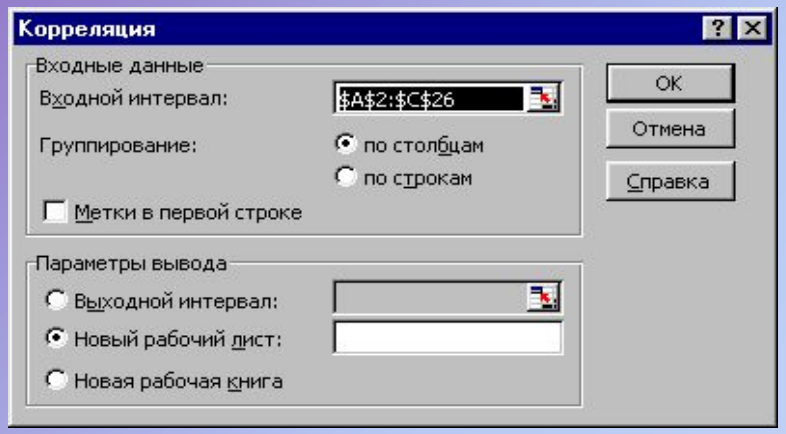


Рисунок 14 – Диалоговое окно **Корреляция**

Результат расчетов представлен в таблице 4.

Таблица 4 – Корреляционная матрица

	Столбец 1	*Столбец 2*	*Столбец 3*
Столбец 1	1		
Столбец 2	0,3395753	1	
Столбец 3	-0,1020202	-0,161494	1