Министерство образования Российской Федерации

Магнитогорский Государственный технический университет
имени Г. И. Носова

Кафедра технологий обработки материалов

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Методические указания

Магнитогорск

2017

1.ЦЕЛИ РАБОТЫ

В реальных условиях функционирования технических объектов и организационно-технических систем зависимость результатов функционирования (откликов) от управляемых и контролируемых воздействий (факторов) проявляется как опосредованная разнообразными случайными причинами (возмущениями). Подобные зависимости принято называть стохастическими.

Корреляционный анализ – это метод математической статистики, который позволяет определить степень взаимосвязи между различными параметрами. Например, между откликом Y и одним из факторов X .

Цели работы:

Изучение методики корреляционного анализа анализа.

Приобретение навыков решения задачи парного и множественного корреляционного анализа в среде электронных таблиц *MS Excel* с применением статистических функций.

Приобретение навыков решения задачи множественного корреляционного анализа в среде *MS Excel* с применением инструмента «Корреляция».

2.КРАТКИЕ СВЕДЕНИЯ ИЗ СТАТИСТИКИ

Стохастическую зависимость, которая проявляется как изменение только математического ожидания отклика, называют корреляционной. Рассматривая корреляционную зависимость отклика от одного фактора, говорят о парной корреляции. Если отклик связан корреляционной зависимостью с несколькими факторами, имеет место множественная корреляция. Характеристикой корреляционной зависимости является статистическая величина, называемая коэффициентом корреляции.

2.1.Парная корреляция

Для парной корреляции, при допущении что и фактор и отклик имеют нормальные распределения, коэффициент корреляции выражается формулой:

$$\rho = \frac{M\left\{\left[x - M(X)\right]\left[y - M(Y)\right]\right\}}{\sqrt{M\left[x - M(X)\right]^2 \; M\left[y - M(Y)\right]^2}} = \frac{K_{XY}}{\sqrt{D(X)D(Y)}}, \qquad (1)$$

где K_{XY} - корреляционный момент. Он представляет собой математическое ожидание произведения отклонений зна-

чений x и y случайных величин X и Y от их математических ожиданий $M(X)$ и $M(Y)$;

$D(X)$ - дисперсия случайной величины X;

$D(Y)$ - дисперсия случайной величины Y.

На практике каждая из случайных величин представляется ограниченным числом значений (выборкой размера n). Поэтому вместо истинного значения коэффициента корреляции ρ моожет быть определена лишь его оценка r, рассчитываемая с использованием выборочных характеристик отклика и фактора:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\,\bar{s}_X\,\bar{s}_Y}; \qquad (2)$$

где $\bar{x}$ и $\bar{y}$ - средние выборочные значения фактора и отклика;

$\bar{s}_X$ и $\bar{s}_Y$ - выборочные стандартные отклонения отклика и фактора;

n - число наблюдений.

Коэффициент корреляции обладает следующими свойствами [4].

1. Он не имеет размерности и поэтому сопоставим для различных статистических рядов.
2. Значение r лежит в интервале от –1 до +1. Если $r = \pm 1$, то зависимость между фактором и откликом является функциональной.
3. Положительное значение коэффициента корреляции указывает на возрастание отклика с увеличением фактора. Отрицательное значение r свидетельствует об убывании Y при возрастании X.
4. Равенство коэффициента парной корреляции нулю не означает отсутствия связи между откликом и фактором. В литературе [1, 3] отмечается, что значение $r = 0$ указывает лишь что эта взаимосвязь не является линейной, но не опровергает возможность существования между ними иной, например экспоненциальной, зависимости.

В *MS Excel* коэффициент парной корреляции можно вычислить с применением статистической функции КОРРЕЛ(). Синтаксис функции:

КОРРЕЛ(*Данные_1*;*Данные_2*),

где *Данные_1* и *Данные_2* - аргументы.

Аргументы должны быть числами, массивами чисел или ссылками на ячейки, содержащие числа. Если аргумент, который является массивом или ссылкой, содержит тексты, логические значения или пустые ячейки, то такие значения игнорируются; однако, ячейки с нулевыми значениями учитываются. Аргументы должны иметь одинаковое количество элементов (точек) данных, в противном случае функция КОРРЕЛ возвращает значение ошибки #Н/Д. Если хотя бы один из аргументов пуст (т. е. не содержит данных), или если стандартное отклонение данных хотя бы в одном из них равно нулю, то функция КОРРЕЛ возвращает значение ошибки #ДЕЛ/0!.

Поскольку коэффициент корреляции вычисляется на основании выборочных данных и является случайной величиной, его значение должно быть проверено на значимость. Смысл проверки состоит в выяснении вопроса: является ли значение $r \neq 0$ случайным событием, или коэффициент корреляции действительно не равен нулю?

Наиболее часто [1-5 и др.] критерием значимости коэффициента парной корреляции принимают условие:

$$t = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2} > t[\alpha; n-2], \qquad (3)$$

где t и $t[\alpha; n-2]$ - рассчитанное и табличное числа Стьюдента. В *MS Excel* для определения табличного числа Стьюдента предусмотрена статистическая функция СТЬЮДРАСПОБР(α; $n-2$).

Возможен также иной подход [6], согласно которому фактическое значение коэффициента парной корреляции r сравнивается с минимальным статистически значимой звеличиной r_{min}:

$$r > r_{min} = \sqrt{\frac{1}{1+\frac{n-2}{(t[\alpha; n-2])^2}}}, \qquad (4)$$

Условия (3) и (4) не являются взаимоисключающими, поскольку получены из одной и той же исходной предпосылки. Если (3) или (4) выполняется, то коэффициент парной корреляции можно считать значимым с доверительной вероятностью $p = 1 - \alpha$.

2.2.Множественная корреляция

Множественная корреляция – обусловленность некоторого признака (например, отклика Y) одновременным действием нескольких других признаков (например, факторов X_1, X_2, … , X_j, … , X_m). При этом возможна парная корреляция среди факторов. Взаимодействия отклика с каждым из факторов и факторов между собой отображают в виде *матрицы корреляции* (рис. 1).

	Y	X_1	...	X_j	...	X_m
Y	1	r_{Y,X_1}	...	r_{Y,X_j}	...	r_{Y,X_m}
X_1	r_{Y,X_1}	1	...	r_{X_1,X_j}	...	r_{X_1,X_m}
...	...	...	1	...	...	...
X_j	r_{Y,X_j}	r_{X_1,X_j}	...	1	....	r_{X_j,X_m}
...	...	...	...	...	1	...
X_m	r_{Y,X_m}	r_{X_1,X_m}	...	r_{X_j,X_m}	...	1

Рис. 1. Матрица корреляции.

Коэффициент множественной корреляции определяют из предположения, что отклик связан с факторами линейной зависимостью. Расчет выполняют по формуле:

$$R = \sqrt{1 - \frac{\Delta_{YX}}{\Delta_{XX}}}, \qquad (5)$$

где Δ_{YX} - определитель матрицы корреляции;

Δ_{XX} - определитель матрицы, получаемой из матрицы корреляции вычеркиванием первой строки и первого столбца.

Значимость множественного коэффициента корреляции проверяют с помощью критерия Фишера (F-критерия):

$$F_p = \frac{R^2}{\left(1 - R^2\right)} \frac{\left(n - m - 2\right)}{m} > F\left[\alpha; m; n - m - 2\right], \qquad (6)$$

где F_p и $F[\alpha; m; n-m-2]$ - рассчитанное и табличное числа Фишера. В *MS Excel* табличное число Фишера можно определить с помощью статистической функции FРАСПОБР($\alpha; m; n-m-2$).

Если условие (6) выполняется, то коэффициент множественной корреляции можно считать значимым с доверительной вероятностью $p = 1-\alpha$.

При анализе степени совместного влияния комплекса факторов на отклик часто используют коэффициент множественной детерминации $D = R^2$. Его значение показывает, на сколько процентов изменчивость отклика обусловлена совместным действием рассматриваемых факторов.

3.ВЫПОЛНЕНИЕ РАБОТЫ

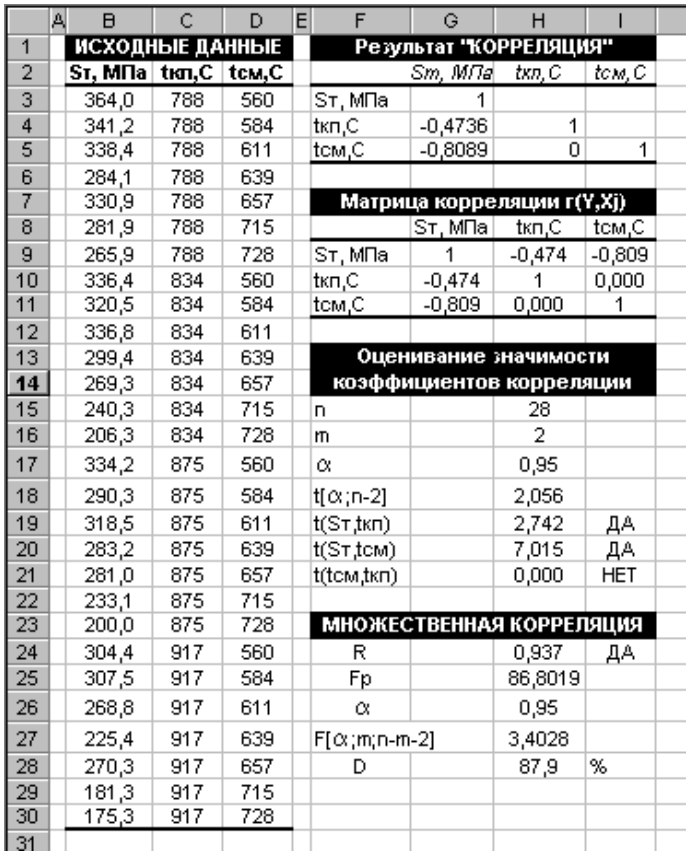
Работа выполняется на отдельном рабочем листе и включает парный и множественный корреляционный анализ как с применением инструмента «Корреляция», так и с использованием статистических функций *MS Excel*. Порядок выполнения работы иллюстрируется на примере анализа взаимосвязи предела текучести металла, прокатанного на широкополосном стане горячей прокатки (ШСГП) от температуры конца прокатки (tкп) и смотки (tсм). Фрагмент рабочего листа с результатами работы представлен на рис.2.

Исходные данные вводятся с клавиатуры (на рис. 2 – в ячейки B3:D30).

3.1.Определение коэффициентов парной корреляции и оценивание их значимости

Применение инструмента «Корреляция». На рис. 2 результаты работы инструмента представлены в ячейках F2:I5. При этом в качестве исходных данных были приняты ячейки B2:D30 (т. е. в область данных включены и обозначения переменных), а в диалоговом окне инструмента «Корреляция» задана опция «Метки данных в первой строке». Более подробно описание работы с инструментом «Корреляция» приведено в приложении.

Использование статистической функции КОРРЕЛ(). На рис. 2 результаты, полученные с использованием функции КОРРЕЛ(), оформлены в виде матрицы корреляции (ячейки F8:I11). Здесь обозначения изучаемых переменных введены с кла-



	B	C	D	F	G	H	I
1	**ИСХОДНЫЕ ДАННЫЕ**			**Результат "КОРРЕЛЯЦИЯ"**			
2	**Sт, МПа**	**tкп,С**	**tсм,С**		*Sm, МПа*	*tкп,С*	*tсм,С*
3	364,0	788	560	Sт, МПа	1		
4	341,2	788	584	tкп,С	-0,4736	1	
5	338,4	788	611	tсм,С	-0,8089	0	1
6	284,1	788	639				
7	330,9	788	657	**Матрица корреляции r(Y,Xj)**			
8	281,9	788	715		Sт, МПа	tкп,С	tсм,С
9	265,9	788	728	Sт, МПа	1	-0,474	-0,809
10	336,4	834	560	tкп,С	-0,474	1	0,000
11	320,5	834	584	tсм,С	-0,809	0,000	1
12	336,8	834	611				
13	299,4	834	639	**Оценивание значимости коэффициентов корреляции**			
14	269,3	834	657				
15	240,3	834	715	n		28	
16	206,3	834	728	m		2	
17	334,2	875	560	α		0,95	
18	290,3	875	584	t[α;n-2]		2,056	
19	318,5	875	611	t(Sт,tкп)		2,742	ДА
20	283,2	875	639	t(Sт,tсм)		7,015	ДА
21	281,0	875	657	t(tсм,tкп)		0,000	НЕТ
22	233,1	875	715				
23	200,0	875	728	**МНОЖЕСТВЕННАЯ КОРРЕЛЯЦИЯ**			
24	304,4	917	560	R		0,937	ДА
25	307,5	917	584	Fp		86,8019	
26	268,8	917	611	α		0,95	
27	225,4	917	639	F[α;m;n-m-2]		3,4028	
28	270,3	917	657	D		87,9	%
29	181,3	917	715				
30	175,3	917	728				
31							

Рис. 2. Пример оформления рабочего листа при корреляционном анализе

виатуры. Для определения коэффициента корреляции между пределом текучести и температурой конца прокатки в ячейках G10 и H9 запрограммировано:

=КОРРЕЛ(B3:B30;C3:C30).

В ячейках I9 и G11 определяется коэффициент парной корреляции между пределом текучести и температурой смотки:

=КОРРЕЛ(B3:B30;D3:D30).

Коэффициент корреляции между температурами конца прокатки и смотки (ячейки I10 и H11) определяется следующим образом:

=КОРРЕЛ(C3:C30;D3:D30).

Оценивание значимости коэффициентов парной корреляции. Для оценивания коэффициентов корреляции необходимо знать число наблюдений n и число факторов m. Число наблюдений определяется в ячейке H15 с использованием статистической функции СЧЕТ():

=СЧЕТ(B3:B30) .

Число факторов задается в ячейку H16 с клавиатуры.

В ячейку H17 с клавиатуры вводится доверительная вероятность α, с которой будет оцениваться значимость результатов парного корреляционного анализа, а в ячейке H18 определяется табличное число Стьюдента:

=СТЬЮДРАСПОБР(1-H17;H15-2) .

В ячейках H19, H20 и H21 определяются числа Стьюдента для соответствующих коэффициентов корреляции, рассчитанные по формуле (3). Например, в ячейке H19:

=ABS(G10)*КОРЕНЬ(H15-2)/КОРЕНЬ(1-G10^2) .

Выводы о значимости коэффициентов корреляции формируются в ячейках G19, G20 и G21 с помощью функции ЕСЛИ(). Например, в G19:

=ЕСЛИ(H19>H18;"ДА";"НЕТ") .

3.2. Определение коэффициента множественной корреляции и его оценивание

Коэффициент множественной корреляции вычисляется в ячейке H24 по формуле (5) с использованием матрицы корреляции, представленной в ячейках F8:I11. В ячейку H24 запрограммировано:

=КОРЕНЬ(1-МОПРЕД(G9:I11)/МОПРЕД(H10:I11)) .

Расчетное число Фишера в ячейке H25 вычисляется по формуле (6):

=H24^28(H15-H16-2)/((1-H24^2)*H16) .

В ячейку H26 с клавиатуры вводится доверительная вероятность α для оценивания значимости результатов множественного корреляционного анализа, а в ячейке H27 определяется табличное число Фишера:

=FРАСПОБР(1-H26;H16;H15-H16-2) .

Вывод о значимости коэффициента множественной корреляции формируются в ячейке G24 с помощью функции ЕСЛИ():

=ЕСЛИ(H25>H27;"ДА";"НЕТ") .

В ячейке G28 рассчитывается коэффициент множественной детерминации:

=100*H24*H24 .

4.СОДЕРЖАНИЕ ВЫВОДОВ ПО РАБОТЕ

В выводах по работе необходимо ответить на следующие вопросы:

1. Связь между какими величинами анализировалась?
2. Какие коэффициенты парной корреляции являются статистически значимыми? О чем это свидетельствует?
3. Является ли значимым коэффициент множественной корреляции? Что это означает?
4. О чем свидетельствует значение коэффициента множественной детерминации?

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА

1. Львовский Е.Н, Статистические методы построения эмпирических формул: Учебн. пособие для втузов, - 2-е изд., перераб. и доп.- М., Высш. шк . , 1988. - 239 с.
2. Математическая статистика: Учебник / Иванова В. М., Калинина В. Н., Нешумова Л. А. и др. – М., Высшая школа, 1981. – 371 с.
3. Четыркин Е. М., Калихман И. Л. Вероятность и статистика. – М., Финансы и статистика, 1982. – 320 с.
4. Ахназарова С.Л,, Кафаров В.В. Организация эксперимента в химии и химической технологии: Учебн. пособие для химико-технологических вузов.- М., Высшая школа , 1978, - 319 с.
5. Замков О. О., Толстопятенко А. В., Черемных Ю. Н. Математические методы в экономике. – М.: МГУ им. М. В. Ломоносова, издательство «ДИС», 1997. – 368 с.
6. Колемаев В. А., Староверов О. В., Турундаевский В. Б. Теория вероятности и математическая статистика. – М.:Высшая школа, 1991. – 400 с.

Содержание

ПРИЛОЖЕНИЕ

ИНСТРУМЕНТ «КОРРЕЛЯЦИЯ» И ЕГО ПРИМЕНЕНИЕ

Инструмент «Корреляция» целесообразно применять для построения матрицы корреляции при проведении множественного корреляционного анализа. В случае парного анализа следует пользоваться статистической функцией КОРРЕЛ().

Запуск инструмента «Корреляция» осуществляется последовательным выбором пунктов из меню различных уровней:

<Сервис> / <Анализ данных> / <Корреляция>.

На экране откроется диалоговое окно (рис. П1).

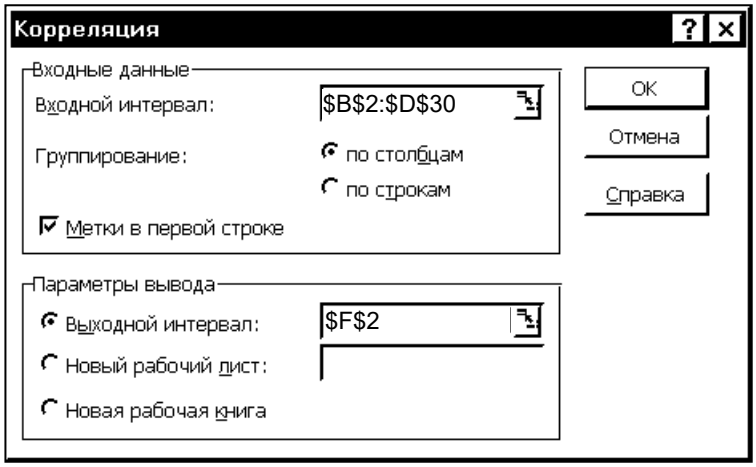


Рис. П1. Диалоговое окно инструмента «Корреляция»

Входной диапазон. Ссылка на ячейки, содержащие анализируемые данные. Ссылка должна состоять как минимум из двух смежных диапазонов данных, организованных в виде столбцов или строк.

Группирование. Указатель особенностей расположения данных. Автоматически предполагается, что данные организованы в виде смежных столбцов (как в примере на рис. 2) и поэтому при запуске инструмента устанавливается «По столбцам». В противном случае необходимо установите переключатель в положение «По строкам».

Метки в первом столбце / Метки в первой строке. Указатель наличия условных обозначений анализируемых величин среди обрабатываемых данных. Если эта опция задана, то данные, записанные в первых ячейках входного диапазона, воспринимаются как условноые обозначения и при анализе не учитываются. Например, при группировке данных по столбцам условными обозначениями будут считаться данные, расположенные в первых строках каждого из них. При группировке по строкам в качестве условных обозначений будут восприниматься данные, расположенные в первых столбцах каждой из них. **Внимание!** Если анализируемые данные не снабжены условным обозначением, то задание опции «**Метки …** » приведет к потере данного из первой ячейки входного диапазона, что повлечет ошибку в расчете выборочных параметров.

Выходной диапазон, Новый лист, Новая книга. Указатели расположения результатов работы инструмента «Описательная статистика». При выборе опции **Выходной диапазон** результаты выводятся на активный рабочий лист активной книги MS Excel. Они будут размещены слева направо и сверху вниз, начиная с ячейки, адрес которой задан в окне (в примере на рис. 2 - начиная с ячейки I3). При выборе опций **Новый лист** или **Новая книга** результаты выводятся в новый рабочий лист активной книги или в первый лист другой книги MS Excel. В обоих случаях результаты будут размещены начиная с ячейки A1.

Результаты работы инструмента «Корреляция» представляют собой матрицу корреляции, усеченную сверху относительно главной диагонали (рис. П2)

	Y	X1	…	Xj	…	Xm
Y	1					
X1	r_{YX1}	1				
…	…	…	…			
Xj	r_{Yxj}	r_{X1Xj}	…	1		
…	…	…	…	…	…	
Xm	r_{YXm}	r_{X1Xm}	…	r_{XjXm}	…	1

Рис. П2. Структура результатов работы инструмента «Корреляция»