Практическая работа 14.
Корреляционный и регрессионный анализ.

Цель работы: Освоение основ проведения корреляционного и регрессионного анализа результатов исследований и навыка их применения для регистрации экспериментальных данных.

Задание:

1 Изучить основы ведения корреляционного и регрессионного анализа.

2 Составить и освоить алгоритм определения параметров линейного однофакторного уравнения регрессии данных.

3 Составить конспект накопления формул по определению величин дисперсии (Д) и среднего квадратичного отклонения (СКО), коэффициентов ковариации и корреляции (КК), и коэффициента детерминации (КД).

4 Описать алгоритм вычисления параметров регрессионного уравнения экспериментальных данных.

1 КРАТКАЯ ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

1.1 Основные понятия корреляционного и регрессионного анализа.

Для решения задач экономического анализа накопленных данных и прогнозирования динамики их изменения во времени часто используются статистические, отчетные или наблюдаемые приемы анализа данных. В таких случаях обычно полагают, что данные являются значениями случайной величины.

Случайной величиной называется переменная величина, которая в зависимости от случая принимает различные значения с некоторой вероятностью. Закон распределения случайной величины показывает частоту ее тех или иных значений в общей их совокупности. При исследовании взаимосвязей между экономическими показателями на основе статистических данных, часто между ними наблюдается стохастическая зависимость. Она проявляется в том, что изменение закона распределения одной случайной величины происходит под влиянием изменения другой. Взаимосвязь между этими двумя величинами могут быть полной (функциональной) и неполной (искаженной другими факторами).

Пример функциональной зависимости – выпуск продукции, и ее потребление в условиях дефицита.

Неполная зависимость наблюдается, например, между стажем рабочих по выпуску продукции и их производительностью труда. Обычно у рабочих с большим стажем работы качество работы, соответственно и изготовленной продукции, значительно отличается относительно молодых рабочих. Но под влиянием дополнительных факторов, таких как образование, здоровье, упорство, быстродействие, такая зависимость может быть искажена.

Раздел математической статистики, посвященный изучению взаимосвязей между случайными величинами, называется корреляционным анализом.

Основная задача корреляционного анализа – это установление характера и тесноты связи между результативными (зависимыми) и факторными (независимыми) показателями (признаками) в рассматриваемом явлении или наблюдаемом процессе. Корреляционную связь можно обнаружить только при массовом сопоставлении фактов.

Характер связи между показателями определяется по корреляционному полю. Если Y - зависимый признак, а X - независимый, то отметив каждый случай $X(i)$ с координатами xi и yi получим корреляционное поле.

Теснота связи определяется с помощью коэффициента корреляции, который рассчитывается специальным образом и лежит в интервалах [-1, +1]. Если значение

коэффициента корреляции лежит в интервале |0, 9| по модулю, то отмечается очень сильная корреляционная зависимость.

В случае, если значение коэффициента корреляции лежит в интервале |0,9, 0,6|, то говорят, что имеет место слабая корреляционная зависимость. Наконец, если значение коэффициента корреляции находится в интервале |-0,6, +0,6|, то говорят об очень слабой корреляционной зависимости или полной ее отсутствии.

Таким образом, корреляционный анализ применяется для нахождения характера и тесноты связи между случайными величинами.

Регрессионный анализ своей целью имеет вывод (выявление или идентификацию) уравнения регрессии, включая статистическую оценку его параметров. Уравнение регрессии позволяет найти значения зависимой переменной, если величина независимых переменных известна.

Практически, речь идет о том, чтобы, анализируя множество точек на множество статистических данных, найти линию, точно отражающую наблюдаемую в этом множестве закономерность (тренд, тенденцию). Эта закономерность обычно в математической статистике называет линией регрессии.

1.2 Виды уравнений регрессии.

По числу факторов различают одно-, двух - и многофакторные уравнения регрессии. По характеру связи однофакторные уравнения регрессии в основном подразделяются на:

а) линейные - $Y = a + bx$;

б) степенные - $Y = a + x^b$;

в) показательные - $Y = a + b^x$.

Обозначениями представлены:

- x - экзогенная (независимая) переменная;
- Y - эндогенная (зависимая, результативная) переменная;
- a, b – параметры.

1.3 Параметры линейного однофакторного уравнения регрессии.

В таблице 6.1 представлены данные о доходах (X) и спрос на некоторый товар (Y) за ряд лет (n).

Таблица 6.1.

Данные для составления
однофакторного уравнения регрессии

Год, n	Доход, X	Спрос, Y
1	x_1	y_1
2	x_2	y_2
3	x_3	y_3
...	...	...
N	x_n	y_n

Предположим, что между X и Y существует линейная взаимосвязь, т.е. $Y = a + bx$. Для того, чтобы найти уравнение регрессии, прежде всего, нужно исследовать тесноту связи между случайными величинами X и Y, т.е. корреляционную зависимость. В таблице 6.1 указаны параметры: x_1, x_2, ... x_n - совокупность значений независимого, факторного признака; y_1, y_2, … y_n – совокупность соответствующих значений зависимого, результативного признака; n – количество наблюдений.

Для нахождения уравнения регрессии вычисляются следующие величины:

1) средние значения: $\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n}$ - для экзогенной переменной,

$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ - для эндогенной переменной;

2) отклонения от средних величин: $\Delta x_i = x_i - \bar{x}$, $\Delta y_i = y_i - \bar{y}$.

Дисперсия и среднеквадратичное отклонение вычисляются по формулам: $D_x = \frac{\sum_{i=1}^{n} \Delta x_i^2}{n-1}$, $D_y = \frac{\sum_{i=1}^{n} \Delta y_i^2}{n-1}$, $\sigma_x = \sqrt{D_x}$, $\sigma_y = \sqrt{D_y}$.

Величины дисперсии и среднего квадратичного отклонения характеризуют разброс наблюдаемых значений вокруг среднего значения. Чем больше дисперсия, тем больше разброс.

1.4 Корреляционный момент (коэффициент ковариации).

Корреляционный момент вычисляется по формуле:

$$K_{xy} = \frac{\Delta x_1 \cdot \Delta y_1 + \Delta x_2 \cdot \Delta y_2 + \ldots + \Delta x_n \cdot \Delta y_n}{n-1} = \frac{\sum_{i=1}^{n} \Delta x_i \cdot \Delta y_i}{n-1}.$$

Корреляционный момент отражает характер взаимосвязи между *x* и *y*. Если $K_{xy} > 0$, то взаимосвязь прямая. Если $K_{xy} < 0$, то взаимосвязь обратная.

1.5 Коэффициент корреляции.

Коэффициент корреляции вычисляется по формуле: $R_{xy} = \frac{K_{xy}}{\sigma_x \sigma_y}$. Доказано, что коэффициент корреляции находится в интервале от минус единицы до плюс единицы ($-1 \le R_{xy} \le 1$).

1.6 Коэффициент детерминации

Коэффициент корреляции в квадрате (R_{xy}^2) называется коэффициентом детерминации.

Если $R_{xy} \ge |0.8|$, то вычисления продолжаются.

1.7 Параметры регрессионного уравнения.

Коэффициент *b* находится по формуле: $b = \frac{K_{xy}}{D_x}$;

После чего легко найти параметр *a*: $a = \bar{y} - b\bar{x}$. Коэффициенты *a* и *b* находятся методом наименьших квадратов, основная идея которого состоит в том, что за меру суммарной погрешности принимается сумма квадратов разности (остатков) между фактическими значениями результативного признака y_i и его расчетными значениями y_{ip}, полученными при помощи уравнения регрессии: $y_{ip} = a + bx_i$.

При этом величины остатков находятся по формуле: $u_i = y_i - y_{ip}$, где y_i - фактическое значение *y*; y_{ip} - расчетное значение *y*.

Пример 1.1. Дано: Пусть будут накоплены статистические данные (таблица 6.2) о доходах (*X*) и спросе (*Y*).

Найти: Корреляционную зависимость между ними и определить параметры уравнения регрессии.

Таблица 6.2.

Данные для составления
однофакторного уравнения регрессии

Год, *n*	Доход, *X*	Спрос, *Y*
1	10	6
2	12	8
3	14	8
4	16	10,3
5	18	10,5
6	20	13

Предположим, что между нашими величинами существует линейная зависимость. Результаты вычислений свести в таблицу 6.2.

Таблица 6.2

Параметры линейного однофакторного уравнения регрессии

Показатели	**Доход, *X***	**Спрос, *Y***
Среднее значение	15 (90/6 = 15)	9,3 (55.8/6=9.3)
Дисперсия	14.6 (73/5 = 14.6)	6,08 (30.43/5 = 6.08)
Среднее квадр. отклонение	3,7417 (3.821)	2,4658
Корреляционный момент	8,96 (9.02)	
Коэффициент корреляции	0,9712 (0.957)	
Параметры	*b* = 0,62	*a* = 0,03

В итоге уравнение регрессии примет вид: $y = -0.03 + 0.62x$. Используя это уравнение, можно найти расчетные значения *Y* и построить график регрессии (рисунок 6.1).

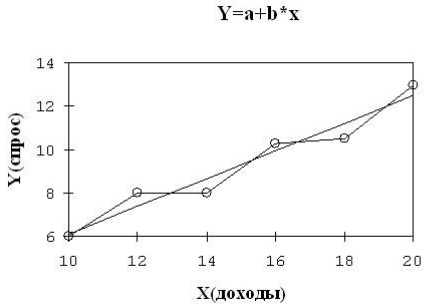

Рисунок 6.1. Фактические и расчетные значения *Y*

Ломаная линия на графике отражает фактические значения *Y*, а прямая линия построена с помощью уравнения регрессии и отражает тенденцию изменения спроса в зависимости от дохода.

2 ПОРЯДОК ВЫПОЛНЕНИЯ РАБОТЫ

Построить уравнение регрессии по данным, представленным в таблице 6.3. Решение задачи представить в письменном виде.

Таблица 6.3.

Данные для составления уравнения регрессии

№№	Индексы розничных цен на продукты питания, *x*	Индекс промышленного производства, *y*
1	100	70
2	105	79
3	108	85
4	113	84
5	118	85
6	118	85
7	110	96
8	115	99
9	119	100
10	118	98
11	120	99
12	124	102
13	129	105
14	132	112

1 Для характеристики зависимости *y* от *x* рассчитать параметры следующих функций: линейной; степенной; равносторонней гиперболы.

2 Для каждой модели рассчитать показатели: тесноты связи и среднюю ошибку аппроксимации.

3 Оценить статистическую значимость параметров регрессии и корреляции.

4 Выполнить прогноз значения индекса промышленного производства *y* при прогнозном значении индекса розничных цен на продукты питания $x = 138$.

3. ОТЧЕТ ДОЛЖЕН СОДЕРЖАТЬ

3.1 Наименование и цель работы.
3.2 Условие задания (полный текст заданий).
3.3 Программные средства, используемые при выполнении работы.
3.4 Описание выполненной работы согласно требованиям преподавателя:
- формулировка решения о наилучшем использовании трудовых ресурсов;
- формулировка решения о максимальном доходе работника;
- формулировка решения о рационе питания работника.
3.5 Сформулированные выводы и составленное заключение о проведении работы.
3.6 Список использованной литературы.

4 КОНТРОЛЬНЫЕ ВОПРОСЫ

4.1 Дайте определение термину «случайная величина».
4.2 Для чего определяется стохастическая зависимость?
4.3 Приведите функциональной и неполной зависимости.
4.4 Дайте определение термину «корреляционный анализ».
4.5 С какой целью проводится корреляционный анализ?
4.6 Как определяется теснота связи?
4.7 Перечислите условия сильной и слабой корреляционной зависимости.
4.9 Какая цель проведения регрессионного анализа?
4.10 Перечислите виды уравнений регрессии.
4.11 Как определяются средние значения для эндогенной и экзогенной переменной?
4.12 Как определяются величины дисперсии (по x и y) и среднего квадратичного отклонения (СКО)?
4.13 Как определяются коэффициенты ковариации и детерминации?
4.15 Как определяются параметры регрессионного уравнения?