МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ
ВЫСШЕГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ»

Ю. М. Фетисов, А. Э. Крупко

МНОГОМЕРНЫЙ КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

Учебное пособие

Воронеж
Издательский дом ВГУ
2015

ВВЕДЕНИЕ

Использование многомерного регрессионно-корреляционного анализа находит место во многих аспектах различных исследованиях и является одним из наиболее употребляемых методов изучения статистических закономерностей. Поэтому это учебное пособие может с успехом использоваться в преподавании курсов «Общая и социально-экономическая статистика», «Математическая статистика», «Математические методы исследования региона», «Информатика» для студентов факультета географии, геоэкологии и туризма Воронежского университета, обучающихся по специальностям «Экология», «Природопользование», «География». Цель пособия – помочь студентам научится осмысленно применять регрессионно-корреляционный анализ в учебных и научных исследованиях, что достигается сочетанием обычного математического и кибернетического подходов (с помощью ЭВМ). Этот анализ подразумевает выявление основных факторов развития. Большое значение имеет выявление статистических связей между явлениями и построения на этой основе моделей. Статистическая связь не проявляется в каждом случае, а как правило, при большом числе наблюдений, что обусловливает использование многомерного анализа. Частным случаем статистической связи является корреляционная связь. Для исследования корреляционной связи необходимо, чтобы выполнялись три условия: 1) наличие данных по достаточно большой совокупности; 2) наблюдалась определенная однородность совокупности; 3) распределение признаков совокупности по нормальному закону распределения. До построения статистической модели анализируются средние величины группировок, вариация признаков. Особое значение имеют среднее квадратическое отклонение и дисперсия. Важным моментом в исследовании, часто имеющее самостоятельное значение, является выявление корреляционных связей. В основе теории корреляции лежит представление о тесноте связи между изучаемыми явлениями. Наиболее

$$y = \frac{1}{\delta\sqrt{2\pi}} e^{-\frac{(x-\lambda)^2}{2\delta^2}},$$

где δ – среднее квадратичное отклонение, e – основание натуральных логарифмов, x – значение переменной, λ – математическое ожидание, или $\bar{x}$, $\pi = 3{,}14$.

Кривую нормального распределения можно использовать для описания большинства распределений, существующих в природе. Вследствие зависимости от величины стандартного отклонения, которое служит удобной мерой рассеяния данных относительно среднего значения, нормальное распределение находит применение в параметрических статистических методах, являясь фундаментом, на котором строятся корреляционный и регрессионный анализы. Любое отклонение данных от нормального закона распределения делает статистические выводы недостоверными. Для «нормализации» кривой можно осуществить преобразование данных с помощью некоторой функции (например, взять логарифм или возвести в квадрат).

Критерий t Стьюдента используется для проверки гипотезы о том, что среднее в выборке $(\bar{x})$ может служить оценкой среднего во всей генеральной совокупности (μ). Критерий определяется по формуле:

$$t = \frac{|\mu - \bar{x}|}{\sigma / \sqrt{n}}.$$

Вычисленные значения t следует сравнить с табличными теоретическими значениями при различных уровнях значимости.

1.2. Вычисление средних величин

Средняя арифметическая величина позволяет выявить наиболее существенные черты, характерные для всей статистической совокупности:

$$\overline{X} = \frac{\sum x}{n},$$

где $\sum$ – сумма, x – отдельное значение признака, n – число наблюдений.

Пример 1.1. Число элементов n ряда составляет 20 объектов: 2, 15, 6, 24, 13, 18, 7, 63, 51, 12, 9, 15, 43, 26, 23, 17, 11, 8, 21, 44. Сумма этого ряда равна 428, разделив на 20, получим $\overline{X} = 21{,}4$

При большом количестве изучаемых показателей среднюю величину проще вычислять по формуле **средневзвешенной**:

$$\overline{\overline{X}} = \frac{\sum_{1}^{n} xm}{\sum_{1}^{n} m} = \frac{x_1 m_1 + x_2 m_2 + \ldots x_n m_n}{m_1 + m_2 + \ldots m_n},$$

где x – центральные значения интервалов, а m – частоты.

Пример 1.2. По данным таблицы 1.1. рассчитаем среднюю величину

Таблица 1.1. Доходы населения с общей численностью 1000 чел.

Среднедушевой доход (тыс. руб)	Число жителей f	Середина интервала x	xf	$x^2 f$
1	2	3	4	5
до 2,0	160	1,5	240	360
2,0 – 4,0	275	3,0	825	2475
4,0 – 6,0	240	5,0	1200	6000
6,0 - 8,0	156	7,0	1092	7644
8,0 - 10,0	109	9,0	981	8829
свыше 10,0	60	11,0	660	7260
Итого	1000	-	4998	32568

Получим $\bar{x} = \frac{4998}{1000} = 5{,}0$; $\sigma = \sqrt{\frac{32568}{1000} - 25} = 2{,}8.$

Если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменной сумму квадратов исходных

величин, то средняя будет являться **квадратической средней** величиной:

$$\overline{x}_{кв.} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}}$$

Пример 1.3. Имеется два участка земли со сторонами квадрата в 1 км и в 5 км.

Найдем величину стороны квадратного участка среднего по площади для этих двух участков. Очевидно, что 3 км не подходит, потому что его площадь будет 9 кв. км., а средняя площадь 13 кв.км. По формуле ср. кв. находим

$$\overline{x}_{кв.} = \sqrt{\frac{1^2+5^2}{2}} = \sqrt{13} = 3{,}61.$$

Пример 1.4. Имеется несколько квадратных участков земли (три из которых имеют стороны по 1 км, два участка со сторонами по 3 км, один участок – 6 км, и пять участков по 4 км). Используя средневзвешенную квадратическую величину, надо найти размер стороны среднего для них по площади участка. По формуле ср. кв. вз. находим $\overline{x}_{кв.} = \sqrt{\frac{1^2 3 + 3^2 2 + 6^2 + 4^2 5}{11}} = \sqrt{12{,}45} = 3{,}53.$

Соответственно, если при замене индивидуальных величин признака на среднюю величину необходимо сохранить неизменным произведение индивидуальных величин, то следует применять **геометрическую среднюю** величину. Ее формула такова:

$$\overline{x}_{геом} = \sqrt[n]{x_1 x_2 \ldots x_n}$$

Пример 1.5. Выпуск в основных ценах хозяйства страны составил за 2010-2013 годы соответственно 105,0%, 104,3%, 103,7% и 101,5%. Средняя арифметическая даст 103,625%, что неверно, потому что сравнению абсолютных показателей даст другой результат. Поэтому необходимо использовать геометрическую среднюю величину. Она будет равна:

$$\overline{x}_{геом} = \sqrt[4]{105} * 104{,}3 * 103{,}7 * 101{,}5 = 103{,}616$$

Если по условиям задачи необходимо, чтобы неизменной оставалась при осреднении сумма величин, обратных индивидуальным значениям признака, то средняя величина является **гармонической**. Формула ее:

$$\overline{x}_{гарм.} = \frac{n}{\sum_{i=1}^{1} \frac{1}{x_i}}$$

Пример 1.6. Автомобиль с грузом движется со скоростью 50 км в час, обратно без груза со скоростью 100 км в час. Расстояние между пунктами равно s. Время поездки $\frac{s}{x_1}+\frac{s}{x_2}=\frac{s}{\overline{x}}+\frac{s}{\overline{x}}$, сократив на s, получим $\frac{1}{x_1}+\frac{1}{x_2}=\frac{1}{\overline{x}}+\frac{1}{\overline{x}}$. Подставляя x_1 и x_2 $\frac{1}{50}+\frac{1}{100}=\frac{2}{\overline{x}}$, получим

$$\overline{x}_{гарм} = \frac{2}{\frac{1}{50}+\frac{1}{100}} = \frac{2\ *100}{3} = 66{,}67 \text{ км/час}$$

Все рассмотренные выше виды средних величин принадлежат к общему типу степенных средних. Различаются они лишь показателем. **Степенная средняя** есть корень к-й степени из частного от деления суммы индивидуальных значений признака в к-й степени на число индивидуальных значений:

$$\overline{x}_{степ.} = \sqrt[k]{\frac{\sum_{i=1}^{n} x_i^k}{n}}$$

При к = 1 получаем арифметическую среднюю, при к = 2 – квадратическую, при к = 3 – кубическую, при к = 0 – геометрическую, при к = –1 – гармоническую среднюю. При этом наблюдается следующее соотношение, которое называется правилом **мажорантности** средних:

$$\bar{X}_{гарм} \leq \bar{X}_{геом} \leq \overline{x}_{арифм.} \leq \overline{x}_{кв.} \leq \overline{x}_{куб.}$$

1.3. Вычисление основных показателей вариации

Для оценки колеблемостей значения изучаемого признака вводятся особыпоказатели – **лимиты**, которые характеризуют максимальное и минимальное значение признаков. Разность между лимитами ряда составляет его **размах**: lim = max. – min.

В то же время более точно степень развития признака выражается таким показателем, как **среднее абсолютное отклонение**:

$$\theta = \frac{\sum_{1}^{n} \left| x - \overline{x} \right|}{n},$$

но чаще используется другой показатель степени разнообразия – **среднее квадратическое отклонение**:

$$\sigma = \sqrt{\frac{\sum_{1}^{n} \left| x - \overline{x} \right|^2}{n}}.$$

При большом числе n можно использовать и более простую формулу:

$$\sigma = \frac{\sum_{1}^{n} x^2}{n} - \overline{x}^2.$$

Вычисление среднего квадратического отклонения при группированных данных производится по формуле **средневзвешенной**:

$$\sigma = \sqrt{\frac{\sum_{1}^{n} \left| x - \overline{x} \right|^2 m}{\sum_{1}^{n} m}},$$

где m – вес (частота).

Дисперсия (D) равна σ^2 и служит мерой рассеяния данных относительно среднего арифметического. Дисперсия может быть простая и взвешенная. **Общая дисперсия** измеряет вариацию признака во всей совокупности под влиянием всех факторов, обусловливающих эту вариацию