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

Ю.М. Фетисов

МЕТОДЫ РЕГРЕССИННОГО И КОРРЕЛЯЦИОННОГО АНАЛИЗА В ГЕОГРАФИИ И ГЕОЭКОЛОГИИ

Учебно-методическое пособие для вузов

Издательский дом ВГУ
2014

ПРЕДИСЛОВИЕ

Важные разделы статистического исследования связаны с корреляционным и регрессионным анализом. Для выявления статистических связей между анализируемыми признаками применяют корреляционный анализ. Общее назначение множественной регрессии состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной (называемой откликом или критерием). Регрессионный анализ позволяет математически описать формулу зависимости между исследуемыми признаками.

В STADIA содержится большой набор процедур множественного корреляционного и регрессионного анализа. Их грамотное использование требует от географа и геоэколога как знания по теории множественного корреляционного и регрессионного анализа, так и владения навыками практической работы в среде этого пакета.

Пособие содержит начальные сведения о пакете STADIA, позволяющие освоить основные приемы работы с массивами данных, а также научиться практическому использованию трех модулей пакета:.множественной линейной регрессии, пошаговой регрессии и общей (нелинейной) регрессии.

Для успешного освоения каждого из этих модулей вначале даётся краткий теоретический материал, а затем рассматривается типовая задача с указателем всех пунктов меню, приводящих к ее решению. С целью проверки знаний и закрепления навыков в пособии приведены задания для самостоятельной работы. Такая структура оказывается очень удобной при приеме зачетов по каждой теме, а так же при использовании пособия для самостоятельного изучения.

4. Диалог. Далее протекает диалог, характерный для выполняемого метода с выдачей числовых результатов анализа и их интерпретации в экранную страницу [*Rez*] текстового редактора, а графиков результатов – в графические страницы [Gr_i], i=1,...15.

Текстовый редактор становится доступным при активизации страницы результатов анализа [*Rez*]. Он поддерживает большинство типичных для подобных редакторов операций: ввод текста с клавиатуры, выделение фрагментов текста, удаление символов (только клавишей *Back Space*) и фрагментов, забор фрагментов в буфер обмена и вставление фрагментов из буфера. Редактор поддерживает также общие операции: чтение и запись в отношении текстовых файлов, изменение шрифта и выдачи результатов на печать.

Для переноса числовых результатов анализа в электронную таблицу необходимо забирать их в *буфер обмена* из страницы результатов, перейти в электронную таблицу и в нужном месте вставить содержимое буфера обмена.

Полученные в графической форме результаты могут быть перенесены в электронную таблицу посредством специальной инструментальной клавиши «*СохрГраф*.». Координаты точек каждого графика переносятся в первые свободные столбцы электронной таблицы.

Система STADIA выпускается в четырех модификациях (учебная, студенческая, базовая и профессиональная), отличающихся только объемом обрабатываемых данных (соответственно 400, 4000, 20000 и 32000 чисел совокупно в матрице данных). Учебная STADIA с файлами примеров свободно доступна по адресу : http://statsoft.msu.ru/Stadia.zip. Это дает возможность студентам использовать данный пакет вне учебных компьютерных классов. STADIA постоянно развивается, появляются более совершенные версии этой системы, открывая новые возможности для пользователей.

2. МНОЖЕСТВЕННЫЙ КОРРЕЛЯЦИОННЫЙ И РЕГРЕССИОННЫЙ АНАЛИЗ

2.1. Теоретические основы.

Множественный корреляционный анализ является одним из методов статистического анализа взаимосвязи нескольких признаков. Он применяется тогда, когда данные наблюдений можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Исходной для анализа является матрица

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1k} \\ x_{21} & x_{22} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

размерностью (n×k), которая представляет собой n наблюдений для каждого из k факторов.

Сначала находят парные коэффициенты корреляции, характеризующие тесноту линейной зависимости между двумя переменными на фоне действия всех остальных показателей, входящих в модель. Они изменяются в пределах от –1 до +1, причем чем ближе коэффициент корреляции к ±1, тем сильнее зависимость между переменными.

Коэффициент парной корреляции вычисляют по формуле

$$r_{jl} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)\cdot(x_{il} - \overline{x}_l)}{s_j \cdot s_l};\ j,l = 0,1,...,k,$$

где $$\overline{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ s_j = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}\ .$$

Здесь r_{jl} – коэффициент корреляции между одним из факторов x_j и фактором x_l (j, l = 1,2, … ,k), r_{ol} – коэффициент корреляции между результативным признаком y и одним из факторов x_l.

Если один из коэффициентов r_{jl} (j, l = 1,2, … ,k) окажется равным 1, то это означает, что факторы x_j и x_l функционально (не вероятностно) связаны

между собой и тогда целесообразно один из них исключить из рассмотрения, причем оставляют тот фактор, у которого коэффициент r_{oi} больше.

После вычисления всех парных коэффициентов корреляции и исключения из рассмотрения того или иного фактора можно построить корреляционную матрицу:

$$R = \begin{pmatrix} 1 & r_{01} & r_{02} & \cdots & r_{0k} \\ r_{10} & 1 & r_{12} & \cdots & r_{1k} \\ r_{20} & r_{21} & 1 & \cdots & r_{2k} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ r_{k0} & r_{k1} & r_{k2} & \cdots & 1 \end{pmatrix}.$$

Матрица R является симметрической и положительно определенной.

Используя корреляционную матрицу, можно вычислить частные коэффициенты корреляции, которые характеризуют тесноту линейной зависимости между двумя переменными при исключении влияния всех остальных показателей, входящих в модель. Например, частный коэффициент корреляции (k-1)–го порядка между y и x_1 равен:

$$r_{01/2,3,\ldots,k} = -\frac{R_{01}}{\sqrt{R_{00} \cdot R_{11}}},$$

где R_{jl} – алгебраическое дополнение элемента r_{jl} корреляционной матрицы R.

Для изучения тесноты связи между результативным признаком y и несколькими факторами $x_1, x_2, \ldots, x_k$ используют множественный коэффициент корреляции r_0. Множественный коэффициент корреляции характеризует тесноту связи между одной результативной переменной и остальными, входящими в модель; r_0 всегда положителен и изменяется от 0 до 1. Множественный коэффициент корреляции также служит и для оценки качества предсказания. Чем больше r_0, тем лучше качество предсказаний данной моделью опытных данных. Квадрат множественного коэффициента корреляции называется множественным коэффициентом детерминации. Он характеризует долю дисперсии, результативной переменной, обусловленной влиянием факторов, входящих в модель.

Множественный коэффициент корреляции определяется по формуле:

$$r_0 = \sqrt{1 - \frac{|R|}{R_{00}}},$$