МИНИСТЕРСТВО ПО РАЗВИТИЮ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ И КОММУНИКАЦИЙ РЕСПУБЛИКИ УЗБЕКИСТАН

ТАШКЕНТСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ ИМЕНИ МУХАММАДА АЛ-ХОРАЗМИЙ

Факультет "Компьютерный инжиниринг"

Кафедра "Компьютерные системы"

А.И.Назаров, С.С.Мирзахалилов, Н.А.Сайфуллаева, С.Б.Довлетова

МЕТОДИЧЕСКИЕ УКАЗАНИЯ

для выполнение практических работ по предмету «Биоинформатика и биомеханика»

ТАШКЕНТ 2020

УДК 519.711.2

Авторы: доцент Назаров А.И., старший преподаватель Мирзахалилов С.С., старший преподаватель Сайфуллаева Н.А., ассистент С.Б.Довлетова «Биоинформатика и биомеханика» / ТУИТ. 44с. Ташкент, 2020

В методических указаниях описаны практические упражнения, выполненные на основе статистических пакетов обработки данных «Statistica». Практические упражнения содержат расчет параметров описательной статистики, корреляционный, дисперсионный, факторный, регрессионный и кластерный анализ.

Приведены методы построения линейных и нелинейных многопараметрических моделей и оценка их точности, а также методы оценки эффективности математических моделей.

Адресовано студентам, обучающимся по направлению «Компьютерный инжиниринг», а также специалистам, которые занимаются статистической обработкой информации.

Рецензенты:

М.Якубов - д.т.н., профессор кафедры «Информационные технологии», ТУИТ.

С.Ташев - начальник отдела по организации учебных курсов Узбекско-Индийского центра ИТ.

Введение

Методические указания предназначены для обработки и интерпретации статистических данных. Многие традиционно изучаемые в статистике понятия, такие как частотные распределения, вероятность, проверка значимости, корреляция, дисперсия, регрессия имеют широкую область применения.

Цель методических указаний научить выявлять связи, скрытые в числовых данных, получать обобщенные характеристики, удобные для интерпретации. Преобразование данных – это радикальный подход к статистическим методам и проблемам. Основной принцип при анализе данных состоит в необходимости добиваться результатов, позволяющих их использовать при последующем математическом моделировании.

В данном методическом указании приводятся практические занятия на основе известных статистических пакетов обработки данных «Statistica». Для облегчения усвоения материала каждая тема освещается с описанием метода и примерами расчетов.

Методическое указание ориентировано на студентов, изучающих применения статистических методов математического моделирования при решении задач технической биоинформатики и биомеханики, а также может быть использовано и студентами других специальностей, изучающих применение статистических пакетов обработки информации для решения биологических, медицинских и экологических задач.

Практическое занятие № 1.
Выборки и их представление

Основные понятия

Выборкой $x_1, ..., x_n$ объема n из совокупности, распределенной по $F(x)$, называется n независимых наблюдений над случайной величиной ξ с функцией распределения $F(x)$.

Вариационным рядом $x_{(1)} \le x_{(2)} \le ... \le x_{(n)}$ называется выборка, записанная в порядке возрастания ее элементов.

Каждому наблюдению из выборки присвоим вероятность, равную $1/n$; получим распределение, которое называют **эмпирическим**; ему соответствует **функция эмпирического распределения**

$$F_n^*(x) \equiv F_n^*(x; x_1, ..., x_n) = \frac{\mu_n(x)}{n}, \qquad (1.1)$$

где: $\mu_n(x)$ - число членов выборки, меньших x. Значение этой функции для статистики определяется тем, что при $n \to \infty$

$$F_n^*(x) \to F(x)$$

Выборки больших объемов труднообозримы и требуют разбивки диапазона значений выборки на равные интервалы и подсчета для каждого интервала **частоты**- количество наблюдений, попавших в него. Частоты, отнесенные к общему числу наблюдений n, называют **относительными частотами**; графическое представление распределения частот по интервалам – **гистограммой**; **накопленной частотой** для данного интервала называют сумму частот данного интервала и всех тех, что левее его [1].

Выполнение в пакете STATISTICA

Генерация выборки

Сгенерируем, например, выборку объема n =50 с показательным распределением со средним значением 5.

Создадим новый файл:

- *File - New Data* - укажем имя файла в окне *File Name: descript.* На экране сетка-таблица, в ее заголовке указаны название и размеры: *10v * 10c* - (10 переменных (variables) - столбцов по 10 наблюдений (cases) - строк.

Преобразуем таблицу к размерам 1×50:

- Кнопка *Vars* (на экране) - *Delete;* окно *Delete Variables:* укажем какие переменные- столбцы убрать : *From variable : var 2, To variable : var 10 - OK;*
-Кнопка *Cases - Add* (добавление) - окно *Add Cases:* укажем, сколько строк добавить и куда : *Number of Cases to Add : 40, Insert after Case :* - OK.

Сгенерируем выборку:

выделим столбец - переменную *Var1* (щелчком мыши по ее заглавию) - нажмем правую клавишу - в открывшемся меню выберем *Variable specs* (спецификации переменной) - в появившемся окне *Variable 1* введем *Name x*, в

нижнем поле *Long name* вводится выражение, определяющее переменную. Ввод можно сделать набором на клавиатуре или с помощью клавиши *Functions*, выбирая в меню *Kategory* и *Name* требуемую функцию и вставляя клавишей *Insert*. Для задания закона распределения следует ввести, например, *=rnd(2)* для $R[0, 2]$,

=Vnormal(rnd(1); 2; 0.5) для $N(2, \sigma^2=0.5^2)$,

=VExpon(rnd(1); 0.2) для *E(5)* со средним $1/0.2=5$; (для нашего примера вместо значения параметра $\lambda=0.2$ можно набрать выражение 1/5).

Такая форма задания определяется способом генерации: с помощью функции, обратной (буква *V*) к функции распределения и генератора случайных чисел $R[0, 1]$ *(rnd(1))*.

Распечатаем выборку командой *Print* меню *File*.

Посмотрим выборку графически:

Graphs - Custom Graphs (настраиваемые графики) - *2D graphs* - в открывшемся окне все можно оставить по умолчанию - *.OK*. Наблюдаемый график (рис. 1.1) распечатаем.

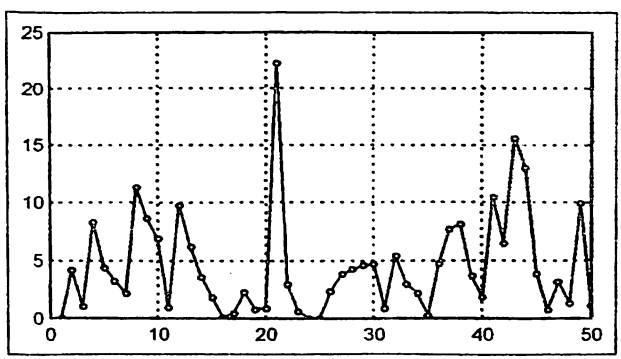

Рис. 1.1. Наблюдения, распределенные по показательному закону со средним 5 (n = 50).

Построение вариационного ряда

Первый способ.

Выделим требуемую переменную (столбец) - нажмем правую клавишу мыши - выберем *Quiq Stats Graphs* (быстрые статистики и графики) - *Values / Stats of Vars* (значения и статистики) - наблюдаем вариационный ряд и выборочное среднее (*mean*) и стандартное отклонение (*SD*).

Второй способ.

Войдем в модуль *Data Menagement* (двойной щелчек левой клавишей мыши на чистом поле и выбор модуля в окне *Module Switcher*; если модуль уже загружен, то *Alt- Tab* до появления модуля) - *Analysis Sort* - устанавлива-

ем имя переменной, тип сортировки: *Ascen* (по возрастанию) или *Desc* (по убыванию) - *OK*.

Функция эмпирического распределения

Graphs - Stats 2D Graphs - Histogram - в появившемся окне установим: *Graph Type : Regular, Cumulative Counts* (накопленные частоты), *Fit Type* (подбираемый тип) : *Exponential* (для нашего примера) или *off* (без подбора). *Variablles: x, Categories* (число интервалов группирования) : 250 - *OK*.

Наблюдаем график функции эмпирического распределения (рис. 1.2). График можно отредактировать: изменить линии, точки, фон, шкалы, надписи; для этого необходимо подвести стрелку в нужное место и дважды щелкнуть левой клавишей мыши. Выведем его на печать или сохраним.

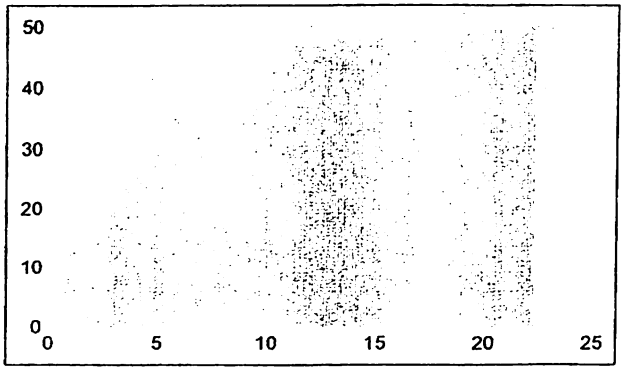

Рис. 1.2. Функция эмпирического распределения

Группирование данных

Analysis Frequency Tables - в окне *Frequency Tables* зададим *No of exact intervals: 10* (10 интервалов группирования; или *Step size: 2, starting at: 0*), в поле *Display options* отметим *Cumulative frequences* (накопленные частоты), *Percentages* (проценты - относительные частоты), *Cumulative Percentages* (накопленные частоты) - *OK*.

Наблюдаем таблицу группированных данных. Выведем ее на печать или сохраним.

Построение гистограммы частот

Graphs - Stats 2D Graphs - Histograms - в появившемся окне устанавливаем: имя переменной, *Graph Type: Regular, Fit Type; off* (без подбора) или нужный тип, число интервалов группирования *Categories:* или *Auto* (автоматический выбор числа интервалов) - *OK*.

Наблюдаем гистограмму (рис.1.3). Отредактируем график, если необходимо. Выведем на печать или сохраним.

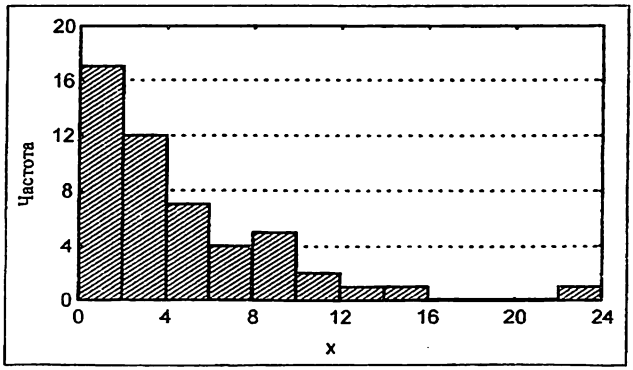

Рис. 1.3. Гистограмма

Вопросы:

1. Параметры объекта и его представление.
2. Шкалы измерений.
3. Методы сбора исходной информации.
4. Ошибка выборки. Виды ошибок выборки.

Практическое занятие № 2.
Параметры описательной статистики

Числовые характеристики эмпирического распределения называются **выборочными характеристиками**: выборочные **среднее** (математическое ожидание), **дисперсия**:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i , \quad s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (2.1)$$

выборочный **момент** порядка k:

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k ; \qquad (2.2)$$

выборочные **квантили** ζ_p порядка p - корни уравнения

$$F(\zeta_p) = p,$$

которыми являются члены вариационного ряда

$$\zeta_{(p)} = \xi_{([np]+1)},$$

где: *[np]* означает целую часть *np*; частным случаем (p = 0.5) является выборочная **медиана** - центральный член вариационного ряда. Значение выборочных характеристик состоит в том, что при $n \to \infty$ они стремятся к истинным значениям распределения $F(x)$ [1,6].

Выполнение в пакете STATISTICA

Выборочные характеристики

Первый способ: на заголовке столбца с выборкой щелкнем правой клавишей мыши - *Quick Basic Stats... - Descriptives of var* - получаем таблицу с характеристиками: *mean* (среднее), *Confid 95%* (доверительные границы нижняя и верхняя с уровнем доверия 0.95), *Sum* (сумма), *Minimum, Maximum, Range* (размах), *Variance* (дисперсия), *Std. Dev.* (стандартное отклонение) и др. Сравним выборочное среднее, медиану и стандартное отклонение с соответствующими теоретическими значениями. Это же можно сделать через меню: *Anflisis - Quick Basic Stats ...*

Второй способ: на заголовке столбца с выборкой щелкнем правой клавишей мыши - *Block Stats Columns* (блок статистик по колонкам) - выделим необходимое или *All.*

Описание двумерных выборок

Ввод данных: зададим новую таблицу 2×32, назовем столбцы *X* и *Y*. Заполним таблицу вручную заданными в табл.2 значениями.

Диаграмма рассеяния:

Graphs - Stats 2D Graphs... - Scatterplots... - вводим значения по осям *X* и *Y* (нажав на кнопку *Variables* и выбрав переменные) - *OK.*

Распечатаем диаграмму (рис. 2.1) или сохраним.

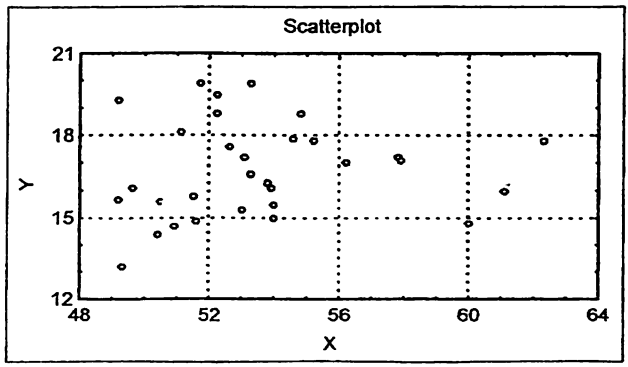

Рис. 2.1. Диаграмма рассеяния

Выборочные характеристики

Выделим те переменные, по которым требуются выборочные характеристики - щелкнем правой клавишей мыши - *Quick Basic Stats - Descriptivs of VARS...* Наблюдаем таблицу выборочных характеристик. Отпечатаем таблицу или сохраним.

Выборочные характеристики можно внести в таблицу данных, в конец соответствующих столбцов. Выделим нужные столбцы, далее см. вторую часть п. Определим корреляционную матрицу:

Analysis - Correlation matrices - Two lists - First list: All - Second list: All - OK - Cancel (отмена предложения на новую матрицу).

Матрицу отпечатаем или сохраним.

Двумерная гистограмма.

Graphs - Stat 3D Sequential Graphs - Bivariate Gistogram - установим по осям *X* и *Y* требуемые переменные (кнопкой *Variables*), зададим число интервалов по каждой оси - *OK*.

Вопросы:

1. Описательная статистика. Мода. Медиана.
2. Среднее значение.
3. Дисперсия, средне-квадратическое отклонение, стандартное ошибка.
4. Основные правила графического оформления.

Практическое занятие №3.
Математическая обработка данных с целью определения структуры и параметров математической модели

Основные определения. Корреляция представляет собой меру зависимости переменных. Наиболее известна корреляция Пирсона. При вычислении корреляции Пирсона предполагается, что переменные измерены, как минимум, в интервальной шкале. Некоторые другие коэффициенты корреляции могут быть вычислены для менее информативных шкал. Коэффициенты корреляции изменяются в пределах от -1.00 до +1.00. Обратите внимание на крайние значения коэффициента корреляции. Значение -1.00 означает, что переменные имеют строгую *отрицательную* корреляцию. Значение +1.00 означает, что переменные имеют строгую *положительную* корреляцию. Отметим, что значение 0.00 означает отсутствие корреляции [2].

Отрицательная корреляция. Две переменные могут быть связаны таким образом, что при возрастании значений одной из них значения другой убывают. Это и показывает отрицательный коэффициент корреляции. Про такие переменные говорят, что они отрицательно коррелированы.

Положительная корреляция. Связь между двумя переменными может быть следующей - когда значения одной переменной возрастают, значения другой переменной также возрастают. Это и показывает положительный коэффициент корреляции. Про такие переменные говорят, что они положительно коррелированы.

Наиболее часто используемый коэффициент корреляции *Пирсона r* называется также *линейной* корреляцией, т.к. измеряет степень линейных связей между переменными.

Простая линейная корреляция (Пирсона r). Корреляция Пирсона (далее называемая просто *корреляцией*) предполагает, что две рассматриваемые переменные измерены, по крайней мере, в интервальной шкале, с которой значения двух переменных "пропорциональны" друг другу. Важно, что значение коэффициента корреляции не зависит от масштаба измерения. Например, корреляция между ростом и весом будет одной и той же, независимо от того, проводились измерения в *дюймах* и *футах* или в *сантиметрах* и *килограммах*. *Пропорциональность* означает просто *линейную зависимость*. Корреляция высокая, если на графике зависимость "можно представить" прямой линией (с положительным или отрицательным углом наклона).

Проведенная прямая называется *прямой регрессии* или прямой, построенной *методом наименьших квадратов*. Последний термин связан с тем, что сумма *квадратов* расстояний (вычисленных по оси Y) от наблюдаемых точек до прямой является минимальной. Заметим, что использование *квадратов* расстояний приводит к тому, что оценки параметров прямой сильно реагируют на выбросы (см. рис.3.1.).

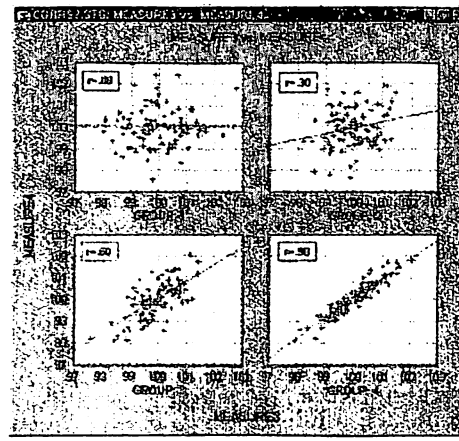

Рис.3.1. Отображение линии регрессии

Как интерпретировать значения корреляций. Коэффициент корреляции Пирсона (*r*) представляет собой меру линейной зависимости двух переменных. Если возвести его в квадрат, то полученное значение коэффициента детерминации r^2) представляет долю вариации, общую для двух переменных (иными словами, "степень" зависимости или связанности двух переменных). Чтобы оценить зависимость между переменными, нужно знать как "величину" корреляции, так и ее *значимость*.

Значимость корреляций. Уровень значимости, вычисленный для каждой корреляции, представляет собой главный источник информации о надежности корреляции. Значимость определенного коэффициента корреляции зависит от объема выборок. Критерий значимости основывается на предположении, что распределение остатков (т.е. отклонений наблюдений от регрессионной прямой) для зависимой переменной *y* является нормальным (с постоянной дисперсией для всех значений независимой переменной *x*). Исследования методом Монте-Карло показали, что нарушение этих условий не является абсолютно критичным, если размеры выборки не слишком малы, а отклонения от нормальности не очень большие. Тем не менее, имеется несколько серьезных опасностей, о которых следует знать, для этого см. следующие разделы.

Выбросы. По определению, выбросы являются нетипичными, резко выделяющимися наблюдениями. Так как при построении прямой регрессии используется сумма *квадратов* расстояний наблюдаемых точек до прямой, то выбросы могут существенно повлиять на наклон прямой и, следовательно, на значение коэффициента корреляции. Поэтому единичный выброс (значение которого возводится в квадрат) способен существенно изменить наклон прямой и, следовательно, значение корреляции (см.3.2.).

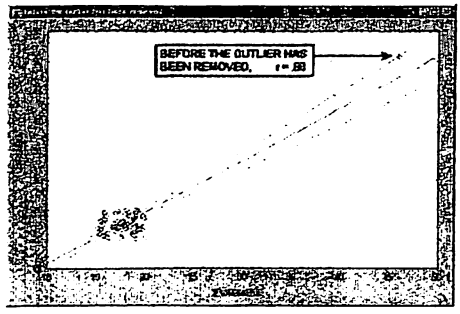

Рис.3.2. Зависимость линии регрессии от выбросов данных

Заметим ,что если размер выборки относительно мал, то добавление или исключение некоторых данных (которые, возможно, не являются "выбросами", как в предыдущем примере) способно оказать существенное влияние на прямую регресии (и коэффициент корреляции). Это показано в следующем примере, где мы назвали исключенные точки "выбросами"; хотя, возможно, они являются не выбросами, а экстремальными значениями.

Обычно считается, что выбросы представляют собой случайную ошибку, которую следует контролировать. К сожалению, не существует общепринятого метода автоматического удаления выбросов (тем не менее, см. следующий раздел). Чтобы не быть введенными в заблуждение полученными значениями, необходимо проверить на *диаграмме рассеяния* каждый важный случай значимой корреляции. Очевидно, выбросы могут не только искусственно увеличить значение коэффициента корреляции, но также реально уменьшить существующую корреляцию (см.3.3.).

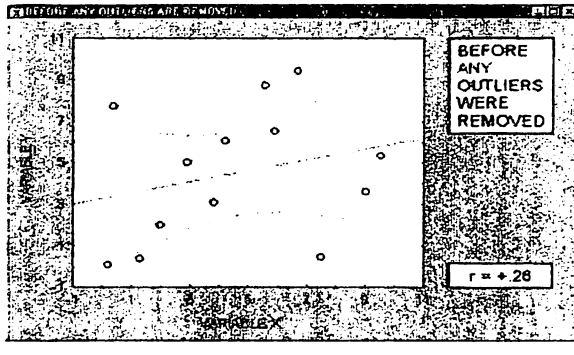

Рис.3.3. Влияние выбросов на корреляцию

Выполнение в пакете STATISTICA

После того как сформирован файл данных. Следует его загрузить в " STATISTICA " и можно приступать к расчетам. Корреляционный анализ запускается (см. рис.3.4.)

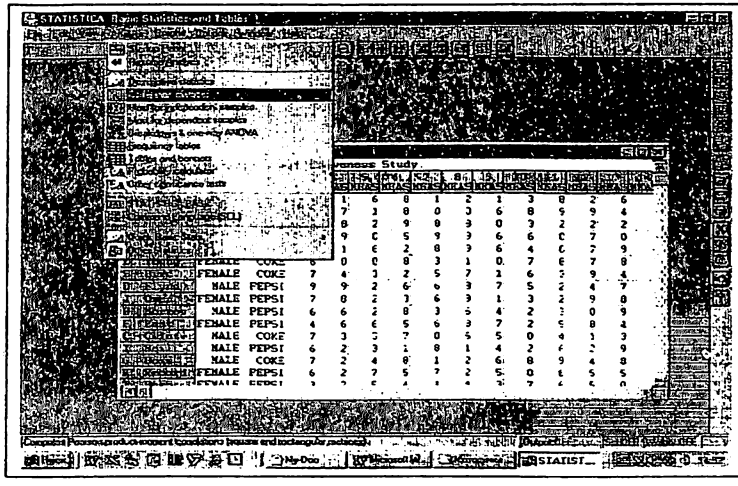

Рис.3.4. Загрузка в статистике корреляционного анализа

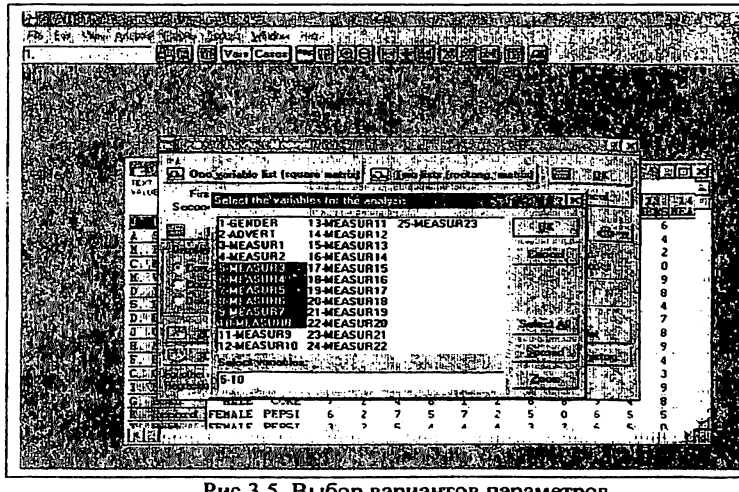

Рис.3.5. Выбор вариантов параметров

Нажимаем опцию «One variable list ». В результате появится окно, в котором следует отметить параметры, корреляционная зависимость которых нас интересует и нажимаем «ОК» (см рис.3.5.). В появившемся окне нажимаем опцию «Correlation» и получаем результат в виде корреляционной матрицы. Достоверные связи помечаются красным цветом.

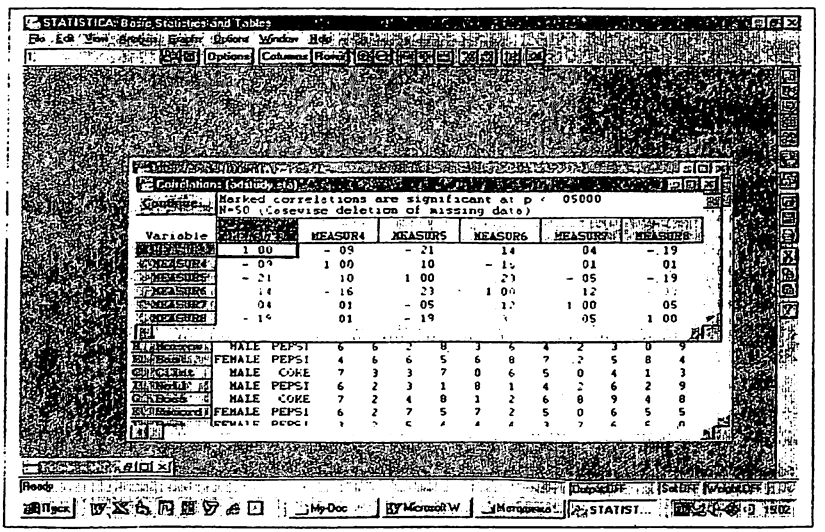

Рис.3.7. Корреляционная матрица

Анализ полученных результатов производиться согласно теории корреляционного анализа и результаты излагаются в схемы.

Вопросы:

1. Основные задачи регрессионного анализа.
2. Построение линейных моделей методом наименьших квадратов.
3. Преимущества оценок метода наименьших квадратов

Практическое занятие № 4.
Однофакторный дисперсионный анализ

Основные соотношения. Изучается влияние, которое оказывает некоторый качественный признак (фактор) на количественный результат (отклик), например, влияние технологии изготовления прибора на его долговечность, влияние способа обработки земли на урожайность и т.д. Пусть фактор имеет k уровней $A_1, A_2, ..., A_k$ и пусть измеряемая величина x есть результат действия фактора и случайной составляющей ε (от фактора не зависящей) [3]:

$$x = f(A) + \varepsilon$$

Будем считать:

1) что при каждом уровне A_j фактора, $j = 1, ..., k$, имеется n_j измерений

$$x_{ij} = a_j + \varepsilon_{ij}, \qquad i = 1, ..., n_j, \tag{4.1}$$

где обозначено $a_j = f(A_j)$.

2) что случайная составляющая ε нормально распределена $N(0, \sigma^2)$ с дисперсией σ^2. Если влияния фактора нет, то все a_j равны. Итак, имеется k выборок объемами $n_1, ..., n_k$, $\sum_{j=1}^{k} n_j = N$. Проверим гипотезу об отсутствии влияния:

$$H: \; a_1 = a_2 = ... = a_k$$

По каждой из выборок методом наибольшего правдоподобия оценим средние a_j и дисперсию σ^2:

$$\hat{a}_j = \bar{x}_j \equiv \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij}, \qquad s_j^2 = \frac{1}{n_j}\sum (x_{ij} - \bar{x}_{\cdot j})^2, \tag{4.2}$$

а затем оценим σ^2 по всем выборкам:

$$\sigma^{2*} = \frac{1}{N-k}\sum_{j=1}^{k} n_j s_j^2 = \frac{1}{N-k}\sum_{j=1}^{k}\sum_{j=1}^{n_j} (x_{ij} - \bar{x}_{\cdot j})^2. \tag{4.3}$$

Эта статистика несмещенно оценивает σ^2 независимо от того, верна или нет гипотеза H.

Другую оценку для σ^2 построим по значениям $\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_k$. Если H верна, то $M\hat{a}_j = a$, $D\hat{a}_j = \sigma^2 / n_j$, $j = 1, ..., k$. Оценки для a и σ^2:

$$\mathcal{E} = \frac{1}{N}\sum_{j=1}^{k} n_j \mathcal{E}_j, \qquad \sigma^{2**} = \frac{1}{k-1}\sum_{j=1}^{k} n_j (\hat{a}_j - \hat{a})^2 \tag{4.4}$$

Из теоремы о совместном распределении оценок среднего и дисперсии нормальной совокупности следует, что статистики $(N-k)\sigma^{2*}$ и $(k-1)\sigma^{2**}$ независимы и распределены как $\sigma^2\chi^2_{N-k}$ и $\sigma^2\chi^2_{k-1}$ соответственно, и потому их отношение

$$F_H = \frac{\sigma^{2**}}{\sigma^{2*}} = \frac{\sigma^2 \chi^2_{k-1}/(k-1)}{\sigma^2 \chi^2_{N-k}/(N-k)}, \qquad (4.5)$$

если гипотеза H верна, имеет F-распределение Фишера.

Если гипотеза не верна, то σ^{2**} имеет тенденцию к увеличению за счет разброса средних a_j, и потому, если F_H имеет слишком большое значение, т.е. если

$$F_H > Q, \qquad (4.6)$$

то гипотеза H об отсутствии влияния фактора A отклоняется, и следует считать, что среди средних a_1, a_2, ..., a_k имеются хотя бы два не равных; здесь $Q = Q(1-\alpha; k-1, N-k)$ - квантиль уровня $1-\alpha$ F-распределения с $k-1$ и $N-k$ степенями свободы, α - выбираемый уровень значимости. Если же (3.6) не выполняется, то это означает, что наблюдения не противоречат гипотезе об отсутствии влияния фактора. Условие (3.6) может быть записано иначе:

$$P\{F \geq F_H\} < \alpha, \qquad (4.7)$$

где F - случайная величина, распределенная по закону Фишера.

Оценка влияния фактора. Отношение $\frac{\hat{a}_j - a}{\sigma^*}\sqrt{n_j}$ подчиняется распределению Стьюдента с $N-k$ степенями свободы, и если $Q = Q(1-\alpha, N-k)$ - квантиль уровня $1-\alpha$ этого распределения, то доверительный интервал для a_j с уровнем доверия $1-2\alpha$:

$$\hat{a}_j \pm \frac{\sigma^*}{\sqrt{n_j}} Q \qquad (4.8)$$

Если гипотеза H о равенстве средних отклоняется. то следует определить, по каким именно уровням фактора средние значимо различаются. Линейная комбинация

$$L = \sum_{j=1}^{k} c_j a_j \qquad (4.9)$$

называется линейным контрастом. Оценка для L :

$$\tilde{L} = \sum_{j=1}^{k} c_j \bar{x}_j, \qquad (4.10)$$

а оценка дисперсии $D\tilde{L}$:

$$S_L^2 = \sigma^{2*} \sum_{j=1}^{k} \frac{c_j^2}{n_j} \qquad (4.11)$$

Зафиксируем произвольное число r контрастов $L^{(1)}, L^{(2)}, \ldots, L^{(r)}$. Можно показать, что одновременно для всех $\tilde{L}^{(1)}, \ldots, \tilde{L}^{(r)}$ выполняются соотношения:

$$\left| L^{(m)} - \bar{L}^{(m)} \right| < S_L^{(m)} \sqrt{kF(1-\alpha, k, N-k)} \qquad (4.12)$$
$m=1, \ldots, r$

с вероятностью 1-α. Это соотношение позволяет сделать вывод о всех интересующих нас контрастах одновременно. В частности, среди разностей $a_j - a_i$ можно выделить те, которые значительно отличаются от нуля на выбранном уровне значимости (метод Шеффе).

Пример. Разработаны две новые методики лечения заболевания T_1 и T_2. Чтобы оценить, какая из них более эффективна они были применены в 10 клиниках, включая традиционную T_0. Результаты хорошей эффективности приведены в табл. 4.1. Проверим гипотезу об отсутствии влияния методик лечения на эффективность выздоровления.

Табл. 4.1

№	T_0	T_1	T_2	№	T_0	T_1	T_2
1	43	71	51	6	43	65	70
2	47	82	62	7	64	74	74
3	67	65	73	8	47	86	65
4	52	68	62	9	56	66	69
5	70	81	48	10	43	60	54

Выполнение в пакете STATISTICA

Будем выполнять в модуле *Basic Statistics and Tables* (можно выполнять также в модуле *ANOVA/MANOVA*). Создадим таблицу с двумя столбцами *P* и *T* и 30 строками; в *P* занесем данные по производительности, в *T* - уровни *T*: технологии *T0*, *T1*, *T2*. Далее выполним:

One - Way ANOVA (Analys Of Variances) - Analysis: Detailed Analysis Of Individual tables, Variabbles: Grouping variabbles (группирующие переменные): *T, Dependent variabbles* (зависимые переменные - отклики): *P - OK - OK* - отметив *Statistics: Number of observations* (количество наблюдений), *Standart deviations* (стандартные отклонения) и *Variances* (дисперсии), получим *Summary table of means* (таблицу средних); видно, как отличаются средние в каждой из групп (при фиксированном уровне фактора *T*) - Возвращаемся в окно *Descriptive Stats and ... Results* и выполняем *Analysis of Variance* - Наблюдаем таблицу: в столбце *SS (Sum of Squares) Effect* указана сумма квадратов, умноженная на (*k* - 1), *df* = 2 = *k* - 1 - число степеней свободы, *MS (Mean Square)* = 839.0 - оценка, *SS* = 2711 - сумма квадратов, умноженная на *(N - k)*, *df* = 27 = *N - k*, *Ms Error* = 100.4 - оценка, *F* = 8.35 - значение статистики, *p* = 0.0015 - вероятность; последняя слишком мала, чтобы поверить в истинность гипотезы *H* об отсутствии влияния фактора *T*. Вывод: фактор *T* (технология) влияет на *P* (производительность).

Возникает вопрос: какой метод лечения можно считать значимо различным? Для ответа на этот вопрос возвращаемся в окно *Descriptive Stats and ... Results* и выполняем *Post - hoc comparasion of means* (сравнение средних) по

методу Шеффе *Sheffe test.* Наблюдаем таблицу, в которой указаны уровни значимости гипотез о равенстве средних для всех пар уровней фактора *T*; видим, что методы лечения *T*0 и *T*1 следует считать различными (вероятность 0.0015 слишком мала, чтобы поверить в равенство средних по *T*0 и *T*1).
Создадим таблицу с тремя столбцами (T_0 – традиционный метод лечения , T_1 - новая методика №1, T_2 - новая методика №2).

Анализ выполняем в модуле *ANOVA/MANOVA:*

Vaariables - Independent Vaariables (factors): A, B Dependent Vaariable list: X - OK - OK -Specific effects (спецификация влияний): выделяем (при двухфакторном анализе) факторы *A* и *B - All effects* - Наблюдаем таблицу *Summery* of *All effects* (итоги по всем влияниям); в столбце *MS Effects* (средние квадраты) оценки σ_{T0}. σ_{T1}. σ_{T2}. Указываются значения статистик Фишера *F* (дисперсионные отношения) и уровни значимости *p*.

Вопросы:

1. Цель дисперсионного анализа (ANOVA).
2. Дисперсия. Средне-квадратическое отклонение. Стандартная ошибка.
3. Основные задачи дисперсионного анализа.
4. Понятие значимого различия (критерии t и Фишера).

Практическое занятие №5.
Использование факторного анализа в биоинформатике

Подтверждающий факторный анализ. Моделирование структурными уравнениями (SEPATH) позволяет проверять частные гипотезы о факторной структуре для множества переменных (подтверждающий факторный анализ) в одной или нескольких выборках (например, вы сможете сравнить факторные структуры разных выборок (опытов).

Анализ соответствий. Анализ соответствий - это описательные/разведочные методы, предназначенные для анализа двух- и многовходовых таблиц, содержащих некоторые взаимосвязи между строками и столбцами. Результаты этого анализа дают информацию, похожую на ту, которую предоставляет факторный анализ, и позволяют изучить структуру категориальных переменных, входящих в таблицу [3].

Объединение двух переменных в один фактор. Зависимость между переменными можно обнаружить с помощью диаграммы рассеяния. Полученная путем подгонки линия регрессии дает графическое представление зависимости. Если определить новую переменную на основе линии регрессии, изображенной на этой диаграмме, то такая переменная будет включить в себя наиболее существенные черты обеих переменных. Итак, фактически, вы сократили число переменных и заменили две одной. Отметим, что новый фактор (переменная) в действительности является линейной комбинацией двух исходных переменных.

Анализ главных компонент. Пример, в котором две коррелированные переменные объединены в один фактор, показывает главную идею факторного анализа или, более точно, анализа главных компонент (это различие будет обсуждаться позднее). Если пример с двумя переменными распространить на большее число переменных, то вычисления становятся сложнее, однако основной принцип представления двух или более зависимых переменных одним фактором остается в силе.

Выделение главных компонент. В основном процедура выделения главных компонент подобна *вращению, максимизирующему дисперсию (варимакс)* исходного пространства переменных. Например, на диаграмме рассеяния вы можете рассматривать линию регрессии как ось *X*, повернув ее так, что она совпадает с прямой регрессии. Этот тип вращения называется *вращением, максимизирующим дисперсию*, так как критерий (цель) вращения заключается в максимизации дисперсии (изменчивости) "новой" переменной (фактора) и минимизации разброса вокруг нее (см. *Стратегии вращения*).

Обобщение на случай многих переменных. В том случае, когда имеются более двух переменных, можно считать, что они определяют трехмерное "пространство" точно так же, как две переменные определяют плоскость. Если вы имеете три переменные, то можете построить 3М диаграмму рассеяния (см.5.1.).

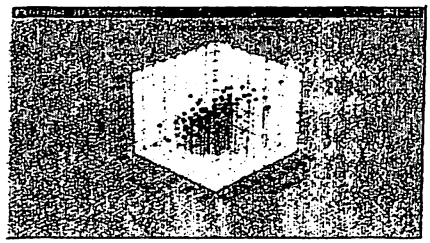

Рис.5.1. 3М диаграмма рассеяния

Для случая более трех переменных, становится невозможным представить точки на диаграмме рассеяния, однако логика вращения осей с целью максимизации дисперсии нового фактора остается прежней.

Несколько ортогональных факторов. После того, как вы нашли линию, для которой дисперсия максимальна, вокруг нее остается некоторый разброс данных. И процедуру естественно повторить. В анализе главных компонент именно так и делается: после того, как первый фактор *выделен*, то есть, после того, как первая линия проведена, определяется следующая линия, максимизирующая остаточную вариацию (разброс данных вокруг первой прямой), и т.д. Таким образом, факторы последовательно выделяются один за другим. Так как каждый последующий фактор определяется так, чтобы максимизировать изменчивость, оставшуюся от предыдущих, то факторы оказываются независимыми друг от друга. Другими словами, некоррелированными или *ортогональными*.

Выполнение в пакете STATISTICA. В пакете ***STATISTICA*** для проведения факторного анализа следует войти в опции "Analisys/Other Statistics . В результате появиться окно вида:

Нажимая на опцию «Factor Analis» можно приступать к анализы данных методами факторного анализа (см.рис.5.2.).

Посмотрим теперь на некоторые стандартные результаты анализа главных компонент. При повторных итерациях можно выделить факторы с все меньшей и меньшей дисперсией. Для простоты изложения считаем, что обычно работа начинается с матрицы, в которой дисперсии всех переменных равны *1.0*. Поэтому общая дисперсия равна числу переменных. Например, если имеется 10 переменных, каждая из которых имеет дисперсию *1*, то наибольшая изменчивость, которая потенциально может быть выделена, равна 10 раз по 1.

Предположим, что при изучении степени удовлетворенности жизнью включено 10 пунктов для измерения различных аспектов удовлетворенности домашней жизнью и работой.

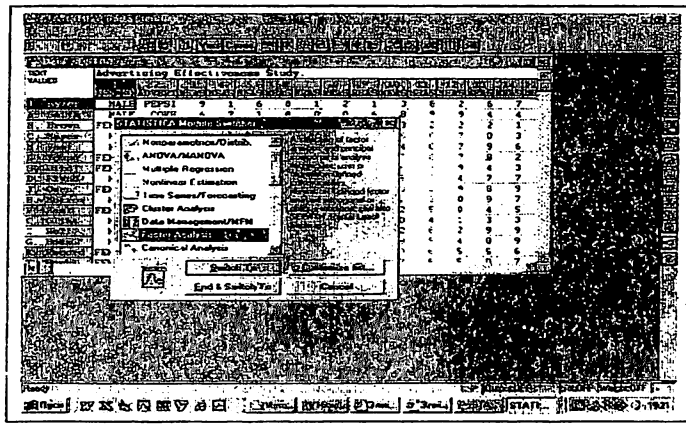

Рис.5.2. Выбор факторного анализа

Дисперсия, объясненная последовательными факторами, представлена в следующей таблице 5.1.

Таблица 5.1. Факторы дисперсии

STATISTICA ФАКТОРНЫЙ АНАЛИЗ	Собственные значения (factor.sta) Выделение: Главные компоненты		
Значение	Собственные значения	% общей дисперсии	Кумулят. соб. знач.
1	6.118369	61.18369	6.11837
2	1.800682	18.00682	7.91905
3	.472888	4.72888	8.39194
4	.407996	4.07996	8.79993
5	.317222	3.17222	9.11716
6	.293300	2.93300	9.41046
7	.195808	1.95808	9.60626
8	.170431	1.70431	9.77670
9	.137970	1.37970	9.91467
10	.085334	.85334	10.00000

Собственные значения. Во втором столбце (*Собственные значения)* таблицы результатов вы можете найти дисперсию нового, только что выделенного фактора. В третьем столбце для каждого фактора приводится процент от общей дисперсии (в данном примере она равна *10*) для каждого фактора. Как можно видеть, первый фактор (значение 1) объясняет 61 процент общей дисперсии, фактор 2 (значение 2) - 18 процентов, и т.д. Четвертый столбец содержит накопленную или кумулятивную дисперсию. Дисперсии, выделяемые факторами, названы *собственными значениями*. Это название происходит из использованного способа вычисления.

Собственные значения и задача о числе факторов. Как только получена информация о том, сколько дисперсии выделил каждый фактор, вы можете возвратиться к вопросу о том, сколько факторов следует оставить. Как говорилось выше, по своей природе это решение произвольно. Однако имеются некоторые общеупотребительные рекомендации, и на практике следование им дает наилучшие результаты.

Критерий Кайзера. Сначала можно отобрать только факторы, с собственными значениями, большими *1*. По существу, это означает, что если фактор не выделяет дисперсию, эквивалентную, по крайней мере, дисперсии одной переменной, то он опускается. Этот критерий предложен Кайзером (Kaiser, 1960), и является, вероятно, наиболее широко используемым. В приведенном выше примере на основе этого критерия вам следует сохранить только 2 фактора (две главные компоненты).

Критерий каменистой осыпи. *Критерий каменистой осыпи* является графическим методом, впервые предложенным Кэттелем. Собственные значения, представленные в таблице ранее, показаны в виде графика (рис.5.3.).

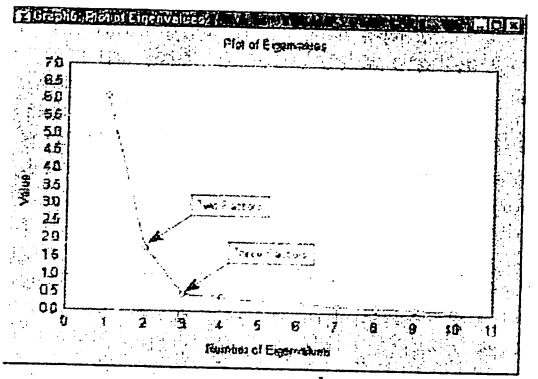

Рис.5.3. Отображение собственных значений

Кэттель предложил найти такое место на графике, где убывание собственных значений слева направо максимально замедляется. Предполагается, что справа от этой точки находится только "факториальная осыпь" - "осыпь" является геологическим термином, обозначающим обломки горных пород, скапливающиеся в нижней части скалистого склона. В соответствии с этим критерием можно оставить в этом примере 2 или 3 фактора.

Какой критерий следует использовать. Теоретически, можно вычислить их характеристики путем генерации случайных данных для конкретного числа факторов. Тогда можно увидеть, обнаружено с помощью используемого критерия достаточно точное число существенных факторов или нет. С использованием этого общего метода первый критерий (*критерий Кайзера*) иногда сохраняет слишком много факторов, в то время как второй критерий (*крите-*

рий каменистой осыпи) иногда сохраняет слишком мало факторов; однако оба критерия вполне хороши при нормальных условиях, когда имеется относительно небольшое число факторов и много переменных. На практике возникает важный дополнительный вопрос, а именно: когда полученное решение может быть содержательно интерпретировано. Поэтому обычно исследуется несколько решений с большим или меньшим числом факторов, и затем выбирается одно наиболее "осмысленное". Этот вопрос далее будет рассматриваться в рамках вращений факторов.

Анализ главных факторов. Прежде, чем продолжить рассмотрение различных аспектов вывода анализа главных компонент, введем анализ главных факторов. Вернемся к примеру вопросника об удовлетворенности жизнью, чтобы сформулировать другую "мыслимую модель". Вы можете представить себе, что ответы субъектов зависят от двух компонент. Сначала выбираем некоторые подходящие общие факторы, такие как, например, "удовлетворение своим хобби", рассмотренные ранее. Каждый пункт измеряет некоторую часть этого общего аспекта удовлетворения. Кроме того, каждый пункт включает уникальный аспект удовлетворения, не характерный для любого другого пункта.

Общности. Если эта модель правильна, то вы не можете ожидать, что факторы будут содержать всю дисперсию в переменных; они будут содержать только ту часть, которая принадлежит общим факторам и распределена по нескольким переменным. На языке факторного анализа доля дисперсии отдельной переменной, принадлежащая общим факторам (и разделяемая с другими переменными) называется *общностью*. Поэтому дополнительной работой, стоящей перед исследователем при применении этой модели, является оценка общностей для каждой переменной, т.е. доли дисперсии, которая является общей для всех пунктов. Доля дисперсии, за которую отвечает каждый пункт, равна тогда суммарной дисперсии, соответствующей всем переменным, минус общность. С общей точки зрения в качестве оценки общности следует использовать множественный коэффициент корреляции выбранной переменной со всеми другими (для получения сведений о теории множественной регрессии сошлемся на раздел *Множественная регрессия*). Некоторые авторы предлагают различные итеративные "улучшения после решения" начальной оценки общности, полученной с использованием множественной регрессии; например, так называемый метод MINRES (метод минимальных факторных остатков; Харман и Джоунс (Harman, Jones, 1966)), который производит испытание различных модификаций факторных нагрузок с целью минимизации остаточных (необъясненных) сумм квадратов.

Вопросы:

1. Цель и задачи факторного анализа.
2. Основная идеология факторного анализа - сокращения числа переменных.
3. Критерий Кайзера.
4. Критерий каменистой осыпи.

Практическое занятие № 6.
Линейный регрессионный анализ

В линейный регрессионный анализ входит широкий круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных [4]:

$$X \equiv (x_1, ..., x_p) \text{ и } Y = (y_1, ..., y_m) \qquad (6.1)$$

Предполагается, что X - независимые переменные (факторы, объясняющие переменные) влияют на значения Y - зависимых переменных (откликов, объясняемых переменных). По имеющимся эмпирическим данным (X_i, Y_i), $i = 1, ..., n$ требуется построить функцию $f(X)$, которая приближенно описывала бы изменение Y при изменении X:

$$Y \approx f(X) \qquad (6.2)$$

Предполагается, что множество допустимых функций, из которого подбирается $f(X)$, является параметрическим:

$$f(X) = f(X, \theta), \qquad (6.3)$$

где θ - неизвестный параметр (вообще говоря, многомерный). При построении $f(X)$ будем считать, что

$$Y = f(X, \theta) + \varepsilon, \qquad (6.4)$$

где первое слагаемое - закономерное изменение Y от X, а второе - ε - случайная составляющая с нулевым средним; $f(X, \theta)$ является условным математическим ожиданием Y при условии известного X и называется ***регрессией Y по X.***

1. Простая линейная регрессия

Пусть X и Y одномерные величины; обозначим их x и y, а функция $f(x, \theta)$ имеет вид $f(x, \theta) = A \cdot bx$, где $\theta = (A, b)$. Относительно имеющихся наблюдений (x_i, y_i), $i = 1, ..., n$, полагаем, что

$$y_i = A - bx_i - \varepsilon_i, \qquad (6.5)$$

где $\varepsilon_1, ..., \varepsilon_n$ - независимые (ненаблюдаемые) одинаково распределенные случайные величины. Можно различными методами подбирать "лучшую" прямую линию. Широко используется ***метод наименьших квадратов.*** Построим оценку параметра $\theta = (A, b)$ так, чтобы величины

$$e_i = y_i - f(x_i, \theta) = y_i - A - bx_i,$$

называемые остатками, были как можно меньше, а именно, чтобы сумма их квадратов была минимальной:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - A - bx_i)^2 = \min \text{ по } (A, b) \qquad (6.6)$$

Чтобы упростить формулы, положим в (4.2) $x_i = x_i - \bar{x} + \bar{x}$; получим:

$$y_i = a + b(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, ..., n,$$

где $\bar{x} = \sum_{i=1}^{n} x_i / n$, $a = A + b\bar{x}$. Сумму $\sum_{i=1}^{n} \left(y_i - a - b(x_i - \bar{x})\right)^2$ минимизируем по (a,b), приравнивая нулю производные по a и b; получим систему линейных уравнений относительно a и b. Ее решение $(\hat{a}, \hat{b})$ легко находится:

$$\hat{a} = \bar{y}, \qquad \text{где } \bar{y} = \sum_{i=1}^{n} y_i / n, \tag{6.7}$$

$$\hat{b} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}. \tag{6.8}$$

Свойства оценок. Нетрудно показать, что если $M\varepsilon_i = 0$, $D\varepsilon_i = \sigma^2$, то

1) $M\hat{a} = a$, $M\hat{b} = b$, т.е. оценки несмещенные;

2) $D\hat{a} = \sigma^2 / n$, $D\hat{b} = \sigma^2 / \sum_{i=1}^{n} (x_i - \bar{x})^2$;

3) $cov(\hat{a}, \hat{b}) = 0$;

если дополнительно предположить нормальность распределения ε_i, то

4) оценки $\hat{a}$ и $\hat{b}$ нормально распределены и независимы;

5) остаточная сумма квадратов

$$Q^2 = \sum_{i=1}^{n} \left[y_i - \hat{a} - \hat{b}(x_i - \bar{x})\right]^2 \tag{6.9}$$

независима от $(\hat{a}, \hat{b})$, а Q^2 / σ^2 распределена по закону хи-квадрат χ^2_{n-2} с n-2 степенями свободы.

Оценка для σ^2 и доверительные интервалы. Свойство 5) дает возможность несмещенно оценивать неизвестный параметр σ^2 величиной

$$s^2 = Q^2 / (n\text{-}2). \tag{6.10}$$

Поскольку s^2 независима от $\hat{a}$ и $\hat{b}$, отношения

$$\sqrt{n}\,\frac{\hat{a} - a}{s} \quad \text{и} \quad \frac{\hat{b} - b}{s_b}, \quad \text{где } s_b = s \Big/ \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

имеют распределение Стьюдента с (n-2) степенями свободы, и потому доверительные интервалы для a и b таковы:

$$|\hat{a} - a| \le t_p \frac{s}{\sqrt{n}}, \quad |\hat{b} - b| \le t_p s_b, \tag{6.11}$$

где t_p - квантиль уровня $(1 + P_д) / 2$ распределения Стьюдента с n - 2 степенями свободы, $P_д$ - коэффициент доверия.

Проверка гипотезы о коэффициенте наклона. Обычно возникает вопрос: может быть, y не зависит от x, т.е. $b = 0$, и изменчивость y обусловлена

только случайными составляющими ε_i? Проверим гипотезу H: $b = 0$. Если 0 не входит в доверительный интервал (4.8) для b, т.е.

$$|b| / s_b > t_p, \tag{6.12}$$

то гипотезу H следует отклонить; уровень значимости при этом $\alpha = 1 - P_Д$.

Другой способ (в данном случае эквивалентный (4.9)) проверки гипотезы H состоит в вычислении статистики

$$F = \frac{\hat{b}^2 / D\hat{b}}{Q^2 / (\sigma^2 (n-2))} = \frac{\hat{b}^2}{s_b^2}, \tag{6.13}$$

распределенной, если H верна, по закону $F(1, n-2)$ Фишера с числом степеней свободы 1 и $n-2$. Если

$$F > F_{1-\alpha}, \tag{6.14}$$

где $F_{1-\alpha}$ - квантиль уровня $1-\alpha$ распределения $F(1, n-2)$, то гипотеза H отклоняется с уровнем значимости α

Вариация зависимой переменной и коэффициент детерминации. Рассмотрим вариацию (разброс) T_{ss} (*total sum of square*) значений y_i относительно среднего значения $\bar{y}$

$$T_{ss} = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{6.15}$$

Обозначим $\tilde{y}_i$ предсказанные с помощью функции регрессии значения y_i: $\tilde{y} = \hat{a} + \hat{b}x_i$. Сумма R_{ss} (*regression sum of square*)

$$R_{ss} = \sum_{i=1}^{n} (\tilde{y}_i - \bar{y})^2 \tag{6.16}$$

означает величину разброса, которая обусловлена регрессией (ненулевым значением наклона $\hat{b}$). Сумма E_{ss} (*error sum of squares*)

$$E_{ss} = \sum_{i=1}^{n} (y_i - \tilde{y}_i)^2 \tag{6.17}$$

означает разброс за счет случайных отклонений от функции регрессии. Оказывается,

$$T_{ss} = R_{ss} + E_{ss}, \tag{6.18}$$

т.е. полный разброс равен сумме разбросов за счет регрессии и за счет случайных отклонений. Величина R_{ss} / T_{ss} - это доля вариации значений y_i, обусловленной регрессией (т.е. доля закономерной изменчивости в общей изменчивости). Статистика

$$R^2 = R_{ss} / T_{ss} = 1 - E_{ss} / T_{ss}$$

называется ***коэффициентом детерминации***. Если $R^2 = 0$, это означает, что регрессия ничего не дает, т.е. знание x не улучшает предсказания для y по сравнению с тривиальным $\tilde{y}_i = \bar{y}$. Другой крайний случай $R^2 = 1$ означает точную подгонку: все точки наблюдений лежат на регрессионной прямой. Чем ближе к 1 значение R^2, тем лучше качество подгонки.

Пример. Исследуется зависимость урожайности y зерновых культур

(ц/га) от ряда факторов (переменных) сельскохозяйственного производства, а именно,

x_1 - число тракторов на 100 га;

x_2 - число зерноуборочных комбайнов на 100 га;

x_3 - число орудий поверхностной обработки почвы на 100 га;

x_4 - количество удобрений, расходуемых на гектар (т/га);

x_5 - количество химических средств защиты растений, расходуемых на гектар (ц/га).

Исходные данные для 20 районов области приведены в табл. 5.1.

Таблица 5.1. Исходные данные

	y	x_1	x_2	x_3	x_4	x_5
1	9.7	1.59	.26	2.05	.32	.14
2	8.4	.34	.28	.46	.59	.66
3	9.0	2.53	.31	2.46	.30	.31
4	9.9	4.63	.40	6.44	.43	.59
5	9.6	2.16	.26	2.16	.39	.16
6	8.6	2.16	.30	2.69	.32	.17
7	12.5	.68	.29	.73	.42	.23
8	7.6	.35	.26	.42	.21	.08
9	6.9	.52	.24	.49	.20	.08
10	13.5	3.42	.31	3.02	1.37	.73
11	9.7	1.78	.30	3.19	.73	.17
12	10.7	2.40	.32	3.30	.25	.14
13	12.1	9.36	.40	11.51	.39	.38
14	9.7	1.72	.28	2.26	.82	.17
15	7.0	.59	.29	.60	.13	.35
16	7.2	.28	.26	.30	.09	.15
17	8.2	1.64	.29	1.44	.20	.08
18	8.4	.09	.22	.05	.43	.20
19	13.1	.08	.25	.03	.73	.20
20	8.7	1.36	.26	.17	.99	.42

Здесь мы располагаем выборкой объема $n = 20$; число независимых переменных (факторов) $k = 5$. Матрица X должна содержать 6 столбцов размерности 20; первый столбец состоит из единиц, а столбцы со 2-го по 6-й представлены соответственно столбцами 3÷7 таблицы (файл *Harvest 2. sta.*). Специальный анализ (здесь не приводимый) технологии сбора исходных данных показал, что допущения (12а) могут быть приняты в качестве рабочей гипотезы, поэтому можем записать уравнения статистической связи между y_i и $X_i = (x_{i1}, x_{i2}, ..., x_{i5})$, $i = 1, ..., n$).

Выполнение в пакете STATISTICA

Работаем в модуле *Multiple Regression* (множественная регрессия).

Ввод данных. Образуем таблицу 6v × 20c с 6 столбцами (*variables* - переменными) и 20 строками (*cases*). Столбцы назовем y, x_1, x_2, ..., x_5. Введем в таблицу исходные данные.

Предварительный просмотр. Предварительно визуально оценим имеющиеся данные, построив несколько диаграмм рассеяния:

Graphs - Stats 2D Graphs - Scatterplots - Variables - X: x1, *Y: y, Graph Type: Regular, Fit* (подбор): *Linear - OK.*

Наблюдаем диаграмму рассеяния с подобранной прямой парной регрессии, параметры которой отражены в заголовке. Повторим это еще 4 раза, заменяя x_1 на другие факторы: x_2, ..., x_5. Иногда такой просмотр позволяет увидеть основную зависимость. В нашем примере этого нет.

Выполнение регрессионного анализа:

Analysis - Startup Panel - кнопка *Variables:* - отбираем зависимую переменную *Dependent var: y* и независимые переменные *Independent var:* $x_1 \div x_5$ (при нажатой клавише *Ctrl*) - *OK - Input file* (входной файл): *Raw Data* (необработанные файлы) - *OK* - в окне *Model Definition* (уточнения) *Metod: Standart, Intercept: Include in model* (постоянную составляющую включить в модель) - *OK.*

В окне *Mult. Regr. Results* имеем основные результаты: коэффициент детерминации $R^2 = 0.517$; для проверки гипотезы H_0 об отсутствии какой бы то ни было линейной связи между переменной y и совокупностью факторов определена статистика $F = 3.00$; это значение соответствует уровню значимости $p = 0.048$ (эквивалент согласно распределению F (5,14) Фишера с $df = 5$ и 14 степенями свободы. Поскольку значение p весьма мало, гипотеза H_0 отклоняется.

Кнопка *Regression summary* - имеем таблицу результатов:

Таблица 5.2. Таблица Результатов

Regression Summary for Dependent Variable: Y				
R = .71923865 *RI* = .51730424 *Adjusted RI* = .34491290 F(5,14) = 3.0008 *p*<.04787 *Std. Error of estimate*: 1.5990				
	B	*St. Err of B*	*t*(14)	*p-level*
Intercpt	3.51460	5.41853	.648625	.527078
X1	-.00613	.93167	-.006580	.994843
X2	15.54246	21.50311	.722800	.481704
X3	.10990	.83254	.132004	.896859
X4	4.47458	1.54345	2.899065	.011664
X5	-2.93251	3.08833	-.949546	.358448

В ее заголовке повторены результаты предыдущего окна; в столбце B указаны оценки неизвестных коэффициентов $\hat{\beta}_j$. Таким образом, оценка $\hat{f}(x)$ неизвестной функции регрессии $f(x)$ в данном случае:

$$\hat{f}(x) = 3.51 - 0.06\,x_1 + 15.5\,x_2 + 0.11\,x_3 + 4.47\,x_4 - 2.93\,x_5 \qquad (6.19)$$

В столбце *St. Err. of B* указаны стандартные ошибки s_j оценок коэффициентов; видно, что стандартные ошибки в оценке всех коэффициентов, кроме β_4 , превышают значения самих коэффициентов, что говорит о статистической ненадежности последних. В столбце *t* -значение статистики Стьюдента для проверки гипотезы о нулевом значении соответствующих коэффициентов; в столбце *p-level* -уровень значимости отклонения этой гипотезы; достаточно малым (0.01) этот уровень является только для коэффициента при x_4 . Только переменная x_4 - количество удобрений, подтвердила свое право на включение в модель. В то же время проверка гипотезы об отсутствии какой бы то ни было линейной связи между y и (x_1 , ..., x_5) с помощью статистики $F = 3.00$, $p = 0.048$, говорит о том, что следует продолжить изучение линейной связи между y и (x_1 , ..., x_5), анализируя как их содержательный смысл, так и матрицу парных корреляций, которая определяется так: возврат в окно *Multi. Regr. Results* - кнопка *Correlations and desc. Stats - Correlations.* Из матрицы видно, что x_1 , x_2 и x_3 (оснащенность техникой)

Correlations (harvest2.sta)

	X1	X2	X3	X4	X5	Y
X1	1.000	.854	.978	.110	.341	.430
X2	.854	1.000	.882	.027	.460	.374
X3	.978	.882	1.000	.030	.278	.403
X4	.110	.027	.030	1.000	.571	.577
X5	.341	.460	.278	.571	1.000	.332
Y	.430	.374	.403	.577	.332	1.000

сильно коррелированы (парные коэффициенты корреляции 0.854, 0.882 и 0.978), т.е. имеет место дублирование информации, и потому, по-видимому, есть возможность перехода от исходного числа признаков (переменных) к меньшему.

Ввод данных. Образуем таблицу 4*v* × 20*c*, назовем ее, например, *Domna.sta.* В первые 2 столбца поместим исходные данные x и y. В третьем столбце поместим значения нового фактора $x2$ квадратов температур, *long name:* = x^2, в четвертом - $x3$ третьих степеней температур x, *long name:* = x^3. Сначала оценим имеющиеся данные визуально, с помощью процедуры *Scatterplot* (диаграмма рассеяния). Видим, что зависимость, возможно, нелинейная. Построим несколько регрессий.

1) Регрессия первой степени: $y = \beta_0 - \beta_1 x$ (*indep. Var.*: x); получим (в скобках указаны стандартные ошибки оценок):

$$y = \underset{(0.98)}{5.36} + \underset{(0.16)}{1.40}\, x$$

$$R^2_{adj} = 0.798,\ s = 2.09.$$

2) Регрессия второй степени: $y = \beta_0 - \beta_1 x - \beta_2 x^2$ (*indep. Var.*: x, $x2$); получим:

$$y = \underset{(1.33)}{9.9} - \underset{(0.57)}{0.88}\,x + \underset{(0.05)}{0.21}\,x^2, \qquad (6.20)$$

$$R^2_{adj} = 0.892,\ s = 1.53,$$

коэффициент $\beta_1 = -0.88$ незначимо отличается от 0. Эта регрессия лучше предыдущей в смысле R^2_{adj} и s. Однако, возможно, регрессия третьей степени окажется лучше?

3) Построим регрессию третьей степени: $y = \beta_o + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ (*indep. Var.*: x, $x2$, $x3$); получим:

$$y = \underset{(2.33)}{11.6} - \underset{(1.74)}{2.35}\,x + \underset{(0.36)}{0.53}\,x^2 - \underset{(0.02)}{0.02}\,x^3$$

$$R^2_{adj} = 0.890,\ s = 1.53,$$

$\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ незначимо отличаются от 0. Поскольку степень увеличилась без увеличения R^2_{adj}, от регрессии третьей степени отказываемся в пользу (22) второй степени. Однако, гипотеза о нулевом значении β_1 в (22) не отклоняется (*p-level* = 0.1), и потому построим

4) регрессию $y = \beta_o - \beta_2 x^2$ без линейного члена (*indep. Var.*: $x2$); получим

$$y = \underset{(0.54)}{8.02} + \underset{(0.01)}{0.13}\,x^2 \qquad (6.21)$$

$$R^2_{adj} = 0.884,\ s = 1.6,$$

Сравнивая ее по R^2_{adj} и s, отдаем предпочтение (22), поскольку ошибка прогноза s меньше.

4. Нелинейная зависимость (обобщение)

Предполагается, что связь между факторами $(x_1, \ldots, x_p)$ и y выражается следующим образом:

$$y = \beta_o - \beta_1 \varphi_1(x_1, \ldots, x_p) - \beta_2 \varphi_2(x_1, \ldots, x_p) - \ldots - \beta_k \varphi_k(x_1, \ldots, x_p) + \varepsilon$$

где $\varphi_j(\,)$, $j = 1, \ldots, k$, - система некоторых функций. Имеется n наблюдений при различных значениях $x \equiv (x_1, \ldots, x_p)$: $x^1, x^2, \ldots, x^n$; имеем:

$$y_i = \beta_o - \sum_{j=1}^{k} \beta_j \varphi_j(x^i) + \varepsilon_i,\ i = 1, \ldots, n,$$

или в матричной форме:

$$y = X\beta + \varepsilon,$$

где X - матрица $n \times (k+1)$, в i-й строке которой $(1, \varphi_1(x^i), \varphi_2(x^i), \ldots, \varphi_k(x^i))$; y, β, ε).

Вопросы:

1. Основные задачи регрессионного анализа.
2. Построение линейных моделей методом наименьших квадратов.
3. Преимущества оценок метода наименьших квадратов

Практическое занятие №7.
Методы построения линейных многопараметрических моделей и оценка их точности

В этом пункте рассмотрим работу с программой "Stepwise Variable Selection" (Пошаговая регрессия) раздела K. "Regression Analysis" (Регрессионный анализ) главного меню. Для выполнения пошагового регрессионного анализа необходимо выполнить следующие действия:

а) Выбрать в главном меню "Regression Analysis" и нажать клавишу "Enter".

б) Выбрать программу "Stepwise Variable Selection" и нажать клавишу "Enter", после чего на экране появится меню программы "Stepwise Variable Selection".

в) Заполнить пустые "окна":

- Dep.vars. -задать имя зависимой переменной;
- Ind.vars. - задать имена независимых переменных;
- Weights - по желанию можно задать имя переменной,содержащей веса наблюдений;
- Constant - выбрать одну из двух моделей множественной линейной регресии: со свободным членом или без него;
- Vertical bars - выбрать один из двух режимов вывода на экран графика предсказанных значений: с отмеченными вертикальными отрезками остатков или без них;
- Conf.level - выбрать коэффициент доверия для построения доверительных интервалов коэффициентов регрессии.

Method - выбрать один из трех методов:

1. Forward - самый распространенный вариант пошагового регрессионного анализа, когда процедура стартует с модели регрессионного анализа не включающей ни одной независимой переменной. На каждом шагу процедуры пошагового регрессионного анализа происходит включение в модель новых независимых переменных на основании статистики F-включения или исключение из модели уже включенных в модель независимых переменных на основании статистики F-исключения.

2. Backward - отличается от предыдущего метода тем, что процедура пошагового регрессионного анализа стартует с модели регрессионного анализа, в которую включены все независимые переменные.

None -стандартная процедура множественного регрессионного анализа.

F-enter - задать граничное значение статистики F-включения (по умолчанию задается значение 4.0);

3. F-remove - задать граничное значение статистики F-исключения (по умолчанию задается значение 4.0). Это значение не должно превосходить граничное значение статистики F-исключения. Max.steps - задать максимально возможное число шагов процедуры пошагового регрессионного анализа.

Control – выбрать один из вариантов представления информации:

"Automatic" - выводится информация только о конечной модели (т.е. результаты последнего шага процедуры пошагового регрессионного анализа), "Manuel" - выводится информация после каждого шага процедуры пошагового регрессионного анализа. В последнем случае переход к новому шагу процедуры пошагового регрессионного анализа осуществляется нажатием клавиши "E.

г) Нажать клавишу "F6". На экране появятся результаты нулевого шага, если выбран режим "Manuel", или последнего шага, если выбран режим "Automatic" пощагового регрессионного анализа.

д) Если выбран режим "Manuel", то нажать клавишу "Enter". На экране появятся результаты первого шага пошагового регрессионного анализа. Переход к последующим шагам пошагового регрессионного анализа осуществляется нажатием клавиши "Enter".

е) На экране дисплея представлены результаты пошагового регрессионного анализа: квадрат множественного коэффициента корреляции (R-squared), квадрат множественного коэффициента корреляции, исправленный на степени свободы (Adjusted), средний квадрат ошибки (MSE), степени свободы (d.f.), список независимых переменных уже включенных в модель (Variables in Model), коэффициенты регрессии (Coeff.), значения статистик F-исключения для независимых переменных уже включенных в модель (F-Remove), список независимых переменных еще не включенных в модель (Variables Not in Model), значения частных коэффициентов корреляции (P.Corr) и значения статистик F-включения (F-Enter) для независимых переменных еще не включенных в модель.

ж) Нажать клавишу "Enter". На экране появятся результаты последнего шага пошагового регрессионного анализа.

з) Нажать клавишу "F10".

и) В появившемся подменю можно выбрать один из режимов вывода результатов на экран (для этого подсветить имя нужной программы и нажать клавишу "Enter"):

- Analysis of variance - вывод на экран таблицы дисперсионного анализа; Conditional sums of sq. -вывод на экран таблицы условных сумм квадратов;
- Plot residuals - вывод на экран диаграммы рассеивания остатков от предсказанных значений, или любой переменной, заданной ее индексом;
- Summarize residuals - вычисление статистик для остатков (среднее, дисперсия, стандартная ошибка среднего, коэффициенты ассиметрии и эксцесса, статистика Дурбина-Ватсона);
- Plot predicted values - вывод на экран диаграммы рассеивания, где по оси абсцисс откладываются предсказанные значения, а по оси ординат - измеренные значения зависимой переменной;
- Probability plot - нормальный вероятностный график для остатков;
- Component effect plot -график эффектов компонент;

Influence measures - меры влияния;

- Correlation matrix - вычисление матрицы корреляций для оценок регрессионных коэффициентов;

- Generate reports вывод на экран любой из следующих одной или всех информаций: наблюдаемые значения (observed values), линиии регрессии (fitted values), остатки (residuals), стандартизованные остатки (standardized residuals), стандартные ошибки для предсказанных значений (standard errors for forecasts), доверительный интервал для единичногопредсказанного значения (confidence limits for individual forecasts), доверительный интервал для среднего предсказанного значения (confidence limits for individual forecasts means);

- Confidence intervals -вывод на экран доверительных интервалов для коэффициентов регрессии;

- Interval plots - вывод на экран доверительных интервалов по предсказанным значениям или по любой переменной, задаваемой своим индексом;

Save results – запись на диск результатов регрессионного анализа.

к) Нажать клавишу "F5". В появившемся подменю можно выбрать:

-Force variable into model-режим принудительного включения переменных в модель;

- Remove variabl.from model-режим принудительного исключения переменных из модели.

Для возвращения в главное меню используйте клавишу "F10".

Вопросы:

1. Метод наименьших квадратов (теория). Основные предположения МНК.
2. Проблемы, возникающие при несоблюдении предположений МНК.
3. Пути решения проблем мультиколлинеарности, автокорреляции и сезонных колебаний.

Практическое занятие №8.
Методы построения нелинейных математических моделей

В этом пункте рассмотрим работу с программой "Nonlinear Regression" (Нелинейная регрессия) раздела K. "Regression Analysis" (Регрессионный анализ) главного меню. Для выполнения нелинейного регрессионного анализа необходимо выполнить следующие действия:

а) Выбрать в главном меню "Regression Analysis" и нажать клавишу "Enter".

б) Выбрать программу "Nonlinear Regression" и нажать клавишу "Enter", после чего на экране появится меню программы "Nonlinear Regression".

в) Заполнить пустые "окна":

- Dep.vars. - задать имя зависимой переменной;
- Parameter vector - задать начальные значения для коэффициентов регрессии;
- Function - задать функцию регрессии на языке APL (коэффициенты регрессии обозначаются через PARM[1], PARM[2] и т.д.), или задать характеристический вектор содержащий функцию регрессии;
- Maximum iterations - задать максимальное число итераций процедуры нелинейного регрессионного анализа;
- Maximum function calls -задать максимальное число вызовов функций регрессии процедуры нелинейного регрессионного анализа (это число должно быть больше максимального числа итераций);
- Stopping cond.on res.ss – задать пороговое значение относительного изменения остаточной суммы квадратов в качестве критерия остановки процедуры нелинейного регрессионного анализа;
- Stopping cond.on estim. – задать пороговое значение относительного изменения оценок коэффициентов регрессии в качестве критерия остановки процедуры нелинейного регрессионного анализа;
- Initial Marquardt param. - задать начальное значение параметра Марквардта;

Initial scaling factor - задать начальное значение маштабного множителя;

Max.value of Marquardt - задать максимальное значение parm параметра Марквардта в качестве критерия остановки процедуры нелинейного регрессионного анализа.

г) Нажать клавишу "F6". На экране появятся результаты регрессионного анализа.

д) Нажать клавишу "F10".

е) В появившемся подменю можно выбрать один из режимов вывода результатов на экран (для этого подсветить имя нужной программы и нажать клавишу "Enter"):

- Analysis of variance - вывод на экран таблицы дисперсионного анализа;
- Plot fitted model - вывод на экран диаграммы рассеивания зависимой переменной от любой выбранной переменной и графика функции регрессии;

- Plot residuals - вывод на экран диаграммы рассеивания остатков от предсказанных значений, или любой переменной, заданной ее индексом;

- Summarize residuals - вычисление статистик для остатков (среднее, дисперсия, стандартная ошибка среднего, коэффициенты ассиметрии и эксцесса, статистика Дурбина-Ватсона);

- Plot predicted values - вывод на экран диаграммы рассеивания, где по оси абсцисс откладываются предсказанные значения, а по оси ординат - измеренные значения зависимой переменной;

- Save residuals - запись на диск остатков;

- Save parameter estimates - запись на диск оценок параметров;

- Save covariance - запись на диск ковариационной матрицы оценок параметров.

Для возвращения в главное меню используйте клавишу "F10".

Вопросы:

1. Основные задачи и виды математических моделей
2. Выдвижение гипотезы о структуре модели
3. Оценивание параметров модели
4. Проверка адекватности модели

Практическое занятие №9.
Методы оценки эффективности математических моделей

Выбор тактики лечения и управление патологическим процессом, как правило, подразумевает наличие четкого представления о месте и роли каждого фактора в исследуемом процессе. Эта задача решается при помощи методов дисперсионного анализа. Рассмотрим методику практического использования возможностей этих методов при помощи пакета статистических программ «Statistica» на примере однофакторного дисперсионного анализа [4,5]. Отметим, что возможности пакета статистических программ «Statistica» позволяют также проводить двух- , трех- факторные анализы.

Основные соотношения. Изучается влияние, которое оказывает некоторый качественный признак (фактор) на количественный результат (отклик), например, влияние технологии изготовления прибора на его долговечность, влияние способа обработки земли на урожайность и т.д. Пусть фактор имеет k уровней $A_1, A_2, ..., A_k$ и пусть измеряемая величина x есть результат действия фактора и случайной составляющей ε (от фактора не зависящей):

$$x = f(A) + \varepsilon$$

Будем считать, 1) что при каждом уровне A_j фактора, $j = 1, ..., k$, имеется n_j измерений

$$x_{ij} = a_j + \varepsilon_{ij}, \qquad i = 1, ..., n_j, \tag{9.1}$$

где обозначено $a_j = f(A_j)$, 2) что случайная составляющая ε нормально распределена $N(0, \sigma^2)$ с дисперсией σ^2. Если влияния фактора нет, то все a_j равны. Итак, имеется k выборок объемами $n_1, ..., n_k$, $\sum_{j=1}^{k} n_j = N$. Проверим гипотезу об отсутствии влияния:

$$H: \ a_1 = a_2 = ... = a_k$$

По каждой из выборок методом наибольшего правдоподобия оценим средние a_j и дисперсию σ^2:

$$\hat{a}_j = \bar{x}_j \equiv \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}, \qquad s_j^2 = \frac{1}{n_j} \sum (x_{ij} - \bar{x}_{\cdot j})^2, \tag{9.2}$$

а затем оценим σ^2 по всем выборкам:

$$\sigma^{2*} = \frac{1}{N-k} \sum_{j=1}^{k} n_j s_j^2 = \frac{1}{N-k} \sum_{j=1}^{k} \sum_{j=1}^{n_j} (x_{ij} - \bar{x}_{\cdot j})^2. \tag{9.3}$$

Эта статистика несмещенно оценивает σ^2 независимо от того, верна или нет гипотеза H.

Другую оценку для σ^2 построим по значениям $\hat{a}_1, \hat{a}_2, ..., \hat{a}_k$. Если H верна, то $M\hat{a}_j = a$, $D\hat{a}_j = \sigma^2 / n_j$, $j = 1, ..., k$. Оценки для a и σ^2:

$$\hat{a} = \frac{1}{N}\sum_{j=1}^{k} n_j \hat{a}_j, \qquad \sigma^{2**} = \frac{1}{k-1}\sum_{j=1}^{k} n_j (\hat{a}_j - \hat{a})^2 \tag{9.4}$$

Из теоремы о совместном распределении оценок среднего и дисперсии нормальной совокупности следует, что статистики $(N-k)\sigma^{2*}$ и $(k-1)\sigma^{2**}$ независимы и распределены как $\sigma^2\chi^2_{N-k}$ и $\sigma^2\chi^2_{k-1}$ соответственно, и потому их отношение

$$F_{II} = \frac{\sigma^{2**}}{\sigma^{2*}} = \frac{\sigma^2\chi^2_{k-1}/(k-1)}{\sigma^2\chi^2_{N-k}/(N-k)}, \tag{9.5}$$

если гипотеза H верна, имеет F-распределение Фишера.

Если гипотеза не верна, то σ^{2**} имеет тенденцию к увеличению за счет разброса средних a_j, и потому, если F_{II} имеет слишком большое значение, т.е. если

$$F_{II} > Q, \tag{9.6}$$

то гипотеза H об отсутствии влияния фактора A отклоняется, и следует считать, что среди средних $a_1, a_2, \ldots, a_k$ имеются хотя бы два не равных; здесь $Q = Q(1-\alpha; k-1, N-k)$ - квантиль уровня $1-\alpha$ F-распределения с $k-1$ и $N-k$ степенями свободы, α - выбираемый уровень значимости. Если же (6) не выполняется, то это означает, что наблюдения не противоречат гипотезе об отсутствии влияния фактора. Условие (6) может быть записано иначе:

$$P\{F \geq F_{II}\} < \alpha, \tag{9.7}$$

где F - случайная величина, распределенная по закону Фишера.

Оценка влияния фактора. Отношение $\frac{\xi_i - a_i}{\sigma^*}\sqrt{n_i}$ подчиняется распределению Стьюдента с $N-k$ степенями свободы, и если $Q = Q(1-\alpha, N-k)$ - квантиль уровня $1-\alpha$ этого распределения, то доверительный интервал для a_j с уровнем доверия $1-2\alpha$:

$$\hat{a}_i \pm \frac{\sigma^*}{\sqrt{n_i}} Q \tag{9.8}$$

Если гипотеза H о равенстве средних отклоняется, то следует определить, по каким именно уровням фактора средние значимо различаются. Линейная комбинация

$$L = \sum_{j=1}^{k} c_j a_j$$

называется линейным контрастом. Оценка для L :

$$\tilde{L} = \sum_{j=1}^{k} c_j \xi_i,$$

а оценка дисперсии $D\tilde{L}$:

$$S_L^2 = \sigma^2 \cdot \sum_{j=1}^{k} \frac{c_j^2}{n_j}$$

Зафиксируем произвольное число r контрастов $L^{(1)}, L^{(2)}, \ldots, L^{(r)}$. Можно показать, что одновременно для всех $\tilde{L}^{(1)}, \ldots, \tilde{L}^{(r)}$ выполняются соотношения:

$$\left| L^{(m)} - \tilde{L}^{(m)} \right| < S_{L^{(m)}} \sqrt{kF(1-\alpha, k, N-k)} \qquad (9.9)$$
$$m=1, \ldots, r$$

с вероятностью 1-α. Это соотношение позволяет сделать вывод о всех интересующих нас контрастах одновременно. В частности, среди разностей $a_j - a_i$ можно выделить те, которые значительно отличаются от нуля на выбранном уровне значимости (метод Шеффе).

Выполнение в пакете STATISTICA

Будем выполнять в модуле *Basic Statistics and Tables* (можно выполнять также в модуле *ANOVA/MANOVA*). Создадим таблицу с двумя столбцами *P* и *T* и 30 строками; в *P* занесем данные по производительности, в *T* - уровни *T*: технологии *T0*, *T1*, *T2*. Далее выполним:

One - Way ANOVA (Analys Of Variances) - Analysis: Detailed Analysis Of Individual tables, Variabbles: Grouping variabbles (группирующие переменные): *T*, *Dependent variabbles* (зависимые переменные - отклики): *P - OK - OK* - отметив *Statistics: Number of observations* (количество наблюдений), *Standart deviations* (стандартные отклонения) и *Variances* (дисперсии), получим *Summary table of means* (таблицу средних); видно, как отличаются средние в каждой из групп (при фиксированном уровне фактора *T*) - Возвращаемся в окно *Descriptive Stats and ... Results* и выполняем *Analysis of Variance* - Наблюдаем таблицу: в столбце *SS (Sum of Squares) Effect* указана сумма квадратов (4), умноженная на $(k - 1)$, $df = 2 = k - 1$ - число степеней свободы, *MS (Mean Square)* = 839.0 - оценка (4), *SS* = 2711 - сумма квадратов (3), умноженная на $(N - k)$, $df = 27 = N - k$, *Ms Error* = 100.4 - оценка (3), $F = 8.35$ - значение статистики (5), $p = 0.0015$ - вероятность в (7); последняя слишком мала, чтобы поверить в истинность гипотезы *H* об отсутствии влияния фактора *T*. Вывод: фактор *T* (технология) влияет на *P* (производительность).

Возникает вопрос: какой метод лечения можно считать значимо различным? Для ответа на этот вопрос возвращаемся в окно *Descriptive Stats and ... Results* и выполняем *Post - hoc comparasion of means* (сравнение средних) по методу Шеффе *Sheffe test.* Наблюдаем таблицу, в которой указаны уровни значимости гипотез о равенстве средних для всех пар уровней фактора *T*; видим, что методы лечения *T0* и *T1* следует считать различными (вероятность 0.0015 слишком мала, чтобы поверить в равенство средних по *T0* и *T1*).

Создадим таблицу с тремя столбцами (T_0 – традиционный метод лечения , T_1 - новая методика №1, T_2 - новая методика №2).

Анализ выполняем в модуле *ANOVA/MANOVA:*

Vaariables - Independent Vaariables (factors): A, B Dependent Vaariable list: X - OK - OK -Specific effects (спецификация влияний): выделяем (при двухфакторном анализе) факторы *A* и *B* - *All effects* - Наблюдаем таблицу *Summery* of *All effects* (итоги по всем влияниям); в столбце *MS Effects* (средние квадраты) оценки σ_{I0}, σ_{I1}, σ_{12}. Указываются значения статистик Фишера *F* (дисперсионные отношения) и уровни значимости *p*.

Вопросы:

1. Оценка точности математических моделей.
2. Оценка уравнения регрессии.
3. Ковариационный анализ.
4. Проблема автокорреляции.

Практическое занятие №10.
Построение кластеров

В этом пункте рассмотрим работу с программой "Cluster analysis" (Кластерный анализ) раздела Q."Multivariate methods"(Многомерные методы) главного меню. Программа предназначена для группировки наблюдений по заданному набору переменных. Для выполнения программы необходимо произвести следующие действия:

а) Выбрать в главном меню "Multivariate methods" и нажать клавишу "Enter".

б) Выбрать программу "Cluster analysis" и нажать клавишу "Enter", после чего на экране появится меню программы.

в) Заполнить пустые "окна":

- Data - указывается имя файла с переменными, матрицей корреляций или матрицей расстояний (могут быть также перечислены отдельные переменные);

- Label - указывается переменная, содержащая метки переменных;

- Method - выбирается неиерархический метод объединения (Seeded) или один из пяти иерархических: средней связи (Average), центроидный (Centroid), дальнего соседа (Furthest), медианной связи (Median), ближнего соседа (Nearest);

- Distance - выбирается один из двух типов входных данных: матрица корреляций или матрица расстояний (Matrix) или матрица данных, по которой вычисляются евклидовы расстояния между наблюдениями (Euclidean);

Standardize - выбирается либо режим стандартизации переменных (Yes), либо ее отсутствия (No);

- X-axis - указывается переменная, играющая роль абсциссы;

- Y-axis - указывается переменная, играющая роль ординаты;

- Z-axis - указывается переменная, играющая роль аппликаты;

- Circle - выбирается режим оконтуривания кластеров (Yes) или его отсутствие (No);

- Codes - указываются коды кластеров.

В случае выбора неиерархического метода (Seeded) дополнительно указываются номера наблюдений, предлагаемые в качестве точек притяжения для будущих кластеров (seed -зерно).

г) Нажать клавишу "F6", после чего на экране появятся результаты счета:

-таблица кластерного анализа.

д) Нажать клавишу "F5",если необходимо выдать результаты на принтер или запомнить их в файле.

е) Нажать клавишу "Esc", после чего появляется меню, предоставляющее следующие возможности:

- Plot clusters - графическое представление кластеров;

- Save cluster numbers - сохранение переменной с номерами кластеров для каждого наблюдения;

- Save distance matrix - сохранение матрицы евклидовых расстояний между наблюдениями.

Для возвращения в главное меню используйте клавишу F10.

Вопросы:

1. Цель и задачи кластерного анализа. Основная идеология кластерного анализа – классификация данных.
2. Метод полных связей.
3. Метод максимального локального расстояния.
4. Метод Варда.

СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ

1. Стратегия действий по пяти приоритетным направлениям развития Республики Узбекистан в 2017—2021 годах. 7.02.2017 г. УП 4947

2. . Innovation in Medicine and Healthcare 2016 - Chen Y.-W., Tanaka S., Howlett R.J., Jain, L.C. – 2016. - 315 c.

3. Smart Technologies in Healthcare - Bouchard Bruno – 2017. - 236 c.

High-performance computing in biomedicine - Bogoslovskiy N.N., Borisov A.V. – 2016.

4. Foundations of Biomedical Knowledge Representation: Methods and Applications - Hommersom A., Lucas P.J.F. – 2015.

5. Цифровая обработка биомедицинских сигналов и изображений - Фролов А.В. и др. – 2016.

6. Основы теории обработки биомедицинских сигналов - Божокин С.В. Суслова И.Б. – 2016.

7.Айвазян С.Ф., Мхитарян В.С. Прикладная статистика и основы эконометрики. Учебник для вузов. М.: ЮНИТИ, 1998. 1022 с.

8. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Методы и модели анализа данных: OLAP и data Mining. – СПб: БХВ – Петербург, 2004. – С.16 – 23.

9. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. М.: Финансы и статистика, 2000. – 352 с.

10. Леденева Т. М. Модели и методы принятия решений: учебное пособие / Т. М. Леденева. — Воронеж: Воронеж, гос. техн. ун–т, 2004. — 189 с.

11. Новицкий П.В., Зограф И.А. Оценка погрешностей результатов измерений. М.: Энергоатомиздат, 1991. – 304с.

12. Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. Робастность в статистике. Подход на основе функций влияния: Пер. с англ. / Хампель Ф., Рончетти Э., Рауссеу П., Штаэль В. – М.: Мир, 1989. – 512с.

СОДЕРЖАНИЕ

Методические указания по предмету «Биоинформатика и биомеханика» для бакалавров по направлению 5330500 – «Компьютерный инжиниринг»

Рассмотрено на заседании кафедры «Компьютерные системы»
от « » 201 г. Протокол №_____

Рассмотрено на заседании факультета «Компьютерный инжиниринг»
от « » 201 г. Протокол №_____

Рассмотрено и рекомендовано к изданию на заседании научно–методического Совета ТУИТ
от «___»___201 г. Протокол № _____

Составители Назаров А.И.
Мирзахалилов С.С.
Сайфуллаева Н.А.
Довлетова С.Б.

Рецензенты: М.Якубов
С.Ташев

Ответственный редактор: Мирзахалилов С.С.

Корректор: Доспанова Д.У.

Формат 60x84 1/16. Печ.лист 2,75.
Заказ №10. Тираж 30.
Отпечатано в «Редакционно издательском» отделе при ТУИТ.
Ташкент ул. Амир Темур, 108.