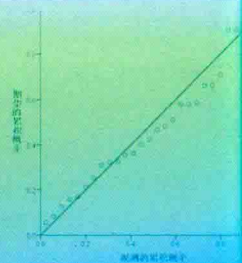


主编  
蔡晶  
魏高文

# 医学统计学 实战进阶



多元线性

Logistic

两分类

多分类

聚类分析

主成分分析

生存分析

生存分析

Cox回归

Meta分析

数据构建与

二分类资料

连续性资料

# 医学统计学实战进阶

主 编 蔡 晶 魏高文

副主编 李国春 武 松 吴建军

编 委 (按姓氏笔画排序)

|     |     |     |     |     |
|-----|-----|-----|-----|-----|
| 王丽梅 | 孔丽娅 | 朱 旭 | 刘志臻 | 许 茜 |
| 孙春阳 | 李国春 | 杨雪梅 | 步怀恩 | 吴建军 |
| 陈 书 | 陈婷婷 | 武 松 | 林有志 | 俞向梅 |
| 徐 刚 | 黄品贤 | 葛 亮 | 韩 梅 | 谢海林 |
| 赖新梅 | 蔡 晶 | 魏高文 | 魏歆然 |     |

学术秘书 葛 亮 许 茜



人民卫生出版社

图书在版编目(CIP)数据

医学统计学实战进阶/蔡晶,魏高文主编. —北京:  
人民卫生出版社,2018

ISBN 978-7-117-27555-2

I. ①医… II. ①蔡…②魏… III. ①医学统计-统  
计学 IV. ①R195.1

中国版本图书馆 CIP 数据核字(2018)第 223581 号

|      |                 |                                |
|------|-----------------|--------------------------------|
| 人卫智网 | www. ipmph. com | 医学教育、学术、考试、健康,<br>购书智慧智能综合服务平台 |
| 人卫官网 | www. pmph. com  | 人卫官方资讯发布平台                     |

版权所有,侵权必究!

医学统计学实战进阶

主 编:蔡 晶 魏高文

出版发行:人民卫生出版社(中继线 010-59780011)

地 址:北京市朝阳区潘家园南里 19 号

邮 编:100021

E - mail: pmph @ pmph. com

购书热线:010-59787592 010-59787584 010-65264830

印 刷:三河市尚艺印装有限公司

经 销:新华书店

开 本:787×1092 1/16 印张:9

字 数:213 千字

版 次:2018 年 11 月第 1 版 2018 年 11 月第 1 版第 1 次印刷

标准书号:ISBN 978-7-117-27555-2

定 价:30.00 元

打击盗版举报电话:010-59787491 E - mail: WQ @ pmph. com

(凡属印装质量问题请与本社市场营销中心联系退换)

# 前 言

《医学统计学实战指导》一书出版后,由于通俗易懂、实战性强,深受学生和教师欢迎,成为畅销书而多次印刷。许多读者向我们建议,编写一本高级统计方法的应用指导,以提升统计分析的“战斗力”。因此,原书的编写人员再接再厉,并在人民卫生出版社的大力支持下,《医学统计学实战进阶》应运而生。

《医学统计学实战进阶》既是《医学统计学实战指导》的姊妹篇,更是在实战指导基础上的提高和进阶,是掌握基础统计后对高级统计方法的学习和运用。本书内容不仅涵盖了高级统计方法中常用的多元线性回归、Logistic 回归、聚类分析、判别分析、主成分分析和因子分析等方法,还涵盖了生存分析和 Cox 回归、Meta 分析等医学研究中常用的高级统计方法。本书继承了《医学统计学实战指导》被广泛好评的优点:通过对范例数据的软件操作和结果解释的详细介绍,让读者通过模仿迅速上手并能在实践中加以使用,从而能够在短时间内提升读者的统计分析能力。本书的编写主要是针对需要进行复杂医学数据处理的医学生和科研工作者,是在掌握一定统计理论和统计方法基础上的进一步提高。

本书的雏形在一年前已基本完成,各位编委做了大量的工作并且付出了非常多的努力最终使本书能够顺利和读者见面。在此对大家的工作表示衷心的感谢!

由于编者学识有限,难免存在不足和纰漏,恳请广大读者批评指正。

蔡晶 魏高文

2018 年 6 月

# 第一章 目录 方法选择

|                           |     |
|---------------------------|-----|
| 第一章 高级统计方法选择 .....        | 1   |
| 第一节 数据分析前的准备工作 .....      | 1   |
| 第二节 多自变量的统计分析方法选择思路 ..... | 2   |
| 第三节 多因变量的统计分析方法选择思路 ..... | 4   |
| 第四节 SPSS 简介 .....         | 5   |
| 第二章 多元线性回归 .....          | 7   |
| 第三章 Logistic 回归 .....     | 23  |
| 第一节 两分类 Logistic 回归 ..... | 23  |
| 第二节 多分类 Logistic 回归 ..... | 32  |
| 第四章 聚类分析和判别分析 .....       | 40  |
| 第一节 聚类分析 .....            | 40  |
| 第二节 判别分析 .....            | 63  |
| 第五章 主成分分析和因子分析 .....      | 74  |
| 第一节 主成分分析 .....           | 74  |
| 第二节 因子分析 .....            | 82  |
| 第六章 生存分析和 Cox 回归 .....    | 99  |
| 第一节 生存分析的常用单因素非参数方法 ..... | 99  |
| 第二节 Cox 回归模型 .....        | 111 |
| 第七章 Meta 分析 .....         | 119 |
| 第一节 Revman 软件下载与安装 .....  | 119 |
| 第二节 数据构建与文献质量评价 .....     | 120 |
| 第三节 二分类资料 Meta 分析 .....   | 128 |
| 第四节 连续性资料 Meta 分析 .....   | 134 |

# 第一章 高级统计方法选择

实践是统计学产生的源泉,认识是统计学发展的动力。随着社会经济与科学技术的发展,医药科研也在与时俱进,传统的单因素、单效应指标的研究模式逐步被淘汰,多因素、多水平、多效应指标的研究模式正成为当代医药科研的主流模式,与此对应的数据统计分析也日益多元化、复杂化。因此,传统经典的单因素、一元统计分析方法的应用越来越受到限制,而多因素、多元统计分析方法也就应运而生。高级统计学(advanced statistics)就是在一元统计分析的基础上,以多维总体为研究对象,以多元统计分析的基本理论和方法为主要研究内容的学科,是处理多维数据的不可缺少的重要工具。特别是20世纪中叶以来,电子计算机技术的发展和普及促进了统计学的飞速发展,高级统计学的应用也越来越广泛,并且方兴未艾,日益显现出其超前的魅力!

## 第一节 数据分析前的准备工作

科研统计的基本流程是:研究设计→搜集资料→整理资料→分析资料→结果报告与结论表达等。统计工作的任一环节发生缺陷,都会影响研究结果的质量,甚至有可能导致错误的结论。如要对搜集到的科研数据进行正确、高效率的统计分析,做到胸有成竹,必须事先做好有关准备工作。

### 一、原始数据的审核

数据(data)又称资料,是由变量及其观测结果(变量值)所组成的。变量(variable)是指观察单位的某种特征或属性,即研究的项目或观察指标。根据变量的属性可分为数值变量(定量变量)、无序分类变量(定性变量或名义变量)和有序分类变量(顺序变量),分别构成计量资料(measurement data)、计数资料(enumeration data)和等级资料(ordinal data)。

1. 原始数据的检查与核对 统计学本身只是一种工具,原始资料的完整、准确和及时性,是正确做出统计分析的前提与基础。因此,在数据分析之前,首先应对原始数据进行审核,以确保原始数据准确无误。

完成搜集资料后,应首先对原始数据的结构、观测数、变量名称与数量、各变量的取值范围、最小值、最大值等进行检查与核对,以及时发现异常值(outlier)。如有错误或遗漏的研究变量取值,应及时采取补救措施,如修正、再次询问、查阅档案、重新检测等。

2. 确定拟分析的自变量和因变量 根据研究目的和资料特征,确定拟分析的自变量(independent variable)和因变量(dependent variable),有助于在数据分析阶段指导数据分析方法的选择。一般将试验因素或分组因素设为自变量,试验效应或观测结果设为因变量。

## 二、原始数据的整理

整理资料(sorting data)就是把搜集到的原始资料,有目的、有计划地进行科学的加工整理、分组和归纳汇总,使其系统化、条理化,以便更好地揭示所研究事物的规律性,进行计算分析。

1. 数据的分组 对于计数资料(无序分类变量)与等级资料(有序分类变量),可按照实际的类别进行分组,必要时可将性质相近的类别合并;对于计量资料(数值变量),则按照专业上的实际意义分组,如四分位数(quartiles)、五分位数(quintiles),也可按使用的方便程度或专业惯例分组,如年龄按每5或10岁一组,等等。

2. 数据的转换 计量资料统计推断的参数检验方法,一般要求数据满足正态性,而实际观测的数据往往不符合正态分布规律,对于正态性(normality)检验发现属于非正态分布的数据,应进行变量转换,如对数、平方根和倒数变换等,使之符合正态分布;分类变量可转换为哑变量进行数值化转换。

## 三、数据录入和管理

随着社会经济与科学技术的飞速发展,医学科研的数据也日益繁杂,大数据的分析也逐步走向前台,传统的手工计算方式已不能满足现代统计分析的需要,电子计算机已经成为统计分析的主要工具。因此,在对数据进行统计分析时,应选择合适的数据管理与分析软件进行数据录入和管理,以提高统计分析的效率,保证分析数据的正确性。

目前常用的数据管理软件主要有:Microsoft FoxPro、Microsoft Access、Microsoft Excel、EpiData等;常用的数据统计分析软件主要有:SAS、SPSS、Epi-Info、STATA等。

## 四、选择数据统计分析模型

医学科研数据的统计分析方法的选择,主要依据研究目的、设计类型、自变量与因变量的类型和数目、数据的分布特征、样本含量等。多元统计分析(multivariate statistical analysis)是从经典统计学中发展起来的综合分析方法,研究客观事物中多个因素与多个效应指标互相关联的统计规律,简称多元分析。本书将重点介绍常用的针对多自变量和(或)多因变量的多元统计分析方法的选择与运用。

### 第二节 多自变量的统计分析方法选择思路

在医药科研时,如果研究目的主要是探讨单因素多水平甚至于多因素多水平的试验效应的差别、试验因素与试验效应间的相互关系,这些情形属于多自变量统计分析的范畴,其数据的统计分析方法选择思路如下。

#### 一、数值型因变量

对于多自变量(或多组)的各个独立的定量观测指标进行分析时,可选择数值型因变量(计量资料)单变量分析方法,常用的方法及其应用条件见表1-1。



表 1-1 计量资料常用的单变量分析方法及其应用条件

| 单变量分析方法               | 分析目的                     | 应用条件                   |
|-----------------------|--------------------------|------------------------|
| 单因素方差分析               | 完全随机设计多组均数比较             | 多样本、正态分布、方差齐性          |
| Kruskal-Wallis $H$ 检验 | 完全随机设计多组平均水平比较           | 多样本、非正态分布或方差不齐         |
| 两因素方差分析               | 随机区组设计多组均数比较             | 多样本、正态分布、方差齐性          |
| Friedman $M$ 检验       | 随机区组设计多组平均水平比较           | 多样本、非正态分布或方差不齐         |
| 多样本协方差分析              | 多样本修正均数比较                | 多样本、正态分布、方差齐性、协变量      |
| 重复测量资料方差分析            | 观测指标多次测量其平均水平的<br>时间变化趋势 | 正态分布、方差齐性、协方差阵球对<br>称性 |
| 拉丁方设计方差分析             | 拉丁方设计多组均数比较              | 多样本、三因素(处理、配伍、序列)      |
| 析因设计方差分析              | 析因设计多组均数比较               | 多样本、处理因素、交互作用          |
| 正交设计方差分析              | 正交设计多组均数比较               | 多样本、处理因素、交互作用          |
| 交叉设计方差分析              | 交叉设计多组均数比较               | 多样本、三因素(处理、配伍、阶段)      |
| 两层次分组方差分析             | 组内分组设计多组均数比较             | 多样本、大组(主)、小组(次)        |

## 二、分类型因变量

对于多自变量(或多组)的各个独立的定性观测指标进行分析时,可选择分类型因变量(计数资料)分析方法,常用的方法及其应用条件见表 1-2。

表 1-2 计数资料常用的分析方法及其应用条件

| 分析方法                            | 分析目的                 | 应用条件                                                                                          |
|---------------------------------|----------------------|-----------------------------------------------------------------------------------------------|
| 成组设计 $R \times C$ 表 $\chi^2$ 检验 | 检验多个独立样本之间的<br>差异    | 不能有 $T < 1$ 的格子,并且表中 $1 \leq T < 5$ 的<br>格子数不能超过格子总数的 $1/5$ ,否则应<br>当用 Fisher 确切概率法直接计算 $P$ 值 |
| Logistic 回归模型与对数线<br>性模型        | 一个分类因变量与多个自<br>变量的关系 | 因变量为分类、自变量为分类或定量变量                                                                            |
| 对应分析                            | 降维(数据简化)             | 分类变量                                                                                          |
| 生存分析                            | 终点事件发生数量与时间          | 单因素分析用 Log-rank 检验,多因素分析<br>可用 Cox 回归                                                         |
| Kruskal-Wallis $H$ 检验           | 多样本各等级差异比较           | 多样本等级资料                                                                                       |
| 等级(秩)相关                         | 两等级变量线性相关方向<br>与程度   | 双变量、等级资料                                                                                      |
| 配对设计方表 $\chi^2$ 检验              | 检验关联性、优势性、一<br>致性    | 配对设计多个等级资料                                                                                    |
| Ridit 分析                        | 两样本或多样本各等级差<br>异比较   | 等级资料、大样本(各组 $n > 50$ )                                                                        |

### 三、多中心试验或多作者研究文献的综合评价

在全球范围内,对于同一研究目的或项目的多中心试验或多作者报告的研究文献,结论可能不完全一致,导致读者难以取舍。对此,可采用 Meta 分析进行综合评价,得出一个总的结论。如果结局变量(因变量)为数值型,当满足同质性时,采用固定效应模型,可选择通用的基于方差的方法( general variance-based 法)等统计分析方法;当不满足同质性时,采用随机效应模型,可选择 DerSimonian and Laird 法等统计分析方法。如果结局变量(因变量)为分类型,当满足同质性时,采用固定效应模型,可选择 Mentel-Haenszel 法、Peto 法、通用的基于方差的方法等统计分析方法;当不满足同质性时,采用随机效应模型,可选择 DerSimonian and Laird 法等统计分析方法。

### 第三节 多因变量的统计分析方法选择思路

现代的医药科研中,经常需要观察多个试验效应指标来分析研究因素的效应。因此,在资料的统计分析时,也就必须对多个试验效应指标进行综合评价。当因变量为分类型因变量(计数或等级资料)时,其数据的统计分析方法选择思路同第二节;当因变量为两个或两个以上有关联的定量观测指标时,可选择多变量分析方法,常用的方法及其应用条件见表 1-3。

表 1-3 计量资料常用的多变量分析方法及其应用条件

| 多变量分析方法                 | 分析目的                | 应用条件                        |
|-------------------------|---------------------|-----------------------------|
| 单样本 Hotelling $T^2$ 检验  | 单样本与总体均数向量比较        | 多元正态分布、多元协方差矩阵相等            |
| 配对样本 Hotelling $T^2$ 检验 | 配对样本均数向量比较          | 多元正态分布、多元协方差矩阵相等            |
| 聚类分析与判别分析               | 多变量无类分类与有类归类        | 多元正态分布、多元协方差矩阵相等            |
| 主成分分析与因子分析              | 多变量降维               | 多样本单因素、多因素、多元正态分布、多元协方差矩阵相等 |
| 两样本 Hotelling $T^2$ 检验  | 两样本单均数向量比较          | 多元正态分布、多元协方差矩阵相等            |
| Wilks $\lambda$ 检验      | 多样本均数向量比较           | 多样本单因素、多因素、多元正态分布、多元协方差矩阵相等 |
| 含协变量的多元方差分析             | 含协变量多变量比较           | 多样本、协变量、多元正态分布、多元协方差矩阵相等    |
| 重复测量的多元方差分析             | 重复测量设计的多变量比较        | 重复测量、多元正态分布、H-F 球对称         |
| 多元相关分析                  | 多个变量间的线性相关关系        | 多变量、多元正态分布                  |
| 多重线性回归分析                | 1 个因变量与多个自变量的数量依存关系 | 多变量、多元正态分布、多元相关             |
| 典型相关分析                  | 多个因变量与多个自变量的相关性     | 多变量、多元正态分布                  |

## 第四节 SPSS 简介

医学统计学是以数理统计理论和方法为基础的,涉及大量的计算公式,如采用传统的手工计算,过程繁杂,且容易出现差错,特别是多元统计分析方法更加复杂,手工计算几无可能。因此,在进行统计分析时,人们常常借助于各种统计软件来解决复杂的计算过程,使统计工作者利用多元统计分析方法来解决实际问题更加简单方便。

SPSS(Statistical Product and Service Solutions)是目前全球最为广泛应用的统计软件之一,也是世界上最早的统计分析软件。SPSS 最初全称为“社会科学统计软件包”(Statistical Package for the Social Sciences),由美国斯坦福大学的三位学生 Norman H. Nie, C. Hadlai (Tex) Hull 和 Dale H. Bent 于 1968 年研究开发,并成立了 SPSS 公司。随着 SPSS 产品服务领域的扩大和服务深度的增加,SPSS 公司于 2000 年正式将全称更改为“统计产品与服务解决方案(Statistical Product and Service Solutions)”,2009 年更名为“预测统计分析软件”(Predictive Analytics Software, PASW),2009 年 10 月 IBM 公司完成收购,2010 年更名为 IBM SPSS,2014 年 SPSS 已升级至 22.0 版本,本教材的各章例题的复杂计算过程,主要采用正版统计软件 SPSS22.0 来完成。

### 一、SPSS 22.0 软件系统要求

为保障 SPSS 软件的正常运行,需要电脑具备一定的软硬件条件,其最低软硬件要求如下。

1. 操作系统 Microsoft Windows XP(32 位版本)、Windows Vista(32 位和 64 位版本)或 Windows 7(32 位和 64 位版本)、Windows 8(32 位和 64 位版本)、Mac。

2. 硬件系统 以 2 千兆赫兹(GHz)或更高频率运行的 Intel 或 AMD 处理器,4GB RAM 或更大,800MB 可用硬盘空间。如果安装一种以上的帮助语言,多出的每种语言将需要 150~170MB 磁盘空间。如果使用物理安装介质,则需要 DVD/CD 驱动器,XGA(1024×768)或更高分辨率的显示器。如果要与 IBM SPSS Statistics 服务器连接,就需要运行 TCP/IP 网络协议的网络适配器。

### 二、主要功能介绍

SPSS 统计分析软件是一款按照功能模块进行配置的软件产品,主要包括 SPSS Statistics Base 模块和其他一系列扩充功能模块,每个独立扩充功能模块是在 SPSS Statistics Base 模块基础上,增加某方面的分析功能。

1. SPSS Statistics Base 为必需的基础模块,管理整个软件平台,管理数据访问、数据处理和输出,并能进行基本统计分析。包括描述统计、行列计算、汇总、计数、交叉分析、分类比较、描述性统计、因子分析、回归分析及聚类分析等。

2. SPSS Advanced Statistics 在分析数据时,除了基本的数据分析外,如果还想建立并分析过程数据,就需要使用高级统计模型(advanced statistic models),可以为等级资料建立更灵活、更成熟的模型,也可以在处理嵌套数据时得到更精确的预测模型,还可以分析事件历史和持续时间数据。具体功能包括:方差成分估计、混合模型、一般线性模型

(GLM)、广义线性模型(GZLMS)、广义估计方程(GEES)、对数线性模型、多因子系统模式的对数线性模型、生存分析(寿命表、Kaplan-Meier 估计、Cox 回归)。

3. SPSS Bootstrapping 为更有效地使用小样本量的数据,通过数据自身重采用的功能,可以模拟大样本情况下的采样结果,从而对数据结构特征和偏差有更直接的认识。当某些参数估计或者假设检验值得怀疑时,也可以使用该功能进行直接采用,从而以一种更加直观的方式来执行结果的检查。

4. SPSS Categories 为对应分析程序,用启发性的二维图和感知图表达数据的关系,从大量变量或二维及多维表格中了解重要讯息。提供非线性主成分分析来描述数据,通过类似传统的回归分析、主成分分析及典型相关分析,处理和了解顺序及名义数据,可视化地探索多变量分类数据。还可以应用齐性分析(homogeneity analysis)来分析分类多变量数据矩阵。

5. SPSS Complex Samples 将抽样设计融入调查分析之中,对复杂抽样数据的总体得到更加有效的统计推论。

6. SPSS Conjoint 包含 3 个相互关联过程,用于进行全特征联合分析。考虑研究时应包括的产品属性、属性水平、产品卡片的数量,用正交设计生成一个包含适量产品卡片的正交主效果片段因子设计。帮助市场研究人员和新产品开发部门了解在消费者心目中什么产品属性是重要的,了解消费者最偏爱的属性水平是什么,进行定价研究,进行品牌价格研究。

7. SPSS Data Preparation 在预处理数据时识别虚假和无效的观测值、变量和数据值,确认可疑的或者残缺的案例,查看数据缺失模式,描述变量分布以备分析,更准确地应用针对于分类变量的算法,迅速找到多元的极端值,执行数据检验,为建模预处理数据。

8. SPSS Decision Trees 基于数据挖掘中发展起来的树结构模型,对分类变量或连续变量进行预测,可以方便、快速地对样本进行细分。

9. SPSS Exact Tests 对于小样本或零星的数据,资料细分到多个类别,或数据变量中有超过 80% 的观测值集中在某一类别,传统方法要得到更好正确的结论,每一单元需要有 5 个以上数据,而 SPSS Exact Tests 则解除了这种限制。超过 30 个精确检验涵盖了小型或大型数据集所有的非参数和分类数据问题,包括独立或相关样本的单样本、两样本和 K-样本检验,拟合度检验,列联表独立性检验和联合测度检验等。

10. SPSS Forecasting 是目前功能最强的时间序列分析工具,能帮助研究人员做更好的预测。包括多重曲线拟合、平滑以及自回归方程估计,可自动从 ARIMA 和指数平滑模型中选择最佳拟合的时间序列和因变量的模型,避免反复选择模型的工作。

11. SPSS Missing Values 缺失数据会带来偏差或错误的分析结果,简单代入法或者简单的回归法都不能正确地填补缺失值。SPSS Missing Values 用 6 种灵活的诊断报告来评估缺失值是否会影响分析结论,通过快捷地诊断缺失值,得到更精确的摘要统计量,方便地用估计值替换缺失值,得到精确的结论。

## 第二章 多元线性回归

多元线性回归是统计分析方法中最常用的方法之一。如果所研究的现象有若干个影响因素,且这些因素对现象的综合影响是线性的,则可以使用多元线性回归的方法建立现象(因变量)与影响因素(自变量)之间的线性函数关系式。由于多元线性回归的计算量比较大,所以有必要应用统计分析软件进行统计分析。这一章将专门介绍 SPSS 软件的多元线性回归分析的操作方法,包括求回归系数,给出回归模型的各项检验统计量值及相应的概率,对输出结果的分析等相关内容。

**例 2-1** 有研究认为血清中高密度脂蛋白降低是引起动脉硬化的一个重要原因,现测量了 28 名被怀疑患有动脉硬化的就诊患者的载脂蛋白 A I、载脂蛋白 B、载脂蛋白 E、载脂蛋白 C 和高密度脂蛋白中的胆固醇含量,资料见表 2-1,试分析 4 种载脂蛋白对高密度脂蛋白中胆固醇含量的影响,并建立线性回归方程。

表 2-1 28 名患者载脂蛋白和高密度脂蛋白中胆固醇含量的测量结果 (mg/dl)

| No. | 载脂蛋白 A I | 载脂蛋白 B | 载脂蛋白 E | 载脂蛋白 C | 高密度脂蛋白 |
|-----|----------|--------|--------|--------|--------|
| 1   | 199      | 112    | 6.9    | 16.7   | 82     |
| 2   | 118      | 136    | 7.1    | 15.7   | 40     |
| 3   | 139      | 94     | 8.6    | 13.6   | 51     |
| 4   | 175      | 160    | 12.1   | 20.3   | 65     |
| 5   | 131      | 154    | 11.2   | 21.5   | 40     |
| 6   | 158      | 141    | 9.7    | 29.6   | 42     |
| 7   | 158      | 137    | 7.4    | 18.2   | 56     |
| 8   | 132      | 151    | 7.5    | 17.2   | 37     |
| 9   | 162      | 110    | 6.0    | 15.9   | 70     |
| 10  | 144      | 113    | 10.1   | 42.8   | 41     |
| 11  | 173      | 123    | 8.7    | 19.0   | 80     |
| 12  | 132      | 131    | 13.8   | 29.2   | 38     |
| 13  | 162      | 137    | 7.2    | 20.7   | 56     |
| 14  | 169      | 129    | 8.5    | 16.7   | 58     |
| 15  | 129      | 138    | 6.3    | 10.1   | 47     |
| 16  | 166      | 148    | 11.5   | 33.4   | 49     |
| 17  | 185      | 118    | 6.0    | 17.5   | 69     |



续表

| No. | 载脂蛋白 A I | 载脂蛋白 B | 载脂蛋白 E | 载脂蛋白 C | 高密度脂蛋白 |
|-----|----------|--------|--------|--------|--------|
| 18  | 155      | 121    | 6.1    | 20.4   | 57     |
| 19  | 175      | 111    | 4.1    | 27.2   | 74     |
| 20  | 136      | 110    | 9.4    | 26.0   | 39     |
| 21  | 153      | 133    | 8.5    | 16.9   | 65     |
| 22  | 110      | 149    | 9.5    | 24.7   | 40     |
| 23  | 160      | 86     | 5.3    | 10.8   | 57     |
| 24  | 112      | 123    | 8.0    | 16.6   | 34     |
| 25  | 147      | 110    | 8.5    | 18.4   | 54     |
| 26  | 204      | 122    | 6.1    | 21.0   | 72     |
| 27  | 131      | 102    | 6.6    | 13.4   | 51     |
| 28  | 170      | 127    | 8.4    | 24.7   | 62     |

SPSS 操作步骤:

1. 数据录入

(1) 点击“变量视图”,在“名称”列下输入“x1(载脂蛋白 A I ,mg/dl)”、“x2(载脂蛋白 B ,mg/dl)”、“x3(载脂蛋白 E ,mg/dl)”、“x4(载脂蛋白 C ,mg/dl)”和“y(高密度脂蛋白中的胆固醇含量,mg/dl)”五个变量。

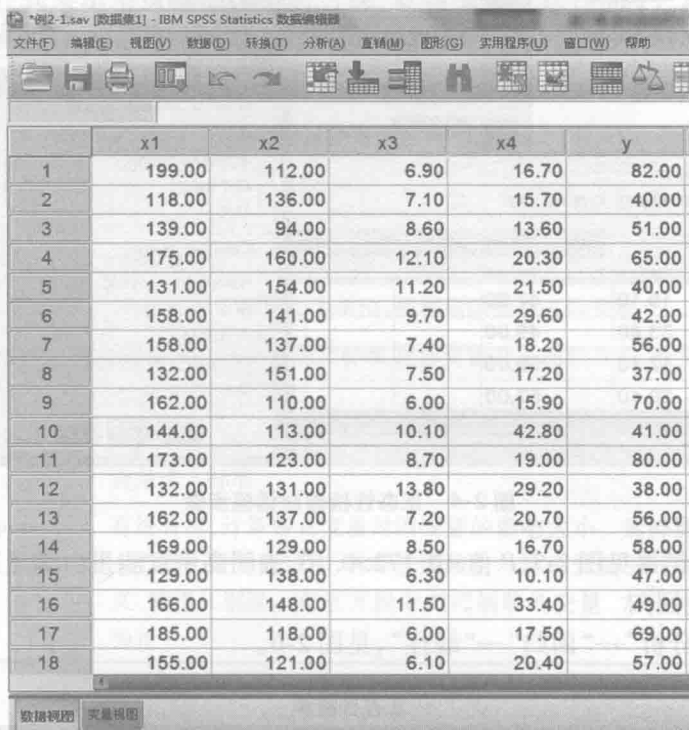
(2) 在“标签”列分别依次输入“载脂蛋白 A I ”,“载脂蛋白 B”,“载脂蛋白 E”,“载脂蛋白 C”,“高密度脂蛋白”,见图 2-1。



图 2-1 定义变量

(3) 点击“数据视图”，按表 2-1 录入全部数据，见图 2-2。

(4) 点击“另存为”，选择合适的保存路径。



|    | x1     | x2     | x3    | x4    | y     |
|----|--------|--------|-------|-------|-------|
| 1  | 199.00 | 112.00 | 6.90  | 16.70 | 82.00 |
| 2  | 118.00 | 136.00 | 7.10  | 15.70 | 40.00 |
| 3  | 139.00 | 94.00  | 8.60  | 13.60 | 51.00 |
| 4  | 175.00 | 160.00 | 12.10 | 20.30 | 65.00 |
| 5  | 131.00 | 154.00 | 11.20 | 21.50 | 40.00 |
| 6  | 158.00 | 141.00 | 9.70  | 29.60 | 42.00 |
| 7  | 158.00 | 137.00 | 7.40  | 18.20 | 56.00 |
| 8  | 132.00 | 151.00 | 7.50  | 17.20 | 37.00 |
| 9  | 162.00 | 110.00 | 6.00  | 15.90 | 70.00 |
| 10 | 144.00 | 113.00 | 10.10 | 42.80 | 41.00 |
| 11 | 173.00 | 123.00 | 8.70  | 19.00 | 80.00 |
| 12 | 132.00 | 131.00 | 13.80 | 29.20 | 38.00 |
| 13 | 162.00 | 137.00 | 7.20  | 20.70 | 56.00 |
| 14 | 169.00 | 129.00 | 8.50  | 16.70 | 58.00 |
| 15 | 129.00 | 138.00 | 6.30  | 10.10 | 47.00 |
| 16 | 166.00 | 148.00 | 11.50 | 33.40 | 49.00 |
| 17 | 185.00 | 118.00 | 6.00  | 17.50 | 69.00 |
| 18 | 155.00 | 121.00 | 6.10  | 20.40 | 57.00 |

图 2-2 数据录入

## 2. 正态性检验

(1) 点击“分析”——“描述统计”——“探索”，弹出“探索”对话框，见图 2-3。



图 2-3 正态性检验菜单操作

(2) 点击“y”，点击箭头，把“y”选入“因变量列表”；点击“绘制”，弹出图对话框，点击“待检验的正态图”，见图 2-4，点击“继续”，点击“确定”。



图 2-4 正态性检验对话框步骤

正态性检验结果见图 2-5,  $P$  值 = 0.172 > 0.10, 表明高密度脂蛋白满足正态性要求。

### 3. 多重线性回归

(1) 点击“分析”—“回归”—“线性”, 见图 2-6。

正态性检验

|        | Kolmogorov-Smirnov <sup>a</sup> |    |      | Shapiro-Wilk |    |      |
|--------|---------------------------------|----|------|--------------|----|------|
|        | 统计量                             | df | Sig. | 统计量          | df | Sig. |
| 高密度脂蛋白 | .138                            | 28 | .182 | .948         | 28 | .172 |

a. Lilliefors 显著水平修正

图 2-5 正态性检验结果



图 2-6 线性回归菜单操作

(2) 在弹出的“线性回归”对话框中,点击“y”,点击箭头,把“y”选入“因变量”;依次点击“x1”、“x2”、“x3”、“x4”,点击箭头,把“x1”、“x2”、“x3”、“x4”选入“自变量”;从“方法”框内下拉式菜单中选择回归分析方法,有强迫进入法(Enter),强制剔除法(Remove),向前选择法(Forward),向后剔除法(Backward)及逐步回归法(Stepwise)5种,各方法说明见表2-2,本例中选择“逐步回归法(Stepwise)”,见图2-7。

表 2-2 自变量的筛选方法

| 自变量的筛选方法        | 说 明                                                                                             |
|-----------------|-------------------------------------------------------------------------------------------------|
| 强迫进入法(Enter)    | 全部自变量纳入回归模型,不做筛选                                                                                |
| 强制剔除法(Remove)   | 只出不进,以 Block 为单位,根据剔除标准去除变量                                                                     |
| 向前选择法(Forward)  | 只进不出,逐一让满足标准的自变量进入回归方程,直至无新的变量可引入为止                                                             |
| 向后剔除法(Backward) | 只出不进,逐一删除对因变量贡献小的自变量,直至进入方程的自变量均满足入选标准                                                          |
| 逐步回归法(Stepwise) | 有进有出,计算各自变量对因变量的影响大小。选择影响最大的变量进入方程。重复此过程,若新变量的引入导致先前引入的变量无统计学意义,就将其剔除。直至方程中无可剔除的变量、方程外无可引入的变量为止 |

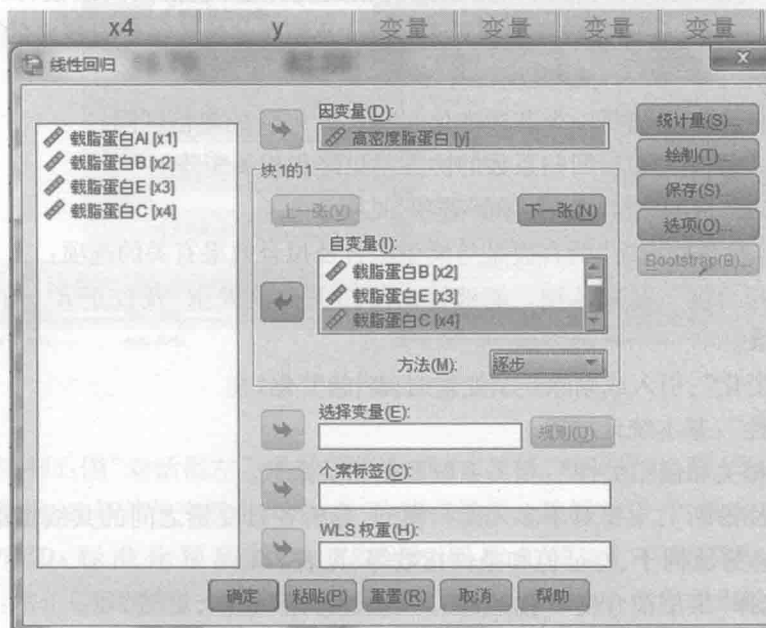


图 2-7 “线性回归”对话框

(3) 单击“统计量”,弹出“线性回归:统计量”对话框,可以选择输出的统计量,见图2-8。

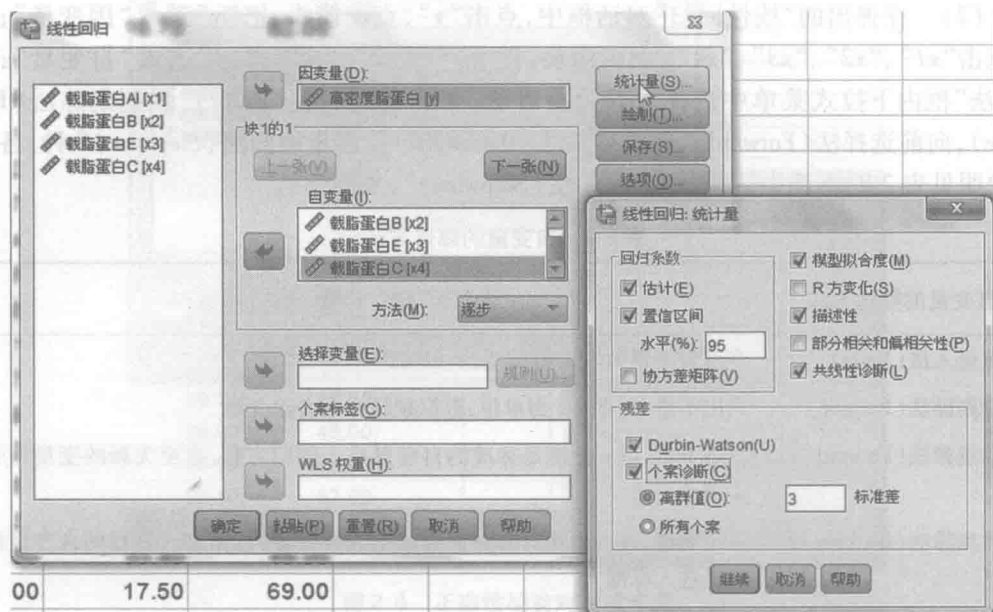


图 2-8 “线性回归:统计量”对话框

下面对“线性回归:统计量”对话框各部分作详细介绍

### 1) 回归系数选项:

①“估计”(系统默认):输出回归系数的相关统计量:包括回归系数、回归系数标准误、标准化回归系数、回归系数检验统计量( $t$ 值)及相应的检验统计量概率的 $P$ 值( $sig.$ )。本例中选择此项。

②“置信区间”:输出每一个非标准化回归系数 95% 的置信区间。

③“协方差矩阵”:输出回归系数的协方差矩阵和相关矩阵。

本例中选择“估计”和“置信区间”选项,见图 2-8。

### 2) 在“回归系数选项”的右边是与模型拟合及拟合效果有关的选项:

①“模型拟合度”:是默认项。能够输出复相关系数 $R$ 、 $R^2$ 及校正 $R^2$ ,估计值的标准误,方差分析表。

②“ $R$ 方变化”:引入或删除一个变量时, $R^2$ 的变化。

③“描述性”:基本统计描述。

④“部分相关和偏相关性”:相关系数及偏相关系数。

⑤“共线性诊断”:主要对于多元回归模型,分析各自变量之间的共线性的统计量,包括容忍度、方差膨胀因子、特征值和条件指数等,见表 2-3。

本例中选择“模型拟合度”“描述性”“共线性诊断”选项,见图 2-8。

### 3) 残差选项:

①Durbin-Watson 检验(简称 DW 检验):DW 检验统计量,用于检验残差的独立性,其取值范围为: $0 < DW < 4$ 。若  $DW = 2$ ,表明相邻两点的残差项相互独立;若  $0 < DW < 2$ ,表明相邻两点的残差项正相关;若  $2 < DW < 4$ ,表明相邻两点的残差项负相关。