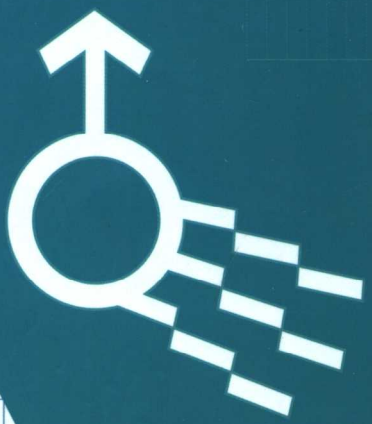
21世纪高等院校经济类与管理类教材

21SHIJI GAODENG YUANXIAO JINGJILEI YU GUANLILEI JIAOCAI

回归分析方法原理及SPSS实际操作



冯力著

中国金融出版社
CHINA FINANCIAL PUBLISHING HOUSE

21世纪高等院校经济类与管理类教材

回归分析方法原理及SPSS实际操作

冯　力　著

中国金融出版社

责任编辑：彭元勋
责任校对：孙　蕊
责任印制：丁淮宾

图书在版编目（CIP）数据

回归分析方法原理及 SPSS 实际操作/冯力著 .—北京：中国金融出版社，2004.5
ISBN 7-5049-3386-4

Ⅰ.回…　Ⅱ.冯…　Ⅲ.①计算机应用—回归分析 ②统计分析—软件包，SPSS　Ⅳ.①0212.1 ②C819

中国版本图书馆 CIP 数据核字（2004）第 040118 号

出版
发行　中国金融出版社
社址　北京市广安门外小红庙南里 3 号
发行部：66024766　读者服务部：66070833　82672183
http://www.chinafph.com
邮编　100055
经销　新华书店
印刷　松源印刷有限公司
尺寸　148 毫米×210 毫米
印张　7.5
字数　223 千
版次　2004 年 5 月第 1 版
印次　2004 年 5 月第 1 次印刷
印数　1—3000
定价　15.00 元

前　言

《回归分析方法原理及SPSS实际操作》的写作目的，是面向非数学专业的在校本科生和实际从事数据分析工作的人员，用尽可能通俗的语言讲解回归分析的方法原理及SPSS的相关操作。

为便于读者对回归分析整个概念体系的融会贯通和SPSS操作方法的实际掌握，本书采用了一个贴近大多数人生活的例子——30名儿童体重、身高、胸围、腰围数据集。这个数据集将贯穿全书的始终，相信您一定会举一反三，将其中所包含的方法原理应用到您所熟悉的实际问题中去。

相信您有很强的理解能力，但阅读本书之前一些基础的统计学知识还是必需的。这里列出一些有关的知识点，供您参考：均值、方差和标准差、概率密度、抽样和抽样分布、区间估计和假设检验。

特别感谢以下诸位老师，他们对本书的写作给予了热情的鼓励和及时的帮助，当然，他们对书中尚存的缺陷没有任何责任。他们是东北财经大学博士后赵进文老师、清华大学博士后宋煦光老师、东北财经大学博士孙玉环老师、东北财经大学博士陈梦根老师、吉通电脑公司刘小娟女士。

本书的写作参阅了以下书目，在此向它们的作者和出版

者一并致谢：《计量经济学》（[美] Danidar N.Gujarati 著，中文版，北京，中国人民大学出版社，1996）；《统计学基本概念与方法》（[美] Gudmund R.Iversen 等著，中文版，北京，高等教育出版社，2000）；《商务与经济统计》，（[美] David R.Anderson 等著，中文版，北京，机械工业出版社，2000）；《SPSS for Windows 10.0 科研统计应用》（贾恩志等主编，南京，东南大学出版社，2001）。

作者简介

冯力，男，1960年1月出生于大连。1984年7月毕业于辽宁财经学院计统系，获经济学学士学位；1997年获经济学硕士学位；博士生。1998年至1999年为西班牙马德里大学访问学者。现任东北财经大学统计学系副教授、硕士生导师。主讲的课程有“统计学”、“SAS统计”、“SPSS统计”等。代表性学术成果有《商业银行信息管理理论与实务》、《信托理论与实务》等著作3部；参加译著《欧洲企业与消费者调查》1部；《论我国统计信息系统结构变迁》、《景气调查误差问题研究》、《统计数字真实性与准确性辨义》等论文10篇。

目　录

第一章 导 言

一、总体

您所感兴趣的所有个体构成了总体。您可能对儿童的健康状况感兴趣，此时，全体儿童就构成一个总体，而某一个儿童就是一个个体；您还可能对股票的价格变动感兴趣，此时，所有的股票就构成一个总体，而某一支股票就是一个个体。无论您的兴趣在哪里，您总要面对某一个总体及其所包含的个体。

总体有大小之分，一个总体所包含的个体数目越多，这个总体就越大。“全体儿童”总体显然远远大于“所有的股票”总体，因为前者比后者包含了更多的个体。

总体还有有限总体与无限总体之分，一个总体所包含的个体数目如果是有限的，这个总体就是一个有限总体；如果所包含的个体数目是无限的，则是一个无限总体。“全体儿童”总体尽管包含了很多的个体，但数目毕竟是有限的，所以它是一个有限总体。如果您是一家饭店的老板，您肯定对您的顾客感兴趣，“您的全体顾客”是您所要面对的总体，这个总体是有限总体还是无限总体呢？事实上它是一个无限总体，因为它包含的个体数目并非是有限的，而是无限的，随时随地都可能有顾客走进您的饭店。

通常情况下，您所面对的多为有限总体。但在统计学家那里，他们更喜欢无限总体，因为，对于无限总体，现成的数学方法可以直接加以引用，而不必为分析结果只是一种“近似”而心有不安。但如果一个有限总体所包含的个体数目足够大，此时将它看作无限总体用数学方法加以处理也应该算作是一种有效的“近似”，结果仍然是可取的。

同一个总体的概念，在数学家看来情形又有所不同，在数学家眼

里没有所谓个体，他们把某一变量的所有可能取值看作是一个总体。

二、变量

变量是我们为总体所包含的个体在某一方面的属性所起的名称。譬如：儿童的性别、年龄、体重、身高、胸围、腰围等您所关心的属性，又可以被称作变量。变量的具体取值最初是体现在总体中的每一个个体身上，譬如：某一儿童性别为男、年龄为 8 岁、体重为 22 公斤、身高为 120 厘米、胸围 57 厘米等、腰围 51 厘米，都是性别、年龄、体重、身高、胸围、腰围这些变量的具体取值。当然这些取值要经过实际的观察和测量才能获得。

变量根据其具体取值的测量尺度不同可区分为三种类型：名义型变量、顺序型变量和尺距型变量。名义型变量的取值表明个体所属的类别，反映个体之间的类别差异。性别就是一个名义型变量，儿童按性别可区分为男、女两类。顺序型变量的取值表明个体之间的等级或顺序差异。譬如：考试成绩区分为优、良、中、及格、不及格，此时的考试成绩就是一个顺序型变量。顺序型变量的取值不仅能够反映个体之间的类别差异，还能够反映类型之间的顺序差异，但它反映不出类别间差异的大小。尺距型变量的取值不仅能够反映个体之间的类别差异和顺序差异，而且还能够反映差异的大小。从反映事物数量方面的精确程度上看，尺距型变量是最高级的变量类型。年龄、体重、身高等都属于尺距型变量。10 岁比 8 岁大 2 岁；22 公斤比 20 公斤重 2 公斤；120 厘米比 115 厘米高 5 厘米。回归分析中所处理的一般都是尺距型变量。

三、样本

您的最终目的是要把握总体在某一个或几个变量上的数字特征，譬如："全体儿童"总体在体重变量上的平均体重是多少？在身高变量上的平均身高是多少？或者在某个体重或身高范围内的人数比重是多少？或者总体中体重与身高两个变量间的相关程度是多少？等等。上述这种就给定的变量由总体所计算得来的数，称作总体参数。显

然，为获得总体参数的具体值，您应该做的第一件事就是面向总体所包含的个体搜集有关方面的数据，接下来再就所搜集得到的数据进行加工计算得到想要的结果。

理想的情况是，就总体所包含的全部个体来搜集数据，进而经过加工计算得到反映总体数字特征的精确的参数值。但由于数据搜集成本的限制，通常情况下很难甚至无法做到这一点。现实的做法是，就总体中的一个样本来进行数据搜集，得到样本数据，由样本数据计算与总体参数对应的统计量，再由统计量的值来推断对应的总体参数。样本统计量是就给定的变量由样本所计算得来的数。

所谓样本就是遵循了随机原则从总体中抽取的一部分个体所构成的集合。譬如：从“全体儿童”总体中随机地抽取 30 名，测量他们的体重、身高等等，这 30 名儿童就构成了问题中的样本。随机原则又可理解为“等可能性”原则，即样本中的每一个个体被抽中的可能性是相等的。违反了这一原则，则不能称其为样本，因而也就不能随意引用概率论与数理统计方法来推断总体。

由于样本所包含的个体数目总是小于总体所包含的个体数目，又由于样本是随机地抽取的，所以，在您抽中某一个样本之前，客观上存在着许许多多的可能样本。譬如：从一个包含 10000 个个体的总体中，抽取一个包含 30 个个体的样本，此时，实际上有 10000^{30} 个可能样本，比总体的个体数目还要多。这 10000^{30} 个可能样本都有被抽中可能，而且可能性是相等的，您最终所抽中的只是其中的一个而已。由此可见，样本的获得是带有随机性的，因而样本统计量也是带有随机性的。但总体参数却始终是一个确定的常数。统计学家们通常所做的事情如果用一句话来概括，就是用带有随机性的样本统计量来推断客观上为某一常数的总体参数。

四、数据集

您从总体中抽取了一个样本，并就某一个或几个您所关心的变量测量了样本中各个个体的具体取值，摆在您面前的就会是一大堆比较零乱的数据，对它们稍作整理就会获得一个数据集。表 1 - 1 是从

“全体儿童”总体中，抽取了一个容量为 30 的样本，测量了 30 名儿童体重、身高、胸围、腰围之后，经过整理所获得的一个数据集。

表 1-1 30 名儿童体重、身高、胸围、腰围数据表

编号	体重	身高	胸围	腰围
1	22.6	119.8	60.5	52.5
2	21.5	121.7	55.5	45.4
3	19.1	121.4	56.5	47.5
4	21.8	124.4	60.5	52.5
5	21.5	120.0	57.7	47.7
6	20.1	117.0	57.0	49.0
7	18.8	118.0	57.1	45.1
8	22.0	118.8	61.7	51.7
9	21.3	124.2	58.4	49.4
10	24.0	124.8	60.8	49.8
11	23.3	124.7	60.0	50.0
12	22.5	123.1	60.0	51.0
13	22.9	125.3	65.2	55.2
14	19.5	124.2	53.7	44.9
15	22.9	127.4	59.5	49.5
16	22.3	128.2	60.1	53.1
17	22.7	126.1	57.4	47.4
18	23.5	128.6	60.4	51.4
19	21.5	129.4	52.0	43.0
20	25.5	126.9	61.5	51.5
21	25.0	126.5	63.9	54.9
22	26.1	128.2	63.0	52.7
23	27.9	131.4	63.1	54.1
24	26.8	130.8	61.5	54.5
25	27.2	133.9	65.8	55.8
26	24.4	130.4	62.6	50.6
27	24.4	131.3	59.5	47.5
28	23.0	130.2	62.5	53.5
29	26.3	136.0	60.0	50.0
30	28.8	138.0	63.7	53.7

表 1－1 数据集中，从列的方向上看，给出了体重、身高、胸围、腰围四个变量，当然，您也可以把编号看作是一个变量，尽管您的兴趣不在编号上面。从行的方向上看，每一行叫做一个观测，每一个观测给出了每一个个体对应于各个变量的观测值。数据集对应的是样本数据，它是任何一项统计分析工作的直接对象。

五、变量间的关系

观察表 1－1 数据集，您可能对其中的体重变量感兴趣，也可能对身高变量感兴趣；进一步您还可能对体重与身高两变量之间的关系感兴趣；或者再进一步您可能对体重、身高、胸围三个变量之间的关系感兴趣。您会问：随着此变量取值的增加，彼变量取值是否也会增加？或者减少？或者不变？此变量增加一个单位，彼变量随之增加或减少几个单位？

您还应该问自己：如果从样本数据上看，变量之间存在着这样或那样的关系，那么这种关系在总体中也存在吗？回归分析就是在回答上述问题的过程中逐步发展起来的一种专门的统计分析方法。

当您打算用此变量来解释彼变量时，此变量就被称作自变量或解释变量，通常用 X 表示；彼变量称作因变量或响应变量，通常用 Y 表示。

变量有名义型、顺序型、尺距型三种类型，因此，变量间的关系会有 9 种不同的组合形式，解决不同组合形式下变量间的关系问题，所采用的统计分析方法有所不同，我们将要讨论的回归分析方法一般来讲只是针对其中自变量和因变量均为尺距型变量的情况。参见表 1－2。

表 1－2 变量类型的各种组合及其对应的统计分析方法

		自变量		
		名义型变量	顺序型变量	尺距型变量
因变量	尺距型变量	方差分析		回归分析
	顺序型变量		秩分析	
	名义型变量	列联分析		Logistic 分析

六、“回归”一词的由来

您不必在“回归”一词上费太多脑筋。英国著名统计学家弗朗西斯·高尔顿（Francis Galton，1822—1911）是最先应用统计方法研究两个变量之间关系问题的人，“回归”一词是由他引入的。他对父母身高与儿女身高之间的关系很感兴趣，并致力于此方面的研究。高尔顿发现，虽然有一个趋势：父母高，儿女也高；父母矮，儿女也矮，但从平均意义上说，给定父母的身高，儿女的身高却趋向于或者说回归于总人口的平均身高。换句话说，尽管父母双亲都异常高或异常矮，儿女身高并非也普遍地异常高或异常矮，而是具有回归于人口总平均高的趋势。更直观地解释，父辈高的群体，儿辈的平均身高低于父辈的身高；父辈矮的群体，儿辈的平均身高高于其父辈的身高。用高尔顿的话说，儿辈身高“回归”到中等身高。这是回归一词的最初由来。

回归一词的现代解释是非常简洁的：回归是研究因变量对自变量

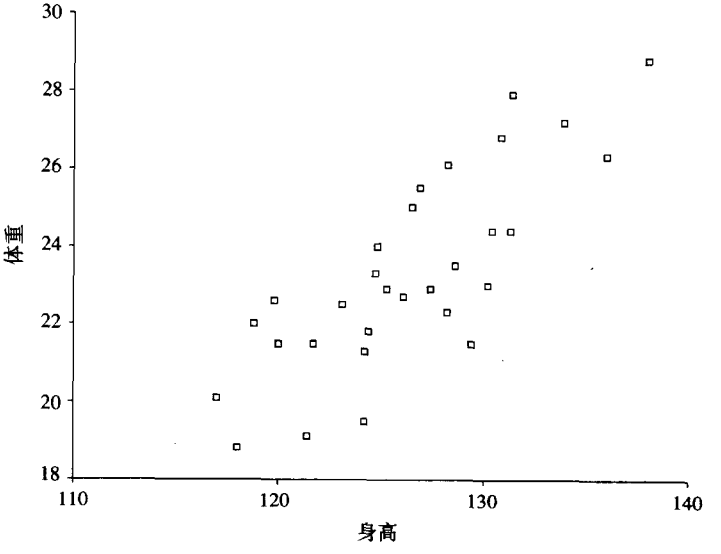


图 1－1　30 名儿童身高与体重散点图

的依赖关系的一种统计分析方法，目的是通过自变量的给定值来估计或预测因变量的均值。

图 1－1 称为散点图，横轴代表自变量身高，纵轴代表因变量体重，图中的各个点，由每一个观测上面的身高和体重的观测值确定。它直观地显示了表 1－1 数据集中 30 名儿童身高与体重之间的关系。观察图形可知，这 30 名儿童身高与体重之间具有一种协变关系，整体上讲，体重随着身高的增加而增加，并且呈直线的趋势。

运用后面将要讨论的线性回归分析的方法，可以为图 1－1 配合出一条直线，参见图 1－2，这条直线叫做回归直线，它给出了任意一个身高上的体重均值。

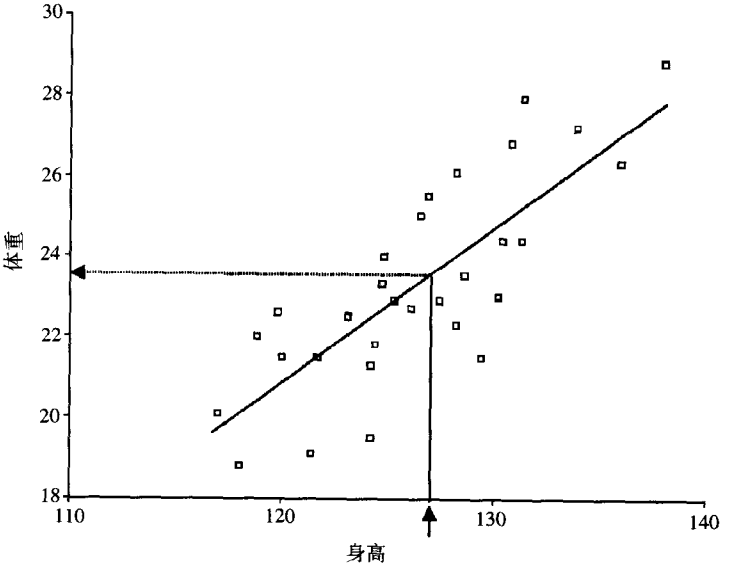


图 1－2 30 名儿童身高与体重散点图及其回归直线

七、用 SPSS 建立数据集

SPSS 是 Statistical Package for Social Science 的缩写，即：社会科学

统计分析软件包，是一种汉化程度比较高的统计分析软件，运用SPSS可以进行几乎所有的统计分析操作，回归分析是其中的一项重要功能。在这里不打算全面叙述SPSS的操作方法，仅就SPSS for Windows 11.0版本，陆续介绍与回归分析有关的操作。如果您从未接触过SPSS，不妨就从回归分析入手，一门深入，融会贯通。

（一）SPSS 11.0 for Windows的安装

SPSS11.0 for Windows可运行于Windows 98、Windows NT 4.0、Windows ME、Windows2000和Windows XP操作系统。作为Windows下的软件产品，具有和其他软件基本相同的安装步骤。具体步骤如下：

1. 启动Windows，将SPSS for Windows 11.0光盘插入光驱；

2. 在“我的电脑”中点击“E”盘或“F”盘，找到SPSS文件夹，点击“setup.exe”，启动安装程序；

3. 根据安装程序的提示向导，依次进行安装，并输入软件系列号码、用户名和单位名称；

4. 退出安装程序；

5. 系统在Windows程序管理器窗口中建立一个SPSS程序组；

6. 用户还可以在桌面上建立一个SPSS的快捷方式，以方便今后的使用。

（二）SPSS11.0 for Windows的启动

在Windows的程序管理器中或桌面的快捷方式上双击SPSS11.0 for Windows的图标，即可启动SPSS11.0。SPSS11.0启动成功后，出现SPSS11.0封面及主窗口，1秒钟后，封面消失，SPSS11.0停留在主窗口，见图1-3、图1-4。

（三）SPSS11.0主窗口

SPSS11.0主窗口的默认标题名称为Untitled - SPSS Data Editor（当您调用其他已经命名的数据集时，标题名称显示为数据集的名称），显示在顶部的左端。顶部的右端为窗口控制钮，点击它可以进行窗口的最小化、还原、关闭等操作。主窗口最底部为系统状态栏，显示系统当前的工作状态。

SPSS11.0主窗口有两个界面，一个是Data View界面，另一个是

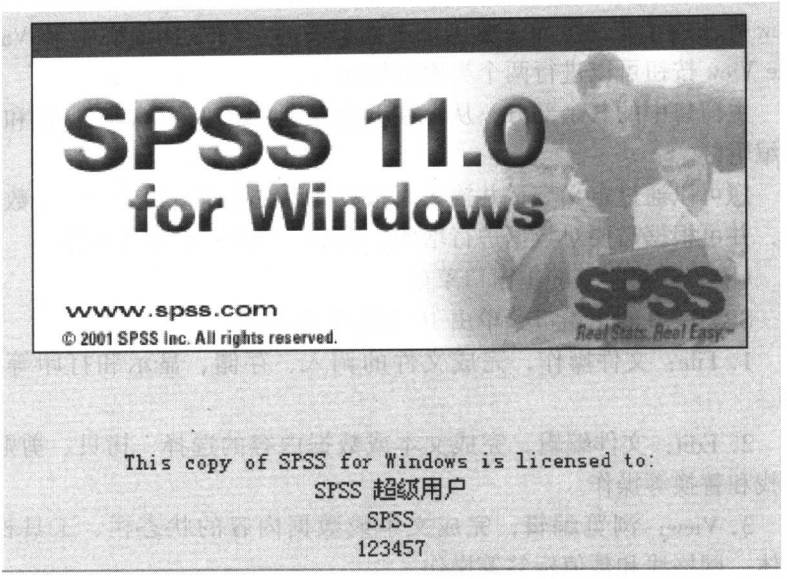


图 1－3 SPSS11.0 封面

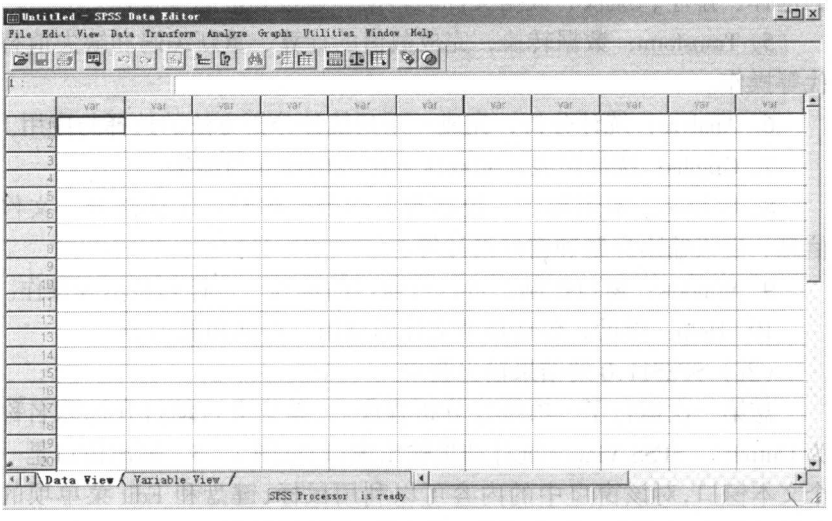


图 1－4 SPSS11.0 主窗口