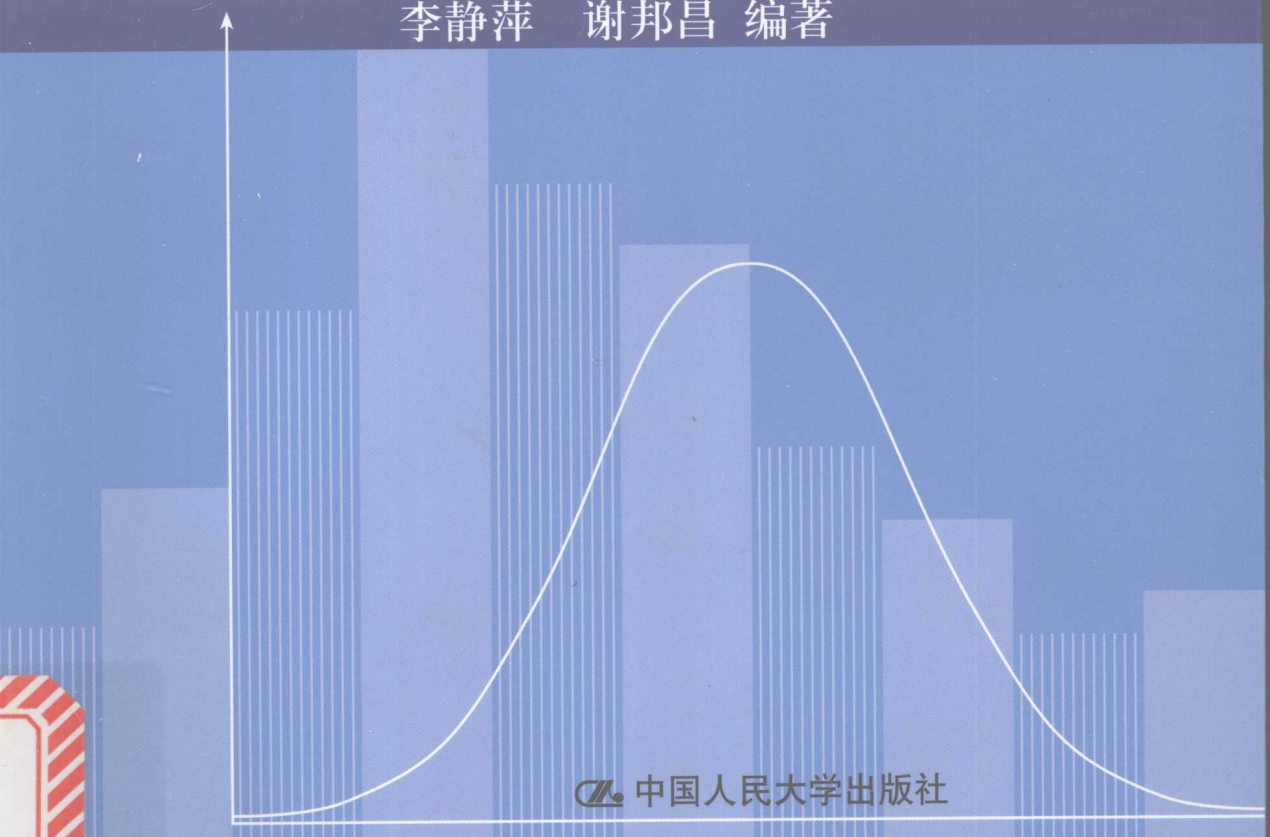
中国人民大学统计咨询研究中心
中国人民大学数据挖掘中心
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心
联合推出

数据分析系列教材

多元统计分析方法与应用

李静萍　谢邦昌　编著



中国人民大学出版社

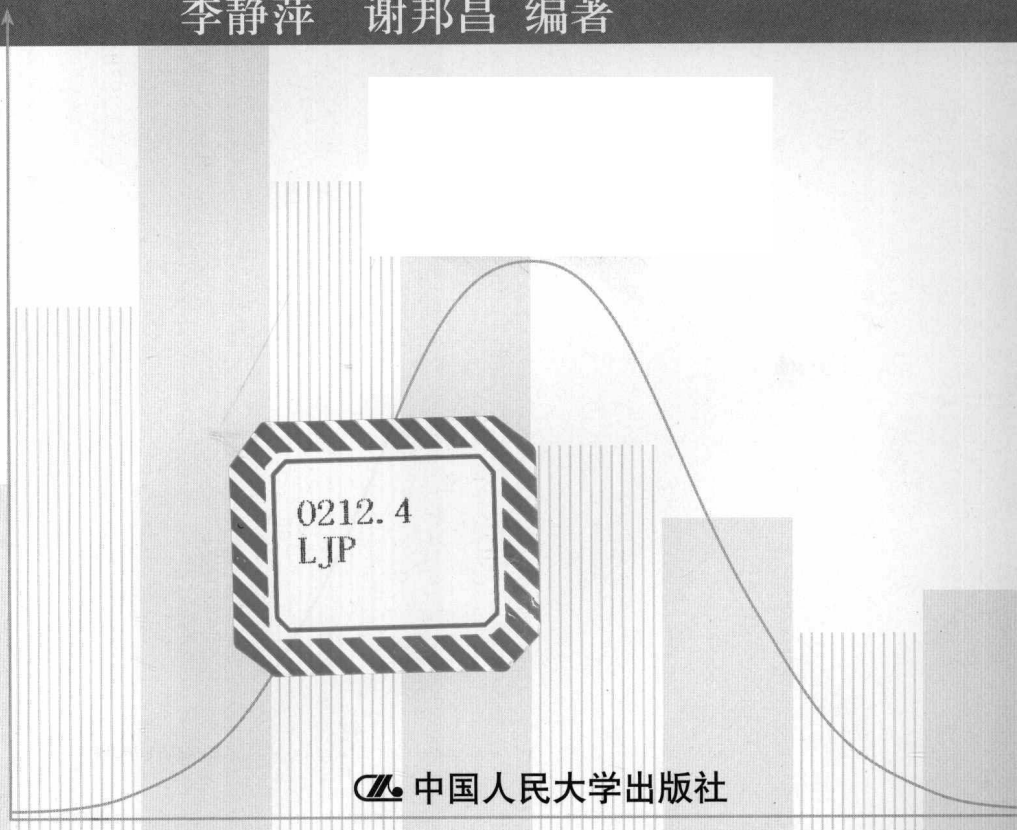
中国人民大学统计学
中国人民大学数[illegible]
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心

联合推出

数据分析系列教材

多元统计分析方法与应用

李静萍　谢邦昌　编著



中国人民大学出版社

图书在版编目(CIP) 数据

多元统计分析：方法与应用/李静萍，谢邦昌编著.
北京：中国人民大学出版社，2008
（数据分析系列教材）
ISBN 978-7-300-09290-4

Ⅰ. 多…
Ⅱ. ①李…②谢…
Ⅲ. 多元分析：统计分析-高等学校-教材
Ⅳ. O212.4

中国版本图书馆 CIP 数据核字（2008）第 062695 号

数据分析系列教材
多元统计分析：方法与应用
李静萍　谢邦昌　编著

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511398（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京丰印诚印务有限公司		
规　　格	170 mm×228 mm　16 开本	**版　　次**	2008 年 6 月第 1 版
印　　张	14.5 插页 1	**印　　次**	2008 年 6 月第 1 次印刷
字　　数	263 000	**定　　价**	29.00 元

《数据分析系列教材》编委会

总　序

随着社会经济的不断发展、科学技术的不断进步，统计方法越来越成为人们必不可少的工具和手段。在教学过程中，老师们也越来越感到运用统计方法解决实际问题的重要，不少人在探索如何运用统计软件介绍和学习统计方法。谢邦昌教授、黄登源教授在多年的教学中，积累了丰富的经验，他们热情倡议，将他们的讲稿提供出来并编写成教材，供更多的人学习和使用。这正与我们的初衷不谋而合。2005 年开始着手这套系列教材的编写，经过不断讨论、反复论证，形成了现在的模式。由于有许多研究生的帮忙，又有几位年轻老师的辛劳，这套书终于问世。

在我们看来，掌握统计方法不仅要理论上弄明白，更重要的在于能够正确有效地运用这些方法，分析说明实际问题。这套书正是试图利用实际数据，通过统计软件的实际操作，将所能够使用的统计方法加以说明，使读者不仅能够了解相应的统计方法，而且能够通过计算机操作学会运用这些方法处理分析实际数据。希望本套书的出版能够为读者提供这样学习的工具。

由于水平有限，难免有不足之处。恳请读者朋友们提出宝贵意见。我们也会循着这样的思路，在教学以及和读者的交流沟通中不断积累、不断提高、不断完善，奉献给读者更多更好的成果。

感谢为这套书的编写付出汗水的研究生，感谢几位认真用心的年轻老师，感谢中国人民大学出版社的大力支持。为方便读者，书中的所有例题数据，都将放在中国人民大学出版社的网站（www. rdjg. com. cn）上，供读者下载并练习。谢谢读者，希望能够加强沟通和联系，为提高统计方法实际运用的能力和水平共同努力。

易丹辉

前　言

多元统计分析是近年来发展迅速的统计分析方法之一，广泛应用于自然科学和社会科学的各个学科，成为各领域研究者和工作者探索多元世界的强有力工具。目前，市面上已经有不少关于多元统计分析方法的教材或专著，其中也不乏从国外引进或翻译的同类教材，对多元统计的方法原理有很深入的介绍。但是，希望在实际研究或工作中应用多元统计方法的研究者，急需一本既能通俗地介绍多元统计方法原理，又能给予切实的操作指南的参考书。

本书面向读者的上述需求，在深入浅出地讲解多元统计方法原理的基础上，侧重于结合实例介绍多元统计方法的应用。在方法的具体实现上，本书采用了在国内广泛使用的统计软件 SPSS（由于本书介绍的个别多元统计分析功能在 SPSS 中没有覆盖，因此部分章节结合 STATISTIC 统计软件进行实例演示），详细介绍多元统计方法在统计软件中的实现以及计算机输出结果的解读。

从内容编排上，本书基本覆盖了常用的多元统计方法；从写作风格上，用浅显的语言阐明各种多元统计方法的功能和原理；从案例应用上，尽可能详尽地介绍统计软件的各种操作选项和输出结果，力求让本书成为读者实际应用多元统计分析方法的好帮手。

本书可作为统计学、经济学、管理学、心理学、生物医学统计等有关专业的高年级本科生或研究生教材或参考书，亦可供市场研究等各个领域的实际工作者实用参考。为方便读者学习，我们将本书案例的数据放在中国人民大学出版社工商管理分社（www. rdjg. com. cn）的网站上，读者可免费下载。

在本书写作过程中，得到了易丹辉教授的悉心指导和大力支持，她对青年教师的关心和扶助是本书得以顺利完成的坚强后盾，在此致以衷心的感谢。此外，研究生陈堰平同学精心准备了各章的案例分析，为本书付出了大量的劳动，在此也向他表示深深的感谢。当然，文责自负，书中难免有疏漏和错误，全部由编著者承担。

编著者

目　录

回归分析

回归分析是研究一个因变量与一个或多个自变量之间相互关系的统计方法。当我们找到变量间的回归关系以后，可以利用这些关系进行下述分析：

1. 描述（description）。如果确定了居民收入与消费总额的回归方程，就可以了解这两个变量之间的关系。

2. 控制（control）。如果找到了商品价格与需要量之间的回归关系，那么通过控制价格，就可以在一定程度上控制需求量。

3. 预测（prediction）。如果找到了居民收入与消费总额的回归关系，就可以根据居民收入估计当年的消费总额。

1.1 一元回归

1.1.1 回归分析概述

最简单的回归分析是一元线性回归，即只包括一个因变量 Y 和一个自变量 X：

$$Y_i=\beta_0+\beta_1 X_i+\varepsilon_i,\quad i=1,2,\cdots,n \tag{1.1.1}$$

其中：Y=因变量（dependent variable；response variable）

X=自变量（independent variable；explanatory variable；regressor variable）

ε=误差项

式（1.1.1）表现的关系称为线性模型（linear model）。其中，(X_i, Y_i) 表示 (X, Y) 的第 i 个观测值，β_0，β_1 是模型中的参数（regression parameters），又叫做回归系数（regression coefficient）。$\beta_0+\beta_1 X_i$ 为反映统计关系直线的分量，ε_i 为反映其他一切随机因素引起的变动。

经典的回归分析假定对于任意 X_i 值有：

1. $Y_i=\beta_0+\beta_1 X_i+\varepsilon_i$，$i=1, 2, \cdots, n$
2. $E(\varepsilon_i)=0$，$E(Y_i)=\beta_0+\beta_1 X_i$
3. $V(\varepsilon_i)=\sigma^2$，$V(Y_i)=\sigma^2$
4. $\mathrm{Cov}(\varepsilon_i, \varepsilon_j)=0$，当 $i\neq j$ 时
   $\mathrm{Cov}(Y_i, Y_j)=0$
5. $\varepsilon_i \sim N(0, \sigma^2)$
   $Y_i \sim N(\beta_0+\beta_1 X_i, \sigma^2)$

1.1.2 参数估计

1. 普通最小二乘估计（ordinary least square estimation，OLSE）

为了由样本数据得到回归参数 β_0 和 β_1 的估计值，可以使用普通最小二乘估计。对每一个样本观测值，考虑观测值 y_i 与其回归值 $E(y_i)=\beta_0+\beta_1 x_i$ 的离差，该离差当然越小越好。综合考虑 n 个离差值，定义离差平方和为：

$$Q(\beta_0, \beta_1)=\sum_{i=1}^{n}(y_i-E(y_i))^2=\sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_i)^2 \tag{1.1.2}$$

所谓最小二乘法，就是寻找 β_0 和 β_1 的估计值 $\hat{\beta}_0$，$\hat{\beta}_1$，使离差平方和达到最小。

利用微积分可证明使式（1.1.2）极小化的 β_0 和 β_1，其估计量形如式（1.1.3）与式（1.1.4）。

$$\hat{\beta}_1=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2}=\frac{\sum x_i y_i-(\sum x_i \sum y_i)/n}{\sum x_i^2-(\sum x_i)^2/n} \tag{1.1.3}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x} \tag{1.1.4}$$

其中：x_i=第 i 个观察值的自变量值
y_i=第 i 个观察值的因变量值
$\bar{x}$=自变量的平均值
$\bar{y}$=因变量的平均值
n=总观察个数

由于式（1.1.3）的第二种形式避开了计算各个$(x_i-\bar{x})$与$(y_i-\bar{y})$的烦琐过程，所以通常均以此式计算 $\hat{\beta}_1$。然而为避免四舍五入误差，在计算时应尽可能保留多位有效数字，建议至少保留四位有效数字。

最小二乘法所提供的回归方程式，使因变量观察值 y_i 与因变量估计值 $\hat{y}_i$ 之间的离差平方和为最小值。在实际当中，最小二乘法的应用最为广泛。

2. 极大似然估计（maximum likelihood estimation，MLE）

除了最小二乘估计外，极大似然估计法也可以作为参数估计的方法。极大似然估计是利用总体的分布密度或概率分布的表达式及样本提供的信息建立似然函数，从而求解未知参数估计量的一种方法。

当总体 X 为连续型分布时，设其分布密度族为 $\{f(x,\theta),\ \theta\in\Theta\}$，假设总体 X 的一个独立同分布的样本为 x_1，x_2，…，x_n，则其似然函数为：

$$L(\theta;x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f(x_i;\theta) \tag{1.1.5}$$

极大似然估计应在一切 θ 中选取使随机样本（X_1，X_2，…，X_n）落在（x_1，x_2，…，x_n）附近的概率最大的 $\hat{\theta}$ 为未知参数 θ 真值的估计值，即极大似然估计量 $\hat{\theta}$ 满足：

$$L(\hat{\theta};x_1,x_2,\cdots,x_n)=\max_{\theta}L(\theta;x_1,x_2,\cdots,x_n)$$

对连续型随机变量，似然函数就是样本的联合分布密度函数，对离散型随机变量，似然函数就是样本的联合概率函数。似然函数的概念不仅仅局限于独立同分布的样本，只要样本的联合密度的形式是已知的，就可以应用极大似然估计。

3. 回归方程的显著性检验

当我们得到一个实际问题的经验回归方程 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1x$ 后，不能直接利用它进行分析，还需要用统计方法对回归方程进行检验。在对回归方程进行检验时，通常需要正态性假设，即假定 ε_i 服从标准正态分布。

具体检验方法有：

（1）t 检验。t 检验是统计推断中常用的一种检验方法，在回归分析中，t 检验用于检验回归系数的显著性，即检验因变量 y 对自变量 x 的影响程度是否显著。

t 检验的原假设是：H_0：$\beta_1=0$，对立假设是 H_1：$\beta_1\neq0$。

若原假设 H_0 成立，则因变量 y 与自变量 x 之间并没有真正的线性关系，也即自变量 x 对因变量 y 没有影响。

t 检验使用的检验统计量为 t 统计量。给定显著性水平 α，双侧检验的临界值为 $t_{\alpha/2}$。当 $|t|\geqslant t_{\alpha/2}$ 时拒绝原假设，认为 β_1 显著不为 0，因变量 y 对自变量 x 的一元线性回归成立；当 $|t|<t_{\alpha/2}$ 时不能拒绝原假设，认为 β_1 与零没有显著差异，因变量 y 对自变量 x 的一元线性回归不成立。

(2) F 检验。对线性回归方程显著性的另一种检验是 F 检验，F 检验根据平方和分解，直接从回归效果检验回归方程的显著性。

总平方和：$\mathrm{SST}=\sum_{i=1}^{n}(y_i-\bar{y}_i)^2$

回归平方和：$\mathrm{SSR}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$

误差平方和：$\mathrm{SSE}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$

容易得到平方和的分解式为：

$$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE} \tag{1.1.6}$$

总平方和 SST 反映因变量的波动程度，回归平方和 SSR 是由回归方程确定的，也就是自变量 x 的波动引起的，误差平方和 SSE 是不能用自变量解释的波动，是由 x 之外的不能控制的因素引起的。显然，回归平方和 SSR 越大，回归的效果越好。

F 检验的假设与 t 检验相同，其检验统计量如下：

$$F=\frac{\mathrm{SSR}/1}{\mathrm{SSE}/(n-2)} \tag{1.1.7}$$

在正态假设下，当原假设成立时，F 服从自由度为 $(1,\ n-2)$ 的 F 分布。当 F 大于临界值 $F_\alpha(1,\ n-2)$ 时，拒绝原假设，说明回归方程显著，x 与 y 有显著的线性关系。

4. 样本决定系数

利用最小二乘法可求出使因变量的观察值 y_i 与因变量的预测值 $\hat{y}_i$ 之间的离差平方和为最小的 β_0 与 β_1 值。y_i 与 $\hat{y}_i$ 之间的差即为以 $\hat{y}_i$ 估计 y_i 所产生的误差；第 i 个观察值的离差为 $y_i-\hat{y}_i$，此差值也称为第 i 个残差 (residual)。因此，最小二乘法中所处理的平方和，常被称为误差平方和或残差平方和，以 SSE 表示。

由前述回归平方和与残差平方和的含义可知，如果在总离差平方和中回归平方和所占的比重越大，则线性回归效果越好，表明回归直线对样本观测值的拟合

优度越好，最理想的情况是各观察值均落在回归直线上，此时 SSE=0；如果残差平方和所占的比重大，则回归直线与样本观测值拟合效果不理想，最坏的情况是 SSR=0，此时估计回归方程式完全无法预测 y。

将回归平方和与总离差平方和之比定义为样本决定系数，记为 r^2，即

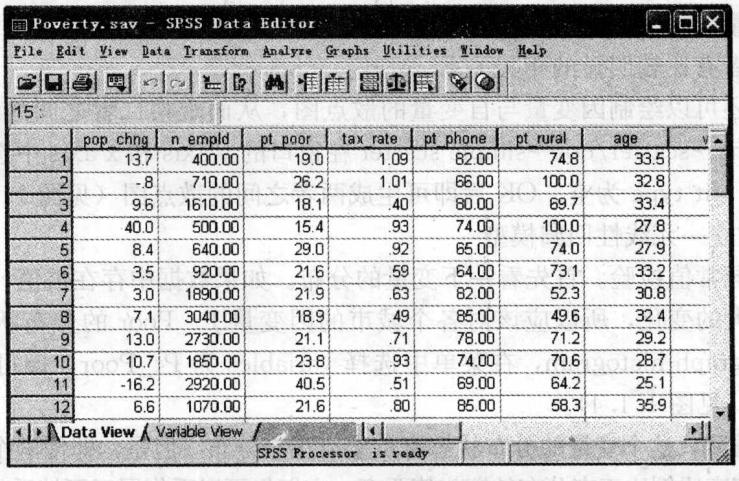
$$r^2 = \frac{\mathrm{SSR}}{\mathrm{SST}} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.1.8}$$

决定系数 r^2 是一个衡量回归直线对样本观测值拟合优度的相对指标，反映了因变量的波动中能用自变量所解释的比例。r^2 的值总是在 0～1 之间，r^2 越接近于 1，拟合优度就越好；反之，说明模型中所给出的 x 对 y 的信息还不充分，回归方程的效果不好，应进行修改，使 x 与 y 的信息得到充分利用。

1.1.3　一元回归应用实例

我们以对数据 Poverty. sav 的分析为例说明回归分析的应用（见图 1.1.1）。该资料是美国 1960—1970 年对随机选择的 30 个城市人口调查结果的比较，按照城市进行分类。在本例中要分析的是有可能与贫困相关的变量以及一个县在贫困线以下的家庭比例。我们看一下已有的变量，根据我们分析的目的选择因变量与自变量，建立一元线性回归模型。

1. 确定因变量与自变量


Poverty.sav - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

15

	pop_chng	n_empld	pt_poor	tax_rate	pt_phone	pt_rural	age
1	13.7	400.00	19.0	1.09	82.00	74.8	33.5
2	-.8	710.00	26.2	1.01	66.00	100.0	32.8
3	9.6	1610.00	18.1	.40	80.00	69.7	33.4
4	40.0	500.00	15.4	.93	74.00	100.0	27.8
5	8.4	640.00	29.0	.92	65.00	74.0	27.9
6	3.5	920.00	21.6	.59	64.00	73.1	33.2
7	3.0	1890.00	21.9	.63	82.00	52.3	30.8
8	7.1	3040.00	18.9	.49	85.00	49.6	32.4
9	13.0	2730.00	21.1	.71	78.00	71.2	29.2
10	10.7	1850.00	23.8	.93	74.00	70.6	28.7
11	-16.2	2920.00	40.5	.51	69.00	64.2	25.1
12	6.6	1070.00	21.6	.80	85.00	58.3	35.9

Data View / Variable View

SPSS Processor is ready

图 1.1.1　Poverty 数据窗体