新编社会学系列教材

多元社会统计分析基础

卢淑华 编著

Multivariate Social Statistical Analysis

北京大学出版社
PEKING UNIVERSITY PRESS

多元社会统计分析基础

卢淑华 编著

Multivariate Social Statistical Analysis

北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

多元社会统计分析基础/卢淑华编著. —北京：北京大学出版社，2017.7
(新编社会学系列教材)
ISBN 978-7-301-28442-1

Ⅰ. ①多… Ⅱ. ①卢… Ⅲ. ①社会统计—多元分析—高等学校—教材
Ⅳ. ①C91-03

中国版本图书馆 CIP 数据核字(2017)第 143064 号

书　　名	多元社会统计分析基础 DUOYUAN SHEHUI TONGJI FENXI JICHU
著作责任者	卢淑华　编著
责任编辑	董郑芳(dzfpku@163.com)
标准书号	ISBN 978-7-301-28442-1
出版发行	北京大学出版社
地　　址	北京市海淀区成府路 205 号　100871
网　　址	http://www.pup.cn
电子信箱	ss@pup.pku.edu.cn
新浪微博	@北京大学出版社　@未名社科—北大图书
电　　话	邮购部 62752015　发行部 62750672　编辑部 62765016
印 刷 者	北京大学印刷厂
经 销 者	新华书店
	730×980 毫米　16 开本　26 印张　449 千字
	2017 年 7 月第 1 版　2017 年 7 月第 1 次印刷
定　　价	68.00 元

前　言

北京大学是改革开放后最早恢复社会学系的，当时在系主任袁方教授的积极倡导下，很注重社会调查方法的研究，一时风光地被同行称作“方法派”，方法课程共分两门：社会调查方法和社会统计学。其中调查方法课程，由王汉生、林彬老师负责；社会统计学课程，包括本科生和研究生的，自恢复以来，一直由我承担，长达十六年之久，直至退休。我为本科生编写的教材《社会统计学》自1989年问世，目前累计印刷已近三十次。供研究生用的多元统计分析教材，一直没有动笔。近年来，我关心了一下当前多元统计方面教材的进展，发现系统介绍多元统计知识的教材，财经类专业新出的教材不少，但社会学专业的教材，几乎没有看到。翻开这些新书，多元分析仍然是围绕三大方法——回归、判别和聚类展开的，这些内容，过去我在课上都讲过，于是萌生了要将过去的讲稿整理成册的想法。我的这个想法，深得北京大学出版社编辑的赞同，因为他们也发现，市场上这方面的教材实在太少了。就这样，本书纳入了北京大学出版社的出版计划，经过了数年的努力，本书在原有讲稿的基础上，吸收了当下所见各家之长，完成了写作。本书有如下几个特点：

第一，本书和其他领域的应用统计有着相同的内容，对于回归、判别和聚类多元三大统计方法，本书都有介绍。但根据社会研究的需要，在内容的重点上，还是有它自己的特点，例如在因果研究中，社会研究并不奢望通过回归方程做出预测，而是希望在找出的原因中，给出哪些因素重要、哪些因素不太重要，也就是回归系数的比较。而影响回归系数比较的原因中，自变量间的相关影响最大，因此本书用相当多的篇幅讨论自变量相关对回归系数的影响，即多重共线的问题。但这样的讨论，并非所有领域都需要，例如气象、地质灾害中的回归分析，关心的是预测准确性，而不是预测中的因素是否相关，因此这些领域在介绍多元回归时，对多重共线并没有那么关心。又如判别分析，实际也是一种预测，它在医学、考古中都有重要的应用，但在社会学中应用不多，因此本书只做简单介绍。

第二，一般来说，多元分析离不开矩阵，而矩阵的表达是比较抽象的，为了减少这些数学工具给读者带来的困难，本书在两个变量就能说明多元复杂性的情况下，尽量用两个变量进行解释，这样在二维空间，不但可以有数字例子，还可以形象地给出平面的几何图形，减少多维空间的抽象性。例如本书第七章“主成分分析”中的特征值和特征向量，都是用二维空间给出解释的。与此同时，每章都有数字例子帮助理解内容和公式，例如回归中的多重共线，本书只用二元回归实例来说明自变量间存在相关带来的复杂性，以及公式中的数量关系。此外，本书有些例子索性只有数，没有赋予具体的内容，这是为了更好地发挥数字本身对理解公式的作用，避免因举例不当引起歧义。可以说，本书是内容解释中运用矩阵符号最少、数字例子最多、最浅显易懂的。

第三，学习统计知识，必须学会统计包，这是毋庸置疑的。没有统计包，统计知识就无法付诸实现。但作为系统介绍统计知识的教材，同时充当统计包教材，是没有必要的。首先，市面上的统计包不止一个（SPSS，SAS，Excel，Matlab等），如果读者手边能用到的统计包，不是书中介绍的，那书中内容岂不浪费了吗？再说即便是书中介绍的，由于版本的不断更新，内容也会过时。况且，这是两种不同的知识体系。正像理科讲授物理有物理教材，讲授怎样做物理实验有物理实验教材一样，没有一本物理教材是同时在讲授怎样做物理实验的。物理实验是为物理内容服务的，所以每个实验都要对实验的物理内容有梗概介绍，正像统计包是为统计内容服务的，所以每一方法都要对方法的统计内容有梗概介绍，但这些梗概介绍都不能替代系统的统计知识。统计知识既不会过时，也不会各书有各书的定义和解释，例如什么是“标准差”，什么是“标准误”，所有教材都有介绍，在所有教材中，它的定义和公式都是相同的。而统计包是资料处理的工具，工具可以不止一个，可以各有各的操作语言，它的学习特点和最好的学习方式是，需要时在使用中学。

本书是系统介绍多元统计知识的，对统计包操作没有介绍，但书中很多例题的计算结果，是以 SPSS 统计包的输出结果给出的，书中对如何读懂统计包的输出有解释。

第四，本书略去了统计量的临界值附表，因为统计包都有统计量对应的概率值。书中由于解释的需要，有时也用到统计量临界值（可以通过扫描书中二维码获得），这时都有说明。

第五，本书在编写过程中，除了尽可能地收集和参考相关的教材外，还参考了 1984 年 5 月 24 日至 6 月 6 日布莱洛克教授来华的讲课内容。当时社会学在我国刚恢复，由中国社科院社会学所主办，有关高校参加，在北京和平宾馆八楼会议厅，布莱洛克教授做了十一次授课，重点介绍了社会统计学中的多元分析，

我在撰写本书时,再次翻阅了当时的笔记,吸收了相关内容。

第六,作为一名多年讲授多元统计分析的教师,我深感目前运用多元分析技术对资料进行处理的效果并不理想,只要翻开任何一篇以多元分析为工具的论文,就可以发现模型的解释力都很低,几乎没有超过30%的,有的甚至只有百分之十几。而即便文章给出了很高的解释力,其结论也只有符合常识,人们才会接受,否则人们宁可相信常识的感觉,也不相信模型给出的结果。如果这样的话,那又何必杀鸡用牛刀呢?实际在布莱洛克来华的讲课中,也有类似的评价,他称社会学是不成熟的学科。在谈到社会学家对待自己的研究成果时,布莱洛克很是风趣地说了两点:一要谦虚,二要幽默。关于第一点,好理解,任何领域的成果都应抱有谦虚的态度;关于第二点,要幽默,更值得玩味,言外之意是对成果别太认真了吧!

正因为这样的原因,本书和其他专业的多元统计不同,多处谈到了测量误差、量度误差理论,引进了无关变量或忽略了有关变量,以及运用拟合而不是显著性进行检验对资料分析结果的影响。所有这些讨论,不是定量研究所特有的,包括定性研究也会遇到。记得布莱洛克曾说起研究控制变量的一个故事,当时由于研究的深入,得出了在多变量的因果模型中,不是模型中每一个变量都可随意拿来作控制变量的,如果用果变量作为控制变量,不仅道理上说不通,而且偏相关会得到一系列不可思议的错误结果。这本是社会研究深化的成果,却被从事定性研究的人,说成是定量研究走入歧途,相当于原本简单的问题,非但没有解决,反而复杂化了。布莱洛克解释说,这样的评价显然是错误的,事实上,从事定性研究同样存在这样的问题,只是还没有意识到里面有问题而已。可见,多元分析,既是资料处理的工具,又是学会深入思考的武器。有了它,我们学会根据因果勾画出现象与现象之间的联系;有了它,我们学会通过表面现象看到隐藏在背后潜在的联系。通过多元统计分析,使我们分析的能力大大提高。随着科技进步和统计知识的普及,一些统计的术语逐渐融入日常用语,例如"标准差""抽样""小概率"几乎随处可见。同样,多元统计中的术语,也并非只出现在资料定量处理的文章中,一般文章中也会用它构建分析问题的思路与术语,这里套用一句现在的流行语,如果没有现代统计知识的积淀,阅读高层次的学术论文时就"out"了。

本书共有十一章,各章内容简述如下:

第一、二章是基础知识,复习本科教材中的一元回归和方差分析,以便和下面的多元统计分析衔接。

第三章介绍三变量的基础知识,以及通过控制变量辨别三变量相关的

类型。

第四章全面介绍多元回归方程的建立;回归系数和一元回归系数的不同;判定系数和复相关系数的意义;多元回归方程和回归系数的检验;以及回归假定的鉴别。在介绍偏相关的部分,在给出运算公式的同时,用例子解释它的含义,运用偏相关对控制变量做进一步讨论,和运用控制变量对假设模型进行验证。本章还涉猎测量误差对相关系数的影响。除了偏相关,还给出了部分偏相关和复偏相关的公式、检验和相互关系。

第五章是多元回归中特有的自变量间相关的讨论,即多重线性问题,由此给出了模型误差,其中包括无关变量和强相关变量的引入、有关变量的忽略,以及测量误差量度误差的影响,最后用 SPSS 统计包输出为例,给出回归方程建立的过程和方法。

第六章,介绍自变量含有定类变量的两种分析方法,一种是以方差分析法为主,将定距变量看作协变量。另一种是以回归分析法为主,将定类变量转换为(0,1)变量,最后对两种方法进行了比较。

第七章开始要用到矩阵,因此在介绍后面各章之前,先对书中要用到的矩阵做一介绍,介绍时都辅以数字例子。其中和后面内容关系最为密切的特征值和特征向量,通过数字例子,给出二维的图形解释。本章主要介绍主成分分析法的意义和求解方法。

第八章介绍因子分析法,因子分析法首先在教育心理学中得到成功的应用。本章介绍了因子分析法的求解方法,由于因子得分是社会研究中因果模型的潜变量,因此本章还介绍了因子得分的求解方法。

第九章介绍聚类分析与判别分析。其中第一节介绍聚类分析,聚类分析就是如何科学分类的问题,分类按最短距离或相关性最强来分是显而易见的,但在聚类过程中,每一类从只有一个个体发展到包含若干个体时,就出现了各种计算类与类之间距离的方法,书中通过数字例子,帮助读者了解各种方法的计算过程。书中给出了各种聚类方法,都有数字例子进行解释。第二节介绍判别分析,它是对一个未知个体如何科学归类的问题。它的道理也是显而易见的,那就是和哪个已知类别相似性越强就归入哪一类。书中列举了三类判别法,判别分析在考古、医学中都有广泛的应用。

第十章是回归分析在定类变量应用中的延伸。本章介绍两种方法,一种称作 Logistic 回归,它是将定类型因变量,通过 Logistic 转换,将定类变量转换为 $[-\infty, +\infty]$ 的连续变量,进而可以按多元回归处理,它和一般回归的不同是,对回归系数的解释,仍要退回到初始定类型因变量的解释,Logistic 回归中的自

变量,可以是定距型,也可是定类型,或是两者兼有的混合型。另一种方法是,对数线性法,它适用于因变量和自变量都是定类型变量的情况,它的做法是对列联表频次取对数,使每一格值可看作是所有自变量包括交互作用的线性叠加,书中对这样分解的道理,通过数字例子做了详细解释,本章用到了对模型检验的拟合概念,它要求不是小概率,而是与之相反,以更大的概率接近被采纳的模型。

第十一章是最后一章,介绍了两种模型的处理方法,第一种是路径分析法,它要求模型中的变量是单向的,称作递归模型,它源于遗传学,社会学中职业流动属于该种模型。另一种是线性结构方程法,它允许模型中的变量是双向的,而且在现成的处理软件中,它要求模型的理论是有充分根据、有说服力的。

在本书编写过程中,始终得到北京大学出版社的鼓励与支持,本书由于图多、表多、公式多,所用字母复杂,其中有英文的、希腊文的,字母还要分大写、小写、粗体、非粗体,外加还要有上标、下标甚至下标的下标,因此排版起来繁琐,有时作者也难免出错,但出版社的编辑和排版老师从无怨言,一遍遍地修改、打印、再打印。因此,没有北京大学出版社全体同仁的辛勤劳动,本书的出版是不可能的。

本书完稿后,曾考虑过请有关老师审阅,但这对审阅人负担太重了,因此是不可行的。好在本书内容给多届同学讲授过,同学学习的过程,一定程度上也是审稿的过程。

感谢北京大学的现任授课老师刘爱玉教授、周飞舟教授、阮桂海教授和林彬教授,他们都看过本书的前言和大纲,并提了宝贵意见,给了我很大的支持与肯定。

还要感谢历届同学,他们认真的学习态度值得称道,至今我还保留了一份同学自己整理的我课堂内容的听课笔记。同学们运用课上学到的统计知识,通过 SPSS 统计包进行处理,完成了毕业论文,本书已将某些同学的成果作为实例,编在内容中。

抛砖引玉,本书欢迎广大师生提出宝贵意见,无论是教材内容,或是印刷错误,都可直接寄北京大学社会学系作者收,或是寄北京大学出版社责任编辑董郑芳收,邮件地址是 ss@ pup. pku. edu. cn。

卢淑华
2016 年 5 月 5 日
于北京大学社会学系

目　录

第一编　基础知识

第二编　多元分析

第一编

基础知识

第一章

回归与相关

第一节　线性回归方程

一、什么是线性回归方程

设有两个变量 x 和 y，当 x 变化时会引起 y 相应的变化，但它们之间的变化关系是不确定的，如果当 x 取得任一可能值 x_i 时，y 相应地服从一定的概率分布，为了研究 x 和 y 之间确定的关系，用 y 概率分布的数字特征（均值）取代分布，于是 x 和均值 y 之间就形成了确定的函数关系。

我们把 $x=x_i$ 条件下，y_i 的均值记作

$$E(y_i)$$

如果它是 x 的函数

$$E(y_i)=f(x_i) \tag{1-1}$$

则表示变量 y 和变量 x 之间存在着相关关系。式(1-1)称作 y 对 x 的回归方程。

当因变量 y 的平均值与自变量 x 呈现线性规律时，称作线性回归方程，这里因为只有一个自变量，又称一元线性回归方程。它的表达式为

$$E(y)=\alpha+\beta x \tag{1-2}$$

其中 α 称作回归常数，β 称作回归系数。

每一个真实 y_i 与回归线的关系是

$$y_i=\alpha+\beta x_i+e_i \tag{1-3}$$

式(1-3)中 y_i 是随机变量，e_i 是随机误差，由于 e_i 的值是非固定的，从而使 x 和 y

呈现非确定性的关系。

二、线性回归方程的建立与最小二乘法

上面所谈变量 x 和变量 y 之间存在线性回归，是指总体而言，对于样本来说，则是通过样本的观测值来估计总体回归直线的 α 和 β。而最小二乘法（Least-squares Criterion）的方法，是通过样本对总体线性回归最好的估计方法。

设从总体中抽取一个样本，其观测值为

$$(x_1, y_1)$$
$$(x_2, y_2)$$
$$\vdots \quad \vdots$$
$$(x_n, y_n)$$

现在围绕这 n 个观测点，画一条直线（图 1-1）。

$$y = a + bx \tag{1-4}$$

可以想象，当 a，b 取不同值时，可以得到无数条直线。那么，在这无数条直线中，哪一条是这 n 个样本点的最佳拟合直线呢？一个很自然的想法，应该是到各点都比较接近的那条直线为最佳。数学上把这样的想法表示为：各点到待估直线铅直距离之和为最小。这就是求回归直线的最小二乘法原理。

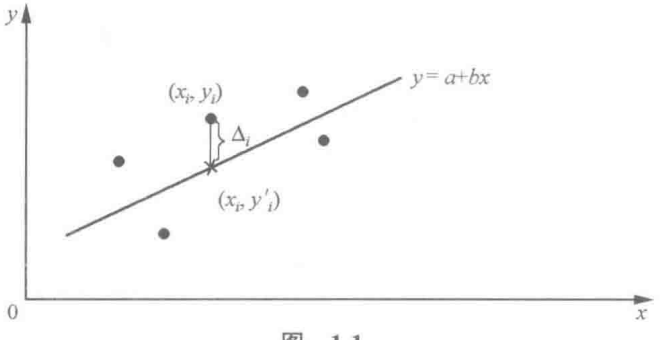

图 1-1

设点 i 的观测值为 (x_i, y_i)，把 x_i 代入待定的直线式（1-4）有

$$y_i' = a + bx_i \tag{1-5}$$

y_i 到待估直线的铅直距离为 y_i 减去 y_i'

$$\Delta_i = y_i - y_i' = y_i - (a + bx_i)$$

n 点铅直距离平方和为

$$Q(a,b) = \sum \Delta_i^2 = \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \tag{1-6}$$

显然，Q 值是 a,b 的函数。根据最小二乘法的原理，就是从不同的 a,b 中求得 $\hat{a},\hat{b}$，使其 $Q(a,b)$ 达最小值

$$\begin{cases}\dfrac{\partial Q}{\partial a}=0\\[2ex]\dfrac{\partial Q}{\partial b}=0\end{cases} \tag{1-7}$$

将式(1-6)代入式(1-7)有

$$\begin{cases}\sum\limits_{i=1}^{n}[y_i-(a+bx_i)]=0\\[2ex]\sum\limits_{i=1}^{n}[y_i-(a+bx_i)]x_i=0\end{cases} \tag{1-8}$$

根据式(1-8)，解二元一次联立方程得

$$a=\bar{y}-b\bar{x} \tag{1-9}$$

$$b=\frac{L_{xy}}{L_{xx}} \tag{1-10}$$

其中

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \tag{1-11}$$

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \tag{1-12}$$

$$L_{xx}=\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \tag{1-13}$$

$$\begin{aligned}L_{xy}&=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\\&=\sum_{i=1}^{n}x_iy_i-\frac{1}{n}\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i\end{aligned} \tag{1-14}$$

为了今后进一步分析的需要，再引入

$$L_{yy}=\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-\frac{1}{n}\left(\sum_{i=1}^{n}y_i\right)^2 \tag{1-15}$$

这样通过最小二乘法所确定的 a,b，代入待估的直线方程式(1-4)得

$$\hat{y}=a+bx \tag{1-16}$$

它将是总体线性回归方程 $y=\alpha+\beta x$ 的最佳估计方程。

第二节　回归方程的假定与检验

一、线性回归模型的基本假定

在第一节中谈到总体的线性回归指的是，当 $x=x_i$ 时，y 的均值 $E(y_i)$ 是 x 的线性函数（式 1-2）：$E(y_i)=\alpha+\beta x_i$。下面就变量及其相互关系给出一些基本假定。

（1）自变量 x 可以是随机变量，也可以是非随机变量。x 值的测量可以认为是没有误差的，或者说误差是可以忽略不计的。

（2）由于 x 和 y 之间存在的是非确定性的相关关系。因此，对于 x 的每一个值 $x=x_i$，y_i 是随机变量，或称作是 y 的子总体。要求 y 的所有子总体 y_1，$y_2,\cdots,y_i,\cdots,y_n$，其方差都相等（equal variance）。

$$D(y_1)=D(y_2)=\cdots=D(y_i)=\cdots=D(y_n)$$

（3）如果 y 的所有子总体，其均值 $E(y_1),E(y_2),\cdots,E(y_i),\cdots,E(y_n)$ 都在一条直线上，则称作线性假定，其数学表达式为

$$E(y_i)=\alpha+\beta x_i$$

由于 α 和 β 对所有子总体都一样，所以 α 和 β 是总体参数。

（4）要求随机变量 y_i 是统计独立的。即 y_1 的数值不影响 y_2 的数值，各 y 值之间都没有关系。

以上称作对总体有关线性、同方差和独立的假定。也可用如下两种数据结构来表达。

① 随机变量 y_i 是统计独立的，且有

均值：$E(y_i)=\alpha+\beta x_i$

方差：$D(y_i)=\sigma^2$

② y_i 与 x_i 有如下关系式：

$$\begin{aligned}
y_1&=\alpha+\beta x_1+\varepsilon_1\\
y_2&=\alpha+\beta x_2+\varepsilon_2\\
&\vdots\quad\quad\vdots\quad\quad\vdots\\
y_n&=\alpha+\beta x_n+\varepsilon_n
\end{aligned}$$

其中 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 是随机变量，它们相互独立，且有

$$E(\varepsilon_i)=0$$

$$D(\varepsilon_i)=\sigma^2$$

当总体具有上述假定时，那么根据样本运用最小二乘法所求得的方程

$$\hat{y}=a+bx \tag{1-17}$$

将是总体线性回归方程

$$E(y)=\alpha+\beta x \tag{1-18}$$

的最佳线性无偏估计方程。式(1-17)中的 a 和 b 将是式(1-18)中 α 和 β 的最佳无偏估计量。

(5) 出于检验的需要，除了上述假定或要求外，还要求 y 值的每一个子总体都满足正态分布。于是综合回归分析中估计和检验两方面的需要，对总体的数据结构有如下的假定

$$\begin{aligned}
y_1&=\alpha+\beta x_1+\varepsilon_1\\
y_2&=\alpha+\beta x_2+\varepsilon_2\\
&\vdots\qquad\quad\vdots\qquad\vdots\\
y_n&=\alpha+\beta x_n+\varepsilon_n
\end{aligned}$$

其中 $\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n$ 是随机变量，它们相互独立，且都服从相同的正态分布 $N(0,\sigma^2)$(σ^2 未知)。

二、回归方程的检验

前面介绍了用最小二乘法求直线回归的方法，它是基于线性回归模型的基本假定进行的。因此在配置回归直线之前，必须对总体变量间是否存在线性相关关系进行检验。否则，对于不存在线性关系的总体，配置回归直线毫无意义。为此，下面要讨论回归方程的检验。

(一) 检验的原假设

根据本节第一部分的讨论，所谓总体变量 x 和变量 y 存在线性关系，指的是存在关系式

$$E(y_i)=\alpha+\beta x_i \tag{1-19}$$

因此，对于总体线性检验的假设可写成如下的形式

$$H_0:\beta=0$$

$$H_1:\beta\neq0$$

有了假设，下面将根据平方和分解求出检验所需的统计量。