YIXUETONGJIXUE
yunyong sanxing lilun jinxing xiandai huigui fenxi

主编　胡良平

医学统计学

运用三型理论进行现代回归分析

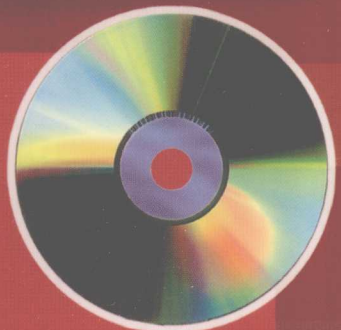
很多与统计学有关的实际问题，均以“表现型”的面貌呈现在人们的面前，表现型常常带有假象，直接依据表现型去盲目套用传统的统计学教科书上的“标准型”，十有八九会出错，因此，要想正确运用统计学，必须弄清反映“表现型”本质的“原型”，将“原型”正确转变成“标准型”后，就很少会出错。这样一种由胡良平创立的可有效解决问题的新理论，被称为“统计学三型理论”。此理论可使统计学思想付诸实施。

统计学三型理论

光盘中SAS程序能方便快捷地实现现代回归分析

光盘中的SAS程序包括：多重线性回归分析、岭回归分析、各种复杂曲线回归分析、主成分回归分析、Poisson回归分析、Probit回归分析、负二项回归分析、配对和非配对设计定性资料多重logistic回归分析、对数线性模型分析、生存分析和时间序列分析。

用户只需用自己的资料替换掉例子中的数据，按一下发送键，就可轻松实现复杂深奥的各种现代回归分析。功能强大的SAS软件具有化繁为简、实用方便之效果。



内附光盘

人民军医出版社
PEOPLE'S MILITARY MEDICAL PRESS

医学统计学

应用三维图谱进行现代回归分析

医学统计学

——运用三型理论进行现代回归分析

YIXUE TONGJIXUE——
YUNYONG SANXING LILUN JINXING XIANDAI HUIGUI FENXI

主　编　胡良平

副主编　高　辉　李长平　葛　毅　胡纯严

编　者　（以姓氏笔画为序）

王　霄　中山医科大学
毛宗福　武汉大学公共卫生学院
刘明华　中山医科大学
刘惠刚　首都医科大学
李子建　济南军区疾病预防控制中心
李长平　天津医科大学
杨业春　中山医科大学
余红梅　山西医科大学
张　熙　中山医科大学
张岩波　山西医科大学
张晋昕　中山医科大学
周诗国　军事医学科学院
胡良平　军事医学科学院
胡纯严　军事医学科学院
柳伟伟　军事医学科学院
高　辉　军事医学科学院
郭东星　山西医科大学
崔　丹　武汉大学公共卫生学院
葛　毅　后勤指挥学院
薛允莲　中山医科大学

人民军医出版社
PEOPLE'S MILITARY MEDICAL PRESS

北　京

图书在版编目(CIP)数据

医学统计学——运用三型理论进行现代回归分析/胡良平主编. —北京:人民军医出版社,2010.7

ISBN 978-7-5091-3976-9

Ⅰ.①医… Ⅱ.①湖… Ⅲ.①医学统计 Ⅳ.①R195.1

中国版本图书馆 CIP 数据核字(2010)第 124235 号

策划编辑:于 岚 **文字编辑**:黄维佳 **责任审读**:刘 平

出 版 人:齐学进

出版发行:人民军医出版社 **经销**:新华书店

通信地址:北京市 100036 信箱 188 分箱 **邮编**:100036

质量反馈电话:(010)51927290;(010)51927283

邮购电话:(010)51927252

策划编辑电话:(010)51927300—8119

网址:www.pmmp.com.cn

印、装:中国农业出版社印刷厂

开本:787mm×1092mm 1/16

印张:18.75 **字数**:452 千字

版、印次:2010 年 7 月第 1 版第 1 次印刷

印数:0001～3000

定价(含光盘):75.00 元

内容提要

本书介绍了现代回归分析方法中的大部分内容，包括多重线性回归分析、岭回归分析、各种复杂曲线回归分析、主成分回归分析、Poisson 回归分析、Probit 回归分析、负二项回归分析、配对和非配对设计定性资料多重 logistic 回归分析、对数线性模型分析、生存分析和时间序列分析。本书既适合未学过 SAS 软件的新用户，也适合多年应用 SAS 软件的老用户；既适合未学过统计学的新读者，又适合从事统计学科研、教学多年的老读者，可供需要运用 SAS 软件解决各种现代回归分析问题的研究生、博士生，以及科研管理人员、临床医师和期刊编辑学习使用。

前　言

笔者在大量的统计咨询、项目评审和为杂志审稿中看到，许多科技人员、管理人员、临床医生、杂志编辑、本科生、研究生和博士生很想快速学会使用 SAS 软件，并将其正确用于自己的科研课题和学位论文的科研设计和统计分析之中，但他们中的很多人都比较盲目，即便他们很有毅力从头至尾学完一本本 SAS 统计分析，仍不知道 SAS 究竟是什么，不知如何使用 SAS 解决常见的统计学问题，更谈不上用 SAS 的高级编程技术去完成自己具有创新性的科研课题；也不知道正确运用统计学的要领是什么，不知道统计学内容的合理划分和正确选用及 SAS 的巧妙实现方法，因而几乎是一用就出错。他们耗费了大量宝贵的时间和精力，学到了许多零零星星、支离破碎、一知半解的知识和技能，就如同一个饥饿的人看到“墙上画着令人垂涎欲滴的烧饼”一样，只能解解“眼馋”，而无法实现“充饥”之目的。每当笔者看到许多人使用 SAS 软件和统计分析方法时所承受的压力、痛苦和无奈，就想把自己通过 20 多年总结出来的知识和经验全部奉献给他们。但一个人的知识、时间和精力都是十分有限的，故笔者诚邀全国多所著名大学从事生物医学统计学的专家共同撰写这部专著，力求以实际工作者为本，集众英才之智慧，献统计学之精品，解实际工作之烦忧，为我国科技质量的提高和科技事业的发展提供优质的技术服务和指导。

本书着重介绍与回归分析有关的内容，回归分析的内容非常丰富。由于自变量和因变量的性质、它们之间的相互关系、分布规律等不同，所以就产生了很多不同的回归分析方法。本书在三型理论指导下，使花样繁多的回归分析变得条理分明，易于理解和接受。其内容包括多重线性回归分析、岭回归分析、各种复杂曲线回归分析、主成分回归分析、Poisson 回归分析、Probit 回归分析、负二项回归分析、非配对设计定性资料的 logistic 回归分析、配对设计定性资料的 logistic 回归分析，还包括通常不纳入回归分析的特殊方法，如对数线性模型分析、生存分析和时间序列分析。书中不仅介绍这些回归分析方法的计算原理，还详细介绍如何用 SAS 软件实现计算及结果解释。为读者解决各种回归分析问题提供了翔实的理论和技术支持。

在本书出版过程中，得到了中山医科大学、山西医科大学、武汉大学公共卫生学院、天津医科大学、首都医科大学、济南军区疾病预防控制中心、后勤指挥学院、军事医学科学院有关教授、副教授和青年学者的大力帮助，他们是张晋昕、张熙、杨业春、薛允莲、王霄、刘明华、张岩波、余红梅、郭东星、毛宗福、崔丹、李长平、刘惠刚、李子建、葛毅、高辉、周诗国、柳伟伟、胡纯严等。在

本书即将出版之际，谨向他们表示真挚的感谢！还要感谢所有为本书默默奉献付出辛劳的人们，特别是本室的在读硕士研究生郭晋、毛玮、陶丽新、王琪、贾元杰、鲍晓蕾和关雪，正是由于他们的积极参与、不懈努力和真心奉献，才使这部专著能够问世。

最后还要提及的是，本书的参编者胡纯严为本书编配了方便快捷调用 SAS 程序的软件，名为“SAS PAL”，为提高读者调用 SAS 程序的准确性和效率贡献了很大的力量。

为了方便广大读者联系开展科研课题合作研究、求助科研设计与数据的统计分析等，特提供笔者的电子邮箱：LPHU812@SINA.COM。

由于参与编写工作的人员多、统稿的工作量大，对书中存在的不妥或疏漏之处，恳请广大读者不吝赐教，以便再版时修订。

胡良平

于北京军事医学科学院生物医学统计学咨询中心

目　录

第1章 概 述

本章将着重介绍三型理论的精神实质，以及如何将此理论与现代回归分析方法建立联系的构想，从而为现代回归分析方法的合理分类提供了依据。以多重回归分析方法为例，很多实际工作者觉得比较复杂，因为需要考虑的情况太多，很难把握，但运用三型理论来解读，却显得十分正常而极易于理解和掌握。读完本章后，读者会觉得多重回归分析方法并不像原先想象的那样复杂，而是有规律可循的。

1.1 何为三型理论

任何具体问题一般都存在3种表现形态，即表现型、原型和标准型，一旦掌握了与具体问题对应的这三种形态，再有的放矢地去解决问题，通常问题便能得到圆满解决，也就可以做到少犯或不犯错误。笔者把这样一种有利于透过事物现象看清其本质的解决问题的思维方法称为三型理论。

笔者于2005年首次提出此理论时，仅针对统计学问题。其实，此理论还可用于其他各领域之中，包括社会学、人文学、心理学、环境学、医学、经济学等。下面举几个实例，以便直观感受三型理论的价值和魅力。

【例1-1】 一个经济学方面的例子。某作坊主雇了5个工人，每人工作8h，每小时生产4个瓷罐，共生产了160个瓷罐，每个瓷罐10元，总共获得1 600元。其中原料和机器磨损消耗支出1 000元，支付给工人的工资600元，即每个工人工资120元。结果作坊主没有获得任何利润。由于生产工具的改进，生产效率提高了，工人仍然工作8h，但每小时生产5个瓷罐，结果生产了200个瓷罐。按原来的价格出售，作坊主共获得2 000元。其中原料和机器磨损消耗支出1 200元，支付给工人的工资仍为600元，即每个工人工资120元。有人认为此作坊主对工人很公道、很善良。请问：这样的认识错在哪里？

分析与解答：与这个问题对应的三型如下。

(1)表现型：工人为作坊主劳动，作坊主支付了工资，属于等价交换，不存在剥削。

(2)原型：作坊主付给工人的工资只是其出卖的劳动力创造价值的一部分，剩余价值被作坊主无偿占有了。

(3)标准型：商品总价值包括生产成本、工人工资和剩余价值三部分，剩余价值来源于资本家对工人劳动力的无偿剥削。

从表面上看，工人工作8h和原来一样，因此作坊主也支付了相应的工资。这似乎是很公道、很合理的。其实，作坊主获得的200元却是工人创造的，因为原料不会创造剩余价值，只有劳动力才能创造价值，这200元就是雇主对工人的剥削。

【例1-2】 一个心理学方面的例子。有些儿童特别好动，无论大人怎样制止他们，几乎无济于事。有人很肯定地认为是其家长溺爱的结果，请利用三型理论来诠释这一现象。

分析与解答：儿童特别好动，与一般的顽皮是有本质区别的。与此问题对应的三型如下。

(1)表现型:学龄及学龄前的孩子看起来都是活泼好动和天真调皮的,但是依据在不同的环境和对待不同的事情做出的截然不同的两种反应,可以将一些孩子归于"多动"之列。这些"多动"的儿童在生活学习中更容易表现出"好动"的迹象。比如,这些儿童无目的性的活动过多。在上课时,调皮的孩子会为了故意引起别人注意,或为了好玩有趣而偶尔做小动作;但特别好动的儿童就不同了,他们好像不受意识支配似的,不停活动,如毫无目的地摇桌子、晃椅子,即使受到老师的提醒、制止或批评,还会马上又不由自主地重复原来的小动作,或改换为乱翻书、东张西望、左顾右盼,以及咬铅笔、切橡皮、招惹邻座的同学。他们的自控能力差,玩得高兴时又喊又叫、又跑又跳、手舞足蹈、莫名兴奋、情不自禁、得意忘形,对大人的厌烦表情和制止行为不能产生约束性心理反应;受到强制性约束的时候,不是安静下来,而是表现出闹脾气、不高兴、发泄沮丧情绪,采取敌意和对抗性行为;令大人既厌烦又无可奈何,令同伴讨厌、害怕和敬而远之,因此不合群,得不到别人尊重。此外,他们伴随运动协调性差,并有知觉、语言、记忆的障碍。如辨认符号和声音费时很久,搞不清含义,语言水平低于同龄儿,记事慢而忘事快等。他们的表现与其他孩子既有相似之处,但又有过之而无不及。家长很容易就误认为顽皮好动是孩子的天性,从而忽视了孩子的身心健康。

(2)原型:这些孩子好动,难以受自我控制的表现,实际是一种常见的儿童心理疾病——儿童多动症的症状,这类孩子智力一般为正常,但存在与实际年龄不相符合的注意力涣散、活动过多、冲动任性、自控能力差的特点。如果不加以纠正,会影响孩子的学习。据调查,多动症的发病率为3%左右,男孩为女孩的4~9倍。儿童多动症的发病机制尚不明,但是国内外学者认为本病是由多种因素引起的,如遗传因素、轻微脑损伤、脑发育不成熟、工业污染、营养因素、家庭和环境因素、药物因素等。

(3)标准型:根据不同的致病因素,对患有"多动症"的儿童进行治疗,其方法有心理治疗和药物治疗。心理治疗是对多动症儿童心理缺陷进行治疗,家长、教师及医务人员启发、诱导、教育儿童以正确的态度对待学习,养成良好的生活、学习习惯,培养、锻炼自我控制能力,此类行为治疗是心理治疗的重要内容之一。药物治疗虽不能代替心理治疗,但没有药物治疗同样达不到治疗目的。药物治疗能为心理治疗提供有利条件,促进儿童集中注意力,加强自我控制能力,专心学习,但要使学习成绩提高、不良行为得到纠正,还需要长期耐心的教育,包括文化知识教育。治疗用药物主要为中枢神经兴奋药,如苯丙胺、哌甲酯、咖啡因、三环类;抗忧郁药,如丙米嗪;抗精神病药,如氯丙嗪等。在我国,中医工作者采用中医中药、针灸、推拿等方法配合治疗儿童多动症,取得了较好效果。目前普遍认为,儿童多动症通过心理、药物治疗及中医中药治疗等,能取得较好的治疗效果。

【例1-3】 第一个医学上的例子。某人最近一段时间以来,不管一天喝多少水,仍然感到口很渴,滴落在地上的尿液会招致一些虫子和蚂蚁。这种现象说明了什么?此人是否患有某种疾病?请问:与此问题对应的三型是什么?

分析与解答:这种现象意味着此人尿液中可能含有某种气味的东西,虫子和蚂蚁想来大饱口福。同时,也提示此人可能患了某种疾病。与此问题对应的三型如下。

(1)表现型:此人多饮,可能也会伴有多尿、多食、消瘦、疲乏,其尿液中可能含有葡萄糖(正常人尿液中不应含有葡萄糖)。

(2)原型:胰岛素缺乏和细胞受体对胰岛素不敏感导致机体糖代谢紊乱,血糖浓度过高,影响肾脏对血糖的滤过作用,导致部分葡萄糖进入尿液中。

(3)标准型:糖尿病的医学诊断标准如下。

①非同日两次空腹血糖>7 mmol/L,其中空腹的定义为禁食8h以上。

②餐后2h血糖>11.1mmol/L。

③具有糖尿病症状并且随机血糖>11.1mmol/L。

具有以上三点特征的人在医学上被定义为患了糖尿病。

【例1-4】 第二个医学上的例子。某人最近一段时间以来,总感到胸部不舒服,其表现型为经常出现咳嗽、气喘、发热、乏力等症状,请问:与此人对应的原型和标准型最可能是什么?

分析与解答:此人很可能肺部出现了问题,其最可能的原型和标准型如下。

(1)原型:此人很可能患了肺炎。

(2)标准型:通过临床上一系列检查,如X线片、痰培养等最终确诊患者的气管和(或)支气管被细菌、病毒、支原体等病原微生物感染而引起肺炎,并具体分型,如大叶性肺炎、支原体肺炎等。

【例1-5】 一个关于调查资料统计分析的实例。某人在北京郊区做了一项关于狗咬伤人情况的调查研究。将被调查者分成若干年龄段,求出每个年龄段(其组中值记为X)上的人被狗咬伤的比例Y,用统计学方法建立Y随X变化的直线回归方程。并声称,可以根据该地区任何一个被调查者的年龄来预测其将被狗咬伤的概率。请运用三型理论来揭示对这项调查资料的统计分析纯粹是在玩弄数字游戏。

分析与解答:与此问题对应的三型如下。

(1)表现型:问题中似乎仅涉及两个变量,其中X(各年龄段上的组中值)为原因,而Y(各年龄段上被狗咬伤的比例,对总体而言应该称概率)为结果,用统计学方法研究Y随X变化的规律时,最简单的方法就是用直线回归方程来描述,所以,此研究者的做法似乎是合情合理的。

(2)原型:该地区的居民是否被狗咬伤与该地区人们养狗的种类、数量、人对狗管理的水平、人们在户外活动的频繁程度、人在受到狗攻击时反应速度和采取应对措施的能力等都有关系。X(各年龄段上的组中值)仅是"人在受到狗攻击时反应速度和采取应对措施的能力"的一个简单度量,它根本不能代表人可能被狗咬伤的全部原因。

(3)标准型:研究一个实际问题中的原因与结果之间的关系不应"盲目"。第一,应依据基本常识和专业知识,尽可能找全对结果可能有影响的重要因素。第二,应调查足够大的样本,注意调查过程中的质量控制,尽可能反映出客观真实情况。第三,应先采取探索性分析方法初步了解各原因变量与结果变量之间可能存在的关系(无关系、线性或非线性关系)。第四,基于探索性分析结果、原因变量的个数和结果变量的性质(定量的、定性的)等,再采取在统计学上最为科学全面的处理方法来研究结果变量随原因变量变化的规律。第五,应再通过调查研究,进一步检查和考核所得出的规律的正确性。

实例不胜枚举,读者可根据三型理论的精髓去分析你关心或感兴趣的事物与现象。

1.2 何为多重回归分析方法

在统计学书籍中,人们经常会看到"多元回归分析"和"多重回归分析"等字样,有时它们之间有区别,有时似乎又代表同一概念。在本书中,若考察的原因变量的个数≥2,而结果变量仅有一个,并且希望研究一个结果变量随多个原因变量变化的依赖关系时,称为多重回归分析;若同时考察的原因变量的个数≥1,而结果变量的个数≥2,并且希望研究多个结果变量随一个或多个原因变量变化的依赖关系时,称为多元回归分析。两者之间的本质区别在于同时考察的结果变量的个数。本书仅讨论多重回归分析问题。

多重回归分析有很多种，其主要区别在于表达结果变量（即因变量）随原因变量（即自变量）依赖关系的方程式中，两类变量的性质、分布特点和表达式中系数的属性。

（1）当表达式中系数的属性为常数（计算之前为待定常数）且模型（对总体而言，若对样本而言，常称为方程）的误差项服从正态分布（人们常简单理解成因变量服从正态分布）时，此时称其为一般线性模型。即便对某些自变量做了对数变换、指数变换或平方根变换，它们前面的系数仍是常数，因此也属于一般线性模型。

（2）当表达式中系数的属性为常数，但模型的误差项不服从正态分布时，通常称为广义线性模型。在拟合方程时，一般都需要对因变量进行适当变换，例如，当因变量为二值变量时，常对其取 logit 变换或 Probit 变换；当因变量服从 Poisson 分布时，常对其取对数变换。

（3）当表达式中系数的属性为某些自变量的函数时，根据因变量的分布情况，又可分为一般非线性模型和广义非线性模型。另外，还可以根据自变量所产生的效应是固定的还是随机的，从而产生出固定效应模型、随机效应模型和混合效应模型。也就是说，关于模型的性质，可有“一般与广义”“线性与非线性”和“固定效应与随机效应”等限定词，当这些限定词都确定后，此模型的性质才被完全确定下来。如广义非线性混合效应模型，就相当复杂了。

（4）特殊情形下的多重回归模型，如描述生存资料的 COX 模型和参数模型，由于资料中含有删失数据，有时还有缺失值，估计模型中参数的方法与常规统计分析方法有所不同。

（5）当因变量之间不满足独立性，即因变量之间具有不同程度的自相关性时，此时的自变量通常是“时间 t”，而将各时间点上的因变量 Y 表示为 t 的函数形式，建立不同时间点上 $Y(t)$ 之间相互依赖的变化关系，此类回归分析称为“时间序列分析”，一般不列入回归分析研究领域之中，直接命名为时间序列分析，但本质上仍然应归属于回归分析。若除了时间因素作为自变量以外，还有其他影响因变量取值的自变量，则此时的时间序列分析就更为复杂了。

总而言之，回归分析通常是研究因变量随自变量变化的依赖关系的，包括研究不同时间点上因变量与其他时间点上因变量之间的关系（即时间序列分析），还包括时间的函数随一系列自变量变化的依赖关系，如 COX 模型中的危险率函数 $h(t)$ 等。

1.3 如何用三型理论来指导多重回归分析方法的合理选用

由三型理论可知，调查或实验得到的数据只是所关心的变量（包括自变量和因变量）在受试对象身上的具体表现，其本质（即问题的原型）可能并没有被完全包含在内，也就是说，研究者可能漏掉了某些对所关心的因变量具有重要影响的自变量，此时，无论你采取什么高级统计分析方法，也很难揭示它们之间的真正依赖关系。换句话说，比合理选择回归分析方法更为重要的是研究设计，在设计中，千万不能遗漏重要的自变量，因变量也应定义和测量准确，不仅要有足够大的样本量，还要特别重视样本的代表性，更要注意受试对象的同质性。应根据基本常识和专业知识，确定变量之间在专业上应有一定的联系，不应把与因变量毫不相干的自变量考虑在内。

在选择具体的回归分析方法之前，最好对每个自变量与因变量之间的关系做一些探索性分析，来决定是否需要对自变量或因变量采取某些变量变换方法，以改善它们之间的关系，有利于满足或接近所选定的回归模型的要求。

至于何时选择一般线性固定效应（或随机效应或混合效应）模型、何时选择一般非线性固定效应（或随机效应或混合效应）模型、何时选择广义线性固定效应（或随机效应或混合效应）模型、何时选择广义非线性固定效应（或随机效应或混合效应）模型，则取决于探索性分析的结果、基本常识和专业知识的掌握及对其运用的熟练程度。自变量的效应是固定的还是随机的，就看

其水平取值是固定的(如性别、血型等)还是从大量水平中随机选取的;模型是线性的还是非线性的合适,取决于自变量与因变量之间的真实关系,比如,因变量与某些自变量是二项型或三项型指数曲线关系,可能就需要选取非线性模型了;选择一般还是广义模型,关键取决于因变量的分布规律以及是否需要对因变量做某种变量变换。

1.4 在使用多重回归分析方法时常犯哪些错误

实际工作中使用得最多的多重线性回归分析方法是以下两种。第一,一般多重线性固定效应模型,常简称为多重线性回归分析。第二,广义线性固定效应模型中的一种,即多重 logistic 回归分析。

人们在使用时常犯的错误有:对因变量有重要影响的自变量往往考虑不周全,很难达到所预期的结果;受试对象的同质性不符合事先规定的研究目的,比如说,目的是研究正常人某指标随某些自变量变化的依赖关系,而受试者中却夹杂着一些有严重疾病的人;筛选变量的策略错误,先进行单变量分析,将 $P<0.05$ 的那些自变量纳入多变量分析;仅依据一种筛选多变量的方法给出的结果下结论,例如仅用前进法、后退法或逐步法之一;对自变量与因变量之间的关系大致属于什么情况未进行探索性分析,一律视为简单线性关系;最终的多重回归方程中仍包含若干个无统计学意义的自变量,即采用不筛选自变量的多重回归分析方法建立回归方程。

下面举两个实例,展示人们在进行回归分析时常犯错误的具体细节。

【例 1-6】 某作者在《174 例原发性肾小球疾病患者血瘀证与临床及病理的相关性分析》一文中运用了多重回归分析方法。原文欲探讨原发性肾小球疾病患者血瘀证程度与临床及其肾脏病理类型之间的关系。采用现场调查,对符合纳入标准的 174 例患者于肾活检前 3d 内行血瘀证和中医虚损证候评分。分析年龄、病程、中医虚损证候、24h 尿蛋白定量(Upro)、高血压及血压控制情况、肾小球滤过率(GFR)、尿酸(UA)、三酰甘油(TG)、胆固醇(CHO)、血红蛋白(Hb)、血浆白蛋白(ALB)等临床指标及不同病理类型与血瘀证积分的关系。计数资料做 Pearson Correlation 相关分析,对相关系数有统计学意义的观察指标做多重逐步回归分析,以确定其在建立回归方程中的必要性;两组间比较采用独立样本的 t 检验、多组间比较采用单因素方差分析。各项临床指标分别与血瘀证积分做 Pearson Correlation 相关分析,结果见表 1-1。本组 174 例患者的血瘀证积分分别与 Upro、CHO、TG、ALB 和虚损证积分具有显著相关性(均 $P<0.001$),与年龄、病程、GFR、UA、Hb 无相关关系($P>0.05$)。对相关系数有统计意义($P<0.05$)的观察指标再行多重逐步回归分析,发现 Upro、TG 和虚损证积分对建立回归方程有统计学意义(均 $P<0.01$)。

表 1-1 174 例患者血瘀证积分与各项临床指标的相关分析

项目	r	P
Upro	0.307	0.000
TG	0.267	0.000
CHO	0.289	0.000
ALB	−0.243	0.001
虚损证积分	0.315	0.000
年龄	−0.060	0.917
病程	−0.064	0.403
UA	0.101	0.184
Hb	−0.084	0.268
GFR	−0.062	0.413

(1)专业结论:血瘀证积分与 Upro 、TG 水平及虚损证积分显著相关,伴有肾小球硬化的局灶增生性肾小球肾炎其血瘀证积分较高,提示血瘀证在一定程度上可以反映原发性肾小球疾病的肾脏慢性化病变,是中医证候中影响肾脏病进展的危险因素之一。

(2)对差错的辨析与释疑:原作者先通过单因素分析,找出 $P<0.05$ 的因素,然后将这些所谓"有统计学意义"的因素(自变量)放在一起,进行不筛选变量的多重 logistic 回归分析,根据计算的结果做出统计学或专业结论。此法的错误之处在于它可能会漏掉某些重要的自变量,这些变量的特点是它们单独对因变量的贡献较小,但是,一旦它们与某些变量同时出现在回归模型中时可能会发挥很大的作用。

(3)筛选变量的正确策略:当自变量不是特别多时,应尽可能使全部的自变量都有机会参与变量筛选过程;当自变量特别多且样本含量又不大时(一般要求样本含量为自变量个数的 5～10 倍或以上,这样回归分析的结果比较稳定),可以将单个自变量筛选中"$P>0.05$"的那些自变量暂时不参与自变量的筛选过程,必要时可从其中随机抽取几个与拟参与自变量筛选的那些自变量一并考察。筛选变量的方法很多,最好用 3～5 种筛选变量的方法处理同一个多重回归分析资料,若计算的结果十分稳定,则可以做出统计学和专业结论,否则,宜考虑采用最优回归子集法筛选变量,并结合最优子集回归模型的评价标准和专业知识,确定 1～2 个最为理想的回归模型。

【例 1-7】 某作者在《正常幼儿语法发育的影响因素研究》一文中运用了单因素和多因素分析方法,研究者欲了解影响幼儿语法发育的有关因素,从而为促进幼儿语言发育提供依据。研究中采用现况定量研究方法,用多阶段分层不等比例抽样方法在北京 4 个城区抽取样本。用"中文早期语言与沟通发展量表"及个人背景问卷,对北京城区 1 056 名 16～30 个月正常幼儿母亲或日间照顾人进行面对面问卷调查。运用 Z 评分法对幼儿语法粗分进行标准化,再用单因素和多重回归分析方法探讨影响幼儿语法发育的因素。为便于分析儿童语言发育的影响因素,特将儿童语法得分粗分按不同月龄、性别进行标准化,分别获得他们的 Z 评分,以消除进行因素分析时月龄和性别的影响。将可能影响幼儿语法发育的 20 个因素分别与幼儿语法得分 Z 评分进行单因素分析。结果显示,母亲或日间照顾人受教育程度、父亲月收入、父母职业、儿童性格外向、气质与情绪随和、儿童身高、居住地区人均收入等因素与幼儿语法得分 Z 评分呈正相关,而儿童开始走路和说话月龄、父亲年龄等因素与幼儿语法得分 Z 评分呈负相关(表 1-2)。由于单因素分析不能消除因素间的混杂作用,故需运用多因素分析方法。将单因素分析有统计学意义的 13 个变量,用后退法(剔除变量的概率为 0.10)与因变量做多重回归分析。共有 4 个变量进入最终模型,幼儿开始说话月龄、父亲月收入和父亲受教育程度的回归系数显著性检验概率小于 0.05(表 1-3)。由各变量标准化回归系数可见,这 4 个变量对语法得分的影响由大到小依次为幼儿开始说话月龄＞父亲月收入＞父亲受教育程度＞幼儿性格。幼儿开始说话月龄与幼儿语法得分 Z 评分呈负相关,其余 3 个因素均与幼儿语法得分 Z 评分呈正相关,即在分别固定了其他 3 个因素后,父亲月收入越高、父亲受教育程度越高、幼儿性格越外向,越利于幼儿语法得分 Z 评分的增加;幼儿开始说话月龄越晚,越不利于幼儿语法得分 Z 评分的增加($r=0.229$,$F=14.487$,$P=0.000$)。

表 1-2 影响幼儿语法发育的因素

因素	r	P
母亲受教育程度	0.092	<0.01
父亲受教育程度	0.104	<0.01
日间照顾人受教育程度	0.070	<0.05
父亲月收入	0.121	<0.001
父亲年龄	−0.100	<0.01
母亲职业	0.078	<0.05
父亲职业	0.070	<0.05
儿童身高	0.083	<0.05
居住地区	0.089	<0.01
儿童性格外向	0.061	<0.05
儿童情绪与气质随和	0.061	<0.05
开始走路年龄	−0.068	<0.05
开始说话年龄	−0.176	<0.001

表 1-3 影响幼儿语法 Z 评分的多重逐步回归分析

变量	回归系数	标准误	标准化回归系数	t	P
截距	−0.432	0.225	—	−1.917	>0.05
父亲受教育程度	0.230	0.094	0.075	2.449	<0.05
幼儿开始说话年龄	−0.326	0.060	−0.165	−5.470	<0.01
父亲月收入	0.226	0.070	0.099	3.206	<0.01
幼儿性格外向	0.117	0.063	0.057	1.878	>0.05

(1)调查结果:调查地区 16～30 个月龄正常幼儿平均语法表达结构得分由 16 个月时的 6 分增加到 30 个月时的 84 分,占总分的 83%。单因素分析和多重回归分析结果显示,父亲受教育程度、父亲月收入、幼儿性格外向是幼儿语法发育的有利因素;幼儿开始说话月龄与幼儿语法发育得分 Z 评分呈负相关。

(2)结论:家庭社会经济状况是影响幼儿语法发育的重要因素。保健人员应重视幼儿语言发育,教育幼儿父母注意与儿童沟通,加强对儿童的早期教育,为儿童创造良好的早期语言环境。

(3)对差错的辨析与释疑:在本例中,原作者犯了与【例 1-6】类似的错误,不再赘述。从表 1-2 最后一列可看出,似乎所列出的 13 个自变量与结果变量(Z 评分)之间都有很密切的关系;但从表 1-2 第 2 列可知,13 个直线相关系数的绝对值都很小,最大的一个绝对值为 0.176,其对结果变量的贡献仅为 $r^2=3.1\%$。由此可见,表 1-2 中的结果主要因 $n=1\,056$ 足够大而得出了统计学上有意义的结果,但其实几乎没有任何实际价值。表 1-3 中包含 2 个无统计学意义的项,表明依据此多重线性回归方程下结论是不可信的。

1.5 本章小结

本章概述了三型理论的精髓,通过 5 个涉及多个领域的实例,诠释了如何用三型理论解决不同领域中问题的技巧;从狭义和广义角度分别解释了何为多重回归分析;进而以三型理论为