

高 等 学 校 教 材

SPSS

统计分析高级教程 (第3版)

张文彤 董 伟 编著

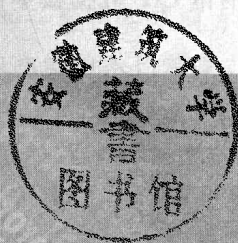
高等教育出版社

高等学校教材

SPSS

统计分析高级教程 (第3版)

张文彤 董伟 编著



高等教育出版社·北京

内容简介

本书全面、系统地介绍了各种多变量统计模型、多元统计分析模型、智能统计分析方法的原理和软件实现,是一本使用 SPSS 进行高级统计分析的实用性很强的指导书和参考书。本书共分 4 个部分,分别是一般线性模型、线性混合模型与广义线性模型,回归模型,多元统计分析方法,以及其他统计分析方法。本书基于 IBM SPSS Statistics 24 中文版,并结合作者多年的统计分析实战经验和 SPSS 行业应用经验,侧重对统计新方法、新观点的讲解,在保证统计理论严谨权威的同时注重叙述的浅显易懂,使本书更加易学易用。

本书可作为高等学校本科生和研究生统计学相关课程教材,也可作为市场营销、金融、财务、人力资源管理等行业中需要做数据分析的人士,或从事咨询、研究、分析等工作的人士的参考书。

图书在版编目(CIP)数据

SPSS 统计分析高级教程 / 张文彤, 董伟编著. --3
版. --北京: 高等教育出版社, 2018. 1
ISBN 978-7-04-049033-6

I. ①S… II. ①张… ②董… III. ①统计分析-统计
程序-高等学校-教材 IV. ①C819

中国版本图书馆 CIP 数据核字(2017)第 302226-号

SPSS Tongji Fenxi Gaoji Jiaocheng

策划编辑 刘艳 责任编辑 刘艳 封面设计 于文燕 版式设计 马敬茹
插图绘制 于博 责任校对 刘莉 责任印制 赵义民

出版发行	高等教育出版社	网 址	http://www.hep.edu.cn
社 址	北京市西城区德外大街 4 号		http://www.hep.com.cn
邮政编码	100120	网上订购	http://www.hepmall.com.cn
印 刷	北京市联华印刷厂		http://www.hepmall.com
开 本	787 mm×1092 mm 1/16		http://www.hepmall.cn
印 张	33.75	版 次	2004 年 9 月第 1 版
字 数	830 千字		2018 年 1 月第 3 版
购书热线	010-58581118	印 次	2018 年 1 月第 1 次印刷
咨询电话	400-810-0598	定 价	65.00 元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换

版权所有 侵权必究

物料号 49033-00

前 言

本书自 2004 年第 1 版出版以来,受到了广大读者的欢迎,被国内数百所高校选为本科生或研究生相关课程教材,在此感谢广大读者的支持与厚爱。

本书第 2 版出版于 2013 年,SPSS 在这几年间已经升级了 4 个版本,而且最新的版本易用性更强,软件功能更丰富,因此需要对全书内容进行有针对性的修订。在第 2 版的基础上,结合 SPSS 版本的更新和读者的反馈,本版以 IBM SPSS Statistics 24 中文版为准,对内容做了如下调整。

1. 内容进一步拓展

本书第 2 版全面覆盖了 SPSS 自身提供的各种高级统计分析功能,但 SPSS 提供的主要是成熟且常用的统计模型,许多独特的新模型并未提供。实际上,SPSS 通过 Python 插件和 R 插件的方式提供了这些模型。为进一步拓展读者的知识面,除介绍 SPSS 新版本直接提供的新模型外,本书还对一些比较重要的用插件方式提供的模型进行了介绍,包括分位数回归、弗斯 Logistic 回归、潜类别分析、支持向量机、随机森林、项响应模型等,并在附录中介绍了相应插件的安装方法,以帮助读者及时跟进统计分析领域的最新进展,并在工作中充分发挥 SPSS 的作用。

2. 更加浅显易懂

本书涉及的统计模型都比较复杂,为降低学习难度,确保读者能够掌握相应的统计分析方法,本书在第 2 版的基础上进一步减少了案例数量。通过对同一案例在不同方法框架下的分析结果进行比较,再辅以对统计理论深入浅出的讲解,降低了初学者的入门难度,最大限度地优化了学习效果,有利于读者学以致用。

本书由张文彤和董伟共同编写,可作为高等学校各专业本科生和研究生的统计学相关课程教材,也可作为市场营销、金融、财务、人力资源管理等行业中需要进行数据分析的人士,或从事咨询、研究、分析等工作的人士的参考书。学习本书的读者需要具备统计分析及 SPSS 操作的基础知识,需要补充这部分知识的读者可以先学习《SPSS 统计分析基础教程》(第 3 版)。对于希望进一步提升统计分析和数据挖掘实战能力的读者,则可以在学习完本书后继续阅读作者的实战案例精粹系列书籍,以进一步提高实战经验。为便于读者交流和使用这套书,读者可以关注微信公众号:统计之星。本书的案例数据文件、拓展资料等可到本书配套的数字课程网站和“医学统计之星”网站上下载。

希望广大读者一如既往地踊跃提出自己的宝贵意见和建议,使得本书再次改版时能够更上一层楼,更好地满足大家的学习和工作需求。

作 者

2017 年 10 月

目 录

第一部分 一般线性模型、线性混合模型与广义线性模型

第 1 章 方差分析模型	3
1.1 模型简介	3
1.1.1 模型入门	3
1.1.2 常用术语	5
1.1.3 适用条件	7
1.2 案例:胶合板磨损深度的比较	8
1.2.1 操作说明	8
1.2.2 结果解释	9
1.2.3 模型参数的估计值	11
1.2.4 两两比较	12
1.2.5 其他常用选项	14
1.3 两因素方差分析模型	15
1.3.1 案例:超市规模、货架位置 与销售量的关系	15
1.3.2 边际平均值与轮廓图	19
1.3.3 拟合劣度检验	21
1.4 因素各水平间的精细比较	22
1.4.1 POSTHOC 子句	22
1.4.2 EMMEANS 子句	22
1.4.3 LMATRIX 和 KMATRIX 子句	23
1.4.4 CONTRAST 子句	25
1.5 方差分析模型进阶	25
1.5.1 含随机因子的方差 分析模型	25
1.5.2 自定义检验使用的 误差项	27
1.5.3 4 类方差分解方法	28
思考与练习	29
参考文献	29
第 2 章 常用的实验设计分析方法	30
2.1 仅研究主效应的实验设计方案	31
2.1.1 完全随机设计	31
2.1.2 随机区组设计	32
2.1.3 交叉设计	32
2.1.4 拉丁方设计	34
2.2 考虑交互作用的实验设计方案	36
2.2.1 析因设计	36
2.2.2 正交设计	38
2.2.3 均匀设计	40
2.3 误差项变动的特殊实验设计 方案	42
2.3.1 嵌套设计	42
2.3.2 重复测量设计	44
2.3.3 裂区设计	45
2.4 协方差分析	45
2.4.1 协方差分析的必要性	45
2.4.2 平行性假定的检验	47
2.4.3 计算和检验修正平均值	48
思考与练习	50
参考文献	50
第 3 章 多元方差分析与重复测量 方差分析	51
3.1 多元方差分析	51
3.1.1 模型简介	51
3.1.2 案例:青少年牙齿发育状况 跟踪	52
3.2 重复测量数据的方差分析	55
3.2.1 模型简介	55
3.2.2 案例:进一步考察年龄对牙齿 发育的影响	57
思考与练习	61
参考文献	62
第 4 章 线性混合模型	63
4.1 模型简介	63

4.1.1	问题的提出	63
4.1.2	模型入门	64
4.2	层次聚集性数据案例	66
4.2.1	拟合基本模型结构	66
4.2.2	在固定效应中加入 自变量	69
4.2.3	在随机效应中加入 自变量	72
4.2.4	更多自变量的引入	73
4.2.5	其他常用选项	74
4.3	重复测量数据案例	75
4.3.1	对数据的初步分析	75
4.3.2	拟合基本模型结构	76
4.3.3	考虑测量间的相关性	79
4.3.4	更改对测量间相关性的 假定	81
4.3.5	模型中可用的相关矩阵 种类	83
4.4	线性混合模型进阶	83

4.4.1	线性混合模型的用途	83
4.4.2	线性混合模型与一般线性 模型的联系	84
	思考与练习	84
	参考文献	84

第5章 广义线性模型、广义估计方程 与广义线性混合模型

5.1	广义线性模型	86
5.1.1	模型简介	86
5.1.2	分析案例	87
5.2	广义估计方程	89
5.2.1	模型简介	89
5.2.2	分析案例	90
5.3	广义线性混合模型	94
5.3.1	模型简介	94
5.3.2	分析案例	94
	思考与练习	98
	参考文献	98

第二部分 回归模型

第6章 多重线性回归模型

6.1	模型简介	101
6.1.1	基本概念	101
6.1.2	分析步骤	102
6.2	案例:销售收入影响因素分析	103
6.2.1	基本分析结果	103
6.2.2	回归模型的假设检验	105
6.2.3	偏回归系数的假设检验	105
6.2.4	标准化偏回归系数	105
6.2.5	衡量回归模型效果的 指标	106
6.3	回归预测与区间估计	108
6.3.1	模型预测值	108
6.3.2	模型的区间估计	109
6.3.3	如何将模型用于预测	110
6.4	残差分析	111
6.4.1	模型的残差	111
6.4.2	利用残差考察模型适用 条件	112

6.5	逐步回归	115
6.5.1	筛选自变量的基本原则	115
6.5.2	常用的逐步回归方法	116
6.5.3	案例:固体垃圾排放量与 土地种类的关系	117
6.6	模型的进一步诊断与修正	119
6.6.1	强影响点的识别与处理	119
6.6.2	多重共线性的识别与 处理	121
6.6.3	回归模型结果解释时 应注意的问题	123
6.7	自动线性建模	124
6.7.1	界面说明	124
6.7.2	案例:生成更高精度的 预测模型	126
	思考与练习	128
	参考文献	128

第7章 线性回归的衍生模型

7.1	非直线趋势的处理:曲线直线化	129
7.1.1	模型简介	129

7.1.2 案例:通风时间和毒物浓度的曲线方程	130	8.3.2 案例:拟合推测胎儿周龄的回归方程	164
7.1.3 使用曲线估算过程分析	131	思考与练习	166
7.2 方差不齐的处理:加权最小二乘法	133	参考文献	166
7.2.1 模型简介	133	第9章 非线性回归模型	167
7.2.2 案例:不等量样品数据的回归方程	134	9.1 模型简介	167
7.2.3 使用 WLS 过程分析	135	9.1.1 问题的提出	167
7.3 共线性的处理:岭回归	137	9.1.2 模型框架	167
7.3.1 模型简介	137	9.2 案例:通风时间和毒物浓度的曲线方程	168
7.3.2 案例:用外形指标推测胎儿周龄	138	9.2.1 操作说明	168
7.4 分类变量的数值化:最优尺度回归	140	9.2.2 结果解释	169
7.4.1 模型简介	140	9.2.3 对模型的进一步分析	170
7.4.2 案例:生育子女数的回归模型	141	9.3 自定义损失函数:最小一乘法	171
7.4.3 应用最优尺度回归方法的注意事项	145	9.3.1 预分析	172
7.5 强影响点的弱化:稳健回归与分位数回归	146	9.3.2 操作说明	172
7.5.1 稳健回归	146	9.3.3 结果解释	173
7.5.2 分位数回归	147	9.4 分段回归模型的拟合	174
7.6 其余回归模型简介	148	9.4.1 预分析	175
7.6.1 断点回归	148	9.4.2 操作说明	176
7.6.2 Tobit 回归	149	9.4.3 结果解释	176
思考与练习	152	9.5 非线性回归模型进阶	177
参考文献	153	9.5.1 参数初始值的设定	177
第8章 路径分析入门	154	9.5.2 模型的拟合方法	178
8.1 两阶段最小二乘法	154	思考与练习	178
8.1.1 模型简介	154	参考文献	178
8.1.2 案例:人口背景资料对收入的影响	154	第10章 二分类 Logistic 回归模型	179
8.1.3 使用 2SLS 过程进行分析	156	10.1 模型简介	179
8.2 路径分析入门	158	10.1.1 模型入门	179
8.2.1 模型简介	158	10.1.2 一些基本概念	181
8.2.2 案例:住院费用影响因素研究	161	10.2 案例:低出生体重儿影响因素研究	182
8.3 偏最小二乘法入门	163	10.2.1 操作说明	182
8.3.1 模型简介	163	10.2.2 结果解释	183
		10.3 分类自变量的定义与比较方法	185
		10.3.1 使用哑变量的必要性	185
		10.3.2 SPSS 中预设的哑变量编码方式	187
		10.3.3 设置哑变量时的注意事项	189

10.4 自变量的筛选方法与逐步回归	189	11.4.2 案例一:与 Logistic 回归 模型比较	216
10.4.1 模型中的假设检验方法	190	11.4.3 案例二:计算 LD50	217
10.4.2 SPSS 中提供的自变量 筛选方法	190	思考与练习	219
10.4.3 案例:低体重儿数据的 逐步回归	191	参考文献	219
10.5 弗斯 Logistic 回归	193	第 12 章 对数线性模型、Poisson 回归	
10.5.1 模型简介	193	模型与潜类别分析	220
10.5.2 案例:骨肉瘤病患预后 分析	194	12.1 对数线性模型简介	220
10.6 Logistic 回归模型进阶	197	12.1.1 模型入门	220
10.6.1 模型拟合效果的判断	197	12.1.2 软件实现	221
10.6.2 拟合优度检验	198	12.2 一般对数线性模型	221
10.6.3 残差分析	200	12.2.1 初步分析	221
10.6.4 多重共线性问题	201	12.2.2 对案例的进一步分析	224
思考与练习	201	12.3 因果关系明确时的对数线性 模型	225
参考文献	201	12.3.1 操作说明	225
第 11 章 多分类、配对 Logistic 回归与 Probit 回归模型	203	12.3.2 结果解释	225
11.1 有序多分类 Logistic 回归模型	203	12.4 对数线性模型的自动筛选	226
11.1.1 模型简介	203	12.4.1 模型的选择策略	226
11.1.2 案例:工作满意度影响 因素分析	204	12.4.2 分析案例	227
11.1.3 模型适用条件的考察	207	12.5 对数线性模型与其他模型的 关系	229
11.2 无序多分类 Logistic 回归模型	208	12.5.1 与方差分析模型的关系	229
11.2.1 模型简介	208	12.5.2 与 Logistic 回归的关系	229
11.2.2 案例:不同背景人群的 选举倾向	208	12.6 Poisson 回归模型	230
11.3 1:1 配对 Logistic 回归	211	12.6.1 模型简介	230
11.3.1 模型简介	211	12.6.2 案例:冠心病死亡与吸烟 的关系	231
11.3.2 案例:雌激素与患子宫 内膜癌的关系	213	12.7 潜类别分析简介	232
11.4 Probit 回归模型	215	12.7.1 模型简介	232
11.4.1 模型简介	215	12.7.2 分析案例	233
		思考与练习	235
		参考文献	235
		第三部分 多元统计分析方法	
第 13 章 主成分分析、因子分析与 多维偏好分析	239	13.1.1 模型简介	239
13.1 主成分分析	239	13.1.2 案例:各地区经济发展 情况综合评价	241
		13.2 因子分析	244
		13.2.1 模型简介	245

13.2.2 案例:对各地区经济数据的进一步分析	246	参考文献	284
13.3 因子分析进阶	253	第 15 章 典型相关分析	285
13.3.1 公因子提取方法	254	15.1 模型简介	285
13.3.2 相关矩阵和协方差	254	15.1.1 基本原理	285
13.3.3 如何确定公因子数量	255	15.1.2 数学描述	286
13.3.4 主成分分析和因子分析 的比较	255	15.2 案例:体力指标和运动能力指标 的相关分析	286
13.4 分类数据的主成分分析 (多维偏好分析)	256	15.2.1 操作说明	287
13.4.1 模型简介	256	15.2.2 典型相关系数	287
13.4.2 界面说明	257	15.2.3 典型结构分析	289
13.4.3 案例:汽车偏好研究	260	15.2.4 典型冗余分析	290
思考与练习	264	15.3 典型相关分析进阶	290
参考文献	264	15.3.1 如何应用典型相关分析	290
第 14 章 对应分析	265	15.3.2 如何理解典型相关分析 的结果	291
14.1 模型简介	265	15.3.3 对应分析与典型相关分析 的等价性	291
14.1.1 问题的提出	265	15.3.4 典型相关分析和因子分析 的关系	291
14.1.2 模型入门	265	15.4 基于最优尺度变换的非线性典型 相关分析	292
14.1.3 软件实现	266	15.4.1 基本原理	292
14.2 案例:头发颜色与眼睛颜色 的关联	266	15.4.2 案例:多重对应分析数据 的再分析	292
14.2.1 预分析	267	思考与练习	295
14.2.2 正式分析	268	参考文献	295
14.2.3 分析结果的正确解释	272	第 16 章 多维尺度分析	296
14.2.4 对案例的进一步分析	272	16.1 不考虑个体差异的多维尺度 分析模型	296
14.3 基于平均值的对应分析	274	16.1.1 模型简介	296
14.3.1 基本原理	275	16.1.2 案例:城市间的地面 距离	297
14.3.2 案例:城市市政工程 建设状况的对应分析	275	16.1.3 距离的各种提供方式	301
14.4 对应分析进阶	278	16.2 考虑个体差异的多维尺度分析 模型	302
14.4.1 特殊类别的处理	278	16.2.1 模型简介	302
14.4.2 对应分析与因子分析 的关系	279	16.2.2 案例:饮料的口味差异 评价	303
14.4.3 对应分析的优势与劣势	279	16.2.3 模型结果的解释与优化	306
14.5 基于最优尺度变换的多重对应 分析	280	16.3 基于最优尺度变换的多维尺度	
14.5.1 基本原理	280		
14.5.2 案例:轿车用户背景资料 的对应分析	280		
思考与练习	283		

分析模型	307	17.6.1 利用标准化来调整	
16.3.1 模型简介	307	聚类模式	339
16.3.2 界面说明	307	17.6.2 如何选择聚类分析	
16.3.3 案例:用 PROXSCAL 过程		方法	340
分析饮料数据	310	17.6.3 距离/相似性测量的	
16.3.4 在模型中考虑更多维度	311	指标体系	340
16.4 多维展开模型	312	17.6.4 基于密度的聚类分析	
16.4.1 模型简介	312	方法简介	341
16.4.2 案例:场景和行为间的		思考与练习	343
匹配关系	312	参考文献	343
思考与练习	315	第 18 章 经典判别分析	344
参考文献	316	18.1 模型简介	344
第 17 章 聚类分析	317	18.1.1 基本原理	344
17.1 模型简介	317	18.1.2 适用条件	345
17.1.1 问题的提出	317	18.1.3 判别效果的评价	346
17.1.2 聚类分析入门	317	18.1.4 分析步骤	347
17.1.3 聚类分析的方法体系	318	18.2 案例:鸢尾花种类判别	347
17.2 K -均值聚类法	319	18.2.1 操作说明	347
17.2.1 基本原理	319	18.2.2 结果解释	348
17.2.2 案例:移动通信客户细分	320	18.2.3 判别结果的图形化展示	350
17.3 聚类结果的验证与自动优化	324	18.2.4 判别效果的验证	352
17.3.1 聚类结果的验证	324	18.2.5 将模型用于新案例分类	353
17.3.2 聚类用变量的调整	325	18.2.6 适用条件的判断	353
17.3.3 聚类结果的自动优化	325	18.3 贝叶斯判别分析	354
17.4 层次聚类法	329	18.3.1 基本原理	354
17.4.1 基本原理	329	18.3.2 软件实现	355
17.4.2 案例:体操裁判打分倾向		18.4 判别分析进阶	356
聚类	329	18.4.1 逐步判别分析	356
17.4.3 各种层次聚类法	333	18.4.2 判别分析和因子分析的	
17.5 两步聚类法	333	相似性和差异	356
17.5.1 基本原理	333	18.4.3 二类判别分析和多重回归	
17.5.2 案例:病例数据的聚类		分析的等价性	356
分析	335	思考与练习	357
17.6 聚类分析进阶	339	参考文献	357
		第四部分 其他统计分析方法	
第 19 章 树模型、随机森林与最近邻		19.1.1 问题的提出	361
元素法	361	19.1.2 模型入门	362
19.1 树模型简介	361	19.1.3 模型特点	365
		19.2 案例:移动客户流失预测	365
		19.2.1 操作说明	365

19.2.2 结果解释	367	20.5 支持向量机简介	410
19.3 对案例的进一步分析	369	20.5.1 基本原理	410
19.3.1 各自变量的重要性	369	20.5.2 分析案例	411
19.3.2 考虑应用模型时的 成本与收益	371	思考与练习	413
19.3.3 考虑进一步细分和剪枝	373	参考文献	413
19.3.4 将模型输出为判别程序	373	第 21 章 信度分析	414
19.4 常见的树模型算法	375	21.1 信度理论入门	414
19.4.1 CHAID 算法和穷举 CHAID 算法	375	21.1.1 真分数测量理论	414
19.4.2 CRT 算法	376	21.1.2 信度与效度	415
19.4.3 QUEST 算法	376	21.1.3 内在信度与外在信度	415
19.4.4 C5.0 算法	377	21.1.4 真分数测量理论的缺陷	415
19.5 随机森林	378	21.2 案例:问卷信度分析	416
19.5.1 模型简介	379	21.2.1 Alpha 信度系数	416
19.5.2 案例:客户风险等级 评估	381	21.2.2 对各项目的进一步分析	417
19.5.3 操作说明	381	21.2.3 对真分数测量理论适用 条件的考察	419
19.5.4 结果解释	382	21.3 其他常用的信度系数	420
19.6 最近邻元素法	386	21.3.1 重测信度	420
19.6.1 模型简介	386	21.3.2 折半信度	421
19.6.2 案例:鸢尾花种类判别	387	21.3.3 Guttman 折半系数	421
19.6.3 k -最近邻元素模型的 本质	390	21.3.4 平行模型的信度系数	422
思考与练习	392	21.3.5 严格平行模型的信度 系数	423
参考文献	392	21.3.6 评分者信度	423
第 20 章 神经网络与支持向量机	393	21.3.7 信度系数总结	425
20.1 模型简介	393	21.4 概化理论简介	425
20.1.1 基本原理	393	21.4.1 概化理论入门	425
20.1.2 注意事项	396	21.4.2 软件实现	426
20.2 案例:对低出生体重儿案例的 重新分析	397	21.5 项目反应理论简介	427
20.2.1 操作说明	397	21.5.1 项目反应理论入门	427
20.2.2 结果解释	398	21.5.2 软件实现	429
20.3 对案例的进一步分析	401	思考与练习	431
20.3.1 模型效果的图形观察	401	参考文献	431
20.3.2 尝试将模型复杂化	403	第 22 章 联合分析	432
20.3.3 纳入更多候选自变量	405	22.1 模型简介	432
20.4 径向基神经网络	407	22.1.1 为什么使用联合 分析	432
20.4.1 基本原理	407	22.1.2 常用术语	433
20.4.2 分析案例	408	22.1.3 分析步骤	434
		22.1.4 软件实现	434

22.2 联合分析的正交设计	435	参考文献	476
22.2.1 生成设计表格	435	第 24 章 生存分析	477
22.2.2 输出设计卡片	437	24.1 生存分析简介	477
22.3 联合分析的数据建模	438	24.1.1 生存分析简史	477
22.3.1 CONJOINT 的过程语法 说明	438	24.1.2 基本概念	478
22.3.2 案例:汽车偏好研究	440	24.1.3 生存分析的基本内容	480
22.3.3 对案例的进一步分析	443	24.1.4 软件实现	480
22.4 联合分析进阶	446	24.2 生存函数的估计和检验	480
22.4.1 适应性联合分析	446	24.2.1 生存函数的基本估计 方法	480
22.4.2 基于选择的联合分析	446	24.2.2 Kaplan-Meier 法	481
思考与练习	447	24.2.3 寿命表法	486
参考文献	447	24.2.4 两种方法的比较	488
第 23 章 时间序列模型	449	24.3 Cox 回归模型	489
23.1 模型简介	449	24.3.1 模型简介	489
23.1.1 基本概念	449	24.3.2 案例:术中放疗效果 分析	490
23.1.2 模型分类	450	24.3.3 模型结果的图形观察	493
23.1.3 分析步骤	450	24.4 含时依协变量的 Cox 回归模型	494
23.1.4 软件实现	450	24.4.1 时依协变量的种类	494
23.2 时间序列的建立和平稳化	451	24.4.2 用时依协变量模型验证 比例风险性	495
23.2.1 填补缺失值	451	24.4.3 用时依协变量模型考察 内在时依协变量的影响	496
23.2.2 定义时间变量	452	24.5 Cox 回归模型进阶	497
23.2.3 时间序列的平稳化	453	24.5.1 生存分析中的分层变量	497
23.3 时间序列的图形化观察	455	24.5.2 用 Cox 回归模型拟合 1:n 配伍 Logistic 回归模型	498
23.3.1 序列图	455	24.5.3 竞争风险 Cox 回归 模型	499
23.3.2 自相关图	456	24.6 加速失效时间模型	499
23.3.3 互相关图	459	24.6.1 模型简介	500
23.4 时间序列的建模与预测	460	24.6.2 案例:对术中放疗案例 拟合参数模型	501
23.4.1 指数平滑模型简介	461	思考与练习	505
23.4.2 ARMA 模型简介	462	参考文献	505
23.4.3 案例:nrc 数据的建模 预测	463	第 25 章 缺失值分析	506
23.5 季节性分解	467	25.1 缺失值理论简介	506
23.5.1 模型简介	468	25.1.1 数据的缺失机制	506
23.5.2 案例:对完整序列 nrc2 进行季节性分解	468		
23.6 时间因果模型	470		
23.6.1 模型简介	470		
23.6.2 案例:KPI 驱动因素 发现	471		
思考与练习	476		

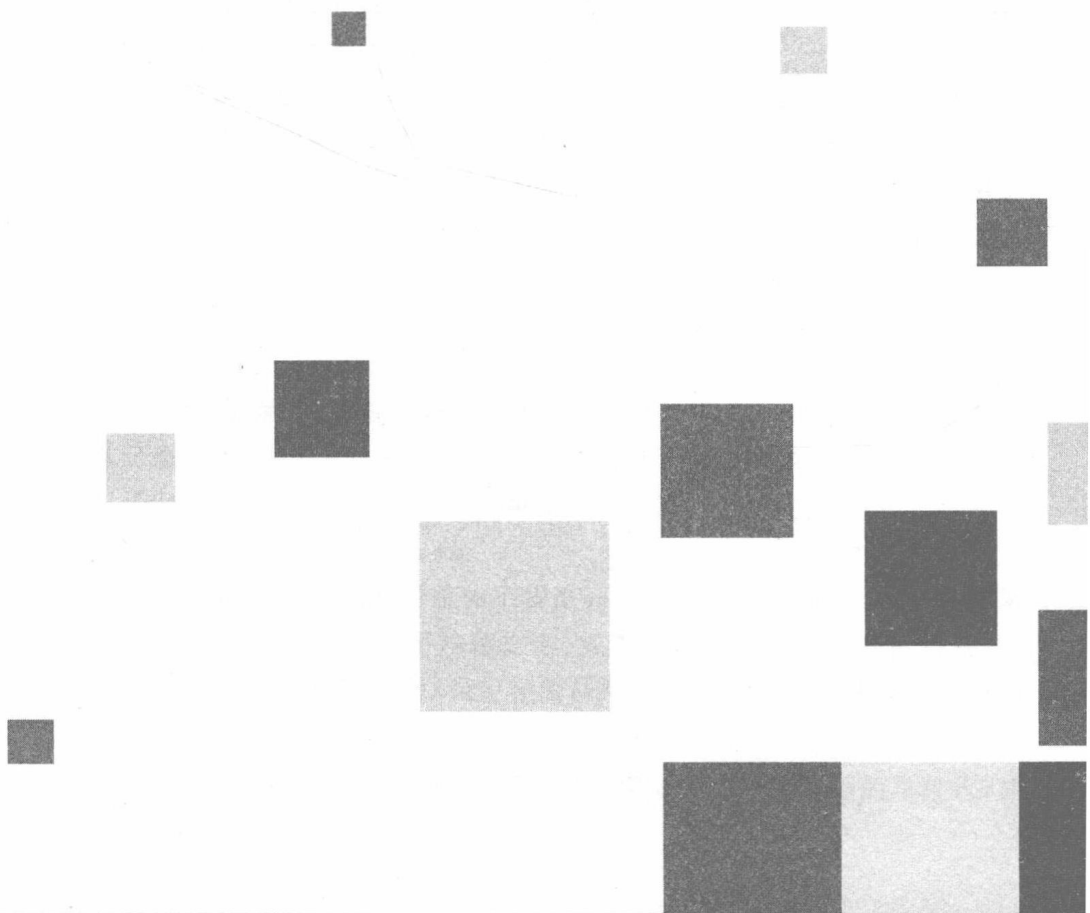
25.1.2 缺失值的处理方法	507	25.3.3 使用 EM 算法进行填补	515
25.2 对缺失情况的基本分析	508	25.4 多重填补	517
25.2.1 生成缺失数据	508	25.4.1 模型简介	517
25.2.2 缺失模式分析	509	25.4.2 缺失模式分析	517
25.2.3 缺失情况的描述统计	511	25.4.3 缺失值的多重填补	519
25.3 缺失值填补技术	512	25.4.4 采用填补后数据建模	520
25.3.1 列表输出	512	思考与练习	521
25.3.2 使用回归算法进行填补	513	参考文献	521
附录 1 常见多变量/多元统计分析方法分类图	522		
附录 2 Python 插件和 R 插件的安装方法	523		



1


第一部分

一般线性模型、线性混合模型 与广义线性模型



第1章 方差分析模型

通过对《SPSS 统计分析基础教程》(第3版)(以下简称基础教程)的学习,读者已经全面掌握了 IBM SPSS Statistics 的基本操作,以及图表绘制、描述统计技术和单因素统计分析方法。但是,真实的世界复杂多变,各种变量间的联系错综复杂,仅仅依靠描述统计或者简单的统计推断方法往往无法满足分析需求,必须依靠更强大的统计模型来解决问题。本书将进一步介绍各种高级统计模型,而本章将要介绍的方差分析模型就是其中最基础和最常用的一种。

 一见模型二字,很多读者就会觉得继续学习下去的心理压力很大。实际上,模型无非是对复杂现实世界的一种简化描述而已,一个出色的模型必然比现实世界更简明、易懂,而模型也不一定是枯燥的公式表达,比如说当其以一种赏心悦目的姿态出现在 T 台上时,中文就将其翻译为“模特”。怎么样,这样解释之后,大家对模型的感觉是不是变好了一点?

1.1 模型简介

在实际项目中,研究者在分析数据时往往需要同时考察多个因素对因变量的影响,如要研究性别对身高的影响,就应当控制年龄、遗传、营养状况等因素的影响。对此单因素统计分析方法是无能为力的,而以方差分析模型为代表的多因素统计分析方法可以在控制其他因素影响的同时研究两个因素之间的关系,分析的效率更高,适用范围更广。

此外,许多时候各个因素之间还存在交互作用。例如,在研究催化剂对化学反应的催化能力时,如果该催化剂只在某个温度范围内效果最佳,则单独研究该催化剂的催化作用并无实际意义,此时催化剂和温度之间的交互作用也应成为研究的重点,即必须要研究在什么温度下该催化剂的催化能力最佳。对交互作用的分析是方差分析模型的特长之一。

1.1.1 模型入门

1. 单因素方差分析模型的结构

首先从一个简单的统计模型开始,假设现在希望描述某个人群的月收入状况,那么根据统计学知识,如果月收入服从正态分布(请读者注意这个前提假设),则平均值能够表示集中趋势,标准差能够表示离散趋势,任何一位受访者 i 的月收入 Y_i 可以被表达为如下形式:

$$Y_i = \mu + \varepsilon_i$$

其中, Y_i 代表第 i 位受访者的月收入。显然,此时总体平均值 μ 是 Y_i 的最佳估计值,而 ε_i 则表示因各种原因导致的第 i 位受访者实际月收入和平均值之差,或者说反映了抽样中的随机误差。由于已经假定月收入服从正态分布,因此模型可以设定 ε_i 服从平均值为 0、标准差为某个定值的正态分布 $N(0, \sigma^2)$ 。

下面开始扩展模型框架。假设现在比较三种职业,即医生、律师和软件工程师的月收入,并

判断其有无差异,那么最简单的做法就是对每种职业的人群都进行随机抽样,得到三组受访者,收集他们的月收入数据,然后进行检验。在此问题中,每一位受访者月收入的平均估计值 Y_{ij} 可以被表达为如下形式:

$$Y_{ij} = \mu_i + \varepsilon_{ij}$$

其中, Y_{ij} 代表第 i 组中第 j 位受访者的月收入。显然,在此表达式中 μ_i 表示某一组月收入的平均值, i 的取值范围为 $1 \sim 3$, 分别代表三种职业之一; 而 ε_{ij} 表示第 i 组的第 j 位受访者的月收入相对于本组月收入平均值的随机误差。

下面来看模型中对 ε_{ij} 的设定,在原始模型框架中,显然各组的 ε_{ij} 可以服从各自的正态分布,但这类研究往往更关心平均值的差异,因此为了简化模型架构,一般会进一步假设各组的 ε_{ij} 服从同一个正态分布,即无论 i 取值是多少, ε_{ij} 均服从同一个平均值为 0、标准差为某个定值的正态分布 $N(0, \sigma^2)$ 。这样一来,如果三种职业的月收入无差异,那么它就应当等于总体平均值(平均水平)再加上一个随机误差项,实际上就变成了同一个变量的正态分布 $N(\mu, \sigma^2)$ 。为了能够对收入水平进行预测,人们又规定 $E(Y) = \mu_i$, 即第 i 组个体的月收入估计值等于该组月收入的平均水平,结合模型结构,这应当不难理解。实际上,如果对应的是样本数据,该估计值就是各组的样本平均值。

为了统计推断的需要,以上模型往往被改写成如下形式:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

其中, μ 表示不考虑职业时月收入的总平均水平; α_i 表示职业 i 的主效应,即职业 i 月收入平均水平的改变情况。例如, $\alpha_1 = 1\,000$, 表明职业 i 月收入的平均水平要比总平均水平高 1 000 元。如果职业 1 和职业 3 的平均月收入不相等,则有 $\alpha_1 \neq \alpha_3$ 。反之,如果三种职业的平均月收入无差异,则因各种职业均不存在主效应而有 $\alpha_1 = \alpha_2 = \alpha_3 = 0$ 。因此,检验职业种类是否对月收入有影响,就是检验如下假设:

$$H_0: \text{对任意的 } i \text{ 取值, 都有 } \alpha_i = 0; H_1: \text{至少有一个 } \alpha_i \neq 0$$

μ 、 α_i 等显然是一个相对的大小,可以有无限多组取值方式。例如,职业 1 比职业 3 的平均月收入高 1 000 元,则当 α_3 为 500 时, α_1 就应当是 1 500; 当 α_3 为 100 时, α_1 就应当是 1 100, 总之加上 1 000 即可。为了能够在实际问题中得到 μ 、 α_i 的具体估计值,模型拟合中又会对它们有一些附加的设定,这被称为模型拟合时的约束条件,相关的详细介绍见后。

在基础教程中,大家已经学习了方差分析的基本思想是变异分解。例如,在单因素方差分析中总变异被分解为如下两部分:

$$\text{总变异} = \text{处理因素导致的变异} + \text{随机变异}$$

对照前面的模型表达式,就会发现实际上 α_i 对应了处理因素导致的变异,而 ε_{ij} 对应了相应的随机变异。在多因素方差分析模型中,这一原理没有任何变化,只是模型中考虑的因素更多而已。

2. 两因素方差分析模型的结构

下面开始对单因素模型进行扩展。同样是上面的问题,有研究者提出:性别也对收入水平有影响,也许正是因为其中一组中男性的比例高于其他组,才导致该组的平均月收入高于后者。为什么不考虑控制性别的影响? 如果要同时考虑性别和职业对收入的影响,则模型扩展为

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \varepsilon_{ijk}$$

该模型对应如下变异分解式: