21世纪统计学系列教材

应用多元统计分析

赵博娟 编著

Applied Multivariate Statistical Analysis

中国人民大学出版社

21世纪统计学系列教材

应用多元统计分析

赵博娟　编著

Applied Multivariate Statistical Analysis

中国人民大学出版社
· 北京 ·

图书在版编目（CIP）数据

应用多元统计分析／赵博娟编著. -- 北京 ：中国人民大学出版社，2019.1
21世纪统计学系列教材
ISBN 978-7-300-26460-8

Ⅰ.①应… Ⅱ.①赵… Ⅲ.①多元分析-统计分析 Ⅳ.①O212.4

中国版本图书馆CIP数据核字（2018）第275190号

21世纪统计学系列教材
应用多元统计分析
赵博娟 编著
Yingyong Duoyuan Tongji Fenxi

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） 010－62511770（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 北京七色印务有限公司
规 格 185 mm×260 mm 16开本 版 次 2019年1月第1版
印 张 11插页1 印 次 2019年1月第1次印刷
字 数 225 000 定 价 29.00元

总序

Preface

改革开放以来, 高等统计教育有了很大的发展. 随着课程设置的不断调整, 有不少教材出版, 同时也翻译引进了一些国外优秀教材. 作为培养我国统计专门人才的摇篮, 中国人民大学统计学系自 1952 年创建以来, 走过了风风雨雨, 一直坚持理论与应用相结合的办学方向, 培养能够理论联系实际、解决实际问题的高层次人才. 随着新知识经济和网络时代的到来, 我们在教学科研的实践中深切地感受到, 无论是自然科学领域、社会科学领域的研究, 还是国家宏观管理和企业生产经营管理, 甚至人们的日常生活, 信息需求量日益增多, 信息处理技术更加复杂, 作为信息技术支柱的统计方法, 越来越广泛地应用于各个领域.

面对新的形势, 我们一直在思索, 课程设置、教材选择、教学方式等怎样才能使学生适应社会经济发展的客观需要. 在反复酝酿、不断尝试的基础上, 我们决定与统计学界的同仁共同编写、出版一套面向 21 世纪的统计学系列教材.

这套系列教材聘请了中科院院士、中国科学技术大学陈希孺教授, 上海财经大学数量经济研究院张尧庭教授, 中国科学院数学与系统科学研究所冯士雍研究员等作为编委. 他们长期任中国人民大学的兼职教授, 一直关心、支持着统计学系的学科建设和应用统计的发展. 中国人民大学应用统计科学研究中心 2000 年已成为国家级研究基地, 这些专家是首批专职或兼职研究人员. 这一开放性研究基地的运作, 将有利于提升我国应用统计科学研究的水平, 也必将进一步促进高等统计教育的发展.

这套教材是我们奉献给新世纪的, 希望它能促进应用统计教育水平的提高. 这套教材力求体现以下特点:

第一, 在教材选择上, 主要面向经济类统计学专业. 选材既包括统计教材也包括风险管理与精算方面的教材. 尽管名为统计学系列教材, 但并不求大、求全, 而是力求精选. 对于目前已有的内容较为成熟、适合教学需要、公认的较好的教材, 并未列入本次出版计划.

第二, 每部教材的内容和写作, 注意广泛吸收国内外优秀教材的成果. 教材力求简明易懂、内容系统和实用, 注重对统计方法思想的阐述, 并结合大量实际数据和实例说明统计方法的特点及应用条件.

第三, 强调与计算机的结合. 为着力提高学生运用统计方法分析解决问题的能力, 教材所涉及的统计计算, 要求运用目前已有的统计软件. 根据教材内容, 选择使用 SAS, SPSS, TSP, STATISTICA, EViews, MINITAB, Excel 等.

感谢中国人民大学出版社的同志们, 他们怀着发展我国应用统计科学的热情和提高统计教育水平的愿望, 经过反复论证, 使这套教材得以出版. 感谢参与教材编写的同行专家、统计学系的教师. 愿大家的辛勤劳动能够结出丰硕的果实. 我们期待着与统计学界的同仁共同创造应用统计辉煌的明天.

易丹辉

于中国人民大学

前言

Foreword

本书从实践的角度, 介绍多元统计分析方法, 并在介绍分析原理的同时, 给出用三种常用统计软件进行编程和运行的操作细节.

书中包括 11 章内容, 依次为数据收集和描述、多元正态分布、线性回归分析案例、聚类分析、主成分分析、因子分析、对应分析、典型相关分析、判别分析、联合分析、广义线性模型简介. 每章的主要内容都包括例子或案例分析, 还有软件输出结果解释, 各章最后介绍的上机实现软件包括可免费下载的 R, 以及商业软件 SPSS 和 SAS. 在使用 SPSS 时, 除了个别章节需要编程, 基本仅需用模块选项点击, 而在使用 R 和 SAS 时, 都需要编程. 不过, 本书涉及的编程都比较简单, 实践中可以模仿例子中的编程, 易于使用.

书中内容力求精练, 以最少的篇幅尽量广阔地概括多元统计分析所涉及的概念和方法, 个别章节也可以根据教学需要略去不讲.

学习本课程所需要的基础课程包括数学分析、高等代数、概率论、数理统计和一般线性回归等. 软件方面不需要很强的编程基础, 但需要了解 R、SPSS 或 SAS 最基本的操作, 这些操作可以在学习和使用中逐渐熟悉.

科学研究离不开数据的收集和分析. 希望学生通过对本课程的学习, 能够比较系统地掌握多元统计分析方法, 熟悉统计软件的使用和对输出结果的解释.

限于笔者水平和经验, 不足之处在所难免, 恳请广大师生在使用中多提宝贵意见和建议.

赵博娟

目录

Contents

第 1 章　数据收集和描述 ····· 1

1.1　数据收集 ····· 1

1.2　两变量图描述和量化分析 ····· 4

1.3　多变量图描述 ····· 11

1.4　上机实现 ····· 14

习题 ····· 16

第 2 章　多元正态分布 ····· 17

2.1　随机向量及其分布和性质 ····· 17

2.2　多元正态分布 ····· 18

2.3　多元正态参数的样本估计 ····· 23

2.4　多元正态参数的假设检验 ····· 25

2.5　正态性检验 ····· 37

2.6　上机实现 ····· 42

习题 ····· 43

第 3 章　线性回归分析案例 **44**

3.1　房地产公司预测房价 44

3.2　Bikeshare 数据 55

3.3　上机实现 60

习题 61

第 4 章　聚类分析 **62**

4.1　点间距离或相似度 63

4.2　分层聚类 68

4.3　K-均值和 K-中心点聚类 77

4.4　确定类的个数 80

4.5　上机实现 81

习题 82

第 5 章　主成分分析 **83**

5.1　基本思想 83

5.2　主成分的定义和计算 84

5.3　降维和解释 86

5.4　上机实现 90

习题 91

第 6 章　因子分析 **92**

6.1　基本思想 92

6.2　因子分析模型 93

6.3　例子和解释 96

6.4　因子分析求解简介 101

6.5　上机实现 102

习题 · · · · · 103

第 7 章 对应分析 · · · · · 104

7.1 基本思想 · · · · · 104

7.2 分析原理 · · · · · 104

7.3 例子和解释 · · · · · 107

7.4 上机实现 · · · · · 110

习题 · · · · · 111

第 8 章 典型相关分析 · · · · · 113

8.1 基本思想 · · · · · 113

8.2 求解原理 · · · · · 114

8.3 有关检验和典型冗余分析 · · · · · 116

8.4 例子和解释 · · · · · 117

8.5 上机实现 · · · · · 123

习题 · · · · · 124

第 9 章 判别分析 · · · · · 125

9.1 基本思想 · · · · · 125

9.2 距离判别 · · · · · 126

9.3 Bayes 判别 · · · · · 127

9.4 Fisher 判别 · · · · · 128

9.5 逐步判别 · · · · · 129

9.6 例子和解释 · · · · · 130

9.7 上机实现 · · · · · 140

习题 · · · · · 142

第 10 章　联合分析 ····· 143

10.1　基本思想 ····· 143

10.2　联合分析 ····· 144

10.3　例子和解释 ····· 145

10.4　上机实现 ····· 150

习题 ····· 152

第 11 章　广义线性模型简介 ····· 153

11.1　广义线性模型简介 ····· 153

11.2　Logistic 回归模型 ····· 154

11.3　Poisson 回归模型 ····· 157

11.4　上机实现 ····· 161

习题 ····· 163

参考文献 ····· 164

应用多元统计分析

第 1 章

数据收集和描述

1.1 数据收集

1.1.1 一手和二手数据

数据有**一手数据** (primary data, 也叫**原始数据**) 和**二手数据** (secondary data) 之分. 前者需要花时间和资源去收集, 后者不需要. 一些组织或个人为了某些特定的应用或研究目的, 必须自己来收集数据, 这样收集的就是一手数据. 如果数据是其他人收集的 (对于他们来说是一手数据), 而你仅仅用它来进行分析或进行数据挖掘, 该数据对你来说就是二手数据. 常见的二手数据包括从期刊和网络上找到的, 由政府、机构、公司、组织专门收集和维护的数据, 比如外汇牌价、房价、人口抽样调查结果等数据. 一般, 在决定开展一手数据收集前, 要先调研一下是否已经有相应的二手数据存在. 在使用二手数据的过程中, 我们要了解数据的来源及其真实和可靠程度, 确定数据是否符合我们的研究目的, 以便恰当地进行分析并解释结果.

数据有**试验数据**和**观测数据**之分. 试验数据中的自变量取值是可以人为改变的. 比如, 通过调节冶炼金属的时间、温度和原料配比, 寻找最佳组合条件, 以炼成满足某些性能指标的金属材料; 又如, 根据各种食品和数量的需要, 通过调节电烤箱的温度和烤制时间, 寻找使食品口感最佳的组合, 以制定自动控制系统菜单, 方便烤箱使用. 这类数据便于分析因果关系. 实践中, 我们遇到的绝大部分二手数据都是观测数据. 如跟

踪观测研究吸烟与罹患癌症的关系等. 这类数据的自变量不能随意调节, 我们不能在身体健康状况完全类似的人中随机抽样, 强迫一部分人吸烟而另一部分不吸, 之后观测他们将来是否罹患癌症.

要获得观测数据, 有两大类抽样方法: **概率抽样方法** (probability sampling method) 和**非概率抽样方法** (nonprobability sampling method). 概率抽样方法包括**简单随机抽样** (simple random sampling, 如常用的抓阄)、**系统抽样** (systematic sampling, 如大家排队报号, 报 7 的倍数的同学出列)、**分层抽样** (stratified sampling, 如本科生和研究生各自随机抽取 100 和 50 人)、**整群抽样** (cluster sampling, 比如某些地区被随机抽中作为某项房屋限购政策试点, 这些地区中的每个人都被选中参加) 和**多级抽样** (multistage sampling, 比如先在全国高校中随机抽几个, 再从抽中的学校中随机抽班级, 再从抽中的班级中随机抽几个同学). 非概率抽样方法包括**方便抽样** (convenience sampling, 比如去食堂门口让去吃饭的学生填写对食堂的满意度调查问卷) 和**目的抽样** (purposive sampling, 也叫**主观抽样** (subjective sampling), 比如因想知道房地产开发商对房价的看法而做的问卷调查) 等. 概率抽样方法的**偏差** (bias) 可以估算, 具体细节可以参考《抽样技术》(*Sampling Techniques*, Cochran (1977)) 等书籍.

数据的收集途径包括电话调查、邮寄问卷、如网络那样的新媒体调查、直接观测和面对面访问等. 数据收集的具体步骤包括: (1) 定义关心的调查问题; (2) 定义关心的总体; (3) 专家开发调查问卷; (4) 小规模试用 (pre-test); (5) 决定样本量和抽样方法; (6) 抽取样本并分析获得的数据. 例如 "全球经理人意见调查", 其问卷由有关专家开发, 而且每年都有更新修改. 各国在开展问卷调查之前, 要将专家的问卷翻译审核, 经小规模试用, 以检查问卷中是否存在文字沟通理解等方面的问题. 样本量和抽样方法由世界经济论坛统一规定, 按公司或企业的规模进行分层抽样, 由抽中企业的几个主要负责人之一填写问卷. 根据各国家或地区的具体情况, 可以同时采用几种问卷收集途径, 如邮寄问卷和面对面访问.

电话调查和邮寄问卷两种方法很常用, 它们各有各的特点. 电话调查花费不太高而且有效率, 但要调查的人不一定能通过电话找到, 因此在某些情况下会产生调查的系统偏差. 比如, Gallup 调查公司在 1948 年杜鲁门 – 杜威总统竞选预测上发生了该公司史上一次大失误, 其原因是当年更倾向支持民主党的穷人很少使用电话. 在进行电话调查时, 有些要注意的事项, 如要解释调查目的, 告诉被访者结果保密, 调查时间要短, 要使用**固定结果问题** (closed-end questions), 以易于在电脑上快速记录受访人的答案, 如 "请问你所属的党派 —— 是共和党、民主党还是其他党派?" 而对于**不固定结果问题** (open-end questions), 如 "你所属的党派?" 受访者可以给出任何答案, 需要时间思考和记录, 因此更适用于邮寄问卷. 邮寄问卷花费少, 花费时间相对长. 在面对面调查的实践中, 也可以由调查员协助解释填写问卷, 以缩短所需时间, 参见 Groebner et al. (2002).

新兴媒体, 比如微信 (WeChat)、Facebook (脸书)、Twitter (推特) 等的影响力已远超传统媒体, 因此, 通过新兴媒体进行民意调查和大数据分析也日渐重要.

值得注意的是, 对某些敏感问题, 有些受访人在回答时可能会撒谎, 也有可能干脆不回答. 此外, 为了节约时间和费用, 有的调查没有通过概率随机抽样产生受访人, 受访人是一些方便样本, 这些数据的分析结果可能会导致一定的系统偏差.

了解数据收集的方法、途径、特点及其存在的问题, 有助于我们对数据分析结果给出正确理解和解释.

1.1.2 数据的度量级别和类型

尽管数据可以来自不同的领域, 但数据的**度量级别** (measurement levels) 和类型都是一样的.

数据度量级别从低到高包括**名义数据** (nominal data)、**有序数据** (ordinal data)、**区间数据** (interval data) 和**比率数据** (ratio data). 名义数据, 也叫**定性数据**或**分类数据** (categorical data), 是最低级形式的数据, 我们可以对数据取值任意编号. 如对婚姻状态, 可以用 1 ~ 4 或 M, S, D 和 O 分别标记, 即 M (已婚), S (未婚), D (离婚) 和 O (其他). 有序数据比名义数据高一级别, 数据的类别是有序的. 如健康状态: 1 (非常健康), 2 (健康), 3 (一般), 4 (不健康) 和 5 (非常不健康). 区间数据是有序的, 而且任意两点的距离是可以精确度量出来的. 如华氏 (Fahrenheit) 和摄氏 (Celsius) 温度. 比率数据有真正有意义的零点, 度量级别最高. 如体重、高度、距离、钱包里的钱数等等. 口袋没钱, 不管是美元还是人民币元, 都是 0.

数据类型可以分为**定量数据** (quantitive data) 和**定性数据** (qualitative data). 定量数据指可以用数字量化的数据, 具体又可以分为连续型数据 (如身高) 和离散型数据 (如某路口每月交通事故次数); 定性数据指取值分类别的数据, 如性别分为男女. 定性数据可以是名义数据, 也可以是有序数据. 定量数据可以是区间数据, 也可以是比率数据.

从数据整体特点来看, 数据类型还可以分为**横截面数据** (cross-sectional data) 和**时间序列数据** (time series data) 及**纵向数据** (longitudinal data). 横截面数据是在某个固定的时间点观测得到的一组数据, 如某校大学生高考入学成绩; 时间序列数据和纵向数据中, 每个对象都有重复观测, 这些重复观测可能是按某种顺序的不同的时间点或不同的状况采集的. 许多社会和医学领域的前瞻群组跟踪研究 (prospective cohort study) 和回顾群组调查研究 (retrospective cohort study) 都在不同时间点有多次观测.

数据的类型和级别不同, 所采用的存放格式和分析方法也不同.

1.1.3 数据存放格式

原始整理的数据多存成矩形 Excel 表格形式 (如 `*.xls`), 如图 1.1 所示. 为了读入

数据, 有些统计分析软件, 如 R, 可以转存成 *.csv 格式. 数据的类型不同, 原始数据存放细节也有所不同.

day2012

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	dteday	season	mnth	holiday	weekday	workingda	weathersi	temp	atemp	hum	windspeed	casual	registered	cnt
2	2012/1/1	1	1	0	0	0	1	0.37	0.375621	0.6925	0.192167	686	1608	2294
3	2012/1/2	1	1	1	1	0	1	0.273043	0.252304	0.381304	0.329665	244	1707	1951
4	2012/1/3	1	1	0	2	1	1	0.15	0.126275	0.44125	0.365671	89	2147	2236
5	2012/1/4	1	1	0	3	1	2	0.1075	0.119337	0.414583	0.1847	95	2273	2368
6	2012/1/5	1	1	0	4	1	1	0.265833	0.278412	0.524167	0.129987	140	3132	3272

图 1.1 Bikeshare 数据的 Excel 格式

截面数据的存放很简单. 以用 Excel 进行数据收集整理为例, 可在第一行存放变量名 (variable), 从第二行开始, 每一行存放一个样品或对象的观测值, 即一行数据对应一个样本观测. 每个变量名所对应的列为各样本点的观测值.

对于时间序列, 前瞻群组跟踪研究数据和回顾群组调查研究数据, 每一个观测对象可以在几个不同时间点有观测值, 这种数据有两种存放方式: (1) 每一个对象有几行观测值, 常称长表格式 (long form); (2) 每一个对象仅有一行观测值, 不同时间观测点用不同的变量名, 常称短表或宽表格式 (short form 或 wide form).

对于一些整合后的列联表数据, 也可以存成矩形表格形式. 可在变量名中添加一个频数 (计数), 将表格中的数值放在这个频数变量所在的列. 这类数据在分析中通常要做加权处理.

1.2 两变量图描述和量化分析

例 1.1

自行车租用数据[①] (day2012.csv) 数据含 2012 年 Capital 自行车租用公司每天租车人次数数据及后来添加的有关当天天气、季节等信息, 见图 1.1. 该数据的具体变量包括 *dteday* (日期), *season* (季节, 1—春, 2—夏, 3—秋, 4—冬), *mnth* (月, $1 \sim 12$), *holiday* (是否节假日, 0—否, 1—是), *weekday* (星期几, $0 \sim 6$), *workday* (是否工作日, 1—不是周末或节假日, 0—是周末或节假日), *weathsit* (天气情况, 1—晴, 无云或少云或局部多云, 2—有雾或多云, 3—小雪或小雨等, 4—大雪或大雨或大雾或冰雹等), *temp* ($0 \sim 1$, 标准化后的摄氏温度), *atemp* ($0 \sim 1$, 标准化后的体感摄氏温度), *hum* ($0 \sim 1$, 标准化后的湿度), *windspd* ($0 \sim 1$, 标准化后的风速), *casual* (临时用户人数), *regist* (注册用户人数), *cnt* (临时用户和注册用户人数之和).

① http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset#.

对于这个数据, 我们一般会关心 *casual* (临时用户人数), *regist* (注册用户人数) 或 *cnt* (临时用户和注册用户人数之和) 受天气和季节影响的情况, 也就是把这三个定量变量之一看成**因变量** (dependent variable), 其他定性或定量变量看成**自变量** (independent variable), 做回归分析. 本节我们先就数据中的不同变量类型, 介绍如何用图描述和简单量化分析了解变量之间的关系.

1.2.1 定量变量与定性变量

1. 图描述

定量变量与定性变量之间的关系, 如 *season* 和 *cnt*, 可用直方图和盒形图展现, 见图 1.2. 直方图和盒形图都显示租车人数与季节大致有关, 春天最少, 秋天最多. 盒形图也叫箱线图 (box-whisker plot 或 box plot), 由中位数和上下四分位点决定盒 (box) 部分, 用上下四分位点差的 1.5 倍作长度, 画上下线 (whisker) 和线端 (the ends of whisker), 若所有观测都在线端内, 以最大和最小值作为线端, 按此画法得到的盒形图也称为 Tukey Boxplot. 以正态分布 $N(\mu,\sigma^2)$ 为例, 中位数为 μ, 上下四分位点距中位数 $0.674\ 5\sigma$, 两线端各距上下四分位点 1.5 倍的四分位距离 ($IQR = 2\times 0.674\ 5\sigma$), 也就是距离中位数 $0.674\ 5\sigma\times 4 = 2.698\ 0\sigma$. 即两线端内相当于 99.30% 置信区间. 因此, 在线端外的点一般看作可能的异常点 (outliers).

图 1.2 是用 SPSS 画的 (操作细节见 1.4 节 "上机实现"). 用下面的 R 语句也可得到

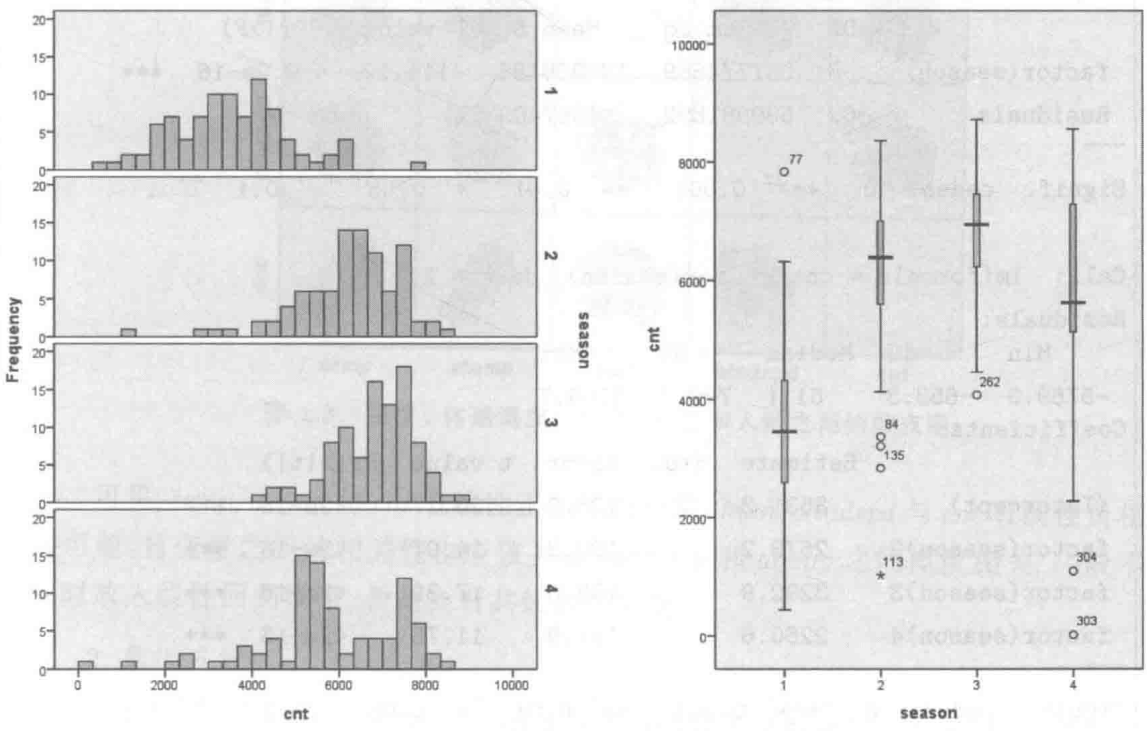

图 1.2 反映 *cnt* 和 *season* 关系的直方图和盒形图

图 1.2 中的直方图和盒形图.

```
X=read.csv("d:\\data\\day2012.csv");par(mfrow=c(1,2));
season=X[,2];cnt=X[,14];seasonid=c("spring","summer","autumn","winter")
par(mfrow=c(4,1),mar=c(1,1,4,1));xmin=min(X[,14]);xmax=max(X[,14]);
for (i in 1:4) hist(cnt[season==i],xlim=c(xmin,xmax),main=seasonid[i])
boxplot(cnt~season,xlab="season",ylab="cnt")
```

除了用于描述数据, 直方图、盒形图和 Q-Q 图 (见第 2 章) 还常用来检验数据是否严重偏离正态.

2. 量化分析

定量变量和定性变量之间的量化分析可以用方差分析进行, 即在正态和方差相等的假设下检验四个季节的人数均值是否一样:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4; \quad H_a: \text{均值不全一样}$$

用 R 语句 `out=lm(cnt~factor(season),data=X);anova(out);summary(out)` 得到下面 `anova(out)` 和 `summary(out)` 的输出:

```
Analysis of Variance Table
Response:  cnt
                  Df     Sum Sq    Mean Sq  F value    Pr(>F)
 factor(season)    3  567774559  189258186   114.19  < 2.2e-16  ***
 Residuals       362  599981892    1657409
---
Signif.  codes:  0  '***' 0.001  '**' 0.01  '*' 0.05  '.' 0.1  ' ' 1

Call:  lm(formula = cnt~factor(season), data = X)
Residuals:
     Min       1Q   Median       3Q      Max
 -5769.9   -653.3     61.1   788.1   4304.7
Coefficients:
                 Estimate  Std.  Error  t value  Pr(>|t|)
 (Intercept)        3531.3         135.0    26.17    <2e-16  ***
 factor(season)2    2678.2         190.3    14.07    <2e-16  ***
 factor(season)3    3292.9         189.3    17.39    <2e-16  ***
 factor(season)4    2260.6         191.9    11.78    <2e-16  ***
---
Signif.  codes:  0  '***' 0.001  '**' 0.01  '*' 0.05  '.' 0.1  ' ' 1
Residual standard error:  1287 on 362 degrees of freedom
```

```
Multiple R-squared:  0.4862, Adjusted R-squared:  0.482
F-statistic:  114.2 on 3 and 362 DF, p-value:  < 2.2e-16
```

其结果显示四个季节人数均值相等的零假设被拒绝, 而且后三个季度各自的平均人数都显著高于第一季度的平均人数.

1.2.2 两个定量变量

1. 图描述

两个定量变量之间的关系, 如温度 *temp*、体感温度 *atemp*、湿度 *hum*、风速 *windspd*、临时和注册用户人数之和 *cnt* 两两变量之间的关系, 可以用矩阵散点图描述, 如图 1.3 所示.

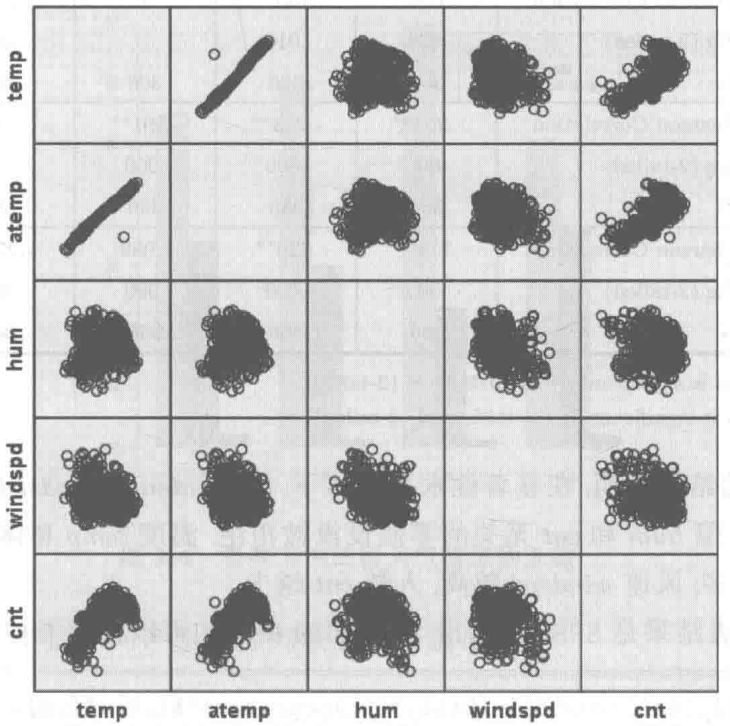

图 1.3 温度、体感温度、湿度、风速和人数之间的散点图

可见, *temp*, *atemp* 与 *cnt* 有线性正相关的迹象, *hum*, *windspd* 与 *cnt* 有线性负相关的可能. 自变量之间的相关性也应该关注, 如 *temp* 和 *atemp* 之间高度相关, 一般不能同时放入线性回归模型, 否则会有共线性的问题.

2. 量化分析

定量变量之间的量化分析可采用变量之间的相关系数进行, 见表 1.1. 在两个变量不相关 (即 $r = 0$) 的零假设下, 统计量