

目 录

前言

第1章 引言 1

- 1.1 什么是回归分析 1
- 1.2 公用数据集 2
- 1.3 回归分析的应用举例 2
- 1.4 回归分析的步骤 9
- 1.5 本书的内容及结构 14
- 习题 14

第2章 简单线性回归 16

- 2.1 引言 16
- 2.2 协方差和相关系数 16
- 2.3 例：计算机的修理数据 20
- 2.4 简单线性回归模型 23
- 2.5 参数估计 23
- 2.6 假设检验 26
- 2.7 置信区间 29
- 2.8 预测 30
- 2.9 拟合效果度量 31
- 2.10 通过原点的回归直线 34
- 2.11 平凡的回归模型 34
- 2.12 文献 35
- 习题 36

第3章 多元线性回归 41

- 3.1 引言 41
- 3.2 数据的描述及模型 41
- 3.3 例：主管人员业绩数据 42
- 3.4 参数估计 43
- 3.5 回归系数的解释 45
- 3.6 最小二乘估计的性质 47
- 3.7 复相关系数 47
- 3.8 单个回归系数的推断 48
- 3.9 线性模型中的假设检验 50
- 3.10 预测 58
- 3.11 小结 58
- 习题 59
- 附录 63

第4章 回归诊断：模型合理性的检测 67

- 4.1 引言 67
- 4.2 标准的回归假定 67
- 4.3 形形色色的残差 69
- 4.4 图形方法 71
- 4.5 拟合模型前的图形工具 73
- 4.6 拟合模型后的图形工具 76
- 4.7 检查线性和正态性假定 76
- 4.8 杠杆、影响及异常 77
- 4.9 影响的各种量度 81
- 4.10 位势-残差图 84
- 4.11 如何处理异常点？ 85
- 4.12 变量在回归方程中的作用 86
- 4.13 添加一个预测变量的效应 90
- 4.14 稳健回归 91
- 习题 91

第5章 定性预测变量 97

- 5.1 引言 97
- 5.2 薪水调查数据 97

5.3	交互作用变量	100
5.4	回归方程系统：两个组的比较	104
5.5	示性变量的其他应用	112
5.6	季节性	112
5.7	回归参数对时间的稳定性	114
	习题	115
第 6 章	变量的变换	122
6.1	引言	122
6.2	线性化变换	123
6.3	X 射线杀菌的数据	125
6.4	方差稳定性变换	129
6.5	异方差误差的诊断	133
6.6	异方差性的消除	135
6.7	加权最小二乘法	136
6.8	数据的对数变换	137
6.9	幂变换	139
6.10	小结	142
	习题	142
第 7 章	加权最小二乘法	146
7.1	引言	146
7.2	异方差模型	147
7.3	两阶段估计	149
7.4	教育经费数据	151
7.5	拟合剂量 - 反应关系曲线	159
	习题	160
第 8 章	相关误差的问题	161
8.1	引言：自相关	161
8.2	消费者支出和货币存量	162
8.3	Durbin-Watson 统计量	164
8.4	通过变换消除自相关	165
8.5	误差自相关时的迭代估计	167
8.6	自相关和变量的缺失	168

- 8.7 住房开工分析 168
- 8.8 Durbin-Watson 统计量的局限性 171
- 8.9 采用示性变量消除季节效应 173
- 8.10 时间序列间的回归 175
- 习题 176

第 9 章 共线性数据的分析 180

- 9.1 引言 180
- 9.2 对推断的影响 181
- 9.3 对预测的影响 186
- 9.4 多重共线性的检测 189
- 9.5 中心化及尺度变换 194
- 9.6 主成分方法 197
- 9.7 附加约束 201
- 9.8 搜寻 β 的线性函数 202
- 9.9 使用主成分作计算 206
- 9.10 文献 207
- 习题 207
- 附录: 主成分 208

第 10 章 回归系数的有偏估计 212

- 10.1 引言 212
- 10.2 主成分回归 212
- 10.3 消除预测变量间的相依性 214
- 10.4 回归系数的约束 216
- 10.5 主成分回归: 注意事项 216
- 10.6 岭回归 218
- 10.7 岭估计 220
- 10.8 岭回归: 几点说明 224
- 10.9 小结 224
- 习题 225
- 附录: 岭回归 227

第 11 章 变量选择的方法 230

- 11.1 引言 230

11.2	问题的归纳	230
11.3	剔除变量的后果	231
11.4	回归方程的用途	232
11.5	评价回归方程的准则	233
11.6	多重共线性和变量选择	234
11.7	评价所有可能的方程	235
11.8	若干变量选择方法	235
11.9	变量选择方法的一般说明	236
11.10	主管人员业绩的研究	237
11.11	共线性数据的变量选择	240
11.12	凶杀数据	240
11.13	运用岭回归选择变量	243
11.14	大气污染研究中的变量选择	244
11.15	拟合回归模型的可能策略	249
11.16	文献	251
	习题	251
	附录: 误设模型的影响	254

第 12 章 Logistic 回归 257

12.1	引言	257
12.2	定性数据的建模	257
12.3	Logit 模型	258
12.4	例子: 破产概率的估计	259
12.5	Logistic 回归诊断	262
12.6	变量选择	264
12.7	Logistic 回归拟合程度的判断	265
12.8	分类问题: 另一种方法	266
	习题	268

附录: 统计用表 270

参考文献 279

索引 286

1

引言

1.1 什么是回归分析

回归分析是研究变量间函数关系的一种方法，其思想非常简单。房地产评估师在考虑房屋的销售价格时，通常将这一价格与该建筑的某些结构特征及购房税（地方税、教育税、县税等）联系起来。我们也会考虑香烟的消费量是否应与一些社会经济变量及人口统计变量联系起来，譬如年龄、受教育程度、收入及香烟的价格等。我们可以用方程或模型来建立响应变量（或称为相依变量）与解释变量（或称为预测变量，可以是一个，也可以为多个）之间的关系。在香烟消费量的例子中，响应变量为香烟的消费量（可用某个州在一年中所销售香烟的包数来度量），解释变量或预测变量是一些社会经济变量和人口统计变量。在房地产评估的例子中，响应变量是房屋的价格，解释变量或预测变量是房屋的结构特征和购买房屋时所缴纳的税。

以 Y 表示响应变量，以 X_1, X_2, \dots, X_p 表示预测变量，其中 p 是预测变量的个数， Y 与 X_1, X_2, \dots, X_p 的真正关系可近似地由下列回归模型刻画：

$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon, \quad (1.1)$$

其中 ε 是随机误差，它是模型不能精确拟合数据的原因。函数 $f(X_1, X_2, \dots, X_p)$ 描述了 Y 与 X_1, X_2, \dots, X_p 之间的关系。最简单的情形莫过于如下的线性回归模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon, \quad (1.2)$$

其中 $\beta_0, \beta_1, \dots, \beta_p$ 称为回归参数或回归系数，它们是未知常数，可通过观测数据来估计。习惯上，回归系数常用希腊字母表示。

预测变量或解释变量有时又称为独立变量、协变量、回归变量、因子或承载变量。独立变量这一名称虽然经常采用，但并不十分可取，因为实际上预测变量之间是很少相互独立的。

1.2 公用数据集

回归分析有许多应用的领域,包括经济、金融、贸易、法律、气象、医学、生物、化学、工程、物理、教育、体育、历史、社会学及其心理学等等。在1.3节中给出了一些应用的例子。回归分析提供了分析读者关注的的数据的最有效的方法。读者不妨考虑一下在工作、研究或感兴趣的领域中能用回归分析处理的问题。当然,首先必须收集相关的数据,然后再将本书随后介绍的回归分析技术应用于这些数据。为了便于读者寻找真实生活中的数据,我们在这一小节中给出大量公用数据集的一些出处链接。

许多数据集可以通过书本或因特网获得。由 Hand et al. (1994) 所著的书中包含了许多领域的数据集,这些数据集的容量均较小,适合于作为练习。而由 Chatterjee, Handcock 和 Simonoff (1995) 所著的书中给出了各个不同领域中的大量数据集,这些数据集被存放在随书附送的光盘中,也可以通过万维网获得^①。

数据集还可以在因特网的其他许多网站上获得。在下面所列的 Web 网站中,有些数据集可以直接拷贝并且粘贴至所选的统计软件包中,而有些需要下载数据文件,然后再输入至统计软件包中。有些网站还包含了与其他数据集或统计网站的进一步链接。

DASL (读作“dazzle”, Data and Story Library) 是最有趣的一个网站,包含了许多数据集,以及每个数据集的背景资料。DASL 是一个例举基本统计方法之应用的数据文件与背景的网上资料库^②,数据集所覆盖的学科范围很广。DASL 还提供了寻找数据文件相关背景资料的强有力的搜索引擎。

另一个包含数据集的 Web 网站是电子数据服务系统 (Electronic Dataset Service)^③,其中的数据集按照所采用的分析方法排列。该网站还包含了与因特网上其他数据资源的进一步链接。

最后,本书有一个 Web 网站 <http://www.ilr.cornell.edu/~hadi/RABE>,它包含了比本书中所出现的所有数据集更多的内容,在该网站可以查获这些数据集与其他数据集。

1.3 回归分析的应用举例

回归分析是应用最广泛的统计工具之一。它提供了建立变量之间函数关系的简便方法,在许多学科领域都有大量的应用。前面所述的香烟消费量和房地产评估就是其中的两例。在这一节中,我们将给出另外一些例子,以说明回归分析在现实生活中的广泛应用。这里所描述的数据集将在本书以后回归技术举例中用到,或在每一章末尾的习题中用到。

① 网址为: <http://www.stern.nyu.edu/~jsimonoff/Casebook>。

② DASL 的 Web 地址为: <http://lib.stat.cmu.edu/DASL/>。

③ 电子数据服务系统的 Web 地址为: <http://www-unix.oit.umass.edu/~statdata/>。

1.3.1 农业科学

纽约州北部地区的 DHI(Dairy Herd Improvement) 公司收集和分析牛奶产量数据, 试图构造合适的模型, 通过一些可度量的变量来预测目前牛奶的产量。响应变量(以磅计量的目前牛奶的产量)和预测变量由表 1.1 给出。样本在产奶时每月抽取一次。母牛产奶的时期称为产奶期, 产奶期数即为母牛产小牛或产奶的次数。被推荐的管理方法是, 让母牛产奶约 305 天, 休息 60 天后再开始下一次产奶。该数据集共有 199 个观测值, 它们来自 DHI 公司的牛奶产量记录。这些牛奶产量数据也可在本书的网站上获得。

表 1.1 牛奶产量数据的变量

变量	定 义
Current	本月牛奶产量 (单位: 磅)
Previous	前月牛奶产量 (单位: 磅)
Fat	牛奶中脂肪的百分比
Protein	牛奶中蛋白质的百分比
Days	本次产奶期开始至今的天数
Lactation	产奶期数
"I79"	示性变量 ("Days< 79" 时为 0, "Days> 79" 时为 1)

表 1.2 劳工就业权法数据中的变量

变量	定 义
COL	一个四人家庭的生活开支
PD	人口密度 (人/平方英里)
URate	1978 年州工会化比率
Pop	1975 年的人口
Taxes	1972 年的财产税
Income	1974 年的人均收入
RTWL	示性变量 (1 表示该州有劳工就业权法, 否则为 0)

1.3.2 劳资关系

1947 年, 美国国会通过了针对 Wagner 法案的 Taft-Hartley 修正案。原来的 Wagner 法案规定, 州法律如不禁止, 则允许工会采用只雇佣工会会员的合同 (a Closed Shop Contract) ^①。而 Taft-Hartley 修正案则规定, 采用只雇佣工会会员的合同非法, 同时也赋予各州禁止入会雇佣 ^② 的权利。这些劳工就业权法在整个劳

① 按照只雇佣工会会员的合约的规定, 所有职员在被雇佣时必须为工会会员, 且必须保持会员身份作为雇佣条件。

② 按照入会雇佣条例, 职员在被雇佣时不必是工会会员, 但必须在两个月内入会, 才允许雇佣方作出雇佣的决定。

工运动中引起了一阵关注。这里我们关心的一个问题是：这些法律对美国一个中等生活水平的四人家庭的生活开支有何影响？为回答这个问题，我们从不同的渠道收集了包括 38 个地区的一个数据集。涉及的变量定义于表 1.2。该劳工就业权法数据列于表 1.3，也能从本书的网站上找到。

表 1.3 劳工就业权法数据

城市	COL	PD	URate	Pop	Taxes	Income	RTWL
Atlanta	169	414	13.6	1790128	5128	2961	1
Austin	143	239	11	396891	4303	1711	1
Bakersfield	339	43	23.7	349874	4166	2122	0
Baltimore	173	951	21	2147850	5001	4654	0
Baton Roige	99	255	16	411725	3965	1620	1
Boston	363	1257	24.4	3914071	4928	5634	0
Buffalo	253	834	39.2	1326848	4471	7213	0
Champaign-Urbana	117	162	31.5	162304	4813	5535	0
Cedar Rapids	294	229	18.2	164145	4839	7224	1
Chicago	291	1886	31.5	7015251	5408	6113	0
Cincinnati	170	643	29.5	1381196	4637	4806	0
Cleveland	239	1295	29.5	1966725	5138	6432	0
Dallas	174	302	11	2527224	4923	2363	1
Dayton	183	489	29.5	835708	4787	5606	0
Denver	227	304	15.2	1413318	5386	5982	0
Detriot	255	1130	34.6	4424382	5246	6275	0
Green Bay	249	323	27.8	169467	4289	8214	0
Hartford	326	696	21.9	1062565	5134	6235	0
Houston	194	337	11	2286247	5084	1278	1
Indianapolis	251	371	29.3	1138753	4837	5699	0
Kansas City	201	386	30	1290110	5052	4868	0
Lancaster,PA	124	362	34.2	342797	4377	5205	0
Los Angeles	340	1717	23.7	6986898	5281	1349	0
Milwaukee	328	968	27.8	1409363	5176	7635	0
Minneapolis,St.Paul	265	433	24.4	2010841	5206	8392	0
Nashville	120	183	17.7	748493	4454	3578	1
New York	323	6908	39.2	9561089	5260	4862	0
Orlando	117	230	11.7	582664	4613	782	1
Philadelphia	182	1353	34.2	4807001	4877	5144	0
Pittsburgh	169	762	34.2	2322224	4677	5987	0
Portland	267	201	23.1	228417	4123	7511	0
St.Louis	184	480	30	2366542	4721	4809	0
San Diego	256	372	23.7	1584583	4837	1458	0
San Francisco	381	1266	23.7	3140306	5940	3015	0
Seattle	195	333	33.1	1406746	5416	4424	0
Washington	205	1073	21	3021801	6404	4224	0
Wichita	206	157	12.8	384920	4796	4620	1
Raleigh-Durham	126	302	6.5	468512	4614	3393	1

1.3.3 历史

在历史学的研究中,非常感兴趣的问题是如何根据与年代有关的特征,去推算历史物件的年代。譬如,表 1.4 中的变量可用来估计古埃及人头盖骨的年代。其中,响应变量为 Year,其他四个变量为可能的预测变量。该数据来源于文献 Thomson and Randall-Maciver(1905),也可在 Hand et al.(1994) 第 299 ~ 301 页中查到。关于这些数据的分析可查阅 Manly(1986)。该数据也可在本书的网站中找到。

表 1.4 古埃及人头盖骨年代数据中的变量

变量	定 义
Year	头盖骨形成的近似年代 (“< 0”为公元前;“> 0”为公元后)
MB	头盖骨的最大宽度
BH	头盖骨最高点的高度
BL	头盖骨基底齿槽冠的长度
NH	头盖骨的鼻梁骨高度

表 1.5 国内移民数据中的变量

变量	定 义
State	州名
NDIR	1990 ~ 1994 年的净国内移民率
Unemp	1994 年的失业率
Wage	1994 年产业工人平均每小时的报酬
Crime	1993 年每 100000 个人的暴力犯罪率
Income	1994 年家庭收入的中位数
Metrop	1992 年居住在大城市的居民占州总人口的比率
Poor	1994 年在贫困线以下人口的比例
Taxes	1993 年人均的全部州税和地方税
Educ	1990 年 25 岁及以上至少有高中学历的人口比例
BusFail	1993 年企业的破产数除以州人口总数
Temp	1993 年该州 12 个月的平均气温 (以华氏度计)
Region	州所在地域 (东北部、南部、中西部和西部)

1.3.4 政府

国内移民 (即某一州或地区的居民移至另一州或地区) 的信息对每个州或地方政府是很重要的。因此人们非常感兴趣构建一个模型,用以预测国内移民的人数或解释为什么人们从一地移至另一地。影响国内移民的因素很多,如气候、犯罪率、税收、就业率等等。我们已取得了 48 个州的一些数据,Alaska 州和 Hawaii 州的数据不在其中。因为这两个州的环境与其他州有着明显的不同,且地理位置

表 1.6 国内移民数据的前六个变量

State	NDIR	Unemp	Wage	Crime	Income	Metrop
Alabama	17.47	6.0	10.75	780	27196	67.4
Arizona	49.60	6.4	11.17	715	31293	84.7
Arkansas	23.62	5.3	9.65	593	25565	44.7
California	-37.21	8.6	12.44	1078	35331	96.7
Colorado	53.17	4.2	12.27	567	37833	81.8
Connecticut	-37.41	5.6	13.53	456	41097	95.7
Delaware	22.43	4.9	13.90	686	35873	82.7
Florida	39.73	6.6	9.97	1206	29294	93.0
Georgia	39.24	5.2	10.35	723	31467	67.7
Idaho	71.74	5.6	11.88	282	31536	30.0
Illinois	-20.87	5.7	12.26	960	35081	84.0
Indiana	9.04	4.9	13.56	489	27858	71.6
Iowa	0.00	3.7	12.47	326	33079	43.8
Kansan	-1.25	5.3	12.14	469	28322	54.6
Kentucky	13.44	5.4	11.82	463	26595	48.5
Louiaiana	-13.94	8.0	13.13	1062	25676	75.0
Maine	-9.77	7.4	11.68	126	30316	35.7
Maryland	-1.55	5.1	13.15	998	39198	92.8
Massachusetts	-30.46	6.0	12.59	805	40500	96.2
Michigan	-13.19	5.9	16.13	792	35284	82.7
Minnesota	9.46	4.0	12.60	327	33644	69.3
Mississippi	5.33	6.6	9.40	434	25400	34.6
Missouri	6.97	4.9	11.78	744	30190	68.3
Montana	41.50	5.1	12.50	178	27631	24.0
Nebraska	-0.62	2.9	10.94	339	31794	50.6
Nevada	128.52	6.2	11.83	875	35871	84.8
New Hampshire	-8.72	4.6	11.73	138	35245	59.4
New Jersey	-24.90	6.8	13.38	627	42280	100.0
New Mexico	29.05	6.3	10.14	930	26905	56.0
New York	-45.46	6.9	12.19	1074	31899	91.7
North Carolina	29.46	4.4	10.19	679	30114	66.3
North Dakota	-26.47	3.9	10.19	82	28278	41.6
Ohio	-3.27	5.5	14.38	504	31855	81.3
Oklahoma	7.37	5.8	11.41	635	26991	60.1
Oregon	49.63	5.4	12.31	503	31456	70.0
Pennsylvania	-4.30	6.2	12.49	418	32066	84.8
Rhode Island	35.32	7.1	10.35	402	31928	93.6
South Carolina	11.88	6.3	9.99	1023	29846	69.8
South Dakota	13.71	3.3	9.19	208	29733	32.6
Tennessee	32.11	4.8	10.51	766	28639	67.7
Texas	13.00	6.4	11.14	762	30775	83.9
Utah	31.25	3.7	11.26	301	35716	77.5
Vermont	3.94	4.7	11.54	114	35802	27.0
Virginia	6.94	4.9	11.25	372	37647	77.5
Washington	44.66	6.4	14.42	515	33533	83.0
West Virginia	10.75	8.9	12.60	208	23564	41.8
Wisconsin	11.73	4.7	12.41	264	35388	68.1
Wyoming	11.95	5.3	11.81	286	33140	29.7

表 1.7 国内移民数据的后六个变量

State	Poor	Taxes	Educ	BusFail	Temp	Region
Alabama	16.4	1553	66.9	0.20	62.77	South
Arizona	15.9	2122	78.7	0.51	61.09	West
Arkansas	15.3	1590	66.3	0.08	59.57	South
California	17.9	2396	76.2	0.63	59.25	West
Colorado	9.0	2092	84.4	0.42	43.43	West
Connecticut	10.8	3334	79.2	0.33	48.63	Northeast
Delaware	8.3	2336	77.5	0.19	54.58	South
Florida	14.9	2048	74.4	0.36	70.64	South
Georgia	14.0	1999	70.9	0.33	63.54	South
Idaho	12.0	1916	79.7	0.31	42.35	West
Illinois	12.4	2332	76.2	0.18	50.98	Midwest
Indiana	13.7	1919	75.6	0.19	50.88	Midwest
Iowa	10.7	2200	80.1	0.18	45.83	Midwest
Kansas	14.9	2126	81.3	0.42	52.03	Midwest
Kentucky	18.5	1816	64.6	0.22	55.36	South
Louisiana	25.7	1685	68.3	0.15	65.91	South
Maine	9.4	2281	78.8	0.31	40.23	Northeast
Maryland	10.7	2565	78.4	0.31	54.04	South
Massachusetts	9.7	2664	80.0	0.45	47.35	Northeast
Michigan	14.1	2371	76.8	0.27	43.68	Midwest
Minnesota	11.7	2673	82.4	0.20	39.30	Midwest
Mississippi	19.9	1535	64.3	0.12	63.18	South
Missouri	15.6	1721	73.9	0.23	53.41	Midwest
Montana	11.5	1853	81.0	0.20	40.40	West
Nebraska	8.8	2128	81.8	0.25	46.01	Midwest
Nevada	11.1	2289	78.8	0.39	48.23	West
New Hampshire	7.7	2305	82.2	0.54	43.53	Northeast
New Jersey	9.2	3051	76.7	0.36	52.72	Northeast
New Mexico	21.1	2131	75.1	0.27	53.37	Midwest
New York	17.0	3655	74.8	0.38	44.85	Northeast
North Carolina	14.2	1975	70.0	0.17	59.36	South
North Dakota	10.4	1986	76.7	0.23	38.53	Midwest
Ohio	14.1	2059	75.7	0.19	50.87	Midwest
Oklahoma	16.7	1777	74.6	0.44	58.36	South
Oregon	11.8	2169	81.5	0.31	46.55	West
Pennsylvania	12.5	2260	74.7	0.26	49.01	Northeast
Rhode Island	10.3	2405	72.0	0.35	49.99	Northeast
South Carolina	13.8	1736	68.3	0.11	62.53	South
South Dakota	14.5	1668	77.1	0.24	42.89	Midwest
Tennessee	14.6	1684	67.1	0.23	57.75	South
Texas	19.1	1932	72.1	0.39	64.40	South
Utah	8.0	1806	85.1	0.18	46.32	West
Vermont	7.6	2379	80.0	0.30	42.46	Northeast
Virginia	10.7	2073	75.2	0.27	55.55	South
Washington	11.7	2433	83.8	0.38	46.93	Midwest
West Virginia	18.6	1752	66.0	0.17	52.25	South
Wisconsin	9.0	2524	78.6	0.24	42.20	Midwest
Wyoming	9.3	2295	83.0	0.19	43.68	West

也对移民带来一定的障碍。响应变量取为 1990 年至 1994 年期间净移民数 (移进人数 - 移出人数) 除以该州的总人口数, 预测变量如表 1.5 所列, 详细数据见表 1.6 和表 1.7, 这些数据也可在本书的网站上找到。

1.3.5 环境科学

1976 年在一项有关水质和土地利用关系的研究中, Haith(1976) 得到了纽约州的 20 条河流流域的有关测量数据 (见表 1.8)。我们感兴趣的问题是, 河流周边地区土地的利用程度对水污染 (用平均氮浓度 (mg/升) 度量) 有何影响。数据见表 1.9, 也可在网站上找到。

表 1.8 纽约州河流水污染研究中的变量

变量	定 义
Y	春、夏、秋各季中定期采集到的样本的平均氮浓度 (mg/升)
X_1	农田覆盖率 (百分比)
X_2	森林覆盖率 (百分比)
X_3	住宅地占土地总面积的百分比
X_4	工业及商业用地占土地总面积的百分比

表 1.9 纽约州河流数据

行数	河流	Y	X_1	X_2	X_3	X_4
1	Olean	1.10	26	63	1.2	0.29
2	Cassadaga	1.01	29	57	0.7	0.09
3	Otaka	1.90	54	26	1.8	0.58
4	Neversink	1.00	2	84	1.9	1.98
5	Hackensack	1.99	3	27	29.4	3.11
6	Wappinger	1.42	19	61	3.4	0.56
7	Fishkill	2.04	16	60	5.6	1.11
8	Honeoye	1.65	40	43	1.3	0.24
9	Susquehanna	1.01	28	62	1.1	0.15
10	Chenango	1.21	28	60	0.9	0.23
11	Tioughnioga	1.33	26	53	0.9	0.18
12	West Canada	0.75	15	75	0.7	0.16
13	East Canada	0.73	6	84	0.5	0.12
14	Saranac	0.80	3	81	0.8	0.35
15	Ausable	0.76	2	89	0.7	0.35
16	Black	0.87	6	82	0.5	0.15
17	Schoharie	0.80	22	70	0.9	0.22
18	Raquette	0.87	4	75	0.4	0.18
19	Oswegatchie	0.66	21	56	0.5	0.13
20	Cohocton	1.25	40	49	1.1	0.13

1.4 回归分析的步骤

回归分析通常由下列几步组成：

- 问题的表述
- 变量的选择
- 数据的收集
- 模型的设定
- 拟合方法的选择
- 模型的拟合
- 模型的论证
- 利用获得的模型给出问题的解释

下面将就这些步骤逐一加以阐述。

1.4.1 问题的表述

通常，回归分析从正确表述问题开始，就是确定要通过分析回答哪些问题。表述问题是回归分析的第一步，并且也许是最重要的一步。其重要性在于，错误地归结问题会导致变量选择的错误、统计方法选择的错误，最终导致徒劳无获。如果我们需考察某一个雇主是否对他的某些雇员有歧视现象，比如是否有歧视妇女的现象。公司的记录可提供有关收入、资历、性别等数据来研究这一问题。文献中有几种用工歧视的定义，譬如，平均来说：(a) 妇女的收入少于同等资历的男性的收入；(b) 与同样收入的男性相比，妇女有更高的资历。为回答问题“平均来说妇女的收入是否少于同等资历男性的收入”，我们选择收入为响应变量，资历、性别为预测变量。但若要回答问题“与同样收入的男性相比，妇女是否有更高的资历”，我们则选择资历为响应变量，收入和性别为预测变量。此时变量扮演的角色发生了改变。

1.4.2 变量的选择

在问题有了明确的表述之后，下一步需要选择一组被该领域的专家认为可用于解释和预测响应变量的变量。响应变量一般用 Y 表示，解释变量或预测变量用 X_1, X_2, \dots, X_p 表示，其中 p 为预测变量的个数。若响应变量取为某地区一套家庭住宅的价格，则有关的预测变量可能为：整套住宅的面积，房屋的面积，房龄，卧室数，盥洗室间数，邻居的类型，住宅的风格和地产税等。

1.4.3 数据的收集

在我们确定了可能有关的变量之后，下一步就是采集在分析中要用到的相关的数据。有时数据可以在受控的情况下采集，这时可以设定某些非主要的因子为常数。但更多的情况下数据是在非试验性条件下采集得到的，调查者几乎不能作任何控制。无论哪种情况，采集的数据由 n 个个体的观测构成，每一个体的观测由有关的各变量的测量值构成，通常以表 1.10 的形式给出。表 1.10 中每一列表示

一个变量；每一行为一个个体（例如一套住宅）的观测，共有 $p+1$ 个值，其中一个值对应于响应变量， p 个值对应于 p 个预测变量。记号 x_{ij} 表示第 j 个变量的第 i 个观测值，即第一个下标表示第 i 个观测，第二个下标表示第 j 个变量。

表 1.10 回归分析中数据的记号

观测 序号	响应变量 Y	预测变量			
		X_1	X_2	\cdots	X_p
1	y_1	x_{11}	x_{12}	\cdots	x_{1p}
2	y_2	x_{21}	x_{22}	\cdots	x_{2p}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	y_n	x_{n1}	x_{n2}	\cdots	x_{np}

表 1.10 中的变量可分为定性和定量两类。定量变量如住宅的价格、卧室数、房龄、税收等，定性变量如邻居的类型（好邻居、坏邻居）、住宅的风格（农场风格、殖民时代风格等）。本书重点讨论响应变量为定量变量的情形。当响应变量为二值变量^①时采用的技巧是Logistic回归，将在第 12 章中介绍。在回归分析中，预测变量可以是定量的也可以是定性的。为了进行计算，定性变量将用示性变量或哑变量进行编码，这部分内容将在第 5 章中讨论。

如果所有的预测变量均为定性变量，则可用方差分析的技巧来分析数据。尽管方差分析可以单独讲解^②，但如在第 5 章中所述，我们可将其看成回归分析的特例。若预测变量既有定性的又有定量的，此时的回归分析称为协方差分析。

1.4.4 模型的设定

响应变量和预测变量之间关系的模型形式，可首先由该领域的专家根据他们的知识或主观的、客观的判断提出。然后通过分析收集到的数据，对假定的模型予以确认或否决。要注意的是：此时我们只需设定模型的形式，模型中仍有若干未知参数。我们需要选择 (1.1) 中函数 $f(X_1, X_2, \cdots, X_p)$ 的适当形式。函数可分为两类：线性和非线性。如

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad (1.3)$$

为线性的，而

$$Y = \beta_0 + e^{\beta_1 X_1} + \varepsilon \quad (1.4)$$

则为非线性的。特别值得注意的是，这里的线性（或非线性）并不是指 Y 和 X_1, X_2, \cdots, X_p 之间的关系是线性的（或非线性的），而是指方程中的回归参数是线性

① 只取两个可能值的变量称为二值变量，比如 yes 或 no, 1 或 0, 成功或失败等。

② 可参阅下列书籍：Scheffe (1959)、Iversen (1976)、Wildth and Ahtola (1978)、Krishnaiah (1980)、Iversen and Norpoth (1987)、Lindman (1992)、Christensen(1996)。

(或非线性)的。如下面的两个模型均为线性的:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon,$$

$$Y = \beta_0 + \beta_1 \log X + \varepsilon.$$

虽然 Y 和 X 之间的关系是非线性的,但是在每一种情况中参数是线性的。在第一个模型中,若记 $X_1 = X$, $X_2 = X^2$,第二个模型中,记 $X_1 = \log X$,则上述两个模型可以表示为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon,$$

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon.$$

在这里我们作了变量变换,这部分内容将在第 6 章中详细阐述。那些可以通过变换转化成线性函数的非线性函数称为可线性化函数。因此,这里讨论的线性模型覆盖面较广,因为它包含了所有可线性化的函数。但并非所有的非线性函数都可以线性化,譬如 (1.4) 就是不可线性化的例子。有些作者将那些不能线性化的非线性函数称为本质非线性函数。

只包含一个预测变量的回归方程称为简单回归方程,若回归方程包含的预测变量多于一个,则称为多元回归方程。譬如,在机器修理中,研究修理时间与需修理的元件个数之间的关系便是简单回归的一个例子。该例子中有一个响应变量(修理时间),一个预测变量(元件个数)。一个非常复杂的多元回归的例子是研究不同地区的经年龄修正后的死亡率(响应变量)与许多环境与社会经济因素(预测变量)之间的关系。本书将讨论这两类问题,上面两个例子在本书中将作较详细的讨论,分别见于第 2 章和第 11 章。

在某些应用中,响应变量可能有多个 Y_1, Y_2, \dots, Y_q , 它们均与同一组预测变量 X_1, X_2, \dots, X_p 有关。比如 Bartlett, Stewart and Abrahamowicz(1998) 给出了 148 个健康人的数据集,共 11 个变量,其中 6 个变量反映的是不同类型的感觉阈值(如颤动、手或脚的温度),另 5 个变量是基准协变量(如年龄、性别、身高、体重等)。这 5 个变量可能对 6 个感觉阈值变量或其中的部分有系统的影响。因此在这一例子中共有 6 个响应变量,5 个预测变量。这一数据集称为 QST (quantitative sensory testing),由于较大(有 148 个观测),因此在此不予列出,但可在本书的网站上找到。关于这一数据集的进一步描述和研究可参阅 Barlett, Stewart and Abrahamowicz(1998)。

当响应变量只有一个时,称这样的回归分析为单变量回归,当响应变量多于一个时,称为多变量回归。但单变量回归和简单回归并非一回事,同样多变量回归也并非多元回归。简单回归和多元回归的差别在于预测变量的个数不同(简单回归只有一个预测变量,而多元回归有两个或更多个预测变量),而单变量回归和多变量回归的不同在于响应变量的个数不同(单变量回归只有一个响应变量,多变量回归有两个或多个响应变量)。本书只涉及单变量回归(既有简单回归也有多元回归,既有线性,也有非线性的),多变量回归的论述可参阅有关多元分析的书

籍, 如 Rencher(1995)、Johnson and Wichern (1992) 以及 Johnson(1998), 本书中的回归即是指单变量回归。

我们将上面讨论的回归分析的分类列于表 1.11 中。

表 1.11 回归分析的分类

回归的类型	条 件
单变量	只有一个定量的响应变量
多变量	有两个或多于两个定量的响应变量
简单	只有一个预测变量
多元	有两个或多于两个预测变量
线性	方程中参数为线性或经变换后为线性
非线性	响应变量与一些预测变量之间的关系是非线性的, 或一些参数非线性且不能通过变换化成线性
方差分析	所有预测变量均为定性的
协方差分析	部分预测变量为定性, 另一部分为定量
Logistic	响应变量为定性

1.4.5 拟合方法

在模型已设定且数据收集到以后, 下一步就要利用收集到的数据对模型中的参数进行估计, 即所谓的参数估计或模型拟合。最常用的估计方法是最小二乘法。在一定的假定条件下 (本书后面将作详细的讨论), 最小二乘估计有良好的性质。本书将主要讨论最小二乘法及其衍生方法 (如加权最小二乘法等)。但在某些情况下 (譬如某些假定条件不成立), 其他方法会优于最小二乘法, 譬如在本书中论及的有最大似然法、岭估计法、主成分估计法等。

1.4.6 模型的拟合

接下来便是用选定的估计方法 (譬如最小二乘法) 去估计回归参数或拟合模型。在 (1.1) 中, 回归参数 $\beta_0, \beta_1, \dots, \beta_p$ 的估计记为 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, 于是估计的回归方程为

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p. \quad (1.5)$$

\hat{Y} (在参数上方加 “^” 表示参数的估计, 读作 Y -hat) 的值称为拟合值。由 (1.5) 可得到 n 个拟合值, 如第 i 个拟合值 \hat{y}_i 为

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \dots, n, \quad (1.6)$$

其中 $x_{i1}, x_{i2}, \dots, x_{ip}$ 是第 i 个观测中 p 个预测变量的值。

(1.5) 还可用于预测对应于任意预测变量值 (不在所观测到的数据中) 的响应变量值, 此时得到的 \hat{Y} 称为预测值。拟合值与预测值的差别在于, 拟合值指的是 n 组观测数据中的某一组预测变量值对应的 \hat{Y} , 而预测值是任一组预测变量值对

应的 \hat{Y} 。一般不建议对远离观测数据的预测变量值预测其对应的响应变量值。如果预测变量值表示的是未来的情况,那么称预测值为预报值。

1.4.7 模型的论证与选择

统计方法(如回归分析)的合理性常依赖于所作的假定,这些假定通常是对数据所作的,或对模型所作的。假定的合理性对分析的准确性以及通过分析所获结论的准确性是至关重要的。在我们利用(1.5)之前,首先要考察假定是否成立。我们有必要提出以下几个问题:

1. 需要作哪些假定?
2. 如何验证每一个假定的合理性?
3. 当某一个或几个假定不成立时,怎么办?

本书的各章将对上述问题作详细的讨论。这里要强调指出的是,假定的合理性的验证,必须在作出任何分析结论之前进行。回归分析可视为一个循环的过程,在这一过程中,回归输出的结果又用于回归诊断、验证、评判,并有可能修正回归输入。这一过程有时需重复若干次,直至获得满意的输出结果。这里所谓满意的输出结果,是指得到的模型满足假定且能很合适地拟合数据。循环的过程可以用图 1.1 表示。

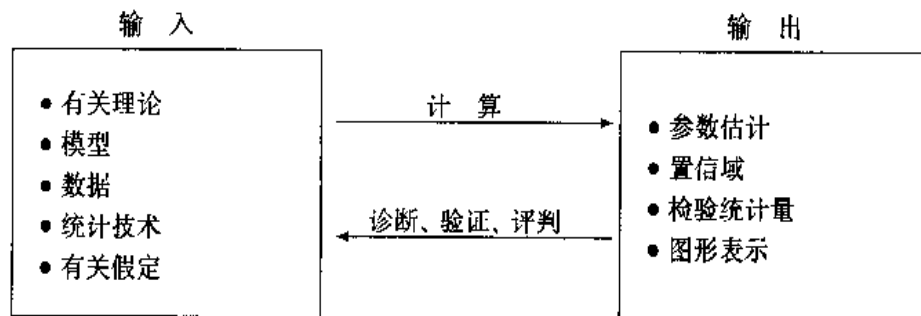


图 1.1 循环回归过程的图形表示

1.4.8 回归分析的目标

确定回归方程是回归分析最重要的结果。回归方程概括了 Y (响应变量)和一组预测变量 X_1, X_2, \dots, X_p 之间的关系。回归方程有许多的用途,可用于评价单个预测变量的重要性,用于分析改变预测变量的值达到的效果,或用于预测一组给定的预测变量值所对应的响应变量值。尽管回归方程是回归分析的最终产品,但同时也可获得许多重要的副产品。我们将回归分析作为有助于我们理解某一特定环境中变量间相互关系的一套数据分析技术。回归分析的任务是通过数据尽可能多地去了解这一环境。需要强调的是,在得到回归方程过程中的发现往往和最终的方程同样有价值。

1.5 本书的内容及结构

本书适合于各类数据分析人员。读者不需要掌握矩阵代数的知识,当然,掌握了矩阵代数的知识无疑有助于理解这一理论。熟悉矩阵代数的读者可参阅附录中的内容。利用矩阵可将回归的结果表达得更为简练,对数学推导也至关重要。

没有矩阵代数的知识也不会影响本书的使用及做回归分析。对于那些不熟悉矩阵代数但又希望了解附录内容的读者,建议先阅读一本由 Hadi (1996) 写的篇幅较短的书《Matrix Algebra as a Tool》,相信绝大多数的读者能独立地或在少量帮助下读完此书。

本书没有数学推导,对数学推导感兴趣的读者可参阅其他的回归分析教材。在本书中我们列出了一些公式,仅供参考。另外,本书中所有的必要的统计计算都可用统计软件包^①完成。

本书的结构如下:

在第2章中首先介绍简单线性回归,第3章将简单线性回归模型推广到多元线性回归模型。在这两章中,将归结模型、指明模型假定,陈述并通过例子解释主要的理论结果。为叙述简便和教学方便,第2章和第3章中的分析和结论均基于标准的回归假定。第4章讨论模型假定的验证,以及模型的诊断及修正。

之后每一章各讨论一类特殊的回归问题。第5章研究部分或全部预测变量为定性变量的情形。第6章讨论数据的变换。第7章讨论最小二乘法的一种衍生方法——加权最小二乘法。第8章研究观测相关的情形,即所谓的自相关问题。第9章和第10章分别讨论一类所谓的共线性问题的诊断与修正。当预测变量间存在较强的相关性时,就产生了共线性问题。

第11章介绍变量选择的方法——选择最好和最吝啬模型的方法。在这一章中,在运用任何变量选择的方法之前,总假定已经作过对模型假定合理性的验证,并已对模型的不合适处作了妥善处理。

第12章也是本书的最后一章,讨论了 Logistic 回归。Logistic 回归用于处理二值响应变量问题,有着重要的实际应用。在此之前的各章中,响应变量均是定量的变量。

尽管本书第5章至第12章还可以按其他次序来介绍,只要第9章在第10章之前,第7章在第12章前即可,但我们还是建议采用本书的次序。

习 题

- 1.1 区分下列变量为定性变量还是定量变量,若为定性变量,指出其可能包含的类别。

^① 许多商业统计软件都包含回归分析。

- | | |
|------------|---------------|
| (a) 地理区域; | (b) 家庭中孩子的个数; |
| (c) 房子的价格; | (d) 种族; |
| (e) 温度; | (f) 燃料的消费; |
| (g) 就业率; | (h) 政党的选择。 |

1.2 在你所感兴趣的领域中, 给出两个能用回归分析处理数据、解决你所感兴趣问题的例子 (与本章前面所提到的例子不同)。对每个例子回答下列问题:

- (a) 你感兴趣的问题是什么?
- (b) 响应变量是什么? 预测变量是什么?
- (c) 每个变量是定性的还是定量的?
- (d) 哪一类回归 (参阅表 1.1) 可用于处理你的数据?
- (e) 给出可能的模型, 并指明参数。

1.3 在下面的一组变量中, 哪些变量可看作响应变量, 哪些可看作预测变量, 加以解释。

- (a) 汽车中的汽缸数和汽油消耗量;
- (b) SAT 考分、平均绩点及大学入学资格;
- (c) 某种商品的供应和需求;
- (d) 公司的资产、股票的回报及净销售额;
- (e) 赛跑的距离、跑完全程的时间及赛跑时的天气状况;
- (f) 体重、是否抽烟、是否得肺癌;
- (g) 孩子的身高体重、父母的身高体重、孩子的年龄、性别。

1.4 对习题 1.3 中的每一个变量,

- (a) 区分是定性的还是定量的;
- (b) 可用什么类型的回归 (参阅表 1.1) 分析数据?

2

简单线性回归

2.1 引言

我们从最简单的情况入手，即考察一个响应变量 Y 与一个预测变量 X_1 之间的关系。由于只有一个预测变量，因此将 X_1 简单地记为 X 。我们先引进协方差和相关系数这两个概念，来度量两个变量之间线性关系的方向及强度。然后构造简单的线性回归模型，并不加推导地给出关键的理论结果，但通过具体的例子加以解释。读者若对其中的数学推导感兴趣，可查阅本章结尾部分提供的文献，其中列出了有关的参考书。

2.2 协方差和相关系数

假定我们已经得到了响应变量 Y 和解释变量 X 的 n 组观测，如表 2.1 所列。我们希望度量 X 和 Y 之间关系的方向和强度，下面讨论协方差和相关系数这两个量。

表 2.1 简单回归及相关分析中使用的数据记号

观测序号	响应变量 (Y)	预测变量 (X)
1	y_1	x_1
2	y_2	x_2
\vdots	\vdots	\vdots
n	y_n	x_n

在 Y 对 X 的散点图上，在 \bar{x} 处作一条垂直线，在 \bar{y} 处作一条水平线，如图 2.1 所示。其中

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n}, \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (2.1)$$

分别是 Y 和 X 的均值。这两条线将图形分成四个象限。对每个 i ，计算下面三个量：

- $y_i - \bar{y}$ ：观测值 y_i 与平均值 \bar{y} 的偏离；
- $x_i - \bar{x}$ ：观测值 x_i 与平均值 \bar{x} 的偏离；
- $(x_i - \bar{x})(y_i - \bar{y})$ ：上述两个量的乘积。

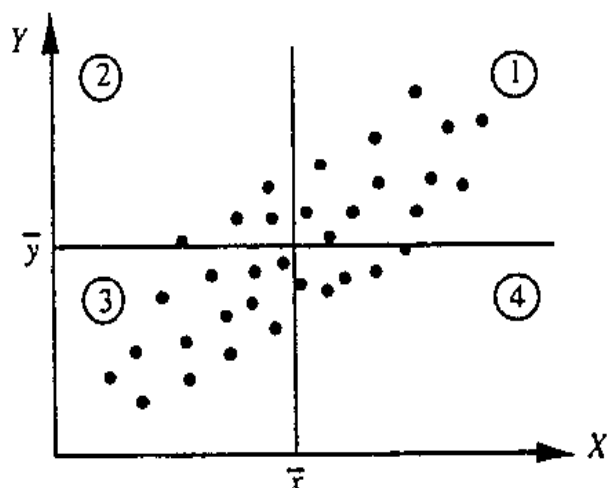


图 2.1 相关系数的图形描述

显然，对第一、第二象限中的每一点， $y_i - \bar{y}$ 为正；而对第三、第四象限中的每一点， $y_i - \bar{y}$ 为负。类似地，对第一、第四象限中的每一点， $x_i - \bar{x}$ 为正；而对第二、第三象限中的每一点， $x_i - \bar{x}$ 为负，见表 2.2。

表 2.2 $x_i - \bar{x}$ 和 $y_i - \bar{y}$ 的符号

象 限	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$
1	+	+	+
2	+	-	-
3	-	-	+
4	-	+	-

如果 Y 与 X 的线性关系是正的（即当 X 增加时， Y 也增加），则落在第一和第三象限的点应超过落在第二和第四象限的点。此时表 2.2 中的最后一列的和更可能为正，因为取正的数多于取负的数。相反，若 Y 与 X 的线性关系是负的（即当 X 增加时， Y 却减少），则落在第二和第四象限的点应超过落在第一和第三象限的点。此时表 2.2 中的最后一列的和取负的可能性更大。因此协方差

$$Cov(Y, X) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (2.2)$$

的符号反映了 Y 与 X 的线性关系的方向。若 $Cov(Y, X) > 0$ ，则表明 Y 与 X 有正的线性关系；相反，若 $Cov(Y, X) < 0$ ，则表明 Y 与 X 有负的线性关系。但

是,遗憾的是 $Cov(Y, X)$ 不能反映 Y 与 X 关系的强度,因为 $Cov(Y, X)$ 的值受测量单位的影响。比如说,我们将 Y 和 (或) X 的测量单位由美元改为千美元,则 $Cov(Y, X)$ 的取值就发生变化。为克服协方差的这一缺点,我们在计算协方差之前,首先将数据标准化。对 Y 的数据的标准化,就是首先将每个 y_i 减去均值,再除以标准差,即我们计算

$$z_i = \frac{y_i - \bar{y}}{s_y}, \quad (2.3)$$

其中

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} \quad (2.4)$$

为 Y 的样本标准差。可以证明, (2.3) 式中标准化后的变量 Z 的均值为 0, 标准差为 1。类似地,可对变量 X 标准化,即将 x_i 减去其均值 \bar{x} , 再除以标准差 s_x 。标准化后 Y 与 X 的协方差就是 Y 与 X 的相关系数,即

$$Cor(Y, X) = \frac{1}{n-1} \sum_{i=1}^n \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{x_i - \bar{x}}{s_x} \right). \quad (2.5)$$

相关系数的另一等价公式为

$$Cor(Y, X) = \frac{Cov(Y, X)}{s_y s_x} \quad (2.6)$$

$$= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 (x_i - \bar{x})^2}}. \quad (2.7)$$

因此, $Cor(Y, X)$ 既可以解释为标准化后变量的协方差,也可以解释为协方差与两个变量标准差的比。由 (2.5) 可以看出,在相关系数的定义中,相关系数是对称的,即 $Cor(Y, X) = Cor(X, Y)$ 。

与 $Cov(Y, X)$ 不同, $Cor(Y, X)$ 不受测量单位的影响。而且, $Cor(Y, X)$ 满足

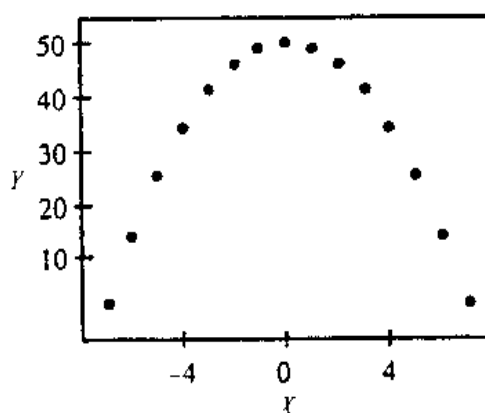
$$-1 \leq Cor(Y, X) \leq 1. \quad (2.8)$$

这些性质说明, $Cor(Y, X)$ 是一个用来度量 Y 与 X 之间相关的方向及强度的非常有用的量。 $Cor(Y, X)$ 的大小可衡量 Y 与 X 线性关系的强弱, $Cor(Y, X)$ 越靠近 1 或 -1, Y 与 X 的线性关系越强。 $Cor(Y, X)$ 的符号可反映 Y 与 X 线性关系的方向,更具体地说,若 $Cor(Y, X) > 0$, 则 Y 与 X 正相关;反之,若 $Cor(Y, X) < 0$, 则 Y 与 X 负相关。

需要注意的是,若 $Cor(Y, X) = 0$, 并不意味着 Y 与 X 没有相关性,而只表明 Y 与 X 没有线性相关性,因为相关系数只度量两个变量的线性关系。换句话说,当 Y 与 X 为非线性相关时, $Cor(Y, X)$ 仍有可能为 0。例如表 2.3 中给出的 Y 与 X 存在着精确的非线性函数关系 $Y = 50 - X^2$ (见图 2.2), 但 $Cor(Y, X) = 0$ 。

表 2.3 Y 与 X 存在完全非线性关系的数据集, 但 $Cor(Y, X) = 0$

Y	X	Y	X	Y	X
1	-7	46	-2	41	3
14	-6	49	1	34	4
25	-5	50	0	25	5
34	-4	49	1	14	6
41	-3	46	2	1	7

图 2.2 表 2.3 中 Y 对 X 的散点图

而且, 与许多别的概述性统计量一样, $Cor(Y, X)$ 还会受数据中一个或多个异常值的影响。为强调这一点, Anscombe(1973) 构造了四个数据集, 每一数据集有不同的模式, 但有相同的概述性统计量 (例如相关系数相同)。其数据和图形见表 2.4 和图 2.3, 这些数据也可在本书网站上查到^①。基于这些统计量所作的分析 (如相关系数), 将不能发现模式的差异。

从图 2.3 可以看出, 仅第一个数据集, 如图 (a), 可由线性模型来刻画; 第二个数据集, 如图 (b), 显然为非线性的, 用二次函数来刻画更好; 第三个数据集, 如图 (c), 存在着一个使拟合线的斜率和截距失真的点; 第四个数据集, 如图 (d), 不适合线性拟合, 所拟合的直线基本上由一个极端的观测确定。综上所述, 在解释 $Cor(Y, X)$ 之前, 考察 Y 与 X 的散点图是非常重要的。

^① <http://www.irl.cornell.edu/~hadi/RABE>.

表 2.4 Anscombe 的具有相同概述性统计量的四个数据集

Y_1	X_1	Y_2	X_2	Y_3	X_3	Y_4	X_4
8.04	10	9.14	10	7.46	10	6.58	8
6.95	8	8.14	8	6.77	8	5.76	8
7.58	13	8.74	13	12.74	13	7.71	8
8.81	9	8.77	9	7.11	9	8.84	8
8.33	11	9.26	11	7.81	11	8.47	8
9.96	14	8.10	14	8.84	14	7.04	8
7.24	6	6.13	6	6.08	6	5.25	8
4.26	4	3.10	4	5.39	4	12.50	19
10.84	12	9.13	12	8.15	12	5.56	8
4.82	7	7.26	7	6.42	7	7.91	8
5.68	5	4.74	5	5.73	5	6.89	8

数据来源: Anscombe(1973)。

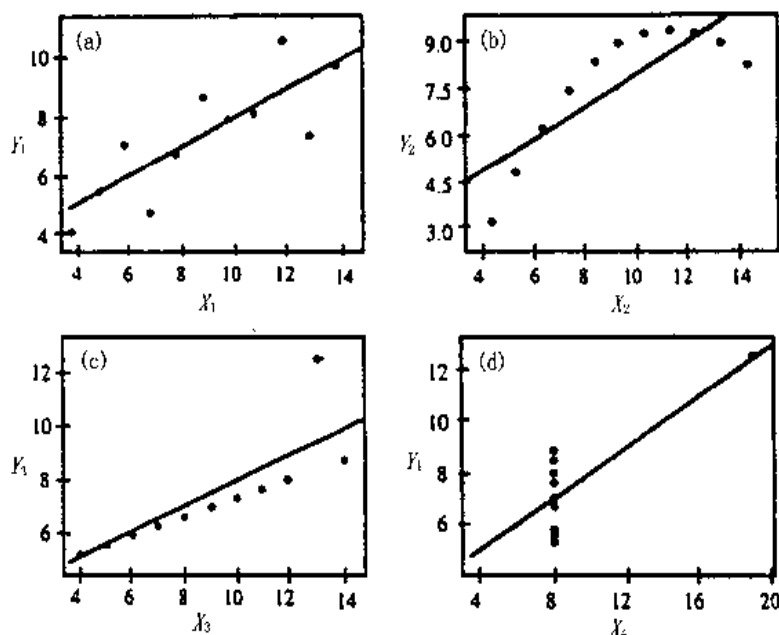


图 2.3 表 2.4 中四个数据集对应的散点图与拟合直线

2.3 例: 计算机的修理数据

以某一销售并修理小型计算机的公司为例, 我们考察修理(或服务)时间与计算机中需要修理或更换的元件个数的关系。取了修理记录的一个样本, 数据在表 2.5 中给出。其中修理时间(以分钟计)为响应变量, 需要更换的元件数为预测变量。这些数据也可在本书网站上查到。我们以这一数据集作为示例, 贯穿本章。

表 2.5 修理时间（以分钟计）及需修理元件个数

行数	修理时间	元件个数	行数	修理时间	元件个数
1	23	1	8	97	6
2	29	2	9	109	7
3	49	3	10	119	8
4	64	4	11	149	9
5	74	4	12	145	9
6	87	5	13	154	10
7	96	6	14	166	10

所要计算的量 \bar{y} 、 \bar{x} 、 $Cov(Y, X)$ 及 $Cor(Y, X)$ 列在表 2.6 中。我们有

$$\bar{y} = \frac{\sum_{i=1}^n y_i}{n} = \frac{1361}{14} = 97.21, \quad \bar{x} = \frac{\sum_{i=1}^n x_i}{n} = \frac{84}{14} = 6,$$

$$Cov(Y, X) = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{n-1} = \frac{1768}{13} = 136,$$

以及

$$Cor(Y, X) = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (x_i - \bar{x})^2}} = \frac{1768}{\sqrt{27768.36 \times 14}} = 0.996.$$

在根据 $Cor(Y, X)$ 的值作出结论之前，首先考察 Y 与 X 的散点图，见图 2.4。同时 $Cor(Y, X) \approx 0.996$ ，非常接近于 1。散点图和 $Cor(Y, X)$ 的值两者非常一致，均表明 Y 与 X 之间存在强线性关系。于是我们可以得出结论：修理时间与需修理的元件数之间存在很强的正相关关系。

虽然在度量变量间线性关系的方向及强度时 $Cor(Y, X)$ 很有用，但却不能用于预测。也就是说，我们不能在已知某一变量时，使用 $Cor(Y, X)$ 预测另一变量的值。而且， $Cor(Y, X)$ 仅仅衡量了两个变量间的关系。但是，回归分析就不同了，它可将一个或多个响应变量与一个或多个预测变量联系起来，也可以用于预测。回归分析是相关分析的一个诱人的推广，因为它通过建立模型，不仅能度量响应变量与预测变量之间关系的方向及强度，且能定量地刻画这一关系。本章此后的部分将讨论简单线性回归模型，第 3 章中讨论多元回归模型。

表 2.6 计算修理时间 Y 和需修理元件个数 X 之间相关系数所需的量

i	y_i	x_i	$(y_i - \bar{y})$	$(x_i - \bar{x})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})(x_i - \bar{x})$
1	23	1	-74.21	-5	5507.76	25	371.07
2	29	2	-68.21	-4	4653.19	16	272.86
3	49	3	-48.21	-3	2324.62	9	144.64
4	64	4	-33.21	-2	1103.19	4	66.43
5	74	4	-23.21	-2	583.90	4	46.43
6	87	5	-10.21	-1	104.33	1	10.21
7	96	6	-1.21	0	1.47	0	0.00
8	97	6	-0.21	0	0.05	0	0.00
9	109	7	11.79	1	138.90	1	11.79
10	119	8	21.79	2	474.62	4	43.57
11	149	9	51.79	3	2681.76	9	155.36
12	145	9	47.79	3	2283.47	9	143.36
13	154	10	56.79	4	3224.62	16	227.14
14	166	10	68.79	4	4731.47	16	275.14
合计	1361	84	0	0	27768.36	114	1768.00

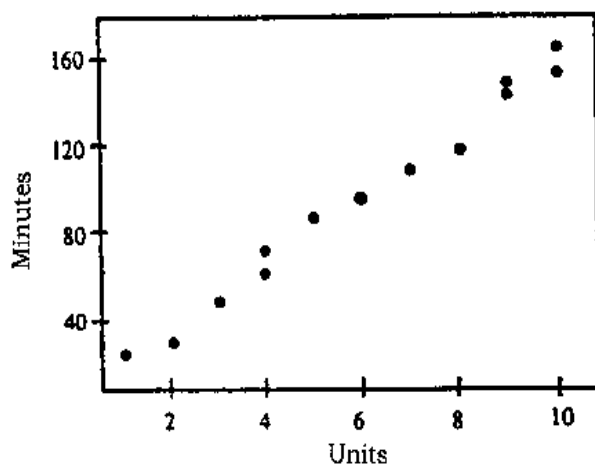


图 2.4 计算机修理数据: 修理时间与修理元件数的散点图

2.4 简单线性回归模型

假定响应变量 Y 与预测变量 X 之间的关系为如下线性^①模型

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (2.9)$$

其中 β_0, β_1 为常数, 称为回归系数或回归参数, ε 为随机干扰或误差。假定在一定的观测范围内, 线性方程 (2.9) 是 Y 与 X 之间真实关系的一个可以接受的近似。即 Y 近似是 X 的线性函数, ε 度量了近似式的偏差, 它不含任何已包含于 X 中的有关 Y 的系统信息。回归系数 β_1 称为斜率, 可解释为 X 变化一个单位时 Y 的变化。回归系数 β_0 称为常数或截距, 它是 X 取 0 时 Y 的预测值。

由 (2.9), 表 2.1 中每一观测值可写为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (2.10)$$

其中 y_i 表示响应变量 Y 的第 i 个观测值, x_i 表示预测变量 X 的第 i 个观测值, ε_i 为表示在 y_i 近似式中的误差。

回归分析与相关分析非常重要的不同点是, 相关系数具有对称性: $Cor(Y, X) = Cor(X, Y)$, 变量 X 与 Y 同等重要。而在回归分析中, 响应变量 Y 是头等重要的, 预测变量 X 的重要性依赖于它对响应变量 Y 的变异性的解释程度而并非其本身, 因此 Y 是头等重要的。

再来看计算机修理的例子。设想该公司需要预测今后几年中需要的修理工程师人数。假定可用如下的线性模型:

$$\text{修理时间 (分钟)} = \beta_0 + \beta_1 \cdot \text{修理的元件数} + \varepsilon \quad (2.11)$$

刻画修理时间与需要修理或更换的电子元件数之间的关系。这一假定的合理性, 可从响应变量关于解释变量的散点图 (图 2.4) 得以证实。图 2.4 表示用线性模型 (2.11) 是合理的。

2.5 参数估计

基于观测数据, 我们希望估计参数 β_0 和 β_1 , 等价地说, 我们要找到能最好地拟合响应变量关于预测变量的散点图 (图 2.4) 中的点的直线。我们采用最常见的最小二乘法, 该方法给出一直线, 使得每个点离开这条直线的纵向距离^②的平方和达到最小。这里纵向距离反映的是响应变量的误差。改写 (2.10) 可得:

$$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i, \quad i = 1, 2, \dots, n. \quad (2.12)$$

^① 这里“线性”有双重含义, 一方面可以解释为变量 Y 和 X 之间的关系是线性的, 另一方面也可解释为回归函数关于参数是线性的, 因此, 如将 (2.9) 改为 $Y = \beta_0 + \beta_1 X^2 + \varepsilon$, 它仍是线性模型, 尽管此时 Y 和 X 之间为平方关系。

^② 也可以用到点到直线的垂直距离 (即最短距离) 替代纵向距离, 那样得到的直线称为正交回归直线。这将在第 10 章中讨论。

于是上述距离的平方和为

$$S(\beta_0, \beta_1) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

使得 $S(\beta_0, \beta_1)$ 达到最小值的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 为

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad (2.13)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \quad (2.14)$$

这里首先给出 $\hat{\beta}_1$ 的公式, 是因为在 $\hat{\beta}_0$ 的公式中要用到 $\hat{\beta}_1$. 由于 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是通过最小二乘法获得的, 因此 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别称为 β_0 和 β_1 的最小二乘估计. 用这一方法给出的直线截距为 $\hat{\beta}_0$, 斜率为 $\hat{\beta}_1$, 每个点与它的纵向距离的平方和达最小. 这条直线也称为最小二乘回归直线, 它可表示为

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \hat{X}. \quad (2.15)$$

最小二乘回归线总是存在的, 因为我们总能找到一条使纵向距离平方和最小的直线. 但在某些情况下不唯一, 这一点我们以后会看到. 不过这种情形在应用中并不常见.

对数据中的每个观测, 我们可以计算

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \hat{x}_i, \quad i = 1, 2, \dots, n. \quad (2.16)$$

这些 \hat{y}_i 称为拟合值. 所以第 i 个拟合值 \hat{y}_i 是相应于 x_i 在最小二乘回归线上的点, 与对应的第 i 个观测的纵向距离为

$$e_i = y_i - \hat{y}_i \quad (2.17)$$

这些纵向距离 e_i 称为普通^①最小二乘残差.

利用计算机维修数据和表 2.6 中的量, 可得

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{1768}{114} = 15.509, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} = 97.21 - 15.509 \times 6 = 4.162. \end{aligned}$$

于是最小二乘回归直线为

$$\text{修理时间 (分钟)} = 4.162 + 15.509 \cdot \text{修理的元件数}. \quad (2.18)$$

^① 以区别于后面给出的其他类型的残差.

图 2.5 将最小二乘回归直线和散点图放在一起, 表 2.7 给出了 (2.16) 中的拟合值和 (2.17) 中的残差。

对 (2.18) 中的系数可作这样的解释: 常数项反映的是修理的起步时间, 约为 4 分钟。斜率反映的是每多修理或调换一个元件所花费的时间, 从数据估计约 16 分钟。譬如说, 需修理 4 个元件, 则将 4 代入 (2.18) 可得时间为 $\hat{y} = 4.162 + 15.509 \times 4 = 66.20$ 。在我们的数据集当中有两个观测, 修理的元件数为 4, 即第 4 和第 5 个, 从表 2.7 可见, 66.198 是这两个观测的拟合值。注意到第 4 和第 5 个观测响应变量的取值不同, 因此残差也不同。

比较 (2.2)、(2.7) 和 (2.13) 可知, $\hat{\beta}_1$ 亦可表示为

$$\hat{\beta}_1 = \frac{Cov(Y, X)}{Var(X)} = Cor(Y, X) \frac{s_y}{s_x}, \quad (2.19)$$

因此 $\hat{\beta}_1$ 与 $Cov(Y, X)$ 、 $Cor(Y, X)$ 有相同的符号。直观上说, 正 (负) 的斜率意味着正 (负) 相关。

表 2.7 计算机修理数据的拟合值 \hat{y}_i , 普通最小二乘残差 e_i

i	元件数	\hat{y}_i	e_i	i	元件数	\hat{y}_i	e_i
1	1	19.67	3.33	8	6	97.21	-0.21
2	2	35.18	-6.18	9	7	112.72	-3.72
3	3	50.69	-1.69	10	8	128.23	-9.23
4	4	66.20	-2.20	11	9	143.74	5.26
5	4	66.20	7.80	12	9	143.74	1.26
6	5	81.71	5.29	13	10	159.25	-5.25
7	6	97.21	-1.21	14	10	159.25	6.75

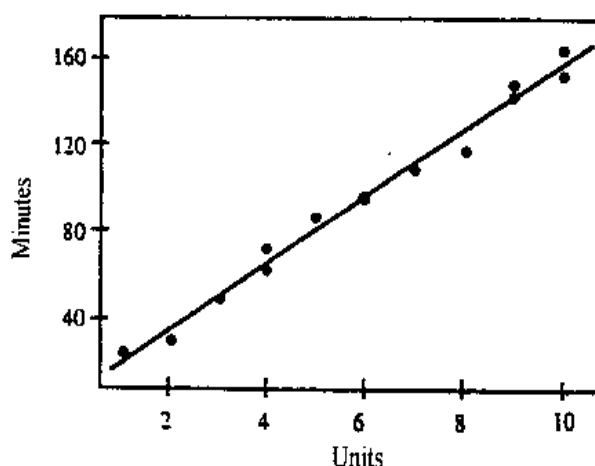


图 2.5 修理时间关于元件个数拟合的最小二乘回归直线

到目前为止, 在我们的分析中, 我们只作了一个假定, 即 Y 与 X 的关系是线性的。这一假定被称为线性假定。这仅仅是关于响应变量和预测变量之间关系的

一个假定。在做分析之前,我们首先应验证这一假定的合理性,即需要确认手头的的数据是否满足线性假定。一个非正式的办法就是考察响应变量与预测变量的散点图,更可取的是在图中添上最小二乘直线(图 2.5)。若散点图提示 Y 与 X 之间为非线性的,则应采取修正措施。譬如可先将数据作适当的变换再作分析。数据变换将放在第 6 章中介绍。

如果散点图类似一条直线,则我们可以认为线性假定是合理的,并可以作进一步的分析。当另外一些假定成立时,最小二乘估计还有其他若干性质,在第 4 章中会叙述这些假定。当然,我们也应首先验证这些假定的合理性。只有在确认了假定是合理的前提下,才可能从分析中得出有意义的结论。在第 4 章也提供了验证假定的方法。利用最小二乘估计的性质,还可以进行统计推断(如区间估计、假设检验、拟合效果的检验等),这些内容将在 2.6 节到 2.9 节中介绍。

2.6 假设检验

如前面所述,作为 Y 的预测变量, X 的作用可以通过相关系数及 Y 对 X 的散点图来反映。而一个更为正规的方式是对回归系数 β_1 作假设检验。若 $\beta_1 = 0$, 则意味着 Y 和 X 之间不存在线性关系。在数这一假设检验之前需作如下的假定:对任意固定的 X , 所有的 ε 之间相互独立,且都服从均值为 0 方差为 σ^2 的正态分布。在这些假定下, $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 分别为 β_0 和 β_1 的无偏估计^①, 它们的方差分别为

$$\text{Var}(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} \right], \quad (2.20)$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}. \quad (2.21)$$

而且, $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的抽样分布均为正态分布, 均值分别为 β_0 、 β_1 , 方差如上。

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差依赖于未知参数 σ^2 , 因此我们还需要利用数据估计 σ^2 。通常可取 σ^2 的无偏估计为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{SSE}{n-2}, \quad (2.22)$$

其中 SSE 为残差平方和。(2.22) 式分母上的 $n-2$ 称为自由度 (df), 它等于观测数减去待估计的回归系数的个数。

用 σ^2 的估计 $\hat{\sigma}^2$ 代入 (2.20) 和 (2.21) 可得 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的方差的无偏估计。一个估计量的标准差的估计称为标准误 ($s.e.$)。于是 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准误为

$$s.e.(\hat{\beta}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}, \quad (2.23)$$

^① 若对于任意 θ 都有 $E\hat{\theta} = \theta$, 则称 $\hat{\theta}$ 是 θ 的无偏估计。

$$s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum(x_i - \bar{x})^2}}, \quad (2.24)$$

其中 $\hat{\sigma}$ 是 (2.22) 式中 $\hat{\sigma}^2$ 的平方根。 $\hat{\beta}_1$ 的标准误差度量了斜率 β_1 的估计精度。标准误差越小, 估计精度越高。

利用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的抽样分布, 可以对预测变量 X 对 Y 的预测作用进行统计分析。在正态性的假定下, 对假设检验问题 $H_0: \beta_1 = 0 \longleftrightarrow H_1: \beta_1 \neq 0$, 常用的检验统计量是 t -检验

$$t_1 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}. \quad (2.25)$$

原假设 $\beta_1 = 0$ 为真时, 统计量 t_1 服从自由度为 $n-2$ 的学生氏 t 分布。具体的检验是比较 t_1 的观测值与某个合适的临界值。该临界值可从本书附录 (表 A.2) 中的 t -分布表中查到, 记为 $t_{(n-2, \frac{\alpha}{2})}$, 其中 α 是显著性水平。考虑到双边的假设检验, 所以将 α 除以 2。若

$$|t_1| \geq t_{(n-2, \frac{\alpha}{2})} \quad (2.26)$$

则在显著性水平 α 下拒绝 H_0 , 其中 $|t_1|$ 是 t_1 的绝对值。另一个与 (2.26) 等价的办法是比较 t -检验的 p -值与 α , 若

$$p(|t_1|) \leq \alpha \quad (2.27)$$

则拒绝 H_0 , 其中被称为 p -值的 $p(|t_1|)$ 是服从自由度为 $n-2$ 的学生氏 t 分布的随机变量的绝对值大于 $|t_1|$ (t -检验观测值的绝对值) 的概率。图 2.6 是 t -分布密度函数的图像, p -值是曲线下两块阴影部分的面积和。在统计软件包的回归输出结果中通常都包含 p -值。注意到拒绝 H_0 就意味着很可能 $\beta_1 \neq 0$, 也就是说预测变量 X 对于响应变量 Y 的预测作用具有统计显著性。

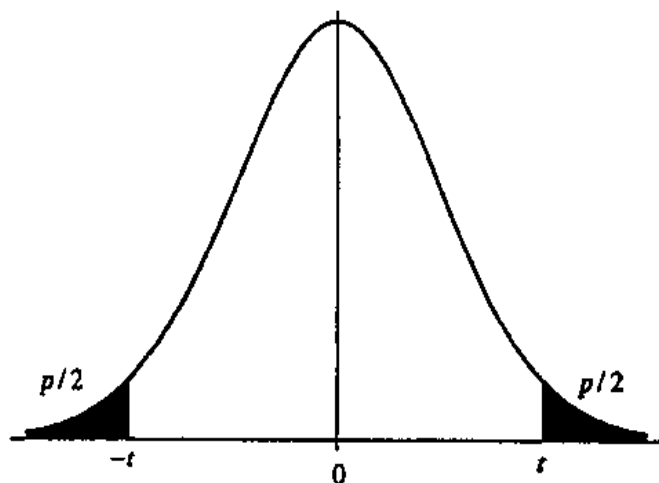


图 2.6 t -分布的概率密度函数。 t -检验的 p -值为曲线下阴影部分的面积

为了完整地描述关于回归参数的假设检验, 这里还将给出在实际应用中经常出现的关于回归参数的另三种类型的假设检验。

检验 $H_0: \beta_1 = \beta_1^0$

上面的 t -检验可以推广到更一般的情况, 检验 $H_0: \beta_1 = \beta_1^0$, 其中 β_1^0 可以是研究者自己选择的某一常数, 对应的双边备择假设为 $H_1: \beta_1 \neq \beta_1^0$ 。此时合适的检验统计量为

$$t_1 = \frac{\hat{\beta}_1 - \beta_1^0}{s.e.(\hat{\beta}_1)}. \quad (2.28)$$

注意到当 $\beta_1^0 = 0$ 时, (2.28) 退化为 (2.25)。 H_0 成立时, (2.28) 中的 t_1 仍服从自由度为 $n-2$ 的学生氏 t 分布。因此当 (2.26) 成立 (或等价地 (2.27) 成立) 时拒绝 $H_0: \beta_1 = \beta_1^0$ 。

下面以计算机修理数据为例, 对上述方法作描述。假定管理者预期每修理或更换一个元件增加的服务时间为 12 分钟, 那么所获得的数据是否支持这一猜测呢? 这一问题可通过检验假设 $H_0: \beta_1 = 12 \longleftrightarrow H_1: \beta_1 \neq 12$ 来回答。检验统计量的值为

$$t_1 = \frac{\hat{\beta}_1 - 12}{s.e.(\hat{\beta}_1)} = \frac{15.509 - 12}{0.505} = 6.948,$$

自由度为 12, 临界值为 $t_{(n-2, \frac{\alpha}{2})} = t_{(12, 0.025)} = 2.18$ 。由于 $t_1 = 6.948 > 2.18$, 结果高度显著, 于是拒绝原假设, 即管理者的估计不能得到数据的支持。这一估计偏低。

表 2.8 回归结果的标准输出 (括号内给出的是相应公式的编号)

变量	系数 (公式)	s.e. (公式)	t-检验 (公式)	p-值
常数	$\hat{\beta}_0$ (2.14)	$s.e.(\hat{\beta}_0)$ (2.23)	t_0 (2.30)	p_0
X	$\hat{\beta}_1$ (2.13)	$s.e.(\hat{\beta}_1)$ (2.24)	t_1 (2.25)	p_1

表 2.9 计算机修理数据的回归结果

变量	系数	s.e.	t-检验	p-值
常数	4.162	3.355	1.24	0.2385
元件个数	15.509	0.505	30.71	< 0.0001

检验 $H_0: \beta_0 = \beta_0^0$

实际问题也可能需要对回归参数 β_0 作检验。假定我们希望检验的问题为: $H_0: \beta_0 = \beta_0^0 \longleftrightarrow H_1: \beta_0 \neq \beta_0^0$, 其中 β_0^0 为研究者选定的某一常数。此时检验用的统计量为

$$t_0 = \frac{\hat{\beta}_0 - \beta_0^0}{s.e.(\hat{\beta}_0)}. \quad (2.29)$$

若 $\beta_0^0 = 0$, 在这一特殊场合, t_0 变为

$$t_0 = \frac{\hat{\beta}_0}{s.e.(\hat{\beta}_0)}, \quad (2.30)$$

对应的检验问题变为 $H_0: \beta_0 = 0 \longleftrightarrow H_1: \beta_0 \neq 0$.

在统计软件包中, 回归系数的最小二乘估计、它们的标准误、检验相应回归系数是否为 0 的 t -统计量值、 p -值等都会在输出结果中给出。这些值常以表格的形式给出, 如表 2.8, 该表称为回归系数表。为方便起见, 表中的值及得到其值的公式 (编号写在括号中) 一起给出了。

仍以计算机修理为例, 利用表 2.5 提供的数据, 表 2.9 给出了回归输出结果的一部分, 譬如 $\hat{\beta}_1 = 15.509$, $s.e.(\hat{\beta}_1) = 0.505$, $t_1 = 15.509/0.505 = 30.71$ 。当 $\alpha = 0.05$ 时, 临界值 $t_{(12, 0.025)} = 2.18$ 。显然 30.71 远大于 2.18, 由 (2.26) 拒绝 $H_0: \beta_1 = 0$, 此时表明预测变量 (修理的元件数) 对于响应变量 (修理时间) 的预测作用是统计显著的。利用 (2.27), 看 p -值 ($p_1 < 0.0001$), 它远小于 0.05, 也可获得同样的结论。

利用相关系数的检验

如上所述, 通过检验 $H_0: \beta_1 = 0 \longleftrightarrow H_1: \beta_1 \neq 0$, 可以判断响应变量和预测变量是否线性相关。我们采用的是 (2.25) 给出的 t -统计量。此外, 我们还可通过对 Y 和 X 的相关系数的检验达到同样的目的。设 Y 和 X 的总体相关系数为 ρ , 若 $\rho \neq 0$, 则 Y 和 X 为线性相关的。写成检验问题即为 $H_0: \rho = 0 \longleftrightarrow H_1: \rho \neq 0$, 检验用的统计量为

$$t_1 = \frac{Cor(Y, X)\sqrt{(n-2)}}{\sqrt{1-[Cor(Y, X)]^2}}, \quad (2.31)$$

其中 $Cor(Y, X)$ 是 Y 和 X 的样本相关系数, 如 (2.6) 所定义, 可用它来估计 ρ 。(2.31) 式的统计量仍服从自由度为 $n-2$ 的学生氏 t 分布。因此, 若 (2.26) 成立 (或等价地 (2.27) 成立), 则拒绝 $H_0: \rho = 0$ 。同样, 若 $H_0: \rho = 0$ 被拒绝, 则认为 Y 和 X 之间存在显著的线性关系。

显然, 若 Y 和 X 之间不存在线性关系, 则 $\beta_1 = 0$, 于是检验问题 $H_0: \beta_1 = 0$ 和 $H_0: \rho = 0$ 必须是一致的。虽然从表面上看 (2.25) 和 (2.31) 不同, 但可以证明它们在本质上是等价的。

2.7 置信区间

为了构造回归参数的置信区间, 我们仍需假定 ε 服从正态分布。如 2.6 节所讨论的那样, 这可以保证 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的抽样分布为正态。这样, β_0 的 $(1-\alpha) \times 100\%$ 置信区间为

$$\hat{\beta}_0 \pm t_{(n-2, \frac{\alpha}{2})} \times s.e.(\hat{\beta}_0), \quad (2.32)$$

其中 $t_{(n-2, \frac{\alpha}{2})}$ 为自由度为 $n-2$ 的 t -分布的 $1-\frac{\alpha}{2}$ 的分位点。类似地, β_1 的 $(1-\alpha) \times 100\%$ 置信区间为

$$\hat{\beta}_1 \pm t_{(n-2, \frac{\alpha}{2})} \times s.e.(\hat{\beta}_1). \quad (2.33)$$

(2.33) 中的置信区间有着通常的解释。这就是说对相同的 X , 用同样大小的样本重复试验, 并用每个样本分别构造斜率参数 β_1 的 95% 置信区间, 那么其中约有 95% 的置信区间包含真参数。

由表 (2.9) 可知, β_1 的 95% 置信区间为

$$15.509 \pm 2.18 \times 0.505 = (14.408, 16.610). \quad (2.34)$$

这说明每损坏一个元件, 需要再花 14 到 17 分钟修复。该例中 β_0 的置信区间的计算留给读者。

注意 (2.32) 和 (2.23) 中的置信区间是对参数 β_0 和 β_1 分别构造的, 不表示这两个参数同时 (或联合) 置信域是矩形。事实上, 联合置信域是椭圆。在第 3 章的附录 (A.13) 中就将更一般的情形, 即多元回归的情形给出置信域, 而此处关于 β_0 和 β_1 的同时置信域只是那里的特例。

2.8 预测

拟合的回归方程可用于预测, 我们将区分如下两种预测:

1. 对于某一选定的预测变量值 x_0 , 预测相应的响应变量 Y 的值;
2. 当 $X = x_0$ 时, 估计响应的均值 μ_0 。

对于第一种情形, 预测值 \hat{y}_0 为

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0, \quad (2.35)$$

预测的标准误为

$$s.e.(\hat{y}_0) = \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}. \quad (2.36)$$

因此, 预测值的置信系数为 $1-\alpha$ 的置信限为

$$\hat{y}_0 \pm t_{(n-2, \frac{\alpha}{2})} s.e.(\hat{y}_0). \quad (2.37)$$

对第二种情形, 响应均值 μ_0 的估计为

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0. \quad (2.38)$$

这一估计的标准误为

$$s.e.(\hat{\mu}_0) = \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum (x_i - \bar{x})^2}}, \quad (2.39)$$

μ_0 的置信系数为 $1 - \alpha$ 的置信限为

$$\hat{\mu}_0 \pm t_{(n-2, \frac{\alpha}{2})} s.e.(\hat{\mu}_0). \quad (2.40)$$

由 (2.35) 和 (2.38) 可见, μ_0 的点估计与响应的预测值 \hat{y}_0 一致。但比较 (2.36) 与 (2.39) 可知, $\hat{\mu}_0$ 的标准误小于 \hat{y}_0 的标准误。这从直观上很容易理解, 因为在取定 $X = x_0$ 时, 预测一个观测值的不确定性大于估计响应均值的不确定性。隐含在响应均值中的“平均”降低了估计的变异性和不确定性。

为区别 (2.37) 和 (2.40), (2.37) 给出的置信限常称为预测限, 而将 (2.40) 给出的称为置信限。

如果我们希望预测当有 4 个元件需要修理时的服务时间, 用 \hat{y}_4 表示其预测值, 则由 (2.35) 得

$$\hat{y}_4 = 4.162 + 15.509 \times 4 = 66.20,$$

由 (2.36) 得到的标准误为

$$s.e.(\hat{y}_4) = 5.392 \sqrt{1 + \frac{1}{14} + \frac{(4-6)^2}{114}} = 5.67.$$

另一方面, 若我们想估计这一修理所需的期望服务时间, 则应分别使用 (2.38) 和 (2.39)。用 μ_4 表示该修理所需的期望服务时间, 则

$$\hat{\mu}_4 = 4.162 + 15.509 \times 4 = 66.20,$$

标准误为

$$s.e.(\hat{\mu}_4) = 5.392 \sqrt{\frac{1}{14} + \frac{(4-6)^2}{114}} = 1.76.$$

利用上述标准误, 可由 (2.37) 及 (2.40) 构造置信区间。

由 (2.36) 可见, 预测的标准误随着 x_0 与预测变量的中心值 \bar{x} 之间的距离增大而增大。因而当所需修理的元件数与观测到的预测变量值偏离较大时, 用来预测修理所需时间应特别当心。这时预测有两种危险。第一, 由于标准误较大, 因而有很大的不确定性。更严重的是: 我们获得的线性关系在 X 的观测范围之外也许并不成立。因此, 对远离预测变量观测范围以外应用所拟合的回归直线要格外小心。譬如我们不会用前面计算机修理例子中获得的线性关系去预测当有 25 个电子元件需修理时需要的服务时间, 因为 25 离 X 的观测数据太远。

2.9 拟合效果度量

当 Y 关于 X 的线性方程拟合完成后, 我们不仅想知道 Y 和 X 之间是否存在线性关系, 而且需要度量拟合的效果。拟合的效果可通过下列方式之一来评判。

1. 当我们使用 (2.25) 或 (2.31) 来做假设检验时, 若拒绝 H_0 , 检验统计量的值 (或相应的 p -值) 的大小, 能够告诉我们 Y 与 X 之间线性关系的强度。 $|t|$ 的

值越大或 p -值越小, 则 Y 与 X 之间的线性关系越强。尽管这些检验是客观的, 但需要前面所作的那些假定, 特别是 ε 的正态性假定。

2. Y 与 X 之间线性关系的强度亦可以通过考察 Y 与 X 的散点图及 (2.6) 中相关系数 $Cor(Y, X)$ 的值直接评估。点离直线越近 (或 $Cor(Y, X)$ 越接近于 1 或 -1), 则线性关系越强。这一方法是不正规的且是主观的, 但只需线性假定。

3. 考察 Y 与 \hat{Y} 的散点图。点离直线越近, 表示 Y 与 X 的线性关系越强。我们还可以计算 Y 与 \hat{Y} 的相关系数

$$Cor(Y, \hat{Y}) = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2 \sum(\hat{y}_i - \bar{\hat{y}})^2}} \quad (2.41)$$

来看该图中线性关系的强度, 其中 \bar{y} 是响应变量 Y 的均值, $\bar{\hat{y}}$ 是拟合值的均值。事实上, Y 对 \hat{Y} 的散点图是多余的, 因为它与 Y 对 X 的散点图是一致的。相应的相关系数满足

$$Cor(Y, \hat{Y}) = |Cor(Y, X)|. \quad (2.42)$$

注意到 $Cor(Y, \hat{Y})$ 不能为负 (为什么?), 但 $Cor(Y, X)$ 可能为正也可能为负。因此在简单线性回归中, Y 对 \hat{Y} 的散点图是多余的。但是, 在多元线性回归中, Y 对 \hat{Y} 的散点图就不再是多余的, 而且这一散点图非常有价值。在第 3 章中我们会看到, 这样的散点图可以用来评判 Y 与一组预测变量 X_1, X_2, \dots, X_p 间的线性关系的强度。

4. 虽然在简单线性回归中 Y 对 \hat{Y} 的散点图和 $Cor(Y, \hat{Y})$ 是多余的, 但无论在简单线性回归还是多元线性回归中, 它们都可以用来反映拟合的效果。而且, $Cor(Y, \hat{Y})$ 还与衡量线性模型对观测数据拟合效果的另一个有用的量度密切相关。下面我们将仔细介绍。当我们获得了线性模型中参数的最小二乘估计后, 可以计算下列三个量:

$$\begin{aligned} SST &= \sum(y_i - \bar{y})^2, \\ SSR &= \sum(\hat{y}_i - \bar{y})^2, \\ SSE &= \sum(y_i - \hat{y}_i)^2, \end{aligned} \quad (2.43)$$

其中 SST 为 Y 中总的离差平方和, SSR 为回归平方和, SSE 为残差平方和。对某一给定点 (x_i, y_i) , 图 2.7 标出了 $y_i - \bar{y}$ 、 $\hat{y}_i - \bar{y}$ 和 $y_i - \hat{y}_i$ 。 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ 是基于所有数据点得到的回归直线, 在 $Y = \bar{y}$ 处画了一条水平线。对每个点 (x_i, y_i) , 有两个点与之对应, 其一为落在拟合直线上的 (x_i, \hat{y}_i) , 其二为落在水平线 $Y = \bar{y}$ 上的 (x_i, \bar{y}) 。

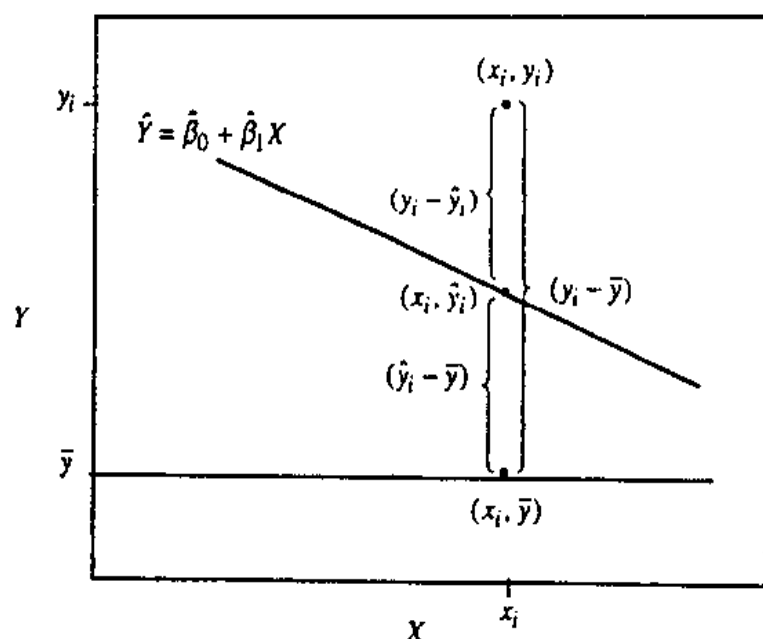


图 2.7 拟合回归直线后计算得到的各种量的图示

无论是简单线性回归还是多元线性回归, 下式总是成立的:

$$SST = SSR + SSE. \quad (2.44)$$

这是因为观测值可表示为

$$\begin{aligned} y_i &= \hat{y}_i + (y_i - \hat{y}_i), \\ \text{观测值} &= \text{拟合值} + \text{残差}. \end{aligned}$$

两边同时减去 \bar{y} , 有

$$\begin{aligned} y_i - \bar{y} &= (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i), \\ \text{观测值与均值的差} &= \text{拟合值与均值的差} + \text{残差}. \end{aligned}$$

因此, Y 的总的离差平方和可以分解成两部分: 第一部分为 SSR , 它度量了 X 的作用, 第二部分为 SSE , 它度量了预测的误差。于是比率 $R^2 = SSR/SST$ 表示在 Y 的总变差中可由预测变量 X 解释的比例。利用 (2.44), R^2 还可表示成

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}. \quad (2.45)$$

另外, 我们还可以证明

$$[Cor(Y, X)]^2 = [Cor(Y, \hat{Y})]^2 = R^2. \quad (2.46)$$

在简单线性回归中, R^2 等于 Y 与 X 的相关系数的平方, 也等于 Y 与 \hat{Y} 的相关系数的平方。因此 (2.45) 提供了相关系数平方的另一种解释。拟合效果指标

R^2 可以解释为在响应变量 Y 的全部变异中可由预测变量 X 解释的比例。由于 $SSE \leq SST$, 故 $0 \leq R^2 \leq 1$ 。若 R^2 靠近 1, 则 Y 的绝大部分变异可由 X 解释。因此 R^2 称为决定系数。 R^2 的大小反映了预测变量 X 对 Y 解释的多少。在多元线性回归的情况下, R^2 有类似的意义。

仍利用计算机修理的数据, 拟合值和残差如表 2.7, 且读者不难计算得到 $Cor(Y, X) = Cor(Y, \hat{Y}) = 0.994$, 于是 $R^2 = 0.994^2 = 0.987$ 。读者可用公式 (2.45) 得到相同的值, 此时 $SST = 27768.348$, $SSE = 348.848$, 因此

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{348.848}{27768.348} = 0.987.$$

这表明响应变量 (服务时间) 近 99% 的变异可由预测变量 (元件数) 解释。 R^2 的值说明服务时间和需修理的元件数有很强的线性关系。

我们再次强调在从统计分析 (如进行假设检验或构造置信区间、预测区间) 中作出任何结论之前, 必须先考察回归假定的合理性, 因为统计分析结论的合理性依赖于这些假定的合理性。第 4 章中给出了一些图形方法用于考察假定是否合理。我们将这些图形方法用于计算机修理数据, 没有任何迹象表明回归分析中所作的假定是不合理的。简单地讲, 这 14 个数据点给出了考察维修时间的信息。因此, 在所观测到的数据范围内, 我们相信我们所作的推断与预测是合理的。

2.10 通过原点的回归直线

前面我们考察了如下的回归模型:

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad (2.47)$$

该回归直线有截距。有时我们也有必要拟合如下的模型:

$$Y = \beta_1 X + \varepsilon, \quad (2.48)$$

显然该直线通过原点。这一模型也称为无截距模型。出于问题本身或其他客观的考虑, 有时可能要求直线必须通过原点。在选择模型 (2.47) 还是模型 (2.48) 时必须多加小心。这两种模型的拟合效果不能通过 R^2 值来比较, 而应该比较残差均方 $\hat{\sigma}^2$ 。因为这两种模型下得到的 R^2 值并不是严格可比的, 因为在第一种情形中, R^2 度量的是与样本均值的偏差, 而在第二种情形中, 度量的是与 0 的偏差。在某些情况下由 (2.48) 所得的 R^2 还可能为负。但是, 在两种模型中残差均方都反映了观测值与预测值的靠近程度。

2.11 平凡的回归模型

在这一节中, 我们给出两个平凡回归模型的例子。所谓平凡回归模型就是没有回归系数的回归方程。第一个例子, 是我们希望根据一个容量为 n 的随机样本

y_1, y_2, \dots, y_n , 对单个随机变量 Y 的均值作检验, 即检验 $H_0: \mu = 0 \longleftrightarrow H_1: \mu \neq 0$. 设 Y 是均值为 μ 方差为 σ^2 的正态随机变量, 则众所周知, 单样本 t -检验统计量

$$t = \frac{\bar{y} - 0}{s.e.(\bar{y})} = \frac{\bar{y}}{s_y/\sqrt{n}} \quad (2.49)$$

可用于检验 H_0 , 其中 s_y 为 Y 的样本标准差. 另外, 上述假设检验也可以写成

$$H_0(\text{模型 } 1): Y = \varepsilon \longleftrightarrow H_1(\text{模型 } 2): Y = \beta_0 + \varepsilon, \quad (2.50)$$

其中 $\beta_0 = \mu_0$, 因此模型 1 对应 $\mu = 0$, 而模型 2 对应 $\mu \neq 0$. 在模型 2 中, β_0 的最小二乘估计为 \bar{y} , 第 i 个拟合值 $\hat{y}_i = \bar{y}$, 第 i 个残差为 $e_i = y_i - \bar{y}$, 因此 σ^2 的估计为

$$\hat{\sigma}^2 = \frac{SSE}{n-1} = \frac{\sum (y_i - \bar{y})^2}{n-1} = s_y^2, \quad (2.51)$$

即为 Y 的样本方差. $\hat{\beta}_0$ 的标准误为 $\frac{\hat{\sigma}}{\sqrt{n}} = \frac{s_y}{\sqrt{n}}$, 即为样本均值 \bar{y} 的标准误. 检验模型 1 对模型 2 的 t -检验统计量为

$$t_1 = \frac{\hat{\beta}_0 - 0}{s.e.(\hat{\beta}_0)} = \frac{\bar{y}}{s_y/\sqrt{n}}, \quad (2.52)$$

显然 (2.51) 与 (2.49) 一致.

第二个例子与配对样本的 t -检验相关. 譬如, 考察某特种饮食方式对减肥是否有效. 随机抽取 n 个人, 每个人在规定的时间内进行. 对每个人在开始饮食前和饮食结束后测量体重, 分别以 Y_1, Y_2 表示. 令 $Y = Y_1 - Y_2$, 这是两个体重之差, 则 Y 为随机变量, 设均值为 μ , 方差为 σ^2 . 于是检验这一特种饮食方案是否有效等价于检验 $H_0: \mu = 0 \longleftrightarrow H_1: \mu > 0$. 根据 Y 的定义, 并假定 Y 为正态变量, 则这一配对样本的 t -检验统计量与 (2.49) 一致. 因此这种情况同样可归结为模型 (2.50), 并如 (2.52) 一样检验特种饮食对减肥是否有效.

上面两个例子表明单样本和配对样本的检验可看作为回归分析中的特例.

2.12 文 献

许多优秀的教材讲解回归分析的理论, 其中还有一些是为某些专门的学科写的. 每本书均提供了完整的理论结果. 如 Snedecor 和 Cochran (1980), Fox (1984) 及 Kmenta (1986) 所写的书, 只用了简单的代数和求和记号. 而 Searle (1971), Rao (1973), Seber (1977), Myers (1990), Sen and Srivastava (1990), Green (1993), Graybill and Lyer (1994) 及 Draper and Smith (1998) 所写的书则以矩阵代数为基础.

习 题

2.1 利用表 2.6 中的数据

- (a) 计算 $Var(Y)$ 和 $Var(X)$;
- (b) 证明 $\sum_{i=1}^n (y_i - \bar{y}) = 0$;
- (c) 证明对任一标准化的随机变量, 其均值为 0, 方差为 1;
- (d) 证明 (2.5)、(2.6) 和 (2.7) 给出的 $Cor(Y, X)$ 的三个公式是等同的;
- (e) 证明 (2.13) 和 (2.19) 中关于 $\hat{\beta}_1$ 的两个公式是等同的。

2.2 指出下列论断是否正确, 并说明理由。

- (a) $Cov(Y, X)$ 和 $Cor(Y, X)$ 可取介于 $-\infty$ 和 $+\infty$ 之间的值;
- (b) 若 $Cov(Y, X) = 0$ 或 $Cor(Y, X) = 0$, 则 Y 和 X 之间不存在任何关系;
- (c) 用最小二乘法对在 Y 关于 \hat{Y} 的散点图中的点进行拟合时, 所得直线的截距为 0, 斜率为 1。

2.3 利用表 2.9, 检验如下的假设, 取 $\alpha = 0.1$,

- (a) $H_0: \beta_1 = 15 \longleftrightarrow H_1: \beta_1 \neq 15$;
- (b) $H_0: \beta_1 = 15 \longleftrightarrow H_1: \beta_1 > 15$;
- (c) $H_0: \beta_0 = 0 \longleftrightarrow H_1: \beta_0 \neq 0$;
- (d) $H_0: \beta_0 = 5 \longleftrightarrow H_1: \beta_0 \neq 5$ 。

2.4 利用表 2.9 构造 β_0 的 99% 置信区间。2.5 利用最小二乘法, 用简单线性回归模型 $Y = \beta_0 + \beta_1 X + \varepsilon$ 拟合一组数据时, 下列任一论断都是正确的。请给出数学推导或用表 2.5 中的数据加以验证。

- (a) 普通最小二乘残差的和为 0;
- (b) (2.25) 和 (2.31) 给出的两个检验是等价的;
- (c) Y 对 \hat{Y} 的散点图和 Y 对 X 的散点图有相同的模式;
- (d) Y 和 \hat{Y} 的相关系数不为负。

2.6 利用表 2.5 中的数据及表 2.7 中的拟合值和残差值验证

- (a) $Cor(Y, X) = Cor(Y, \hat{Y}) = 0.994$;
- (b) $SST = 27768.348$;
- (c) $SSE = 348.848$ 。

2.7 对表 2.4 中的四个数据集, 验证下列几个量对应的值相同:

- (a) $\hat{\beta}_0$ 和 $\hat{\beta}_1$;
- (b) $Cor(Y, X)$;
- (c) R^2 ;
- (d) t -检验。

2.8 当用最小二乘法对一组数据拟合简单线性回归模型 $Y = \beta_0 + \beta_1 X + \varepsilon$ 时, 假定 $H_0: \beta_1 = 0$ 未被拒绝, 这就意味着模型可简单地写成 $Y = \beta_0 + \varepsilon$ 。此时 β_0 的最小二乘估计 $\hat{\beta}_0 = \bar{y}$ (你能证明这一点吗?)。

- (a) 此时普通最小二乘残差是什么?
- (b) 证明残差的和为 0。

- 2.9 设 Y 和 X 分别表示 1972 年和 1968 年在美国 19 个城市妇女的劳动就业率, 对这一数据集作回归, 输出结果如表 2.10, 且 $SSR = 0.0358$, $SSE = 0.0544$. 假设模型 $Y = \beta_0 + \beta_1 X + \varepsilon$ 满足通常的回归假定.

表 2.10 作 Y 关于 X 回归时的结果输出

变量	系数	s.e.	t-检验	p-值
常数	0.203311	0.0976	2.08	0.0526
X	0.656040	0.1961	3.35	< 0.0038
$n = 19$	$R^2 = 0.397$	$R_a^2 = 0.362$	$\hat{\sigma} = 0.0566$	d.f. = 17

- (a) 计算 $Var(Y)$ 和 $Cor(Y, X)$;
- (b) 若某一城市 1968 年的妇女劳动就业率为 45%, 那么该城市 1972 年的妇女劳动就业率估计为多少?
- (c) 若进一步假定 1968 年妇女劳动就业率的均值和方差分别为 0.5 和 0.005, 请对习题 2.9(b) 中估计的量构造 95% 的置信区间;
- (d) 构造回归直线斜率 β_1 的 95% 的置信区间;
- (e) 显著性水平为 5% 时, 对假设检验问题 $H_0: \beta_1 = 1 \longleftrightarrow H_1: \beta_1 > 1$ 作检验.
- (f) 若将 Y 和 X 的位置对调作回归, 你认为 R^2 的值会怎样?
- 2.10 也许有人会问: 是否身高相似的人易结为夫妻? 为此, 选择了若干对新婚夫妻为样本, 以 X 表示丈夫的身高, Y 表示妻子的身高, 如表 2.11 所示, 这些数据也可以在本书的网站上找到.
- (a) 计算丈夫身高和妻子身高的协方差;
- (b) 若身高计量单位以英寸取代厘米, 该协方差会如何变化?
- (c) 计算丈夫身高和妻子身高的相关系数;
- (d) 若身高计量单位以英寸取代厘米, 相关系数会如何变化?
- (e) 若每一男子均和一位比自己矮 5 厘米的女子结婚, 则相关系数为多少?
- (f) 若我们希望用一回归模型来拟合丈夫和妻子的身高, 你会选择哪一变量作为响应变量? 请给出理由.
- (g) 使用上题 (2.10(f)) 中选定的响应变量, 检验模型中斜率是否为 0;
- (h) 再次使用 (2.10(f)) 中选定的响应变量, 检验模型中截距是否为 0;
- (i) 假定用 (2.10(f)) 中选定的响应变量, 同时检验模型中的截距和斜率是否为 0;
- (j) 上述几个假设检验问题中, 你选择哪一个假设检验用以检验身高相似的人易结婚这一问题? 请给出结论.
- (k) 如果你认为上面的几个假设检验用于检验身高相似的人易结婚这一问题均不合适, 请给出合适的假设检验, 并给出你对这一假设检验的结论.
- 2.11 使用最小二乘法, 用过原点的简单线性回归模型 $Y = \beta_1 X + \varepsilon$ 对一组数据进行拟合, 此时 β_1 的最小二乘估计为 $\hat{\beta}_1 = \frac{\sum y_i x_i}{\sum x_i^2}$ (你能证明吗?).

- (a) 给出一个例子, 从理论或其他角度考虑表明用模型 (2.48) 来拟合是合理的;
 - (b) 证明残差 e_1, e_2, \dots, e_n 之和未必为 0;
 - (c) 给出一个数据集 Y 与 X 的例子说明用模型 (2.48) 拟合数据时, 所得 R^2 为负;
 - (d) 可以用什么指标来比较模型 (2.48) 和模型 (2.47) 的拟合效果?
- 2.12** 为研究某一大都市报开设周日版的可行性, 获得了 34 种报纸的平日和周日的发行量信息 (以千为单位)。数据如表 2.12 所示, 也可在本书的网站上获得。(数据来源: Gale Directory of Publications, 1994)
- (a) 构造周日发行量关于平日发行量的散点图, 该散点图是否提示两者之间存在线性关系? 你认为这种关系可能吗?
 - (b) 拟合一条回归直线, 由平日发行量去预测周日发行量。
 - (c) 给出 β_0, β_1 的 95% 置信区间。
 - (d) 周日发行量和平日发行量之间有显著关系吗? 用统计检验回答这一问题, 并写明你所作的假设检验以及结论。
 - (e) 周日发行量的变化中有多大比例可通过平日发行量来解释?
 - (f) 当平日发行量为 500000 时, 给出报纸周日发行量均值的 95% 置信区间。
 - (g) 某一正在考虑提供周日版的报纸, 平日发行量为 500000。给出该报纸周日发行量的 95% 预测区间。
 - (h) 另一正在申请出周日版的报纸, 平日发行量为 2000000, 给出该报周日版发行量的 95% 预测区间。如何将这一区间与 (2.12(g)) 中的区间估计作比较? 你认为这一预测区间精确吗?

表 2.11 丈夫 (H) 和妻子 (W) 的身高 (厘米)

行数	H	W	行数	H	W	行数	H	W
1	186	175	33	180	166	65	181	175
2	180	168	34	188	181	66	170	169
3	160	154	35	153	148	67	161	149
4	186	166	36	179	169	68	188	176
5	163	162	37	175	170	69	181	165
6	172	152	38	165	157	70	156	143
7	192	179	39	156	162	71	161	158
8	170	163	40	185	174	72	152	141
9	174	172	41	172	168	73	179	160
10	191	170	42	166	162	74	170	149
11	182	170	43	179	159	75	170	160
12	178	147	44	181	155	76	165	148
13	181	165	45	176	171	77	165	154
14	168	162	46	170	159	78	169	171
15	162	154	47	165	164	79	171	165
16	188	166	48	183	175	80	192	175
17	168	167	49	162	156	81	176	161
18	183	174	50	192	180	82	168	162
19	188	173	51	185	167	83	169	162
20	166	164	52	163	157	84	184	176
21	180	163	53	185	167	85	171	160
22	176	163	54	170	157	86	161	158
23	185	171	55	176	168	87	185	175
24	169	161	56	176	167	88	184	174
25	182	167	57	160	145	89	179	168
26	162	160	58	167	156	90	184	177
27	169	165	59	157	153	91	175	158
28	176	167	60	180	162	92	173	161
29	180	175	61	172	156	93	164	146
30	157	157	62	184	174	94	181	168
31	170	172	63	185	160	95	187	178
32	186	181	64	165	152	96	181	170

表 2.12 报纸平日的发行量与周日发行量(千份)

报纸名称	平日发行量	周日发行量
Baltimore Sun	391.952	488.506
Boston Globe	516.981	798.198
Boston Herald	355.628	235.084
Charlotte Observer	238.555	299.451
Baltimore Sun	391.952	488.506
Chicago Sun Times	537.780	559.093
Chicago Tribune	733.775	1133.249
Cincinnati Enquirer	198.832	348.744
Denver Post	252.624	417.779
Des Moines Register	206.204	344.522
Hartford Courant	231.177	323.084
Houston Chronicle	449.755	620.752
Kansas City Star	288.571	423.305
Los Angeles Daily News	185.736	202.614
Los Angeles Times	1164.388	1531.527
Miami Herald	444.581	553.479
Minneapolis Star Tribune	412.871	685.975
New Orleans Times-Picayune	272.280	324.241
New York Daily News	781.796	983.240
New York Daily Times	1209.225	1762.015
Newsday	825.512	960.308
Omaha World Herald	223.748	284.611
Orange County Register	354.843	407.760
Philadelphia Inquirer	515.523	982.663
Pittsburgh Press	220.465	557.000
Portland Oregonian	337.672	440.923
Providence Journal-Bulletin	197.120	268.060
Rochester Democrat and Chronicle	133.239	262.048
Rocky Mountain News	374.009	482.052
Sacramento Bee	273.844	338.355
San Francisco Chronicle	570.364	704.322
St. Louis Post-Dispatch	391.286	585.681
St. Paul Pioneer Press	201.860	267.781
Tampa Tribune	321.626	408.343
Washington Post	838.902	1165.567

3

多元线性回归

3.1 引言

本章将讲述一般的多元线性回归模型，给出回归分析的标准结果。但对这些结果没有数学推导，而是通过数值例子来描述。读者若对数学推导感兴趣，可参阅本书第 2 章末尾提供的参考文献。

3.2 数据的描述及模型

设有一个响应变量 Y 和 p 个预测（或解释）变量 X_1, X_2, \dots, X_p 的 n 组观测数据，通常如表 3.1 所列。 Y 和 X_1, X_2, \dots, X_p 之间的关系用如下的线性模型来刻画：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon_i \quad (3.1)$$

其中 $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ 为常数，称为偏回归系数（为简单起见称为回归系数）， ε 为随机干扰或误差。假定对任意一组固定的、落在数据范围内的 X_1, X_2, \dots, X_p ，(3.1) 较好地描述了 Y 和 X_1, X_2, \dots, X_p 间的真实关系（即 Y 近似地是 X_1, X_2, \dots, X_p 的线性函数，而 ε 为近似的偏差）。而 ε 不包含任何已含于诸 X 中的有关 Y 的系统信息。

表 3.1 多元回归分析中的数据

观测 序号	响应变量	预测变量			
	Y	X_1	X_2	\dots	X_p
1	y_1	x_{11}	x_{12}	\dots	x_{1p}
2	y_2	x_{21}	x_{22}	\dots	x_{2p}
3	y_3	x_{31}	x_{32}	\dots	x_{3p}
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	y_n	x_{n1}	x_{n2}	\dots	x_{np}

由 (3.1), 表 3.1 中的每一组观测可写成

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \cdots, n, \quad (3.2)$$

其中 y_i 为响应变量 Y 的第 i 个观测值, $x_{i1}, x_{i2}, \cdots, x_{ip}$ 是预测变量的第 i 组观测值 (表 3.1 的第 i 行), ε_i 为误差项。

多元线性回归是简单线性回归的推广, 因此本章所给出的结果本质上是第 2 章中结果的推广。我们也可以把简单线性回归看成是多元线性回归的特例, 因为简单线性回归的所有结果都可以通过取 $p = 1$, 从多元回归的结果得到。譬如, 当 $p = 1$ 时, (3.1) 和 (3.2) 分别简化为 (2.9) 和 (2.10)。

3.3 例: 主管人员业绩数据

贯穿本章, 我们将使用一项工业心理学研究中的数据来诠释那些标准的回归结果。最近在一个大型金融机构中作了一项关于雇员对其主管满意度的调查。其中一个问题设计为对主管的工作业绩的综合评价, 另外若干个问题涉及主管与其雇员间相互关系的具体方面。该研究试图解释主管的性格与雇员对其整体满意度之间的关系。起初选择了 6 个调查项目作为可能的解释变量, 表 3.2 给出了这些变量。从该表中不难发现, 这 6 个解释变量有两个主要类型: 变量 X_1, X_2 和 X_5 反映的是雇员和主管人员之间直接的人际关系, X_3 和 X_4 主要和工作有关, 变量 X_6 不是对主管的直接评价, 而是对雇员自己把握晋升机会的一般评价。

表 3.2 主管人员业绩数据的变量描述

变量	定 义
Y	对主管工作情况的总体评价
X_1	处理雇员的抱怨
X_2	不允许特权
X_3	学习新知识的机会
X_4	依据工作业绩升职
X_5	对不良表现过于吹毛求疵
X_6	提升到更好工作的速度

分析用的数据是对每个雇员通过问卷调查获得的。对每个问题的响应从“非常满意”到“非常不满意”分别给 1 至 5 分, 再将其分为两类: $\{1, 2\}$ 归为一类, 认为是“肯定”, $\{3, 4, 5\}$ 归为另一类, 认为“否定”。在该公司中随机抽取了 30 个部门, 每个部门有 35 个左右的雇员和一个主管。表 3.3 给出了每个部门中对每一问题回答“肯定”的员工的比例, 共 7 个变量的 30 个观测, 每个观测表示一个部门。我们将这一数据集称为“主管人员业绩数据”, 这一数据集可在本书的网站上找到^①。

^① <http://www.ilr.cornell.edu/~hadi/RABF>。

表 3.3 主管人员业绩数据

行数	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	43	51	30	39	61	92	45
2	63	64	51	54	63	73	47
3	71	70	68	69	76	84	48
4	61	63	45	47	54	84	35
5	81	78	56	66	71	83	47
6	43	55	49	44	54	49	34
7	58	67	42	56	66	68	35
8	71	75	50	55	70	66	41
9	72	82	72	67	71	83	31
10	67	61	45	47	62	80	41
11	64	53	53	58	58	67	34
12	67	60	47	39	59	74	41
13	69	62	57	42	55	63	25
14	68	83	83	45	59	77	35
15	77	77	54	72	79	77	46
16	81	90	50	72	60	54	36
17	47	85	64	69	79	79	63
18	65	60	65	57	55	80	60
19	65	70	46	57	75	85	46
20	50	58	68	54	64	70	52
21	50	40	33	34	43	64	33
22	64	61	52	62	66	80	41
23	53	66	52	50	63	80	37
24	40	37	42	58	50	57	49
25	63	54	42	48	66	75	33
26	66	77	66	63	88	74	72
27	78	75	58	74	80	78	49
28	48	57	44	45	51	83	38
29	85	85	71	71	77	74	55
30	82	82	39	59	64	78	39

假定变量 Y 和 6 个解释变量之间可通过下列线性模型联系起来:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + \varepsilon. \quad (3.3)$$

关于这一假定及其他有关假定的合理性的验证方法将在第 4 章中给出。

3.4 参数估计

我们希望根据可得到的数据估计 $\beta_0, \beta_1, \cdots, \beta_6$ 。如第 2 章简单线性回归一样, 我们采用最小二乘法, 即最小化误差平方和。由 (3.2), 误差可写为

$$\varepsilon_i = y_i - \beta_0 x_{i1} - \cdots - \beta_p x_{ip}, \quad i = 1, 2, \cdots, n. \quad (3.4)$$

于是误差平方和为

$$S(\beta_0, \beta_1, \cdots, \beta_p) = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 x_{i1} - \cdots - \beta_p x_{ip})^2.$$

由微积分原理可知, 使 $S(\beta_0, \beta_1, \dots, \beta_p)$ 最小化的最小二乘估计 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, 是下列方程组的解:

$$\begin{aligned} s_{11}\hat{\beta}_1 + s_{12}\hat{\beta}_2 + \dots + s_{1p}\hat{\beta}_p &= s_{y1} \\ s_{12}\hat{\beta}_1 + s_{22}\hat{\beta}_2 + \dots + s_{2p}\hat{\beta}_p &= s_{y2} \\ &\vdots \\ s_{1p}\hat{\beta}_1 + s_{2p}\hat{\beta}_2 + \dots + s_{pp}\hat{\beta}_p &= s_{yp}, \end{aligned} \quad (3.5)$$

其中

$$\begin{aligned} s_{ij} &= \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j), \quad i, j = 1, 2, \dots, p, \\ s_{yj} &= \sum_{a=1}^n (y_a - \bar{y})(x_{aj} - \bar{x}_j), \quad j = 1, 2, \dots, p, \\ \bar{x}_j &= \frac{\sum_{a=1}^n x_{aj}}{n}, \quad \bar{y} = \frac{\sum_{a=1}^n y_a}{n}, \end{aligned}$$

且

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}_1 - \hat{\beta}_2\bar{x}_2 - \dots - \hat{\beta}_p\bar{x}_p.$$

上述方程组称为正规方程组。 $\hat{\beta}_0$ 为截距或常数项的估计, $\hat{\beta}_j$ 为预测变量 X_j 的回归系数的估计。

假定上述方程组可解且有唯一解。对于熟悉矩阵的读者可参阅本章附录, 那里给出了显式解。在这里, 对如何解正规方程组不作更多介绍, 我们假定可用计算机软件能给出精确的数值解。

利用回归系数的估计 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, 可以写出用最小二乘法拟合的回归方程:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p. \quad (3.6)$$

对数据中的每一个观测, 可计算

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \dots, n. \quad (3.7)$$

它们称为拟合值。相应的普通最小二乘残差为

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n. \quad (3.8)$$

σ^2 的一个无偏估计为

$$\hat{\sigma}^2 = \frac{SSE}{n - p - 1}, \quad (3.9)$$

其中

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2 \quad (3.10)$$

为残差平方和。(3.9) 式中的分母 $n - p - 1$ 称为自由度(*d.f.*), 它等于观测数减去过待估回归系数的个数。

在一定的假定下, 最小二乘估计有许多优良的性质。至于如何验证这些假定的合理性, 我们将在第 4 章中讨论。但我们已将那些验证假定的方法用于主管人员业绩的例子, 并且未发现任何模型误设的证据。因此, 在本章以后的讨论中, 还将继续采用该例子。

最小二乘估计的性质将在 3.6 节中讨论。利用这些性质, 可以进行一些合适的统计推断(譬如区间估计、假设检验、拟合效果的检验等)。这些内容我们将在 3.7 节到 3.10 节中一一讲述。

3.5 回归系数的解释

在多元回归方程中, 回归系数的解释常常引起混乱。简单回归方程表示一条直线, 而多元回归方程表示一个平面(当预测变量为两个时)或超平面(预测变量多于两个时)。在多元回归中, 称为常系数的 β_0 , 其意义与简单回归中的一样, 即是当 $X_1 = X_2 = \cdots = X_p = 0$ 时的 Y 值。回归系数 β_j , $j = 1, 2, \cdots, p$, 有好几种解释。一种解释是: 当 X_j 变化一个单位而其他预测变量不变时, Y 相应的改变量。这一改变量与其他预测变量固定于什么值无关。但在实际中, 预测变量之间往往是关联的。固定某些预测变量同时变化其他预测变量有时是不可能的。

回归系数 β_j 也称为偏回归系数, 这是因为 β_j 反映的是预测变量 X_j 经其他预测变量调整后对响应变量 Y 的贡献。此处“调整”做何解释呢? 不失一般性, 仅就只有两个预测变量的情形加以说明。当 $p = 2$ 时模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon. \quad (3.11)$$

若变量 X_1, X_2 取为主管人员业绩数据中对应的变量, 则运用统计软件包可得如下回归方程

$$\hat{Y} = 15.3276 + 0.7803X_1 - 0.0502X_2. \quad (3.12)$$

X_1 的系数表示当 X_1 的值增加一个单位而 X_2 不变时, Y 增加 0.7803。如下面所阐述的, 这一值是经 X_2 “调整”后 X_1 的贡献。类似地, X_2 的系数表示当 X_2 的值增加一个单位而 X_1 不变时, Y 减少 0.0502。这也是经 X_1 调整后 X_2 的贡献。

这一解释是容易理解的, 若我们注意到这样的事实: 多元回归方程可通过若干简单回归方程得到。如在 (3.12) 中, X_2 的系数可以用如下的方法得到:

1. 拟合 Y 关于 X_1 的简单回归模型, 相应的残差记为 $e_{Y \cdot X_1}$, 其中 \cdot 前的量为响应变量, \cdot 后的量为预测变量, 此时回归方程为

$$\hat{Y} = 14.3763 + 0.754610X_1. \quad (3.13)$$

2. 拟合 X_2 (暂将 X_2 当作响应变量) 关于 X_1 的简单回归模型, 相应的残差记为 $e_{X_2 \cdot X_1}$, 回归方程为

$$\hat{X}_2 = 18.9654 + 0.513032X_1. \quad (3.14)$$

残差 $e_{Y \cdot X_1}$ 和 $e_{X_2 \cdot X_1}$ 如表 3.4 所示。

表 3.4 偏残差

行数	$e_{Y \cdot X_1}$	$e_{X_2 \cdot X_1}$	行数	$e_{Y \cdot X_1}$	$e_{X_2 \cdot X_1}$
1	-9.8614	-15.1300	16	-1.2912	-15.1383
2	0.3287	-0.7995	17	-4.5182	1.4269
3	3.8010	13.1224	18	5.3471	15.2527
4	-0.9167	-6.2864	19	-2.1990	-8.8776
5	7.7641	-2.9819	20	-8.1437	19.2787
6	-12.8799	1.8178	21	5.4393	-6.4867
7	-6.9352	-11.3385	22	3.5925	1.7397
8	0.0279	-7.4428	23	-11.1806	-0.8255
9	-4.2543	10.9660	24	-2.2969	4.0524
10	6.5925	-5.2604	25	7.8748	-4.6691
11	9.6294	6.8439	26	-6.4813	7.5311
12	7.3471	-2.7473	27	7.0279	0.5572
13	7.8379	6.2266	28	-9.3891	-4.2082
14	-9.0089	21.4529	29	6.4818	8.4269
15	4.5187	-4.4689	30	5.7457	-22.0340

3. 拟合上述两个残差间的简单回归模型, 响应变量取为 $e_{Y \cdot X_1}$, 预测变量取为 $e_{X_2 \cdot X_1}$, 回归方程为

$$\hat{e}_{Y \cdot X_1} = 0 - 0.0502e_{X_2 \cdot X_1}. \quad (3.15)$$

非常有趣的是, 在这一回归方程中 $e_{X_2 \cdot X_1}$ 前的系数与方程 (3.12) 中 X_2 前的系数相同, 均为 -0.0502 。事实上, 它们的标准误也相同。这在直观上如何解释呢? 在上述第一步中, 我们建立了 Y 与 X_1 的线性关系, 回归的残差是 Y 中除去 X_1 的线性贡献后剩余的部分。换句话说, 残差是 Y 中和 X_1 没有线性关系的部分。在第二步中, 用 X_2 取代 Y 做回归, 因此, 残差是 X_2 中和 X_1 没有线性关系的部分。第三步寻找 Y 的残差和 X_2 的残差的线性关系, 获得的回归系数反映了去除 X_1 对 Y 及对 X_2 的作用后, X_2 对 Y 的作用。

回归系数 β_j 是偏回归系数, 因为它反映的是 X_j 对响应变量 Y 的贡献大小, 但这一贡献是变量 Y 和 X_j 都经其他预测变量作过线性调整以后的贡献 (参见习题 3.4)。

在 (3.15) 中, 回归方程截距的估计为 0, 因为两组残差的均值都为 0 (因为残差的和为 0)。用同样的方法可以得到 (3.12) 中 X_1 的回归系数, 只要将上述三个步骤中互换 X_1 和 X_2 的位置即可, 留给读者作为练习。

由上面的讨论可以看到, 简单线性回归的系数和多元线性回归的系数是不同的, 除非各预测变量间是不相关的。在非试验数据或观测数据中, 预测变量很少

是不相关的;相反,在试验设计中,预测变量值是由研究者设计的,因而可将解释变量设置为不相关的。因此,若样本是在设计好的试验中获得的,预测变量间往往不相关,因而由此样本得到的简单线性回归的系数往往和多元线性回归的系数相同。

3.6 最小二乘估计的性质

在一些标准的回归假定下(如第4章中所述),最小二乘估计有如下的性质,熟悉矩阵代数的读者可以参阅本章附录,用矩阵的记号可以将这些性质表达得更简洁。

1. $\hat{\beta}_j, j = 0, 1, \dots, p$ 为 β_j 的无偏估计,其方差为 $\sigma^2 c_{jj}$, c_{jj} 是校正的平方和与乘积和矩阵的逆矩阵 C 的第 j 个对角元。 $\hat{\beta}_i$ 和 $\hat{\beta}_j$ 的协方差是 $\sigma^2 c_{ij}$, c_{ij} 是矩阵 C 的第 i 行第 j 列元素。在所有的线性无偏估计中,最小二乘估计具有最小方差,因此称为最佳线性无偏估计(BLUE, Best Linear Unbiased Estimator)。
2. $\hat{\beta}_j, j = 0, 1, \dots, p$ 服从均值为 β_j 方差为 $\sigma^2 c_{jj}$ 的正态分布。
3. $W = SSE/\sigma^2$ 服从自由度为 $n - p - 1$ 的 χ^2 分布,且 $\hat{\beta}_j (j = 0, 1, \dots, p)$ 和 σ^2 独立。
4. 向量 $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)$ 服从均值为 $\beta = (\beta_0, \beta_1, \dots, \beta_p)$, 协方差阵为 $\sigma^2 C$ 的 $p + 1$ 元正态分布。

这些结果可用于对单个回归系数作假设检验和构造区间估计。这些将在 3.8 节中讨论。

3.7 复相关系数

当用线性模型对一组数据进行拟合之后,需要这一拟合的适合性作评价。对这一部分的讨论可以像前一章 2.9 节中一样进行,只需推广到多元的情形即可,在此不再赘述。

变量 Y 与预测变量 X_1, X_2, \dots, X_p 之间线性关系的强弱可以通过考察 Y 和 \hat{Y} 的散点图及 Y 与 \hat{Y} 的相关系数来评价。 Y 与 \hat{Y} 的相关系数由下式给出

$$Cor(Y, \hat{Y}) = \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum (y_i - \bar{y})^2 \sum (\hat{y}_i - \bar{\hat{y}})^2}}, \quad (3.16)$$

其中 \bar{y} 是响应变量 Y 的均值, $\bar{\hat{y}}$ 是拟合值的均值。和简单线性回归的情形一样(如(2.45)), 决定系数 $R^2 = [Cor(Y, \hat{Y})]^2$ 由下式给出

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}. \quad (3.17)$$

R^2 表示响应变量 Y 的全部变差中可由 X_1, X_2, \dots, X_p 解释的部分所占的比例。

在多元回归中, $R = \sqrt{R^2}$ 称为复相关系数, 因为它度量了 Y 与 X_1, X_2, \dots, X_p 之间线性关系的强弱。

由表 3.5, 在主管人员业绩的例子中, $R^2 = 0.73$, 表示主管人员工作情况综合评价中 73% 的变差可由 6 个变量解释。

当模型较好地拟合数据时, R^2 显然很接近于 1, 因为此时观测值和预测值很接近, 于是 $\sum(y_i - \hat{y}_i)^2$ 很小。反之, 若 Y 与预测变量 X_1, X_2, \dots, X_p 不存在线性关系, 则模型不能很好地拟合数据, 观测值 y_i 最好的预测值为 \bar{y} 。即当 Y 与预测变量没什么关系时, Y 最好的预测为样本均值, 因为样本均值使离差平方和达到最小, 此时 R^2 为 0。 R^2 的大小常被用作衡量线性模型对数据拟合效果的尺度。但正如第 2 章中所指出的那样, R^2 值较大并不一定意味着模型对数据拟合得很好, 在 3.9 节中我们会作更详细的分析。

另一个和 R^2 联系紧密的量是 R_a^2 , 即所谓修正的 R^2 , 也常用它来度量拟合效果的好坏。它定义为

$$R_a^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}, \quad (3.18)$$

它可将 (3.17) 中的 SSE 与 SST 分别除以其自由度后得到。由 (3.17) 和 (3.18), 有

$$R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2). \quad (3.19)$$

通常用 R_a^2 来比较预测变量个数不同的模型的拟合效果 (在第 11 章中讨论)。通过对 R^2 的修正, R_a^2 更适合于比较变量个数不同的模型, 但 R_a^2 不能解释为 Y 的总变差中被预测变量解释的部分所占的比例。许多回归软件包会同时提供 R^2 和 R_a^2 的值。

3.8 单个回归系数的推断

利用 3.6 节中最小二乘估计的性质, 可对回归系数进行统计推断。检验 $H_0: \beta_j = \beta_j^0 \longleftrightarrow H_1: \beta_j \neq \beta_j^0$ 的统计量为

$$t_j = \frac{\hat{\beta}_j - \beta_j^0}{s.e.(\hat{\beta}_j)}, \quad (3.20)$$

其中 β_j^0 是调查者给定的某一常数。原假设 H_0 成立时, t_j 服从自由度为 $n-p-1$ 的 t -分布。通过比较上述 t_j 的观测值和合适的临界值 $t_{(n-p-1, \frac{\alpha}{2})}$ 的大小, 便可实施上述检验。临界值可查本书附录 (表 A.2) 中提供的 t -分布表得到, 其中 α 为显著性水平, α 之所以除以 2, 是因为该检验问题中备择假设是双边的。若

$$|t_j| \geq t_{(n-p-1, \frac{\alpha}{2})} \quad (3.21)$$

则在显著性水平 α 下拒绝 H_0 。等价地, 也可以比较 p -值。若

$$p(|t_j|) \leq \alpha \quad (3.22)$$

则拒绝 H_0 , 其中 $p(|t_j|)$ 是检验的 p -值, 它是服从自由度为 $n-p-1$ 的 t -分布的随机变量的绝对值大于 $|t_j|$ 的概率, 如图 2.6。许多统计软件包的回归输出都包含 p -值。

我们常常检验的是 $H_0: \beta_j^0 = 0$, 此时 t -检验可简化为

$$t_j = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}, \quad (3.23)$$

它是 $\hat{\beta}_j$ 与其标准误的比, 本章附录 (A.8) 给出了 $\hat{\beta}_j$ 的标准误 $s.e.(\hat{\beta}_j)$ 。许多统计软件包都计算回归系数的标准误, 且包含在回归输出中。

注意到拒绝 $H_0: \beta_j = 0$ 意味着 β_j 不为 0, 因此在经其他变量调整后 X_j 对响应变量 Y 来讲为统计显著的解释变量。

区间估计是另一个统计推断问题。 β_j 的置信系数为 $1-\alpha$ 的置信限为

$$\hat{\beta}_j \pm \hat{\sigma} \sqrt{c_{jj}} t_{(n-p-1, \frac{\alpha}{2})}, \quad (3.24)$$

其中 $t_{(n-p-1, \alpha)}$ 是自由度为 $n-p-1$ 的 t -分布的 $1-\alpha$ 分位点。(3.24) 给出的置信区间是单个系数 β_j 的置信区间, 回归系数的联合置信域在本章附录 (A.13) 中给出。

注意到 $p=1$ 时, (3.23)、(3.21) 和 (3.22) 分别退化到 (2.25)、(2.26) 和 (2.27) 式, 即取 $p=1$ 时, 由多元回归的结果可获得简单回归的结果。

实际中还有许多与多元回归有关的其他统计推断问题, 这些将在后面几节中继续讨论。

例: 主管人员业绩数据 (续)

现在我们以主管人员业绩数据为例来描述 t -检验。表 3.5 给出了用 6 个解释变量建立的线性模型的结果。拟合得到的回归方程为:

$$\hat{Y} = 10.787 + 0.613X_1 - 0.073X_2 + 0.320X_3 + 0.081X_4 + 0.038X_5 - 0.217X_6. \quad (3.25)$$

表 3.5 中的 t -检验值可用于检验假设: $H_0: \beta_j = 0 \longleftrightarrow H_1: \beta_j \neq 0, j = 1, 2, \dots, p$ 。从表 3.5 可以看出, 只有 X_1 的回归系数显著不为 0, 而 X_3 的回归系数接近于显著不为 0, 其他变量均不显著。各参数置信区间的构造留给读者作为练习。

表 3.5 主管人员业绩数据的回归输出

变量	系数	s.e.	t-检验	p-值
常数	10.787	11.5890	0.93	0.3616
X_1	0.613	0.1610	3.81	0.0009
X_2	-0.073	0.1357	-0.54	0.5956
X_3	0.320	0.1685	1.90	0.0699
X_4	0.081	0.2215	0.37	0.7155
X_5	0.038	0.1470	0.26	0.7963
X_6	-0.217	0.1782	-1.22	0.2356
$n = 30$	$R^2 = 0.73$	$R_a^2 = 0.66$	$\hat{\sigma} = 7.068$	$d.f. = 23$

注意到这一模型中常数项统计上也是不显著的(t -值为 0.93, p -值为 0.3616)。尽管如此, 常数项仍应该保留在模型中, 除非有非常明确的理由才能剔除它, 因为常数项反映了响应变量的基准水平。一般来说, 不显著的预测变量应从模型中剔除, 但常数项却应保留。

3.9 线性模型中的假设检验

除了单个 β 的假设检验外, 对线性模型的分析中还有若干别的假设检验问题。需考察的最常见的假设有如下几种:

1. 所有预测变量的回归系数均为 0;
2. 某几个回归系数为 0;
3. 某几个回归系数相等;
4. 回归系数满足某些特定的约束。

关于回归系数的这些不同的假设可用统一的方法来检验。我们首先讲述这个一般的方法, 然后用主管人员业绩数据来讲解具体的检验。

模型 (3.1) 称为全模型(FM, full model)。原假设是某些回归系数等于指定的值。用这些指定的值代入全模型即得到简化模型(RM, reduced model)。简化模型中待估参数的个数少于全模型中待估参数的个数。我们希望检验

$$H_0: \text{简化模型是充分的} \longleftrightarrow H_1: \text{全模型是充分的}.$$

注意到简化模型是嵌套的。所谓一组模型嵌套是指这些模型是某一较大模型的特例。关于这些嵌套模型的假设检验实质上是拿原假设对应的简化模型的拟合效果与全模型的拟合效果作比较。若简化模型与全模型有相同的拟合效果, 定义简化模型(通过设定某些 β_j 的值)的那个原假设不被拒绝。具体步骤如下:

设 \hat{y}_i 和 \hat{y}_i^* 分别为全模型和简化模型对 y_i 的预测值。模型对数据的失拟可用残差平方和来度量。全模型的残差平方和记为 $SSE(FM)$, 即

$$SSE(FM) = \sum (y_i - \hat{y}_i)^2. \quad (3.26)$$

类似地, 简化模型的残差平方和记为 $SSE(RM)$, 即

$$SSE(RM) = \sum (y_i - \hat{y}_i^*)^2. \quad (3.27)$$

在全模型中有 $p+1$ 个待估回归参数 $(\beta_0, \beta_1, \dots, \beta_p)$, 假设在简化模型中有 k 个待估参数。注意到 $SSE(RM) \geq SSE(FM)$, 两者之差为用简化模型拟合时导致残差平方和增加的量, 若它很大, 则简化模型是不充分的。我们用如下的比值来判断简化模型是否充分:

$$F = \frac{[SSE(RM) - SSE(FM)] / (p+1-k)}{SSE(FM) / (n-p-1)}. \quad (3.28)$$

这一比值称为 F -检验。在 (3.28) 中分子、分母中的残差平方和分别除以相应的自由度是为了保证所得的检验统计量具有标准的统计分布。全模型有 $p+1$ 个参数, 因此 $SSE(FM)$ 的自由度为 $n-p-1$; 类似地, 简化模型有 k 个参数, 因此 $SSE(RM)$ 的自由度为 $n-k$ 。从而 $SSE(RM) - SSE(FM)$ 的自由度为 $(n-k) - (n-p-1) = p+1-k$ 。上述 F -比在原假设为真时服从自由度为 $p+1-k$ 和 $n-p-1$ 的 F -分布。

若 F 的观测值大于自由度为 $p+1-k$ 和 $n-p-1$ 的 F -分布的 $1-\alpha$ 分位点, 则在显著性水平 α 下, 拒绝 H_0 , 即简化模型不合适。读者若对上述结果的证明感兴趣, 可参阅 Rao(1973), Searle(1971), Seber(1977) 或 Graybill(1976)。

于是, 若

$$F \geq F_{(p+1-k, n-p-1; \alpha)} \quad (3.29)$$

或等价地, 若

$$p(F) \leq \alpha \quad (3.30)$$

则拒绝 H_0 。其中 (3.29) 中的 F 是 (3.28) F -检验的观测值, $F_{(p+1-k, n-p-1; \alpha)}$ 是从 F -分布表中查得的临界值 (见本书附表 A.4 和 A.5), α 为显著性水平。 $p(F)$ 为 F -检验的 p -值, 即服从自由度为 $p+1-k$ 和 $n-p-1$ 的 F -分布的随机变量大于 (3.28) 中 F 的观测值的概率。许多统计软件包的回归输出中都包含这样的 p -值。

在本节的剩余部分中, 我们将以主管人员业绩数据为例, 对 (3.28) 这一一般的 F -检验给出若干具体的例子。

3.9.1 所有回归系数为 0 的检验

(3.28) 式中 F -检验的一个重要特例为检验是否所有的预测变量都没有解释力, 即是否它们所有的回归系数都为 0。此时简化模型和全模型分别为

$$RM: H_0: Y = \beta_0 + \epsilon, \quad (3.31)$$

$$FM: H_1: Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon. \quad (3.32)$$

全模型的残差平方和 $SSE(FM) = SSE$ 。在简化模型中, β_0 的最小二乘估计为 \bar{y} , 于是简化模型的残差平方和为 $SSE(RM) = \sum (y_i - \bar{y})^2 = SST$ 。简化模型只有—

个回归参数, 而全模型有 $p+1$ 个回归参数, 因此 (3.28) 式为

$$\begin{aligned} F &= \frac{[SSE(RM) - SSE(FM)] / (p+1-k)}{SSE(FM) / (n-p-1)} \\ &= \frac{[SST - SSE] / p}{SSE / (n-p-1)} \end{aligned} \quad (3.33)$$

因为 $SST = SSR + SSE$, 于是用 SSR 代替 $SST - SSE$ 有

$$F = \frac{SSR/p}{SSE/(n-p-1)} = \frac{MSR}{MSE}, \quad (3.34)$$

其中 MSR 为回归均方, MSE 为残差均方。(3.34) 中的 F -检验可用于检验所有预测变量的回归系数均为 0 这一假设 (常数项除外)。

表 3.6 多元回归的方差分析表 (ANOVA)

来源	平方和	d.f.	均方	F-检验
回归	SSR	p	$MSR = \frac{SSR}{p}$	$F = \frac{MSR}{MSE}$
残差	SSE	$n-p-1$	$MSE = \frac{SSE}{n-p-1}$	

(3.34) 式中的 F -检验也可用样本的复相关系数表示。事实上, 总体的回归系数均为 0 与总体的复相关系数为 0 是等价的。若以 R_p 表示样本的复相关系数, R_p 可以通过 p -个预测变量的模型 (即估计 p 个回归系数和一个截距) 对观测数据的拟合得到, 因此检验 $H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0$ 的 F -统计量可表示为

$$F = \frac{R_p^2/p}{(1-R_p^2)/(n-p-1)}, \quad (3.35)$$

其中自由度为 p 和 $n-p-1$ 。

与 F -检验有关的量一般都列在方差分析表 (ANOVA) 中, 如表 3.6。表中第一列表示响应变量的变差的来源有两类。 Y 的总的离差平方和有分解式 $\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$, 因此 Y 总变差的一部分可由预测变量解释, 即 $SSR = \sum (\hat{y}_i - \bar{y})^2$, 另一部分不可解释, 即 $SSE = \sum (y_i - \hat{y}_i)^2$ 。这平方和的分解列于表的第二列。表中第三列为相应的自由度, 第四列是均方, 即平方和除以各自的自由度, 最后一列为 (3.34) 中的 F -检验。有些统计软件包还提供 p -值 $p(F)$ 。

再回到主管人员业绩的数据。虽然回归系数的 t -检验已显示某些回归系数 (β_1 和 β_3) 显著不为 0, 但这里纯粹出于举例说明的目的, 我们再来检验是否六个解释变量均无解释能力, 即检验 $\beta_1 = \beta_2 = \cdots = \beta_6 = 0$ 。此时对应于 (3.31) 和 (3.32) 的简化模型和全模型分别为

$$RM: H_0: Y = \beta_0 + \varepsilon, \quad (3.36)$$

$$FM: H_1: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_6 X_6 + \varepsilon. \quad (3.37)$$

表 3.7 主管人员业绩数据的方差分析表 (ANOVA)

来源	平方和	d.f.	均方	F-检验
回归	3147.97	6	524.661	10.5
残差	1149.00	23	49.9565	

对全模型, 我们需要估计 7 个参数, 其中 6 个为回归系数, 一个为截距 β_0 。表 3.7 为方差分析表, $SSE(FM) = SSE = 1149$ 。在原假设成立即回归系数均为 0 的条件下, 对应的简化模型需要估计的参数只有 1 个 (β_0), 简化模型的残差平方和为

$$SSE(RM) = SST = SSR + SSE = 3147.97 + 1149 = 4296.97.$$

该值即为 $\sum(y_i - \bar{y})^2$ 的值。此时 $F = 10.5$ 。F-值也可由 (3.35) 式计算得到:

$$F = \frac{R_p^2/p}{(1 - R_p^2)/(n - p - 1)} = \frac{0.7326/6}{(1 - 0.7326)/23} = 10.50.$$

在附表 A.5 中查自由度为 6 和 23 的 F-分布表, 1% 的临界值为 3.71(3.71 通过插值得到)。因为 $F = 10.50 > 3.71$, 因此拒绝 H_0 , 即并非所有的 β 均为 0。这完全在我们的预料之中, 因为某些 t-值很大。

若单个回归系数的 t-检验是显著的, 则该 F-检验通常也是显著的。但有一种情况会令人困惑不解, 即单个回归系数的 t-检验均不显著, 但 (3.35) 给出的 F-检验是显著的。这就意味着虽然每个预测变量单独地没有显著的解释能力, 但这些预测变量作为一个整体确能显著地对响应变量作出解释。如果这种情况发生, 则需要对数据作认真的分析。有可能某些解释变量之间有很强的相关性, 即出现所谓的多重共线性现象。我们将在第 9 章中详细讨论这个问题。

3.9.2 某些回归系数为 0 的检验

到目前为止, 在主管人员业绩数据的例子中, 我们始终试图用 6 个变量 X_1, X_2, \dots, X_6 对 Y 作解释。(3.34) 的 F-检验提示我们不可能所有的回归系数均为 0, 因此这些解释变量中有一个或几个与 Y 相关。现在我们感兴趣的问题是: 能否用较少的预测变量充分地解释 Y 呢? 回归分析的重要目标之一就是用尽可能少的、有意义的变量充分地解释响应变量。这有两方面的好处: 其一, 可以让我们分离出最重要的变量; 其二, 对我们所研究的过程有更为简洁的描述, 从而让我们更容易理解这一过程。“表述简约化”或所谓的“吝啬原则”是回归分析中应遵循的一个重要的指导性原则。

研究变量 Y 可否用较少的变量解释, 也就是检验某些回归系数是否为 0 的问题。至于哪些变量应选入回归方程, 如果没有过硬的理论依据, 那么前面讲过的在表 3.5 中给出的 t-检验会给我们一些提示。仍看主管人员业绩数据的例子, 假如我们希望用两个变量来解释响应变量。其中一个取自 X_1, X_2, X_5 , 它们反映的

是主管与雇员间的人际关系,另一个取自 X_3, X_4, X_6 , 不是个人性格方面的因素。在 t -检验中, X_1 和 X_3 是显著的, 因此我们考察仅用 X_1 和 X_3 是否足以解释变量 Y 。此时简化模型为

$$\text{RM: } Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon. \quad (3.38)$$

该模型对应的假设为

$$H_0: \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0. \quad (3.39)$$

包含方差分析表和系数表的回归输出如表 3.8。

表 3.8 Y 关于 X_1 和 X_3 的回归输出

方差分析表 (ANOVA)				
来源	平方和	d.f.	均方	F-检验
回归	3042.32	2	1521.1600	32.7
残差	1254.65	27	46.4685	
系数表				
变量	系数	s.e.	t-检验	p-值
常数	9.8709	7.0610	1.40	0.1735
X_1	0.6435	0.1185	5.43	< 0.0001
X_3	0.2112	0.1344	1.57	0.1278
$n = 30$	$R^2 = 0.708$	$R_a^2 = 0.686$	$\hat{\sigma} = 6.817$	d.f. = 27

简化模型的残差平方和 $SSE(\text{RM}) = 1254.65$ 。由表 3.7, 全模型的残差平方和 $SSE(\text{FM}) = 1149.00$ 。因此 (3.28) 的 F -检验的观测值为

$$F = \frac{[1254.65 - 1149]/4}{1149/23} = 0.528, \quad (3.40)$$

自由度为 4 和 23。临界值为 $F_{(4,23,0.05)} = 2.8$, 因此 F -检验并不显著, 不能拒绝 H_0 , 即认为 X_1 和 X_3 足以解释 Y 。再考察类似于第 4 章中的残差图, 以考察剔除 X_2, X_4, X_5 和 X_6 是否会引起对模型假定的偏离。在这一例子中, 残差图令人满意。于是我们得到这样的结论: 剔除变量 X_2, X_4, X_5 和 X_6 并不会影响模型的解释能力。

我们作下述注解以总结本节:

1. F -检验也可以用样本的复相关系数表述。记 R_p 为有 p -个变量的全模型对应的样本复相关系数, R_q 为有 q -个变量的简化模型对应的样本复相关系数。原假设为 $p - q$ 个指定变量的回归系数为 0。检验这一假设的 F -检验为

$$F = \frac{(R_p^2 - R_q^2)/(p - q)}{(1 - R_p^2)/(n - p - 1)}, \quad d.f. = p - q, n - p - 1. \quad (3.41)$$

在我们目前讨论的例子中, 由表 3.7 及 3.8, $n = 30, p = 6, q = 2, R_6^2 = 0.7326, R_2^2 = 0.7080$ 。将这些量代入 (3.41) 可得 $F = 0.528$, 与前述一致。

2. 当简化模型只比全模型少一个预测变量时, 比如检验 $\beta_j = 0$, (3.28) 中的 F -检验的自由度为 1 和 $n-p-1$ 。此时可以证明 (3.28) 给出的 F -检验等价于 (3.21) 给出的 t -检验。更精确地说,

$$F = t_j^2. \quad (3.42)$$

即自由度为 1 和 $n-p-1$ 的 F -分布等同于自由度为 $n-p-1$ 的 t -分布的平方。在统计理论中, 这是一个众所周知的结果 (读者可以参考本书附录中的表 A.2、A.4 和 A.5 知 $F(1, v) = t^2(v)$)。

3. 在简单回归中, 预测变量的个数是 $p = 1$, 将多元回归方差分析表 (表 3.6) 中的 p 用 1 取代, 便得到简单回归的方差分析表 (表 3.9)。表 3.9 中的 F -检验可用来检验原假设: X_1 对响应变量无解释功效, 即回归系数为 0。这与 2.6 节中 (2.25) 的 t -检验一致, 即

$$t_1 = \frac{\hat{\beta}_1}{s.e.(\hat{\beta}_1)}. \quad (3.43)$$

因此, 在简单线性回归中, F -检验与 t -检验一致, 两检验统计量间的联系为:

$$F = t_1^2. \quad (3.44)$$

表 3.9 简单回归中的方差分析表 (ANOVA)

来源	平方和	d.f.	均方	F -检验
回归	SSR	1	$MSR = \frac{SSR}{1}$	$F = \frac{MSR}{MSE}$
残差	SSE	$n - 2$	$MSE = \frac{SSE}{n - 2}$	

3.9.3 某些回归系数等同的检验

利用 3.9 节中的一般方法, 可以检验某一回归模型中的两个回归系数是否相等。在我们的例子中, 欲检验预测变量 X_1 和 X_3 的回归系数是否相等。它在假定 X_2, X_4, X_5 和 X_6 的回归系数均为 0 下进行, 于是原假设为:

$$H_0: \beta_1 = \beta_3 | (\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0). \quad (3.45)$$

此时全模型为 (已假定 $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$):

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon. \quad (3.46)$$

在原假设的条件下, 设 $\beta_1 = \beta_3 = \beta'_1$, 则简化模型可写为:

$$Y = \beta'_0 + \beta'_1 (X_1 + X_3) + \varepsilon. \quad (3.47)$$

实施这一检验的较为简便的方法是先利用模型 (3.46) 拟合数据, 回归结果在表 3.8 中给出。再用简化模型 (3.47) 拟合数据, 若记 $W = X_1 + X_3$, 即用如下的模型

$$Y = \beta'_0 + \beta'_1 W + \varepsilon \quad (3.48)$$

拟合数据, 获得 β'_0 和 β'_1 的最小二乘估计、样本复相关系数 (此时为 Y 与 W 的简单相关系数, 因为只有一个变量)。拟合的方程为

$$\hat{Y} = 9.988 + 0.444W,$$

$R_1^2 = 0.6685$ 。由 (3.41), 用于检验原假设的 F -值为

$$F = \frac{(R_p^2 - R_q^2)/(p - q)}{(1 - R_p^2)/(n - p - 1)} = \frac{(0.7080 - 0.6685)/(2 - 1)}{(1 - 0.7080)/(30 - 2 - 1)} = 3.65,$$

自由度为 1 和 27, 对应的临界值为 $F_{(1,27,0.05)} = 4.21$ 。从而 F -检验不显著, 不能拒绝原假设。该方程的残差分布也较满意 (这里没给出)。

方程

$$\hat{Y} = 9.988 + 0.444(X_1 + X_3)$$

并未和给定的数据不一致。因此, 我们的结论是: 在决定雇员对主管人员的满意度上, 变量 X_1 和 X_3 有相同的正影响。这一假设亦可用 t -检验来做^①:

$$t = \frac{\hat{\beta}_1 - \hat{\beta}_3}{s.e.(\hat{\beta}_1 - \hat{\beta}_3)},$$

自由度为 27。用 t -检验和 F -检验是一致的, 因为自由度为 $(1, p)$ 的 F -分布与自由度为 p 的 t -分布的平方相等。

就上述例子, 我们已经讨论了如何一步步地建立模型, 讨论了在其他回归系数为 0 的假定下, β_1 和 β_3 是否相等。我们还可以检验其他更为复杂的原假设, 譬如 β_1 与 β_3 相等, 而 $\beta_2, \beta_4, \beta_5, \beta_6$ 均为 0。这一假设 H'_0 可表示为:

$$H'_0: \beta_1 = \beta_3, \beta_2 = \beta_4 = \beta_5 = \beta_6 = 0. \quad (3.49)$$

(3.45) 与 (3.49) 的差别在于, (3.45) 中假定 $\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0$, 而 (3.49) 则需要检验这一假定。我们可以很容易地检验 (3.49)。在 H'_0 的假定下, 简化模型为 (3.47)。但它不是与模型 (3.46) 作比较, 而是与含 6 个变量的全模型比较。此时 F -检验的观测值为

$$F = \frac{(0.7326 - 0.6685)/5}{0.2674/23} = 1.10, \quad d.f. = 5, 23.$$

与前面一样, 结果并不显著。但第一个仅对系数是否相同的检验更为敏感。(为什么?)

^① $s.e.(\hat{\beta}_i - \hat{\beta}_j) = \sqrt{Var(\hat{\beta}_i) + Var(\hat{\beta}_j) - 2Cov(\hat{\beta}_i, \hat{\beta}_j)}$, 可参阅本章附录。

3.9.4 有约束条件下回归参数的估计和检验

有时对一组数据拟合回归方程时,希望对参数值加一些约束。一个常用的约束是回归系数的和等于某个值,譬如 1 等。这些约束条件往往有理论上的原因,因为变量间可能存在一定的物理关系。在我们讨论的例子中,虽然没有明显的约束,但为作演示,我们考虑约束 $\beta_1 + \beta_3 = 1$ 。假定 (3.46) 式的模型为真,那么可进一步讨论当 X_1 或 X_3 增加一定的量时, Y 是否也增加同一个量。原假设 H_0 可表示为:

$$H_0: \beta_1 + \beta_3 = 1 | (\beta_2 = \beta_4 = \beta_5 = \beta_6 = 0). \quad (3.50)$$

或等价地 $\beta_3 = 1 - \beta_1$ 。在 H_0 下,模型简化为

$$H_0: Y = \beta_0 + \beta_1 X_1 + (1 - \beta_1) X_3 + \varepsilon \quad (3.51)$$

或

$$H_0: Y - X_3 = \beta_0 + \beta_1 (X_1 - X_3) + \varepsilon.$$

若记 $Y' = Y - X_3$, $V = X_1 - X_3$, 则模型又可写成

$$H_0: Y' = \beta_0 + \beta_1 V + \varepsilon.$$

因此,为得到约束条件下 β_1 、 β_3 的最小二乘估计,只要拟合以 V 为预测变量, Y' 为响应变量的回归方程。该方程为

$$\hat{Y}' = 1.166 + 0.694V,$$

从而得到用简化模型拟合的方程

$$\hat{Y} = 1.166 + 0.694X_1 + 0.306X_3,$$

其 $R^2 = 0.6905$ 。

H_0 的 F -检验值为

$$F = \frac{(0.7080 - 0.6905)/1}{0.2920/27} = 1.62, \quad d.f. = 1, 27.$$

它并不显著,因此不能拒绝 X_1 和 X_3 的回归系数之和为 1 的假设。

我们已检验过关于 β_1 和 β_3 的两种假设, $\beta_1 = \beta_3$ 和 $\beta_1 + \beta_3 = 1$ 。检验的结果是两个假设都不能被拒绝。这提示我们 β_1 和 β_3 可能都等于 0.5。我们也可以利用前面的方法直接去检验 $H_0: \beta_1 = \beta_3 = 0.5$ 。

在前面的例子中,我们检验了 $\beta_1 = \beta_3$, 这一检验可看成有约束条件下的假设检验的特例,其约束条件为 $\beta_1 - \beta_3 = 0$ 。全部或部分回归系数为 0 的检验也可看作约束条件下回归系数检验的特例。

从上面的讨论可见,一组给定的数据,可能可以用多个模型适当描述。此时,应全面地考察这些模型,这很重要。其中的某些模型可能比另一些更有意义(是

否有意义应根据应用背景及研究的主题来判断), 最后其中的某一个模型可能会被采纳。审视对数据的多种描述比仅研究一种描述会给出更深入的了解。

回归方程中究竟选择哪些变量是一个很复杂的问题, 将在第 11 章中详细讨论。这里提供两个注意点, 后面各章对此会作详细讨论。

1. 若回归系数的估计不是显著地不为 0, 则往往就取为 0。这将简化模型并减小预测方差。
2. 对于某个给定问题, 由于其理论上的重要性, 常常会在回归方程中保留某个或某些变量, 尽管其回归系数是统计不显著的。也就是说, 虽然回归系数的估计值并不显著地非 0, 但在回归方程中仍然保留相应的变量。这些被保留的变量应对过程给出有意义的描述, 相应的回归系数也有助于评价诸 X 对响应变量 Y 的贡献。

3.10 预测

给定预测变量的值, 可以用拟合的回归方程对响应变量作预测。若 $\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0p})$, 则对应于 \mathbf{x}_0 的预测值 \hat{y}_0 为

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \dots + \hat{\beta}_p x_{0p}, \quad (3.52)$$

其标准误 $s.e.(\hat{y}_0)$ 在本章附录中由 (A.10) 给出 (用矩阵记号), 可由统计软件包计算。置信系数为 $1 - \alpha$ 的预测限为

$$\hat{y}_0 \pm t_{(n-p-1, \frac{\alpha}{2})} s.e.(\hat{y}_0).$$

如同简单线性回归所述, 给定 \mathbf{x}_0 后也可给出相应的响应均值的估计。若记对应于 \mathbf{x}_0 的响应均值为 μ_0 , 其估计值 $\hat{\mu}_0$ 与 (3.52) 相同:

$$\hat{\mu}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_{01} + \hat{\beta}_2 x_{02} + \dots + \hat{\beta}_p x_{0p},$$

其标准误 $s.e.(\hat{\mu}_0)$ 也在本章附录中由 (A.12) 给出 (用矩阵记号)。 μ_0 的置信系数为 $1 - \alpha$ 的置信限为

$$\hat{\mu}_0 \pm t_{(n-p-1, \frac{\alpha}{2})} s.e.(\hat{\mu}_0).$$

3.11 小结

我们已经讲述了与线性模型有关的各类假设检验, 给出了作这些假设检验的一般方法, 还看到了各种不同检验采用的统计量也可由相应的样本复相关系数表示。这里特别强调的是, 在作任何假设检验之前, 首先应检查模型的假定是否恰当。正如在第 4 章中将会看到的, 残差图是实现上述目标的最简便直观的方式。若假设检验所基于的假定不成立, 则假设检验的结论就不合理。若根据检验结果选择了一个新的模型, 那么在分析结束之前, 必须考察这一新模型对应的残差。只有仔细注意细节才能成功地分析数据。

习题

- 3.1 利用主管人员业绩数据验证 (3.12) 获得的拟合方程 $\hat{Y} = 15.3276 + 0.7803X_1 - 0.0502X_2$ 中 X_1 的系数, 可通过若干简单回归方程得到。请仿照 3.5 节中获得 X_2 的系数的办法。
- 3.2 构造一个含有两个预测变量和一个响应变量的数据集, 使得下列两个方程中 X_1 的回归系数相同, 两方程为 $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1$ 和 $\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2$ 。提示: 两个预测变量应不相关。
- 3.3 表 3.10 给出了 22 个学生在统计课程中的考试成绩, 其中 F 为期末考试成绩, P_1, P_2 为两次预考成绩。该数据在本书的网站上可查到。
- (a) 分别用下列三个模型拟合数据:
- 模型 1: $F = \beta_0 + \beta_1 P_1 + \varepsilon$;
- 模型 2: $F = \beta_0 + \beta_2 P_2 + \varepsilon$;
- 模型 3: $F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \varepsilon$ 。
- (b) 对三个模型分别检验 $\beta_0 = 0$;
- (c) 用 P_1 和 P_2 对 F 进行预测时, 哪一个预测得更好?
- (d) 若某一学生两次预考成绩分别为 78 和 85, 请问你会选择哪一个模型预测他的期终成绩? 预测值为多少?
- 3.4 如果我们比较下列几个回归方程, 可以找到简单回归与多元回归系数之间的关系:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2; \quad (3.53)$$

$$\hat{Y} = \hat{\beta}'_0 + \hat{\beta}'_1 X_1; \quad (3.54)$$

$$\hat{Y} = \hat{\beta}''_0 + \quad + \hat{\beta}'_2 X_2; \quad (3.55)$$

$$\hat{X}_1 = \hat{\alpha}_0 + \quad + \hat{\alpha}_2 X_2; \quad (3.56)$$

$$\hat{X}_2 = \hat{\alpha}'_0 + \hat{\alpha}_1 X_1. \quad (3.57)$$

利用表 3.10 中的数据, 并取 $Y = F, X_1 = P_1, X_2 = P_2$, 验证

- (a) $\hat{\beta}'_1 = \hat{\beta}_1 + \hat{\beta}_2 \hat{\alpha}_1$, 即 Y 关于 X_1 的简单回归系数是多元回归中 X_1 的系数与 X_2 的系数乘以 X_2 关于 X_1 的回归系数之和;
- (b) $\hat{\beta}'_2 = \hat{\beta}_2 + \hat{\beta}_1 \hat{\alpha}_2$, 即 Y 关于 X_1 的简单回归系数是多元回归中 X_1 的系数与 X_2 的系数乘以 X_2 关于 X_1 的回归系数之和。
- 3.5 对给定的 20 组数据, 拟合响应变量 Y 关于预测变量 X_1 的简单回归模型。表 3.11 给出了回归输出结果, 但有些值被抹掉了。请补全 13 个缺失的值, 并计算 $\text{Var}(Y)$ 和 $\text{Var}(X_1)$ 。
- 3.6 对给定的 18 组数据, 拟合响应变量 Y 关于预测变量 X_1 的简单回归模型。表 3.12 给出了回归输出结果, 但有些值被抹掉了。请补全 13 个缺失的值, 并计算 $\text{Var}(Y)$ 和 $\text{Var}(X_1)$ 。
- 3.7 利用表 3.5 的回归输出结果分别构造 β_1 和 β_2 的 95% 置信区间。

表 3.10 考试成绩数据: 期末成绩 F , 第一次预考 P_1 , 第二次预考 P_2

行数	F	P_1	P_2	行数	F	P_1	P_2
1	68	78	73	12	75	79	75
2	75	74	76	13	81	89	84
3	85	82	79	14	91	93	97
4	94	90	96	15	80	87	77
5	86	87	90	16	94	91	96
6	90	90	92	17	94	86	94
7	86	83	95	18	97	91	92
8	68	72	69	19	79	81	82
9	55	68	67	20	84	90	83
10	69	69	70	21	65	70	66
20	91	91	89	22	83	79	81

表 3.11 Y 对 X_1 回归的输出结果 (20 个观测)

方差分析表 (ANOVA)				
来源	平方和	$d.f.$	均方	F -检验
回归	1848.76	-	-	-
残差	-	-	-	-
系数表				
变量	系数	$s.e.$	t -检验	p -值
常数	-23.4325	12.74	-	0.0824
X_1	-	0.1528	8.32	< 0.0001
$n =$	$R^2 =$	$R_a^2 =$	$\hat{\sigma} =$	$d.f. =$

表 3.12 Y 对 X_1 回归的输出结果 (18 个观测)

方差分析表 (ANOVA)				
来源	平方和	$d.f.$	均方	F -检验
回归	-	-	-	-
残差	-	-	-	-
系数表				
变量	系数	$s.e.$	t -检验	p -值
常数	3.43179	-	0.265	0.7941
X_1	-	0.1421	-	< 0.0001
$n =$	$R^2 = 0.716$	$R_a^2 =$	$\hat{\sigma} = 7.342$	$d.f. =$

3.8 解释为什么检验 (3.45) 式的假设比检验 (3.49) 式的假设更为敏感。

3.9 利用主管人员业绩数据, 在下列两个模型中分别检验假设 $H_0: \beta_1 = \beta_3 = 0.5$ 。

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon;$

(b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon.$

3.10 人们想知道是否身高相近的男女更容易结为夫妻, 为此, 对新婚夫妇做了抽

样调查。设 X 为丈夫的身高, Y 为妻子的身高, 见表 2.11。请采用你在第 2 章练习 2.10(f) 中选择的响应变量来检验截距和斜率均为 0 的假设。

- 3.11 为判定某一公司是否对妇女歧视, 从该公司中收集了一些数据: 年薪 (以千美元为单位), 资历, 性别 (“1” 为男性, “0” 为女性)。用两个线性模型对数据进行拟合, 回归输出结果见表 3.13。假设通常的回归假定成立。
- 男性是否比同等资历的女性有更高的收入?
 - 与女性有同等收入的男性是否资历较低?
 - 上述结果有否不一致? 并解释之。
 - 如果你是被告的辩护律师, 你会选择哪一个模型, 为什么?

表 3.13 薪水歧视数据的回归输出结果

模型 1: 预测变量为年薪				
变量	系数	s.e.	t-检验	p-值
常数	20009.5	0.8244	24271	< 0.0001
资历	0.935253	0.0500	18.7	< 0.0001
性别	0.224337	0.4681	0.479	0.6329
模型 2: 预测变量为资历				
变量	系数	s.e.	t-检验	p-值
常数	-16744.4	896.4	-18.7	< 0.0001
性别	0.850979	0.4349	1.96	0.0532
年薪	0.836991	0.0448	18.7	< 0.0001

表 3.14 薪水关于四个预测变量的回归输出结果

方差分析表 (ANOVA)				
来源	平方和	d.f.	均方	F-检验
回归	23665352	4	5916338	22.98
残差	22657938	88	257477	
系数表				
变量	系数	s.e.	t-检验	p-值
常数	3526.4	327.7	10.76	0.000
性别	722.5	117.8	6.13	0.000
教育	90.02	24.69	3.65	0.000
经历	1.2690	0.5877	2.16	0.034
本公司经历	23.406	5.201	4.50	0.000
$n = 93$	$R^2 = 0.515$	$R_a^2 = 0.489$	$\hat{\sigma} = 507.4$	$d.f. = 88$

- 3.12 表 3.14 给出了多元回归模型的回归输出结果。响应变量为某公司雇员的薪水 (以美元计), 预测变量有:

性别 示性变量 (1 为男性, 0 为女性)
 教育 受雇时在校学习的年数

经历 来公司前工作的时间 (以月计)

本公司经历 在本公司工作时间 (以月计)

对下面的 (a)、(b) 两题, 设定原假设与备择假设, 采用 5% 的显著性水平。

- (a) 构造 F -检验以检验回归总的拟合情况;
- (b) 考虑了性别、教育、本公司经历对薪水的影响后, 薪水和经历之间还存在正线性关系吗?
- (c) 对一个受过 12 年教育, 有 10 个月经历和 15 个月本公司经历的男士, 请你预测一下他的薪水。
- (d) 对受过 12 年教育, 有 10 个月经历和 15 个月本公司经历的男士们, 请你预测一下他们的平均薪水。
- (e) 对受过 12 年教育, 有 10 个月经历和 15 个月本公司经历的女士们, 请你预测一下她们的平均薪水。

表 3.15 薪水关于教育的回归的方差分析表 (ANOVA)

方差分析表 (ANOVA)				
来源	平方和	d.f.	均方	F -检验
回归	7862535	1	7862535	18.60
残差	38460756	91	422646	

3.13 表 3.14 给出了一个全回归模型的输出结果。现在考虑只取教育一个预测变量的简化模型, 表 3.15 给出了这一模型拟合的方差分析表。构造一个假设检验用以比较全模型和简化模型, 从假设检验的结果可得什么结论 ($\alpha = 0.05$)?

3.14 香烟消费数据: 一个国家的保险组织希望研究 50 个州和哥伦比亚特区的香烟消费模式。这一研究中选择的变量见表 3.16, 数据见表 3.17, 各州按字母次序排列 (数据也可在本书的网站上找到)。对下面的 (a)、(b), 给出原假设和备择假设、检验方法及在 5% 显著性水平下的结论。

- (a) 检验在销量关于 6 个预测变量的线性回归模型中, 女性比例这一变量是不必要的。
- (b) 检验上述回归方程中女性比例和 HS 这两个变量都不必要。
- (c) 给出收入变量的回归系数的 95% 置信区间。
- (d) 当收入变量从上述回归模型中去除时, 销量变量变差可以被其他预测变量解释的百分比是多少? 请叙述理由。
- (e) 销售变量变差可以被下面三个其他预测变量解释的百分比是多少: 价格、年龄、收入? 请给出理由。
- (f) 当仅用收入一个变量为预测变量时, 销量变量变差可以由收入变量解释的百分比是多少?

表 3.16 在表 3.17 的香烟消费数据中采用的变量

变量	定义
年龄	在一个州中居民年龄的中位数
HS	在一个州中年龄超过 25 岁且读完中学的居民的百分比
收入	州人均收入 (单位: 美元)
黑人比例	在一个州的居民中黑人的百分比
女性比例	在一个州的居民中女性占的百分比
价格	在一个州销售的每包香烟价格的加权平均 (以美分计)
销量	州中销售香烟的人均包数

3.15 考虑如下两个模型:

$$RM: H_0: Y = \varepsilon,$$

$$FM: H_1: Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

- 对上述假设进行 F -检验。
- 若 $p = 1$, 构造一个关于 Y 和 X_1 的数据集, 使得 H_0 在 5% 的显著性水平下不被拒绝。
- 叙述原假设的含义。
- 计算适当的、联系上述两个模型的 R^2 。

附录

这里我们用矩阵记号来表述多元回归分析中的标准结果。记

$$\mathbf{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{10} & x_{11} & \cdots & x_{1p} \\ x_{20} & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

则 (3.1) 式的线性模型可表示为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (\text{A.1})$$

其中对任意的 i , $x_{i0} = 1$ 。最小二乘估计中关于 $\boldsymbol{\varepsilon}$ 的假定为:

$$E(\boldsymbol{\varepsilon}) = 0, \quad \text{Var}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \mathbf{I}_n,$$

表 3.17 香烟消费数据 (1970)

州	年龄	HS	收入	黑人比例	女性比例	价格	销量
AL	27.0	41.3	2948.0	26.2	51.7	42.7	89.9
AK	22.9	66.7	4644.0	3.0	45.7	41.8	121.3
AZ	26.3	58.1	3665.0	3.0	50.8	38.5	115.2
AR	29.1	39.9	2878.0	18.3	51.5	38.8	100.3
CA	28.1	62.6	4493.0	7.0	50.8	39.7	123.0
CO	26.2	63.9	3855.0	3.0	50.7	31.1	124.8
CT	29.1	56.0	4917.0	6.0	51.5	45.5	120.0
DE	26.8	54.6	4524.0	14.3	51.3	41.3	155.0
DC	28.4	55.2	5079.0	17.1	53.5	32.6	200.4
FL	32.3	52.6	3738.0	15.3	51.8	43.8	123.6
GA	25.9	40.6	3354.0	25.9	51.4	35.8	109.9
HI	25.0	61.9	4623.0	1.0	48.0	36.7	82.1
ID	26.4	59.5	3290.0	0.3	50.1	33.6	102.4
IL	28.6	52.6	4507.0	12.8	51.5	41.4	124.8
IN	27.2	52.9	3772.0	6.9	51.3	32.2	134.6
IA	28.8	59.0	3751.0	1.2	51.4	38.5	108.5
KS	28.7	59.9	3853.0	4.8	51.0	38.9	114.0
KY	27.5	38.5	3112.0	7.2	50.9	30.1	155.8
LA	24.8	42.2	3090.0	29.8	51.4	39.3	115.9
ME	28.0	54.7	3302.0	0.3	51.3	38.8	128.5
MD	27.1	52.3	4309.0	17.8	51.1	34.2	123.5
MA	29.0	58.5	4340.0	3.1	52.5	41.0	124.3
MI	26.3	52.8	4180.0	11.2	51.0	39.2	128.6
MN	26.8	57.6	3859.0	0.9	51.0	40.1	104.3
MS	25.1	41.0	2626.0	36.8	51.6	37.5	93.4
MO	29.4	48.8	3781.0	10.3	51.8	36.8	121.3
MT	27.1	59.2	3500.0	0.3	50.0	34.7	111.2
NB	28.6	59.3	3789.0	2.7	51.2	34.7	108.1
NV	27.8	65.2	4563.0	5.7	49.3	44.0	189.5
NH	28.0	57.6	3737.0	0.3	51.1	34.1	265.7
NJ	30.1	52.5	4701.0	10.8	51.6	41.7	120.7
NM	23.9	55.2	3077.0	1.9	50.7	41.7	90.0
NY	30.3	52.7	4712.0	11.9	52.5	41.7	119.0
NC	26.5	38.5	3252.0	22.2	51.0	29.4	172.4
ND	26.4	50.3	3086.0	0.4	49.5	38.9	93.8
OH	27.7	53.2	4020.0	9.1	51.5	38.1	121.6
OK	29.4	51.6	3387.0	6.7	51.3	39.8	108.4
OR	29.0	60.0	3917.0	1.3	51.0	29.0	157.0
PA	30.7	50.2	3971.0	8.0	52.0	44.7	107.3
RI	29.2	46.4	3959.0	2.7	50.9	40.2	123.9
SC	24.8	37.8	2990.0	30.5	50.9	34.3	103.6
SD	27.4	53.3	3123.0	0.3	50.3	38.5	92.7
TN	28.1	41.8	3119.0	15.8	51.6	41.6	99.8
TX	26.4	47.4	3606.0	12.5	51.0	42.0	106.4
UT	23.1	67.3	3227.0	0.6	50.6	36.6	65.5
VT	26.8	57.1	3468.0	0.2	51.1	39.5	122.6
VA	26.8	47.8	3712.0	18.5	50.6	30.2	124.3
WA	27.5	63.5	4053.0	2.1	50.3	40.3	96.7
WV	30.0	41.6	3061.0	3.9	51.6	41.6	114.5
WI	27.2	54.5	3812.0	2.9	50.9	40.2	106.4
WY	27.2	62.9	3815.0	0.8	50.0	34.4	132.2

其中 $E(\varepsilon)$ 是 ε 的期望值 (均值), \mathbf{I}_n 是 n 阶单位阵, ε^T 是 ε 的转置。即 $\varepsilon_i, i = 1, 2, \dots, n$, 两两不相关, 均值为 0, 方差为常数。故

$$E(\mathbf{Y}) = \mathbf{X}\beta.$$

β 的最小二乘估计 $\hat{\beta}$ 可以通过最小化观测值与期望值的偏差平方和获得, 即最小化 $S(\beta)$, 其中

$$S(\beta) = \varepsilon^T \varepsilon = (\mathbf{Y} - \mathbf{X}\beta)^T (\mathbf{Y} - \mathbf{X}\beta).$$

最小化 $S(\beta)$, 导出下列方程组

$$(\mathbf{X}^T \mathbf{X})\hat{\beta} = \mathbf{X}^T \mathbf{Y},$$

该方程为 (3.5) 中的正规方程组。假定 $(\mathbf{X}^T \mathbf{X})$ 可逆, 则最小二乘估计 $\hat{\beta}$ 可明确地写为:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y},$$

$\hat{\beta}$ 是 \mathbf{Y} 的线性函数。 \mathbf{Y} 的拟合值 $\hat{\mathbf{Y}}$ 为

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{P}\mathbf{Y}, \quad (\text{A.2})$$

其中

$$\mathbf{P} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (\text{A.3})$$

称为帽子矩阵或投影阵。残差向量为

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{P}\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y}. \quad (\text{A.4})$$

最小二乘估计量的性质如下:

1. $\hat{\beta}$ 是 β 的无偏估计 (即 $E\hat{\beta} = \beta$), 协方差阵 $\text{Var}(\hat{\beta})$ 为

$$\text{Var}(\hat{\beta}) = E(\hat{\beta} - \beta)(\hat{\beta} - \beta)^T = \sigma^2(\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 \mathbf{C},$$

其中

$$\mathbf{C} = (\mathbf{X}^T \mathbf{X})^{-1}. \quad (\text{A.5})$$

在 β 的所有线性无偏估计中, 最小二乘估计具有最小方差, 因此 $\hat{\beta}$ 称为 β 的最佳线性无偏估计 (BLUE, Best Linear Unbiased Estimator).

2. 残差平方和可表示成

$$\mathbf{e}^T \mathbf{e} = \mathbf{Y}^T (\mathbf{I}_n - \mathbf{P})^T (\mathbf{I}_n - \mathbf{P}) \mathbf{Y} = \mathbf{Y}^T (\mathbf{I}_n - \mathbf{P}) \mathbf{Y}. \quad (\text{A.6})$$

最后一个等式成立是因为 $\mathbf{I}_n - \mathbf{P}$ 为对称幂等阵。

3. σ^2 的一个无偏估计为

$$\hat{\sigma}^2 = \frac{\mathbf{e}^T \mathbf{e}}{n - p - 1} = \frac{\mathbf{Y}^T (\mathbf{I}_n - \mathbf{P}) \mathbf{Y}}{n - p - 1}. \quad (\text{A.7})$$

若再假定 ε 服从多元正态分布, 进一步我们可以得到下列结果:

4. $\hat{\beta}$ 服从 $p+1$ 维正态分布, 均值向量为 β , 协方差阵为 $\sigma^2 \mathbf{C}$. $\hat{\beta}_j$ 的边缘分布是均值为 β_j 方差为 $\sigma^2 c_{jj}$ 的正态分布, 其中 c_{jj} 是 (A.5) 式中 \mathbf{C} 的第 j 个对角元. $\hat{\beta}_j$ 的标准误为

$$s.e.(\hat{\beta}_j) = \hat{\sigma} \sqrt{c_{jj}}. \quad (\text{A.8})$$

且 $\hat{\beta}_i$ 和 $\hat{\beta}_j$ 的协方差为 $\text{Cov}(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij}$.

5. $W = \mathbf{e}^T \mathbf{e} / \sigma^2$ 服从自由度为 $n-p-1$ 的 χ^2 分布.

6. $\hat{\beta}$ 和 $\hat{\sigma}^2$ 相互独立.

7. 拟合值向量 $\hat{\mathbf{Y}}$ 服从 n 维退化正态分布, 均值 $E(\hat{\mathbf{Y}}) = \mathbf{X}\beta$, 协方差阵 $\text{Var}(\hat{\mathbf{Y}}) = \sigma^2 \mathbf{P}$.

8. 残差向量 \mathbf{e} 服从 n 维退化正态分布, 均值 $E(\mathbf{e}) = 0$, 协方差阵 $\text{Var}(\mathbf{e}) = \sigma^2(\mathbf{I}_n - \mathbf{P})$.

9. 观测值 $\mathbf{x}_0 = (x_{00}, x_{01}, \dots, x_{0p})^T$ 对应的预测值 \hat{y}_0 为

$$\hat{y}_0 = \mathbf{x}_0^T \hat{\beta}, \quad (\text{A.9})$$

其中 $x_{00} = 1$. \hat{y}_0 的标准误为

$$s.e.(\hat{y}_0) = \hat{\sigma} \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}. \quad (\text{A.10})$$

\mathbf{x}_0^T 对应的响应均值的估计 $\hat{\mu}_0$ 为

$$\hat{\mu}_0 = \mathbf{x}_0^T \hat{\beta}, \quad (\text{A.11})$$

标准误为

$$s.e.(\hat{\mu}_0) = \hat{\sigma} \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}. \quad (\text{A.12})$$

10. 回归参数 β 的 $100(1-\alpha)\%$ 的联合置信域为

$$\left\{ \beta : \frac{(\beta - \hat{\beta})^T (\mathbf{X}^T \mathbf{X}) (\beta - \hat{\beta})}{\hat{\sigma}^2 (p+1)} \leq F_{(p+1, n-p-1, \alpha)} \right\} \quad (\text{A.13})$$

它是一个以 $\hat{\beta}$ 为中心的椭球体.

4

回归诊断：模型合理性的检测

4.1 引言

在第 2、3 章中，我们陈述了对简单回归模型及多元线性回归模型作推断的基本结果。这些结果建立在一些概述性统计量基础之上，这些统计量由数据来计算。对给定的一批数据拟合模型时，我们希望保证这种拟合不过分依赖于其中个别的或少数几个观测。只有当标准的回归假定满足时，第 2、3 章所述的那些关于分布、置信区间和假设检验的理论才是合理的、有意义的。本章 4.2 节叙述了这些假定。当这些假定不满足时，前文所述的那些标准结果将不成立，使用这些结果将导致严重的错误。我们再次重申，本书最关注的是检测线性模型的一些基本假定是否合理，及如何修正这些假定所遭到的破坏，并以此作为全面、有效地分析数据的一种手段。本章阐述了核实这些假定的方法。我们主要借助于图形方法而不是严格的数值规则去检查。

4.2 标准的回归假定

前两章我们给出了回归参数的最小二乘估计及其性质。最小二乘估计的性质以及第 2、3 章中所述的统计分析基于如下一些假定：

1. 关于模型形式的假定：我们假定联系响应 Y 和预测变量 X_1, X_2, \dots, X_p 的模型关于回归参数 $\beta_0, \beta_1, \dots, \beta_p$ 是线性的，即

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon. \quad (4.1)$$

相应地，第 i 个观测数据可表示为

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad i = 1, 2, \dots, n, \quad (4.2)$$

我们称此为线性假定。对于简单回归，我们很容易通过考察 Y 关于 X 的散点图来判断线性假定是否合理。如果图中数据点散布呈直线状，那么线性假

定是合理的。对于多元回归，由于数据是高维的，核实线性假定比较困难。本章下文提供的一些图形工具可以用于检查多元回归的线性假定。在线性假定不成立时，有时候可以通过数据变换得到线性。数据变换将在第6章讨论。

2. 关于误差的假定：(4.2) 中的误差 $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ 被假定为独立同分布的、均值为 0、方差为 σ^2 的正态随机变量。注意，这蕴涵着四个方面的假定：

- 误差 $\varepsilon_i, i = 1, 2, \dots, n$ 服从正态分布。我们称此为正态性假定。正态性假定不容易验证，尤其在预测变量取值不重复时。正态性假定的合理性可以通过一些合适的残差图来评价，这将在本章下文中讨论。
- 误差 $\varepsilon_i, i = 1, 2, \dots, n$ 具有 0 均值。
- 误差 $\varepsilon_i, i = 1, 2, \dots, n$ 具有相同的（但未知）方差 σ^2 。这称为等方差假定，也称为方差齐性假定。该假定不成立时，就是所谓的异方差问题，我们在第7章讨论。
- 误差 $\varepsilon_i, i = 1, 2, \dots, n$ 相互独立（此时，两两间的协方差为 0）。我们称之为误差独立假定。该假定不成立时，就是自相关问题，这将在第8章讨论。

3. 关于预测变量的假定：关于预测变量的假定有三个：

- 预测变量 X_1, X_2, \dots, X_p 是非随机的，也就是说，假定它们的取值 $x_{1j}, x_{2j}, \dots, x_{nj}; j = 1, 2, \dots, p$ 是固定的或事先选定的。只有当试验者能够将预测变量的值设置于预定的水平时，该假定才满足。显然，对于非试验的或观测的情况，该假定不满足。这时尽管第2、3章的理论结果依然成立，但对它们的解释应作修改。当预测变量为随机变量时，所有的推断是关于观测数据的条件推断。应当注意到，条件推断的观点和本书中所述的数据分析方法是一致的。我们的主要目标也是从现有的数据中提取最多的信息。
- 取值 $x_{1j}, x_{2j}, \dots, x_{nj}; j = 1, 2, \dots, p$ 是无误差地测得的。该假定几乎总是不成立。测量误差将会影响到误差方差、复相关系数以及单个回归系数的估计。影响程度的大小取决于多个因素，其中最重要的因素是测量误差的标准差及其相关结构。测量误差将增大误差方差，并降低所观察到的复相关系数的数量级。测量误差对于单个回归系数的影响更难以估量，一个变量的回归系数的估计不仅受它本身的测量误差的影响，而且还受方程中其他变量的测量误差的影响。

即使在所有测量误差均不相关的这种最简单的情况下，要修正测量误差对于回归系数估计的影响也需要知道各变量的测量误差的方差与随机误差的方差之比。而能够知道这些数据的场合是难得一遇的。在社会科学领域，这问题尤为突出。所以决不能奢望在估计回归系数时彻底剔除测量误差的影响。与随机误差相比，只要测量误差不太大，那么其影响也就无足轻重了。在解释系数时，应当记住这一点。尽管当变量有误差时回归系数的估计存在某些问题，但回归方程仍然可以用于预测。不过预测

的精度会下降。对于该问题的更广泛的讨论,读者可以参考 Full (1987), Chatterjee and Hadi (1988), 和 Chu-Lu and Van Ness (1999)。

- 预测变量 X_1, X_2, \dots, X_p 之间是线性无关的。这个假定保证最小二乘解 (即正规方程组 (3.5) 的解) 的唯一性。该假定不满足导致的问题称为共线性问题,我们在第 9 和第 10 章讨论。

上面关于预测变量的前两个假定无法核实,因此它们在分析中不起主要作用。但它们的确影响回归结果的解释。

4. 关于观测的假定: 所有观测是同样可靠的,它们对于回归结果与结论,作用大致相同。

最小二乘法有一个特点,即略微偏离潜在的前提假定不会严重影响分析、推断的合理性。然而,显著地偏离模型假定会严重扭曲结论。因此,通过各种图形来审视残差和数据模式的结构是非常重要的—件事。

4.3 形形色色的残差

在回归分析中,检测模型缺陷的一个简单而有效的方法是研究残差图。残差图能够指明哪个或哪些标准假定不成立。更重要的是,残差分析可能引导我们发现数据中的结构,也可能指出那些蕴含在数据中的、在只用一些概述性统计分析时容易被疏漏的信息。这些启发或线索可能帮助我们更好地理解所研究的问题,或者找到更好的模型。实践表明,对残差进行缜密的图形分析往往是回归分析最重要的一部分工作。

正如我们在第 2、3 章所见,若采用线性模型 (4.1) 对一组数据进行最小二乘拟合,我们得到如下拟合值

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \dots + \hat{\beta}_p x_{ip}, \quad i = 1, 2, \dots, n. \quad (4.3)$$

相应的普通最小二乘残差为

$$e_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n. \quad (4.4)$$

(4.3) 中的拟合值可以换一种形式表示为

$$\hat{y}_i = p_{i1}y_1 + p_{i2}y_2 + \dots + p_{in}y_n, \quad i = 1, 2, \dots, n, \quad (4.5)$$

其中诸 p_{ij} 仅依赖于预测变量的取值 (不涉及响应变量)。(4.5) 式直接地表明了观测数据与预测值之间的关系。在简单回归中, p_{ij} 为

$$p_{ij} = \frac{1}{n} + \frac{(x_i - \bar{x})(x_j - \bar{x})}{\sum (x_i - \bar{x})^2}. \quad (4.6)$$

在多元回归中, p_{ij} 是所谓帽子矩阵或投影矩阵的元素,该矩阵在第 3 章附录 (A.3) 中定义。

若 $i = j$, 则 p_{ii} 是投影矩阵 \mathbf{P} 的对角元素。对于简单回归,

$$p_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2}. \quad (4.7)$$

p_{ii} 称为第 i 个观测的杠杆值, 因为从 (4.5) 可见, \hat{y}_i 是 Y 的所有观测值的加权和, 而 p_{ii} 正是 y_i 对于第 i 个拟合值 \hat{y}_i 的权重 (参见 Hoaglin and Welsch, 1978)。因此, 我们总共有 n 个杠杆值, 记之为

$$p_{11}, p_{22}, \dots, p_{nn}. \quad (4.8)$$

杠杆值在回归分析中扮演着重要角色, 我们会经常碰到它们。

当 4.2 节所述的假定成立时, (4.4) 式定义的普通残差 e_1, e_2, \dots, e_n 之和为 0, 但它们的方差不同, 因为

$$\text{Var}(e_i) = \sigma^2(1 - p_{ii}), \quad (4.9)$$

其中 p_{ii} 是 (4.8) 式中的第 i 个杠杆值, 依赖于 $x_{i1}, x_{i2}, \dots, x_{ip}$ 。为解决方差不等的问题, 我们拿 e_i 除以它的标准差使之标准化, 得到

$$z_i = \frac{e_i}{\sigma\sqrt{1 - p_{ii}}}. \quad (4.10)$$

这称为第 i 个标准化残差, 因为它的均值为 0 标准差为 1。标准化残差依赖于未知的 σ , 即 ε 的标准差。 σ^2 的一种无偏估计为

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n - p - 1} = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1} = \frac{SSE}{n - p - 1}, \quad (4.11)$$

其中 SSE 为普通残差的平方和, 分母 $n - p - 1$ 称为自由度 ($d.f.$), 它等于观测个数 n 减去待估回归系数的个数 $p + 1$ 。

σ^2 的另一种无偏估计为

$$\hat{\sigma}_{(i)}^2 = \frac{SSE_{(i)}}{(n - 1) - p - 1} = \frac{SSE_{(i)}}{n - p - 2}, \quad (4.12)$$

其中 $SSE_{(i)}$ 是用除第 i 个观测之外的 $(n - 1)$ 个观测拟合模型时的残差平方和。 $\hat{\sigma}^2$ 与 $\hat{\sigma}_{(i)}^2$ 都是 σ^2 的无偏估计。

在 (4.10) 中用估计 $\hat{\sigma}$ 替代 σ , 得

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - p_{ii}}}, \quad (4.13)$$

若用估计 $\hat{\sigma}_{(i)}$ 替代 σ , 则得

$$r_i^* = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1 - p_{ii}}}. \quad (4.14)$$

形如 (4.13) 的残差称为内学生化残差, 形如 (4.14) 的残差称为外学生化残差, 因为 $\hat{\sigma}_{(i)}$ 没有涉及 e_i 。为简化术语便于表达, 我们将学生化残差称为标准化残差。

标准化残差之和不等于 0, 但它们方差相同。外标准化残差近似地[†]服从自由度为 $n-p-2$ 的 t -分布, 但内标准化残差不是。当样本容量较大时, 这些残差都近似地服从标准正态分布。标准化残差之间不是严格地相互独立, 但当观测的数量较大时, 可以近似地认为它们相互独立。

这两种残差之间有如下关系

$$r_i^* = r_i \sqrt{\frac{n-p-2}{n-p-1-r_i^2}} \quad (4.15)$$

可见, 其中的一个是另一个的单调变换。因此, 就作残差图而言, 使用哪种标准化残差差异不大, 无须加以区别。今后, 我们将内标准化残差用于各种残差图。我们将采用多种残差图来核实回归假定。

4.4 图形方法

图形方法在数据分析中起着重要作用。在对数据拟合线性模型时, 图形方法尤为重要。正如 Chambers et al. (1983, p1) 所说: “没有哪件统计工具能像一张精选出来的图形一样有威力。”图形方法可以被视为探索性工具, 同时也是验证分析或统计推断不可或缺的一部分。Huber (1991) 说, “目测具有正规诊断方法所没有的洞察力。”最能够说明这一点的一个例子是 Anscombe 的四组数据, 即第 2 章表 2.4 给出的四个数据集。Anscombe(1973) 用如下办法构造了这四个数据集: 各数据集集中的数据对 (Y, X) 具有相同的概述性统计量 (同样的相关系数、回归直线及相同的标准误, 等等), 但它们的散点图完全不一样 (图 4.1)。

散点图 4.1(a) 显示线性模型可能是合理的, 图 4.1(b) 则暗示了某种非线性模型 (可能可以线性化)。图 4.1(c) 中除一个点外其他点能很好地用一个线性模型拟合, 而这个点明显地远离那条直线。这可能是一个异常点, 在下结论之前应当仔细研究。图 4.1(d) 表明, 试验设计有缺陷或者是收集到的样本很糟糕。对于 $X = 19$ 这一点, 读者可以验证: (i) 无论相应的 Y 值多大多小, 这点上的残差总是零 (是不变的); (ii) 去掉这点, 用剩下的点得到的最小二乘估计不唯一 (除了垂直于横轴的直线, 任何一条通过剩下的点之平均值的直线都是最小二乘直线!)。过度影响回归结果的观测称为强影响点。因而 $X = 19$ 这点是一个极端的强影响点, 因为它独自就决定了回归直线的截距和斜率。

[†] 译注: 原文无“近似”的意思, 疑有误, 参见: 陈希孺和王松桂著。《近代回归分析——原理方法及应用》, 安徽教育出版社, 1987 年, 第 102 页第 4 行。

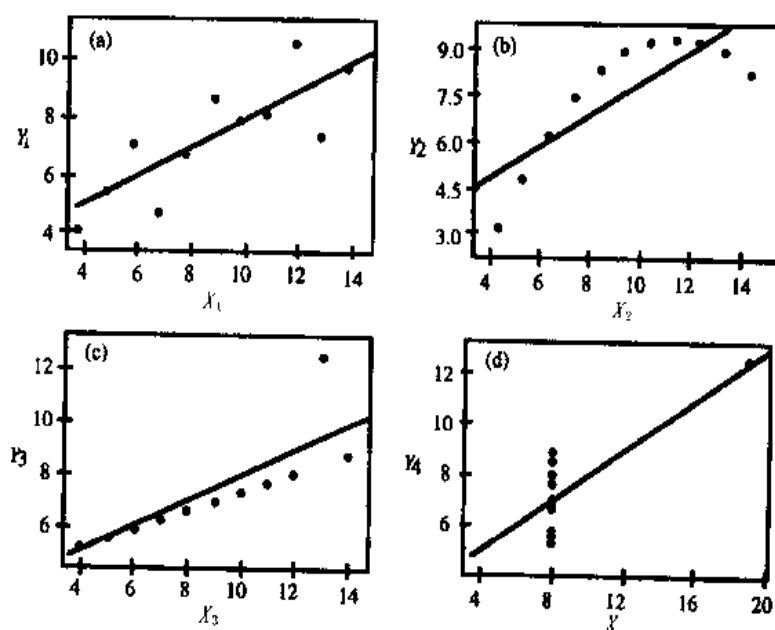


图 4.1 Anscombe 的四组数据 (X, Y) 及其最小二乘直线拟合图

这里我们把散点图用作探索性工具, 但你也可以在验证分析中将图形方法用于弥补数值方法的不足。假如我们希望检验 Y 与 X 是否正相关, 或者等价地, 是否可以对 Y 和 X 拟合一条斜率为正的回归直线。读者可以验证, 上述四个数据集具有相同的相关系数 ($Cor(Y, X) = 0.80$), 相同的回归直线 ($Y = 3 + 0.5X$), 以及回归系数有相同的标准误。因此, 根据这些数值结果, 人们可能得出错误的结论, 也即这四批数据可用相同的模型描述。此结论潜在的假定是 Y 和 X 的关系是线性的, 但这里该假定并不成立, 比如图 4.1(b) 中的数据集。因此, 这样的检验是不合理的。如同其他统计方法, 相关性检验[†]也是基于某些潜在假定的。仅当潜在的假定成立, 那么由这些方法得出的结论方才合理。从上面的例子清楚地看到, 仅仅采用数值结果进行分析往往会得出错误的结论。

图形方法在许多方面都很有用:

1. 发现数据中的错误 (例如, 一个异常点可能是印刷错误);
2. 辨别数据中的模式 (例如, 密集群, 异常点, 明显的差距等);
3. 探索变量间的关系;
4. 发现新现象;
5. 确认或否定各项假定;
6. 评价拟合的模型是否充分;
7. 建议修正措施 (例如, 数据变换, 重新设计试验, 收集更多数据, 等等);
8. 整体上完善数值分析。

本章介绍一些在回归分析中有用的图形工具。我们这里讨论的图形工具可以

[†] 译注: 此处原文为 “the test for linear relationship”, 但联系上下文来看, 似乎 “相关性检验” 更妥贴。

分为两类（并不互相排斥）：

- 拟合模型前的图形，这些图形用于数据纠错，模型选择等方面；
- 拟合模型后的图形，这些图形对于核实各项假定、评价拟合效果特别有用。

我们的介绍主要取材于 Hadi (1993) 以及 Hadi and Son (1997)。在研究一个具体的图形前，首先想想当假定成立时这图形看起来会是怎样的，然后仔细考察它是否与想像中的一致。这就能帮助我们确认或推翻假定。

4.5 拟合模型前的图形工具

描述响应变量与预测变量间关系的模型形式应当根据理论背景或者欲待检验的假设来定。但如果没有任何关于模型形式的先验信息可用，那么可以根据数据来定。在拟合一个模型之前，应当对数据作详细的检查。拟合模型前的图形可用作探索性工具，有如下四类可能的图形：

1. 一维图
2. 二维图
3. 旋转图
4. 动态图

4.5.1 一维图

数据分析之初，往往要逐个地审视待研究的变量。目的是对每个变量各自的分布做到心里有数。下列一些图形可用于研究单个变量：

- 直方图
- 茎叶图
- 点图
- 箱线图

一维图形有两项主要功能。它们可以显示单个变量的分布，显示该变量是对称的还是偏斜的。若某个变量严重偏斜，那就需要作变换。对于一个偏斜度很高的变量，我们建议作对数变换。单变量图是采用原始变量还是采用变换后的变量作进一步研究提供指导。

单变量图也可以指明各变量中是否有异常值。我们应当检查这些异常值，看是否由于抄写错误造成的。在这一阶段，我们不应该删除任何观测。在后面的分析中一旦它们显得难以处理时，我们就要注意这些观测。

4.5.2 二维图

理想地，在面对多维数据时，我们希望有同样维数的图形工具来考察数据。显然，这只有在变量数目很少时方可行。但是，我们可以审视数据集中两两变量之间的散点图，以此来考察各变量对之间的关系并作初步的模式识别。

当变量数目较少时，可以将这些变量对的散点图排成一个矩阵，有时候称之为打样图或者图矩阵。图 4.2 即是一例，是一个响应变量、两个预测变量之间的图

矩阵。各变量对的散点图被置于图矩阵的上三角部位。我们也可以将两两变量之间的相关系数排成一个矩阵，并置之于此图矩阵的下三角部位。这样的安排便于考察这些散点图。而对各变量对的相关系数的解释也总是和相应的散点图联系在一起。这么做有两层理由：(a) 相关系数度量的仅仅是线性关系；(b) 相关系数并不稳健，其值有可能严重地受一两个观测的影响。

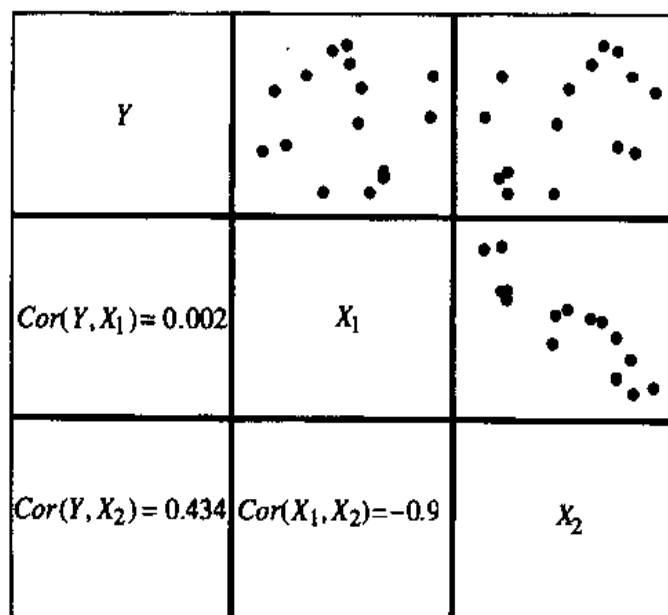


图 4.2 Hamilton 数据的图矩阵

我们希望图矩阵中的各散点图看上去是怎样的呢？对于简单回归，我们预期 Y 与 X 之间呈现某种直线模式。但对于多元回归， Y 与各自变量之间的散点图可能呈直线状，也可能不呈直线状。在线性模式较为肯定的场合，这些散点图的非线性状态并不说明线性模型不正确。下述一例。

表 4.1 Hamilton (1987) 的数据

Y	X ₁	X ₂	Y	X ₁	X ₂
12.37	2.23	9.66	12.86	3.04	7.71
12.66	2.57	8.94	10.84	3.26	5.11
12.00	3.87	4.40	11.20	3.39	5.05
11.93	3.10	6.64	11.56	2.35	8.51
11.06	3.39	4.91	10.83	2.76	6.59
13.03	2.83	8.52	12.63	3.90	4.90
13.13	3.02	8.04	12.46	3.16	6.96
11.44	2.14	9.05			

例：Hamilton 的数据

Hamilton (1987) 根据这样的思路生成了几个数据集： Y 同时依赖于各个预测变量

而不是单个。其中一个数据集列于表 4.1。从这批数据的图矩阵 (图 4.2) 可见, Y 与 X_1 之间 ($R^2 = 0$) 及 Y 与 X_2 之间 ($R^2 = 0.19$) 都不存在线性关系。然而, 作 Y 关于 X_1 与 X_2 两个变量的回归时, 拟合程度几近完美。读者可以验证下列回归方程:

$$\begin{aligned}\hat{Y} &= 11.989 + 0.004X_1; & t\text{-test} &= 0.009; & R^2 &= 0.0; \\ \hat{Y} &= 10.632 + 0.195X_2; & t\text{-test} &= 1.74; & R^2 &= 0.188; \\ \hat{Y} &= -4.515 + 3.097X_1 + 1.032X_2; & F\text{-test} &= 39222; & R^2 &= 1.0.\end{aligned}$$

前两个方程显示, Y 既不依赖于单个的 X_1 也不依赖于单个的 X_2 , 然而同时用 X_1 、 X_2 几乎能完美地预测 Y 。顺便指出, 第一个方程修正的 R^2 是负的, $R_a^2 = -0.08$ 。

在滤去所有其他预测变量的线性效应之后所作的 Y 与单个预测变量的散点图才呈直线状。在 4.12.1 节, 我们介绍两类这样的图形, 即所谓的添加变量图和残差加分量图。

因为我们假定预测变量之间是线性无关的, 所以预测变量对的散点图不应该呈现直线状, 更理想地, 我们希望从中看不出任何可辨识的模式, 无论是线性的还是非线性的。但在上述 Hamilton 的数据中, 该假定并不成立, 从图 4.2 可见, X_1 与 X_2 有明显的线性关系。这里我们要小心, 这些散点图不呈直线状还不能说明全部预测变量之间是线性无关的, 因为线性关系可能存在于多个预测变量之间。变量对的散点图无法揭示这种多变量关系。多重共线性问题将在第 9、10 章中讨论。

4.5.3 旋转图

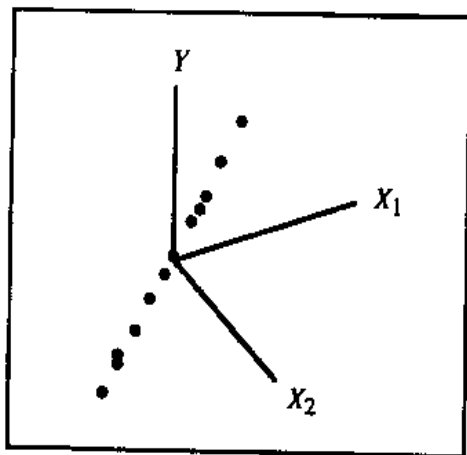


图 4.3 Hamilton 数据的旋转图

近年来电脑硬件、软件技术的发展已使得画三维甚至高维的数据图成为可能。最简单的是三维旋转图。旋转图是一种三个变量的散点图, 图中的点可以按不同方向旋转以致于能够明显反映出三维结构。语言不足以充分形容旋转图, 只有在

电脑屏幕上观察运动着的旋转图才能感受到旋转的真正威力。当图形转到某个合适的角度时，我们可以停止旋转。譬如，我们可以从 Y 关于 X_1 、 X_2 的旋转图确认 Hamilton 的数据中 Y 与 X_1 、 X_2 的完美关系。在图形转动时，我们发现所有数据点几乎落在一个平面上。图 4.3 给出了一个合适的方向，从这个角度看该平面，似乎所有的数据点都散布在一条直线上。

4.5.4 动态图

动态图是用于探索多维数据之结构关系的非凡工具。动态图给予数据分析者完全不同于静态图的感觉。人们可以操纵这些图形，任何变化能够即时地反映在电脑屏幕上。譬如，可以同时作两个或多个三维旋转图，然后使用动态图技术考察三维以上的结构关系。有不少关于该主题的文章和书籍，许多统计软件也含有动态图工具（例如，旋转、擦除、连接等）。有兴趣的读者可以参考 Becker, Cleveland, and Wilks (1987)，以及 Velleman (1999)。

4.6 拟合模型后的图形工具

前一节介绍的图形是数据检查、模型构建阶段的有力工具。对数据拟合了某个模型之后画的一些图形有助于检查假定是否合适、评价拟合的充分性。这些图形大致分为以下几类：

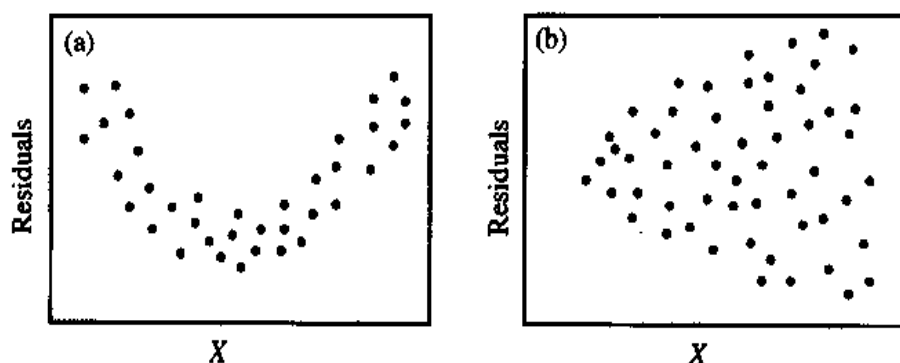
1. 检查线性、正态性假定的图形；
2. 检测异常点、强影响点的图形；
3. 变量效应的诊断图。

4.7 检查线性和正态性假定

变量数目较少时，我们可以通过前面介绍的交互和动态的图形操作来检查线性假定。当变量数较多时，检查线性假定就变得困难了。不过，我们可以通过研究拟合模型后的残差来检查线性和正态性假定。

下列标准化残差图可用于检查线性和正态性假定。

1. 标准化残差的正态概率图：这是排序后的标准化残差关于所谓的正态得分的散点图。正态得分是来自标准正态分布的容量为 n 的样本的次序统计量之期望值。如果残差是正态分布的，那么排序后应近似地与排序后的正态得分相同。因此，在正态性假定下，该散点图看来应像一条截距为 0、斜率为 1 的直线。（截距和斜率分别表示标准化残差的均值和标准差。）



(a) 非线性的模式;

(b) 异方差模式。

图 4.4 两张残差关于 X 的散点图, 示意使模型假定不满足的两种模式

2. 标准化残差关于每个预测变量的散点图: 在标准假定下, 标准化残差与每个预测变量都是不相关的。如果这些假定成立, 图中的数据点应当是随机散布的。图中出现可辨识的模式则表明某些假定可能不成立。若线性假定不成立, 看到的图有可能像图 4.4(a) 那样。在这种情况下, 为保证模型是线性的, 可能有必要对 Y 或对相应的那个预测变量或对两者同时作变换。类似图 4.4(b) 的散点图则表明方差不齐性。此时, 对数据进行方差稳定性变换可能是必要的。用于修正模型缺陷的几种变换在第 6 章介绍。
3. 标准化残差关于拟合值的散点图: 在标准假定下, 标准化残差与拟合值也是不相关的, 因此图中的点也应当随机散布。对于简单回归, 标准化残差关于 X 的散点图与关于拟合值的散点图是一样的。
4. 标准化残差序列图: 这是按观测顺序来显示标准化残差的诊断图。如果各观测的排列顺序是不重要的, 那这种诊断图就没有必要。但如果顺序是重要的 (譬如, 各观测是按时间顺序获得的或者各观测间有某种空间顺序), 那么这种图可用于检查误差独立性假定。在误差独立性假定下, 图中的点应当在一个以 0 为中心的水平带形区域内随机散布。

4.8 杠杆、影响及异常

在对一批数据拟合模型时, 我们希望拟合情况不过度地取决于一个或少数几个观测。回想一下, 在 Anscombe 的四组数据中, 图 4.1(d) 中的直线完全取决于一个点。如果去掉这个极端的点, 那么就会得到一条非常不同的直线。在多变量的情况下, 不可能用图形来检测这种情形。但是, 我们仍然希望知道这样的点是否存在。应当指出, 此时残差图也无济于事, 因为该点的残差为 0! 因此这并非一个异常点, 而是一个强影响点。

如果一个点是强影响点, 那么, 单独地删除它或者将它与其他两三个点一起删除会导致拟合的模型发生本质的改变 (譬如, 系数估计, 拟合值, t -检验等等)。一般来说, 删除任何一点都会改变拟合, 但我们只关心删除后会导致重大改变的

那些点（即它们有着不恰当的影响）。我们举一个例子来说明这一点。

例：纽约州的河流数据

考虑 1.3.5 节中描述的纽约州的河流数据，数据在表 1.9 中给出。我们对氮的平均浓度 Y 与四个表示土地使用状况的预测变量拟合一个线性模型：

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \varepsilon. \quad (4.16)$$

表 4.2 列出了用数据的三个子集获得的回归系数[†]以及检验系数显著性的 t -检验值。表中第二列给出的是基于全部 20 个观测（河流）的回归结果。第三列中是删去 Neversink 河（第 4 号观测）后的结果。第四列中是删去 Hackensack 河（第 5 号观测）后的结果。

表 4.2 纽约州的河流数据：单个系数的 t -检验

检验	删除的观测		
	未删除	Neversink	Hackensack
t_0	1.40	1.21	2.08
t_1	0.39	0.92	0.25
t_2	-0.93	-0.74	-1.45
t_3	-0.21	-3.15	4.08
t_4	1.86	4.45	0.66

请注意用这三个数据子集得到的回归结果之间的巨大差异！尽管这三个数据子集相互之间仅仅相差一个观测。譬如，看 β_3 的 t -检验值，使用全部数据时该检验是不显著的，删去 Neversink 河时显著为负，删去 Hackensack 河时显著为正。可见，仅一个观测就能导致根本不同的结果和结论！Neversink 和 Hackensack 河称为强影响观测，因为它们对回归结果的影响远远强于其他观测。审视表 1.9 中的原始数据，一眼就能发现 Hackensack 河，因其 X_3 （住宅地面积比例）的值突出地大。这是因为 Hackensack 河是数据中唯一的一条市区河，地理位置紧邻人口密度高的纽约市。其他河流都在农村。尽管 Neversink 河也是强影响点，但从原始数据来看，它与其他河流的区别不明显。

因此，鉴别数据中可能存在的强影响观测是重要的。我们介绍几种检测强影响观测的方法。通常，强影响观测或者其响应变量取值异常或者其预测变量取值异常。

4.8.1 响应变量取值异常

标准化残差大的观测其响应变量的取值异常，因为在 Y -方向上它们远离拟合的回归方程。由于各标准化残差近似服从标准正态分布，那么标准化残差之绝对值大于 2 或 3 的点称为异常点。异常点显示在这些点上模型拟合的失败。它们可

[†]译注：原文表 4.2 中并未列出回归系数。

用一些正规的检验程序(比如可以参考 Hawkins (1980), Barnett and Lewis (1994), Hadi and Simonoff (1993), Hadi and Velleman (1997) 以及这些文献中的参考文献)或者一些合适的残差图来鉴别。这里我们采用后一类方法。各点残差的图形模式比其数值更重要。一旦模型有不合适的地方,残差图常常能够将它们全面地暴露出来。研究残差图是我们采用的一种主要的分析工具。

4.8.2 预测变量取值异常

异常点也可能出现在预测变量(X -空间)中。它们同样也会影响回归结果。以前所述的杠杆值 p_{ii} 可用于度量观测在 X -空间中的异常程度。从(4.7)给出的简单回归情形下 p_{ii} 的公式就可以看出这一点。该公式显示,一个点偏离 \bar{x} 越远,那么相应的 p_{ii} 值越大。多元回归亦是如此, p_{ii} 值大的观测是 X -空间的异常观测(与预测变量空间中其他点相比而言)。因此, p_{ii} 可用作在 X -空间中异常程度的量度。在 X -空间中异常的观测(例如图 4.1(d) X_4 值最大的那个点)称为高杠杆点,以区别于响应变量异常的观测(它们具有大的标准化残差)。

杠杆值有几个有趣的性质(Dodge and Hadi (1999), 以及 Chatterjee and Hadi (1988) 第2章,有综合的论述)。譬如,它们介于 $0 \sim 1$ 之间,平均值为 $(p+1)/n$ 。 p_{ii} 值大于 $2(p+1)/n$ (均值的两倍)的点通常被认为是高杠杆点(Hoaglin and Welsch, 1978)。

不管作什么分析,必须把高杠杆点标出来,并审视它们是否强影响点。杠杆值图(比如序列图、点图或箱线图)可以暴露出可能存在的高杠杆点。

4.8.3 伪装与淹没的问题

标准化残差为核实线性、正态性假定以及鉴别异常值提供了颇有价值的信息。然而,鉴于以下原因,单单建立在残差基础上的分析可能无法检测出异常观测和强影响观测:

1. 高杠杆点的存在: 普通残差 e_i 和杠杆值 p_{ii} 有下列关系

$$p_{ii} + \frac{e_i^2}{SSE} \leq 1 \quad (4.17)$$

其中 SSE 是残差平方和。该不等式表明,高杠杆点(p_{ii} 大的点)残差较小。比如,图 4.1(d) 中 $X = 19$ 的那个点,尽管其残差为 0,但影响极端地强。因此,除了用标准化残差来检查异常点外,最好还用杠杆值来鉴别那些惹麻烦的点。

2. 伪装与淹没问题: 伪装是指数据中有异常点但我们没能发现它们。这是因为某些异常点可能被数据中的其他一些异常点掩藏起来了。淹没[†]指我们错误地将某些非异常点判断为异常点。这是因为异常点倾向于将回归方程往它们身边拉近,因而使得其他点远离拟合的方程。这样,伪装是错误的否定判断,

[†] 译注: 原文这里是 Masking, 但据上下文, 这里疑为 Swamping, 故译为淹没。

而淹没则是错误的肯定。下面给出的一个数据集就是存在伪装与淹没问题的例子。在 Hadi and Simonoff (1993) 及其参考文献中给出了一些方法，相对于标准化残差与杠杆值而言，这些方法不容易受伪装与淹没问题的干扰。

由于上面的原因，我们需要别的工具去度量各观测的影响。在介绍这些方法之前，我们采用一个实际的例子来说明上面这些概念。

例：纽约州的河流数据

我们来看纽约州的河流数据，但这次为说明不同的问题。我们考虑拟合简单回归模型

$$Y = \beta_0 + \beta_4 X_4 + \varepsilon \tag{4.18}$$

来刻画平均的氮浓度 Y 与工业或商业用地面积的百分比 X_4 之间的关系。 Y 关于 X_4 的散点图及相应的用最小二乘法拟合的直线在图 4.5 中给出。相应的标准化残差 r_i 及杠杆值 p_{ii} 列于表 4.3，它们各自的序列图显示于图 4.6。在标准化残差序列图中，所有的残差都较小，表明数据中没有异常点。但这是个错误的结论，因为数据中有两个明显的异常点，这可从图 4.5 中的散点图上看出来。因此，伪装发生了！正因为 (4.17) 中杠杆值与残差之间的关系，Hackensack 河的杠杆值 $p_{ii} = 0.67$ 很大，因而残差很小。尽管残差值小是理想的，但这里残差值小的原因不是因为拟合得好，而是因为第 5 号观测是一个高杠杆点，还有第 4 号观测也是，它们把回归直线拉向自己了。

表 4.3 纽约州的河流数据：拟合模型 (4.18) 时的标准化残差 r_i 及杠杆值 p_{ii} 。

行	r_i	p_{ii}	行	r_i	p_{ii}
1	0.03	0.05	11	0.75	0.06
2	-0.05	0.07	12	-0.81	0.06
3	1.95	0.05	13	-0.83	0.06
4	-1.85	0.25	14	-0.83	0.05
5	0.16	0.67	15	-0.94	0.05
6	0.67	0.05	16	-0.48	0.06
7	1.92	0.08	17	-0.72	0.06
8	1.57	0.06	18	-0.50	0.06
9	-0.10	0.06	19	-1.03	0.06
10	0.38	0.06	20	0.57	0.06

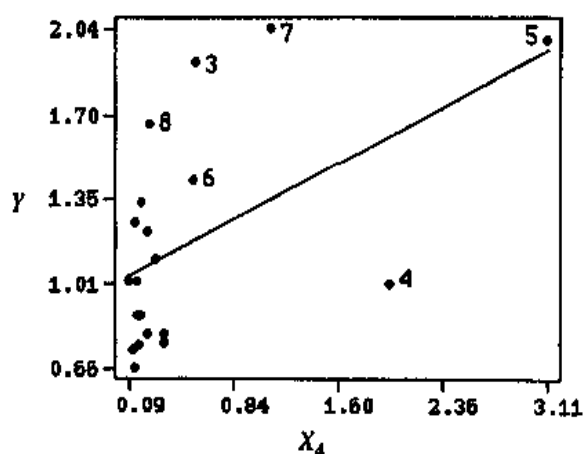
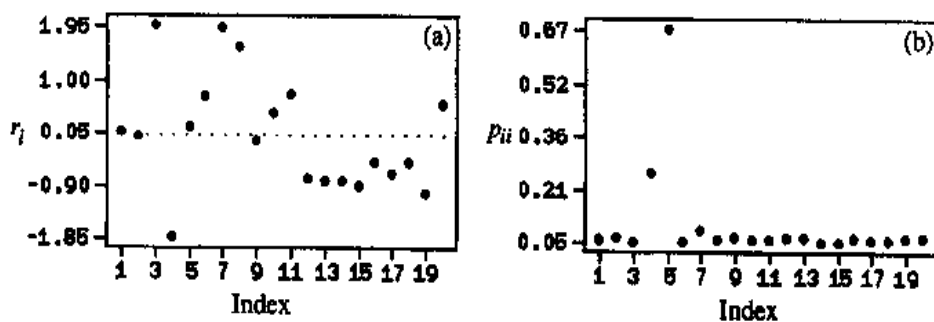
图 4.5 纽约州的河流数据: Y 关于 X_4 的散点图(a) 标准化残差 r_i 的序列图(b) 杠杆值 p_{ii} 的序列图

图 4.6 纽约州的河流数据

对于 p_{ii} , 一个常用的阈值为 $2(p+1)/n = 0.2$ (Hoaglin and Welsch, 1978)。因此, 将我们前面在散点图 4.5 中发现的两个突出的点 (Hackensack, $p_{ii} = 0.67$, 和 Neversink, $p_{ii} = 0.25$) 标记为高杠杆点, 这两个点也可在图 4.6(b) p_{ii} 的序列图中看出来, 它们远离其他点。这个例子清楚地说明, 单单看残差图是不够的。

4.9 影响的各种量度

一个观测的影响, 由在拟合过程中删除它所导致的后果来度量。这时, 一般每次删除一个点。删去第 i 个观测后 ($i = 1, 2, \dots, n$) 得到的回归系数记为 $\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}, \dots, \hat{\beta}_{p(i)}$ 。类似地, 剔除第 i 个观测后的预测值与残差均方记为 $\hat{y}_{1(i)}, \hat{y}_{2(i)}, \dots, \hat{y}_{n(i)}$ 与 $\hat{\sigma}_{(i)}^2$ 。注意, 根据第 i 个观测删去后拟合的方程, 第 m 个观测的拟合值为

$$\hat{y}_{m(i)} = \hat{\beta}_{0(i)} + \hat{\beta}_{1(i)}x_{m1} + \dots + \hat{\beta}_{p(i)}x_{mp} \quad (4.19)$$

各种影响量度, 看的都是诸如 $\hat{\beta}_j - \hat{\beta}_{j(i)}$ 或 $\hat{y}_j - \hat{y}_{j(i)}$ 等数量的差异。文献中有众多的影响量度, 读者可从下面这些书中的某一本书中查阅到详细资料: Belsley, Kuh,

and Welsch (1980), Cook and Weisberg (1982), Atkinson (1985) 及 Chatterjee and Hadi (1988)。这里给出其中的三种量度。

4.9.1 Cook 距离

Cook (1977) 提出的一种影响量度被广泛使用。Cook 距离衡量的是由全部数据得到的回归系数与删去第 i 个观测得到的回归系数之间的差异, 或者等价地, 是由全部数据得到的拟合值与删去第 i 个观测得到的拟合值之间的差异。相应地, Cook 距离衡量第 i 个观测的影响用的是

$$C_i = \frac{\sum_{j=1}^n (\hat{y}_j - \hat{y}_{j(i)})^2}{\hat{\sigma}^2(p+1)}, \quad i = 1, 2, \dots, n. \quad (4.20)$$

可以证明 C_i 可表达为

$$C_i = \left(\frac{r_i^2}{p+1} \right) \left(\frac{p_{ii}}{1-p_{ii}} \right), \quad i = 1, 2, \dots, n. \quad (4.21)$$

因此, Cook 距离实际上是两个基本量的乘积函数。前一个量是 (4.13) 中定义的标准残差 r_i 的平方, 后一个量是所谓的位势函数 $p_{ii}/(1-p_{ii})$, 其中 p_{ii} 是前面引进的第 i 个观测的杠杆值。如果一个点是强影响的, 删除它会导致巨大变化, 且 C_i 的值会很大。因此, C_i 的值大表明该点是强影响的。有人建议, 若 C_i 值大于自由度为 $p+1$ 和 $n-p-1$ 的 F 分布之 50% 分位点, 那么相应的点就可作为强影响点。一种实际的操作规则是, 将 C_i 值大于 1 的点归为强影响点。我们建议通过图形考察所有的 C_i 值, 而不是采用某个严格的截止规则。 C_i 的点图或序列图是一种有力的图形工具。当所有 C_i 值都几乎相同时, 不必采取任何措施。另一方面, 若有些数据点的 C_i 值比其余点突出, 那么该对这些点打上标记, 并作检查。接着可能要去掉这些讨厌点重新拟合模型, 来看这些点的效应。

4.9.2 Welsch-Kuh 量度

一种类似于 Cook 距离的量度由 Welsch and Kuh (1977) 提出, 并命名为 $DFITS_i$ 。其定义为

$$DFITS_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)} \sqrt{p_{ii}}}, \quad i = 1, 2, \dots, n. \quad (4.22)$$

因此, $DFITS_i$ 是用全部数据获得的与用删去第 i 个观测后的数据获得的第 i 个观测的两种拟合值之间的差异, 再按比例 $\hat{\sigma}_{(i)} \sqrt{p_{ii}}$ 折算。可以证明, $DFITS_i$ 可写成

$$DFITS_i = r_i^* \sqrt{\frac{p_{ii}}{1-p_{ii}}}, \quad i = 1, 2, \dots, n, \quad (4.23)$$

其中 r_i^* 为 (4.14) 中定义的标准残差。 $DFITS_i$ 相当于 $\sqrt{C_i}$, 但其中用的是 $\hat{\sigma}_{(i)}$ 而不是 $\hat{\sigma}$ 。通常, $|DFITS_i|$ 大于 $2\sqrt{(p+1)/(n-p-1)}$ 的点被视为强影响点。同样, 我们也可借助序列图、点图或箱线图等图形工具, 根据这一量度, 挑出与其他

点相比影响程度异常高的点来,而不采用某个严格的阈值。使用 C_i 或 $DFITS_i$, 没有多大区别,它们给出的答案相似,因为它们都是残差和杠杆值的函数。绝大多数电脑软件提供这两种量度或其中一种,看其中一种就足够了。

4.9.3 Hadi 影响量度

Hadi (1992) 根据如下事实,提出一种衡量第 i 个观测之影响的量度: 强影响观测要么在响应变量上异常,要么在预测变量上异常,或者两者皆异常。于是,可用

$$H_i = \frac{p_{ii}}{1 - p_{ii}} + \frac{p + 1}{1 - p_{ii}} \frac{d_i^2}{1 - d_i^2}, \quad i = 1, 2, \dots, n \quad (4.24)$$

衡量第 i 个观测的影响,其中 $d_i = e_i / \sqrt{SSE}$ 是所谓的正规化残差。(4.24) 的右边第一项为位势函数,衡量的是在 X -空间的异常程度。第二项是残差的函数,衡量响应变量上的异常程度。可以看到,如果观测在响应变量上、或者(并且)在预测变量上异常,那么它们会有大的 H_i 值。也就是说,如果它们有较大的 r_i 值或 p_{ii} 值或两者皆是,则 H_i 大。 H_i 这个量度不针对于某个具体的回归结果,它可被看作是一种综合的、一般的影响量度,揭示的是对至少一个回归结果有强影响的观测。

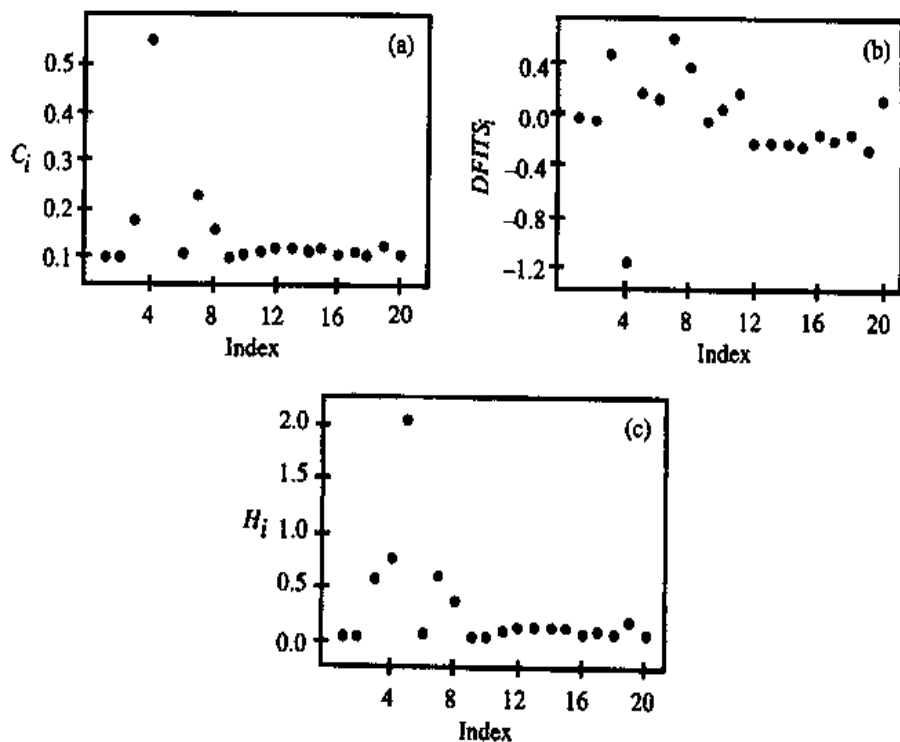
注意, C_i 和 $DFITS_i$ 是残差和杠杆值的乘积函数,而 H_i 是和函数。如同 Cook 距离和 Welsch-Kuh 量度一样,也最好用图形方法来考察影响量度 H_i 。

例: 纽约州的河流数据

再次考虑拟合 (4.18) 中关于平均的氮浓度 Y 与商业或工业用地面积的百分比 X_4 之间关系的简单回归模型。 Y 关于 X_4 的散点图及相应的最小二乘回归直线在图 4.5 中给出。图 4.5 中,第 4 号观测 (Neversink 河) 与第 5 号观测 (Hackensack 河) 与其他大块数据点离得很远。第 7, 3, 8, 6 号观测也比较稀疏地散布在图的左上部。由拟合模型 (4.18) 导出的上述三种影响量度列于表 4.4, 相应的序列图见图 4.7。没有一个 C_i 值超过其阈值 1 的。但是, C_i 的序列图 4.7(a) 明确显示第 4 号观测 (Neversink) 应被标记为强影响观测。该观测还超过了其 $DFITS_i$ 的阈值 $2\sqrt{(p+1)/(n-p-1)} = 2/3$ 。正如从图 4.7 所见,第 5 号观测 (Hackensack) 没有被 C_i 或 $DFITS_i$ 标记出来。这是因为它的高杠杆值使其残差很小,还因为这两个量度的乘积性质。 H_i 的序列图 4.7(c) 显示,第 5 号观测 (Hackensack) 是影响最强的,其次是第 4 号观测 (Neversink),这与散点图 4.5 是相符的。

表 4.4 纽约州的河流数据。拟合模型 (4.18) 导出的影响量度: Cook 距离 C_i , Welsch-Kuh 量度 $DFITS_i$, 与 Hadi 影响量度 H_i .

行	C_i	$DFITS_i$	H_i	行	C_i	$DFITS_i$	H_i
1	0.00	0.01	0.06	11	0.02	0.19	0.13
2	0.00	-0.01	0.07	12	0.02	-0.21	0.14
3	0.10	0.49	0.58	13	0.02	-0.22	0.15
4	0.56	-1.14	0.77	14	0.02	-0.19	0.13
5	0.02	0.22	2.04	15	0.02	-0.22	0.16
6	0.01	0.15	0.10	16	0.01	-0.12	0.09
7	0.17	0.63	0.60	17	0.02	-0.18	0.12
8	0.07	0.40	0.37	18	0.01	-0.12	0.09
9	0.00	-0.02	0.07	19	0.04	-0.27	0.19
10	0.00	0.09	0.08	20	0.01	0.15	0.11



(a) Cook 距离 C_i (b) Welsch-Kuh 量度 $DFITS_i$ (c) Hadi 的影响量度 H_i

图 4.7 纽约州的河流数据: 各影响量度的序列图

4.10 位势 - 残差图

(4.24) 中 H_i 的公式使人想到一个简单的图形工具, 可帮助区分不寻常的观测是高杠杆点, 异常点, 或两者皆是。该图形称为位势 - 残差图 (P-R 图) (Hadi, 1992), 因为它是以下两者的散点图:

位势函数

$$\frac{p_{ii}}{1 - p_{ii}}$$

关于

残差函数

$$\frac{p+1}{1 - p_{ii}} \frac{d_i^2}{1 - d_i^2}$$

P-R 图与 Gray (1986) 及 McCulloch and Meeter (1983) 建议的 L-R 图 (杠杆 - 残差图) 是有关的。L-R 图是 p_{ii} 关于 d_i^2 的散点图。这两种图形的比较参见 Hadi (1992)。

作为示例, 由拟合模型 (4.18) 获得的 P-R 图见图 4.8。第 5 号观测, 是一个高杠杆点, 正位于该图的左上角。四个异常观测 (3, 7, 4, 8) 位于该图的右下部。

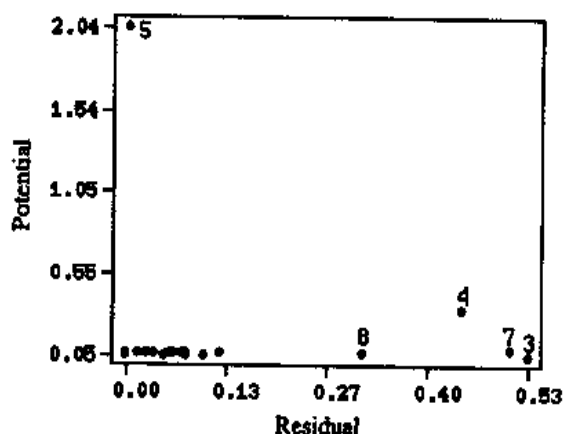


图 4.8 纽约州的河流数据: 位势 - 残差图

现在清楚了, 某些特别的点可被标记为异常点, 高杠杆点, 或强影响点。杠杆值与各种影响量度的主要用途是, 它们为分析者全面、生动地描述了不同点在整个拟合过程中所起的作用。对于落入其中某一类的任何一个观测, 我们应当仔细检查其准确度 (过失误差、抄写错误), 关联度 (它是否属于该数据集), 以及特殊含义 (非正常条件、罕见的情况)。异常点总应该仔细检查。高杠杆点、但非强影响点不会带来什么问题。应对高杠杆点且强影响点作调查, 因为这些点就预测变量而言是异常的, 而且它们也影响拟合。为了了解所作分析对于这些点的敏感性, 应当去掉这些讨厌的点后拟合模型, 再考察获得的系数。

4.11 如何处理异常点?

异常点和强影响观测不应该机械被删除或自动降低权重, 因为它们不一定是坏的观测。相反, 如果它们是准确的, 它们就可能是数据中含信息最多的点。比如, 他们可能指出数据并非来自正态总体, 或者模型不是线性的。我们采用下例描述的指数增长数据来说明异常点及强影响观测可能是数据中含信息最多的点。

例: 指数增长数据

图 4.9 是两变量的散点图, 其一是某群体的大小 Y , 另一个是时间 X 。正如在散点图上看到的, 数据的主体显示了群体大小与时间之间的某种线性关系, 如

图 4.9 中直线所示。按此模型来看，右上角的 22 和 23 两点是异常点。然而，如果这两点是正确的，那么它们则是数据集中仅有的、显示着这批数据可能服从某种非线性（譬如，指数）模型的观测，正如图中所示。把这想像为一个细菌的群体，它在一段时间内增长非常缓慢，但过了某个时间上的临界点之后，迅猛增长。

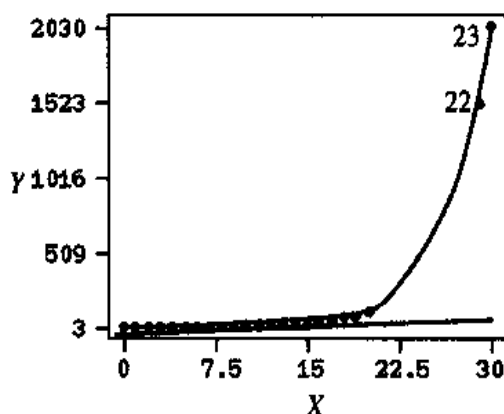


图 4.9 群体大小 Y 关于时间 X 的散点图。曲线是由对全部数据拟合一个指数函数得到的。直线是删去 22 和 23 号观测后的最小二乘直线。

一旦鉴别出了异常点和强影响观测之后，如何处理它们呢？因为异常点和强影响观测可能是数据集中信息最丰富的观测，因而不应该不加说明、自动地抛弃它们。相反，应当通过考察，判断它们为何是异常的或强影响的。根据这些考察才可能采取合适、正确的措施。这些正确的措施包括：改正数据中的错误，删除异常点或降低它们的权重，变换数据，考虑不同的模型，以及重新设计试验或抽样调查、收集更多数据。

4.12 变量在回归方程中的作用

正如我们已指出过的，变量可以依次、逐个地被引进回归方程。在实践中常常提出的一个问题是：给定一个目前包含 p 个预测变量的回归模型，那么从该模型中删除某个变量或对该模型添加某个变量的效应如何呢？通常，是对模型中的每个变量计算 t -检验值来回答该问题的。若 t -检验的绝对值很大，则变量被保留，否则变量被删除。这种做法仅当潜在的假定成立时才是有效的。因此，应当结合合适的图形来解释 t -检验值。人们提出了两种图，它们能直观地给出这方面的信息，而且常常很有启发性。在决定一个回归方程中某个变量是该保留还是剔除时，它们可作为 t -检验的补充。第一种图称为添加变量图，第二种称为残差加分量图。

4.12.1 添加变量图

添加变量图，由 Mosteller and Tukey (1977) 提出，能让我们形象地看到，一个将考虑被引进的新变量，其回归系数的量级大小。图中各点的最小二乘直线的

斜率等于该新变量之回归系数的估计。同时,该图也标示出在决定量级大小时起关键作用的数据点。我们可以对每个预测变量 X_j 建立一个添加变量图。 X_j 的添加变量图本质上是两组不同残差的散点图。第一组是 Y 关于除 X_j 之外的所有预测变量的回归之残差。我们称该组为 Y -残差。第二组残差由 X_j 关于所有其他预测变量的回归获得,这时暂且把 X_j 当作一个响应变量。我们称这一组为 X_j -残差。这样, X_j 的添加变量图不过就是 Y -残差关于 X_j -残差的散点图。因此,如果我们有 p 个预测变量,我们能够建立 p 张添加变量图,每个变量一张。

注意, X_j 的添加变量图中 Y -残差反映的是 Y 中未被除 X_j 外的所有预测变量解释的那个部分。类似地, X_j -残差反映 X_j 中未被其他预测变量解释的那个部分。若对 X_j 的添加变量图中的数据点拟合最小二乘回归直线,该直线的斜率为 $\hat{\beta}_j$,即在作 Y 关于包括 X_j 在内的所有预测变量的回归时, X_j 之回归系数的估计。这是对我们 3.5 节中谈到的偏回归系数的一种有启发性的、但等价的解释。

假如变量被引进方程的话,图中数据点的斜率给出了其回归系数的量级大小。因此,添加变量图中线性关系越强,则 X_j 对于已包含了其他预测变量的回归方程的额外贡献越大。若散点图显示无显著斜率,那么该变量在模型中不太可能有用。该散点图也将直观地显示出,在决定这斜率及其相应的 t -检验值时,哪些数据点影响最强。添加变量图也被称为偏回归图。我们顺便提一下,实际上不必要去做拟合。这些残差可以很简单地在对 Y 关于预测变量全集作拟合的计算过程中获得。详细讨论可参阅 Veleman and Welsch (1981) 与 Chatterjee and Hadi (1988)。

4.12.2 残差加分量图

残差加分量图由 Ezekiel (1924) 提出,是回归分析中最早的图形方法之一。由 Larsen and McCleary (1972) 重新起用,他们称之为偏残差图。按照 Wood (1973) 的提法,我们称之为残差加分量图,因为这个名字更切合其含义。

X_j 的残差加分量图是 $(e + \hat{\beta}_j X_j)$ 关于 X_j 的散点图,其中 e 为 Y 关于所有预测变量的普通最小二乘回归的残差, $\hat{\beta}_j$ 为该回归中 X_j 的系数。注意, $\hat{\beta}_j X_j$ 是第 j 个预测变量对于拟合值的贡献(分量)。就像在添加变量图中一样,该图中数据点的斜率为 $\hat{\beta}_j$,即 X_j 的回归系数。除了能图示斜率外,该图还能显示 Y 与 X_j 之间是否存在非线性关系。因而该图能够建议可能的线性化变换。然而,在添加变量图中无法显示非线性关系,因为其横轴表示的并非变量本身。这两种图都很有用,但在发现欲引进模型的变量中的非线性性方面,残差加分量图比添加变量图更敏感。而添加变量图更易于解释,且能指出强影响观测。

例: 苏格兰山地赛马数据

苏格兰山地赛马数据是 1984 年在苏格兰举行的 35 场赛马的纪录,包括一个响应变量(最高时间纪录,单位:秒)及两个解释变量(里程,以英里为单位,及攀登高度,以英尺为单位)。数据集在表 4.5 中给出。因为该数据集是三维的,让我们首先拿数据的三维旋转图作为一种探索性工具来作考察。该旋转图的一个有意思的方向显示于图 4.10。图中标出了五个观测。明显地,第 7、18 号观测是异

表 4.5 苏格兰山地赛马数据

行 Row	赛事 Race	时间 Time	里程 Distance	高度 Climb
1	Greenmantle New Year Dash	965	2.5	650
2	Carnethy	2901	6	2500
3	Craig Dunain	2019	6	900
4	Ben Rha	2736	7.5	800
5	Ben Lomond	3736	8	3070
6	Goatfell	4393	8	2866
7	Bens of Jura	12277	16	7500
8	Cairnpapple	2182	6	800
9	Scolty	1785	5	800
10	Traprain Law	2385	6	650
11	Lairig Ghru	11560	28	2100
12	Dollar	2583	5	2000
13	Lomonds of Fife	3900	9.5	2200
14	Cairn Table	2648	6	500
15	Eildon Two	1616	4.5	1500
16	Cairngorm	4335	10	3000
17	Seven Hills of Edinburgh	5905	14	2200
18	Knock Hill	4719	3	350
19	Black Hill	1045	4.5	1000
20	Creag Beag	1954	5.5	600
21	Kildoon	957	3	300
22	Meall Ant-Suiche	1674	3.5	1500
23	Half Ben Nevis	2859	6	2200
24	Cow Hill	1076	2	900
25	North Berwick Law	1121	3	600
26	Creag Dubh	1573	4	2000
27	Burnswark	2066	6	800
28	Largo	1714	5	950
29	Criffel	3030	6.5	1750
30	Achmony	1257	5	500
31	Ben Nevis	5135	10	4400
32	Knockfarrel	1943	6	600
33	Two Breweries Fell	10215	18	5200
34	Cockleroi	1686	4.5	850
35	Moffat Chase	9590	20	5000

常点，它们（在时间方向上）远离其他大部分点所喻示的那个平面。第 7 号观测在高度方向上偏离较远。图中第 33、31 号观测也是异常点，但程度轻些。尽管第 11、31 号观测邻近其他大部分点喻示的那个平面，但它们与该平面上的其他点相距颇远。（第 11 号观测主要在里程方向上、第 31 号观测在高度方向上偏离较远。）旋转图清楚地显示了数据中含有不寻常的点（异常点、高杠杆点与（或）强影响观测）。

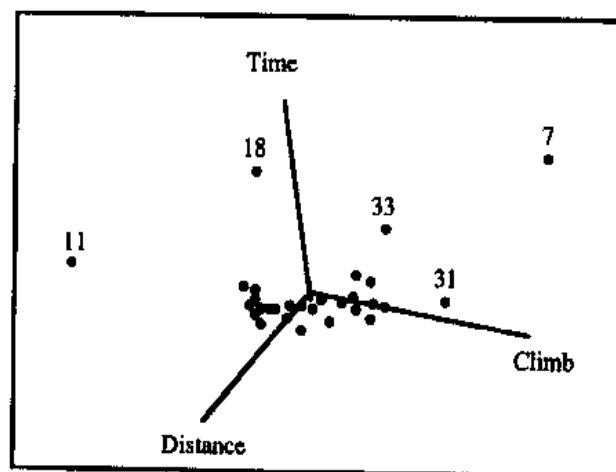


图 4.10 苏格兰山地赛马数据的旋转图

拟合的方程为

$$\text{Time} = -539.483 + 373.073\text{Distance} + 0.662888\text{Climb}. \quad (4.25)$$

我们希望处理的问题是：一个预测变量在另一个变量被纳入模型时作用是否显著？两个变量的 t -检验值分别为 10.3 和 5.39，表明显著性很高。这意味着对这两个变量而言，上述问题的回答都是肯定的。该结论的合理性可通过对相应的添加变量图与残差加分量图的考察而强化。这两个图分别在图 4.11 与图 4.12 给出。譬如，在图 4.11(a) 关于里程的添加变量图中，纵轴是时间关于高度（另一个预测变量）的回归中得到的残差，横轴是里程关于高度的回归中得到的残差。类似地，在关于高度的添加变量图中，用来作图的量是时间关于里程回归的残差以及高度关于里程回归的残差。

可以发现，四个图中均有很强的线性趋势，支持由上面的 t -检验得出的结论。但这些图也表明存在一些点可能会影响我们的结果和结论。第 7、11、18 场赛事明显地突出。这些点的序号已标于图上。第 31、33 场赛事也值得怀疑但程度较轻。根据对由前面拟合的方程所获得的 P-R 图（图 4.13）的考察，可将第 11 场赛事判为高杠杆点，第 18 场赛事为异常点，第 7 场赛事两者皆是。在继续展开分析前，这些点应被细细地审查。

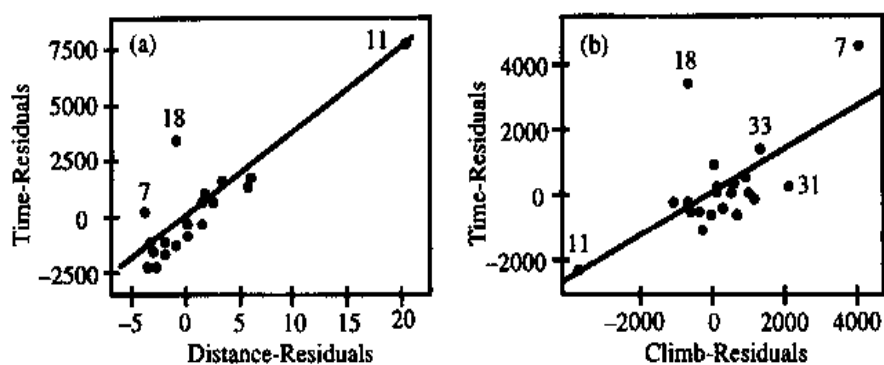


图 4.11 苏格兰山地赛马数据: 添加变量图 (a) 里程 (b) 高度。

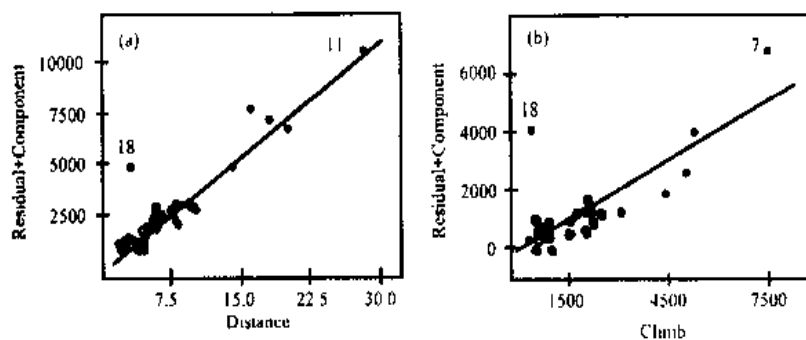


图 4.12 苏格兰山地赛马数据: 残差加分量图 (a) 里程 (b) 高度。

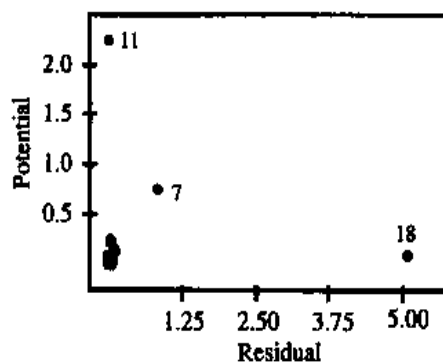


图 4.13 苏格兰山地赛马数据: 位势 - 残差图

4.13 添加一个预测变量的效应

我们来对在一个回归方程中引进一个新变量的效应作一般的讨论。应当提出两个问题: (a) 新变量的回归系数是否显著? (b) 新变量的引进是否本质地改变了

方程中原有变量的回归系数? 根据对上述两个问题的回答, 在一个方程中引进一个新变量时, 有四种可能的结果:

- 状况 A: 新变量的回归系数不显著, 且其余的回归系数与前次相比没有大的改变。在这些条件下, 新变量不应该被引进回归方程, 除非有其他外部条件 (譬如, 出于理论或宗旨的考虑) 强行将之纳入。
- 状况 B: 新变量的回归系数显著, 且以前引进的变量的回归系数起了本质变化。在这种状况下, 新变量应该保留, 但应对共线性^①作考察。如果没有共线性的证据, 该变量应被纳入方程, 接着应再考察添加别的变量的可能性。另一方面, 如果变量间有共线性关系, 那么应该采取第 10 章中概括的那些修正措施。
- 状况 C: 新变量的回归系数显著, 且以前引进的变量的回归系数没有本质改变。这是最理想的状况, 它出现在新变量与以前引进的变量间不相关的时候。此时, 新变量应被保留在方程中。
- 状况 D: 新变量的回归系数不显著, 但以前引进的变量的回归系数起了本质变化。这是明显的共线性的证据, 必须采取修正措施, 方能解决新变量在回归方程中的去留问题。

根据这些讨论, 显而易见, 一个变量对于回归方程的效应决定着其被纳入方程的适宜性。本章出现的结果影响着变量选择策略的设计。变量选择方法在第 11 章中介绍。

4.14 稳健回归

识别异常点和强影响观测, 另一类有用的方法 (这里没有讨论) 是稳健回归。这是一类拟合方法, 对高杠杆点赋以较小的权重。有大量关于稳健回归的文献。有兴趣的读者可以参阅以下一些书籍: Huber (1981), Hampel et al. (1986), Rousseeuw and Leroy (1987), Staudte and Sheather (1990), Birkes and Dodge (1993)。我们还必须提到以下论文: Krasker and Welsch (1982), Coakley and Hettmansperger (1993), Chatterjee and Mächler (1997), 以及 Billor, Chatterjee, and Hadi (1999), 这些文章采用了在拟合中限制影响和杠杆的想法。

习 题

4.1 分别对下列每个数据集, 检查标准回归假定是否合适:

- (a) 1.3.1 节中描述的牛奶产量数据。
- (b) 1.3.2 节中描述的、表 1.3 给出的劳动就业权法数据。
- (c) 1.3.3 节中描述的埃及人头盖骨数据。

^① 共线性源于预测变量间的高度相关。这问题在第 9、10 章中讨论。

- (d) 1.3.4 节中描述的国内移民数据。
- (e) 1.3.5 节中描述的、表 1.9 中给出的纽约州的河流数据。
- 4.2 找一个可用回归分析来回答其中感兴趣问题的数据集。然后：
- (a) 检查通常的多元回归假定是否合理。
- (b) 用至今所学的回归分析方法分析数据，回答感兴趣的问题。
- 4.3 考虑在 2.3 节中讨论过的计算机修理问题。在第二个抽样时期内，获得了变量修理时间与需修理元件数的另 10 个观测。因为所有观测都是按同样的方法从某固定的环境中收集来的，所有 24 个观测合起来形成一个数据集。数据列在表 4.6 中。
- (a) 拟合一个修理时间关于元件数的线性回归方程。
- (b) 检查每一条标准回归假定，指出哪一条或哪几条看上去不成立。

表 4.6 扩充的计算机修理时间数据：修理时间（分钟数）与修理元件数

行	元件数	修理时间	行	元件数	修理时间
1	1	23	13	10	154
2	2	29	14	10	166
3	3	49	15	11	162
4	4	64	16	11	174
5	4	74	17	12	180
6	5	87	18	12	176
7	6	96	19	14	179
8	6	97	20	16	193
9	7	109	21	17	193
10	8	119	22	18	195
11	9	149	23	18	198
12	9	145	24	20	205

- 4.4 为寻找一个回归数据集中的不寻常点，分析者考察了 P-R 图（见图 4.14）。请指明图中每个不寻常点的类型。

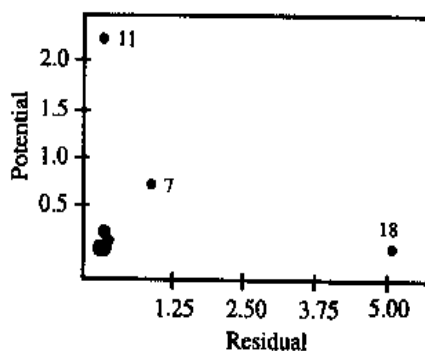


图 4.14 习题 4.4 中用到的 P-R 图

4.5 请指出哪种图或哪些图能用于核实下列每一条假定。对每张图,简述一个相应假定合适的例子和相应假定明显不合适的例子。

- (a) 响应与各预测变量之间存在线性关系。
- (b) 各观测之间相互独立。
- (c) 误差项方差相同。
- (d) 误差项互不相关。
- (e) 误差项正态分布。
- (f) 各观测对最小二乘结果影响相同。

4.6 下列图形常用于核实 Y 关于 X_1, X_2, \dots, X_p 的普通最小二乘回归中的某些假定:

- 1. Y 与每个预测因子 X_j 的散点图。
- 2. 变量 X_1, X_2, \dots, X_p 间的散点图矩阵。
- 3. 内标准化残差的正态概率图。
- 4. 残差关于拟合值的散点图。
- 5. 位势 - 残差图。
- 6. Cook 距离的序列图。
- 7. Hadi 影响量度的序列图。

对每种图形:

- (a) 可以核实什么假定?
- (b) 画一张样图,其中假定看似没有不妥。
- (c) 画一张表明假定不成立的样图。

4.7 再次考虑习题 3.14 描述的、表 3.17 给出的香烟消费数据。

- (a) 你预期销售额分别与每个解释变量之间有怎样的关系(譬如正、负等)?请作解释。
- (b) 计算两两变量间的相关系数矩阵,作相应的散点图矩阵。
- (c) 在两两变量间的相关系数与相应的散点图矩阵之间是否存在不一致?
- (d) 你在 (a) 中的预期与你在相关系数矩阵及相应的散点图矩阵中之所见是否存在差异?
- (e) 作销售额关于六个预测变量的回归。你在 (a) 中的预期与你从所有预测变量的回归系数中所看到的是否存在差异?若存在,请解释这种不一致。
- (f) 你将如何解释各回归系数与销售额和单个预测变量的相关系数这两者之间的差异?
- (g) 对习题 3.14 你所作的检验与得到的结论是否有问题?

4.8 再次考虑习题 3.12 描述的、表 3.14 给出的薪水数据。

- (a) 画 P-R 图。鉴定所有不同寻常的观测,并将之归为异常点、高杠杆点和(或)强影响观测。
- (b) 在习题 3.12 中你所作的检验与得到的结论是否有问题?

4.9 对下列两条陈述,或作数学证明,或用习题 3.14 中描述的、表 3.17 中给出的香烟消费数据从数值上论证其正确性:

- (a) 普通最小二乘回归的残差之和为零。
 (b) $\hat{\sigma}^2$ 与 $\hat{\sigma}_{(i)}^2$ 的关系为

$$\hat{\sigma}_{(i)}^2 = \hat{\sigma}^2 \left[\frac{n-p-1-r_i^2}{n-p-2} \right]. \quad (4.26)$$

4.10 鉴别表 4.7 中的数据集中的不寻常观测。

表 4.7 习题 4.10 用的数据

行	Y	X	行	Y	X
1	8.11	0	7	9.60	19
2	11.00	5	8	10.30	20
3	8.20	15	9	11.30	21
4	8.30	16	10	11.40	22
5	9.40	17	11	12.20	23
6	9.30	18	12	12.90	24

4.11 考虑表 4.5 中的苏格兰山地赛马数据。选择某个观测序号 i (譬如, $i = 33$ 表示选中了第 33 号观测), 并建立一个示性变量 (哑变量) U_i 。 U_i 的第 i 个值取 1, 其余值均取 0。现在考虑比较下列两个模型:

$$H_0: \text{Time} = \beta_0 + \beta_1 \text{Distance} + \beta_2 \text{Climb} + \varepsilon, \quad (4.27)$$

$$H_1: \text{Time} = \beta_0 + \beta_1 \text{Distance} + \beta_2 \text{Climb} + \beta_3 U_i + \varepsilon. \quad (4.28)$$

记 r_i^* 为由拟合模型 (4.27) 获得的第 i 个外标准化残差。证明 (或用一个例子证实)

- (a) 对模型 (4.28) 中 $\beta_3 = 0$ 的 t -检验值与从模型 (4.27) 获得的第 i 个外标准化残差是一样的, 即 $t_3 = r_i^*$ 。
 (b) 用于比较模型 (4.27) 与 (4.28) 的 F -检验值可简化为第 i 个外标准化残差之平方, 即 $F = r_i^{*2}$ 。
 (c) 去掉第 i 个观测后, 对苏格兰山地赛马数据拟合模型 (4.27)。
 (d) 证明模型 (4.28) 中 $\beta_0, \beta_1, \beta_2$ 的估计值与 (c) 中得到的一样。因此, 添加一个相应于第 i 个观测的示性变量等价于删去该观测!
- 4.12 考虑表 4.8 中的数据, 其中包括一个响应变量 Y 与六个预测变量。数据可从本书网站上获得。首先考虑拟合一个 Y 关于全部六个 X -变量的线性模型。
 (a) 哪些最小二乘假定看上去是不合适的 (假如有的话)?
 (b) 计算 $r_i, C_i, DFITS_i$ 和 H_i 。
 (c) 作 $r_i, C_i, DFITS_i$ 和 H_i 的序列图, 以及位势-残差图。
 (d) 鉴别数据中所有不寻常的观测, 并区分类型 (即异常点、高杠杆点等等)。
- 4.13 再次考虑表 4.8 中的数据。现在假定我们拟合一个 Y 关于前三个 X -变量的线性模型。回答下列问题, 并用适当的添加变量图解释理由。

表 4.8 习题 4.12-4.14 用的数据

行	Y	X_1	X_2	X_3	X_4	X_5	X_6
1	443	49	79	76	8	15	205
2	290	27	70	31	6	6	129
3	676	115	92	130	0	9	339
4	536	92	62	92	5	8	247
5	481	67	42	94	16	3	202
6	296	31	54	34	14	11	119
7	453	105	60	47	5	10	212
8	617	114	85	84	17	20	285
9	514	98	72	71	12	-1	242
10	400	15	59	99	15	11	174
11	473	62	62	81	9	1	207
12	157	25	11	7	9	9	45
13	440	45	65	84	19	13	195
14	480	92	75	63	9	20	232
15	316	27	26	82	4	17	134
16	530	111	52	93	11	13	256
17	610	78	102	84	5	7	266
18	617	106	87	82	18	7	276
19	600	97	98	71	12	8	266
20	480	67	65	62	13	12	196
21	279	38	26	44	10	8	110
22	446	56	32	99	16	8	188
23	450	54	100	50	11	15	205
24	335	53	55	60	8	0	170
25	459	61	53	79	6	5	193
26	630	60	108	104	17	8	273
27	483	83	78	71	11	8	233
28	617	74	125	66	16	4	265
29	605	89	121	71	8	8	283
30	388	64	30	81	10	10	176
31	351	34	44	65	7	9	143
32	366	71	34	56	8	9	162
33	493	88	30	87	13	0	207
34	648	112	105	123	5	12	340
35	449	57	69	72	5	4	200
36	340	61	35	55	13	0	152
37	292	29	45	47	13	13	123
38	688	82	105	81	20	9	268
39	408	80	55	61	11	1	197
40	461	82	88	54	14	7	225

来源: Chatterjee and Hadi (1988).

- (a) 我们是否应将 X_4 添进上面的模型? 如果是, 则将 X_4 留在模型中。
- (b) 我们是否应将 X_5 添进上面的模型? 如果是, 则将 X_5 留在模型中。
- (c) 我们是否应将 X_6 添进上面的模型?
- (d) 你会推荐哪个或哪些模型作为对 Y 可能最好的描述。利用上述结果, 若必要, 作补充分析。

4.14 考虑对表 4.8 中的数据拟合模型 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ 。记 u 为作 Y 关于 X_1, X_2 的回归得到的残差, v 为作 X_3 关于 X_1 的回归得到的残差。证明 (或用表 4.8 中的数据作为例子证实):

- (a) $\hat{\beta}_3 = \sum_{i=1}^n u_i v_i / \sum_{i=1}^n v_i^2$ 。
- (b) $\hat{\beta}_3$ 的标准误为 $\hat{\sigma} / \sqrt{\sum_{i=1}^n v_i^2}$ 。

5

定性预测变量

5.1 引言

在回归分析中，定性变量，或分类变量作为预测变量可能非常有用。如性别，婚姻状况，或政治派别等定性变量可以用示性变量或虚拟变量来表示。这些变量只取两个值，一般是 0 和 1，表明观测属于两个可能类别中的一个。示性变量的取值大小并不反映类别之间有定量的序，只是用来标识类别。例如，在对计算机程序员薪水的分析，可能将教育，工作年数，性别作为预测变量。性别变量可以量化：如是女性就取 1，如是男性就取 0。在回归方程中示性变量也可用以区分 3 个或 3 个以上的类别。例如，上面描述的回归分析可以用一个示性变量表示观测是来自系统程序员还是应用程序员。由性别和程序员的类型所决定的四种情况可以由两个示性变量结合起来表示，这一点将在本章的后面看到。

示性变量有多种用法。每当数据分析中出现定性变量，就可以考虑采用示性变量。我们将举例说明示性变量的一些应用并将建议一些别的应用。希望读者能从例题中认识到这种技术的普遍适用性。在第一个例子里，我们考察前面提到的薪水调查数据，并采用示性变量来表示影响回归关系的各种分类变量。第二个例子采用示性变量来分析和检验回归关系在总体的各个子集中的等价性。

我们仍然假定响应变量是定量连续变量，但预测变量可能是定量变量也可能是分类变量。有关响应变量是示性变量的情形将在第 12 章处理。

5.2 薪水调查数据

薪水调查数据来自对一家大公司计算机专业人员的薪水调查。调查的目的是明确并且量化那些决定薪水差异的变量。另外，这批数据也可用来检验公司的薪水管理制度是否得到遵行。数据见表 5.1 并可从本书的网站上得到^①。响应变量是薪水 (S)，预测变量是：(1) 工作年数 (X)，用年来衡量；(2) 教育 (E)，用 1 表

^① <http://www.ilr.cornell.edu/~hadi/RABE>

表 5.1 薪水调查数据

行	S	X	E	M	行	S	X	E	M
1	13876	1	1	1	24	22884	6	2	1
2	11608	1	3	0	25	16978	7	1	1
3	18701	1	3	1	26	14803	8	2	0
4	11283	1	2	0	27	17404	8	1	1
5	11767	1	3	0	28	22184	8	3	1
6	20872	2	2	1	29	13548	8	1	0
7	11772	2	2	0	30	14467	10	1	0
8	10535	2	1	0	31	15942	10	2	0
9	12195	2	3	0	32	23174	10	3	1
10	12313	3	2	0	33	23780	10	2	1
11	14975	3	1	1	34	25410	11	2	1
12	21371	3	2	1	35	14861	11	1	0
13	19800	3	3	1	36	16882	12	2	0
14	11417	4	1	0	37	24170	12	3	1
15	20263	4	3	1	38	15990	13	1	0
16	13231	4	3	0	39	26330	13	2	1
17	12884	4	2	0	40	17949	14	2	0
18	13245	5	2	0	41	25685	15	3	1
19	13677	5	3	0	42	27837	16	2	1
20	15965	5	1	1	43	18838	16	2	0
21	12336	6	1	0	44	17483	16	1	0
22	21352	6	3	1	45	19207	17	2	0
23	13839	6	2	0	46	19346	20	1	0

示高中毕业 (H.S.), 2 表示学士学位 (B.S.), 3 表示更高学位; (3) 管理 (M), 1 表示此人在管理职位, 而 0 表示其他职位。我们将采用回归分析来衡量这三个变量对于薪水的影响。

薪水和工作年数之间将采用线性关系。我们假定工作年数每增加一年就增加一定的薪水。教育也用线性关系处理。如果在回归方程中教育变量用原始形式, 那就是假定教育程度每提高一级就增加一定的薪水。也就是说, 如果其他变量保持为常数, 则薪水与教育之间的关系是线性的。这样的解释是可能的, 但或许过于严格。取而代之, 我们将教育视为分类变量并定义两个示性变量表示三种教育水平。这两个变量使我们可以衡量受教育水平对薪水差异的影响, 而不论这种影响是否是线性的。管理变量也可以采用示性变量来表示, 1 表示受观测者在管理职位上, 0 表示在普通职位上。

当用示性变量表示一组类别时, 示性变量所需要的个数是分类个数减 1。例如, 前面提到的 3 个教育水平类别, 我们只需采用两个示性变量 E_1 和 E_2 来处

表 5.2 教育和管理 6 个组合类别的回归方程

类	E	M	回归方程	
1	1	0	$S = (\beta_0 + \gamma_1)$	$+\beta_1 X + \varepsilon$
2	1	1	$S = (\beta_0 + \gamma_1 + \delta_1)$	$+\beta_1 X + \varepsilon$
3	2	0	$S = (\beta_0 + \gamma_2)$	$+\beta_1 X + \varepsilon$
4	2	1	$S = (\beta_0 + \gamma_2 + \delta_1)$	$+\beta_1 X + \varepsilon$
5	3	0	$S = \beta_0$	$+\beta_1 X + \varepsilon$
6	3	1	$S = (\beta_0 + \delta_1)$	$+\beta_1 X + \varepsilon$

理, 其中

$$E_{i1} = \begin{cases} 1 & \text{如果第 } i \text{ 个人属于 H.S. 类,} \\ 0 & \text{否则,} \end{cases}$$

和

$$E_{i2} = \begin{cases} 1 & \text{如果第 } i \text{ 个人属于 B.S. 类,} \\ 0 & \text{否则.} \end{cases}$$

正如前面所描述的那样, 这两个变量结合起来可以唯一地表示三个类别。对于 H.S., $E_1 = 1, E_2 = 0$; 对于 B.S., $E_1 = 0, E_2 = 1$; 对于更高学位, $E_1 = 0, E_2 = 0$ 。此外, 如果有第三个变量 E_3 , 依第 i 个人是否属于更高学位这一类, 定义其取值 E_{i3} 为 1 或 0, 则对于每一个人我们有 $E_1 + E_2 + E_3 = 1$ 。那么 $E_3 = 1 - E_1 - E_2$, 这表明其中的一个变量是多余的^①。类似地, 用来区分两个管理类别的示性变量只需要一个。示性变量取值为 0 的类别称为基础类别(base category) 或对照组(control group), 因为示性变量的回归系数是相对于对照组的增量。

表 5.3 薪水调查数据的回归分析

变量	系数	标准误	t -检验	p -值
常数	11031.800	383.2	28.80	< 0.0001
X	564.184	30.5	17.90	< 0.0001
E_1	-2996.210	411.8	-7.28	< 0.0001
E_2	147.825	387.7	0.38	0.7049
M	6883.530	313.9	21.90	< 0.0001
$n = 46$	$R^2 = 0.957$	$R_a^2 = 0.953$	$\hat{\sigma} = 1027$	$d.f. = 41$

采用前面定义的示性变量, 回归模型为

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \varepsilon. \quad (5.1)$$

^① 如果同时使用 E_1, E_2 和 E_3 , 则在预测变量之间存在一个完全的线性关系, 这是共线性的一个极端情形, 我们将在第 9 章讲述。

根据 (5.1) 中的示性变量不同的取值, 6 种类别 (3 种教育类别, 2 种管理类别) 的每一类有不同的回归方程, 如表 5.2 所示。根据提出的模型, 我们看到, 在经工作年数调整之后, 示性变量有助于将基本薪水水平表达为教育和管理的函数。

模型 (5.1) 的回归结果见表 5.3。薪水差异可以由模型来解释的程度很高 ($R^2 = 0.957$)。分析进行到此时, 我们应该研究诸残差的模式来检查模型设定是否合理。但是, 我们暂时将模型的检查推后并假定模型是令人满意的, 以便于讨论对回归结果的解释。后面, 我们将反过来对残差进行分析并发现这个模型需要改动。

X 的系数是 546.16。也就是说, 每增加一年的工作, 估计年薪就会增加 \$546。其他系数可以通过表 5.2 来解释。管理示性变量的系数 δ_1 的估计为 6883.50。依表 5.2 我们将这一量解释为管理职位的年薪比普通职位年薪的平均增加值。对于教育变量, γ_1 度量的是 H.S. 与更高学位的薪水差别, γ_2 度量的是 B.S. 与更高学位的薪水差别。差 $\gamma_2 - \gamma_1$ 度量的是 H.S. 与 B.S. 的薪水差别。从回归分析的结果看, 对于计算机专业人员, 一个拥有更高学位的职员可以比拥有高中文凭的职员平均多挣 \$2996, 而一个学士学位的职员和一个更高学位的职员薪水平均相差 \$148 (这一差别不具有统计显著性, $t = 0.38$), 一个学士学位比一个具有高中文凭的人多挣 \$3144。这些薪水差别对有相同工作年数的人都成立。

5.3 交互作用变量

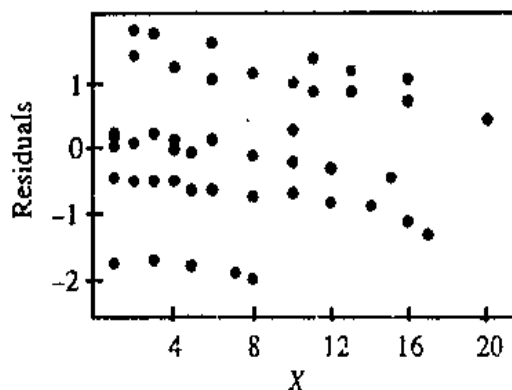


图 5.1 对工作年数 (X) 的标准化残差图

现在反过来讨论模型的设定问题, 考虑图 5.1, 这是对 X 画的残差图。这幅图显示存在三个或更多的残差水平。前面定义的示性变量可能并不足以解释教育和管理的薪水差异的影响。实际上, 每一个残差可以由教育和管理的六个组合之一识别出来。为了清楚地看出这一点, 我们将相对于新的分类变量 (一个新的分类变量, 它对于六种组合中每一种分别取一个不同的值) 的残差图画出来。见图 5.2。从中可以看出残差的大小按新的分类变量聚类。因为教育变量和管理变量的组合并未在原模型中得到满意的处理, 在六种组合的每一种组合中, 残差或者几乎全

为正或者几乎全为负。这表明 (5.1) 中给出的模型并未充分解释薪水和工作年数, 教育, 以及管理变量之间的关系。该图指出数据中一些隐含的数据结构并未被挖掘出来。

这幅图明示着教育和管理对于薪水的影响是不可加的。但注意到在模型 (5.1) 及表 5.2 对它的进一步解释中, 这两个变量的效应都是可加的。例如, 管理的影响是由 δ_1 衡量的, 与教育达到的水平无关。这些变量的非可加性效应, 即所谓的乘积效应或交互效应, 可以通过构造另外的变量来衡量。交互作用变量定义为已有示性变量的乘积 ($E_1 \cdot M$) 和 ($E_2 \cdot M$)。将这两个变量包含在 (5.1) 的右边, 从而导出一个关于教育和管理不可加, 但却承认这两个变量有交互效应的模型。

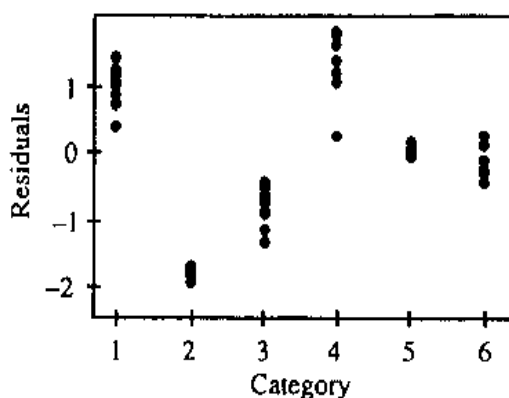


图 5.2 对教育 - 管理组合类别变量的标准化残差图

扩展后的模型为

$$S = \beta_0 + \beta_1 X + \gamma_1 E_1 + \gamma_2 E_2 + \delta_1 M + \alpha_1 (E_1 \cdot M) + \alpha_2 (E_2 \cdot M) + \varepsilon, \quad (5.2)$$

回归结果见表 5.4。扩展模型对 X 的标准化残差图见图 5.3。注意到第 33 个观测是异常点。此模型过高预测了这一点上的薪水。检查原数据集中这个观测, 发现这个人比其他特征相似的人年薪低几百元。为了确保这一个观测不过度影响回归估计, 将它去掉重新作回归。新的结果在表 5.5 中给出。

回归系数基本没有改变。但是, 误差标准差的估计减少为 \$67.28, 而且模型所能解释的变异比例达到 0.9998。对 X (图 5.4) 的残差图与原可加模型类似的残差图相比更令人满意。另外, 从教育 - 管理组合变量的残差图 (图 5.5) 来看, 每个组合类的残差关于 0 的分布对称。因此, 交互作用项的引入更加准确地反映了薪水变异。所以模型 (5.2) 较为充分地描述了薪水和工作年数, 教育以及管理之间的关系。

表 5.4 薪水数据的回归分析：扩展模型

变量	系数	标准误	<i>t</i> -检验	<i>p</i> -值
常数	11203.50	79.07	142.0	< 0.0001
<i>X</i>	496.98	5.57	89.3	< 0.0001
<i>E</i> ₁	-1730.69	105.30	-16.4	< 0.0001
<i>E</i> ₂	-349.03	97.57	-3.6	0.0009
<i>M</i>	7047.32	102.60	68.7	< 0.0001
<i>E</i> ₁ · <i>M</i>	-3065.99	149.30	-20.5	< 0.0001
<i>E</i> ₂ · <i>M</i>	1836.58	131.20	14.0	< 0.0001
<i>n</i> = 46	<i>R</i> ² = 0.999	<i>R</i> _a ² = 0.999	$\hat{\sigma}$ = 173.8	<i>d.f.</i> = 39

表 5.5 薪水数据的回归分析：扩展模型，观测 33 被删除

变量	系数	标准误	<i>t</i> -检验	<i>p</i> -值
常数	11199.70	30.54	367.0	< 0.0001
<i>X</i>	498.41	2.15	232.0	< 0.0001
<i>E</i> ₁	-1741.28	40.69	-42.8	< 0.0001
<i>E</i> ₂	-357.00	37.69	-9.5	< 0.0001
<i>M</i>	7040.49	39.63	178.0	< 0.0001
<i>E</i> ₁ · <i>M</i>	-3051.72	57.68	-52.9	< 0.0001
<i>E</i> ₂ · <i>M</i>	1997.62	51.79	38.6	< 0.0001
<i>n</i> = 45	<i>R</i> ² = 1.0	<i>R</i> _a ² = 1.0	$\hat{\sigma}$ = 67.13	<i>d.f.</i> = 38

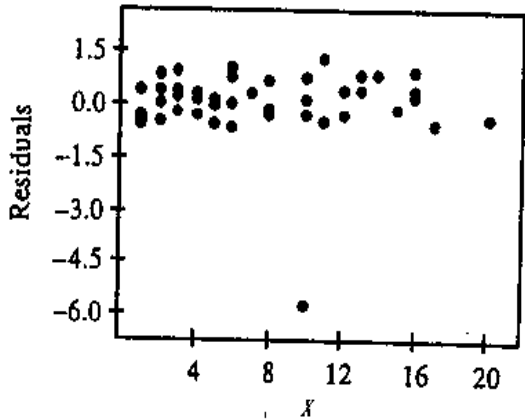


图 5.3 对工作年数的标准化残差图：扩展模型

误差的标准差估计为 \$67.28，至此，我们猜想已经揭示出公司现行并被仔细执行的薪水公式。采用 95% 置信区间，每一年工作年数的价值估计在 \$494.08 和 \$502.72 之间。这些大约 \$500 的增加值被加在与六个教育 - 管理组合类相应的起薪上。由于最后的回归模型不是可加的，直接对示性变量的系数进行解释是相当困难的。为了看到定性变量如何影响薪水差异，我们用这些系数形成六种组合类

表 5.6 采用 (5.2) 中的不可加模型对基本薪水的估计

类	E	M	系数	基本薪水的估计	标准误 (元)	95% 置信区间
1	1	0	$\beta_0 + \gamma_1$	9459	31	(9398, 9520)
2	1	1	$\beta_0 + \gamma_1 + \delta + \alpha_1$	13448	32	(13385, 12511)
3	2	0	$\beta_0 + \gamma_2$	10843	26	(10792, 10894)
4	2	1	$\beta_0 + \gamma_2 + \delta + \alpha_2$	19880	33	(19815, 19945)
5	3	0	β_0	11200	31	(11139, 11261)
6	3	1	$\beta_0 + \delta$	18240	29	(18183, 18297)

的每一类的基本薪水估计。这些结果和标准误以及置信区间在表 5.6 中给出。标准误的计算应用的是第 3 章附录中的式 (A.10)。

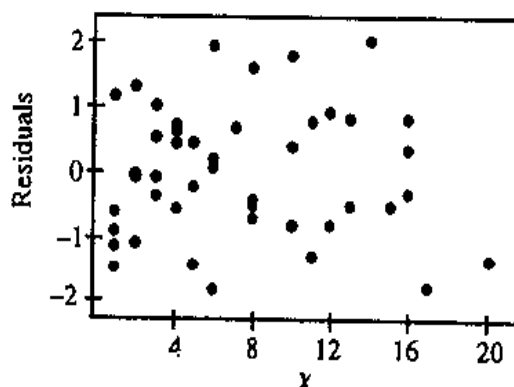


图 5.4 对工作年数的标准化残差图: 扩展模型, 观测 33 被去除

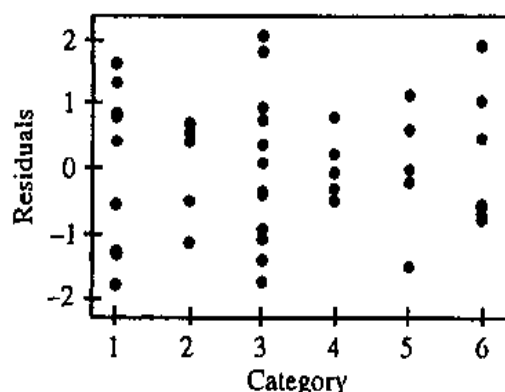


图 5.5 对教育 - 管理组合变量的标准化残差图: 扩展模型, 观测 33 被去除

采用带有示性变量和其交互项的回归模型, 已经将这次调查中所选择的计算

机专业人员的薪水变异几乎全部解释出来了。该模型对于数据解释的准确水平是罕见的！我们只能推想这家公司采用了精确的薪水管理制度并且得到严格执行。

回顾前面，我们可以得到一个等价的带有一组不同示性变量和回归参数的模型。我们可以定义五个变量，每一个取值 0 或 1，对应于六种教育 - 管理组合类别中的五种。表 5.6 中基本薪水的估计值和标准误将会是相同的。但以前的模型有以下的优势：它允许我们将三组预测变量 (1) 教育，(2) 管理，(3) 教育 - 管理的交互作用效应分离开。模型中的交互项是在我们发现原可加模型不能对薪水变异做出满意的解释之后添加进去的。一般来讲，我们从简单模型开始并在需要的情况下再循序渐进到较复杂的模型。我们总是希望保留具有可接受误差结构的最简单模型。

5.4 回归方程系统：两个组的比较

一组数据可能包含两个或多个不同的子集，其中每一组需要各自的回归方程。如果用 - 个回归关系去拟合合并的数据集，则会发生严重的偏差。可以采用示性变量完成关于这一问题的分析。对每个数据子集各自回归方程的分析可以应用于截面数据和时间序列数据。下面的例子讨论如何处理截面数据。对于时间序列数据的应用将在第 5.5 节中讨论

两组数据的模型可以在所有的方面都不同或只在某些方面有差异。本节我们讨论三种不同情形：

1. 每组数据都有各自不同的回归模型。
2. 模型具有相同的截距但是斜率不同。
3. 模型具有相同的斜率但是截距不同。

后面我们在只有一个定量预测变量的情况下举例说明如上的情形。这些思想可以直接推广到具有多个定量预测变量的情况。

5.4.1 具有不同斜率和不同截距的模型

我们用就业中的机会平等问题来举例说明这种情形。许多大型公司和政府机构在雇佣前实施一些测试筛选申请者。这一测试意在衡量申请者的职业才能，测试结果会影响是否雇佣的决定。联邦政府已经规定^① 这些测试：(1) 必须衡量与申请的工作直接相关的能力；(2) 绝不能在种族和国籍上区别对待。(1)、(2) 两项规定的可操作性定义是很难把握的，我们不试图解决那些可操作问题。我们把种族分成白种人和少数人种两组来研究，检验这两组中测试得分与工作业绩之间是否存在不同的回归关系，从而考察雇佣中是否存在种族歧视问题。

^① 1964 年公民权利法令，第七章的 Tower 修正案。

令 Y 代表工作业绩, 令 X 为雇佣前的测试得分。我们想比较

$$\begin{aligned} \text{模型 1(合并的): } y_{ij} &= \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, j = 1, 2, i = 1, 2, \dots, n_j, \\ \text{模型 2(少数人种): } y_{i1} &= \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}, \\ \text{模型 2(白种人): } y_{i2} &= \beta_{02} + \beta_{12} x_{i2} + \varepsilon_{i2}. \end{aligned} \quad (5.3)$$

图 5.6 描绘了这两个模型。在模型 1 中, 种族差异被忽略了, 数据被合并在一起, 只有一条回归线。在模型 2 中, 两个子集分别有各自的回归关系, 每一个都有不同的回归系数。我们假定每个子集中误差项的方差是相同的。

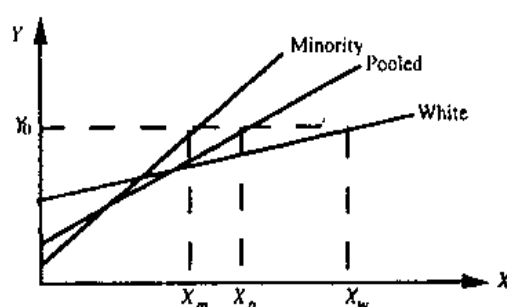


图 5.6 对雇佣者在雇佣前测试中的要求

在分析数据之前, 让我们简要地考虑一下在解释和应用结果时可能出现的错误类型。如图上所示, 设 Y_0 为对工作业绩的最低要求, 则若采用模型 1, 在测试中被接受的得分需大于 X_p 。但是, 如果实际上模型 2 是正确的, 则对于白种人要求的测试得分是 X_w , 而对于少数人种是 X_m 。采用 X_p 代替 X_w 和 X_m 表示对于白种人放宽了测试的要求而对于少数人种则加严了。如果用错误的模型来设定分数线, 则可能导致最终选择的不平等性, 所以需要仔细地检查数据。必须确定两组数据是存在不同的关系, 还是存在相同的关系。用一个估计的方程就可以充分表达合并数据。注意到不论选择模型 1 还是模型 2, X_m , X_w , 和 X_p 的估计值总有一定的抽样误差, 应当和适当的置信区间结合起来使用。(置信区间的构造在后面考虑。)

针对这一分析, 我们采用了一个特定的雇佣程序收集数据。20 个申请者试用 6 个星期。一个星期用来培训。剩下的 5 个星期用于工作。受观测者是采用一种与雇佣前测试得分无关的方法从一群申请者中挑选出来的。在培训期结束时做一次测验, 六个星期结束时进行工作业绩的评价。这两次得分结合起来形成工作业绩指标。(那些在六个星期期末表现令人不满意的雇员被除去。) 这些数据见表 5.7 并可在本书的网站上得到。我们将这些数据称为雇佣前测试数据。

我们想检验原假设 $H_0: \beta_{11} = \beta_{12}, \beta_{01} = \beta_{02}$, 对立假设为: 这些参数存在实质上的差别。这个检验可以通过示性变量实施。令 z_{ij} 取值 1 若 $j = 1$, 取值 0 若 $j = 2$ 。也就是说, Z 是一个新的变量, 对于少数人种申请者它取值为 1, 对于白种

表 5.7 雇佣前测试数据

行	测试	种族	JPERF	行	测试	种族	JPERF
1	0.28	1	1.83	11	2.36	0	3.25
2	0.97	1	4.59	12	2.11	0	5.30
3	1.25	1	2.97	13	0.45	0	1.39
4	2.46	1	8.14	14	1.76	0	4.69
5	2.51	1	8.00	15	2.09	0	6.56
6	1.17	1	3.30	16	1.50	0	3.00
7	1.78	1	7.53	17	1.25	0	5.85
8	1.21	1	2.03	18	0.72	0	1.90
9	1.63	1	5.00	19	0.42	0	3.85
10	1.98	1	8.04	20	1.53	0	2.95

人申请者它取值为 0。我们考虑两个模型,

$$\text{模型 1: } y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij},$$

$$\text{模型 3: } y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \delta(z_{ij} \cdot x_{ij}) + \varepsilon_{ij}. \quad (5.4)$$

变量 $(z_{ij} \cdot x_{ij})$ 代表组(种族)变量 Z 和雇佣前测试得分 X 的交互作用。注意到模型 3 和模型 2 等价。如果我们观测少数人种这一组, 可以看成 $x_{ij} = x_{i1}$ 和 $z_{ij} = 1$; 此模型 3 变成

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \gamma + \delta x_{i1} + \varepsilon_{i1} \\ &= (\beta_0 + \gamma) + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \\ &= \beta_{01} + \beta_{11} x_{i1} + \varepsilon_{i1}. \end{aligned}$$

这和模型 2 中针对少数人种的模型相同, 其中 $\beta_{01} = \beta_0 + \gamma$ 和 $\beta_{11} = \beta_1 + \delta$ 。类似地, 对于白种人这一组, 我们有 $x_{ij} = x_{i2}$, $z_{ij} = 0$, 模型 3 变成

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \varepsilon_{i2}.$$

这和模型 2 中针对白种人的模型相同, 其中 $\beta_{02} = \beta_0$ 和 $\beta_{12} = \beta_1$ 。因此, 模型 1 和 2 的比较等价于模型 1 和 3 的比较。模型 3 可以看成是一个完全模型 (FM), 模型 1 是受约束的模型 (RM), 因为模型 1 是由模型 3 令 $\gamma = \delta = 0$ 得到的。因此, 我们的原假设 H_0 变成 $H_0: \gamma = \delta = 0$ 。对这个假设可构造像第 3 章中两个模型比较那样的 F -检验统计量去检验。在本例中, 检验统计量是

$$F = \frac{[SSE(RM) - SSE(FM)]/2}{SSE(FM)/16},$$

它的自由度是 2 和 16。(为什么?) 继续对数据进行分析, 模型 1 和模型 3 的回归结果在表 5.8 和表 5.9 给出。两种情形下关于预测变量的残差图 (图 5.7 和图 5.8) 似乎都可以接受。模型 1 中右下方的一个残差需要做进一步分析。

表 5.8 回归结果，雇佣前测试数据：模型 1

变量	系数	标准误	t- 检验	p- 值
常数	1.03	0.87	1.19	0.2486
测试 (X)	2.36	0.54	4.39	0.0004
$n = 20$	$R^2 = 0.52$	$R_a^2 = 0.49$	$\hat{\sigma} = 1.59$	$d.f. = 18$

表 5.9 回归结果，雇佣前测试数据：模型 3

变量	系数	标准误	t- 检验	p- 值
常数	2.01	1.05	1.91	0.0736
测试 (X)	1.31	0.67	1.96	0.0677
种族 (Z)	-1.91	1.54	-1.24	0.2321
种族 · 测试 (X · Z)	2.00	0.95	2.09	0.0527
$n = 20$	$R^2 = 0.664$	$R_a^2 = 0.601$	$\hat{\sigma} = 1.41$	$d.f. = 16$

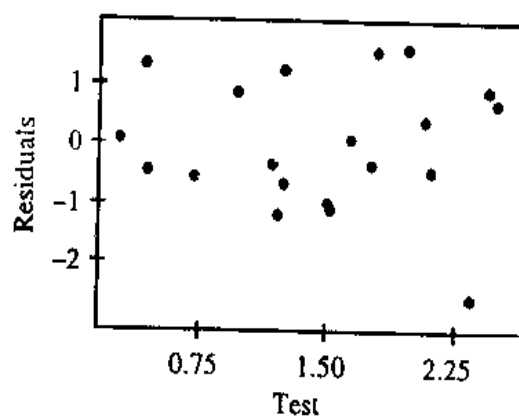


图 5.7 对测试得分的标准化残差图：模型 1

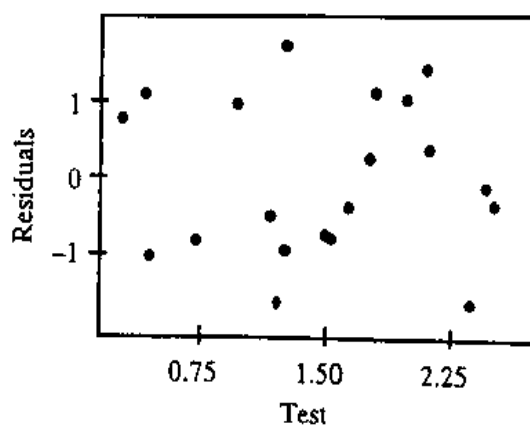


图 5.8 对测试得分的标准化残差图：模型 3

为了评价假设, 我们计算前面提出的 F -比, 它等于

$$F = \frac{(45.51 - 31.81)/2}{31.81/16} = 3.4,$$

它在略高于 5% 的水平下是显著的。因此, 基于这个检验, 我们可以得出结论: 这两组数据可能存在不同的回归关系。特别地, 对于少数人种我们有

$$Y_1 = 0.10 + 3.31X_1,$$

对于白种人我们有

$$Y_2 = 2.01 + 1.32X_2.$$

这个结果和我们在讨论有歧视问题时图 5.6 中描绘的情形非常相似。表示少数人种组关系的直线比白种人组的直线有较大的斜率和较小的截距。如果使用合并模型, 会出现在图 5.6 中讨论过的歧视问题。

尽管采用示性变量的正规方法导出了可能的结论是: 两组数据存在不同的回归关系, 但是我们对每一组数据并没有仔细分析。回想我们曾假定两组的方差相同。在这样的假定要求下两组样本间唯一的区别是回归系数的不同。图 5.9 给出了对示性变量的残差图。这两组残差看上去并没有差别。我们再仔细看一下每一组样本, 表 5.10 分别给出了每组样本的回归系数。图 5.10 和图 5.11 是残差图。回归系数当然是与模型 3 得到的一样。少数人种组和白种人组的误差标准差估计分别是 1.29 和 1.51。这两种情形下关于测试得分的残差图都是可接受的。在前面的分析中没有发现的一个有趣的现象是: 雇佣前测试得分可以解释少数人种样本组的大部分变异, 但是对于白种人样本组只是勉强有用。

表 5.10 分别的回归结果

样本	$\hat{\beta}_0$	$\hat{\beta}_1$	t_1	R^2	$\hat{\sigma}$	$d.f.$
少数人种	0.10	3.31	5.31	0.78	1.29	8
白种人	2.01	1.31	1.82	0.29	1.51	8

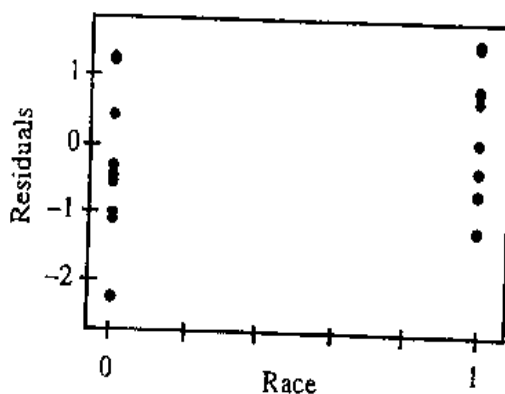


图 5.9 对种族的标准残差图: 模型 1

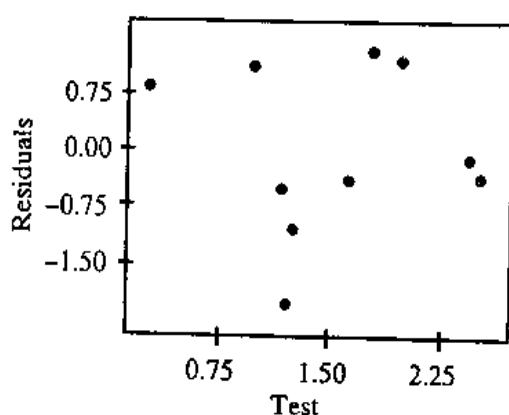


图 5.10 对测试得分的标准化残差图：只对少数人种作的回归

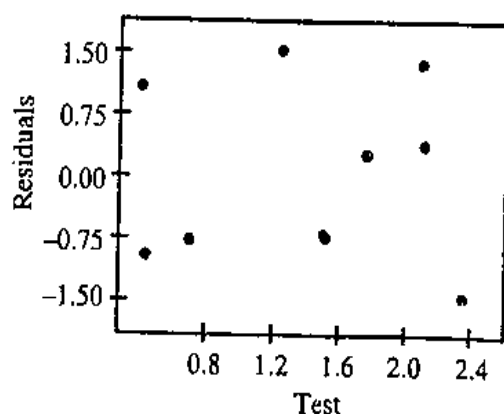


图 5.11 对测试得分的标准化残差图：只对白种人作的回归

我们前面的结论依然合理。这两个回归方程是不同的。不仅回归系数不同，而且残差均方也略显不同。更重要的是， R^2 有很大的不同。对于白种人样本组， $R^2 = 0.29$ 太小 ($t = 1.82$; 2.306 是要求的显著值) 以致于雇佣前测试得分不能视为工作业绩的一个充分的预测变量。这一发现和我们最初的目的有关，因为比较两组样本回归关系的前提是单独考虑每一组样本时这些关系应该较为合适。考虑雇佣前测试的合法性，我们的结论是，如果按照法律规定来应用，那么测试对不同种族不应该区别对待，但现在它将对两个种族群体造成有偏的结果。而且，基于这一发现，我们有理由说这个测试对于筛选白种人申请者是没有价值的。

最后，我们讨论在采用雇佣前测试的情况下，如何适当决定分数线。只考虑关于少数人种样本组的结果。如果设 Y_0 为可以接受的成功的最低工作业绩，那么由回归方程得 (再参见图 5.6)

$$X_m = \frac{Y_0 - \hat{\beta}_0}{\hat{\beta}_1},$$

其中 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是回归系数的估计值。 X_m 是为达到 Y_0 的要求可以接受的最低测

试得分。由于 X_m 是用带抽样变异的量定义的, 故 X_m 也有抽样变异。这种变异可以最容易通过构造 X_m 的置信区间来刻画。近似 95% 置信水平的置信区间的形式为 (Scheffé, 1959, p.52)

$$X_m \pm \frac{t_{(n-2, \alpha/2)}(\hat{\sigma}/n)}{\hat{\beta}_1},$$

其中 $t_{(n-2, \alpha/2)}$ 是 t -分布的分位点。 $\hat{\sigma}^2$ 是 σ^2 的最小二乘估计。如果 Y_0 定为 4, 则 $X_m = (4 - 0.10)/3.13 = 1.18$, 测试分数线 95% 的置信区间为 (1.09, 1.27)。

5.4.2 具有相同斜率和不同截距的模型

在前一小节, 我们处理的是两组样本具有截然不同模型的情形, 两个模型的系数完全不同, 如 (5.3) 中给出的模型 1 和模型 2, 亦如图 5.6 所画。现在假定存在一个理由令人相信这两组样本存在相同的斜率 β_1 , 我们希望检验假设: 两组样本具有相同截距, 也就是说, $H_0: \beta_{01} = \beta_{02}$ 。这种情形下, 我们比较

$$\begin{aligned} \text{模型 1(合并的): } y_{ij} &= \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, j = 1, 2; i = 1, 2, \dots, n_j, \\ \text{模型 2(少数人种): } y_{i1} &= \beta_{01} + \beta_1 x_{i1} + \varepsilon_{i1}, \\ \text{模型 2(白种人): } y_{i2} &= \beta_{02} + \beta_1 x_{i2} + \varepsilon_{i2}. \end{aligned} \quad (5.5)$$

这两个模型具有相同的斜率 β_1 但是截距 β_{01} 和 β_{02} 不同。采用前面定义的示性变量 Z , 我们可以将模型 2 写为

$$\text{模型 3: } y_{ij} = \beta_0 + \beta_1 x_{ij} + \gamma z_{ij} + \varepsilon_{ij}. \quad (5.6)$$

(5.6) 的模型 3 中没有交互作用变量 ($z_{ij} \cdot x_{ij}$)。如果像 (5.4) 那样有这一项, 将会产生斜率和截距都不相同的两个模型。

模型 2 和模型 3 的等价性可以从下面看出: 当 $x_{ij} = x_{i1}$ 和 $z_{ij} = 1$ 时, 模型 3 成为

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \gamma + \varepsilon_{i1} \\ &= (\beta_0 + \gamma) + \beta_1 x_{i1} + \varepsilon_{i1} \\ &= \beta_{01} + \beta_1 x_{i1} + \varepsilon_{i1}, \end{aligned}$$

这和模型 2 中对于少数人种取 $\beta_{01} = \beta_0 + \gamma$ 是相同的。类似地, 对于白种人模型 3 变成

$$y_{i2} = \beta_0 + \beta_1 x_{i2} + \varepsilon_{i2}.$$

因此模型 2(或者等价地, 模型 3) 表示截距为 $\beta_0 + \gamma$ 和 β_0 的两条平行线^①(相同斜率)。因此, 我们的原假设隐含着对模型 3 中的 γ 的一个限制, 即 $H_0: \gamma = 0$ 。

^① 一般, 若模型含 X_1, \dots, X_p 及一个示性变量 Z , 那么模型 3 表示两个截距不同的平行的平面或超平面。

我们用 F -检验去检验这个假设

$$F = \frac{[SSE(RM) - SSE(FM)]/1}{SSE(FM)/17},$$

其中自由度是 1 和 17。等价地, 我们可以用 t -检验去检验模型 3 中的假设 $\gamma = 0$, 即

$$t = \frac{\hat{\gamma}}{s.e.(\hat{\gamma})},$$

其自由度是 17。同样, 从这些检验得出任何结论之前应该对模型 3 假定的有效性进行评估。对于这个例子, 我们将以上检验的计算和在此基础上得到的结论, 留给读者作为练习。

5.4.3 具有相同截距和不同斜率的模型

现在我们来处理第三种情形, 两个样本组具有相同的截距 β_0 , 我们希望检验假设: 两个组具有相同斜率, 即 $H_0: \beta_{11} = \beta_{12}$ 。在这个例子中, 我们比较

$$\begin{aligned} \text{模型 1(合并的): } y_{ij} &= \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, j = 1, 2; i = 1, 2, \dots, n_j, \\ \text{模型 2(少数人种): } y_{i1} &= \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1}, \\ \text{模型 2(白种人): } y_{i2} &= \beta_0 + \beta_{12} x_{i2} + \varepsilon_{i2}. \end{aligned} \quad (5.7)$$

这两个模型具有相同的截距 β_0 但是不同的斜率 β_{11} 和 β_{12} 。利用前面定义的示性变量 Z , 我们可以将模型 2 写成

$$\text{模型 3: } y_{ij} = \beta_0 + \beta_1 x_{ij} + \delta(z_{ij} \cdot x_{ij}) + \varepsilon_{ij}. \quad (5.8)$$

模型 3 中有交互项 $(z_{ij} \cdot x_{ij})$ 但是没有变量 Z 的单独项。模型 2 和模型 3 的等价性可以通过观察看出, 对少数人种, 当 $x_{ij} = x_{i1}$ 和 $z_{ij} = 1$, 模型 3 成为

$$\begin{aligned} y_{i1} &= \beta_0 + \beta_1 x_{i1} + \delta x_{i1} + \varepsilon_{i1} \\ &= \beta_0 + (\beta_1 + \delta) x_{i1} + \varepsilon_{i1} \\ &= \beta_0 + \beta_{11} x_{i1} + \varepsilon_{i1}. \end{aligned}$$

这和模型 2 中对于少数人种取 $\beta_{11} = \beta_1 + \delta$ 是一样的。类似地, 模型 3 对于白种人样本组成为

$$y_{i2} = \beta_0 + \beta_{12} x_{i2} + \varepsilon_{i2}.$$

因此, 我们的原假设隐含着对于模型 3 中的 δ 作限制, 即 $H_0: \delta = 0$ 。我们用 F -检验去检验这个假设

$$F = \frac{[SSE(RM) - SSE(FM)]/1}{SSE(FM)/17},$$

其中自由度是 1 和 17。等价地, 我们可以用 t -检验去检验模型 3 中 $\delta = 0$, 即

$$t = \frac{\hat{\delta}}{s.e.(\hat{\delta})},$$

其自由度是 17。模型 3 假定有效性的评估, 以上检验量的计算, 以及在此基础上的结论都作为习题留给读者。

5.5 示性变量的其他应用

如 5.4 节中所描述的示性变量的那些应用, 可以被推广应用于各种各样的问题 (例如, 可参阅 Fox(1984) 和 Kmenta(1986))。例如, 我们想比较 $k \geq 2$ 个总体或组的均值。这里通常使用的方法是众所周知的方差分析(ANOVA)。从第 $j(j = 1, 2, \dots, k)$ 个总体中抽取一个容量为 n_j 的随机样本。我们对响应变量就有了总共 $n = n_1 + \dots + n_k$ 个观测值。令 y_{ij} 为第 j 个样本的第 i 个响应。那么, y_{ij} 可以模型化为

$$y_{ij} = \mu_0 + \mu_1 x_{i1} + \dots + \mu_p x_{ip} + \varepsilon_{ij}. \quad (5.9)$$

这个模型中有 $p = k - 1$ 个示性预测变量 x_{i1}, \dots, x_{ip} 。如果其对应的响应来自于第 j 个总体, 那么 x_{ij} 取 1, 否则取 0。剩下的那个总体是熟知的对照组。对于对照组所有示性变量取值都为 0。因此, 对于对照组, (5.9) 变成

$$y_{ij} = \mu_0 + \varepsilon_{ij}. \quad (5.10)$$

在 (5.9) 和 (5.10) 中, 假定随机误差 ε_{ij} 为具有均值 0, 方差为常数 σ^2 的独立正态变量。常数项 μ_0 表示对照组的均值, 回归系数 μ_j 可以被解释为对照组和第 j 组之间的均值之差。如果 $\mu_j = 0$, 那么, 对照组和第 j 组的均值是相等的。所有组具有相同均值的原假设 $H_0: \mu_1 = \dots = \mu_p = 0$ 可以用 (5.10) 中的模型表示。对立假设: 至少有一个 μ_j 与 0 有显著差异可以用 (5.9) 中的模型表示。将 (5.9) 和 (5.10) 中的模型分别看成完全的和简化的模型。因此可以通过 (3.33) 中的 F -检验来检验 H_0 。所以, 采用示性变量使我们可以将 ANOVA 方法表达为回归分析的一种特殊的情形。

上面讨论的例子是基于截面数据的。示性变量也可用于时间序列数据。另外, 也有一些生长过程的模型将示性变量用作因变量。这些模型, 被称为 logistic 回归模型, 将在第 12 章里讨论。

在 5.6 节、5.7 节中, 我们将讨论示性变量在时间序列数据中的应用。特别地, 将讨论季节性和参数对时间的稳定性这两个概念。我们将阐述问题, 并提供数据, 而分析留给读者。

5.6 季节性

这里, 用作例子的数据集称为滑雪撬销售数据, 见表 5.11, 也可从本书的网站

表 5.11 1964-1973 年可支配收入与滑雪橇销售量

行	日期	销售量	PDI	行	日期	销售量	PDI
1	Q1/64	37.0	109	21	Q1/69	44.9	153
2	Q2/64	33.5	115	22	Q2/69	41.6	156
3	Q3/64	30.8	113	23	Q3/69	44.0	160
4	Q4/64	37.9	116	24	Q4/69	48.1	163
5	Q1/65	37.4	118	25	Q1/70	49.7	166
6	Q2/65	31.6	120	26	Q2/70	43.9	171
7	Q3/65	34.0	122	27	Q3/70	41.6	174
8	Q4/65	38.1	124	28	Q4/70	51.0	175
9	Q1/66	40.0	126	29	Q1/71	52.0	180
10	Q2/66	35.0	128	30	Q2/71	46.2	184
11	Q3/66	34.9	130	31	Q3/71	47.1	187
12	Q4/66	40.2	132	32	Q4/71	52.7	189
13	Q1/67	41.9	133	33	Q1/72	52.2	191
14	Q2/67	34.7	135	34	Q2/72	47.0	193
15	Q3/67	38.8	138	35	Q3/72	47.8	194
16	Q4/67	43.7	140	36	Q4/72	52.8	196
17	Q1/68	44.2	143	37	Q1/73	54.1	199
18	Q2/68	40.4	147	38	Q2/73	49.5	201
19	Q3/68	38.4	148	39	Q3/73	49.5	202
20	Q4/68	45.4	151	40	Q4/73	54.3	204

上得到。数据包含两个变量：销售量 S ，即生产滑雪橇或相关器械的一个公司在 1964~1973 年的销售量（以百万计），和个人可支配收入 PDI （购买潜力的综合量度）。这些变量都是按季度统计的。我们在第 8 章中用这些数据来说明相关误差的问题。这个模型是 S 关于 PDI 的方程，

$$S_t = \beta_0 + \beta_1 \cdot PDI_t + \varepsilon_t,$$

其中， S_t 是第 t 个时期的以百万计的销售量， PDI_t 是相应的个人可支配收入。我们假定对销售量存在以季度为基础的季节性影响。为了衡量这一影响，我们定义刻画季节性的示性变量。由于有 4 个季度，所以定义 3 个示性变量， Z_1, Z_2 和 Z_3 ,

$$z_{t1} = \begin{cases} 1 & \text{如果第 } t \text{ 个时期是第一个季度,} \\ 0 & \text{否则,} \end{cases}$$

$$z_{t2} = \begin{cases} 1 & \text{如果第 } t \text{ 个时期是第二个季度,} \\ 0 & \text{否则,} \end{cases}$$

$$z_{t3} = \begin{cases} 1 & \text{如果第 } t \text{ 个时期是第三个季度,} \\ 0 & \text{否则.} \end{cases}$$

关于数据的分析和解释留给读者。笔者分析过这批数据后发现实际上只存在两季(见第8章对于这批数据的讨论,分析时只用了一个示性变量,表示两个季节)。关于用示性变量来分析季节性的进一步讨论可见 Kementa(1986)。

5.7 回归参数对时间的稳定性

示性变量也可用于分析回归系数对于时间的稳定性或检验结构的变化。当数据是跨越时间的截面观测时,我们考虑回归系统问题的一种推广。我们的目标是分析回归关系随时间变化的稳定性。这里描述的方法适用于时间之间和空间之间的比较。我们采用表 5.12-5.14 给出的教育经费数据来介绍此方法。对 50 个州测量的变量为:

- Y : 在公共教育方面的人均开支
- X_1 : 人均个人收入
- X_2 : 低于 18 岁的人口比例 (%)
- X_3 : 城市居民的人口比例 (%)

表示地理区域的变量是分类变量 (1= 东北部, 2= 中北部, 3= 南部, 4= 西部)。第 7 章将使用这个数据集去说明处理多元回归中异方差的方法,并分析区域变量对于回归关系的影响。这里,我们关注关于教育经费的回归关系随时间变化的稳定性。

数据为前面描述的 4 个变量在每个州于 1960, 1970 和 1975 年的取值。假定这三年中各年的回归关系形式相同^①,那么稳定性的分析就可以通过评估回归系数的估计关于时间的变化来实现。对于合并数据集的 150 个观测 (50 个州 3 年的数据),我们定义两个示性变量, T_1 和 T_2 , 其中

$$T_{i1} = \begin{cases} 1 & \text{如果第 } i \text{ 个观测来自 1960 年,} \\ 0 & \text{否则,} \end{cases}$$

$$T_{i2} = \begin{cases} 1 & \text{如果第 } i \text{ 个观测值来自 1970 年,} \\ 0 & \text{否则.} \end{cases}$$

用 Y 表示人均教育开支,则模型的形式为

$$\begin{aligned} Y = & \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \gamma_1 T_1 + \gamma_2 T_2 + \delta_1 T_1 \cdot X_1 \\ & + \delta_2 T_1 \cdot X_2 + \delta_3 T_1 \cdot X_3 + \alpha_1 T_2 \cdot X_1 + \alpha_2 T_2 \cdot X_2 \\ & + \alpha_3 T_2 \cdot X_3 + \varepsilon. \end{aligned}$$

^① 形式相同指每个方程含相同的变量及相同的变量变换形式。该假定应有经验上的依据。

表 5.12 教育经费数据 (1960 年)

行	州	Y	X ₁	X ₂	X ₃	区域
1	ME	61	1704	388	399	1
2	NH	68	1885	372	598	1
3	VT	72	1745	397	370	1
4	MA	72	2394	358	868	1
5	RI	62	1966	357	899	1
6	CT	91	2817	362	690	1
7	NY	104	2685	341	728	1
8	NJ	99	2521	353	826	1
9	PA	70	2127	352	656	2
10	OH	82	2184	387	674	2
11	IN	84	1990	392	568	2
12	IL	84	2435	366	759	2
13	MI	104	2099	403	650	2
14	WI	84	1936	393	621	2
15	MN	103	1916	402	610	2
16	IA	86	1863	385	522	2
17	MO	69	2037	364	613	2
18	ND	94	1697	429	351	2
19	SD	79	1644	411	390	2
20	NB	80	1894	379	520	2
21	KS	98	2001	380	564	2
22	DE	124	2760	388	326	3
23	MD	92	2221	393	562	3
24	VA	67	1674	402	487	3
25	WV	66	1509	405	358	3
26	NC	65	1384	423	362	3
27	SC	57	1218	453	343	3
28	GA	60	1487	420	498	3
29	FL	74	1876	334	628	3
30	KY	49	1397	594	377	3
31	TN	60	1439	346	457	3
32	AL	59	1359	637	517	3
33	MS	68	1053	448	362	3
34	AR	56	1225	403	416	3
35	LA	72	1576	433	562	3
36	OK	80	1740	378	610	3
37	TX	79	1814	409	727	3
38	MT	95	1920	412	463	4
39	ID	79	1701	418	414	4
40	WY	142	2088	415	568	4
41	CO	108	2047	399	621	4
42	NM	94	1838	458	618	4
43	AZ	107	1932	425	699	4
44	UT	109	1753	494	665	4
45	NV	114	2569	372	663	4
46	WA	112	2160	386	584	4
47	OR	105	2006	382	534	4
48	CA	129	2557	373	717	4
49	AK	107	1900	434	379	4
50	HI	77	1852	431	693	4

表 5.13 教育经费数据 (1970 年)

行	州	Y	X ₁	X ₂	X ₃	区域
1	ME	189	2828	351	508	1
2	NH	169	3259	346	564	1
3	VT	230	3072	348	322	1
4	MA	168	3835	335	846	1
5	RI	180	3549	327	871	1
6	CT	193	4256	341	774	1
7	NY	261	4151	326	856	1
8	NJ	214	3954	333	889	1
9	PA	201	3419	326	715	2
10	OH	172	3509	354	753	2
11	IN	194	3412	359	649	2
12	IL	189	3981	349	830	2
13	MI	233	3675	369	738	2
14	WI	209	3663	361	659	2
15	MN	262	3341	365	664	2
16	IA	234	3265	344	572	2
17	MO	177	3257	336	701	2
18	ND	177	2730	369	443	2
19	SD	187	2876	369	446	2
20	NB	148	3239	350	615	2
21	KS	196	3303	340	661	2
22	DE	248	3795	376	722	3
23	MD	247	3742	364	766	3
24	VA	180	3068	353	631	3
25	WV	149	2470	329	390	3
26	NC	155	2664	354	450	3
27	SC	149	2380	377	476	3
28	GA	156	2781	371	603	3
29	FL	191	3191	336	805	3
30	KY	140	2645	349	523	3
31	TN	137	2579	343	588	3
32	AL	112	2337	362	584	3
33	MS	130	2081	385	445	3
34	AR	134	2322	352	500	3
35	LA	162	2634	390	661	3
36	OK	135	2880	330	680	3
37	TX	155	3029	369	797	3
38	MT	238	2942	369	534	4
39	ID	170	2668	368	541	4
40	WY	238	3190	366	605	4
41	CO	192	3340	358	785	4
42	NM	227	2651	421	698	4
43	AZ	207	3027	387	796	4
44	UT	201	2790	412	804	4
45	NV	225	3957	385	809	4
46	WA	215	3688	342	726	4
47	OR	233	3317	333	671	4
48	CA	273	3968	348	909	4
49	AK	272	4146	440	484	4
50	HI	212	3513	383	831	4

表 5.14 教育经费数据 (1975 年)

行	州	Y	X ₁	X ₂	X ₃	区域
1	ME	235	3944	325	508	1
2	NH	231	4578	323	564	1
3	VT	270	4011	328	322	1
4	MA	261	5233	305	846	1
5	RI	300	4780	303	871	1
6	CT	317	5889	307	774	1
7	NY	387	5663	301	856	1
8	NJ	285	5759	310	889	1
9	PA	300	4894	300	715	2
10	OH	221	5012	324	753	2
11	IN	264	4908	329	649	2
12	IL	308	5753	320	830	2
13	MI	379	5439	337	738	2
14	WI	342	4634	328	659	2
15	MN	378	4921	330	664	2
16	IA	232	4869	318	572	2
17	MO	231	4672	309	701	2
18	ND	246	4782	333	443	2
19	SD	230	4296	330	446	2
20	NB	268	4827	318	615	2
21	KS	337	5057	304	661	2
22	DE	344	5540	328	722	3
23	MD	330	5331	323	766	3
24	VA	261	4715	317	631	3
25	WV	214	3828	310	390	3
26	NC	245	4120	321	450	3
27	SC	233	3817	342	476	3
28	GA	250	4243	339	603	3
29	FL	243	4647	287	805	3
30	KY	216	3967	325	523	3
31	TN	212	3946	315	588	3
32	AL	208	3724	332	584	3
33	MS	215	3448	358	445	3
34	AR	221	3680	320	500	3
35	LA	244	3825	355	661	3
36	OK	234	4189	306	680	3
37	TX	269	4336	335	797	3
38	MT	302	4418	335	534	4
39	ID	268	4323	344	541	4
40	WY	323	4813	331	605	4
41	CO	304	5046	324	785	4
42	NM	317	3764	366	698	4
43	AZ	332	4504	340	796	4
44	UT	315	4005	378	804	4
45	NV	291	5560	330	809	4
46	WA	312	4989	313	726	4
47	OR	316	4697	305	671	4
48	CA	332	5438	307	909	4
49	AK	546	5613	386	484	4
50	HI	311	5309	333	831	4

由 T_1 和 T_2 的定义, 上面的模型等价于

$$\begin{aligned} 1960 \text{ 年: } Y &= (\beta_0 + \gamma_1) + (\beta_1 + \delta_1)X_1 + (\beta_2 + \delta_2)X_2 \\ &\quad + (\beta_3 + \delta_3)X_3 + \varepsilon, \\ 1970 \text{ 年: } Y &= (\beta_0 + \gamma_2) + (\beta_1 + \alpha_1)X_1 + (\beta_2 + \alpha_2)X_2, \\ &\quad + (\beta_3 + \alpha_3)X_3 + \varepsilon, \\ 1975 \text{ 年: } Y &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \varepsilon. \end{aligned}$$

正如前面的注释, 这种分析方法隐含着各年的回归函数形式相同的假定。我们感兴趣的一个假设是

$$H_0: \gamma_1 = \gamma_2 = \delta_1 = \delta_2 = \delta_3 = \alpha_1 = \alpha_2 = \alpha_3 = 0,$$

它意味着在 1960–1975 年间回归系统没有变化。

这个例子的数据, 我们称为教育经费数据, 列在表 5.12, 表 5.13 和表 5.14 中, 也可从本书的网站上得到。请读者将以上的分析作为练习。

习 题

5.1 采用 (5.6) 中定义的模型:

- 检查通常的最小二乘假定是否成立。
- 采用 F -检验去检验假设 $H_0: \gamma = 0$ 。
- 采用 t -检验去检验假设 $H_0: \gamma = 0$ 。
- 验证上面两个检验的等价性。

5.2 采用 (5.8) 中定义的模型:

- 检查通常的最小二乘假定是否成立。
- 采用 F -检验去检验假设 $H_0: \delta = 0$ 。
- 采用 t -检验去检验假设 $H_0: \delta = 0$ 。
- 验证上面两个检验的等价性。

5.3 采用 5.6 节中的想法对表 5.11 中的滑雪撬销售数据进行全面分析。

5.4 采用 5.7 节中提出的想法对表 5.12, 表 5.13 和表 5.14 中的教育经费数据进行全面分析。

5.5 试验三种类型的肥料以观察哪一种肥料会有助于收获较多的谷物。试验在 40 块类似的试验田中实施。这 40 块实验田被随机地分成 4 个组, 每组 10 块。肥料 1 被用于第 1 组的 10 块地里。类似地, 肥料 2 和 3 分别用于第 2 组和第 3 组中。第 4 组的谷物没有施任何肥料; 它被作为对照组。表 5.15 给出了 40 块地的每块地谷物的产量 y_{ij} 。

- 构造 3 个示性变量 F_1, F_2, F_3 , 分别对应于 3 种肥料。
- 拟合模型 $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \varepsilon_{ij}$ 。

表 5.15 按照化学肥料种类分组的谷物产量

肥料 1	肥料 2	肥料 3	对照组
31	27	36	33
34	27	37	27
34	25	37	35
34	34	34	25
43	21	37	29
35	36	28	20
38	34	33	25
36	30	29	40
36	32	36	35
45	33	42	29

(c) 检验假设: 平均来说, 3 种类型的化肥均对谷物产量没有影响。明确要检验的假设, 采用的检验, 以及在 5% 显著性水平下的结论。

(d) 检验假设: 平均来说, 3 种类型的化肥均对于谷物产量有相同的影响, 但是与对照组显著不同。明确需检验的假设, 采用的检验, 以及在 5% 显著性水平下的结论。

(e) 三种肥料中哪一种对于谷物产量有最大的效应?

5.6 在一门统计课程中, 用于学生的个人信息被收集来用于课堂分析。学生们的年龄 (岁), 身高 (英寸), 和体重 (英镑) 的数据在表 5.16 中给出, 且可从本书的网站上得到。每个学生的性别也被标出了, 且女生标为 1 男生为 0。我们想研究学生们的体重和身高的关系。体重作为响应变量, 身高作为预测变量。

(a) 你认为变量的作用应该颠倒过来吗?

(b) 一个方程能够充分描绘男女两组学生身高和体重的关系吗? 考察对合并数据拟合模型得到的标准化残差图, 将男女生对应的残差区分开来。

(c) 请采用交互作用变量以及本章描述的方法, 找到描绘学生体重和身高关系的最优模型。

(d) 你认为我们应该将年龄包含进来作为预测体重的变量吗? 对你的答案给出直观的解释。

5.7 总统选举数据 (1916–1996 年)[†]: 表 5.17 中的数据是由耶鲁大学 Ray Fair 教授提供的。他发现美国总统选举中总统候选人的支持率可以通过三个宏观经济变量、执政政党, 以及一个反映选举是否在一次战争结束后举行的示性变量来精确预测。被考虑的变量在表 5.18 中给出。所有的增长率都是以百分比计的年增长率。考虑对数据用下面的初始模型拟合

$$V = \beta_0 + \beta_1 \cdot I + \beta_2 \cdot D + \beta_3 \cdot W + \beta_4 \cdot (G \cdot I) + \beta_5 \cdot P + \beta_6 \cdot N + \varepsilon. \quad (5.11)$$

(a) 是否有必要在上面的模型中保留变量 I ?

[†] 译注: 本题的变量解释、数据及相关的研究, 可从网站 <http://fairmodel.ecom.yale.edu/vote2004/indexz.htm> 上找到。

- (b) 是否有必要在上面的模型中保留交互作用变量 ($G \cdot I$)?
- (c) 请尝试不同的模型从而得到能更好地预测未来总统选举的模型。如果需要, 可以包含交互项。

表 5.16 一班学生的年龄(岁)、身高(英寸)、体重(磅)和性别(1=女, 0=男)数据

年龄	身高	体重	性别	年龄	身高	体重	性别
19	61	180	0	19	65	135	1
19	70	160	0	19	70	120	0
19	70	135	0	21	69	142	0
19	71	195	0	20	63	108	1
19	64	130	1	19	63	118	1
19	64	120	1	20	72	135	0
21	69	135	1	19	73	169	0
19	67	125	0	19	69	145	0
19	62	120	1	27	69	130	1
20	66	145	0	18	64	135	0
19	65	155	0	20	61	115	1
19	69	135	1	19	68	140	0
19	66	140	0	21	70	152	0
19	63	120	1	19	64	118	1
19	69	140	0	19	62	112	1
18	66	113	1	19	64	100	1
18	68	180	0	20	67	135	1
19	72	175	0	20	63	110	1
19	70	169	0	20	68	135	0
19	74	210	0	18	63	115	1
20	66	104	1	19	68	145	0
20	64	105	1	19	65	115	1
20	65	125	1	19	63	128	1
20	71	120	1	20	68	140	1
19	69	119	1	19	69	130	0
20	64	140	1	19	69	165	0
20	67	185	1	19	69	130	0
19	60	110	1	20	70	180	0
20	66	120	1	28	65	110	1
19	71	175	0	19	55	155	1

表 5.17 总统选举数据 (1961–1996 年)

年份	V	I	D	W	G	P	N
1916	0.5168	1	1	0	2.229	4.252	3
1920	0.3612	1	0	1	-11.463	16.535	5
1924	0.4176	-1	-1	0	-3.872	5.161	10
1928	0.4118	-1	0	0	4.623	0.183	7
1932	0.5916	-1	-1	0	-14.901	7.069	4
1936	0.6246	1	1	0	11.921	2.362	9
1940	0.5500	1	1	0	3.708	0.028	8
1944	0.5377	1	1	1	4.119	5.678	14
1948	0.5237	1	1	1	1.849	8.722	5
1952	0.4460	1	0	0	0.627	2.288	6
1956	0.4224	-1	-1	0	-1.527	1.936	5
1960	0.5009	-1	0	0	0.114	1.932	5
1964	0.6134	1	1	0	5.054	1.247	10
1968	0.4960	1	0	0	4.836	3.215	7
1972	0.3821	-1	-1	0	6.278	4.766	4
1976	0.5105	-1	0	0	3.663	7.657	4
1980	0.4470	1	1	0	-3.789	8.093	5
1984	0.4083	-1	-1	0	5.387	5.403	7
1988	0.4610	-1	0	0	2.068	3.272	6
1992	0.5345	-1	-1	0	2.293	3.692	1
1996	0.5474	1	1	0	2.918	2.268	3

表 5.18 表 5.17 中 1916–1996 年总统选举数据的变量

变量	定义
年份	选举年份
V	民主党在两党总统选举中的份额
I	示性变量 (如选举时是民主党执政就取 1, 是共和党执政则取 -1)
D	示性变量 如是民主党现任总统参与竞选则取 1, 如是共和党现任总统参与竞选则取 -1, 否则取 0)
W	示性变量 (对 1920, 1944 和 1948 年取 1, 否则取 0)
G	选举年前三季度人均实际 GDP 增长率
P	现任总统任期的前 15 个季度, GDP 价格指数增长率的绝对值
N	现任总统任期的前 15 个季度中人均实际 GDP 增长率大于 3.2% 的季度数

6

变量的变换

6.1 引言

有时候数据的形式并不适合于直接作分析，在分析之前常常不得不先对变量作变换。作变换是为了达到线性性，正态性，或者等方差等目的。在实践中，通常变换后的变量比原变量更适合用线性回归模型来拟合。在本章中，我们将讨论在什么情形下需要作数据变换，可选择的变换，以及对变换后的数据进行分析。

下面我们主要以简单回归为例来说明数据变换。在多元回归中有多个预测变量，可能其中的一些变量需要变换，而另一些则不需要。尽管同样的技术可以应用于多元回归，但是多元回归的数据变换需要更加细心。

数据需要作变换是由于原始变量、或者用原始变量表达的模型，违背了回归的一个或多个标准假定。最常见的是违背了模型的线性和误差方差为常数的假定。正如在第2章和第3章提到的那样，只要参数以线性形式出现在模型中，即使预测变量的形式是非线性的，这个回归模型仍为线性的。例如，下面4个模型都是线性的：

$$\begin{aligned}Y &= \beta_0 + \beta_1 X + \varepsilon, \\Y &= \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon, \\Y &= \beta_0 + \beta_1 \log X + \varepsilon, \\Y &= \beta_0 + \beta_1 \sqrt{X} + \varepsilon.\end{aligned}$$

因为模型关于参数 $\beta_0, \beta_1, \beta_2$ 是线性的。另一方面，

$$Y = \beta_0 + e^{\beta_1 X} + \varepsilon$$

是非线性的，因为参数 β_1 并不是以线性的形式出现在模型中。为了满足标准的回归模型的假定，我们常常对变换后的变量进行分析，而不是对原始变量。须作变换的原因有多种。

1. 理论上可以确定两个变量之间的关系不是线性的。但是, 经过适当变换后, 新的变量之间的关系呈线性。考虑一个学习理论 (实验心理学) 中普遍使用的学习模型, 它将第 i 次完成一项任务所花的时间 (T_i) 刻画为

$$T_i = \alpha\beta^i, \quad \alpha > 0, 0 < \beta < 1. \quad (6.1)$$

在 (6.1) 中 T_i 和 i 之间的关系是非线性的, 我们不能直接使用线性回归的方法。但是, 如果我们在等式两边同时取对数, 就得到

$$\log T_i = \log \alpha + i \log \beta, \quad (6.2)$$

这表明 $\log T_i$ 和 i 是线性关系。该变换使我们能使用标准回归方法。尽管原始变量之间的关系并不是线性的, 但是变换后的变量之间的关系是线性的。所以变换可以实现拟合模型的线性性。

2. 被分析的响应变量 Y , 它的概率分布的方差可能与其均值有关。如果其均值与预测变量 X 的取值有关, 则 Y 的方差将随 X 而变化, 而不会是一个常数。在这种情况下, Y 的分布通常是非正态的。非正态性使得标准显著性检验不合理 (尽管在大样本情况下影响不大), 因为这些检验是建立在正态性假定基础上的。误差方差不相等虽然仍会产生无偏的估计, 但不再是最小方差意义下的最优估计。在这些情形下, 我们经常变换数据以保证正态性和误差同方差性。在实践中, 常选择可以保证同方差性的变换 (方差稳定性变换)。凑巧的是方差稳定性变换也往往是好的正态性变换。

3. 既没有先验的理论也没有概率方面的原因要求作数据变换。其变换的依据来自对于用原始变量拟合的线性回归模型的残差分析。

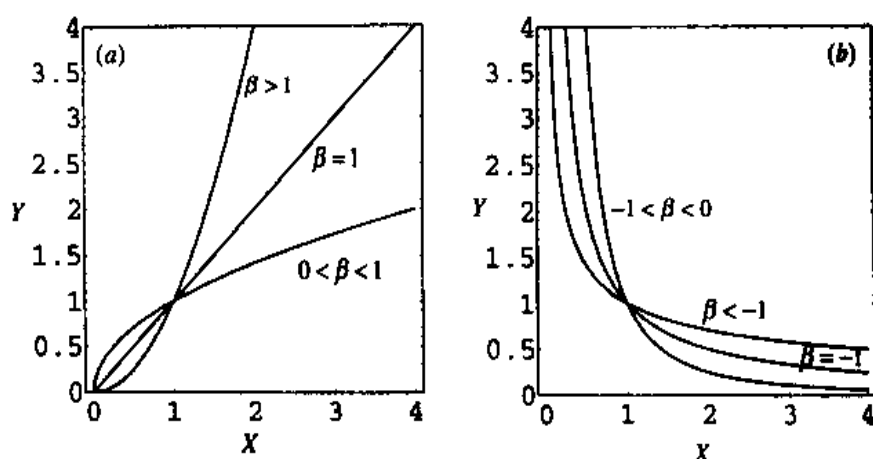
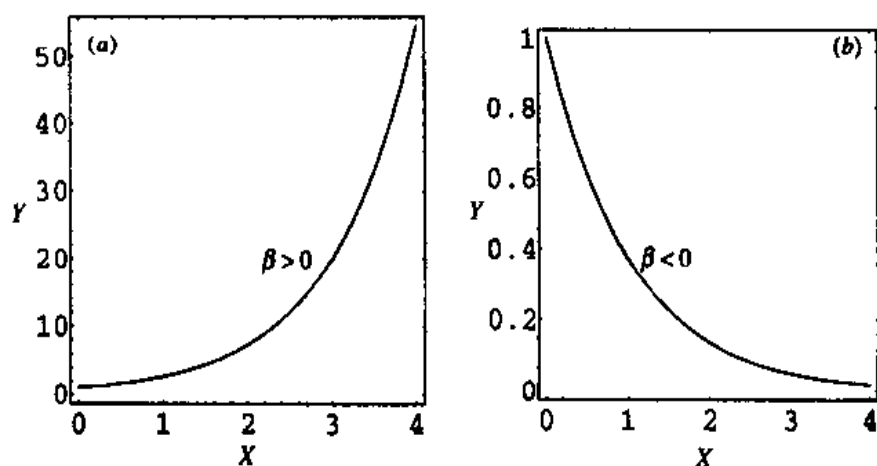
下面各节将对以上需要作变量变换的各种情形分别举例说明。

6.2 线性化变换

在回归分析当中一个标准的假定是描绘数据的模型是线性的。但是从理论考虑, 或者从 Y 相对于预测变量 X_j 的散点图上考察, Y 和 X_j 的关系有时呈现非线性。然而有许多简单非线性回归, 经过适当的变换可以变成线性。我们在表 6.1 中列出了一些可线性化的曲线。相应的图见图 6.1 到图 6.4。

表 6.1 可线性化的简单回归函数及相应的变换

函数	变换	线性形式	图
$Y = \alpha X^\beta$	$Y' = \log Y, X' = \log X$	$Y' = \log \alpha + \beta X'$	图 6.1
$Y = \alpha e^{\beta X}$	$Y' = \ln Y$	$Y' = \ln \alpha + \beta X$	图 6.2
$Y = \alpha + \beta \log X$	$X' = \log X$	$Y' = \alpha + \beta X'$	图 6.3
$Y = \frac{X}{\alpha X - \beta}$	$Y' = \frac{1}{Y}, X' = \frac{1}{X}$	$Y' = \alpha - \beta X'$	图 6.4(a)
$Y = \frac{e^{\alpha + \beta X}}{1 + e^{\alpha + \beta X}}$	$Y' = \ln \frac{Y}{1 - Y}$	$Y' = \alpha + \beta X$	图 6.4(b)

图 6.1 可线性化函数 $Y = \alpha X^\beta$ 图 6.2 可线性化函数 $Y = \alpha e^{\beta X}$

在 (Y, X) 的散点图中观察到弯曲现象时, 可以选择图 6.1~ 图 6.4 中给出的可线性化的曲线来表示这些数据。但是, 也有许多非线性模型不能线性化。例如, 一条经修正的指数曲线 $Y = \alpha + \beta\theta^X$, 以及

$$Y = \alpha_1 e^{\theta_1 X} + \alpha_2 e^{\theta_2 X},$$

它是两个指数函数的和。严格的非线性模型 (即不能通过变量变换线性化的) 需要用完全不同的方法去拟合。我们在本书中不讨论它们, 有兴趣的读者可参见 Bates and Watts (1998), Seber and Wild (1989) 以及 Ratkosky (1990)。

在下面的例子中, 理论上导出的模型不是线性的。但是, 这个模型可以被线性化。我们对其进行适当的分析。

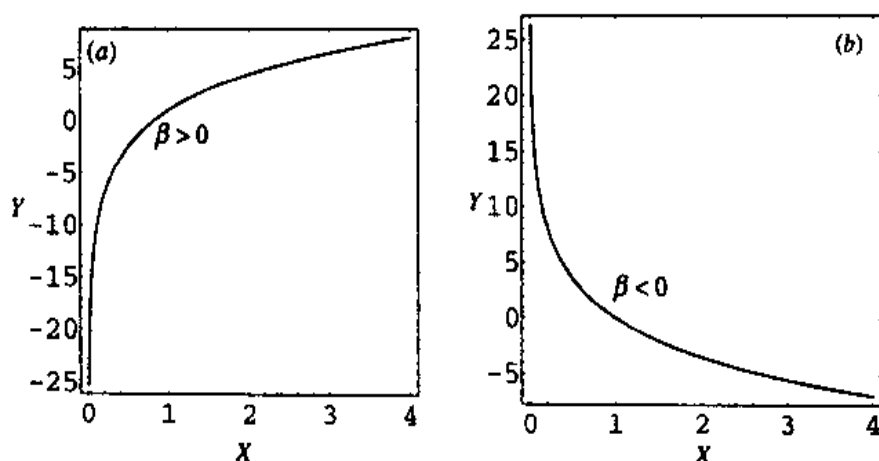
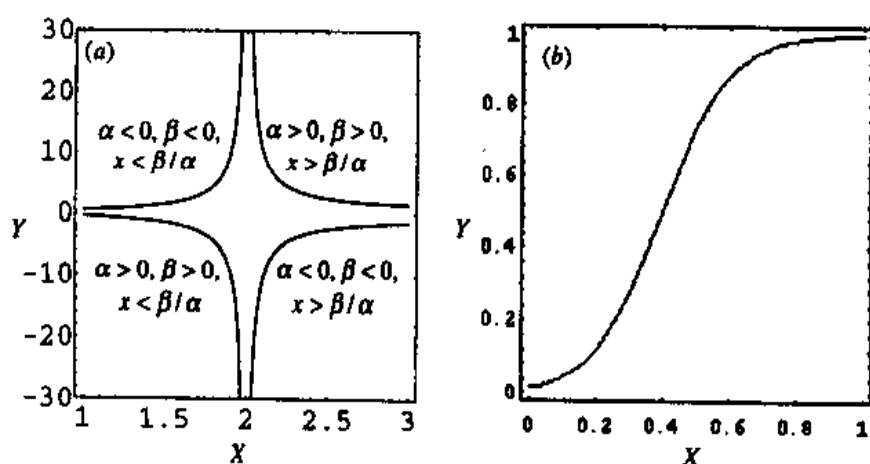
图 6.3 可线性化函数 $Y = \alpha + \beta \log X$ 

图 6.4 可线性化函数

(a) $Y = X/(\alpha X - \beta)$, 和 (b) $Y = (e^{\alpha + \beta X})/(1 + e^{\alpha + \beta X})$

6.3 X 射线杀菌的数据

在一个实验中, 用 200 千伏 X 射线照射海洋细菌数次, 次数 $t = 1 \sim 15$, 每次照射 6 分钟用平板计数法估计尚存活的细菌数, 数据列在表 6.2 中 (单位 100 个)。此数据也可以在本书的网站上找到^①。响应变量 n_t 表示经过 t 次照射后仍然存活的细菌个数。试验是为了检验在恒定的辐射区域内 X 射线杀菌作用的单个击中 (single-hit) 假设。根据这个理论, 每一个细菌都存在一个致命中心, 一条射线必定击中该中心才会致使细菌失去活性或死亡。用于研究的特殊细菌不会形成团或链, 所以细菌的个数可以用平板计数法来估计。

① <http://www.ilr.cornell.edu/~hadi/RABE>

如果这一理论合适, 则 n_t 和 t 之间应当有如下关系

$$n_t = n_0 e^{\beta_1 t}, \quad t \geq 0, \quad (6.3)$$

其中 n_0 和 β_1 是参数。这些参数有简单的物理解释: n_0 是实验开始时细菌的个数, β_1 是死亡(或衰变)的速率。我们对 (6.3) 式的两边取对数, 得到

$$\ln n_t = \ln n_0 + \beta_1 t = \beta_0 + \beta_1 t, \quad (6.4)$$

其中 $\beta_0 = \ln n_0$, $\ln n_t$ 是 t 的线性函数。如果我们引入 ε_t 作为随机误差, 则模型变成

$$\ln n_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad (6.5)$$

现在就可以应用标准的最小二乘法了。

为了使 ε_t 在变换后的模型 (6.5) 中是可加的, 原始模型 (6.3) 中的误差项必须以乘积形式出现。正确的模型表达形式应该为

$$\ln n_t = \ln n_0 e^{\beta_1 t \varepsilon'_t}, \quad (6.6)$$

其中 ε'_t 是倍增的随机误差。比较 (6.5) 和 (6.6), 可以看出 $\varepsilon_t = \ln \varepsilon'_t$ 。标准最小二乘法中 ε_t 应该服从正态分布, 也就意味着 ε'_t 服从对数正态分布^①。在实践中, 对变换后的数据进行模型拟合以后, 通常我们会通过观察其残差来判断模型假定是否成立, 而不会对原始模型中的随机成分 ε'_t 作研究。

表 6.2 存活的细菌数目 (单位 100)

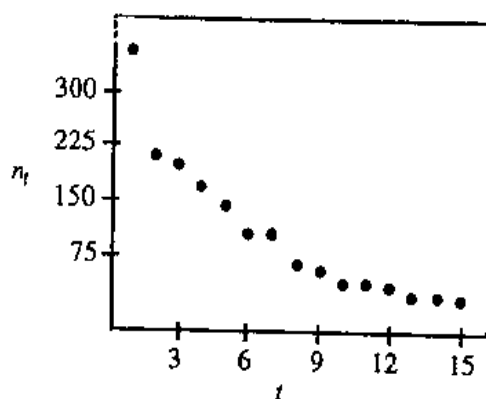
t	n_t	t	n_t	t	n_t
1	355	6	106	11	36
2	211	7	104	12	32
3	197	8	60	13	21
4	166	9	56	14	19
5	142	10	38	15	15

① 如果 $\ln Y$ 具有正态分布, 则称随机变量 Y 具有对数正态分布。

表 6.3 模型 (6.7) 中的回归系数估计

变量	系数	标准误	t -检验	P -值
常数	-259.58	22.73	11.42	< 0.0001
时间 (t)	-19.46	2.50	-7.79	< 0.0001
	$n = 15$	$R^2 = 0.823$	$\hat{\sigma} = 41.83$	$d.f. = 13$

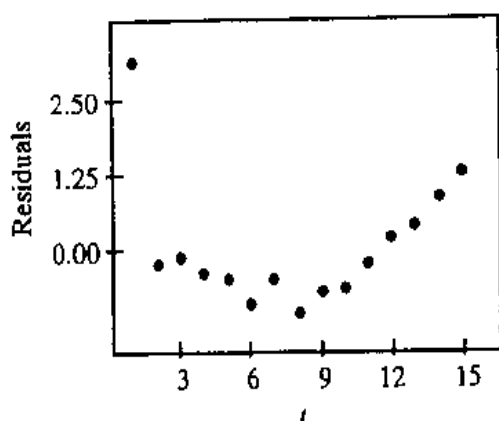
6.3.1 线性模型的不适用

图 6.5 n_t 对 t 的散点图

数据分析的第一步是将原始数据 n_t 关于 t 的散点图画出来。散点图 6.5 揭示出 n_t 和 t 之间的关系是非线性的。但是，我们仍采用简单线性模型对其进行拟合，并研究误设模型带来的影响。模型为

$$n_t = \beta_0 + \beta_1 t + \varepsilon_t, \quad (6.7)$$

其中 β_0 和 β_1 是常数，随机误差 ε_t 的均值为 0、方差相等，并且互不相关。表 6.3 给出了 β_0, β_1 的估计，它们的标准误及相关系数的平方。尽管时间变量的回归系数是显著的，并且有很高的 R^2 ，但是对这批数据用上面的线性模型来拟合是不合适的。 n_t 对 t 的散点图表明当 t 取较大值时偏离线性性（图 6.5）。如果观察标准化残差对时间的散点图（图 6.6），我们会看得更清楚。残差分布有一个特殊的模式：从 $t = 2$ 到 11 都是负的，从 $t = 12$ 到 15 都是正的，而 $t = 1$ 时的残差看似一个异常值。这种系统偏离的模式进一步证实线性模型 (6.7) 不适合拟合这批数据。

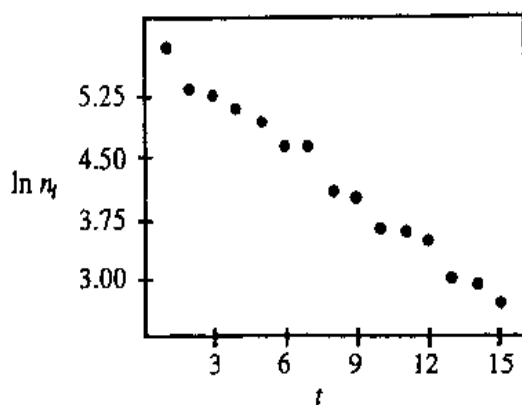
图 6.6 由 (6.7) 得出的对时间 t 的标准化残差散点图

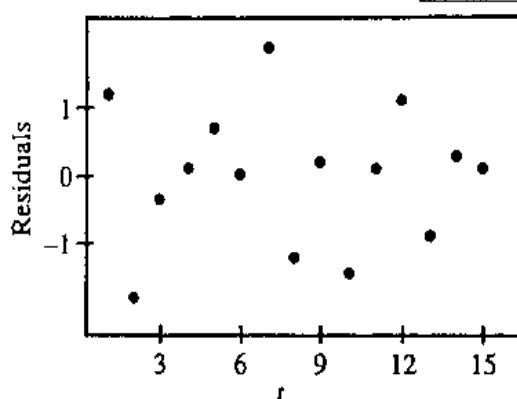
6.3.2 为得到线性性而进行对数变换

表 6.4 当 $\ln n_t$ 对时间 t 作回归时的回归系数估计

变量	系数	标准误	t -检验	p -值
常数	5.973	0.0598	99.9	< 0.0001
时间 (t)	-0.218	0.0066	-33.2	< 0.0001
$n = 15$		$R^2 = 0.988$	$\hat{\sigma} = 0.11$	$d.f. = 13$

n_t 和 t 之间的关系明显是非线性的, 理论上的考虑以及图 6.7 提示我们应该对变换后的变量 $\ln n_t$ 进行分析。从 $\ln n_t$ 对 t 的散点图可以看出它们之间呈线性关系, 表明对数变换是合适的。从表 6.4 列出的 (6.5) 的拟合结果, 可以看出回归系数的高度显著性, 标准误是合理的, 以及数据中近 99% 的变异可以由此模型来解释。并且从标准化残差对 t 的散点图 (图 6.8) 可以看到, 残差分布并没有系统的模式, 该图是令人满意的。所以数据证实了 $\ln n_t$ 与 t 之间的线性关系, 从而也证实了 X 射线杀菌作用的单个击中假设。

图 6.7 $\ln n_t$ 对时间 t 的散点图

图 6.8 变换后对时间 t 的标准化残差散点图

在对变换后的数据进行分析时,必须非常注意模型参数的估计。在我们的例子中, β_1 的点估计是 -0.218 , 其 95% 的置信区间是 $(-0.232, -0.204)$ 。方程中常数项的估计是 $\ln n_0$ 的最优线性无偏估计。如果令 $\hat{\beta}_0$ 表示为 β_0 的估计,那么, $e^{\hat{\beta}_0}$ 就可以作为 n_0 的估计。当 $\hat{\beta}_0 = 5.973$ 时, $e^{\hat{\beta}_0} = 392.68$ 。这个估计不是 n_0 的无偏估计;也就是说在实验开始时细菌的真实个数可能比 392.68 小些。可作适当修正来减少 n_0 估计的偏倚。 $\exp[\hat{\beta}_0 - \frac{1}{2}\text{Var}(\hat{\beta}_0)]$ 是 n_0 的近似无偏估计。在我们现在的例子中, n_0 的修正估计是 381.11。值得注意的是估计 n_0 时的偏倚对于射线理论的假设检验或者衰变速率的估计都没有影响。

一般来讲,如果遇到非线性情况,那么它就会表现在数据的散点图上。如果数据散点图近似于图 6.1 到 6.4 中的某一个,那么可以在对数据作适当变换以后再加以拟合。变换后模型的合适性问题可以用第 4 章中的方法来研究。

6.4 方差稳定性变换

前面我们已经讨论了采用变量变换来达到回归函数的线性性。变量变换也可用于稳定误差的方差,也就是说,使得所有观测的误差方差为常数。误差方差为常数是**最小二乘法**的标准假定之一。它通常被称为**方差齐性(homoscedasticity)**假定。当不是所有观测的误差方差为常数时,误差就被称为**异方差的(heteroscedatic)**。异方差性通常可以通过适当的残差图来识别,例如标准化残差对拟合数据或对每个预测变量的散点图。图 6.9 就是一个例子,其残差分布呈漏斗形状,随着 X 值的增加或者作扇形散开或者收拢。

如果存在异方差性,而没有对其进行修正,那么直接将 OLS 应用于原始数据将会导致系数估计在理论上缺乏精确性。回归参数估计的标准误差通常会被低估,从而对其准确性会产生错觉。

我们可以通过适当的变换消除异方差性。下面描述一种方法可以 (a) 检测异方差性和它对于数据分析的影响, (b) 通过变换消除数据的异方差性。

回归问题中的响应变量 Y 也许会服从一个概率分布: 其方差是其均值的函数。正态分布具有一个其他许多分布不具有的性质: 它的均值和方差相互独立, 也

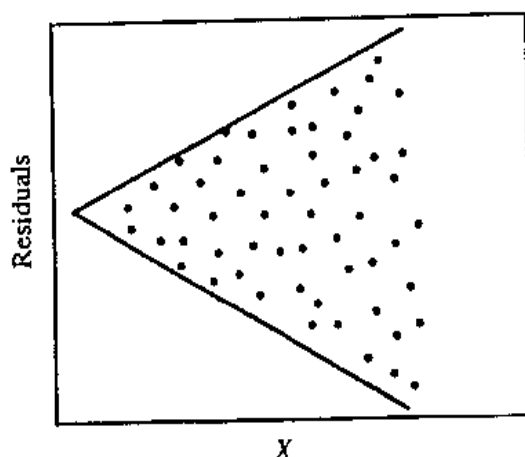


图 6.9 异方差残差的一个例子

就是说它们彼此都不是对方的函数。二项分布和普松分布是不具有此性质的两个普通概率分布的例子。例如, 我们知道, 若一个变量服从参数为 n 和 π 的二项分布, 则其均值为 $n\pi$, 方差为 $n\pi(1-\pi)$ 。同样, 众所周知, 服从普松分布的随机变量的均值和方差是相等的。当知道了-一个随机变量的均值和方差的关系之后, 就可能找到这个变量的某种简单变换, 使得方差近似为常数 (稳定方差)。为了方便检索, 对于通常碰到的、且方差是其均值函数的概率分布, 我们在表 6.5 中列出了相应的方差稳定性变换。表 6.5 列出的变换不仅能稳定方差, 而且能使变换后的变量近似于正态分布。因此, 这些变换具有双重目的: 使变量正态化以及使得其方差独立于其均值。

表 6.5 稳定方差的变换

Y 的概率分布	用其均值 μ 表达的 $Var(Y)$	变换	产生的方差
普松分布 ^a	μ	\sqrt{Y} 或 $(\sqrt{Y} + \sqrt{Y+1})$	0.25
二项分布 ^b	$\mu(1-\mu)/n$	$\sin^{-1} \sqrt{Y}$ (度)	$821/n$
		$\sin^{-1} \sqrt{Y}$ (弧度)	$0.25/n$
负二项分布	$\mu + \lambda^2 \mu^2$	$\lambda^{-1} \sinh^{-1}(\lambda \sqrt{Y})$ 或 $\lambda^{-1} \sinh^{-1}(\lambda \sqrt{Y} + 0.5)$	0.25

注: a. 对于较小的 Y 值, 通常建议采取 $\sqrt{Y+0.5}$ 。

b. n 表示样本容量, 对于 $Y = r/n$, 一个较好的变换是 $\sin^{-1} \sqrt{(r+3/8)/(n+3/4)}$ 。

举一个例子, 考虑下面的情形: 令 Y 为车间里发生事故的次数, X 为车间里车床的运行速度。我们希望研究发生事故的次数 Y 与车床运行速度 X 之间的关系。假如 X 与 Y 两者之间存在线性关系且可表示为

$$Y = \beta_0 + \beta_1 X + \varepsilon,$$

其中 ε 是随机误差。 Y 的均值看上去随着 X 而增加。从经验观察得知, 不常发生的事件 (发生几率很小的事件) 通常服从普松分布。我们假定 Y 服从普松分布。

由于 Y 的均值和方差是相等的^①，因而 Y 的方差是 X 的函数。因此方差齐性的假定不成立。从表 6.5 可以看到普松分布随机变量的平方根 \sqrt{Y} 的方差与均值独立，近似为 0.25。为了确保方差齐性，我们用 \sqrt{Y} 对 X 进行回归。这里选择的变换是用来稳定方差的，变换形式是由响应变量的分布假定所决定的。接下来介绍一个数据分析案例。

各航线的损伤事故

表 6.6 损伤事故数 Y 及占航班总数的比例 N

行	Y	N	行	Y	N	行	Y	N
1	11	0.0950	4	19	0.2078	7	3	0.1292
2	7	0.1920	5	9	0.1832	8	1	0.0503
3	7	0.0750	6	4	0.0540	9	3	0.0629

表 6.6 给出了某一年中从纽约州起飞的美国 $9(n=9)$ 条主要航线的损伤事故数和占航班总数的比例，并在图 6.10 中画出散点图。令 f_i 和 y_i 分别表示该年第 i 条航线的航班数和损伤事故数。第 i 条航线占航班总数的比例为

$$n_i = \frac{f_i}{\sum f_i}.$$

如果所有航线是同样安全的，则损伤事故数可以用以下的模型来解释，

$$y_i = \beta_0 + \beta_1 n_i + \varepsilon_i,$$

其中 β_0 和 β_1 是常数， ε_i 是随机误差。

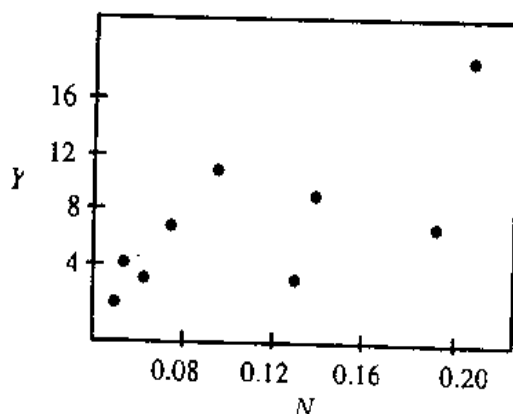


图 6.10 Y 对 N 的散点图

① 普松随机变量 Y 的概率函数为 $Pr(Y=y) = e^{-\lambda} \lambda^y / y!$; $y=0, 1, \dots$ ，其中 λ 是参数。普松随机变量的均值和方差都等于 λ 。

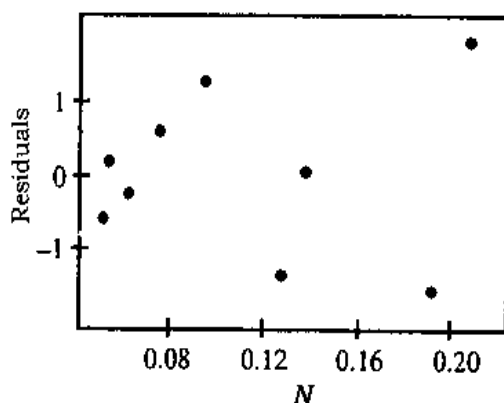
表 6.7 回归系数估计 (Y 对 N 作回归)

变量	系数	标准误	t -检验	p -值
常数	-0.14	3.14	-0.045	0.9657
N	64.98	25.20	2.580	0.0365
$n = 9$		$R^2 = 0.487$	$\hat{\sigma} = 4.201$	$d.f. = 7$

模型拟合的结果在表 6.7 中。图 6.11 给出了对 n_i 的残差图。可以看到在图 6.11 中残差随 n_i 而增加, 因此, 方差齐性的假设似乎被破坏了。这并不奇怪, 由于损伤事故数服从普松分布, 其方差和其均值成比例。为了确保方差齐性的假定, 我们做平方根变换。我们不再分析 Y 而是对 \sqrt{Y} 作分析, 其方差大约为 0.25, 并且比原始变量更近似于正态分布。

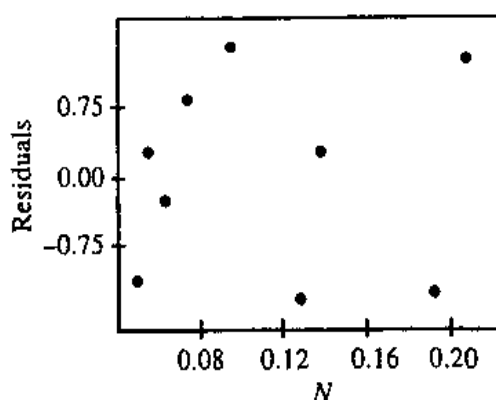
因此, 我们拟合的模型是

$$\sqrt{y_i} = \beta_0' + \beta_1' n_i + \varepsilon_i. \quad (6.8)$$

图 6.11 对 N 的标准化残差散点图表 6.8 当 $\sqrt{Y_i}$ 对 N 作回归时的回归系数估计

变量	系数	标准误	t -检验	p -值
常数	1.169	0.578	2.02	0.0829
N	11.856	4.638	2.56	0.03678
$n = 9$		$R^2 = 0.483$	$\hat{\sigma} = 0.773$	$d.f. = 7$

式 (6.8) 的拟合结果在表 6.8 中给出。(6.8) 中的对于 n_i 的残差图在图 6.12 中给出。变换后的模型残差不再随着 n_i 而增大。这表明变换后的模型没违背方差齐性的假定。现在可以采用标准方法对由 $\sqrt{y_i}$ 和 n_i 表达的模型进行分析。该回归是显著的 (通过 t 统计量来判断) 但并不很强。航线损伤事故的变异只有 48% 被其航班数的变异所解释。看来, 为了更好地解释损伤事故还应该考虑其他因素。

图 6.12 $\sqrt{p_i}$ 对 n_i 作回归后得到的对 N 的标准化残差图

前面的例子中, 响应变量(损伤事故数)的性质喻示着相对于拟合曲线的误差方差不是常数。考虑平方根转换根据的是经验事实: 意外事故发生的次数倾向于服从普松分布。而对于普松分布的随机变量, 平方根变换是合适的(表 6.5)。但是, 也存在一些情形, 误差方差不是常数, 也没有先验的理由去怀疑这个事实。对这种情形, 我们可以通过实证分析揭示出异方差性的存在, 并且采用适当的变换消除这一影响。如果没有检测到不相等的误差方差且未消除之, 则估计将会有较大的标准误, 但仍将是无偏的。这会对参数导出较宽的置信区间, 且会使检验的灵敏度较低。在下一个例子中, 我们将说明分析这种异方差模型的方法。

6.5 异方差误差的诊断

表 6.9 27 个工业企业中工人数及主管人数

行	X	Y	行	X	Y	行	X	Y
1	294	30	10	697	78	19	700	106
2	247	32	11	688	80	20	850	128
3	267	37	12	630	84	21	980	130
4	358	44	13	709	88	22	1025	160
5	423	47	14	627	97	23	1021	97
6	311	49	15	615	100	24	1200	180
7	450	56	16	999	109	25	1250	112
8	534	62	17	1022	114	26	1500	210
9	438	68	18	1015	117	27	1650	135

在一项对 27 个不同规模的工业企业的研究中, 记录了每个企业中工人人数(X)和主管人数(Y)(表 6.9)。这批数据可以在本书的网站上查到。现欲研究这两个变量之间的关系, 以下述线性模型为起点

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i. \quad (6.9)$$

表 6.10 主管人数 (Y) 对工人数 (X) 作回归时的回归系数估计

变量	系数	标准误	t-检验	p-值
常数	14.448	9.562	1.51	0.1350
N	0.105	0.011	9.30	< 0.0001
$n = 27$		$R^2 = 0.776$	$\hat{\sigma} = 21.73$	$d.f. = 25$

Y 相对于 X 的散点图提示可以将一元线性模型作为研究的起点 (图 6.13)。该线性模型拟合的结果由表 6.10 给出。

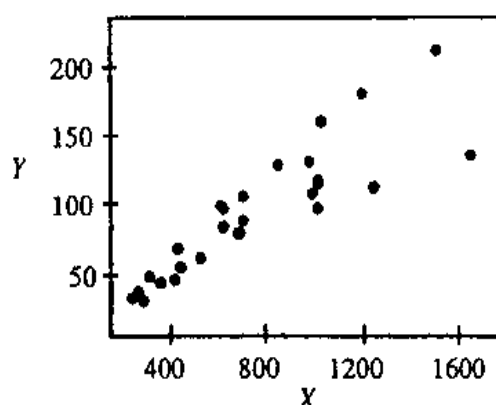


图 6.13 主管人数 (Y) 对工人数 (X) 的散点图

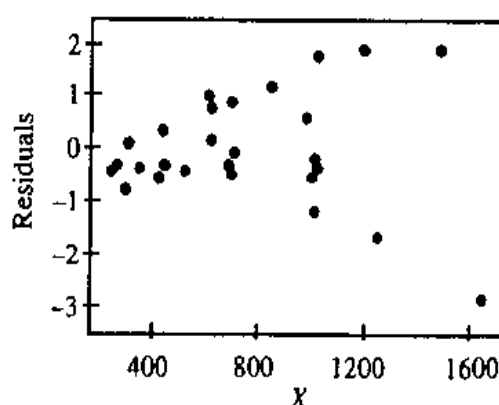


图 6.14 主管人数 (Y) 对工人数 (X) 作回归后得到的对 X 的标准化残差散点图

对 X 的标准化残差散点图 (图 6.14) 显示残差的方差随着 X 而增加。残差散布在一个沿着 X 轴的方向发散的范围。一般来讲, 如果残差散布的范围随着 X 增加而发散 (即变宽), 则误差方差也随着 X 的增加而增加。反之, 如果范围收缩 (即变窄), 则误差方差随着 X 的增加而减小。如果包含残差的范围由两条与

X 轴平行的直线围成, 则没有异方差的迹象。所以标准化残差关于预测变量的散点图指出了异方差的存在。从图 6.14 可以看出, 在我们目前的例子中, 残差随着 X 的增加而增加。

6.6 异方差性的消除

在许多工业, 经济和生物的应用中, 当遇到误差方差不相等时, 通常会发现误差标准差随着预测变量的增加而增加。基于这一经验, 在当前的例子中我们假定误差的标准差与 X 成比例 (这方面的证据可以从残差图 6.14 中得到):

$$\text{Var}(\varepsilon_i) = k^2 x_i^2, k > 0. \quad (6.10)$$

在 (6.9) 式的两边都除以 x_i , 我们得到

$$\frac{y_i}{x_i} = \frac{\beta_0}{x_i} + \beta_1 + \frac{\varepsilon_i}{x_i}. \quad (6.11)$$

现在, 定义一组新的变量和系数,

$$Y' = \frac{Y}{X}, \quad X' = \frac{1}{X}, \quad \beta'_0 = \beta_1, \quad \beta'_1 = \beta_0, \quad \varepsilon' = \frac{\varepsilon}{X}.$$

采用新的变量 (6.11) 简化为

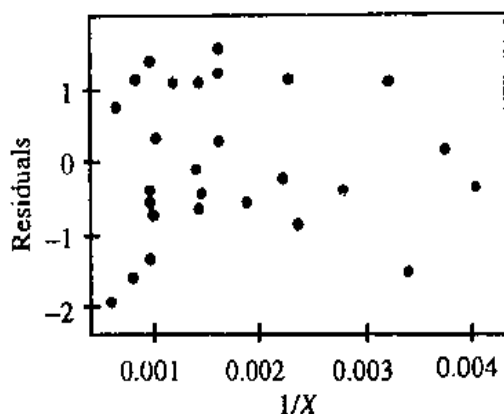
$$y'_i = \beta'_0 + \beta'_1 x'_i + \varepsilon'_i. \quad (6.12)$$

注意到在变换后的模型中, $\text{Var}(\varepsilon'_i)$ 是常数并且等于 k^2 。如果我们在 (6.10) 中给出的关于误差项的假定成立, 为了适当地拟合模型我们必须采用变换后的变量: $\frac{Y}{X}$ 和 $\frac{1}{X}$ 分别作为响应变量和预测变量。如果对于变换后数据的拟合模型是 $\hat{\beta}'_0 + \hat{\beta}'_1/X$, 则采用原始变量的拟合模型为

$$\hat{Y}' = \hat{\beta}'_1 + \hat{\beta}'_0 X. \quad (6.13)$$

变换后模型中的常数项是原模型中的 X 的回归系数, 反之亦然。这一点可以通过比较 (6.11) 和 (6.12) 看出。

变换后模型的拟合残差相对于预测变量的散点图见图 6.15。可以看到残差是随机分布的且大致散布于和横轴平行的带状区域中。变换后的模型没有明显的异方差迹象。残差分布没有表现出明显的模式, 我们可以得出结论: 变换后的模型是合适的。我们关于误差项的假定看来是正确的; 变换后的模型误差具有同方差性, 最小二乘法理论的标准假定成立。关于 $\frac{Y}{X}$ 和 $\frac{1}{X}$ 的拟合结果导出可以被原模型采用的 β'_0 和 β'_1 的估计。

图 6.15 Y/X 对 $1/X$ 作回归后得到的标准化残差关于 $1/X$ 的散点图

变换后变量的方程为 $Y/X = 0.121 + 3.083/X$ 。采用原始变量，我们有 $\hat{Y} = 3.083 + 0.121X$ 。结果总结在表 6.11 中。通过比较表 6.10 和表 6.11，我们看到通过变量变换可以降低标准误。斜率估计的方差减少了 33%。

表 6.11 当用变换后的变量 Y/X 和 X 进行拟合时，原方程的回归系数估计

变量	系数	标准差	t-检验	p-值
X	3.803	4.570	0.832	0.4131
N	0.121	0.009	13.44	< 0.0001
$n = 27$		$R^2 = 0.758$	$\hat{\sigma} = 22.577$	$d.f. = 25$

6.7 加权最小二乘法

误差项具有异方差性的线性回归模型也可以用一种所谓加权最小二乘(WLS)的方法来拟合，其中参数的估计可以通过最小化加权误差平方和得到，而权重和误差的方差成反比。这和普通的最小二乘法 (OLS) 形成对照，参数的 OLS 估计是通过最小化相同权重的误差平方和得到的。在前面的例子中，WLS 估计是通过最小化

$$\sum \frac{1}{x_i^2} (y_i - \beta_0 - \beta_1 x_i)^2 \quad (6.14)$$

得到，与之相对，OLS 估计最小化的是

$$\sum (y_i - \beta_0 - \beta_1 x_i)^2. \quad (6.15)$$

可以证明 WLS 等价于对变换后的变量 Y/X 和 $1/X$ 作 OLS。我们将此留给读者作为练习。

在第 7 章中将更详细地讨论加权最小二乘估计法。

6.8 数据的对数变换

在回归分析中,对数变换是一种应用最广泛的变换。即并不直接研究数据,而是对数据的对数进行统计分析。当被分析变量的标准差相对于其均值而言较大时,这种变换尤其有用。采用对数变换,经常会减小原始数据的波动程度和非对称性。这变换对于消除异方差性也是有效的。我们利用表 6.9 给出的工业数据说明这一点,其中已经检测出异方差性。除了演示采用对数变换消除异方差性,我们还在这个例子中证明:对于给定的一组数据会存在几种适当的描述(模型)。

我们现在不是拟合 (6.9) 中给出的模型,而是拟合模型

$$\ln y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad (6.16)$$

(即不是 Y 对于 X 作回归,而是 $\ln Y$ 对于 X 作回归)。相应的散点图由图 6.16 给出。(6.16) 的拟合结果见表 6.12。系数是显著的,这里 R^2 的值 (0.77) 可以与拟合模型 (6.9) 得到的结果相媲美的。

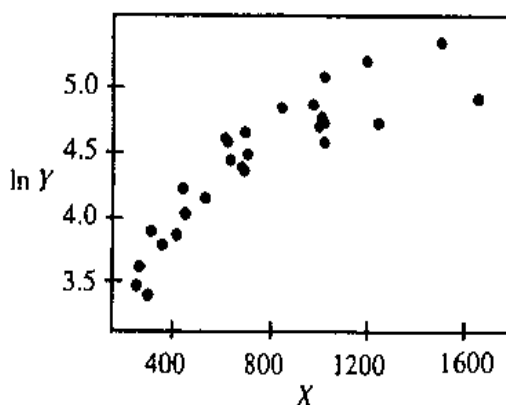


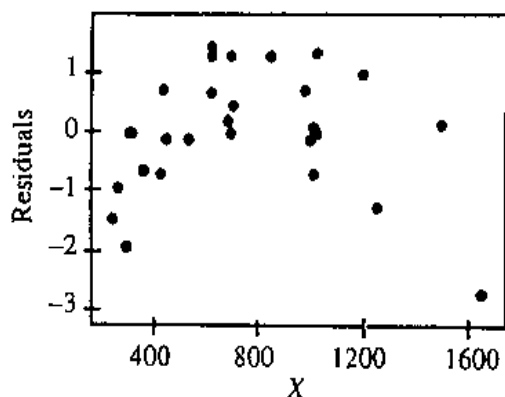
图 6.16 $\ln Y$ 相对于 X 的散点图

表 6.12 当 $\ln Y$ 对 X 作回归时的回归系数估计

变量	系数	标准误	t -检验	p -值
常数	3.515	0.1110	31.65	< 0.0001
X	0.0012	0.0001	9.15	< 0.0001
$n = 27$		$R^2 = 0.77$	$\hat{\sigma} = 0.252$	$d.f. = 25$

残差关于 X 的散点图见图 6.17。该散点图很有启发性。异方差性已被消除了,但是该图还呈现明显的非线性。残差显示出二次效应,提示对这批数据更适当的模型也许是

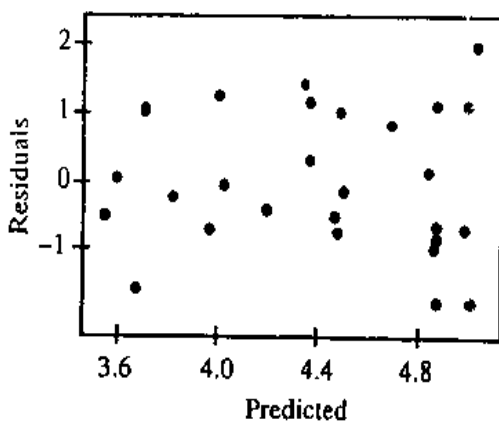
$$\ln y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i. \quad (6.17)$$

图 6.17 $\ln Y$ 对 X 作回归后得到的对 X 的标准化残差图

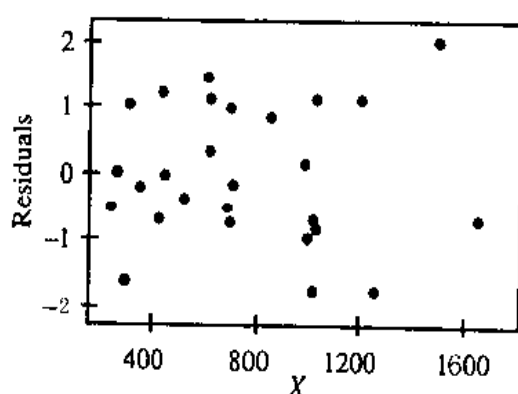
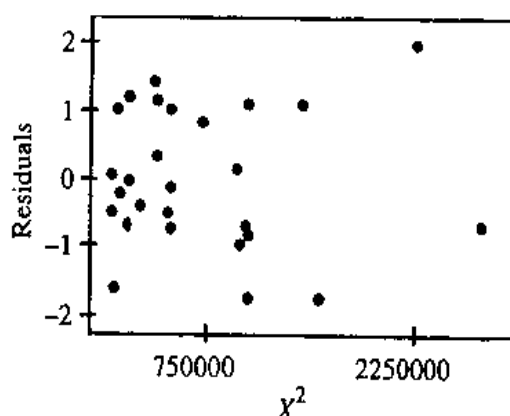
方程 (6.17) 是多元回归模型, 因为它有两个预测变量, X 和 X^2 。正如在第 4 章中讨论的那样, 残差图也可用于多元回归中检测模型的缺陷。为说明残差图用于检测模型缺陷的有效性及揭示可能采取的修正的能力, 我们将模型 (6.17) 拟合的结果总结在表 6.13 中。标准化残差对拟合值和每一个预测变量 X 和 X^2 的散点图分别见图 6.18–6.20^①。

表 6.13 $\ln Y$ 对 X 和 X^2 作回归时的回归系数估计

变量	系数	标准误	t- 检验	p- 值
常数	2.8516	0.1566	18.2	< 0.0001
X	3.11267E-3	0.0004	7.80	< 0.0001
X^2	-1.10226E-6	0.220E-6	-4.93	< 0.0001
$n = 27$		$R^2 = 0.886$	$\hat{\sigma} = 0.1817$	$d.f. = 24$

图 6.18 $\ln Y$ 对 X 和 X^2 作回归后对于拟合值的标准化残差图

① 回想我们在第 4 章的讨论, 在简单回归中残差对拟合值和对预测变量 X_1 的散点图是相同的; 因此, 我们只需要考察两个散点图中的一个。在多元回归中, 残差对拟合值的散点图与对每个预测变量的散点图都不同。

图 6.19 $\ln Y$ 对 X 和 X^2 作回归后对 X 的标准化残差图图 6.20 $\ln Y$ 对 X 和 X^2 作回归后对 X^2 的标准化残差图

包含二次项的模型残差似乎令人满意。残差没有异方差或非线性迹象。现在我们对于同一批数据有两个同样可以接受的模型。表 6.13 给出的模型也许较好，因为它有较高的 R^2 。而表 6.11 给出的模型较易解释，因为它是基于原始变量建立的。

6.9 幂变换

前面，我们采用了几种形式的变换（例如倒数变换 $1/Y$ ，平方根变换 \sqrt{Y} ，和对数变换 $\ln Y$ ）。这些变换是根据理论或经验来选择的，其目的是使模型线性化、正态化和（或）稳定误差的方差。这些变换使人想到一般的幂变换。在幂变换中，我们提高响应变量和（或）一些预测变量的幂次。例如，不用 Y 而用 Y^λ ，其中 λ 是指数，它如何选择要根据理论或经验。当 $\lambda = -1$ ，我们得到倒数变换，当 $\lambda = 0.5$ ，我们得到平方根变换，当 $\lambda = 0$ ，我们得到对数变换^①。 $\lambda = 1$ 表示不需

^① 注意，若 $\lambda = 0$ ，则 $Y^\lambda = 1, \forall Y$ 。故采用变换 $(Y^\lambda - 1)/\lambda$ 来避免这一问题。可以证明，当 $\lambda \rightarrow 0$ 时， $(Y^\lambda - 1)/\lambda \rightarrow \ln Y$ 。这种变换称为 Box-Cox 幂变换。更详细的讨论请参阅 Carroll and Ruppert(1988)。

要作变换。

如果不能根据理论决定 λ , 那么可以用数据来决定合适的 λ 。这可以用数值方法实现。在实践中, 可以试验几个 λ , 然后选择最优的。通常用来尝试的 λ 值是 2, 1.5, 1.0, 0.5, 0, -0.5, -1, -1.5, -2。选择这些值作变换是由于它们易于解释。它们被称为变换阶梯。下面举例说明。

例: 脑数据

表 6.14 中的数据是取自一个大的数据集的一个样本。这批数据也可以在本书的网站上找到。数据的出处见 Jerison (1973)。Rousseeuw and Leroy (1987) 也曾分析过这批数据。数据记录了 28 种动物两个变量的测量值: 脑重量的平均值 Y (以克计), 身体重量的平均值 X (以千克计)。数据分析的一个目的是确定一个较重的身体是否需要一个较重的脑来支配。另一个目的是看看能否可以用脑重量与身体重量之比来衡量智力水平。数据的散点图 (图 6.21) 并没有表现出这其中存在明显的联系。这主要是因为存在非常大的动物 (例如, 两个大象数据和 3 个恐龙数据)。让我们将幂变换同时应用于 Y 和 X 。 Y^λ 对 X^λ 在 λ 取变换阶梯上几个值时的散点图由图 6.22 给出。可以看到 $\lambda = 0$ (对应取对数变换) 是最合适的值。在 $\lambda = 0$ 时, 散点图呈线性, 但是 3 个恐龙的数据不符合由其他数据点揭示出的线性模式。这幅图给人的感觉是或者恐龙的脑重量被低估了。或者其身体重量被高估了。

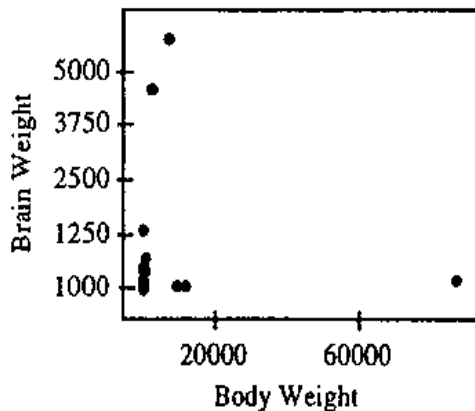
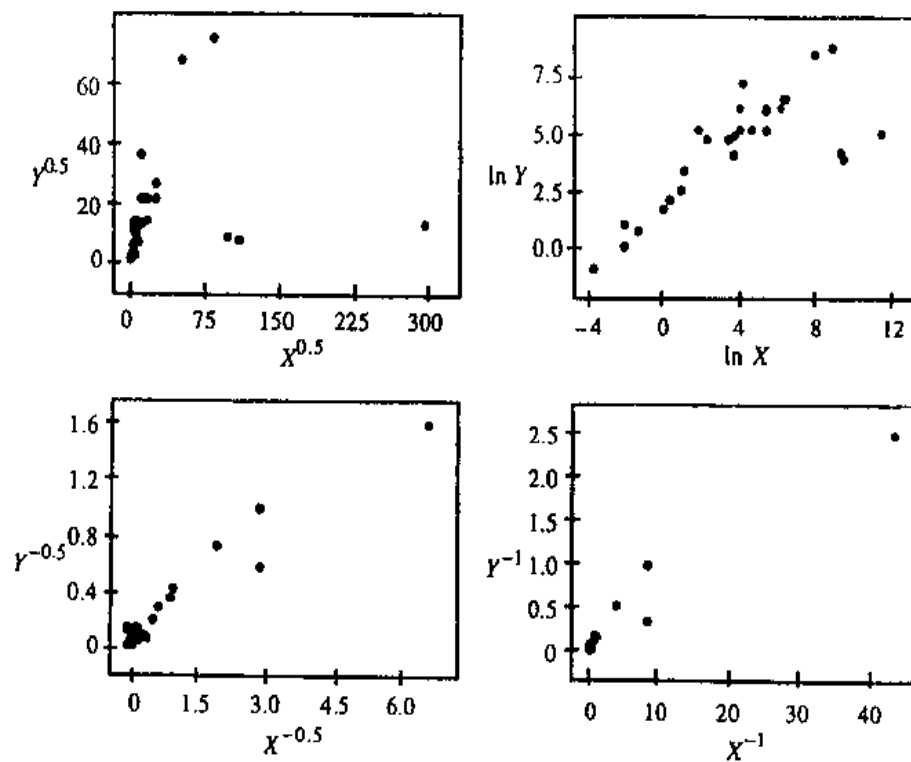


图 6.21 脑数据: 脑重量对身体重量的散点图

注意到在这个例子中, 我们对响应变量和预测变量都实施变换, 并且对两个变量选取相同的幂。在其他的应用中, 也许对每个变量用不同的幂或只对一个变量作变换更为合适。有关数据变换更详细的讨论可以参见 Carroll and Ruppert (1988) 和 Atkinson (1985)。

表 6.14 脑数据: 脑重量(克)及身体重量(千克)

名	脑重量	身体重量	名	脑重量	身体重量
Mountain beaver	8.1	1.35	African elephant	5712.0	6654.00
Cow	423.0	465.00	Triceratops	70.0	9400.00
Gray Wolf	119.5	36.33	Rhesus monkey	179.0	6.80
Goat	115.0	27.66	Kangaroo	56.0	35.00
Guinea pig	5.5	1.04	Hamster	1.0	0.12
Diplodocus	50.0	11700.00	Mouse	0.4	0.02
Asian elephant	4603.0	2547.00	Rabbit	12.1	2.50
Donkey	419.0	187.10	Sheep	175.0	55.50
Horse	655.0	521.00	Jagular	157.0	100.00
Potar monkey	115.0	10.00	Chimpanzee	440.0	52.16
Cat	25.6	3.30	Brachiosaurus	154.5	87000.00
Giraffe	680.0	529.00	Rat	1.9.0	0.28
Gorilla	406.0	207.00	Mole	3.0	0.12
Human	1320.0	62.00	Pig	180.0	192.00

图 6.22 对应各种 λ 值的 Y^λ 对 X^λ 散点图

6.10 小结

在拟合线性模型之后,应该考察残差是否呈现出异方差性。如果残差随着预测变量的取值的增减而增减,就存在异方差性,这可以很方便地从残差图检测出来。如果存在异方差性,在拟合模型时需要注意这一点,否则最小二乘的估计结果将不会有最大的精度(最小的方差)。异方差性可以通过变量变换去除。可以利用变换后模型的参数估计导出原模型中的相应参数估计。经过适当变换后的模型的残差应该不再显示出异方差性。

习 题

6.1 杂志广告: 在一项广告收入的研究中,收集了 1986 年 41 种杂志的数据(表 6.15)。观测的变量是广告的页数和广告收入,并列出了杂志名。

- 将广告收入对广告的页数拟合线性回归方程。说明这个拟合是不充分的。
- 对数据作适当的变换,然后对变换后的数据作拟合。评价拟合的效果。
- 你不应该对存在大量的异常值而感到惊讶,因为杂志是不同类的,期望有一个简单的关系来将他们联系到一起是不现实的。将异常值去掉,然后得到一个可以接受的回归方程。

6.2 风寒指数: 表 6.16 给出了在各种无风实际温度(T)和风速(V)条件下,由于风寒作用而产生的实际温度(W)。当人行走在静止空气中时(明显的风是 4 英里/小时(mph)),无风条件可以作为寒冷的评价。原始数据是国家气象局公布的,从波士顿科学博物馆的出版物中收集到。温度以华氏度($^{\circ}\text{F}$)来衡量,风速以 mph 来计。

- 表 6.16 中的数据格式并不能直接应用于回归分析。你需要另构建一个包含 3 个列的表,分别对应变量 W , T 和 V 。这个表可以在本书的网站上找到。
- 拟合 W , T 和 V 之间的线性关系。残差的模式应该显示出线性模型不合适。
- 在用 T 修正 W (比如,保持 T 不变)后,考察 W 和 V 之间的关系。它们之间的关系是否为线性?
- 在用 V 修正 W 后,考察 W 与 T 之间的关系,这关系是否线性?
- 拟合模型

$$W = \beta_0 + \beta_1 T + \beta_2 V + \beta_3 \sqrt{V} + \varepsilon. \quad (6.18)$$

这个模型是否合适? W 的值由国家气象局根据公式

$$W = 0.0817(3.71\sqrt{V} + 5.81 - 0.25V)(T - 9.14) + 91.4 \quad (6.19)$$

算得的。上面的公式是否是 W 的一个精确的数值描述?

- 你可以给出一个比 (6.18) 和 (6.19) 更好的模型吗?

表 6.15 1986 年中 41 种杂志的广告页数 (P , 以百计) 以及销售收入 (R , 以百万元计)

杂志	P	R	杂志	P	R
Cosmopolitan	25	50.0	Town and country	1	7.0
Redbook	15	49.7	True Story	77	6.6
Glamour	20	34.0	Brides	13	6.2
Southern Living	17	30.7	Book Digest Magazine	5	5.8
Vogue	23	27.0	W	7	5.1
Sunset	17	26.3	Yankee	13	4.1
House and Garden	14	24.6	Playgirl	4	3.9
New York Magazine	22	16.9	Saturday Review	6	3.9
House Beautiful	12	16.7	New Woman	3	3.5
Mademoiselle	15	14.6	Ms.	6	3.3
Psychology Today	8	13.8	Cuisine	4	3.0
Life Magazine	7	13.2	Mother Earth News	3	2.5
Smithsonian	9	13.1	1001 Decorating Ideas	3	2.5
Rolling Stone	12	10.6	Self	5	2.3
Modern Bride	1	8.8	Decorating & Craft Ideas	4	1.8
Parents	6	8.7	Saturday Evening Post	4	1.5
Architectural Digest	12	8.5	McCall's Needleweek and Craft	3	1.3
Harper's Bazaar	9	8.3	Weight Watchers	3	1.3
Apartment Life	7	8.2	High Times	4	1.0
Bon Appetite	9	8.2	Soap Opear Digest	2	0.3
Gourmet	7	7.3			

表 6.16 在各种风速 V (英里/小时) 和温度 T ($^{\circ}\text{F}$) 条件下的风寒指数

V	50	40	30	20	10	0	-10	-20	-30	-40	-50	-60
5	48	36	27	17	5	-5	-15	-25	-35	-46	-56	-66
10	40	29	18	5	-8	-20	-30	-43	-55	-68	-80	-93
15	35	23	10	-5	-18	-29	-42	-55	-70	-83	-97	-112
20	32	18	4	-10	-23	-34	-50	-64	-79	-94	-108	-121
25	30	15	-1	-15	-28	-38	-55	-72	-88	-105	-118	-130
30	28	13	-5	-18	-33	-44	-60	-76	-92	-109	-124	-134
35	27	11	-6	20	-35	-48	-65	-80	-96	-113	-130	-137
40	26	10	-7	-21	-37	-52	-68	-83	-100	-117	-135	-140
45	25	9	-8	-22	-39	-54	-70	-86	-103	-120	-139	-143
50	25	8	-9	-23	-40	-55	-72	-88	-105	-123	-142	-145

6.3 参见表 5.17 中总统选举的数据, 其中响应变量 V 是美国总统候选人的得票率。由于响应变量是一个比率, 它的取值介于 0 和 1 之间。变换 $Y = \log(V/(1-V))$ 将取值介于 0 和 1 之间的 V 变成取值介于 $-\infty$ 和 $+\infty$ 之间

的 Y 。所以,更有理由期望 Y 比 V 更满足正态性的假定。

(a) 考虑拟合模型

$$Y = \beta_0 + \beta_1 \cdot I + \beta_2 \cdot D + \beta_3 \cdot W + \beta_4 \cdot (G \cdot I) + \beta_5 \cdot P + \beta_6 \cdot N + \varepsilon, \quad (6.20)$$

这和 (5.11) 中的模型一致,只是以 Y 代替 V 。

(b) 对每个模型,考察在第4章中讨论过的适当的残差图以确定哪个模型更满足标准假定,是原始变量 V 的模型呢还是变换后的变量 Y 的模型。

(c) (6.20) 隐含的原始变量 V 与预测变量的模型形式是什么?即寻找函数形式 f ,使

$$V = f(\beta_0 + \beta_1 \cdot I + \beta_2 \cdot D + \beta_3 \cdot W + \beta_4 \cdot (G \cdot I) + \beta_5 \cdot P + \beta_6 \cdot N + \varepsilon). \quad (6.21)$$

[提示:这是个非线性函数,称为 logistic 函数,将在第12章里讨论。]

6.4 石油产量数据:表 6.17 中的数据是 1880-1988 年间以百万桶计的世界原油年产量。数据取自 Moore 和 McCabe(1993), p.147。

表 6.17 世界原油年产量,以百万桶计(1880-1988 年)

年份	OIL	年份	OIL	年份	OIL
1880	30	1940	2150	1972	18584
1890	77	1945	2595	1974	20389
1900	149	1950	3803	1976	20188
1905	215	1955	5626	1978	21922
1910	328	1960	7674	1980	21722
1915	432	1962	8882	1982	19411
1920	689	1964	10310	1984	19837
1925	1069	1966	12016	1986	20246
1930	1412	1968	14104	1988	21338
1935	1655	1970	16690		

(a) 构造一个石油产量 (OIL) 对年份的散点图,并观察该图像中点的散布呈非线性状。为了对这些数据拟合线性模型,必须对 OIL 作变换。

(b) 构造 $\log(\text{OIL})$ 对年度的散点图。从 1880 到 1973 年的点呈一条直线。中东石油产区的政治骚乱影响了 1973 年后石油产量的模式。

(c) 将 $\log(\text{OIL})$ 对年份作线性回归。评价模型的拟合效果。

(d) 构造标准化残差的序列图。这幅图清楚地显示出标准假定中的一条假定被违反了。哪一条?

6.5 计算机行业一项显著的技术进步是在硬盘上高密度地存储信息的能力。存储的成本持续下降。表 6.18 给出了 1988-1998 年以美元计存储每兆字节的平均价格。

表 6.18 存储每兆字节的平均价格 (1988–1998 年)

年份	价格	年份	价格
1988	11.54	1994	0.705
1989	9.30	1995	0.333
1990	6.86	1996	0.179
1991	5.23	1997	0.101
1992	3.00	1998	0.068
1993	1.46		

- (a) 是否能用线性时间趋势描绘数据? 用编码方法定义一个新的变量 t , 将 1988 取值为 1, 1989 取值为 2, 等等。
- (b) 拟合模型 $P_t = P_0 e^{\beta t}$, 其中 P_t 是在时期 t 的价格。这个模型是否合适地描述了数据?
- (c) 引入一个示性变量, 对 1988–1991 年取 0, 对其他年度取 1。拟合 $\log(P_t)$ 关于时间 t , 示性变量, 以及时间与示性变量的乘积的模型。解释拟合模型的系数。

7

加权最小二乘法

7.1 引言

至此，我们在回归分析的讨论中，都假定回归模型的形式是

$$y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i, \quad (7.1)$$

其中 ε_i 是随机误差，它们相互独立，且同分布，均值为 0，方差为 σ^2 。可利用各种残差图来验证这一假定（见第 4 章）。如果残差不满足这一假定，则表明或许回归方程的形式不恰当，或许需添加其他变量，或许数据中有些观测是异常点。

但也有例外，以 6.5 节中主管人员数据为例，原模型的误差不是 i.i.d. 的。更确切地说，这些误差方差不是常数。对于这批数据，可以应用变换去修正这一情形，从而可以得到原模型参数较好的估计（比普通的最小二乘法（OLS）好）。

在这一章和第 8 章，我们研究误差不是 i.i.d. 的情形。本章处理异方差问题，即误差方差不相同，第 8 章处理自相关问题，即误差不是相互独立的。

第 6 章介绍了用变量变换来处理异方差性以稳定方差，而加权最小二乘法（WLS）等价于对变换后的变量实施 OLS。这里介绍的 WLS 方法不仅是一种处理异方差性的手段，其本身也是一种估计方法。例如，在拟合剂量 - 反应曲线（7.5 节）和 logistic 模型中（7.5 节和第 12 章），WLS 比 OLS 更好。

在本章中，不假定误差的方差相等。因此只假定 ε_i 相互独立，均值为 0 和 $\text{Var}(\varepsilon_i) = \sigma_i^2$ 。在这种情形下，我们用 WLS 方法去估计 (7.1) 中的回归系数。所谓 WLS 估计即指通过最小化

$$\sum_{i=1}^n \omega_i (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})^2$$

得到 $\beta_0, \beta_1, \dots, \beta_p$ 的估计，其中 ω_i 是权重，与误差的方差成反比（即 $\omega_i = 1/\sigma_i^2$ ）。不难看出在利用 WLS 方法估计 $\beta_0, \beta_1, \dots, \beta_p$ 时，任何具有小权重的观测的影响

将会被极大地削弱。最极端的情况是当 $\omega_i = 0$ 时, WLS 估计会将第 i 个观测忽略掉。

在采用的 WLS 方法时, 我们应用了数据产生过程中的先验信息, 以及在用 OLS 拟合产生的残差去检测异方差性时所发现的线索。如果权重是未知的, 通常的解决办法是两阶段法。在第一阶段, 采用 OLS 的结果来估计权重。在第二阶段, 应用第一阶段估计的权重实施 WLS。本章后面将会举例说明。

7.2 异方差模型

有三种情形可能产生异方差。前两种情形中, 一旦识别出异方差的来源, 估计可以在一阶段内完成。第三种情形较为复杂并需要采用前面提到的两阶段估计法。第一种情形的例子可以在第 6 章里找到, 这里也将回顾一下。对第二种情形将只加以描述, 但没有数据分析的例子。对第三种情形将用两个例子说明。

7.2.1 主管人员数据

6.5 节给出了 27 个工业企业中每个企业工人数 (X) 和管理人员数 (Y) 的数据。假定的回归模型为

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad (7.2)$$

ε_i 的方差依赖于用 x_i 度量的企业的规模, 即 $\sigma_i^2 = k^2 x_i^2$, 其中 k 是正的常数 (详见 6.5 节)。这种类型异方差的经验证据可以从标准化残差对 X 的散点图得到。图 7.1 是这种情形的代表。残差分布呈漏斗形状, 随着 X 的值增大或者散开或者收拢。如果没有采取修正措施而直接对原数据实施 OLS, 那么在理论上, 系数的估计将会缺乏精确性。此外, 这类异方差通常会低估回归系数的标准差, 从而对精度形成错误的认识。而这个问题可以通过第 6 章描述的加权最小二乘法来解决。

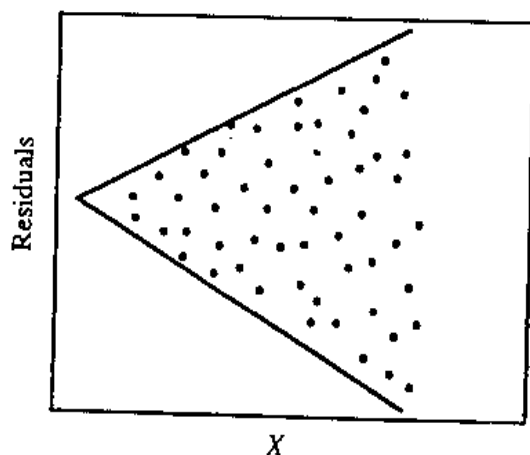


图 7.1 异方差残差的一个例子

在多元回归中也会考虑这种处理异方差性的方法。在(7.1)中误差的方差也许只受一个预测变量的影响。(方差是多个预测变量函数的情形将在后面讨论。)标准化残差对可疑变量的散点图提供了经验证据。例如,如果对于(7.1)给出的模型,发现其标准化残差对 X_2 的散点图产生出与图7.1类似的形状,那么就可以假定 $Var(\varepsilon_i)$ 与 x_{i2}^2 成比例,即 $Var(\varepsilon_i) = k^2 x_{i2}^2$,其中 $k > 0$ 。这时可以通过最小化

$$\sum_{i=1}^n \frac{1}{x_{i2}^2} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

得到参数的估计。如果采用的软件配有加权最小二乘法的程序,我们可以将权重变量定为 $1/x_{i2}^2$ 。而如果软件只能运行OLS,我们就按第6章描述的方法将数据进行变换。即把(7.1)两边除以 x_{i2} 从而得到

$$\frac{y_i}{x_{i2}} = \beta_0 \frac{1}{x_{i2}} + \beta_1 \frac{x_{i1}}{x_{i2}} + \cdots + \beta_p \frac{x_{ip}}{x_{i2}} + \frac{\varepsilon_i}{x_{i2}},$$

变量 $1/X_2$ 的系数的OLS估计是 β_0 的WLS估计。变量 X_j/X_2 的系数是 β_j 的估计, $j \neq 2$ 。拟合中的常数项是 β_2 的估计。简单回归方法的详细讨论可参见第6章。

7.2.2 大学支出数据

第二种异方差情形多见于大规模调查,其观测是从划分好的组或类中抽取的若干抽样单元的平均值。一般,会记录从每个类中抽取的抽样单元的个数及其平均值。在有些情况下,也会记录衡量差异程度的量,如标准差或极差。

表 7.1 教育费用调查中的变量

变量名	描述
Y	年度总费用(超出学费)
X_1	学校所处的城镇规模
X_2	到最近的市中心的距离
X_3	学校类型(公立或私立)
X_4	学生规模
X_5	新生的毕业比例
X_6	离家距离

例如,考虑一个针对大学生的调查,其目的是为了估计每年与大学教育相关的总支出,并考察这些支出与其就读学院的特征之间的关系。表7.1给出一组用来解释支出的变量。可以用模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_6 X_6 + \varepsilon \quad (7.3)$$

作回归分析来研究这种关系。在本例中,一个类别就是一个学校,一个抽样单元就是一个学生。数据是通过随机选择一组学校,然后在每个学校随机采访一定数

量的学生收集的。(7.3) 中的响应变量 Y 是第 i 个学校学生的平均支出。预测变量是这个学校的一些特征。该校这些变量的数值通过公布的官方统计资料得到。

平均支出的精度与样本容量的平方根成比例, 即 \bar{y}_i 的标准差是 $\sigma/\sqrt{n_i}$, 其中 n_i 表示第 i 个学校中接受采访的学生人数, σ 是学生总体年支出的标准差。那么, 模型 (7.1) 中 ε_i 的标准差是 $\sigma_i = \sigma/\sqrt{n_i}$ 。采用 WLS 估计回归系数, 其中权重为 $\omega_i = 1/\sigma_i^2$ 。由于 $\sigma_i^2 = \sigma^2/n_i$, 通过最小化加权误差平方和

$$S = \sum_{i=1}^n n_i (y_i - \beta_0 - \sum_{j=1}^6 \beta_j x_{ij})^2 \quad (7.4)$$

得到回归系数的估计。这种做法默认为, 采访了较多学生的学院的数据比只采访少量学生的学院的数据更可靠, 它们在决定回归系数时占的权重应较大。对不同观测具有不同精度的情况采用这种权重机制是合理的。

系数的估计和描述统计量可以通过专门的 WLS 计算机程序, 或对变换后的数据采用 OLS 而得到。将 (7.1) 式的两边同时乘以 $\sqrt{n_i}$, 我们得到新模型

$$y_i \sqrt{n_i} = \beta_0 \sqrt{n_i} + \beta_1 x_{i1} \sqrt{n_i} + \cdots + \beta_6 x_{i6} \sqrt{n_i} + \varepsilon_i \sqrt{n_i}. \quad (7.5)$$

(7.5) 式中的误差项 $\varepsilon_i \sqrt{n_i}$ 满足等方差的假定。将 $y_i \sqrt{n_i}$ 关于 $\sqrt{n_i}$ 以及 6 个做过变换的预测变量 $x_{ji} \sqrt{n_i}$ 组成的 7 个新变量作回归。应用 OLS 将会得到回归系数的估计和它们的标准误。注意到 (7.5) 中的回归模型有 7 个预测变量, 一个新变量 $\sqrt{n_i}$ 和 6 个原始预测变量乘以 $\sqrt{n_i}$ 。又注意到在 (7.5) 式中没有常数项, 因为原模型中的常数项 β_0 现在是 $\sqrt{n_i}$ 的系数。因此对变换后的变量作回归必须限制常数项为 0。也就是说, 我们拟合一个没有截距的模型。有关这一点更多的细节在 7.4 节中给出。

7.3 两阶段估计

前面的两个问题中, 一开始就预料到异方差性的存在。第一个问题中, 被研究过程的本身揭示出误差方差会随着预测变量的增大而增大。第二个例子中, 数据收集的方法隐含着异方差性。这两个例子中, 方差齐性可通过变量变换得到。变换是根据原始数据中所含的信息直接构造来的。本节所描述的问题中, 也会存在方差不相等的先验暗示。但是, 异方差的准确结构是通过经验决定的。所以, 回归参数的估计需要两个阶段。

在多元回归中检测到异方差性不是件容易的事。通常是依靠分析者非常好的直觉发现的。对于多元回归模型, 标准化残差对拟合值以及对每个预测变量的散点图可作为检测异方差的第一步。如果残差的大小随着 \hat{y}_i 或 x_{ij} 系统地变动, 就表明异方差存在。但是, 散点图并不能解释为什么方差会不同。(见下面的例子)。

当一组给定的预测变量值对应的响应变量有重复的测量时, 有一种方法可以直接检测异方差。例如, 在一个预测变量的情形下, 我们在 x_1 上有测量值

$y_{11}, y_{21}, \dots, y_{n_1 1}$, 在 x_2 上有测量值 $y_{12}, y_{22}, \dots, y_{n_2 2}$, 在 x_k 上有测量值 $y_{1k}, y_{2k}, \dots, y_{n_k k}$. 为举例说明, 取 $k=5$, 数据的散点图见图 7.2. 拥有如此丰富的数据, 没有必要在异方差的性质上作限制性的假定. 从图上可以清楚地看出异方差并没有遵循一个简单形式, 例如 $\text{Var}(\varepsilon_i) = k^2 x_i^2$. 方差首先随着 x 的增加而降低直至 x_4 , 然后在 x_5 上有一个跳跃. 回归模型可以描述为

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij}, \quad i = 1, 2, \dots, n_j; \quad j = 1, 2, 3, 4, 5. \quad (7.6)$$

其中 $\text{Var}(\varepsilon_{ij}) = \sigma_j^2$.

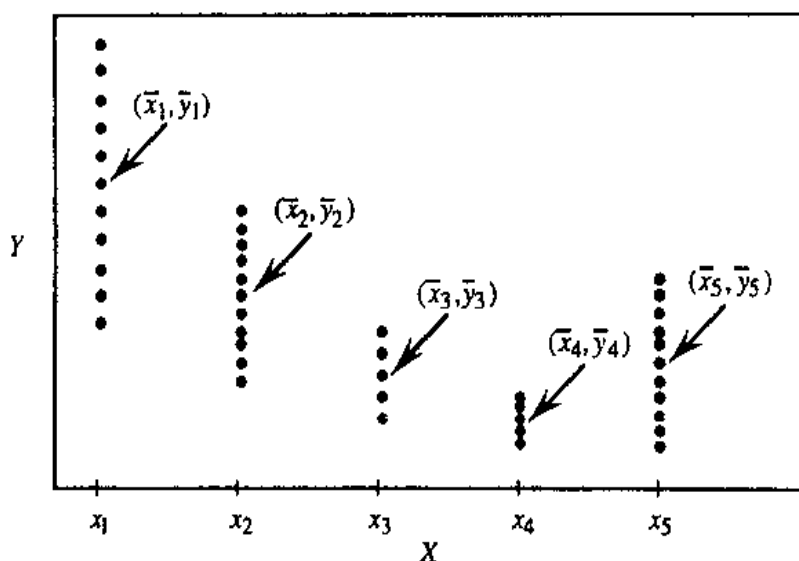


图 7.2 具有重复观测值的非常数方差

在第 j 类或第 j 组的第 i 个观测值的残差是 $e_{ij} = y_{ij} - \hat{y}_{ij}$. 加上再减去响应变量在第 j 类中的均值 \bar{y}_j , 我们得到

$$e_{ij} = (y_{ij} - \bar{y}_j) + (\bar{y}_j - \hat{y}_{ij}). \quad (7.7)$$

这表明残差由两部分组成, y_{ij} 与 \bar{y}_j 的差以及 \bar{y}_j 与回归线上的点 \hat{y}_{ij} 的差. 这第一部分被称为纯误差. 第二部分衡量的是拟合优度. 可根据纯误差^①作出异方差性的评价. WLS 的权重被估计为 $\omega_{ij} = 1/s_j^2$, 其中

$$s_j^2 = \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 / (n_j - 1)$$

是第 j 组响应变量的方差.

^① 纯误差的概念也可被用于拟合优度检验 [参见, 例, Draper and Smith (1998)].

在可控制的实验室环境下收集数据时,研究人员可以在预测变量的任意值上进行重复观测。但是在非实验室环境下收集数据时,对 X 的某个给定的值响应变量有重复观测值的情况并不常见。当只有一个预测变量时,一些重复观测可能会发生。但是当有多个预测变量时,难以想像碰巧两个观测的所有预测变量值完全相同。但是,根据预测变量近似相等而对响应变量进行分类则有可能形成伪重复观测。读者可以参阅 Daniel 和 Wood(1980),其中对这些方法有详细的讨论。多元回归中一个较可行的研究异方差的方法是根据先验的、自然的以及有意义的联系等对观测值进行归类。作为例子,我们分析州教育经费的数据。这些数据在第 5 章中用过。

7.4 教育经费数据

教育经费数据在 5.7 节中使用过,在那里将这些数据看成是跨时期的(这批数据分别是在 1965, 1970 和 1975 得到的)以检测系数的稳定性。这里我们采用这批数据来说明多元回归中处理异方差的方法,并分析区域特征对于回归关系的影响。我们将只使用 1975 年的数据作分析。目的是使用 50 个州的数据来得到教育经费与其他变量之间关系的最佳表示。数据根据地理区域自然分组。我们的假定是,对于不同区域尽管回归关系在结构上是相同的,但是系数和误差方差可以不同。这些不同的方差形成了在分析中能直接处理的异方差问题。变量名称和定义见表 7.2,数据见表 7.3,也可在本书的网站上查到^①。模型为

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon. \quad (7.8)$$

表 7.2 各州教育经费,变量列表

变量	描述
Y	1975 年人均教育费用
X_1	1973 年的人均收入
X_2	1974 年每 1000 人中年龄在 18 岁以下的居民数
X_3	1970 年每 1000 人中居住在城市居民数

基于地域同质性的假定,可以将这些州分成不同的地域,采用四个范围较广的地理分组:(1) 东北,(2) 中北部,(3) 南部和(4) 西部。为了考察地域影响或检验不同地域回归关系的等同性,可以采用示性变量去分析这批数据。但是,这里我们的目的是寻找一个对所有的地域和州都适用的最好的关系。通过加权最小二乘法的一种推广,将地域间的差异考虑在内,便可实现这一目标。

假定四个区域各自有不同的误差方差,分别记为 $(c_1\sigma)^2, (c_2\sigma)^2, (c_3\sigma)^2$, 以及 $(c_4\sigma)^2$, 其中 σ 是相同的部分, c_j 对于不同的区域是不同的。根据加权最小二乘

^① <http://www.irl.cornell.edu/~hadi/RABE>

表 7.3 教育经费数据

行	州	Y	X ₁	X ₂	X ₃	地域	行	州	Y	X ₁	X ₂	X ₃	地域
1	ME	235	3944	325	508	1	2	NH	231	4578	323	564	1
3	VT	270	4011	328	322	1	4	MA	261	5233	305	846	1
5	RI	300	4780	303	871	1	6	CT	317	5889	307	774	1
7	NY	387	5663	301	856	1	8	NJ	285	5759	310	889	1
9	PA	300	4894	300	715	2	10	OH	221	5012	324	753	2
11	IN	264	4908	329	649	2	12	IL	308	5753	320	830	2
13	MI	379	5439	337	738	2	14	WI	342	4634	328	659	2
15	MN	378	4921	330	664	2	16	IA	232	4869	318	572	2
17	MO	231	4672	309	701	2	18	ND	246	4782	333	443	2
19	SD	230	4296	330	446	2	20	NB	268	4827	318	615	2
21	KS	337	5057	304	661	2	22	DE	344	5540	328	722	3
23	MD	330	5331	323	766	3	24	VA	261	4715	317	631	3
25	WV	214	3828	310	390	3	26	NC	245	4120	321	450	3
27	SC	233	3817	342	476	3	28	GA	250	4243	339	603	3
29	FL	243	4647	287	805	3	30	KY	216	3967	325	523	3
31	TN	212	3946	315	588	3	32	AL	208	3724	332	584	3
33	MS	215	3448	358	445	3	34	AR	221	3680	320	500	3
35	LA	244	3825	355	661	3	36	OK	234	4189	306	680	3
37	TX	269	4336	335	797	3	38	MT	302	4418	335	534	4
39	ID	268	4323	344	541	4	40	WY	323	4813	331	605	4
41	CO	304	5046	324	785	4	42	NM	317	3764	366	698	4
43	AZ	332	4504	340	796	4	44	UT	315	4005	378	804	4
45	NV	291	5560	330	809	4	46	WA	312	4989	313	726	4
47	OR	316	4697	305	671	4	48	CA	332	5438	307	909	4
49	AK	546	5613	386	484	4	50	HI	311	5309	333	831	4

表 7.4 回归结果: 各州教育经费 (n=50)

变量	系数	标准差	t-检验	p-值
常数	-556.568	123.200	-4.52	< 0.0001
X ₁	0.072	0.012	6.24	< 0.0001
X ₂	1.552	0.315	4.93	< 0.0001
X ₃	-0.004	0.051	-0.08	0.9342
n = 50	R ² = 0.591	R _a ² = 0.565	$\hat{\sigma}$ = 40.47	d.f. = 46

法的原理, 回归系数应当由最小化

$$S_w = S_1 + S_2 + S_3 + S_4$$

来确定。其中

$$S_j = \sum_{i=1}^{n_j} \frac{1}{c_j^2} (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \beta_3 x_{i3})^2, \quad j = 1, 2, 3, 4. \quad (7.9)$$

S_1 到 S_4 分别对应一个地域, 只有属于这个地域的州的数据才累加上去。因子 $1/c_j^2$ 是权重, 它决定每个观测对于估计回归系数有多大的影响。权重方案在直觉上是合理的, 因为那些最不稳定的 (误差方差最大的) 观测应该在系数确定中具有最小的影响。

WLS 估计的合理性还有另一个理由。我们的目标是作数据变换使模型的参数不受影响, 但使变换后的模型中误差方差是常数。所采取的变换是将每个观测值除以适当的 c_j , 结果产生 Y/c_j 关于 $1/c_j, X_1/c_j, X_2/c_j, X_3/c_j$ 的回归^①。那么, 在概念上, 误差项也是被 c_j 除, 结果产生的误差具有相同的方差 σ^2 , 从而由此估计的系数具有标准最二乘法的所有性质。

c_j 的值是未知的, 必须与 σ^2 和 β 一样加以估计。我们建议两阶段估计法。在第一阶段, 按模型 (7.8) 用原始数据实施回归。将残差根据地域分组, 然后计算地域误差方差的估计。例如, 在东北部, 计算 $\hat{\sigma}_1^2 = \sum e_i^2/9$, 其中求和是将东北部 9 个州对应的 9 个残差的平方累加起来。以同样的方式计算 $\hat{\sigma}_2^2, \hat{\sigma}_3^2, \hat{\sigma}_4^2$ 。在第二个阶段, 用 $\hat{\sigma}_j^2$ 替代 (7.9) 式中的 c_j^2 。

表 7.4 给出的是使用 50 个州的数据做出的第一个阶段 (OLS) 的回归结果。用两个残差图来考察模型设定: 一个是标准化残差对拟合数据的散点图 (图 7.3), 另一个是标准化残差对代表不同地域的分类变量的散点图 (图 7.4)。图 7.3 的目的是探查残差大小、变化与拟合值之间的函数关系模式。观测到的散点图具有漏斗形状, 意味着异方差的存在。图 7.4 中残差的散布程度对于不同的地域是不同的, 这也意味着不同地域的方差是不相同的。标准化残差对每个预测变量的散点图 (图 7.5 至图 7.7) 显示出误差的方差随着 X_1 取值的增加而增加。

考察这个例子中的标准化残差及影响度量是很有启示意义的。读者可以验证观测 49 (阿拉斯加) 是一个异常点, 其标准化残差值为 3.28。这个观测的标准化残差在图 7.3 中与其他残差值是分离的。观测点 44 (犹他州) 和 49 (阿拉斯加) 是高杠杆点, 其杠杆值分别为 0.29 和 0.44。在考察影响度量时, 我们发现只有观测 49 是一个强影响点, 其 Cook's 距离值为 2.13, DFIT 值为 3.30。犹他州是高杠杆点但没有强影响。而阿拉斯加既是高杠杆点也是强影响点。与其他州相比, 阿拉斯加代表一种非常特殊的情形: 一个人口非常少但石油收入丰厚的州。这一年是 1975 年! 因此阿拉斯加的教育预算与其他州并不具有严格的可比性。所以,

① 如果用两个下标, i 和 j , 表示一个变量, j 表示地域, i 表示在这个地域中的观测, 那么在地域 j 中的一个观测的每个变量都被 c_j 除。注意到 β_0 是与变量 $1/c_j$ 的系数。变换后模型为

$$\frac{y_{ij}}{c_j} = \beta_0 \frac{1}{c_j} + \beta_1 \frac{x_{1ij}}{c_j} + \beta_2 \frac{x_{2ij}}{c_j} + \beta_3 \frac{x_{3ij}}{c_j} + \epsilon'_{ij},$$

ϵ'_{ij} 的方差为 σ^2 。注意到变换后的模型与原模型具有相同的系数。变换后的模型也是无截距模型。

在后面的分析中剔除这个观测（阿拉斯加），因为它对回归结果有相当大的影响，因而会扭曲整体状况。

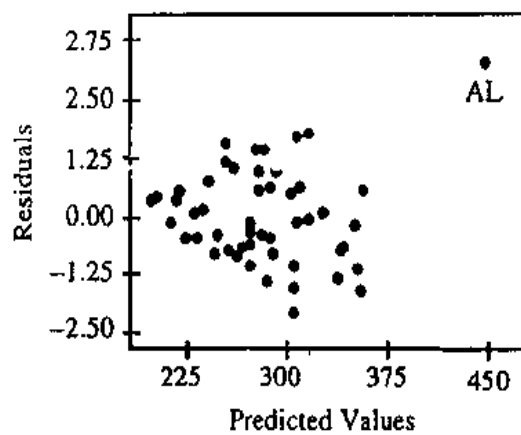


图 7.3 对拟合值的标准化残差图

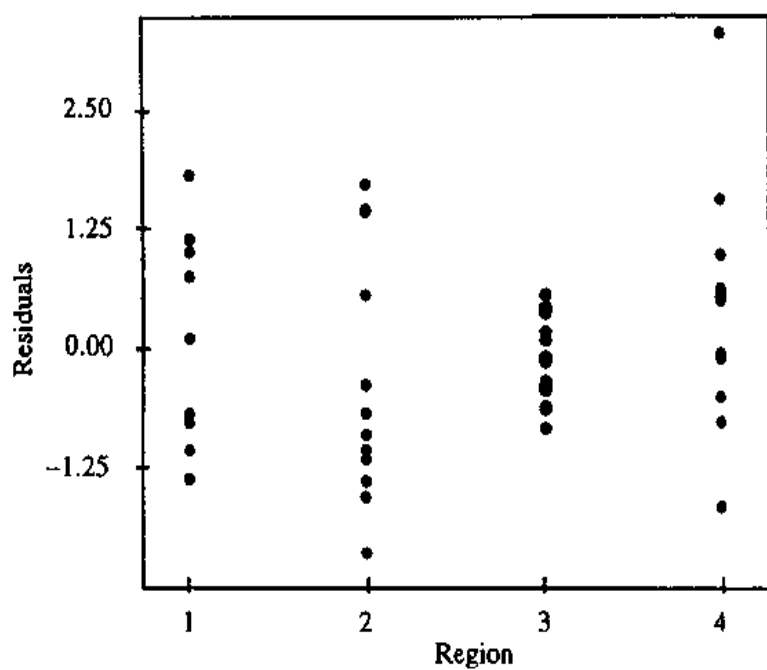
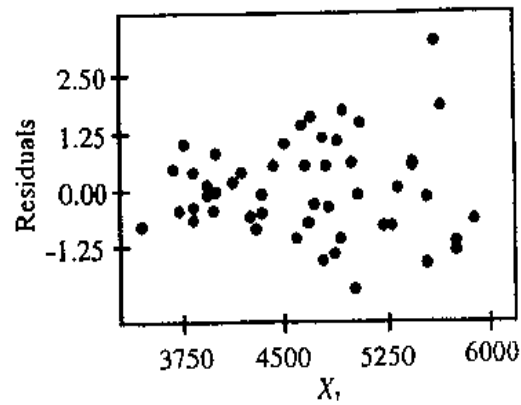
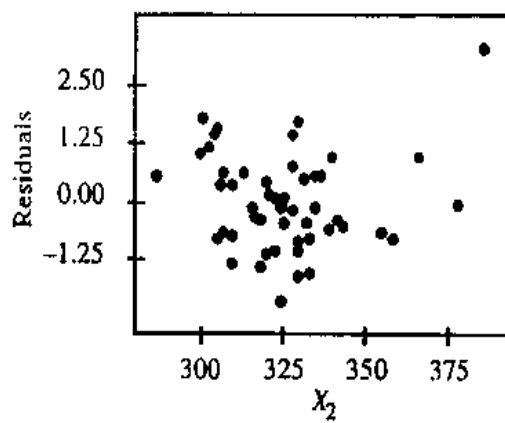
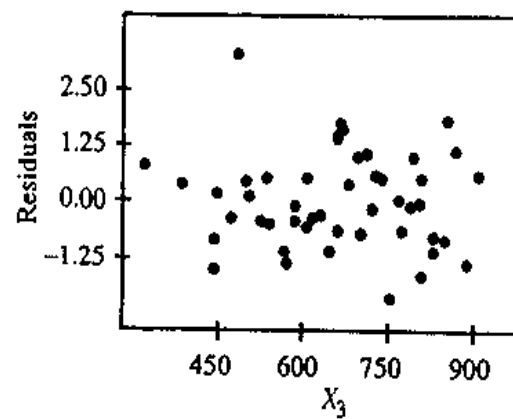


图 7.4 对区域变量的标准化残差图

图 7.5 对预测变量 X_1 的标准化残差图图 7.6 对预测变量 X_2 的标准化残差图图 7.7 对预测变量 X_3 的标准化残差图

阿拉斯加的数据对回归系数的估计有不恰当的影响。为了检查这种影响，在

剔除阿拉斯加的情况下重新计算回归。系数估计值的改变非常显著。见表 7.5。所以在后面的分析中将阿拉斯加这个观测剔除。类似于图 7.3 和 7.4 的散点图作于图 7.8 和图 7.9。剔除阿拉斯加后，图 7.8 和图 7.9 仍显示出异方差性。

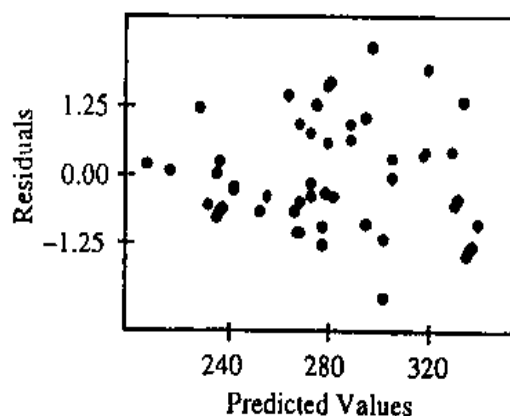


图 7.8 对拟合值的标准化残差图（剔除阿拉斯加）

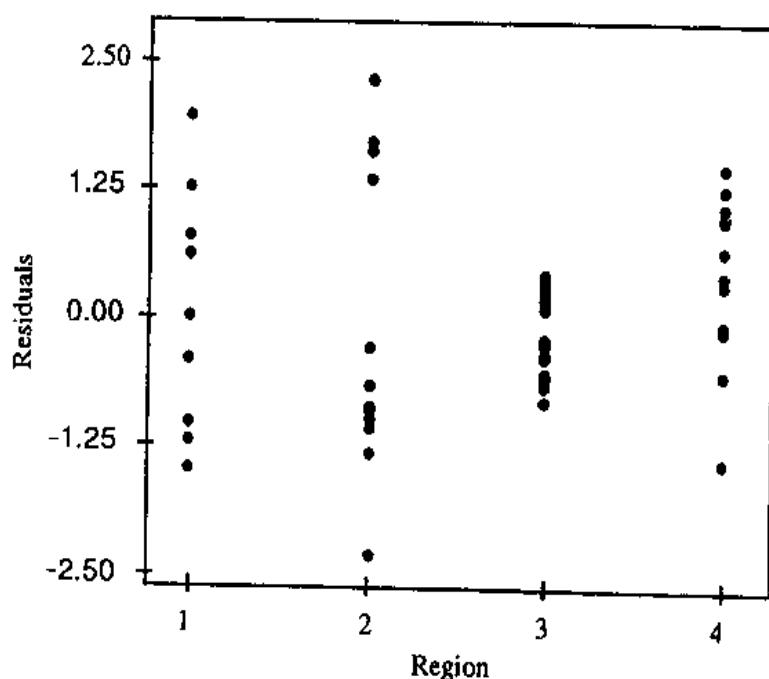


图 7.9 对地域变量的标准化残差散点图（剔除阿拉斯加）

为继续进行分析，我们必须得到权重。它们是采用上面描述的方法通过 OLS 残差计算出来的，结果见表 7.6。WLS 的回归结果见表 7.7，表中同时列出 OLS 的结果作比较。变换后的模型的标准残差图见图 7.10 和图 7.11。标准化残差对拟

表 7.5 回归结果: 各州教育经费 ($n=49$), 剔除阿拉斯加

变量	系数	标准误	t-检验	P-值
常数	-277.577	132.400	-2.10	0.0417
X_1	0.048	0.012	3.98	0.0003
X_2	0.887	0.331	2.68	0.0103
X_3	0.067	0.049	1.35	0.1826
$n = 49$	$R^2 = 0.497$	$R_a^2 = 0.463$	$\hat{\sigma} = 35.81$	$d.f. = 45$

合值的散点图没有明显的模式 (图 7.10)。而且, 从图 7.11 来看, 不同地域的残差散布程度与图 7.4 和图 7.9 相比显得平均些。WLS 的结果优于 OLS 的结果。从表 7.7 我们看到, 当将 $\hat{\sigma}$ 或 R^2 作为拟合效果的指标时, WLS 的解对于原始数据的拟合不如 OLS 的好^①。这在意料之中, 因为 OLS 一个重要的性质是它提供了一个具有最小 $\hat{\sigma}$ 或等价地, 最大 R^2 的解。我们选择 WLS 解是因为残差的模式。对地域的标准化残差图 (比较图 7.9 和图 7.11) 显示出在处理异方差问题上 WLS 是成功的。

表 7.6 加权最小二乘法的权重 c_j

区域 j	n_j	$\hat{\sigma}_j^2$	c_j
西北	9	1451.11	1.110
中西部	12	2436.98	1.439
南部	16	249.43	0.460
西部	12	950.42	0.898

表 7.7 对于教育数据 ($n=49$), OLS 与 WLS 的系数估计, 剔除阿拉斯加

变量	OLS			WLS		
	系数	标准差	t	系数	标准差	t
常数	-277.577	132.40	-2.10	-315.531	78.15	-4.04
X_1	0.048	0.01	3.98	0.062	0.01	7.92
X_2	0.887	0.33	2.68	0.874	0.20	4.37
X_3	0.067	0.05	1.35	0.029	0.03	0.86
	$R^2 = 0.497$		$\hat{\sigma} = 35.81$	$R^2 = 0.477$		$\hat{\sigma} = 36.50$

① WLS 解的 $\hat{\sigma}$ 是

$$\hat{\sigma}^2 = \frac{1}{45} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

的平方根, 其中 $\hat{y}_i = -315.531 + 0.062x_{i1} + 0.874x_{i2} + 0.0297x_{i3}$ 是用 WLS 估计的系数计算出的拟合值, 权重为 c_j ; 权重在计算 $\hat{\sigma}$ 时没有更多的作用。

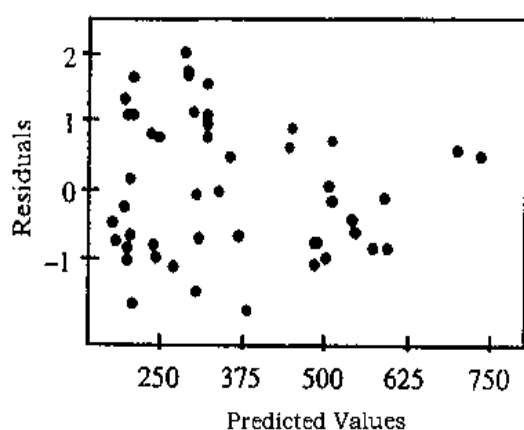


图 7.10 由 WLS 得到的对拟合值的标准化残差图

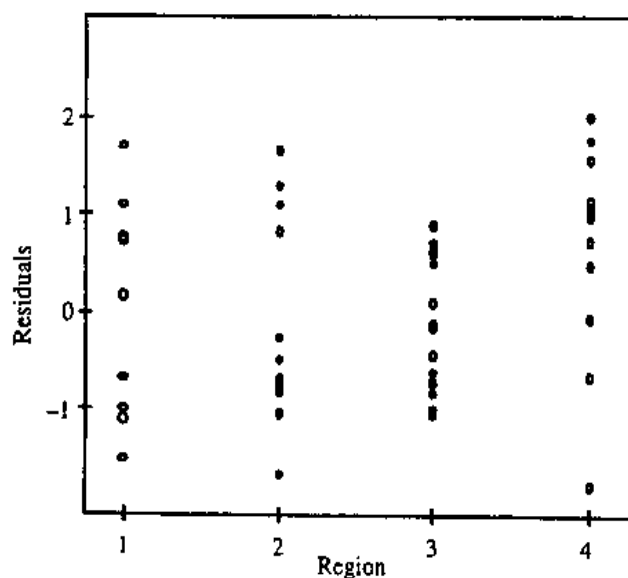


图 7.11 由 WLS 得到的对地域变量的标准化残差图

因为目前尚未得到关于求 WLS 解的两阶段方法的精确分布理论, 因此不可能作任何精确的显著性检验。如果权重事先知道而不是由数据估得, 那么, 基于 WLS 方法的统计检验将会是精确的。当然, 事先知道权重是困难的。但是, 前面的分析清楚地表明: 权重是必需。此外, 由于 Y 中被解释的变异小于 50% ($R^2 = 0.447$), 那么必须继续寻找其他的因素。建议读者引进代表四个地区的示性变量来分析这批数据。正如在第 5 章中指出的那样, 任何具有四个类别的模型只需要三个示性变量。异方差常常可以通过引入表示数据类别的示性变量消除。

7.5 拟合剂量 - 反应关系曲线

加权最小二乘分析的一个重要的应用领域是在响应变量 Y 为比例数据时 (值在 0 和 1 之间) 拟合线性回归。考虑下面的情形: 试验者可以将刺激控制在不同的水平。对各被试分别随机地指定一个刺激水平, 然后记录下每个被试的二值响应结果。根据这组观测, 可以构建刺激与对刺激的反应比例之间的关系模型。一个常用的例子是药理学领域中生物鉴定问题, 其中刺激水平可以代表药物或毒药的不同剂量, 二值响应是死亡或生存。另一个例子是消费行为的研究, 其中刺激是提供的折扣, 而二值响应是对一些商品的购买与否。

假定选择 k 个不同的剂量试用某种杀虫剂。设有 n_j 个昆虫接受第 j 个剂量水平 x_j , 令 r_j 为其中被杀虫剂杀死的个数 ($j = 1, 2, \dots, k$)。我们想估计剂量与死亡比例的关系。样本比例 $p_j = r_j/n_j$ 是二项分布的随机变量, 其均值为 π_j , 方差为 $\pi_j(1 - \pi_j)/n_j$, 其中 π_j 是一个接受剂量为 x_j 的被试的总体死亡概率。假定 π 与 X 的关系如下

$$\pi = f(X), \quad (7.10)$$

其中函数 $f(\cdot)$ 随着 X 增加而增加 (或至少非减) 并且介于 0 和 1 之间。此函数必须满足这些性质, 因为 (1) π 是介于 0 和 1 之间的概率, (2) 如果杀虫剂是有毒的, 高剂量将会减少被试存活的机会 (或增加死亡的机会)。这些考虑实际上排除了线性模型

$$\pi_j = \alpha + \beta x_j + \epsilon_j, \quad (7.11)$$

因为那样 π_j 将会是无界的。

刺激 - 反应之间的关系通常是非线性的。人们发现, 下列非线性函数可以准确地反映剂量 x_j 和死亡比例之间的关系

$$\pi_j = \frac{e^{\beta_0 + \beta_1 x_j}}{1 + e^{\beta_0 + \beta_1 x_j}}. \quad (7.12)$$

式 (7.12) 被称为 *logistic 响应函数* 且具有如图 7.12 所示的形状。可以看出 logistic 函数介于 0 和 1 之间, 且是单调的。基于临界值概念的物理方面的考虑为使用式 (7.12) 来表示刺激 - 反应关系提供了一个启发式的论据 (Cox, 1989)。

上面描述的机制与我们其他的例子有很大的不同。在目前的情况中, 试验者可以控制剂量或刺激, 且可以利用重复观测来估计响应变量在每个剂量水平上的变异。这是一个设计好的试验研究, 不像其他例子, 是观测性的或非试验性的。

这类分析的目的在于不仅仅是为确定剂量 - 反应之间关系的本质, 并且要估计引起特定反应水平的剂量。尤其感兴趣的是会导致总体的 50% 产生反应的剂量 (中位剂量)。

logistic 模型 (有时称为 *logit 模型*) 在生物学和流行病学的研究中已经得到广泛的应用。对于用二值响应数据来分析比例的问题, 这是很有吸引力的模型且易于拟合。

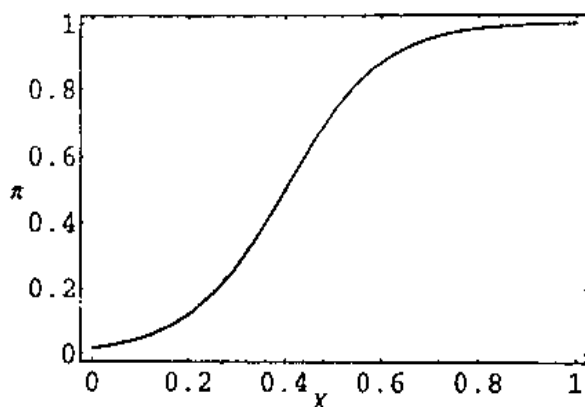


图 7.12 logistic 响应函数

另一个可以选用的模型是用正态概率分布的累积分布函数来表示响应函数。正态分布的累积曲线具有和 logistic 函数类似的形状。这个模型被称为 *probit* 模型，详细讨论建议读者参见 Finney(1964)。

除了医学和药理学之外，logistic 模型已经被应用于风险分析，学习理论，消费者行为（选择模型）研究和市场促销研究中。

由于图 7.12 的响应函数是非线性的，我们可以对变换后的变量进行研究。变换的选择为使响应函数线性化。但是，变换后的变量将会具有异方差。那么，我们必须采用加权最小二乘法去拟合变换后的数据。

我们将会用一整章的内容（第 12 章）来讨论 logistic 回归模型，因为此模型具有非常重要且多种多样的实际应用。关于 logistic 模型的适用性及拟合问题也将会在那里讨论。

习 题

- 7.1 采用表 5.12 中教育经费的数据重复在 7.4 节中的分析。
- 7.2 采用表 5.13 中教育经费的数据重复在 7.4 节中的分析。
- 7.3 对于表 7.3 中 Y 关于三个预测变量 X_1, X_2, X_3 的回归模型，计算杠杆值，标准化残差值，Cook's 距离以及 DFIT。对每个量度画一个适当的图。根据图形验证 Alaska 和 Utah 是高杠杆点，但只有 Alaska 是强影响点。
- 7.4 采用表 7.3 中的教育经费数据，拟合 Y 关于三个预测变量 X_1, X_2, X_3 以及表示地域的示性变量的线性回归模型。比较模型拟合的结果与在 7.4 节中得到的 WLS 结果。检验不同地域中回归关系的等同性。
- 7.5 用表 5.12 中的数据重复前面的练习。

8

相关误差的问题

8.1 引言：自相关

回归模型中的标准假定之一是：与第 i 个和第 j 个观测值对应的误差项 ε_i 和 ε_j 是不相关的。如果误差项之间存在相关性，则表明数据中还有其他的信息没有被目前的模型表达出来。当观测具有某种自然顺序时，其误差之间的相关性被称为自相关。

有多种原因可以导致自相关。例如在时间与空间上相邻的误差往往是相似的。经济时间序列中，连续的误差倾向于正相关。即若误差是较大的正值，则其紧接的误差也是正的；若误差是较大的负值，则其紧接的误差也是负的。从相邻的试验田地或区域中抽取的观测，其误差也倾向于有相关性，这是由于它们受类似外部环境的影响。

有时自相关症状的出现是由于回归方程右边有一个变量被忽略了而造成的。如果被忽略变量的相邻值是相关的，那么，估计模型中的误差就会呈现相关性。当方程中加入此变量时，明显的自相关问题就消失了。自相关的存在对于数据分析有以下几种影响：

1. 回归系数的最小二乘估计是无偏的，但它们不再具有最小方差意义上的有效性。
2. σ^2 和回归系数的标准差可能被严重低估；也就是说，从数据中估计出的标准差会比它们的实际值小得多，从而给出一个精确性的假象。
3. 置信区间和通常采用的各种显著性检验严格来说将不再合理。

由于上述原因，自相关性的存在性应是一个非常值得重视的问题，而不应该被忽视。

我们区分两种类型的自相关并描述处理它们的方法。第一种类型的自相关只是一种表象。这是由于忽略了应该包含在模型中的一个变量引起的。一旦找到这个变量，自相关问题就解决了。第二种类型的自相关称为纯自相关。校正纯自相关的方法涉及数据的变换。这些方法的正式推导可以在 Johnson(1984) 和 Kmenta(1986) 中找到。

8.2 消费者支出和货币存量

表 8.1 给出了从 1952 到 1956 年的消费者支出 (Y) 和货币存量 (X) 的季度数据, 都是以十亿美元计。这些数据也可以在本书的网站上找到^①。

货币定量理论采用的最简单模型为

$$y_t = \beta_0 + \beta_1 x_t + \varepsilon_t, \quad (8.1)$$

其中 β_0 和 β_1 是常数, ε_t 是误差项。经济学家对于估计 β_1 和它的标准误很有兴趣; β_1 称为乘数, 是财政和货币政策中的一个非常重要的工具。由于观测是以时间排序的, 因此有理由猜测自相关可能是存在的。回归结果被总结在表 8.2 中。

回归系数是显著的; 斜率系数的标准误是 0.115。货币供应量一个单位的变化, 导致的总的消费者支出变化的 95% 置信区间为 $2.30 \pm 2.10 \times 0.115 = (2.06, 2.54)$ 。 R^2 的值显示消费者支出中大约 96% 的变动能被货币存量的变异所解释。如果基本的回归假定是合理的, 这个数据分析就完成了。为了检验模型的假定, 我们考察其残差。如果残差存在自相关的迹象, 应该在消除自相关性以后再对模型进行估计。

表 8.1 消费者支出与货币存量

年份	季度	消费者支出	货币存量	年份	季度	消费者支出	货币存量
1952	1	214.6	159.3	1954	3	238.7	173.9
	2	217.7	161.2		4	243.2	176.1
	3	219.6	162.8	1955	1	249.4	178.0
	4	227.2	164.6		2	254.3	179.1
1953	1	230.9	165.9	1956	3	260.9	180.2
	2	233.3	167.9		4	263.3	181.2
	3	234.1	168.3	1956	1	265.6	181.6
	4	232.3	169.7		2	268.2	182.5
1954	1	233.7	170.5		3	270.4	183.3
	2	236.5	171.6		4	275.6	184.3

来源: Friedman and Meiselman (1963), p.266

表 8.2 消费者支出对货币存量 X 作回归时的结果

变量	系数	标准误	t-检验	p-值
常数	-154.72	19.850	-7.79	< 0.0001
X	2.30	0.115	20.08	< 0.0001
$n = 20$	$R^2 = 0.957$	$R_a^2 = 0.955$	$\hat{\sigma} = 3.983$	$d.f. = 18$

对时间序列数据而言, 一个非常有用的散点图是序列图 (标准化残差对时间的散点图)。见图 8.1。该图揭示了残差的模式并呈现出误差相关情形的特征, 即

① <http://www.ilr.cornell.edu/~hadi/RABF>

符号相同的残差成群成批地出现。这种模式的特征有几个连续的残差是正的,接着几个是负的,如此等等。从图 8.1 我们可以看到头 7 个残差是正的,接着 7 个负的,最后 6 个正的。这种模式表明模型中的误差项是相关的,因此需要进一步的分析。

这种视觉的印象可以通过计数残差符号图中的游程数而正式确认,这里残差按观测的顺序排列。这类图形称为顺序图。本例中,残差符号的顺序图为

++++++- - - - - ++++++

它表明有 3 个游程。当 n_1 个残差为正, n_2 个残差为负时,在随机性假定下,游程的期望数 μ 和它的方差 σ^2 为

$$\mu = \frac{2n_1n_2}{n_1 + n_2} + 1,$$

$$\sigma^2 = \frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}.$$

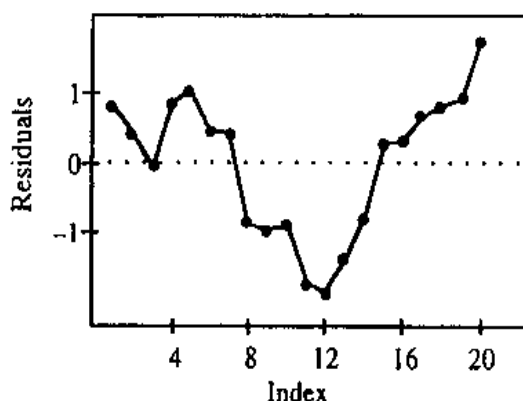


图 8.1 标准残差的序列图

本例中 $n_1 = 13, n_2 = 7$, 得出游程数的期望值为 8.1 以及标准差为 1.97。而观测到的游程数为 3, 与游程数的期望值之差为 5.1, 是标准差的 2 倍还多, 这表明残差序列显著地不随机。这种规范的游程检验方法只是确认了目测看出的结论: 残差中存在某种模式。

许多计算机软件包都提供游程检验。因此很容易实施这个近似的游程检验。但是我们所描述的游程检验不适用于较小的 n_1 和 n_2 (小于 10)。对于较小的 n_1 和 n_2 需要精确的概率表来判断显著性。游程检验的详细讨论, 读者可以参阅如 Lehmann (1975), Gibbons (1993) 和 Hollander 及 Wolfe (1999) 等非参数统计书。除了可由游程检验加以确认的图解分析方法之外, 自相关误差也可以通过 Durbin-Watson 统计量检验到。

8.3 Durbin-Watson 统计量

Durbin-Watson 统计量是回归分析中一种常用的自相关检验方法的基础。这个检验基于这样的假定: 相邻的误差是相关的, 即

$$\varepsilon_t = \rho\varepsilon_{t-1} + \omega_t, \quad |\rho| < 1, \quad (8.2)$$

其中 ρ 是 ε_t 与 ε_{t-1} 之间的相关系数, ω_t 独立地服从均值为 0 和方差相同的正态分布。在这种情形中, 误差被称为具有一阶自回归结构或一阶自相关。在大多数情况下, 误差 ε_t 可能具有更加复杂的相关结构。式 (8.2) 中给出的一阶相依结构, 可以认为是对实际误差结构的一个简单近似。

Durbin-Watson 统计量定义为

$$d = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2},$$

其中 e_t 是第 t 个普通最小二乘 (OLS) 残差。统计量 d 用于检验原假设 $H_0: \rho = 0$ 对备择假设 $H_1: \rho > 0$ 。注意到在 (8.2) 中当 $\rho = 0$ 时, ε 是不相关的。

由于 ρ 是未知的, 我们用 $\hat{\rho}$ 估计 ρ , 即

$$\hat{\rho} = \frac{\sum_{t=2}^n e_t e_{t-1}}{\sum_{t=1}^n e_t^2}. \quad (8.3)$$

d 和 $\hat{\rho}$ 之间的近似关系为

$$d \doteq 2(1 - \hat{\rho}),$$

(\doteq 表示近似相等) 这表明 d 的范围是 0 到 4。由于 $\hat{\rho}$ 是 ρ 的估计, 很清楚当 $\rho = 0$ 时, d 接近于 2, 而当 $\rho = 1$ 时, d 接近于 0。 d 的样本值越接近 2, 越有确凿的证据表明误差之间没有自相关性。 d 与 2 的偏离标示着自相关性的存在。对于正相关性的正式检验可以按如下步骤进行: 计算样本统计量 d 。然后, 如果

1. $d < d_L$, 拒绝 H_0 。
2. $d > d_U$, 不拒绝 H_0 。
3. $d_L < d < d_U$, 无法下结论。

对应于不同分位点的 (d_L, d_U) 的值已经由 Durbin 和 Watson (1951) 制成表格。这张表附在本书末的附录中 (表 A.6 和 A.7)。

负的自相关性的检验很少用。然而, 如果检验的话, 不是用 d , 而是用 $4 - d$, 检验的过程与检验正自相关性相同。

在我们的消费者支出和货币存量数据中, d 的值是 0.328。查表 A.6, 当 $n = 20, p = 1$ (预测变量的个数), 显著水平为 0.05 时, 我们有 $d_L = 1.20$ 及 $d_U = 1.41$ 。由于 $d < d_L$, 我们的结论是 d 在 5% 的水平上是显著的, 所以 H_0 被拒绝, 说明自相关性是存在的。这实质上再次证实我们以前的结论, 即通过观察残差序列图得到的结论。

如果 d 比 $d_U = 1.41$ 大, 自相关不再是个问题, 也就没有进一步分析的必要。当 $d_L < d < d_U$ 时, 对方程的进一步分析可做可不做。我们建议在 Durbin-Watson 统计量落在无法下结论的区域时, 采用下面描述的方法重新估计方程, 观察是否发生较大的变化。

如前面指出的那样, 相关性误差的存在使标准误, 置信区间以及统计检验失真, 因此我们应该重新估计方程。当表明误差具有自相关时, 可以采用两个方法: (1) 采用变换后的变量, (2) 引进具有时间顺序效应的其他变量。我们以货币存量数据为例说明第一种方法。第二种方法在 8.6 节中举例说明。

8.4 通过变换消除自相关

当残差图和 Durbin-Watson 统计量显示误差存在相关性时, 回归方程应该在考虑自相关性的情况下重新拟合。一个修正模型的方法是采用涉及未知自相关系数 ρ 的变换。 ρ 的引入使模型成为非线性。因此直接应用最小二乘法是不可能的。但是, 有许多方法可以解决这种非线性问题 (Johnston, 1984)。我们采用 Cochrane 和 Orcutt (1949) 的方法。

从模型 (8.1) 得到, ε_t 和 ε_{t-1} 可以被表达为

$$\varepsilon_t = y_t - \beta_0 - \beta_1 x_t,$$

$$\varepsilon_{t-1} = y_{t-1} - \beta_0 - \beta_1 x_{t-1}.$$

将以上两式代入式 (8.2) 式中, 我们得到

$$y_t - \beta_0 - \beta_1 x_t = \rho(y_{t-1} - \beta_0 - \beta_1 x_{t-1}) + \omega_t.$$

将上面方程中的各项重新安排, 我们得到

$$\begin{aligned} y_t - \rho y_{t-1} &= \beta_0(1 - \rho) + \beta_1(x_t - \rho x_{t-1}) + \omega_t, \\ y_t^* &= \beta_0^* + \beta_1^* x_t^* + \omega_t, \end{aligned} \quad (8.4)$$

其中

$$y_t^* = y_t - \rho y_{t-1},$$

$$x_t^* = x_t - \rho x_{t-1},$$

$$\beta_0^* = \beta_0(1 - \rho),$$

$$\beta_1^* = \beta_1.$$

由于诸 ω_t 是不相关的, 方程 (8.4) 为一个具有不相关误差的线性模型。将 y_t^* 作为响应变量以及 x_t^* 作为预测变量, 这就使得我们可以采用普通最小二乘法作回归。原方程中参数的估计为

$$\hat{\beta}_0 = \frac{\hat{\beta}_0^*}{1 - \hat{\rho}}, \quad \hat{\beta}_1 = \hat{\beta}_1^*. \quad (8.5)$$

因此,当模型(8.1)中的误差具有式(8.2)中给出的自回归结构时,我们可以对方程两边作变换,从而使得变换后的变量满足误差不相关的假定。

因为 ρ 是未知的,所以必须用数据来估计。Cochrane and Orcutt (1949) 提出一个迭代方法。操作步骤如下:

1. 用模型(8.1)拟合数据,计算 β_0 和 β_1 的OLS估计。
2. 计算残差并采用式(8.3)估计 ρ 。
3. 分别将 $y_t - \hat{\rho}y_{t-1}$ 和 $x_t - \hat{\rho}x_{t-1}$ 作为响应变量和预测变量,采用式(8.4)中的方程来拟合,并由(8.5)式得到 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 。
4. 考察新拟合方程的残差。如果新的残差仍然显示出自回归性,将 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 作为 β_0 和 β_1 的估计代替原始的最小二乘估计来重复整个过程。另一方面,如果新的残差显示没有自相关性,过程停止。对原始数据拟合的方程为

$$\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_t.$$

在实践中,我们建议如果首次应用 Cochrane-Orcutt 方法没有消除误差之间的自相关性,那么应该寻找另外的方法来处理。我们对表 8.1 中的数据应用 Cochrane-Orcutt 方法。

原始数据中 d 的值为0.328,是高度显著的。 $\hat{\rho}$ 为0.751。在对变量 $(y_t - 0.751y_{t-1})$ 和 $(x_t - 0.751x_{t-1})$ 拟合回归方程时,我们的 d 值为1.43。在5%的显著性水平上,当 $n = 19, p = 1$ 时, d_U 为1.40。因此,不能拒绝 $H_0: \rho = 0$ ^①。拟合的方程为

$$\hat{y}_t^* = -53.64 + 2.64x_t^*,$$

由(8.5)得到用原变量表示的方程为

$$\hat{y}_t = -215.4 + 2.64x_t.$$

斜率的标准误为0.31,而用最小二乘法得到的原始方程 $y_t = -154.7 + 2.3x_t$ 的斜率标准误为0.115。新估计的标准误是原来的近3倍。变量变换后拟合的方程的残差图见图8.2。残差图显示相邻残差同符号较少,可见 Cochrane-Orcutt 方法已经发挥了作用。

^① 检验的显著水平是不精确的,因为在估计过程中使用了 $\hat{\rho}$ 。 d 的值为1.43,与以前的0.328相比,表明自相关程度有了很大改善。

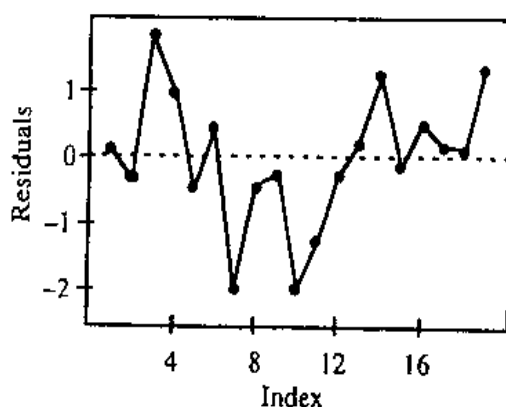


图 8.2 经过 Cochrane-Orcutt 方法迭代一次后的标准残化残差序列图

8.5 误差自相关时的迭代估计

Cochrane-Orcutt 方法的一个优点是可以通过标准最小二乘法得到参数估计。尽管需要两个步骤，但是方法相对简单。一个更直接的方法是同时估计 ρ, β_0, β_1 的值。模型的表述和前面一样需要构造变换变量 $y_t - \rho y_{t-1}$ 和 $x_t - \rho x_{t-1}$ 。参数估计通过最小化误差平方和得到，即使

$$S(\beta_0, \beta_1, \rho) = \sum_{t=2}^n [y_t - \rho y_{t-1} - \beta_0(1 - \rho) - \beta_1(x_t - \rho x_{t-1})]^2$$

达到最小。如果 ρ 已知， β_0 和 β_1 的值可以很容易地通过 $y_t - \rho y_{t-1}$ 对 $x_t - \rho x_{t-1}$ 作回归而得到。最终的估计可以通过在许多的 ρ 值中寻找，直到找到一个使得 $S(\rho, \beta_0, \beta_1)$ 达到最小的 ρ, β_0 和 β_1 的组合。搜寻过程可以用一个标准的回归计算程序来完成，但是采用自动搜寻方法会使整个过程更有效率。这个方法由 Hildreth 和 Lu(1960) 给出。关于估计方法和估计性质的讨论见 Kmenta(1986)。

表 8.3 回归估计的比较

方法	$\hat{\rho}$	$\hat{\beta}_0$	$\hat{\beta}_1$	$s.e.(\hat{\beta}_1)$
OLS	—	-154.700	2.300	0.115
Cochrane-Orcutt	0.874	-324.440	2.758	0.444
迭代法	0.824	-235.509	2.753	0.436

一旦得到使 $S(\rho, \beta_0, \beta_1)$ 达到最小的 $\hat{\rho}, \hat{\beta}_0, \hat{\beta}_1$ 时， β_1 的标准误可以用第 2 章中 (2.24) 来近似。使用该公式时就假定在 ρ 已知情况下对 $y_t - \rho y_{t-1}$ 关于 $x_t - \rho x_{t-1}$ 作回归；即 $\hat{\beta}_1$ 的标准误为

$$s.e.(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum [x_t - \hat{\rho}x_{t-1} - \bar{x}(1 - \hat{\rho})]^2}}$$

其中 $\hat{\sigma}$ 是 $S(\hat{\rho}, \hat{\beta}_0, \hat{\beta}_1)/(n-2)$ 的平方根。若有适当的计算工具可以轻易地完成迭代计算，那么建议采用后一种方法。但是，采用迭代方法和两阶段 Cochrane-

Orcutt 方法得到的估计和标准误一般不会有很大差别。对于表 8.1 中的数据, 采用 OLS, Cochrane-Orcutt 和迭代法三种方法得到的估计列在表 8.3 中, 以便比较。

8.6 自相关和变量的缺失

在揭示自相关性的同时, 回归残差的特性也可以显示模型设定其他方面的缺陷。在前面的例子中, 根据残差序列图和基于 Durbin-Watson 统计量的检验得出了误差为自相关的结论。而误差自相关只是残差图成群状或 Durbin-Watson 值较小的可能解释之一。

一般来说, 残差相对于任何一个潜在预测变量的散点图也许会揭示出可以用来解释响应变量变异的其他信息。当残差序列图的形状呈现出前面例子中描绘的情形时, 有理由怀疑这可能是由于遗漏了随时间变化的变量。当然, 当残差呈现出在 0 均值线上下交替聚集时, 当估计的自相关系数非常大及 Durbin-Watson 统计量显著时, 那么自相关的存在似乎得到了压倒多数的支持。但是, 我们将会看到这个结论也许是不正确的。也许将观测到的症状解释为模型形式误设的表现会更好。

应该考虑所有可能的修正方法。事实上, 在认为误差具有某种自回归结构之前, 最好全面地研究是否还有其他一些预测变量的可能性。如果增加一个变量能够解释模型误差自相关表象, 那将更令人满意, 也可能更有用。因为那样可以估计那个变量的边际效果并可用来提供信息。用变换去修正纯自相关可被视为是没有办法的办法。

8.7 住房开工分析

作为一个由于遗漏了一个预测变量而呈现自相关假象的例子, 考虑下面这个由中西部建筑行业协承建工程。协会想分析住房开工与人口增长之间的关系, 从而预测建筑业的发展。他们的做法是, 收集本区域住房开工的年度数据, 并寻找这些数据与潜在的住房购买人数之间的联系。因为精确地估算潜在的房屋购买人数几乎是不可能的, 研究人员只得拿区域内 22 至 44 岁的人群规模作为一个反映潜在房屋购买者的规模的变量。经过努力, 他们将本区域内 25 年的历史数据汇集起来 (见表 8.4)。表 8.4 中的数据也可以从本书的网站上得到。他们的目的是得到一个住房开工和人口规模之间的简单回归关系

$$H_t = \beta_0 + \beta_1 P_t + \varepsilon_t, \quad (8.6)$$

然后通过人口变化估计出对于新房子的需求。建筑协会也清楚人口与住房开工之间的关系会非常复杂。甚至有理由认为住房会影响人口增长 (通过移民) 而不是反过来。尽管这个模型毫无疑问是天真的, 但可以成为他们分析问题的一个起点。

分析

表 8.4 住房开工 (H), 人口规模 (P) 以百万计, 以及抵押资金可用性指数

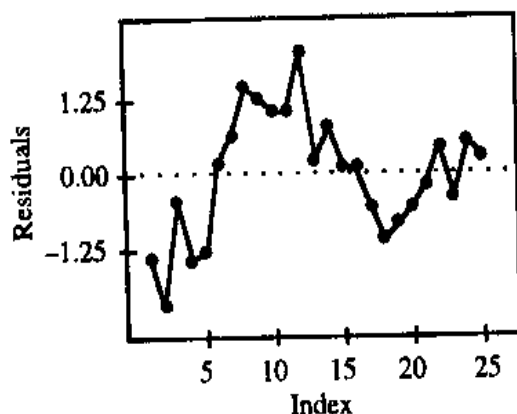
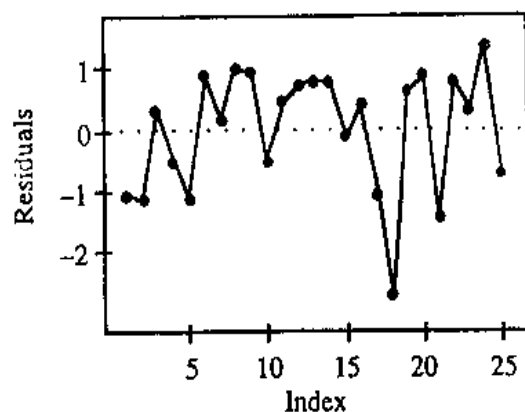
t	H	P	D
1	0.09090	2.200	0.03635
2	0.08942	2.222	0.03345
3	0.09755	2.244	0.03870
4	0.09550	2.267	0.03745
5	0.09678	2.280	0.04063
6	0.10327	2.289	0.04237
7	0.10513	2.289	0.04715
8	0.10840	2.290	0.04883
9	0.10822	2.299	0.04836
10	0.10741	2.300	0.05160
11	0.10751	2.300	0.04879
12	0.11429	2.340	0.05523
13	0.11048	2.386	0.04770
14	0.11604	2.433	0.05282
15	0.11688	2.482	0.05473
16	0.12044	2.532	0.05531
17	0.12125	2.580	0.05898
18	0.12080	2.605	0.06267
19	0.12368	2.631	0.05462
20	0.12679	2.658	0.05672
21	0.12996	2.684	0.06674
22	0.13445	2.711	0.06451
23	0.13325	2.738	0.06313
24	0.13863	2.766	0.06573
25	0.13964	2.793	0.07229

采用模型 (8.6) 对 25 年数据拟合的回归结果见表 8.5。 H 的变异中被 P 解释的比例是 $R^2 = 0.925$ 。我们也看到人口增长 1 百万会导致住房开工增加约 71000。Durbin-Watson 统计量和残差序列 (图 8.3) 揭示出误差之间存在强自相关性。但是, 可以推测有其他一些变量可以进一步解释住房开工, 并且它们可能是造成误差自相关的原因。这些变量包括失业率, 婚姻和家庭的社会趋势, 政府的房屋方案, 建筑和抵押资金的供给。第一选择是衡量本区域抵押资金供应情况的指标。将这个变量加入到方程中去, 模型变成

$$H_t = \beta_0 + \beta_1 P_t + \beta_2 D_t + \varepsilon_t.$$

表 8.5 住房开工规模 (H) 对人口规模 (P) 作回归的结果

变量	系数	标准误	t -检验	p -值
常数	-0.0609	0.0104	-5.85	< 0.0001
P	0.0714	0.0042	16.90	< 0.0001
$n = 25$	$R^2 = 0.925$	$d = 0.621$	$\hat{\sigma} = 0.0041$	$d.f. = 23$

图 8.3 对于住房开工数据, 作 H_t 对 P_t 的回归后得到的标准化残差序列图图 8.4 对于住房开工数据, 作 H_t 对 P_t 和 D_t 的回归后得到的标准化残差序列图

增加一个预测变量具有消除误差自相关的效果。从表 8.6 可以看出 Durbin-Watson 统计量的值为 1.852, 在接受域内。残差序列图 (图 8.4) 也有改进。回归系数和它们相应的 t -值表明存在一个显著的人口规模效果, 但是在第一个方程中它被夸大了近 2 倍。在某种意义上, 在固定的人口规模水平下抵押资金供应的变化比人口规模变化产生的影响更大。

如果回归方程中的每一个变量用其标准化形式代替 (变换后的变量均值为 0,

表 8.6 住房开工规模 (H) 对人口规模 (P) 及指标 (D) 作回归的结果

变量	系数	标准误	t -检验	p -值
常数	-0.0104	0.0103	-1.01	0.3220
P	0.0347	0.0064	5.39	< 0.0001
D	0.7605	0.1216	6.25	< 0.0001
$n = 25$	$R^2 = 0.973$	$d = 1.85$	$\hat{\sigma} = 0.0025$	$d.f. = 22$

方差为 1), 则产生的回归方程为

$$\tilde{H}_t = 0.4668\tilde{P}_t + 0.5413\tilde{D}_t,$$

其中 \tilde{H} 表示 H 的标准值, $\tilde{H} = (H - \bar{H})/s_H$ 。 \tilde{P}_t 增长一个标准化单位等价于 H_t 增长 0.4668 个标准化单位; 即, 如果人口增长一个标准差, 那么 H_t 增长 0.4668 个标准差。类似的, 如果 D_t 增加一个标准差, 那么 H_t 增加 0.5413 个标准差。因此, 模型用标准化变量表示后, 可以清楚地看出抵押资金指标比人口规模更重要 (更有影响)。

住房开工的例子说明了重要的两点。首先, 较大的 R^2 并不意味着数据已经被拟合、解释得很好了。任意一对呈现时间趋势的变量通常都是高度相关的。较大的 R^2 并不一定意味着两个变量之间的相互关系被充分刻画了。其次, Durbin-Watson 统计量以及残差图可能显示出误差之间存在自相关性, 但实际上误差是独立的, 只是遗漏了一个或多个变量导致了这一现象。尽管 Durbin-Watson 统计量被设计用来检测一阶自相关, 但是当模型的其他假定违背时, 例如错误确定包含在模型中的变量时, 它也会是显著的。一般来说, Durbin-Watson 统计量的显著值应该被解释为存在问题, 遗漏变量的可能性以及误差自相关性的存在都应该被考虑到。

8.8 Durbin-Watson 统计量的局限性

在前面消费者支出与货币存量和住房开工与人口规模的例子中, 从原回归方程得到的残差表明模型误设与时间相依性有关。在这两个情形中, Durbin-Watson 统计量小得足够得出存在正自相关的结论。残差序列图进一步证实了存在一个与时间相依的误差项。在这两个例子中我们对自相关性作了不同的处理。在一个例子中 (住房开工), 我们发现增加一个预测变量可以消除误差的自相关性, 在另外一个例子中 (货币存量), 我们用 Cochrane-Orcutt 方法处理所谓的纯自相关问题。两种情形下从残差中观测到的时间相依性都是一阶相关。Durbin-Watson 统计量和残差模式都表明相邻时间区间的误差存在相依性。如果时间相依性不是一阶的, 残差的散点图仍然有效。但是, Durbin-Watson 统计量却不能度量高阶时间相依性, 可能不会提供较多有价值的信息。

作为一个例子, 我们考虑美国一个生产和销售滑雪器械公司所做的, 以获得季度销售量和一个最主要的经济指标之间关系的研究。选择的指标是个人可支配

表 8.7 滑雪撬销售量对于 PDI

变量	系数	标准误	t-检验	p-值
常数	12.3921	2.539	4.9	< 0.0001
PDI	0.1979	0.016	12.4	< 0.0001
$n = 40$	$R^2 = 0.801$	$d = 1.968$	$\hat{\sigma} = 3.019$	$d.f. = 38$

收入 PDI, 以十亿美元计。初始的模型是

$$S_t = \beta_0 + \beta_1 \text{PDI}_t + \varepsilon_t,$$

其中 S_t 是在 t 时期以十亿美元计的滑雪撬销售量, PDI_t 是同期的个人可支配收入。可得到 10 年 (40 个季度) 的数据 (表 5.11)。此数据也可从本书的网站上得到。回归结果见表 8.7, 残差序列图见图 8.5。

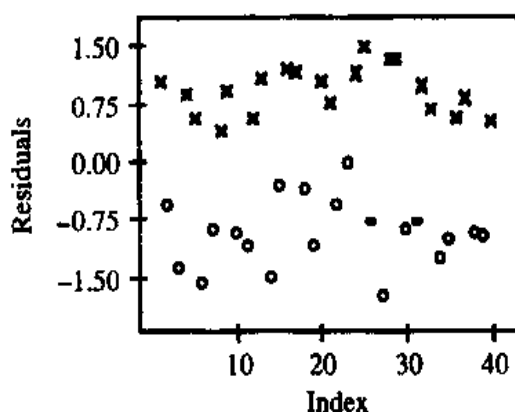


图 8.5 标准化残差序列图

(1 和 4 季度用叉表示, 2 和 3 季度用圆圈表示)

乍一看, 表 8.7 的结果令人鼓舞。销售量的变异被 PDI 解释的比例为 0.80。PDI 增加一个单位对销售量的边际贡献介于 \$165,420 和 \$230,380 之间 ($\hat{\beta}_1 = 0.1979$), 置信系数为 95%。另外, Durbin-Watson 统计量为 1.968, 表示没有一阶自相关性。

PDI 可以解释销售变量变异的大部分是在意料之中, 因为两个变量都是随时间而增加的。因此, 尽管 R^2 的值 0.80 是很好了, 但是不能作为这个模型的最终评价。而且, 尽管 Durbin-Watson 统计量在一个可以接受的范围内, 但是从图 8.5 可以清楚地看到残差存在某种时间相依性。我们注意到, 对所有的年度第一和第四季度的残差是正的, 而第二和第三季度的残差是负的。由于滑雪活动受天气条件的影响, 我们怀疑一个季节效应被忽视了。残差的模式表明有两个与滑雪撬销售量有关的季节: 第二和第三季度, 对应温暖的季节, 第四和第一季度对应的是寒冷季节, 正是滑雪旺季。该季节效应可以简单地用一个对寒冷季节取 1, 对温暖季节取 0 的示性 (哑) 变量来刻画 (见第 5 章), 扩充的数据集见表 8.8 中并可从本书的网站上得到。

表 8.8 1964-1973 年可支配收入、滑雪橇销售量, 以及季节变量数据

季度	销售量	PDI	季节
Q1/64	37.0	109	1
Q2/64	33.5	115	0
Q3/64	30.8	113	0
Q4/64	37.9	116	1
Q1/65	37.4	118	1
Q2/65	31.6	120	0
Q3/65	34.0	122	0
Q4/65	38.1	124	1
Q1/66	40.0	126	1
Q2/66	35.0	128	0
Q3/66	34.9	130	0
Q4/66	40.2	132	1
Q1/67	41.9	133	1
Q2/67	34.7	135	0
Q3/67	38.8	138	0
Q4/67	43.7	140	1
Q1/68	44.2	143	1
Q2/68	40.4	147	0
Q3/68	38.4	148	0
Q4/68	45.4	151	1
Q1/69	44.9	153	1
Q2/69	41.6	156	0
Q3/69	44.0	160	0
Q4/69	48.1	163	1
Q1/70	49.7	166	1
Q2/70	43.9	171	0
Q3/70	41.6	174	0
Q4/70	51.0	175	1
Q1/71	52.0	180	1
Q2/71	46.2	184	0
Q3/71	47.1	187	0
Q4/71	52.7	189	1
Q1/72	52.2	191	1
Q2/72	47.0	193	0
Q3/72	47.8	194	0
Q4/72	52.8	196	1
Q1/73	54.1	199	1
Q2/73	49.5	201	0
Q3/73	49.5	202	0
Q4/73	54.3	204	1

8.9 采用示性变量消除季节效应

增加季节变量后, 模型被扩展为

$$S_t = \beta_0 + \beta_1 \text{PDI}_t + \beta_2 Z_t + \varepsilon_t, \quad (8.7)$$

其中 Z_t 为上面描述的 0-1 变量, β_2 是衡量季节效应的参数。注意到模型 (8.7) 可以用两个模型表示 (一个对应寒冷季节, $Z_t = 1$; 另一个对应温暖季节, $Z_t = 0$):

$$\text{冬季: } S_t = (\beta_0 + \beta_2) + \beta_1 \text{PDI}_t + \varepsilon_t,$$

$$\text{夏季: } S_t = \beta_0 + \beta_1 \text{PDI}_t + \varepsilon_t.$$

因此, 这个模型假定是销售量可以用 PDI 的线性函数来近似, 对冬季是一条直线, 对夏季是另一条直线。两条线是平行的; 即 PDI 变化的边际效应对两个季节是一样的。截距反映了销售水平对不同的季节是不同的 (见图 8.6)。

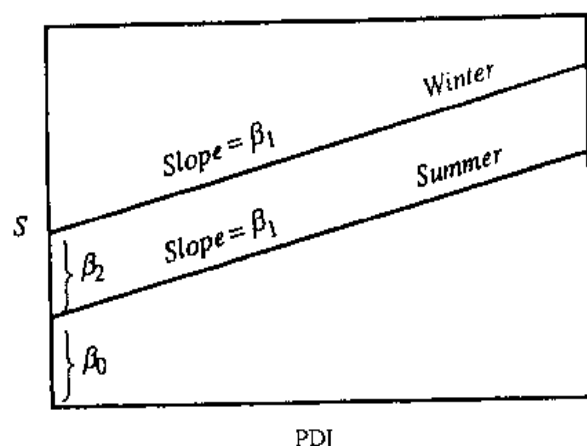


图 8.6 用季节修正后的滑雪橇销售量关于 PDI 的模型

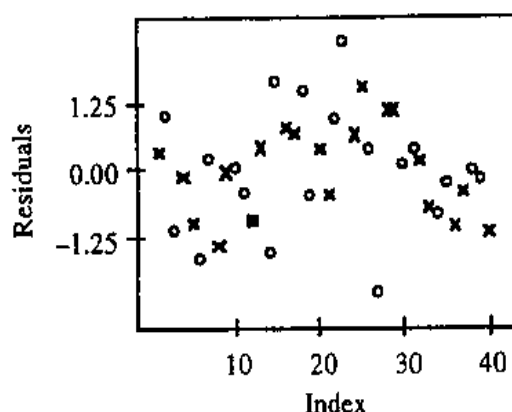


图 8.7 含季节变量回归方程的标准化残差序列图

(1 和 4 季度用叉表示, 2 和 3 季度用圆圈表示)

回归结果总结在表 8.9 中, 标准化残差序列图见图 8.7。我们看到所有季节模式的迹象都已经被消除了。而且, PDI 边际效应的估计的精确度提高了。现在的置信区间为 \$186,520 到 \$210,880。另外, 季节效应被量化了, 我们可以说, 对于给定的 PDI 水平, 冬季将会带来比夏季高出 \$4,734,109 到 \$6,194,491 的销售量 (95% 置信度)。

滑雪橇数据说明了有关自相关的重要的两点。首先, Durbin-Watson 统计量只对发生于相邻观测之间的相关误差是敏感的 (一阶自相关)。对于滑雪橇数据, 一阶相关系数是 -0.001, 二阶, 四阶, 六阶和八阶相关系数分别是 -0.81, 0.76,

表 8.9 滑雪橇销售量关于 PDI 及季节变量的回归

变量	系数	标准误	t-检验	p-值
常数	9.5402	0.9748	9.79	0.3220
PDI	0.1987	0.0060	32.90	< 0.0001
Z	5.4643	0.3597	15.20	< 0.0001
$n = 40$	$R^2 = 0.972$	$d = 1.772$	$\hat{\sigma} = 1.137$	$d.f. = 37$

-0.71 和 0.73。对这种情形, Durbin-Watson 检验并不显著。有另外的检验可以应用于高阶自相关性的检测(见 Box and Pierce(1970))。但是,在所有的情形中,如果误差项存在时间相依性,残差图将会反映出来。

其次,当自相关性被揭示出以后,应该重新拟合模型。误差自相关的出现通常是由于模型忽略了一个与时间相依的变量。将遗漏的变量包含进来就会消除观测到的自相关性。然而,有时候不存在这样的变量。那就不得不对原始变量实施差分变换以消除自相关性。

如果观测不是按时间排序的,那么严格讲,使用 Durbin-Watson 统计量不是很恰当。但是,这个统计量仍然是个有用的诊断工具。如果数据是按照另外的标准排序的,例如字母顺序, Durbin-Watson 统计量将会靠近 2.0。小的值值得怀疑,应该仔细检查。

许多数据集是按照与研究有关的顺序排列的。城市或公司也许是按规模排序。Durbin-Watson 统计量的值较小表示存在显著规模效应。因此,度量规模的变量应该作为一个预测变量放入模型。在这些情况下,差分法或 Cochrane-Orcutt 差分法将是不适合的。

8.10 时间序列间的回归

本章分析的数据集有个共同的特征,他们都是时间序列数据(即观测出现于连续的时期)。这和前面几章研究的数据集不同(除了第 6 章的细菌数据),那里所有的观测是在同一个时点上产生的。这些例子中的观测是同时期的,从而产生截面数据。当观测是同时产生的(与单个时期有关),我们就有截面数据。时间序列和截面数据集的区别可以通过比较本章中讨论的滑雪橇销售数据(数据按时间顺序产生),以及 3.3 节中与一个时间点有关的主管人员业绩数据看出。

一个时间序列相对于另一个时间序列的回归分析在经济、贸易、公共健康和其他社会科学领域有广泛的应用。时间序列数据具有许多截面数据没有的特性。我们提醒大家注意这些特性,并建议一些处理它们的技术。

在截面数据中,自相关的概念是无关紧要的。观测的排序经常是随意的。因此,相邻误差的相关性是人为造成的。然而,对于时间序列数据,自相关通常是一个重要的因素。自相关的存在表示数据存在着还没有发现的隐含结构(通常是与时间有关的)。另外,绝大多数时间序列数据表现出季节性,研究者应该寻找季

节性模式。残差中(如滑雪橇数据)有规律的时间模式通常表示存在季节性。对于季度或月度数据,像已经指出的那样,引入示性变量是一个令人满意的解决方法。对于季度数据,需要4个示性变量但在分析中只用3个(见第4章的讨论)。对于月度数据,我们会需要12个示性变量但为了避免共线性只用11个(见第5章的讨论)。并不是所有的示性变量都是显著的,他们中的一些将会在最后的分析阶段被剔除。

为了找到 y_t 与 $x_{1t}, x_{2t}, \dots, x_{pt}$ 之间的关系,可以把预测变量的滞后值也包含进来以扩大预测变量集。如模型

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{1t-1} + \beta_3 x_{2t} + \varepsilon_t$$

在分析时间序列数据时是有意义的,但是对于截面数据是没有意义的。上面给出的模型隐含着在一个给定期间的 Y 的取值不仅受当期的 X_1 和 X_2 的影响而且受 X_1 前一期值的影响(即 X_1 对于 Y 有一期滞后的影响)。滞后多期的变量也是有可能的,也可以包含在预测变量集里。

时间序列数据也有可能包含一定的趋势。分析可能有时间趋势的数据时通常包含时间(t)的直接函数作为预测变量。例如在预测变量中包含 t, t^2 等变量。它们用于解释可能的线性或二次趋势。简单的一阶差分($y_t - y_{t-1}$),或更复杂的如 Cochrane-Orcutt 方法中的滞后型($y_t - \alpha y_{t-1}$)也是可能的。对于其全面的讨论,读者可以参阅时间序列分析方面的书籍,例如 Shumway(1988), Hamilton(1994)。

总的来说,对于时间序列间的回归分析必须非常小心数据中经常出现的自相关性和季节效应,也应该研究采用滞后预测变量的可能性。

习 题

8.1 用模型(8.6)拟合表8.4中的数据。

(a) 计算 Durbin-Watson 统计量 d 。从 d 中你可得出关于自相关性存在的什么结论?

(b) 当采用模型(8.6)拟合表8.4中的数据时,比较游程数与它们的期望值和标准差。从这个比较中你可以得出有关自相关性的什么结论?

8.2 石油产量数据:参见表6.17的石油产量数据。采用 $\log(\text{OIL})$ 对年份作线性回归拟合后得到的残差序列图显示出明显的周期模式。

(a) 计算 Durbin-Watson 统计量 d 。从 d 中你可得出关于自相关性存在的什么结论?

(b) 比较游程数与它们的期望值和标准差。从这个比较中你可以得出有关自相关性存在的什么结论?

8.3 参见表5.17的总统选举数据。由于数据是按时间顺序获取的(1916-1996年间的选举年),在用(5.11)中的模型拟合数据时,也许会怀疑存在自相关问题。

(a) 你同意吗?请作解释。

(b) 增加时间趋势作为另外一个预测变量将会改善还是加剧自相关? 解释原因。

8.4 道琼斯工业平均指数 (DJIA): 表 8.10 和表 8.11 包含 1996 年所有交易日每天的 DJIA。此数据可以在本书的网页上找到。DJIA 是非常著名的金融指数, 反映纽约股票交易所股票价格的水平。该指数由 30 种股票组成。变量 Day 表示这一年的交易日。1996 年有 262 个交易日, 所以变量 Day 从 1 到 262。

- 采用 1996 年所有 262 个交易日的数据, 拟合一个 DJIA 关于变量 Day 的线性回归模型。线性趋势的模型是否充分? 检查残差的时间相依性。
- 将 $DJIA_{(t)}$ 对 $DJIA_{(t-1)}$ 作回归, 即对 DJIA 作关于它自己滞后一个时期的回归。这个模型是否充分? 残差中是否有自相关的迹象?
- 由于每日的 DJIA 的变化 (波动性) 较大, 为了调节这种现象, 将对 DJIA 的对数展开分析。采用 $\log(DJIA)$ 而不是 DJIA 重复上面的练习。你的结论是相似的吗? 你是否发现有些不同?

8.5 再次参见练习 8.4 中数据。

- 采用练习 8.4 中你发现的充分的模型重新拟合模型, 但是只用 1996 年前 6 个月的数据 (表 8.10 中 130 天的数据)。计算残差均方。
- 采用上面的模型去预测 1996 年 7 月前 15 个交易日的日 DJIA (表 8.11)。比较你的预测值与表 8.11 中实际的 DJIA 值, 从而计算预测误差, 即 1996 年 7 月前 15 天的 DJIA 实际值与模型给出的相应预测值之间的差异。
- 计算预测误差平方的平均值并与残差均方作比较。
- 重复前面的练习, 但是采用模型去预测这年后半年的每日 DJIA (132 天)。
- 采用 DJIA 相对于 Day 的散点图解释你前面得到的结论。

8.6 继续对练习 8.4 和 8.5 中的数据建模。所谓的股票价格随机游动模型的一个简单的表述为第 t 天股票价格指数的最优估计是其在第 $t-1$ 天的值。在回归模型中这意味着对于练习 8.4 和练习 8.5 中拟合的模型, 常数项为 0, 回归参数为 1。

- 实施适当的显著性统计检验。(分别检验系数的值, 然后再同时检验。) 哪个检验是合适的: 单独的或同时的?
- 随机游动理论意味着指数的一阶差分 (相连值的差) 应该具有独立正态分布, 均值为 0, 方差为常数。检验 DJIA 和 $\log(DJIA)$ 的一阶差分, 看这假设是否成立?
- DJIA 很容易获得。收集最近的数据, 看看从 1996 年数据得到的结论对于最近的时期是否成立。

表 8.10 1996 年上半年的 DJIA 数据

Day	日期	DJIA	Day	日期	DJIA	Day	日期	DJIA
1	1/1/96	5117.12	45	3/1/96	5536.56	89	5/2/96	5498.27
2	1/2/96	5177.45	46	3/4/96	5600.15	90	5/3/96	5478.03
3	1/3/96	5194.07	47	3/5/96	5642.42	91	5/6/96	5464.31
4	1/4/96	5173.84	48	3/6/96	5629.77	92	5/7/96	5420.95
5	1/5/96	5181.43	49	3/7/96	5641.69	93	5/8/96	5474.06
6	1/8/96	5197.68	50	3/8/96	5470.45	94	5/9/96	5475.14
7	1/9/96	5130.13	51	3/11/96	5581.00	95	5/10/96	5518.14
8	1/10/96	5032.94	52	3/12/96	5583.89	96	5/13/96	5582.60
9	1/11/96	5065.10	53	3/13/96	5568.72	97	5/14/96	5624.71
10	1/12/96	5061.12	54	3/14/96	5586.06	98	5/15/96	5625.44
11	1/15/96	5043.78	55	3/15/96	5584.97	99	5/16/96	5635.05
12	1/16/96	5088.22	56	3/18/96	5683.60	100	5/17/96	5687.50
13	1/17/96	5066.90	57	3/19/96	5669.51	101	5/20/96	5748.82
14	1/18/96	5124.35	58	3/20/96	5655.42	102	5/21/96	5736.26
15	1/19/96	5184.68	59	3/21/96	5626.88	103	5/22/96	5778.00
16	1/22/96	5219.36	60	3/22/96	5636.64	104	5/23/96	5762.12
17	1/23/96	5192.27	61	3/25/96	5643.86	105	5/24/96	5762.86
18	1/24/96	5242.84	62	3/26/96	5670.60	106	5/27/96	5762.86
19	1/25/96	5216.83	63	3/27/96	5626.88	107	5/28/96	5709.67
20	1/26/96	5271.75	64	3/28/96	5630.85	108	5/29/96	5673.83
21	1/29/96	5304.98	65	3/29/96	5587.14	109	5/30/96	5693.41
22	1/30/96	5381.21	66	4/1/96	5637.72	110	5/31/96	5643.18
23	1/31/96	5395.30	67	4/2/96	5671.68	111	6/3/96	5624.71
24	2/1/96	5405.06	68	4/3/96	5689.74	112	6/4/96	5665.71
25	2/2/96	5373.99	69	4/4/96	5682.88	113	6/5/96	5697.48
26	2/5/96	5407.59	70	4/5/96	5682.88	114	6/6/96	5667.19
27	2/6/96	5459.61	71	4/8/96	5594.37	115	6/7/96	5697.11
28	2/7/96	5492.12	72	4/9/96	5560.41	116	6/10/96	5687.87
29	2/8/96	5539.45	73	4/10/96	5485.98	117	6/11/96	5668.66
30	2/9/96	5541.62	74	4/11/96	5487.07	118	6/12/96	5668.29
31	2/12/96	5600.15	75	4/12/96	5532.59	119	6/13/96	5657.95
32	2/13/96	5601.23	76	4/15/96	5592.92	120	6/14/96	5649.45
33	2/14/96	5579.55	77	4/16/96	5620.02	121	6/17/96	5652.78
34	2/15/96	5551.37	78	4/17/96	5549.93	122	6/18/96	5628.03
35	2/16/96	5503.32	79	4/18/96	5551.74	123	6/19/96	5648.35
36	2/19/96	5503.32	80	4/19/96	5535.48	124	6/20/96	5659.43
37	2/20/96	5458.53	81	4/22/96	5564.74	125	6/21/96	5705.23
38	2/21/96	5515.97	82	4/23/96	5588.59	126	6/24/96	5717.79
39	2/22/96	5608.64	83	4/24/96	5553.90	127	6/25/96	5719.27
40	2/23/96	5630.49	84	4/25/96	5566.91	128	6/26/96	5682.70
41	2/26/96	5565.10	85	4/26/96	5567.99	129	6/27/96	5677.53
42	2/27/96	5549.21	86	4/29/96	5573.41	130	6/28/96	5654.63
43	2/28/96	5506.21	87	4/30/96	5569.08			
44	2/29/96	5485.62	88	5/1/96	5575.22			

表 8.11 1996 年下半年的 DJIA 数据

Day	日期	DJIA	Day	日期	DJIA	Day	日期	DJIA
131	7/1/96	5729.98	175	8/30/96	5616.21	219	10/31/96	6029.38
132	7/2/96	5720.38	176	9/2/96	5616.21	220	11/1/96	6021.93
133	7/3/96	5703.02	177	9/3/96	5648.39	221	11/4/96	6041.68
134	7/4/96	5703.02	178	9/4/96	5656.90	222	11/5/96	6081.18
135	7/5/96	5588.14	179	9/5/96	5606.96	223	11/6/96	6177.71
136	7/8/96	5550.83	180	9/6/96	5659.86	224	11/7/96	6206.04
137	7/9/96	5581.86	181	9/9/96	5733.84	225	11/8/96	6219.82
138	7/10/96	5603.65	182	9/10/96	5727.18	226	11/11/96	6255.60
139	7/11/96	5520.50	183	9/11/96	5754.92	227	11/12/96	6266.04
140	7/12/96	5510.56	184	9/12/96	5771.94	228	11/13/96	6274.24
141	7/15/96	5349.51	185	9/13/96	5838.52	229	11/14/96	6313.00
142	7/16/96	5358.76	186	9/16/96	5889.20	230	11/15/96	6348.03
143	7/17/96	5376.88	187	9/17/96	5888.83	231	11/10/96	6840.01
144	7/18/96	5464.18	188	9/18/96	5877.36	232	11/19/96	6397.60
145	7/19/96	5426.82	189	9/19/96	5867.74	233	11/20/96	6430.02
146	7/22/96	5390.94	190	9/20/96	5888.46	234	11/21/96	6418.47
147	7/23/96	5346.55	191	9/23/96	5894.74	235	11/22/96	6471.76
148	7/24/96	5354.69	192	9/24/96	5874.03	236	11/25/96	6547.79
149	7/25/96	5422.01	193	9/25/96	5877.36	237	11/26/96	6528.41
150	7/26/96	5473.06	194	9/26/96	5868.85	238	11/27/96	6499.34
151	7/29/96	5434.59	195	9/27/96	5872.92	239	11/28/96	6499.34
152	7/30/96	5481.93	196	9/30/96	5882.17	240	11/29/96	6521.70
153	7/31/96	5528.91	197	10/1/96	5904.90	241	12/2/96	6521.70
154	8/1/96	5594.75	198	10/2/96	5933.79	242	12/3/96	6442.69
155	8/2/96	5679.83	199	10/3/96	5932.85	243	12/4/96	6422.94
156	8/5/96	5674.28	200	10/4/96	5992.86	244	12/5/96	6437.10
157	8/6/96	5696.11	201	10/7/96	5979.81	245	12/6/96	6381.94
158	8/7/96	5718.67	202	10/8/96	5966.77	246	12/9/96	6463.94
159	8/8/96	5713.49	203	10/9/96	5930.62	247	12/10/96	6473.25
160	8/9/96	5681.31	204	10/10/96	5921.67	248	12/11/96	6402.52
161	8/12/96	5704.98	205	10/11/96	5969.38	249	12/12/96	6303.71
162	8/13/96	5647.28	206	10/14/96	6010.00	250	12/13/96	6304.87
163	8/14/96	5666.88	207	10/15/96	6004.78	251	12/16/96	6268.35
164	8/15/96	5665.78	208	10/16/96	6020.81	252	12/17/96	6308.33
165	8/16/96	5689.45	209	10/17/96	6059.20	253	12/18/96	6346.77
166	8/19/96	5699.44	210	10/18/96	6094.23	254	12/19/96	6473.64
167	8/20/96	5721.26	211	10/21/96	6090.87	255	12/20/96	6484.40
168	8/21/96	5689.82	212	10/22/96	6061.80	256	12/23/96	6489.02
169	8/22/96	5733.47	213	10/23/96	6036.46	257	12/24/96	6522.85
170	8/23/96	5722.74	214	10/24/96	5992.48	258	12/25/96	6522.85
171	8/26/96	5693.89	215	10/25/96	6007.02	259	12/26/96	6546.68
172	8/27/96	5711.27	216	10/28/96	5972.73	260	12/27/96	6560.91
173	8/28/96	5712.38	217	10/29/96	6007.02	261	12/30/96	6549.37
174	8/29/96	5647.65	218	10/30/96	5993.23	262	12/31/96	6448.27

9

共线性数据的分析

9.1 引言

多元回归方程的解释，隐含地依赖于预测变量间无强相关性的假定。通常将一个回归系数解释为，当相应的预测变量增加一个单位、且所有其他预测变量保持不变时，响应变量的变化量。但如果预测变量间存在很强的线性关系，那这种解释就不一定合适。虽然，增加回归方程中某个变量的值，同时使其他变量保持恒定，这在概念上总是可能的。但是，在估计数据中可能没有任何关于这种操作结果的信息。而且，在被研究的那个过程中，变动一个变量而保持其他变量不变，这也可能做不到。在这些情况下，将回归系数简单地解释为边际效应是不合适的。

若预测变量间完全不存在线性关系，则称它们为正交的[†]。在绝大多数回归应用中，预测变量是不正交的。通常正交性的欠缺程度还不至于严重到影响分析。但是，在有些情况下，预测变量间的相关性非常强，以致于回归结果都模糊不清了。典型的情况是，无法估计回归方程中单个变量独自的效应。系数的估计值对于数据的微小变动、对于在方程中增、删变量非常敏感。回归系数的抽样误差很大，对推断及由此回归模型所作的预测都有影响。

严重的非正交性情况也称作共线性数据问题，或多重共线性问题。该问题可能很难察觉。它不是一种可以通过考察回归残差来揭示的模型设定错误。事实上，多重共线性也并非模型设定错误，而是一种数据有缺陷的状况。不管怎样，弄清楚多重共线性何时出现、并了解其可能的后果，这是很重要的。我们建议，在多重共线性出现时，应当非常谨慎地对待由回归分析得到的所有结论。

本章着重讨论三个问题：

- 多重共线性如何影响统计推断和预测？
- 如何检测多重共线性？

[†] 译注：本章及第 10 章中，原文将“正交”解释为“完全不存在线性关系”，这不准确，而且对“正交”这概念的使用也不恰当。因涉及多处，不便改动，故我们仍照原文直译。据译者的理解，这两章中所谓的“变量正交”实际指的是“变量不相关”。若变量都已经中心化，那么“正交”与“不相关”是同一回事。

• 怎么解决多重共线性带来的困难?

在分析数据时,不可能分开来回答这些问题。若多重共线性有可能出现,那必须同时处理这三个问题。

我们从两个例子出发展开讨论,它们分别选来说明多重共线性对于推断与预测的影响。接着讨论检测多重共线性的几种方法。在结束本章时,我们罗列了一系列解决多重共线性问题的办法。一个显而易见的处方是收集更好的数据,我们对此作了考虑,但主要还是讨论如何改进对已有数据的解释。第10章中讨论了在存在多重共线性时不同于普通最小二乘估计的一些有效的方法。

9.2 对推断的影响

第一个例子说明,欲从一组线性相依的预测变量中挑出重要的预测变量,结果可能是模棱两可的。本例的背景是公共教育中机会均等情况的研究,如 Coleman et al.(1966), Mosteller and Moynihan (1972) 等的报告。

配合 1964 的《民权法》,美国国会下令进行一项“关于在公共教育机构中由于种族、肤色、宗教或民族血统等原因造成的、提供给个人的教育机会的不均等情况”的调查。数据从全国若干个有代表性的学区中收集而来。除了报告诸如学生学业成就水平、学校设施等变量的综合统计量外,还试图用回归分析来找到决定学业成就最重要的因子。本例的数据由 1965 年随机选取的 70 个学校的测量值组成,其中包括衡量学生学业成就的、学校设施的及教师资历的变量。目的是评价学校投入对于学业成就的效应。

假定现已形成了一个可以接受的、用于衡量学校环境情况的指数,猜测该指数对学业成就有影响。该指数包括对设施、教学资料、特定的教学方案、教师的培训与激励机制等等方面的综合评价。学业成就可用一个由各项标准化测试的得分构造的指数来衡量。之外,还有一些变量可能影响学校投入与学业成就之间的关系。学生的表现可能受到家庭环境及学校中同伴的影响。在分析中,必须重视这些变量,才能评价学校投入的效应。我们假定已对这些变量构造了一些符合我们要求的指数。数据在表 9.1 和 9.2 中给出,也可从本书的网站^①上找到。

可用下列回归模型来修正那两个基本变量(学业成就与学校)之间的关系

$$ACHV = \beta_0 + \beta_1 \cdot FAM + \beta_2 \cdot PEER + \beta_3 \cdot SCHOOL + \varepsilon. \quad (9.1)$$

可由 β_3 的 t -值来检验学校变量的贡献。回想一下, β_3 的 t -值检验的是,在变量 FAM 与 PEER 被纳入方程之后,变量 SCHOOL 的必要性。实际上,可拿上述模型与

$$ACHV - \beta_1 \cdot FAM - \beta_2 \cdot PEER = \beta_0 + \beta_3 \cdot SCHOOL + \varepsilon \quad (9.2)$$

相比较,即在用 FAM 与 PEER 修正之后再去评价学校变量的贡献。将方程 (9.1) 的形式改为

$$ACHV - \beta_1 \cdot FAM - \beta_2 \cdot PEER = \beta_0 + \beta_3 \cdot SCHOOL + \varepsilon,$$

^① <http://www.ilr.cornell.edu/~hadi/RABE>

表 9.1 均等教育机会 (EEO) 数据的前 50 个观测; 各标准化的指数

行	ACHV	FAM	PEER	SCHOOL
1	-0.43148	0.60814	0.03509	0.16607
2	0.79969	0.79369	0.47924	0.53356
3	-0.92467	-0.82630	-0.61951	-0.78635
4	-2.19081	-1.25310	-1.21675	-1.04076
5	-2.84818	0.17399	-0.18517	0.14229
6	-0.66233	0.20246	0.12764	0.27311
7	2.63674	0.24184	-0.09022	0.04967
8	2.35847	0.59421	0.21750	0.51876
9	-0.91305	0.61561	-0.48971	-0.63219
10	0.59445	0.99391	0.62228	0.93368
11	1.21073	1.21721	1.00627	1.17381
12	1.87164	0.41436	0.71103	0.58978
13	-0.10178	0.83782	0.74281	0.72154
14	-2.87949	-0.75512	-0.64411	-0.56986
15	3.92590	-0.37407	-0.13787	-0.21770
16	4.35084	1.40353	1.14085	1.37147
17	1.57922	1.64194	1.29229	1.40269
18	3.95689	-0.31304	-0.07980	-0.21455
19	1.09275	1.28525	1.22441	1.20428
20	-0.62389	-1.51938	-1.27565	-1.36598
21	-0.63654	-0.38224	-0.05353	-0.35560
22	-2.02659	-0.19186	-0.42605	-0.53718
23	-1.46692	1.27649	0.81427	0.91967
24	3.15078	0.52310	0.30720	0.47231
25	-2.18938	-1.59810	-1.01572	-1.48315
26	1.91715	0.77914	0.87771	0.76496
27	-2.71428	-1.04745	0.77536	-0.91397
28	-6.59852	-1.63217	-1.47709	-1.71347
29	0.65101	0.44328	0.60956	0.32833
30	-0.13772	-0.24972	0.07876	-0.17216
31	-2.43959	-0.33480	-0.39314	-0.37198
32	-3.27802	-0.20680	-0.13936	0.05626
33	-2.48058	-1.99375	-1.69587	-1.87838
34	1.88639	0.66475	0.79670	0.69865
35	5.06459	-0.27977	0.10817	-0.26450
36	1.96335	-0.43990	-0.66022	-0.58490
37	0.26274	-0.05334	-0.02396	-0.16795
38	-2.94593	-2.06699	-1.31832	-1.72082
39	-1.38628	-1.02560	-1.15858	-1.19420
40	-0.20797	0.45847	0.21555	0.31347
41	-1.07820	0.93979	0.63454	0.69907
42	-1.66386	-0.93238	-0.95216	-1.02725
43	0.58117	-0.35988	-0.30693	-0.46232
44	1.37447	-0.00518	0.35985	0.02485
45	-2.82687	-0.18892	-0.07959	0.01704
46	3.86363	0.87271	0.47644	0.57036
47	-2.64141	-2.06993	-1.82915	-2.16738
48	0.05387	0.32143	-0.25961	0.21632
49	0.50763	-1.42382	-0.77620	-1.07473
50	0.64347	-0.07852	-0.21347	-0.11750

表 9.2 均等教育机会 (EEO) 数据的后 20 个观测; 各标准化的指数

行	ACHV	FAM	PEER	SCHOOL
51	2.49414	-0.14925	-0.03192	-0.36598
52	0.61955	0.52666	0.79149	0.71369
53	0.61745	-1.49102	-1.02073	-1.38103
54	-1.00743	-0.94757	-1.28991	-1.24799
55	-0.37469	0.24550	0.83794	0.59596
56	-2.52824	-0.41630	-0.60312	-0.34951
57	0.02372	1.38143	1.54542	1.59429
58	2.51077	1.03806	0.91637	0.97602
59	-4.22716	-0.88639	-0.47652	-0.77693
60	1.96847	1.08655	0.65700	0.89401
61	1.25668	-1.95142	-1.94199	-1.89645
62	-0.16848	2.83384	2.47398	2.79222
63	-0.34158	1.86753	1.55229	1.80057
64	-2.23973	-1.11172	-0.69732	-0.80197
65	3.62654	1.41958	1.11481	1.24558
66	0.97034	0.53940	0.16182	0.33477
67	3.16093	0.22491	0.74300	0.66182
68	-1.90801	1.48244	1.47079	1.54283
69	0.64598	2.05425	1.80369	1.90066
70	-1.75915	1.24058	0.64484	0.87372

就获得了对于修正概念的另一种观点。左边是一种修正的学业成就指数, 修正是通过减去 FAM 和 PEER 的线性贡献来实现的。该方程是以修正的学业成就得分关于 SCHOOL 变量的回归形式表示的。这种表示仅仅出于解释的需要。各 β 的估计值由方程 (9.1) 中给出的原始模型获得。回归结果总结在表 9.3 中, 残差关于 ACHV 的预测值的散点图见图 9.1。

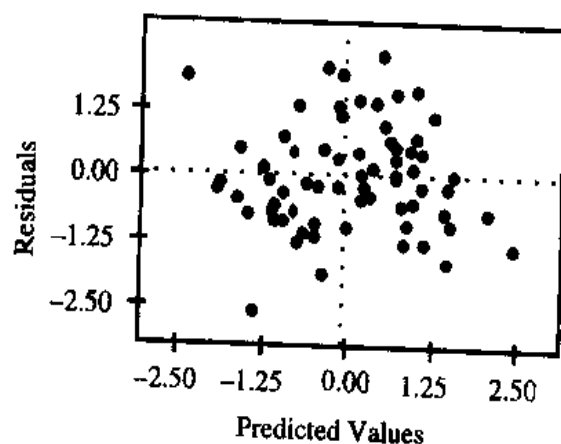


图 9.1 标准化残差关于 ACHV 之拟合值的散点图

首先考察残差图, 我们看到其中没有明显的模型误设的迹象。位于图左下方那个点的残差值离开均值 0 大约 2.5 个标准差, 应引起重视。但是, 若将它从样本中删除, 回归结果几乎没有改变。因此, 该观测在分析时依然保留。

表 9.3 EEO 数据: 回归结果

变量	系数	标准误	<i>t</i> -检验	<i>p</i> -值
常数	-0.070	0.251	-0.28	0.7810
FAM	1.101	1.411	0.78	0.4378
PEER	2.322	1.481	1.57	0.1218
SCHOOL	-2.281	2.220	-1.03	0.3080
$n = 70$	$R^2 = 0.206$	$R_a^2 = 0.170$	$\hat{\sigma} = 2.07$	$d.f. = 66$

从表 9.3 我们看到, 学业成就得分的约 20% 的变异可由三个预测因子共同来解释 ($R^2 = 0.206$)。F-值为 5.72, 其自由度为 3 和 66, 在高于 0.01 的水平下显著。因此, 尽管被解释的总变异估计才为 20%, 但可以认为, FAM, PEER, SCHOOL 是合适的预测变量。然而, 单个 *t*-值都较小。总之, 各概述性统计量告诉我们, 这三个预测变量放在一起是重要的, 但从各 *t*-值来看, 假如保留另两个的话, 任何一个预测因子都可以从模型中剔除。

这些结果是存在严重多重共线性的典型表现。这些预测变量高度相关, 以致于每一个都可以在回归方程中替代另两个, 而几乎不影响总的解释能力。*t*-值小进一步证实任何一个预测变量都可被剔除出方程。因此, 回归分析在评价学校投入对学业成就的重要性方面无法提供任何信息。罪魁祸首显然是多重共线性。这三个预测变量两两之间的相关系数及相应的散点图 (图 9.2) 都表明, 两两预测变量之间都呈现着很强的线性关系, 各相关系数都很高。在各散点图中, 所有观测都密集于穿过相应变量之均值点的直线旁边。

本例中, 多重共线性可想而知。每一个变量都由另两个决定、或帮助决定另两个, 这是这三个变量的本质。事实只有一个变量而没有三个变量, 得出这样的结论也不无道理。遗憾的是, 该结论无助于回答学校设施对于学业成就的效应这个原始问题。有两种可能性: 第一, 多重共线性可能是由于样本数据不充分导致的, 可以通过补充观测得以改善。第二, 多重共线性可能是由于调查过程的内在性质造成的。下面我们讨论这两种情况。

对于第一种情况, 收集样本应保证预测变量间的相关性不太大。譬如, 从 FAM 关于 SCHOOL 的散点图 (图 9.2 中右上角的图) 中可以发现, 样本中没有数值位于左上或右下区域的学校。因此样本中没有 FAM 值高但 SCHOOL 值低、或 FAM 值低但 SCHOOL 值高时的学业成就信息。然而, 只有在这两种条件下收集了数据, 才能确定 FAM 和 SCHOOL 单个变量对于 ACHV 的效应。例如, 假定在图的左上象限中有一些观测的话, 那么至少有可能在 FAM 固定时, 比较 SCHOOL 取高值和低值时 ACHV 的平均值。

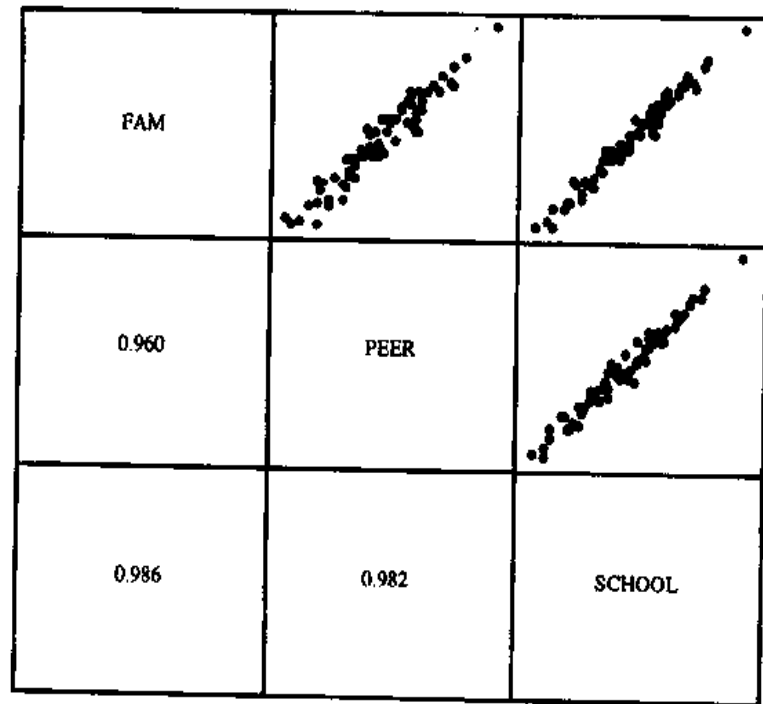


图 9.2 三个预测变量 FAM、PEER、SCHOOL 之间成对的散点图及相应的相关系数

因为模型中有三个预测变量，那么在样本中应当包含八种不同的数据组合。若用 + 表示大于均值的值，用 - 表示小于均值的值，则表 9.4 中列出了八种可能性。

表 9.4 三个预测变量的数据组合

组合	变量		
	FAM	PEER	SCHOOL
1	+	+	+
2	+	+	-
3	+	-	+
4	-	+	+
5	+	-	-
6	-	+	-
7	-	-	+
8	-	-	-

分析中发现的强相关性意味着仅 1, 8 两种组合出现在数据中。如果样本是偶然呈现这种情况的，那么解决多重共线性问题的处方为再增补一些其他组合下的数据。譬如，组合 1 和 2 下的数据可用于评价 FAM 和 PEER 均固定在高于均值的水平时，SCHOOL 对于 ACHV 的效应。假如数据中仅含这两种组合，则分析将由 ACHV 关于 SCHOOL 的简单回归构成，这时仅能得到局部的结论，即在 FAM

和 PEER 均高于均值时学校情况与学业成就之间关系的评估。

增补数据解决多重共线性问题的处方不是万试万灵的。常常因为预算、时间、人员等的限制而不可能去收集更多的数据。一般,较好的做法是事先就防备到可能会发生数据不充分的情况,尽可能地按照设计去收集数据。遗憾的是,事先的设计不总是可行的。在调查或观测研究中,如本例,通常一直要到样本单元选定了、一些费钱费时的测量做了之后,才知道预测变量的值。若按照这套程序进行的话,很难保证能得到一个平衡的样本。

出现多重共线性的第二个可能的原因是,变量间的关系是实施抽样的那个过程的一种固有特性。如果在总体中,变量 FAM, PEER, SCHOOL 仅以表 9.4 中的 1, 8 两种数据组合形式存在,那么,估计这些变量单个地对于学业成就的效应是不可能的。进一步分析这些效应的唯一对策是,寻找可以解释预测变量间内在关系的根本原因。经过这一过程,人们可能会发现其他变量才是影响均等教育机会及学业成就的更基本的决定因素。

9.3 对预测的影响

我们接着考察,在用多元回归方程作预测时,多重共线性对于预测的影响。我们用一个历史数据集去估计回归系数,其中的观测按时间顺序排列。我们根据回归方程中各预测变量的未来值,得到响应变量的预测。预测变量的未来值必须已知或者可根据其他数据和模型预测。在我们的讨论中,不考虑预测变量之预报值中的不确定性,假定预测变量的未来值是给定的。

我们选择了一个例子,有关法国经济中进口活动的总量数据。Malinvaud(1968)已经分析过这批数据。我们接着他的叙述进行讨论。各变量为进口额 (IMPORT), 国内总产值 (DOPROD), 存储量 (STOCK), 与国内消费总额 (CONSUM), 单位均为 10 亿法郎。数据从 1949 年至 1966 年,列于表 9.5 中,也可从本书的网站获得。考虑的模型为

$$\text{IMPORT} = \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_3 \cdot \text{CONSUM} + \varepsilon. \quad (9.3)$$

回归结果置于表 9.6 中。残差序列图 (图 9.3) 呈现出特别的模式,这说明设定的模型不合适。这时,尽管多重共线性看来是出现了 ($R^2 = 0.973$ 但所有 t -值较小),但不应该在此模型下继续追究。只有当模型设定满意时,才有必要去追究多重共线性。这个模型的困难在于,欧洲共同市场从 1960 年起开始运行,从而导致了进出口关系的变化。因为我们本章的目的是研究多重共线性的效应,我们准备去刻划 1959 年后的活动而把模型复杂化。我们假定,现在是 1960 年,只看 1949-1959 这 11 年的情况。这些数据的回归结果汇总于表 9.7。此时,残差图是令人满意的 (图 9.4)。

表 9.5 法国经济数据

YEAR	IMPORT	DOPROD	STOCK	CONSUM
49	15.9	149.3	4.2	108.1
50	16.4	161.2	4.1	114.8
51	19.0	171.5	3.1	123.2
52	19.1	175.5	3.1	126.9
53	18.8	180.8	1.1	132.1
54	20.4	190.7	2.2	137.7
55	22.7	202.1	2.1	146.0
56	26.5	212.4	5.6	154.1
57	28.1	226.1	5.0	162.3
58	27.6	231.9	5.1	164.3
59	26.3	239.0	0.7	167.6
60	31.1	258.0	5.6	176.8
61	33.3	269.8	3.9	186.6
62	37.0	288.4	3.1	199.7
63	43.3	304.5	4.6	213.9
64	49.0	323.4	7.0	223.8
65	50.3	336.8	1.2	232.0
66	56.6	353.9	4.5	242.9

来源: Malinvaud (1968)。

表 9.6 进口数据 (1949-1966 年): 回归结果

ANOVA 表				
来源	平方和	d.f.	均方	F-检验
回归	2576.92	3	858.974	168
残差	71.39	14	5.099	
系数表				
变量	系数	标准误	t-检验	p-值
常数	-19.725	4.125	-4.78	0.0003
DOPROD	0.032	0.187	0.17	0.8656
STOCK	0.414	0.322	1.29	0.2195
CONSUM	0.243	0.285	0.85	0.4093
$n = 18$	$R^2 = 0.973$	$R_a^2 = 0.967$	$\hat{\sigma} = 2.258$	$d.f. = 14$

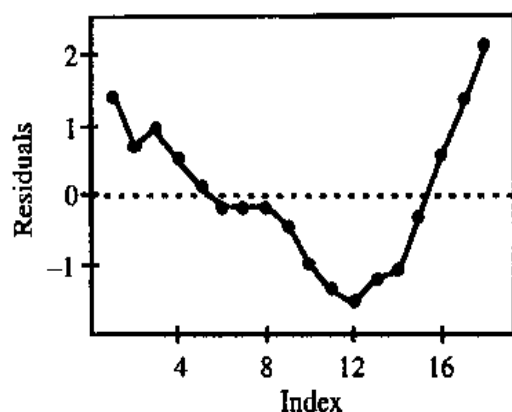


图 9.3 进口数据 (1949—1966 年): 标准化残差序列图

表 9.7 进口数据 (1949—1959 年): 回归结果

ANOVA 表				
来源	平方和	<i>d.f.</i>	均方	<i>F</i> -检验
回归	204.776	3	68.2587	286
残差	1.673	7	0.2390	
系数表				
变量	系数	标准误	<i>t</i> -检验	<i>p</i> -值
常数	-10.128	1.212	-8.36	< 0.0001
DOPROD	-0.051	0.070	-0.73	0.4883
STOCK	0.587	0.095	6.20	0.0004
CONSUM	0.287	0.102	2.81	0.0263
<i>n</i> = 11	$R^2 = 0.992$	$R_a^2 = 0.988$	$\hat{\sigma} = 0.4889$	<i>d.f.</i> = 7

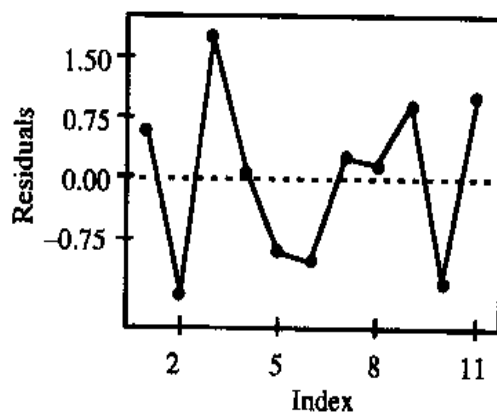


图 9.4 进口数据 (1949—1959 年): 标准化残差序列图

$R^2 = 0.99$ 是高的。但 DOPROD 的系数为负, 且无统计显著性, 这与事先的预计相矛盾。我们相信, 若 STOCK 和 CONSUM 保持恒定, DOPROD 的增长

会导致 IMPORT 的增长,可能是各种原材料、制造加工设备的进口增长。这里,有可能是多重共线性引起了矛盾,而事实也正如此。CONSUM 和 DOPROD 间的简单相关系数为 0.997。进一步的调查发现,在这 11 年中,CONSUM 大约是 DOPROD 的三分之二。这两个量之间的关系估计为

$$\text{CONSUM} = 6.259 + 0.686 \cdot \text{DOPROD}.$$

尽管存在如此严重的多重共线性关系,但该回归方程仍然可以作出很好的预报。由表 9.7,预报方程为

$$\text{IMPORT} = -10.13 - 0.051 \cdot \text{DOPROD} + 0.587 \cdot \text{STOCK} + 0.287 \cdot \text{CONSUM}.$$

回顾一下,模型对历史数据拟合得非常好,残差的变动看上去也是纯随机的。为作预报,我们必须确信,该整体关系的特征、强度在未来阶段都将保持不变。这个信心问题是所有预报模型共有的问题,无论多重共线性是否出现。出于本例的目的,我们假定该整体关系确实在未来阶段保持不变^①。DOPROD 和 CONSUM 之间的关系也隐含在该假定中。只要 DOPROD, STOCK 和 CONSUM 的未来值保持着 CONSUM 大约等于 DOPROD 的三分之二的关系,那么预测将是准确的。

例如,让我们预报下一年进口额的变化,假定 DOPROD 增长 10 个单位,而 STOCK 和 CONSUM 保持现在的水平不变。那么预报结果为

$$\text{IMPORT}_{1960} = \text{IMPORT}_{1959} - 0.051(10),$$

意味着 IMPORT 将下降 0.51 个单位。但是,如果 DOPROD 与 CONSUM 间的关系未受影响,那么,CONSUM 将增加 $10(2/3)$ 个单位,而预报结果为

$$\text{IMPORT}_{1960} = \text{IMPORT}_{1959} - 0.51 + 0.287 \cdot 10(2/3) = \text{IMPORT}_{1959} + 1.4.$$

实际上 IMPORT 增长了 1.4 个单位,这是个更令人满意的结果并且可能是个更佳的预报。DOPROD 单独增加的情况,改变了用来估计模型参数的数据的基本结构,因而不可能得出有意义的预报。

总之,上述两个例子说明了多重共线性数据可能会严重限制回归分析在推断和预报中的作用。当怀疑有多重共线性存在的情况下,解释回归结果务必格外小心。9.4 节我们讨论检测预测变量间严重共线性关系的各种方法。

9.4 多重共线性的检测

在前面的例子中,已经介绍了一些检测多重共线性的想法。本节中,我们再回顾一下那些想法,并补充介绍一些能指出共线性的准则。多重共线性往往与不稳

^① 出于阐述方便的目的,我们忽略了我们前面的发现所带来的困难,即自 1960 年起,欧洲共同市场的建立改变了该关系。但我们不得不忠告读者,即使历史的拟合非常好,然而结构上的变化往往会使得预测变成一桩吃力不讨好的事。

定的回归系数估计联系在一起。这是由预测变量间存在的强线性关系导致的, 不是一个模型误设的问题。因此, 对共线性数据导致的问题的调查, 应当在满意地设定了模型之后开始。然而, 在为搜索理想的模型而对变量或数据点作增、删、变换的过程中, 也可能出现一些多重共线性的征兆。使系数估计不稳定的多重共线性征兆为:

- 增、删某个变量时, 系数的估计有大的改变。
- 改变或删除一个数据点时, 系数的估计有大的改变。
- 且残差图显示设定的模型已令人满意, 则下列情况下可能存在多重共线性:
 - 系数估计的代数符号与事先预期的不相符; 或者
 - 预计比较重要的变量其系数的标准误很大 (t -值小)。

对于前面讨论的进口数据, DOPROD 的系数是负的且不显著。这都与事先预期相矛盾。增、删一个变量的效应可见表 9.8。从中我们发现, 某些变量的存在与否对其他系数有着重大影响。对于 EEO 数据 (表 9.1 与表 9.2), 系数的代数符号都正确, 但它们的标准误很大以致没有一个系数在统计意义下是显著的。而预计它们都是重要的。

表 9.8 进口数据 (1949–1959 年): 所有可能回归的回归系数

回归	变量			
	常数	DOPROC	STOCK	CONSUM
1	-6.558	0.146	-	-
2	19.611	-	0.691	-
3	-8.013	-	-	0.214
4	-8.440	0.145	0.622	-
5	-8.884	-0.109	-	0.372
6	-9.743	-	0.596	0.212
7	-10.128	-0.051	0.587	0.287

多重共线性的存在也可由预测变量间的相关系数的大小来反映。某对预测变量间的相关系数大则表明那对变量间存在较强的线性关系。EEO 数据中各对预测变量间的相关系数 (图 9.2) 都很大。在进口数据中, DOROD 和 CONSUM 间的相关系数为 0.997。

多重共线性的来源可能比两个变量间的简单相关更复杂。一个线性关系可能涉及多个预测变量。通过简单相关系数不一定能发现这种关系。举个例子, 我们来看, 广告费用 (A_t)、促销费用 (P_t) 及销售费用 (E_t) 对于一个公司在时期 t 内总销售额的效应的分析。数据反映的是该公司 23 年的情况, 这期间公司的运营条件相当稳定。数据列于表 9.9 中, 也能从本书的网站获得。

提出的回归模型为

$$S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon_t, \quad (9.4)$$

其中 A_{t-1} 和 P_{t-1} 为滞后一年的变量。回归结果在表 9.10 中给出。残差关于预测

表 9.9 广告、促销、销售费用与销售额的年度数据 (百万美元)

行	S_t	A_t	P_t	E_t	A_{t-1}	P_{t-1}
1	20.11371	1.98786	1.0	0.30	2.01722	0.0
2	15.10439	1.94418	0.0	0.30	1.98786	1.0
3	18.68375	2.19954	0.8	0.35	1.94418	0.0
4	16.05173	2.00107	0.0	0.35	2.19954	0.8
5	21.30101	1.69292	1.3	0.30	2.00107	0.0
6	17.85004	1.74334	0.3	0.32	1.69292	1.3
7	18.87558	2.06907	1.0	0.31	1.74334	0.3
8	21.26599	1.01709	1.0	0.41	2.06907	1.0
9	20.48473	2.01906	0.9	0.45	1.01709	1.0
10	20.54032	1.06139	1.0	0.45	2.01906	0.9
11	26.18441	1.45999	1.5	0.50	1.06139	1.0
12	21.71606	1.87511	0.0	0.60	1.45999	1.5
13	28.69595	2.27109	0.8	0.65	1.87511	0.0
14	25.83720	1.11191	1.0	0.65	2.27109	0.8
15	29.31987	1.77407	1.2	0.65	1.11191	1.0
16	24.19041	0.95878	1.0	0.65	1.77407	1.2
17	26.58966	1.98930	1.0	0.62	0.95878	1.0
18	22.24466	1.97111	0.0	0.60	1.98930	1.0
19	24.79944	2.26603	0.7	0.60	1.97111	0.0
20	21.19105	1.98346	0.1	0.61	2.26603	0.7
21	26.03441	2.10054	1.0	0.60	1.98346	0.1
22	27.39304	1.06815	1.0	0.58	2.10054	1.0

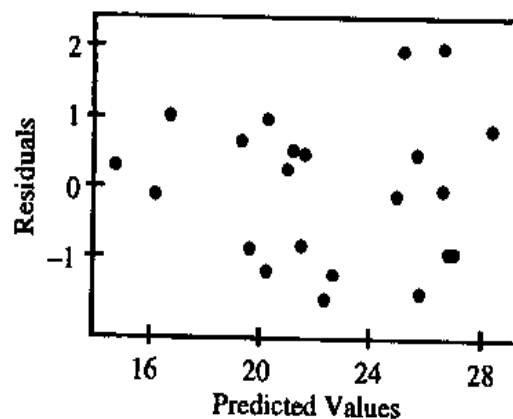


图 9.5 标准化残差关于销售额拟合值的散点图

值的散点图、残差序列图 (图 9.5 和图 9.6), 以及残差关于预测变量的散点图 (未列出) 没有显示出任何模型误设的问题。而且, 预测变量间的相关系数也都很小 (表 9.11)。然而, 如果为检查系数的稳定性, 我们做个小小的试验, 从模型中去掉当期广告变量 A_t , 结果就大相径庭了。 P_t 的系数从 8.37 降到了 3.70; 滞后广告费

用 A_{t-1} 与滞后促销费用 P_{t-1} 的系数的符号改变了。但销售费用的系数稳定, 且 R^2 没有大的变化。

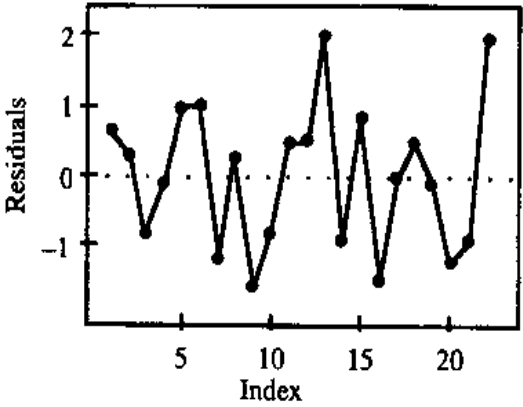


图 9.6 标准化残差的序列图

表 9.10 广告数据的回归结果

ANOVA 表				
来源	平方和	<i>d.f.</i>	均方	<i>F</i> -检验
回归	307.572	5	61.514	35.3
残差	27.879	16	1.742	
系数表				
变量	系数	标准误	<i>t</i> -检验	<i>p</i> -值
常数	-14.194	18.715	-0.76	0.4592
<i>A</i>	5.361	4.028	1.33	0.2019
<i>P</i>	8.372	3.586	2.33	0.0329
<i>E</i>	22.521	2.142	10.51	< 0.0001
A_{t-1}	3.855	3.578	1.08	0.2973
P_{t-1}	4.125	3.895	1.06	0.3053
$n = 22$	$R^2 = 0.917$	$R_a^2 = 0.891$	$\hat{\sigma} = 1.320$	<i>d.f.</i> = 16

表 9.11 广告数据的逐对相关系数

	A_t	P_t	E_t	A_{t-1}	P_{t-1}
A_t	1.000				
P_t	-0.357	1.000			
E_t	-0.129	0.063	1.000		
A_{t-1}	-0.140	-0.316	-0.166	1.000	
P_{t-1}	-0.496	-0.296	0.208	-0.358	1.000

该证据表明, 当期、滞后的广告费用及促销费用各变量之间存在某种关系。 A_t

关于 P_t, A_{t-1}, P_{t-1} 的回归之 R^2 为 0.973。方程的形式为

$$\hat{A}_t = 4.63 - 0.87P_t - 0.86A_{t-1} - 0.95P_{t-1}.$$

根据对该公司运营情况的进一步调查发现,在这平稳的 23 年中,公司对经费预算进行了严格的控制。特别地,对预算有一条粗略的硬性规定,即每两年中, A_t, A_{t-1}, P_t 及 P_{t-1} 之和应大约控制在 5 个单位左右。关系式

$$A_t + P_t + A_{t-1} + P_{t-1} \doteq 5$$

是多重共线性的缘由。

对多重共线性的全面考察,将涉及对每个预测变量关于所有其他预测变量的回归之 R^2 的考察。预测变量间的关系可由所谓的方差膨胀因子(VIF) 这个量来判断。记 R_j^2 为预测变量 X_j 关于所有其他预测变量作回归得到的复相关系数之平方。那么 X_j 的方差膨胀因子为

$$\text{VIF}_j = \frac{1}{1 - R_j^2}, \quad j = 1, \dots, p, \quad (9.5)$$

其中 p 为预测变量的个数。很明显,若 X_j 与其他预测变量有较强的线性关系,那么 R_j^2 将接近于 1,从而 VIF_j 就会很大。方差膨胀因子的值大于 10,常被视为数据有共线性问题的一种信号。

若预测变量间不存在任何线性关系(也即各预测变量是正交的)[†],那么 R_j^2 将为 0,而 VIF_j 就将为 1。 VIF_j 的值偏离 1,说明数据偏离正交性而倾向于共线性。 VIF_j 的值也度量了,由于 X_j 与其他预测变量的线性关系而导致的第 j 个回归系数的方差比 X_j 与其他变量不相关时的方差增大的倍数。这解释了此诊断量的得名。

当 R_j^2 趋于 1,即表明预测变量间存在线性关系时, $\hat{\beta}_j$ 的 VIF 趋于无穷。这意味着,VIF 超过 10 是多重共线性可能对估计带来麻烦的一个信号。

回归系数的普通最小二乘估计的精度由其方差来衡量,与回归模型中误差项的方差 σ^2 成比例,这比例常数就是 VIF。因而,可用各 VIF 获得各 OLS 估计与其真实值的距离平方之期望的表达式。记该距离平方为 D^2 ,可以证明,平均起来,

$$D^2 = \sigma^2 \sum_{j=1}^p \text{VIF}_j.$$

该距离是最小二乘估计精度的另一种量度。该距离越小,则估计越精确。若预测变量正交,则 VIF 将都为 1,而 D^2 将为 $p\sigma^2$ 。因而,比率

$$\frac{\sigma^2 \sum_{i=1}^p \text{VIF}_i}{p\sigma^2} = \frac{\sum_{i=1}^p \text{VIF}_i}{p} = \overline{\text{VIF}}$$

[†] 译注:参见 9.1 节译注。

即各 VIF 之均值, 衡量的是 OLS 估计的均方误差相对于数据正交时的倍数。因此, \overline{VIF} 也可作多重共线性的指标[†]。

目前, 绝大多数电脑软件包例行地计算各 VIF_j 值。有些还内置了对较高的 VIF_j 值报警的信息。在任何回归分析中, 总应该检查 VIF_j 的值, 以避免用最小二乘法对共线性数据拟合回归模型时的隐患。

在上述三个例子 (EEO 数据, 进口数据及广告数据) 中, 我们都看到了共线性的迹象。这些数据集中各 VIF_j 的值及其均值在表 9.12 中给出。EEO 数据中各 VIF_j 值的变化从 30.2 至 83.2, 说明全部三个变量之间强相关, 去掉某个变量还不能消除共线性。VIF 的均值 50.3 表明 OLS 估计的均方误差是预测变量正交时的 50 倍。

进口数据中, OLS 估计的均方误差是预测变量正交时的 313 倍。但是, 各 VIF_j 值表明, 国内产值与消费强相关, 但与 STOCK 变量不相关。一个包含 CONSUM 或 DOPROD 以及 STOCK 的回归方程将消除共线性。

广告数据中, (变量 E 的) VIF_E 为 1.1, 表明该变量与其余预测变量不相关。但其他四个变量的 VIF_j 很大, 变化范围从 26.6 至 44.1。这表明这四个变量间存在较强的线性关系, 这是我们已经注意到的一个事实。这里, 解决办法可为, 作销售额 S_t 关于 E_t 及 $(A_t, P_t, A_{t-1}, P_{t-1})$ 中的三个变量的回归, 然后考察由此导出的各 VIF_j 值来看共线性是否被消除。

表 9.12 三个数据集的方差膨胀因子

EEO 数据		进口数据		广告数据	
变量	VIF	变量	VIF	变量	VIF
FAM	37.6	DOPROD	469.7	A_t	37.4
PEER	30.2	STOCK	1.0	P_t	33.5
SCHOOL	83.2	CONSUM	469.4	E_t	1.1
				A_{t-1}	26.6
				P_{t-1}	44.1
均值	50.3	均值	313.4	均值	28.5

9.5 中心化及尺度变换

至今我们描述的检测多重共线性的指标都可以通过标准的回归计算得到。另外还有一种更统一的分析多重共线性的办法, 这需要用到一些在通常标准回归软件包中没有的计算。该分析依据这样一个事实: 每一个线性回归方程都可换用一组正交的预测变量来重新表述。这些新变量是以原始预测变量的线性组合的形式获得的, 称为预测变量集的主成分 (Seber, 1984; Johnson and Wichern, 1992)。

[†] 译注: 这一段中的诸公式都是在各预测变量已经中心化与长度单位化条件下才成立的, 参见 9.5 节。

为逐步阐明主成分方法, 我们可能先要对变量进行中心化且(或)尺度变换。前面我们主要讨论的是带有常数项 β_0 的回归模型

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon. \quad (9.6)$$

但也碰到过需拟合无截距模型

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \quad (9.7)$$

的情况(比如, 见第3章和第7章)。在讨论带常数项的模型时, 对变量作中心化和尺度变换是合适的, 但在讨论无截距模型时, 我们仅需要对变量作尺度变换。

9.5.1 有截距模型中的中心化和尺度变换

如果我们拟合的是如同式(9.6)那样有截距的模型, 那我们需要对变量作中心化和尺度变换。从每个观测中减去所有观测的均值, 便可获得中心化的变量。比如, 中心化的响应变量为 $Y - \bar{y}$, 中心化的第 j 个预测变量为 $X_j - \bar{x}_j$ 。中心化变量的均值为 0。中心化变量还可以作尺度变换。常用两种类型的尺度变换: 长度单位化和标准化。响应变量 Y 及第 j 个预测变量 X_j 的长度单位化变换依下法获得:

$$\begin{aligned} \tilde{Z}_y &= \frac{Y - \bar{y}}{L_y}, \\ \tilde{Z}_j &= \frac{X_j - \bar{x}_j}{L_j}, \quad j = 1, \cdots, p, \end{aligned} \quad (9.8)$$

其中 \bar{y} 是 Y 的均值, \bar{x}_j 是 X_j 的均值, 而

$$L_y = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}, \quad L_j = \sqrt{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}, \quad j = 1, \cdots, p. \quad (9.9)$$

L_y 称为中心化变量 $Y - \bar{y}$ 的长度, 因为它衡量的是 $Y - \bar{y}$ 中各观测值的大小。类似地, L_j 衡量的是变量 $X_j - \bar{x}_j$ 的长度。(9.8) 中变量 \tilde{Z}_y, \tilde{Z}_j 均值都为 0 长度都是 1, 因此这类尺度变换称为长度单位化。另外, 长度单位化还具有如下性质:

$$Cor(X_j, X_k) = \sum_{i=1}^n z_{ij} z_{ik}. \quad (9.10)$$

即原始变量 X_j 与 X_k 的相关系数可方便地由尺度变换后的版本 Z_j 与 Z_k 的乘积和来计算。

第二类尺度变换称为标准化, 定义为

$$\begin{aligned} \hat{Y} &= \frac{Y - \bar{y}}{s_y}, \\ \hat{X}_j &= \frac{X_j - \bar{x}_j}{s_j}, \quad j = 1, \cdots, p, \end{aligned} \quad (9.11)$$

其中

$$s_y = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n-1}} \text{ 与 } s_j = \sqrt{\frac{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}{n-1}}, j = 1, \dots, p \quad (9.12)$$

分别为响应和第 j 个预测变量的标准差。(9.11) 中标准化变量 \hat{Y} 与 \hat{X}_j 均值都为 0, 标准差都是 1。

因为对数据作平移或尺度变换不影响其相关性, 因此, 拿经长度单位化或标准化后的变量作分析是既充分又方便。 p 个变量 X_1, \dots, X_p 的方差和协方差可整齐地用一个名为矩阵的数字阵列来表示。该矩阵称为方差-协方差矩阵。矩阵从左上到右下对角线上的元素称为对角元。方差-协方差矩阵的对角元为方差, 非对角元为协方差^①。进口数据 (1949-1959 年) 三个预测变量的方差-协方差矩阵为:

$$\begin{array}{c} \text{DOPROD} \quad \text{STOCK} \quad \text{CONSUM} \\ \text{DOPROD} \left(\begin{array}{ccc} 899.971 & 1.279 & 617.326 \\ 1.279 & 2.720 & 1.214 \\ 617.326 & 1.214 & 425.779 \end{array} \right) \\ \text{STOCK} \\ \text{CONSUM} \end{array}$$

譬如, $\text{Var}(\text{DOPROD}) = 899.971$ 是第一个对角元, $\text{Cov}(\text{DOPROD}, \text{CONSUM}) = 617.326$ 为第一行与第三列交点处的值 (或第三行与第一列)。

类似地, 两两变量间的相关系数也可用矩阵表示, 称为相关矩阵。进口数据中三个预测变量的相关矩阵为

$$\begin{array}{c} \text{DOPROD} \quad \text{STOCK} \quad \text{CONSUM} \\ \text{DOPROD} \left(\begin{array}{ccc} 1.000 & 0.026 & 0.997 \\ 0.026 & 1.000 & 0.036 \\ 0.997 & 0.036 & 1.000 \end{array} \right) \\ \text{STOCK} \\ \text{CONSUM} \end{array} \quad (9.13)$$

这与标准化的预测变量的方差-协方差矩阵相同。譬如, $\text{Cor}(\text{DOPROD}, \text{CONSUM}) = 0.997$, 表明这两个变量高度相关。注意, 相关矩阵的所有对角元都等于 1。

回想一下, 一组变量之间如果不存在线性关系, 则称为是正交的[†]。如果标准化的预测变量是正交的, 那么其方差-协方差矩阵之对角元为 1、非对角元为 0。

9.5.2 无截距模型中的尺度变换

如果我们拟合的是如同 (9.7) 那样的无截距模型, 我们不对数据中心化, 因为中心化会在模型中引进一个常数项。这可从下式看出:

$$Y - \bar{y} = \beta_1(X_1 - \bar{x}_1) + \dots + \beta_p(X_p - \bar{x}_p) + \varepsilon. \quad (9.14)$$

^① 不熟悉矩阵代数的读者可参阅 Hadi(1996), *Matrix Algebra As a Tool*.

[†] 译注: 参见 9.1 节译注。

整理各项, 我们得到

$$\begin{aligned} Y &= \bar{y} - (\beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p) + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \\ &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon \end{aligned} \quad (9.15)$$

其中 $\beta_0 = \bar{y} - (\beta_1 \bar{x}_1 + \cdots + \beta_p \bar{x}_p)$ 。尽管常数项未直接地在 (9.14) 中出现, 但可明显地从 (9.15) 中看出来。因此, 当我们处理无截距模型时, 我们只须对数据作尺度变换。经尺度变换的变量定义为:

$$\begin{aligned} \tilde{Z}_y &= \frac{Y}{L_y}, \\ \tilde{Z}_j &= \frac{X_j}{L_j}, \quad j = 1, \cdots, p, \end{aligned} \quad (9.16)$$

其中

$$L_y = \sqrt{\sum_{i=1}^n y_i^2}, \quad L_j = \sqrt{\sum_{i=1}^n x_{ij}^2}, \quad j = 1, \cdots, p. \quad (9.17)$$

(9.16) 中经尺度变换的变量均为单位长度的, 但均值不一定为 0。它们也不满足 (9.10), 除非原始变量都是 0 均值的。

这里, 我们应当提到, 中心化 (在合适的场合) 和 (或) 尺度变换是不失一般性的做法, 因为原始变量的回归系数可从变换后的变量的回归系数恢复出来。譬如, 如果我们对中心化的数据拟合了一个回归模型, 得到的回归系数 $\hat{\beta}_1, \cdots, \hat{\beta}_p$ 与对原始数据拟合模型得到的估计是一样的。用中心化数据, 常数项的估计总是 0。有截距模型的常数项的估计可如此获得:

$$\hat{\beta}_0 = \bar{y} - (\hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_p \bar{x}_p).$$

然而, 尺度变换会改变回归系数的估计值。比如, 由原始数据获得的估计 $\hat{\beta}_1, \cdots, \hat{\beta}_p$ 与由标准化数据获得的估计 $\hat{\theta}_1, \cdots, \hat{\theta}_p$ 之间的关系为

$$\begin{aligned} \hat{\beta}_j &= (s_y/s_j)\hat{\theta}_j, \quad j = 1, \cdots, p, \\ \hat{\beta}_0 &= \bar{y} - \sum_{j=1}^p \hat{\beta}_j \bar{x}_j. \end{aligned} \quad (9.18)$$

在采用长度单位化而不是标准化时, 也可获得类似的公式。

在本章后文及第 10 章中, 我们将广泛地使用经中心化且 (或) 尺度变换后的变量。

9.6 主成分方法

正如我们在前一节中提到的, 用主成分方法检测多重共线性依据如下事实: 任何一组 p 个变量均可变换为一组 p 个正交的变量。新的正交的变量称为主成

分, 记为 C_1, \dots, C_p 。每个变量 C_j 都是 (9.11) 中变量 $\tilde{X}_1, \dots, \tilde{X}_p$ 的一种线性组合。即

$$C_j = v_{1j}\tilde{X}_1 + v_{2j}\tilde{X}_2 + \dots + v_{pj}\tilde{X}_p, \quad j = 1, \dots, p. \quad (9.19)$$

这些线性组合选得使变量 C_1, \dots, C_p 正交^①。各主成分的方差-协方差矩阵形为

$$\begin{matrix} & C_1 & C_2 & \cdots & C_p \\ \begin{matrix} C_1 \\ C_2 \\ \vdots \\ C_p \end{matrix} & \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix} \end{matrix}.$$

因为各主成分正交, 所以所有非对角元均为 0。第 j 个对角元的值 λ_j 是第 j 个主成分 C_j 的方差。主成分的排列要使得 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$, 即第一主成分方差最大, 最后一个主成分方差最小。各 λ 称为预测变量 X_1, \dots, X_p 的相关矩阵的特征根。(9.19) 中用来构造 C_j 的系数可整齐地排列成如下一个列

$$\begin{pmatrix} v_{1j} \\ v_{2j} \\ \vdots \\ v_{pj} \end{pmatrix},$$

这是属于第 j 个特征根 λ_j 的特征向量。只要有一个 λ 精确地等于 0, 那么这些原始变量之间就存在一种完全的线性关系, 这是多重共线性的一种极端情况。若某个 λ 比其他小很多 (且接近于 0), 那就存在多重共线性。接近于 0 的 λ 的个数, 就等于数据中存在的不同多重共线性的组数。即若只有一个 λ 接近于 0, 则只有一组多重共线性; 若有两个接近于 0 的 λ , 则有两组不同的多重共线性; 依此类推。

(9.13) 中相关矩阵的特征根为 $\lambda_1 = 1.999, \lambda_2 = 0.998, \lambda_3 = 0.003$ 。相应的特征向量为

$$\begin{pmatrix} 0.706 \\ 0.044 \\ 0.707 \end{pmatrix}, \quad \begin{pmatrix} -0.036 \\ 0.999 \\ -0.026 \end{pmatrix}, \quad \begin{pmatrix} -0.707 \\ -0.007 \\ 0.707 \end{pmatrix}.$$

因此, 1949-1959 年间进口数据的各主成分为:

$$\begin{aligned} C_1 &= +0.706 \tilde{X}_1 + 0.044 \tilde{X}_2 + 0.707 \tilde{X}_3, \\ C_2 &= -0.036 \tilde{X}_1 + 0.999 \tilde{X}_2 - 0.026 \tilde{X}_3, \\ C_3 &= -0.707 \tilde{X}_1 - 0.007 \tilde{X}_2 + 0.707 \tilde{X}_3. \end{aligned} \quad (9.20)$$

^① 这一技术的矩阵描述在本章附录中给出。

这些主成分列于表 9.13。新变量的方差 - 协方差矩阵为

$$\begin{matrix} & C_1 & C_2 & C_3 \\ \begin{matrix} C_1 \\ C_2 \\ C_3 \end{matrix} & \begin{pmatrix} 1.999 & 0 & 0 \\ 0 & 0.998 & 0 \\ 0 & 0 & 0.003 \end{pmatrix} \end{matrix}.$$

这些主成分没有简单的解释,因为在某种意义上,每一个主成分都是所有原始变量的混合。但是,这些新变量提供了一种统一的、获得多重共线性信息的方法,也是第 10 章中所述的另一种估计技术的基础。

表 9.13 进口数据的主成分 (1949—1959 年)

年份	C_1	C_2	C_3
49	-2.1258	0.6394	-0.0204
50	-1.6189	0.5561	-0.0709
51	-1.1153	-0.0726	-0.0216
52	-0.8944	-0.0821	0.0110
53	-0.6449	-1.3064	0.0727
54	-0.1907	-0.6591	0.0266
55	0.3593	-0.7438	0.0427
56	0.9726	1.3537	0.0627
57	1.5600	0.9635	0.0233
58	1.7677	1.0146	-0.0453
59	1.9304	-1.6633	-0.0809

对于进口数据,较小的值 $\lambda_3 = 0.003$ 指出了多重共线性。本章考虑的其他数据集同样也有提供这种信息的特征根。对于 EEO 数据, $\lambda_1 = 2.952, \lambda_2 = 0.040, \lambda_3 = 0.008$ 。对于广告数据, $\lambda_1 = 1.701, \lambda_2 = 1.288, \lambda_3 = 1.145, \lambda_4 = 0.859, \lambda_5 = 0.007$ 。在每个例子中,出现较小的特征根就表明存在多重共线性。

通过计算相关矩阵的条件数,就可获得各变量总的多重共线性的一种量度。条件数定义为

$$\kappa = \sqrt{\frac{\text{相关矩阵的最大特征根}}{\text{相关矩阵的最小特征根}}} = \sqrt{\frac{\lambda_1}{\lambda_p}}.$$

条件数总是不小于 1 的。较大的条件数则表明强共线性。当条件数超过 15 时(意味着 λ_1 超过 λ_p 的 225 倍),数据中的共线性的负面效应变得很强。EEO、进口和广告三个数据集的条件数分别为 19.20, 25.81 和 15.59。阈值 15 并非出于任何理论上的考虑,而是经验的观察结论。当相关矩阵的条件数超过 30 时,总应该采取修正措施。

判断多重共线性的另一个经验准则为各特征根的倒数和, 即

$$\sum_{j=1}^p \frac{1}{\lambda_j}. \quad (9.21)$$

假如这和大于预测变量数目的 5 倍, 则多重共线性存在。

此类分析还额外提供了一条信息。因为 λ_j 是第 j 个主成分的方差, 若 λ_j 近似为 0, 则相应的主成分 C_j 也就近似为常数。因此, 定义该主成分的方程对导致多重共线性的预测变量间的关系提供了信息。譬如, 在进口数据中, $\lambda_3 = 0.003 \doteq 0$ 。因此, C_3 近似为常数。该常数为 C_3 的均值 0。主成分的均值都为 0, 因为它们都是标准化变量的线性函数, 而每个标准化变量均值都为 0。因此,

$$C_3 = -0.707\tilde{X}_1 - 0.007\tilde{X}_2 + 0.707\tilde{X}_3 \doteq 0.$$

整理各项得

$$\tilde{X}_1 \doteq \tilde{X}_3, \quad (9.22)$$

其中 \tilde{X}_2 的系数 (-0.007) 已被近似为 0。方程 (9.22) 体现了存在于 CONSUM 和 DOPROD 两个变量的标准化形式之间的近似关系。该结果与我们先前看到的、这两者之间有很高的简单相关系数 ($r = 0.997$), 结论是一致的。(读者可考察 CONSUM 关于 DOPROD 的散点图来验证这很高的 r 值。) 因为 λ_3 是唯一一个很小的特征根, 主成分分析告诉我们, 数据中反映出的预测变量间的相关结构不会比方程 (9.22) 中给出的 CONSUM 与 DOPROD 间的简单关系更为复杂。

对于广告数据, 最小的特征根为 $\lambda_5 = 0.007$ 。相应的主成分为

$$C_5 = 0.514\tilde{X}_1 + 0.489\tilde{X}_2 - 0.010\tilde{X}_3 + 0.428\tilde{X}_4 + 0.559\tilde{X}_5. \quad (9.23)$$

令 C_5 为 0, 解出 \tilde{X}_1 , 即导出下列近似关系

$$\tilde{X}_1 \doteq -0.951\tilde{X}_2 - 0.833\tilde{X}_4 - 1.087\tilde{X}_5, \quad (9.24)$$

其中 \tilde{X}_3 的系数被近似为 0。这方程反映了我们早先发现的 A_t, P_t, A_{t-1} 与 P_{t-1} 间的关系。另外, 因为 $\lambda_4 = 0.859$, 且其他 λ 都较大, 我们可以确信, (9.24) 给出的 A_t, P_t, A_{t-1} 与 P_{t-1} 间的关系是数据中多重共线性的唯一来源。

整个这一节, 都是根据各指标量 (或者相关系数或者特征根) 的大小来研究多重共线性是否存在的。尽管我们谈论大或小, 但没办法决定这些阈值。大小是相对的, 也只是给出一个指示: 一切看来正常呢还是有些问题。判断大小唯一合理的准则就是看所发现的多重共线性导致的含糊结果对研究关心的问题是否至关重要。

这里我们也应该小心, 我们分析的数据可能含有一个或少量观测, 它们对共线性的各种量度 (比如, 相关系数, 特征根或条件数) 可能有过度的影响。这些观测称为共线性影响观测。更详细的内容读者可参阅 Hadi (1988)。

9.7 附加约束

我们已经说过,多重共线性是与数据缺陷有关的一种状况,不是模型误设引起的。在考虑多重共线性问题前,我们假定模型形式是经仔细构造的,且残差也是可接受的。因为改善数据通常不切实际,也常常不可能做到,所以我们将注意力集中在那些能比直接运用最小二乘法更好地解释现有数据的方法上。本节中,我们将试图对回归系数的那些有用的线性组合作出鉴别和估计,而不是尽力去解释单个的回归系数。单个回归系数的其他一些估计方法在第10章中讨论。

在转到讨论寻找回归系数的有用线性组合的问题前,关于模型设定的另一点必须先讨论。在设定关系到多重共线性的某关系时,一个精细的步骤是确定回归系数间存在的理论关系。比如,在进口数据的模型

$$\text{IMPORT} = \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_3 \cdot \text{CONSUM} + \varepsilon \quad (9.25)$$

中,人们可能认为 DOPROD 和 CONSUM 的边际效应是相同的。即根据经济学的推理,在看到数据之前,我们就确定了 $\beta_1 = \beta_3$, 或等价地, $\beta_1 - \beta_3 = 0$ 。如同 3.9.3 节中所述, (9.25) 中的模型变为

$$\begin{aligned} \text{IMPORT} &= \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_1 \cdot \text{CONSUM} + \varepsilon \\ &= \beta_0 + \beta_2 \cdot \text{STOCK} + \beta_1(\text{CONSUM} + \text{DOPROD}) + \varepsilon. \end{aligned}$$

这样, β_1 与 β_3 共同的值由 IMPORT 关于 STOCK 及新变量 $\text{NEWVAR} = \text{DOPROD} + \text{CONSUM}$ 的回归来估计。该新变量仅是提取 β_1 与 β_3 的共同估计值一种技术手段,没有别的意义。回归结果列于表 9.14 中。STOCK 和 NEWVAR 这两个预测变量的相关系数为 0.0299, 特征根为 $\lambda_1 = 1.030$ 和 $\lambda_2 = 0.970$ 。其中不再有多重共线性的迹象。残差关于时间、关于拟合值的图(分别为图 9.7 与图 9.8)表明也没有其他模型误设问题。估计的模型为

$$\text{IMPORT} = -9.007 + 0.086 \cdot \text{DOPROD} + 0.612 \cdot \text{STOCK} + 0.086 \cdot \text{CONSUM}.$$

表 9.14 进口数据 (1949–1959 年) 带约束 $\beta_1 = \beta_3$ 的回归结果

变量	系数	标准误	t-检验	p-值
常数	-9.007	1.245	-7.23	< 0.0001
STOCK	0.612	0.109	5.60	0.0005
NEWVAR	0.086	0.004	24.30	< 0.0001
$n = 11$	$R^2 = 0.987$	$R_a^2 = 0.984$	$\hat{\sigma} = 0.5693$	$d.f. = 8$

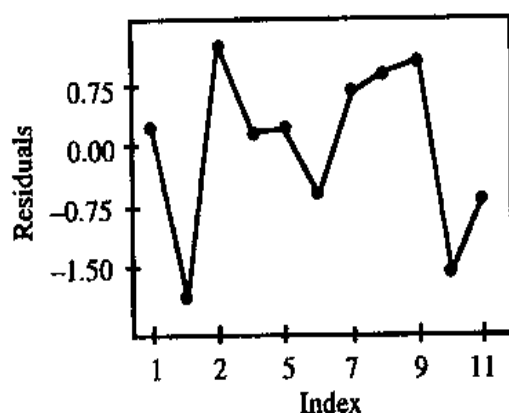


图 9.7 标准化残差的序列图。进口数据 (1949–1959 年) 带约束 $\beta_1 = \beta_3$

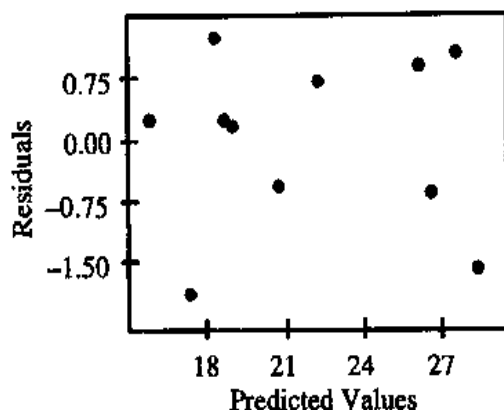


图 9.8 标准化残差关于拟合值的散点图。进口数据 (1949–1959 年) 带约束 $\beta_1 = \beta_3$

注意, 根据 3.9.3 节中概括的方法, 也可能把约束 $\beta_1 = \beta_3$ 作为一个假设来检验。尽管论点 $\beta_1 = \beta_3$ 是建立在已有理论的基础上的, 但评估一下该约束对于全模型解释能力的影响还是有意思的。全模型与约束模型的 R^2 值分别为 0.992 与 0.987。检验 $H_0(\beta_1 = \beta_3)$ 的 F -比为 3.36, 自由度为 1, 8。这两个结果都说明该约束与数据是相符的。

当然, $\beta_1 = \beta_3$ 仅是在设定回归模型时可用的众多约束中的一个例子。所有可能的约束可在第 3 章描述的线性约束集中找到。通常约束的合理性是以潜在的理论为基础的。它们常常可以解决多重共线性带来的问题。另外, 任何一个具体的约束都可视为一个可检验的假设, 并可用第 3 章中描述的方法检验。

9.8 搜寻 β 的线性函数

我们假定模型

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$$

是经精心设定的,因此,其中出现的回归系数在作政策分析与作决策时是首先感兴趣的。我们已经看到,多重共线性的出现可能使单个的 β 无法精确估计出来。然而,如下文所述,精确地估计 β 的某些线性函数常常是做得到的 (Silvey, 1969)。显然的问题有:哪些线性函数可被估计?在那些可估的线性函数中,哪些在分析中有意义?本节我们用数据来帮助鉴别那些能被精确估计的、同时在分析中又有价值的线性函数。

首先,我们间接地说明,总有些 β 的线性函数可被精确地估计^①。再次考虑进口数据。我们已经表明,历史上 CONSUM 和 DOPROD 之间存在近似为 $\text{CONSUM} = (2/3)\text{DOPROD}$ 的关系。把原始模型中的 CONSUM 替换掉,则

$$\text{IMPORT} = \beta_0 + \left(\beta_1 + \frac{2}{3}\beta_3\right) \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \varepsilon. \quad (9.26)$$

等价地说,从方程中去掉 CONSUM 我们能够得到 $\beta_1 + (2/3)\beta_3$ 与 β_2 的精确估计。多重共线性不再存在。DOPROD 与 STOCK 的相关系数为 0.026。结果在表 9.15 中给出。 R^2 几乎没变,残差图(这里未展示)也是令人满意的。在这个案例中,我们使用的数据之外的额外信息论证了在 IMPORT 关于 DOPROD 和 STOCK 的回归中,DOPROD 的系数为线性组合 $\beta_1 + (2/3)\beta_3$ 。同时,我们也说明了,尽管数据中存在多重共线性,但这一线性函数可以被精确地估计。当然, $\beta_1 + (2/3)\beta_3$ 的值是否是有用这是另一个问题,至少知道该回归中 DOPROD 系数之估计衡量的不纯粹是 DOPROD 的边际效应,还包括部分 CONSUM 的效应,这很重要。

表 9.15 对进口数据 (1949–1959 年) 拟合模型 (9.26) 的回归结果

变量	系数	标准误	<i>t</i> -检验	<i>p</i> -值
常数	-8.440	1.432	-5.88	0.0004
DOPROD	0.145	0.007	20.70	< 0.0001
STOCK	0.622	0.128	4.87	0.0012
$n = 11$	$R^2 = 0.983$	$R_a^2 = 0.978$	$\hat{\sigma} = 0.667$	$d.f. = 8$

上面的例子间接地说明了总有些 β 的线性函数可被精确地估计。然而,有一种构造性的方法用来鉴别可被精确估计的 β 的线性组合。我们将用 9.4 节中介绍的广告数据来说明这一方法。其中的概念与本章其他章节中的相比不太直观。我们力求简单。对该问题的规范讨论见本章附录。

我们从 9.6 节中介绍的将标准化的预测变量转换为一组新的正交变量的线性变换入手。五个预测变量的标准化形式记为 $\hat{X}_1, \dots, \hat{X}_5$ 。标准化的响应变量,即

^① 对本问题的进一步的处理请参阅本章附录。

销售额, 记为 \tilde{Y} 。将 X_1, \dots, X_5 转换为·组新的正交变量 C_1, \dots, C_5 的变换为

$$\begin{aligned} C_1 &= +0.532\tilde{X}_1 - 0.232\tilde{X}_2 - 0.389\tilde{X}_3 + 0.395\tilde{X}_4 - 0.595\tilde{X}_5, \\ C_2 &= -0.024\tilde{X}_1 + 0.825\tilde{X}_2 - 0.022\tilde{X}_3 - 0.260\tilde{X}_4 - 0.501\tilde{X}_5, \\ C_3 &= -0.668\tilde{X}_1 + 0.158\tilde{X}_2 - 0.217\tilde{X}_3 + 0.692\tilde{X}_4 - 0.057\tilde{X}_5, \\ C_4 &= +0.074\tilde{X}_1 - 0.037\tilde{X}_2 + 0.895\tilde{X}_3 + 0.338\tilde{X}_4 - 0.279\tilde{X}_5, \\ C_5 &= -0.514\tilde{X}_1 - 0.489\tilde{X}_2 + 0.010\tilde{X}_3 - 0.428\tilde{X}_4 - 0.559\tilde{X}_5. \end{aligned} \quad (9.27)$$

定义 C_1 的方程中, 各系数是属于预测变量相关矩阵的最大特征根的特征向量的分量。类似地, 定义 C_2 至 C_5 的系数依次是属于其余从大到小的特征根的特征向量的分量。如前面 9.6 节所述, 变量 C_1, \dots, C_5 是预测变量之标准化形式的主成分。

如 (9.4) 那样, 以原始变量给出的回归模型为

$$S_t = \beta_0 + \beta_1 A_t + \beta_2 P_t + \beta_3 E_t + \beta_4 A_{t-1} + \beta_5 P_{t-1} + \varepsilon_t. \quad (9.28)$$

用标准化的变量方程可写为

$$\tilde{Y} = \theta_1 \tilde{A}_t + \theta_2 \tilde{P}_t + \theta_3 \tilde{E}_t + \theta_4 \tilde{A}_{t-1} + \theta_5 \tilde{P}_{t-1} + \varepsilon', \quad (9.29)$$

其中 \tilde{A}_t 表示变量 A_t 的标准化形式。方程 (9.29) 中的回归系数通常称为 *beta* 系数。它们代表各预测变量变动一个标准差时的边际效应。譬如, θ_1 表示, 广告费用 (A) 增加一个标准差时相应的销售额 (S) 的增量, 单位是销售额的标准差。

在对数据拟合模型 (9.28) 时, 记 $\hat{\beta}_j$ 为 β_j 的最小二乘估计。类似地, 记 $\hat{\theta}_j$ 为拟合模型 (9.29) 时得到的 θ_j 的最小二乘估计。 $\hat{\beta}_j$ 与 $\hat{\theta}_j$ 之间的关系为

$$\begin{aligned} \hat{\beta}_j &= (s_y/s_j)\hat{\theta}_j, \quad j = 1, \dots, 5, \\ \hat{\beta}_0 &= \bar{y} - \sum_{j=1}^5 \hat{\beta}_j \bar{x}_j, \end{aligned} \quad (9.30)$$

其中 \bar{y} 为 Y 的均值, s_y, s_j 分别为响应变量和第 j 个预测变量的标准差。

方程 (9.29) 有一个等价的形式

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \alpha_5 C_5 + \varepsilon'. \quad (9.31)$$

方程 (9.29) 与 (9.31) 的等价性是由方程 (9.27) 中各 \tilde{X} 与 C 间的关系导出的, 而各 α 与 θ 的关系以及它们的估计值 $\hat{\alpha}$ 与 $\hat{\theta}$ 的关系如下

$$\begin{aligned} \hat{\theta}_1 &= +0.532\hat{\alpha}_1 - 0.024\hat{\alpha}_2 - 0.668\hat{\alpha}_3 + 0.074\hat{\alpha}_4 - 0.514\hat{\alpha}_5, \\ \hat{\theta}_2 &= -0.232\hat{\alpha}_1 + 0.825\hat{\alpha}_2 + 0.158\hat{\alpha}_3 - 0.037\hat{\alpha}_4 - 0.489\hat{\alpha}_5, \\ \hat{\theta}_3 &= -0.389\hat{\alpha}_1 - 0.022\hat{\alpha}_2 - 0.217\hat{\alpha}_3 + 0.895\hat{\alpha}_4 + 0.010\hat{\alpha}_5, \\ \hat{\theta}_4 &= +0.395\hat{\alpha}_1 - 0.260\hat{\alpha}_2 + 0.692\hat{\alpha}_3 + 0.338\hat{\alpha}_4 - 0.428\hat{\alpha}_5, \\ \hat{\theta}_5 &= -0.595\hat{\alpha}_1 - 0.501\hat{\alpha}_2 - 0.057\hat{\alpha}_3 - 0.279\hat{\alpha}_4 - 0.559\hat{\alpha}_5. \end{aligned} \quad (9.32)$$

注意该变换涉及的权重与定义方程 (9.27) 的相同。变换后的模型优点在于各主成分正交的。用各 $\hat{\alpha}$ 的方差表示的回归系数估计的精度很容易计算。 $\hat{\alpha}_j$ 的方差估计为 $\hat{\sigma}^2/\lambda_j$, 与第 j 个特征根成反比。除 $\hat{\alpha}_5$ 外其他 α 均可被精确估计, 因为只有 λ_5 较小。(回顾一下, $\lambda_1 = 1.701, \lambda_2 = 1.288, \lambda_3 = 1.145, \lambda_4 = 0.859, \lambda_5 = 0.007$ 。)

我们对于各 $\hat{\alpha}$ 的兴趣仅在于拿它们作为分析 $\hat{\theta}$ 的工具。根据方程 (9.32), 计算、分析各 $\hat{\theta}_j$ 的方差及标准误是件简单的事。 $\hat{\theta}_j$ 之方差为

$$Var(\hat{\theta}_j) = \sum_{i=1}^p v_{ij}^2 Var(\hat{\alpha}_i), \quad j = 1, \dots, p, \quad (9.33)$$

其中 v_{ij} 为式 (9.32) 第 j 个方程中 $\hat{\alpha}_i$ 的系数。因为 $\hat{\alpha}_i$ 的方差估计为 $\hat{\sigma}^2/\lambda_i$, $\hat{\sigma}^2$ 为残差均方, 所以式 (9.33) 的估计为

$$\widehat{Var}(\hat{\theta}_j) = \hat{\sigma}^2 \sum_{i=1}^p \frac{v_{ij}^2}{\lambda_i}. \quad (9.34)$$

譬如, $\hat{\theta}_1$ 的方差估计为

$$\hat{\sigma}^2 \left[\frac{(+0.532)^2}{\lambda_1} + \frac{(-0.024)^2}{\lambda_2} + \frac{(-0.668)^2}{\lambda_3} + \frac{(+0.074)^2}{\lambda_4} + \frac{(-0.514)^2}{\lambda_5} \right]. \quad (9.35)$$

记得有 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_5$ 且仅 λ_5 很小 (为 0.007)。因此, 方差表达式中仅最后一项大且可能破坏 $\hat{\theta}_1$ 的精度。因为其他 $\hat{\theta}_j$ 的方差表达式类似于公式 (9.35), 所以要求方差小等价于要求 $1/\lambda_5$ 的系数小。细看把 $\{\hat{\alpha}_i\}$ 变换至 $\{\hat{\theta}_j\}$ 的各个方程, 我们发现 $\hat{\theta}_3$ 是最精确的估计, 因为其方差表达式中 $1/\lambda_5$ 的系数为 $(-0.01)^2 = 0.0001$ 。

把这一类型的分析推而广之, 就有可能鉴别出各 θ 的有意义的线性函数, 它们的估计精度可以比单个的 θ 更高。譬如, 我们可能对估计 $\theta_1 - \theta_2$ 比估计单个的 θ_1, θ_2 更感兴趣。在销售额模型中, $\theta_1 - \theta_2$ 度量的是当年的广告预算 X_1 增加一个单位、同时当年的促销预算 X_2 削减一个单位, 相应的销售额的增量。换句话说, $\theta_1 - \theta_2$ 表示当年资源使用分配变动的效应。 $\theta_1 - \theta_2$ 的估计为 $\hat{\theta}_1 - \hat{\theta}_2$ 。该估计的方差可如此方便地获得: 拿式 (9.32) 中 $\hat{\theta}_1$ 与 $\hat{\theta}_2$ 的方程相减, 再像前面那样根据算得的各 $\hat{\alpha}$ 的系数去算方差。即

$$\hat{\theta}_1 - \hat{\theta}_2 = 0.764\hat{\alpha}_1 - 0.849\hat{\alpha}_2 - 0.826\hat{\alpha}_3 + 0.111\hat{\alpha}_4 - 0.025\hat{\alpha}_5.$$

由此可知 $\hat{\theta}_1 - \hat{\theta}_2$ 的方差为

$$\begin{aligned} & (0.764)^2 Var(\hat{\alpha}_1) + (-0.849)^2 Var(\hat{\alpha}_2) + (-0.826)^2 Var(\hat{\alpha}_3) \\ & + (0.111)^2 Var(\hat{\alpha}_4) + (-0.025)^2 Var(\hat{\alpha}_5), \end{aligned} \quad (9.36)$$

方差的估计为

$$\hat{\sigma}^2 \left[\frac{(0.764)^2}{\lambda_1} + \frac{(-0.849)^2}{\lambda_2} + \frac{(-0.826)^2}{\lambda_3} + \frac{(0.111)^2}{\lambda_4} + \frac{(-0.025)^2}{\lambda_5} \right]. \quad (9.37)$$

$1/\lambda_5$ 的系数较小, 使得可能把 $\theta_1 - \theta_2$ 估计准。将此程序一般化, 我们看到, θ 的任何线性函数, 只要其方差表达式中 $1/\lambda_5$ 的系数较小, 那都可能被准确地估计。

9.9 使用主成分作计算

作此项分析需要的计算,除了标准的最小二乘计算程序外还涉及别的程序。原始数据须经一个主成分子程序来处理,该子程序对预测变量的相关矩阵计算特征根以及如同式(9.32)中的变换权重。绝大多数回归软件包将各 beta 系数的估计作为标准输出的一部分。

表 9.16 拟合模型 (9.29) 获得的回归结果

变量	系数	标准误	t-检验	p-值
\bar{X}_1	0.583	0.438	1.33	0.2019
\bar{X}_2	0.973	0.417	2.33	0.0329
\bar{X}_3	0.786	0.075	10.50	< 0.0001
\bar{X}_4	0.395	0.367	1.08	0.2973
\bar{X}_5	0.503	0.476	1.06	0.3053
$n = 22$	$R^2 = 0.917$	$R_a^2 = 0.891$	$\hat{\sigma} = 0.3303$	$d.f. = 16$

表 9.17 拟合模型 (9.31) 获得的回归结果

变量	系数	标准误	t-检验	p-值
C_1	-0.346	0.053	-6.55	< 0.0001
C_2	0.418	0.064	6.58	< 0.0001
C_3	-0.151	0.067	-2.25	0.0391
C_4	0.660	0.078	8.46	< 0.0001
C_5	-1.220	0.846	-1.44	0.1683
$n = 22$	$R^2 = 0.917$	$R_a^2 = 0.891$	$\hat{\sigma} = 0.3303$	$d.f. = 16$

对于广告数据,估计值 $\hat{\theta}_1, \dots, \hat{\theta}_5$ 可由两种等价的方法来计算。它们可以直接由方程(9.29)中的标准化变量的回归获得。该回归结果在表 9.16 中给出。另外,我们可以用最小二乘法拟合(9.31)中标准化的响应变量关于 5 个主成分的回归模型,得到估计值 $\hat{\alpha}_1, \dots, \hat{\alpha}_5$ 。该回归的结果呈现在表 9.17 中。于是,我们再用式(9.32)获得 $\hat{\theta}_1, \dots, \hat{\theta}_5$ 。譬如,

$$\begin{aligned}\hat{\theta}_1 &= (0.532)(-0.346019) + (-0.024)(0.417889) + (-0.668)(-0.151328) \\ &\quad + (0.074)(0.659946) + (-0.514)(-1.22026) = 0.5830.\end{aligned}$$

使用式(9.32)中的系数,可算得 $\hat{\theta}_1, \dots, \hat{\theta}_5$ 的标准误。譬如, $\hat{\theta}_1$ 的方差估计为

$$\begin{aligned}& (0.532 \times s.e.(\hat{\alpha}_1))^2 + (-0.024 \times s.e.(\hat{\alpha}_2))^2 + (-0.668 \times s.e.(\hat{\alpha}_3))^2 \\ & \quad + (0.074 \times s.e.(\hat{\alpha}_4))^2 + (-0.514 \times s.e.(\hat{\alpha}_5))^2 \\ &= (0.532 \times 0.0529)^2 + (-0.024 \times 0.0635)^2 + (-0.668 \times 0.0674)^2 \\ & \quad + (0.074 \times 0.0780)^2 + (-0.514 \times 0.8456)^2 = 0.1918,\end{aligned}$$

因而 $\hat{\theta}_1$ 的标准误为

$$s.e.(\hat{\theta}_1) = \sqrt{0.1918} = 0.438.$$

应该注意, 检验 $\beta_j = 0$ 与 $\theta_j = 0$ 的 t -值是相同的。beta 系数 θ_j 与 β_j 是成比例的。在用 $\hat{\beta}_j/s.e.(\hat{\beta}_j)$ 或 $\hat{\theta}_j/s.e.(\hat{\theta}_j)$ 构造 t -值时, 比例因子消掉了。

$\theta_1 - \theta_2$ 的估计值为 $0.583 - 0.973 = -0.390$ 。 $\hat{\theta}_1 - \hat{\theta}_2$ 的方差估计值可由公式 (9.36) 算得, 为 0.008。 $\theta_1 - \theta_2$ 的一个 95% 置信区间为 $-0.390 \pm 2.12\sqrt{0.008}$ 或 $(-0.58, -0.20)$ 。即若当年将一个单位的促销开支改为广告, 其效果是销售额损失 0.20 至 0.58 个标准化的单位。

还有其他可被精确地估计的线性函数, 任何方差表达式中 $1/\lambda_5$ 的系数较小的线性函数都是。比如, 方程组 (9.32) 暗示着 $\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_4, \hat{\theta}_5$ 两两之间的差异均可被估准。然而, 在这个问题中, 有些差异是有意义的, 而另一些没意义。比如, 如前面所述, 差异 $\theta_1 - \theta_2$ 是有意义的。它表示把当期一个单位的促销开支改为广告的效应。但差异 $\theta_1 - \theta_4$ 显然没有意义。它表示把当年一个单位的广告开支放到前一年的广告开支中去的效应。资源的分配在时间上逆转是不可能的。因此, 尽管 $\theta_1 - \theta_4$ 可被估准, 但它在销售额的分析中没有意义。

一般地, 当列出了方程组 (9.32) 中的权重并知道了相应的特征根, 总是可能通过审视其权重, 来鉴别出那些可被估准的原始回归系数的线性函数。在可估准的线性函数中, 只有一部分对于研究的问题有意义。

综上所述, 在有多重共线性的征兆且不可能补充数据的场合, 仍然可能估准某些回归系数和某些线性函数。为调查哪些系数和哪些线性函数可估, 我们推荐刚刚描述的 (变换为主成分的) 分析。这种分析方法不能克服多重共线性, 假如它存在的话。仍有一些回归系数及回归系数的函数不可估。但这里推荐的分析将指明那些可估的函数, 并指出预测变量间的结构上的相依性。

9.10 文献

本章所用的主成分技术在绝大多数多元统计分析的书书中都有推导。应当注意, 主成分分析涉及的仅仅是预测变量。分析目的在于刻画及识别预测变量间的相依性 (假如存在的话)。关于主成分的全面讨论, 读者可参阅 Johnson and Wichern (1992) 或 Seber (1984)。目前一些商业化的统计软件包提供了本章所述的分析。

习 题

9.1 在 9.4 节对广告数据的分析中, 我们建议, 销售额 S_t 关于 E_t 及其余四个变量 ($A_t, P_t, A_{t-1}, P_{t-1}$) 中的三个的回归可能解决共线性问题。作这四个我们建议的回归, 对其中的每一个, 检查各 VIF_j , 看是否消除了共线性。

9.2 汽油消耗: 为研究决定汽车的汽油消耗的因素, 收集了 30 种型号的汽车数据。其中包括每辆汽车的汽油消耗量 (Y), 以英里/加仑为单位, 以及另 11 个

反映物理、机械特征的变量。表 9.19 中数据的来源是 1975 年的 *Motor Trend* 杂志。变量的定义在表 9.18 中给出。我们希望判断该数据集是否共线性。

- (a) 计算预测变量 X_1, \dots, X_{11} 的相关矩阵, 并作相应的两两散点图。识别任何共线性的证据。
- (b) 计算相关矩阵的特征根、特征向量以及条件数。数据中是否存在多重共线性?
- (c) 考察属于较小特征根的特征向量, 识别多重共线性与哪些变量有关。
- (d) 作 Y 关于 11 个预测变量的回归, 计算每个预测因子的 VIF。哪些预测因子受到了共线性的影响?

表 9.18 表 9.19 汽油消耗数据中的变量

变量	定义
Y	英里/加仑
X_1	排气量 (立方英寸)
X_2	马力
X_3	扭矩 (英尺·磅)
X_4	压缩比
X_5	后轴动力比
X_6	化油器 (筒形)
X_7	变速档数
X_8	整体长度 (英寸)
X_9	宽度 (英寸)
X_{10}	重量 (磅)
X_{11}	传动类型 (1 = 自动; 0 = 手动)

9.3 参阅表 5.17 中的总统选举数据, 考虑拟合一个 V 关于所有变量 (包括一个表示选举年份的时间趋势) 以及尽可能多的二因子、三因子交互作用项。

- (a) 对这些数据你能拟合的线性回归模型中项 (系数) 数最多为几个? [提示: 考虑数据中观测的个数。]
- (b) 考察上述模型中预测变量间是否存在多重共线性。(计算相关矩阵、条件数及各 VIF。)
- (c) 识别与共线性有关的变量子集。尝试通过删去与多重共线性有关的变量中的某一些来解决多重共线性问题。
- (d) 拟合 V 关于你所发现的无多重共线性的预测变量集的回归模型。

表 9.19 汽油消耗与汽车变量

Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
18.9	350.0	165	260	8.00	2.56	4	3	200.3	69.9	3910	1
17.0	350.0	170	275	8.50	2.56	4	3	199.6	72.9	3860	1
20.0	250.0	105	185	8.25	2.73	1	3	196.7	72.2	3510	1
18.3	351.0	143	255	8.00	3.00	2	3	199.9	74.0	3890	1
20.1	225.0	95	170	8.40	2.76	1	3	194.1	71.8	3365	0
11.2	440.0	215	330	8.20	2.88	4	3	184.5	69.0	4215	1
22.1	231.0	110	175	8.00	2.56	2	3	179.3	65.4	3020	1
21.5	262.0	110	200	8.50	2.56	2	3	179.3	65.4	3180	1
34.7	89.7	70	81	8.20	3.90	2	4	155.7	64.0	1905	0
30.4	96.9	75	83	9.00	4.30	2	5	165.2	65.0	2320	0
16.5	350.0	155	250	8.50	3.08	4	3	195.4	74.4	3885	1
36.5	85.3	80	83	8.50	3.89	2	4	160.6	62.2	2009	0
21.5	171.0	109	146	8.20	3.22	2	4	170.4	66.9	2655	0
19.7	258.0	110	195	8.00	3.08	1	3	171.5	77.0	3375	1
20.3	140.0	83	109	8.40	3.40	2	4	168.8	69.4	2700	0
17.8	302.0	129	220	8.00	3.00	2	3	199.9	74.0	3890	1
14.4	500.0	190	360	8.50	2.73	4	3	224.1	79.8	5290	1
14.9	440.0	215	330	8.20	2.71	4	3	231.0	79.7	5185	1
17.8	350.0	155	250	8.50	3.08	4	3	196.7	72.2	3910	1
16.4	318.0	145	255	8.50	2.45	2	3	197.6	71.0	3660	1
23.5	231.0	110	175	8.00	2.56	2	3	179.3	65.4	3050	1
21.5	360.0	180	290	8.40	2.45	2	3	214.2	76.3	4250	1
31.9	96.9	75	83	9.00	4.30	2	5	165.2	61.8	2275	0
13.3	460.0	223	366	8.00	3.00	4	3	228.0	79.8	5430	1
23.9	133.6	96	120	8.40	3.91	2	5	171.5	63.4	2535	0
19.7	318.0	140	255	8.50	2.71	2	3	215.3	76.3	4370	1
13.9	351.0	148	243	8.00	3.25	2	3	215.5	78.5	4540	1
13.3	351.0	148	243	8.00	3.26	2	3	216.1	78.5	4715	1
13.8	360.0	195	295	8.25	3.15	4	3	209.3	77.4	4215	1
16.5	350.0	165	255	8.50	2.73	4	3	185.2	69.0	3660	1

附录：主成分

在本附录中，我们采用矩阵记号介绍主成分方法去检测多重共线性。

A. 模型

回归模型可表达为

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (\text{A.1})$$

其中 \mathbf{Y} 为 $n \times 1$ 的响应变量的观测向量， $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_p)$ 是一个 $n \times p$ 阶的矩

阵, 包括 p 个预测变量的 n 个观测, θ 是 $p \times 1$ 的回归系数向量, ε 是 $n \times 1$ 的随机误差向量。假定 $E(\varepsilon) = 0, E(\varepsilon\varepsilon^T) = \sigma^2\mathbf{I}$, 其中 \mathbf{I} 为 n 阶单位阵。不失一般性, 再假定 \mathbf{Y}, \mathbf{Z} 均已经中心化和尺度变换, 以致 $\mathbf{Z}^T\mathbf{Z}$ 与 $\mathbf{Z}^T\mathbf{Y}$ 都是相关系数矩阵。

存在方阵 \mathbf{A} 和 \mathbf{V} , 满足^①

$$\mathbf{V}^T(\mathbf{Z}^T\mathbf{Z})\mathbf{V} = \mathbf{A} \text{ 与 } \mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}. \quad (\text{A.2})$$

\mathbf{A} 是一个对角阵, 对角线上依次为 $\mathbf{Z}^T\mathbf{Z}$ 的从大到小的特征根。这些特征根记为 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ 。 \mathbf{V} 的各列依次为属于 $\lambda_1, \cdots, \lambda_p$ 的正则化的特征向量。因为 $\mathbf{V}\mathbf{V}^T = \mathbf{I}$, 所以 (A.1) 中的回归模型用主成分可表达为

$$\mathbf{Y} = \mathbf{Z}\mathbf{V}\mathbf{V}^T\theta + \varepsilon = \mathbf{C}\alpha + \varepsilon, \quad (\text{A.3})$$

其中

$$\mathbf{C} = \mathbf{Z}\mathbf{V}, \quad \alpha = \mathbf{V}^T\theta. \quad (\text{A.4})$$

矩阵 \mathbf{C} 含 p 个列 $\mathbf{C}_1, \cdots, \mathbf{C}_p$, 每个列都是预测变量 $\mathbf{Z}_1, \cdots, \mathbf{Z}_p$ 的线性组合。 \mathbf{C} 的各列是正交的, 称为预测变量 $\mathbf{Z}_1, \cdots, \mathbf{Z}_p$ 的主成分。 \mathbf{C} 的各列满足 $\mathbf{C}_i^T\mathbf{C}_j = \lambda_j$, 及 $i \neq j$ 时, $\mathbf{C}_i^T\mathbf{C}_j = 0$ 。

这些主成分和特征根可用于检测并分析预测变量中的共线性。方程 (A.3) 中给出的回归模型, 是用正交的预测变量对方程 (A.1) 的重新参数化。各 λ 可视为各主成分的样本方差。若 $\lambda_i = 0$, 则第 i 个主成分的所有观测也都为 0。因为第 j 个主成分是 $\mathbf{Z}_1, \cdots, \mathbf{Z}_p$ 的线性函数, 所以, 当 $\lambda_j = 0$ 时, 预测变量间存在一种精确的线性相依关系。于是, 当 λ_j 很小 (近似等于 0) 时, 预测变量间存在一种近似的线性关系。即一个小的特征根是多重共线性的一种标示。另外, 从公式 (A.4) 我们得到

$$\mathbf{C}_j = \sum_{i=1}^p v_{ij}\mathbf{Z}_i.$$

这确定了导致多重共线性的线性关系的精确形式。

B. $\hat{\theta}$ 的线性函数的精度

分别记 $\hat{\alpha}, \hat{\theta}$ 为 α 与 θ 的最小二乘估计。可以证明 $\hat{\alpha} = \mathbf{V}^T\hat{\theta}$, 以及反过来, $\hat{\theta} = \mathbf{V}\hat{\alpha}$ 。由 $\hat{\alpha} = (\mathbf{C}^T\mathbf{C})^{-1}\mathbf{C}^T\mathbf{Y}$ 知, $\hat{\alpha}$ 的方差-协方差矩阵为 $V(\hat{\alpha}) = \mathbf{A}^{-1}\sigma^2$, 及 $\hat{\theta}$ 的方差-协方差为 $V(\hat{\theta}) = \mathbf{V}\mathbf{A}^{-1}\mathbf{V}^T\sigma^2$ 。记 \mathbf{L} 为任一 $p \times 1$ 常数向量。线性函数 $\delta = \mathbf{L}^T\theta$ 具有最小二乘估计 $\hat{\delta} = \mathbf{L}^T\hat{\theta}$, 其方差为

$$\text{Var}(\hat{\delta}) = \mathbf{L}^T\mathbf{V}\mathbf{A}^{-1}\mathbf{V}^T\mathbf{L}\sigma^2. \quad (\text{A.5})$$

记 \mathbf{V}_j 为 \mathbf{V} 之第 j 列。于是选择合适的常数 r_1, \cdots, r_p , \mathbf{L} 可表为

$$\mathbf{L} = \sum_{j=1}^p r_j\mathbf{V}_j.$$

^① 譬如, 可参阅 Strang (1988) 或 Hadi (1996)。

于是 (A.5) 变为 $Var(\hat{\delta}) = \mathbf{R}^T \mathbf{\Lambda}^{-1} \mathbf{R} \sigma^2$, 或等价地,

$$Var(\hat{\delta}) = \left(\sum_{j=1}^p \frac{r_j^2}{\lambda_j} \right) \sigma^2, \quad (\text{A.6})$$

其中 $\mathbf{R} = (r_1, \dots, r_p)^T$ 。

总之, $\hat{\delta}$ 的方差是各特征根倒数的线性组合。因此, 如果没有特征根接近于 0 或在 λ_j 较小时, r_j^2 与 λ_j 有相同的数量级, 那么 $\hat{\delta}$ 会有很好的精度。而且, 总是可能选择一个向量 \mathbf{L} , 使 $\hat{\theta}$ 的一个线性函数 $\mathbf{L}^T \hat{\theta}$, 消除了较小的特征根的效应, 从而有小的方差。对这些概念更完整的描述, 可参阅 Silvey (1969)。

10

回归系数的有偏估计

10.1 引言

在第 9 章我们论证了，当预测变量间存在多重共线性关系时，单个回归系数的普通最小二乘估计就会趋于不稳定，并可能导致错误的推断。本章介绍另两种估计方法，在出现多重共线性的情况下，与普通最小二乘方法 (OLS) 相比，它们能更为有效地分析数据。尽管这里讨论的估计量是有偏的，但精度（用均方误差来度量）却往往优于 OLS 估计（参见 Draper and Smith (1998), McCallum (1970) 和 Hoerl and Kennard (1970)）。这些方法拟合数据不如 OLS 法好，残差平方和没那么小、复相关系数也没那么大。然而，这两种方法的潜力在于，得到的系数估计精度更高、在对新的数据作预测时预测误差更小。

遗憾的是，判断这些方法得到的结果是否优于 OLS 法的准则依赖于模型中真实却未知的回归系数值。也就是说，没有完全客观的办法去判断究竟何时该用这其中的某一种方法去替代 OLS 法。然而，在怀疑共线性存在时，我们建议采用这些方法。得到的回归系数的估计可能会对数据给出一种新的解释，相应地，也能对所研究的过程给出更好的理解。

这里我们讨论的用于替代 OLS 的两种方法是：(1) 主成分回归；(2) 岭回归。主成分分析在第 9 章中介绍过，我们假定读者已熟悉这部分内容。接着将要阐述的是，可从两个角度解释主成分估计方法：一种解释与预测变量的非正交性有关；另一种解释与对回归系数所加的约束有关。岭回归也与对系数所加的约束有关。岭方法在本章中介绍，在第 11 章又将被应用于变量选择问题。我们将采用第 9 章中分析的法国进口数据来考察主成分回归和岭回归这两种方法。

10.2 主成分回归

我们考虑的模型为

$$\text{IMPORT} = \beta_0 + \beta_1 \cdot \text{DOPROD} + \beta_2 \cdot \text{STOCK} + \beta_3 \cdot \text{CONSUM} + \varepsilon, \quad (10.1)$$

这些变量的定义见 9.3 节。令 \bar{y} 、 \bar{x}_j 分别为 Y 和 X_j 的均值, 再令 s_y 、 s_j 分别为 Y 和 X_j 的标准差。模型 (10.1) 可用标准化的变量表示为 (见 9.5 节)

$$\tilde{Y} = \theta_1 \tilde{X}_1 + \theta_2 \tilde{X}_2 + \theta_3 \tilde{X}_3 + \varepsilon', \quad (10.2)$$

其中 $\tilde{Y} = (y_i - \bar{y})/s_y$ 是响应变量 (IMPORT) 的标准化, $\tilde{X}_j = (x_j - \bar{x}_j)/s_j$ 是第 j 个预测变量的标准化。许多回归软件包既给出 (10.1) 中的常规的回归系数估计值, 也给出 (10.2) 中标准化的回归系数估计值。系数估计值满足

$$\begin{aligned} \beta_j &= (s_y/s_j)\theta_j, \quad j = 1, 2, 3, \\ \beta_0 &= \bar{y} - \beta_1 \bar{x}_1 - \beta_2 \bar{x}_2 - \beta_3 \bar{x}_3. \end{aligned} \quad (10.3)$$

标准化预测变量的主成分为 (见方程 (9.20))

$$\begin{aligned} C_1 &= 0.706\tilde{X}_1 + 0.044\tilde{X}_2 + 0.707\tilde{X}_3, \\ C_2 &= -0.036\tilde{X}_1 + 0.999\tilde{X}_2 - 0.026\tilde{X}_3, \\ C_3 &= -0.707\tilde{X}_1 - 0.007\tilde{X}_2 + 0.707\tilde{X}_3. \end{aligned} \quad (10.4)$$

这些主成分已在表 9.13 中给出。模型 (10.2) 可用主成分表示为

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \varepsilon'. \quad (10.5)$$

(10.2) 与 (10.5) 是等价的, 因为各 α 与各 θ 之间是一一对应的。具体地

$$\begin{aligned} \alpha_1 &= 0.706\theta_1 + 0.044\theta_2 + 0.707\theta_3, \\ \alpha_2 &= -0.036\theta_1 + 0.999\theta_2 - 0.026\theta_3, \\ \alpha_3 &= -0.707\theta_1 - 0.007\theta_2 + 0.707\theta_3. \end{aligned} \quad (10.6)$$

反过来,

$$\begin{aligned} \theta_1 &= 0.706\alpha_1 - 0.036\alpha_2 - 0.707\alpha_3, \\ \theta_2 &= 0.044\alpha_1 + 0.999\alpha_2 - 0.007\alpha_3, \\ \theta_3 &= 0.707\alpha_1 - 0.026\alpha_2 + 0.707\alpha_3. \end{aligned} \quad (10.7)$$

这些关系对各 α 、 θ 的最小二乘估计 $\hat{\alpha}$ 、 $\hat{\theta}$ 也同样成立。因此, 各 $\hat{\alpha}$ 、 $\hat{\theta}$ 可由 Y 关于主成分 C_1, C_2, C_3 的回归得到, 或者由 Y 关于各标准化变量的回归得到。对进口数据拟合模型 (10.2) 与 (10.5) 的回归结果见表 10.1 及表 10.2。从表 10.1 知, $\theta_1, \theta_2, \theta_3$ 的估计值分别为 $-0.339, 0.213, 1.303$ 。类似地, 由表 10.2 知, $\alpha_1, \alpha_2, \alpha_3$ 的估计值分别为 $0.690, 0.191, 1.160$ 。这两个表中的任意一个可根据 (10.6) 或 (10.7) 由另一个表得到。

尽管方程 (10.2) 和 (10.5) 是等价的, 但 (10.5) 中各 C 变量是正交的。而且, 应注意到, 用主成分 (方程 (10.5)) 给出的回归关系是不容易解释的。那个模型中的预测变量是原始预测变量的线性组合。不像各 θ 那样, 各 α 不再能简单地解释为各原始预测变量的边际效应。因此, 我们仅仅把主成分回归用作分析多重共线性问题的一种手段。最终的估计结果通常还是转化为用各 θ 来表述, 以便于解释。

表 10.1 对进口数据 (1949–1959 年) 拟合模型 (10.2) 得到的回归结果

变量	系数	标准误	t -检验	p -值
\bar{X}_1	-0.339	0.464	-0.73	0.4883
\bar{X}_2	0.213	0.034	6.20	0.0004
\bar{X}_3	1.303	0.464	2.81	0.0263
$n = 11$	$R^2 = 0.992$	$R_a^2 = 0.988$	$\hat{\sigma} = 0.034$	$d.f. = 7$

表 10.2 对进口数据 (1949–1959 年) 拟合模型 (10.5) 得到的回归结果

变量	系数	标准误	t -检验	p -值
C_1	0.690	0.024	28.70	< 0.0001
C_2	0.191	0.034	5.62	0.0008
C_3	1.160	0.656	1.77	0.1204
$n = 11$	$R^2 = 0.992$	$R_a^2 = 0.988$	$\hat{\sigma} = 0.034$	$d.f. = 7$

10.3 消除预测变量间的相依性

前面提到过, 主成分回归有两种解释。我们先用主成分技术消除估计数据中的多重共线性。只使用部分主成分去解释响应变量的变异, 就能达到消除共线性的目的。注意到, 当所有三个主成分都使用时, OLS 解就可以通过方程组 (10.7) 被精确地恢复出来。

各 C 变量分别具有样本方差 $\lambda_1 = 1.999, \lambda_2 = 0.998, \lambda_3 = 0.003$ 。回想一下, 各 λ 是 DOPROD, STOCK 和 CONSUM 三个变量的相关矩阵的特征根。因为 C_3 的方差为 0.003, 所以定义 C_3 的线性函数接近于 0, 这也就是数据中多重共线性的来源。我们去掉 C_3 , 考虑 \tilde{Y} 与单个 C_1 的回归以及与 C_1, C_2 的回归, 即考虑两种可能的回归模型

$$\tilde{Y} = \alpha_1 C_1 + \varepsilon \quad (10.8)$$

与

$$\tilde{Y} = \alpha_1 C_1 + \alpha_2 C_2 + \varepsilon. \quad (10.9)$$

由这两个模型都可导出对所有三个原始系数 $\theta_1, \theta_2, \theta_3$ 的估计。这些估计是有偏的, 因为在这两种情况下, 都有一些信息被丢掉了 (方程 (10.9) 中丢掉了 C_3 , 方程 (10.8) 中丢掉了 C_2 和 C_3)。

α_1 或 α_1, α_2 的估计值可依次由 \tilde{Y} 关于 C_1 以及关于 C_1, C_2 的回归获得。然而, 利用 C_1, C_2, C_3 之间的正交性, 可获得一种更简单的计算方法^①。譬如, α_1 的估计值可由 (10.5), (10.8) 或 (10.9) 的回归得到。类似地, α_2 的估计值可由 (10.5) 或 (10.9) 的回归得到。另外, 如果有了各 θ 的 OLS 估计值, 那么各 α 的估计值可由方程 (10.6) 得到。于是, 相应于 (10.8)、(10.9) 的各 θ 的主成分回归估计, 可反

① 对于任何回归方程, 若所考虑的全部可能的预测变量之间是正交的, 那么, 其中的部分变量引进或删除时, 回归系数的估计值都不会改变。

过来由方程 (10.7) 算得, 只要将其中相应的一些 α 的值置为 0。下列例子说明了这个步骤。

在方程 (10.7) 中代入 $\alpha_1 = 0.690, \alpha_2 = \alpha_3 = 0$, 就可算出只用第一主成分作回归得到的各 θ 的估计, 即

$$\begin{aligned}\hat{\theta}_1 &= 0.706 \times 0.690 = 0.487, \\ \hat{\theta}_2 &= 0.044 \times 0.690 = 0.030, \\ \hat{\theta}_3 &= 0.707 \times 0.690 = 0.487.\end{aligned}\quad (10.10)$$

于是得到回归方程

$$\tilde{Y} = 0.487\tilde{X}_1 + 0.030\tilde{X}_2 + 0.487\tilde{X}_3.$$

像 (10.9) 那样使用前两个主成分得到的估计, 亦可如法炮制, 将 $\alpha_1 = 0.690, \alpha_2 = 0.191, \alpha_3 = 0$ 代入 (10.7) 算出。方程 (10.1) 中原始变量之回归系数 $\beta_0, \beta_1, \beta_2, \beta_3$ 的估计可将 $\theta_1, \theta_2, \theta_3$ 代入 (10.3) 得到。

表 10.3 对进口数据 (1949—1959 年) 用数目不同的主成分获得的标准化变量与原始变量的回归系数的估计值

变量	第一主成分 模型 (10.8)		第一、第二主成分 模型 (10.9)		所有主成分 模型 (10.5)	
	标准化	原始	标准化	原始	标准化	原始
常数	0	-7.735	0	-9.106	0	-10.130
DOPROD	0.487	0.074	0.480	0.073	-0.339	-0.051
STOCK	0.030	0.083	0.221	0.609	0.213	0.587
CONSUM	0.487	0.107	0.483	0.106	1.303	0.287
σ	0.232		0.121		0.108	
R^2	0.952		0.988		0.992	

采用三个主成分模型得到的标准化及原始回归系数的估计值列于表 10.3。明显地, 使用主成分的数目不同, 给出的结果有本质不同。前面已经论证过, OLS 估计是不尽人意的。 \tilde{X}_1 (DOPROD) 的系数为负, 这出乎意料、也难以解释。而且, 有一个严重的共线性关系通过主成分 C_3 混入其中。该变量方差几乎为 0 ($\lambda_3 = 0.003$), 因此接近于 0。余下的两个主成分中, 很明显, 第一主成分与 DOPROD 和 CONSUM 的组合效应有关, 第二主成分仅与 STOCK 有关。从表 10.3 中可以清楚地看出这一结论。DOPROD 和 CONSUM 的系数几乎完全由 IMPORT 关于单个 C_1 的回归决定, 当使用 C_2 时这些系数不会改变。 C_2 的添加导致 STOCK 的系数从 0.083 增到 0.609。同时, R^2 也从 0.952 增加到 0.988。选用基于前两个主成分的模型, 得到的用原始单位表述的方程为

$$\text{IMPORT} = -9.106 + 0.073 \cdot \text{DOPROD} + 0.609 \cdot \text{STOCK} + 0.106 \cdot \text{CONSUM}. \quad (10.11)$$

这对 IMPORT 的关系提供了一种不同于 OLS 结果的、更为合理的描述。另外,分析也导出了预测变量间线性相依性的一种直接的量化(以标准化变量的形式表示的)。我们得到 $C_3 = 0$ 或等价地(由方程(10.4))

$$-0.707\tilde{X}_1 - 0.007\tilde{X}_2 + 0.707\tilde{X}_3 \doteq 0,$$

说明 DOPROD 和 CONSUM 的标准化值本质上是相等的。如果方程(10.11)用于预测或政策决策分析,无论从定性的还是从定量的角度而言,这一信息可能很有用。

10.4 回归系数的约束

对主成分回归方程的结果还有第二种解释。这种解释跟第9章中介绍的对各 θ 加约束的概念有关。对方程(10.9)的估计可通过在方程组(10.7)中令 α_3 为0获得。由(10.6), $\alpha_3 = 0$ 意味着

$$-0.707\theta_1 - 0.007\theta_2 + 0.707\theta_3 = 0 \quad (10.12)$$

或者 $\theta_1 \doteq \theta_3$ 。使用原始的单位,方程(10.12)变为

$$-6.60\beta_1 + 4.54\beta_3 = 0 \quad (10.13)$$

或者 $\beta_1 = 0.69\beta_3$ 。因此,由 C_1, C_2 回归得到的估计,也应该能用 OLS 获得,就像第9章那样,对系数加一个由方程(10.13)给出的线性约束。

回想一下,第9章中,我们假定 $\beta_1 = \beta_3$ 是加于系数的一个先验的约束。我们曾指出,该约束是根据对所研究的过程的了解所作的定性判断。不是看了数据之后加的。现在,通过数据,我们发现,主成分 C_1, C_2 的回归给出的结果等价于加上(10.13)这个约束。结果显示,国内总产值对于进口额的边际效应大概是国内消费总额对于进口额的边际效应的69%。

总之,主成分回归方法既给出了回归系数的另一种估计,还提供了其他一些有关产生数据的潜在过程的有用信息。预测变量间的线性相关结构可被清晰地表示出来。方差(特征值)小的主成分指出了原始变量间的线性关系,那就是多重共线性的来源。通过从回归中去掉一个或多个主成分来消除多重共线性等价于对回归系数加一些约束。这提供了一种构造性的方法去识别那些与所提模型一致的约束,以及发现数据中包含的信息。

10.5 主成分回归: 注意事项

在第9章中我们已经发现主成分分析是检测多重共线性的一种有效工具。本章中,我们采用主成分方法替代最小二乘法,在多重共线性存在的场合获得回归系数的估计。在进口数据例子中,该方法表现得对我们很有利,其中三个主成分中的前两个已经成功地捕捉到了响应变量中的绝大部分波动(见表10.3)。但这种

分析不一定对所有数据集都管用。实际上,在解释响应变量的波动时,主成分回归可能会失败。Hadi and Ling (1998) 用一个名为 Hald 的数据集和一个人为构造的响应变量 U 来说明这一点。原始数据可从 Draper and Smith (1998) p348 找到,也可从本书的网站^①上找到。该数据集含四个预测变量。响应变量 U 和相应于四个预测变量的四个主成分 C_1, \dots, C_4 列于表 10.4。变量 U 已经标准化。这四个主成分的样本方差为 $\lambda_1 = 2.2357, \lambda_2 = 1.5761, \lambda_3 = 0.1866, \lambda_4 = 0.0016$ 。条件数 $\kappa = \sqrt{\lambda_1/\lambda_4} = \sqrt{2.236/0.002} = 37$ 很大,表明原始数据中存在多重共线性。

表 10.4 响应变量 U 与四个预测变量的一组主成分

U	C_1	C_2	C_3	C_4
0.955	1.467	1.903	-0.530	0.039
-0.746	2.136	0.238	-0.290	-0.030
-2.323	-1.130	0.184	-0.010	-0.094
-0.820	0.660	1.577	0.179	-0.033
0.471	-0.359	0.484	-0.740	0.019
-0.299	-0.967	0.170	0.086	-0.012
0.210	-0.931	-2.135	-0.173	0.008
0.558	2.232	-0.692	0.460	0.023
-1.119	0.352	-1.432	-0.032	-0.045
0.496	-1.663	1.828	0.851	0.020
0.781	1.641	-1.295	0.494	0.031
0.918	-1.693	-0.392	-0.020	0.037
0.918	-1.746	-0.438	-0.275	0.037

表 10.5 对 Hald 数据采用全部四个主成分得到的回归结果

变量	系数	标准误	t -检验	p -值
C_1	-0.002	0.001	-1.45	0.1842
C_2	-0.002	0.002	-1.77	0.1154
C_3	0.002	0.005	0.49	0.6409
C_4	24.761	0.049	502.00	< 0.0001
$n = 13$	$R^2 = 1.00$	$R_a^2 = 1.00$	$\hat{\sigma} = 0.0069$	$d.f. = 8$

对这些数据拟合模型

$$U = \alpha_1 C_1 + \alpha_2 C_2 + \alpha_3 C_3 + \alpha_4 C_4 + \varepsilon, \quad (10.14)$$

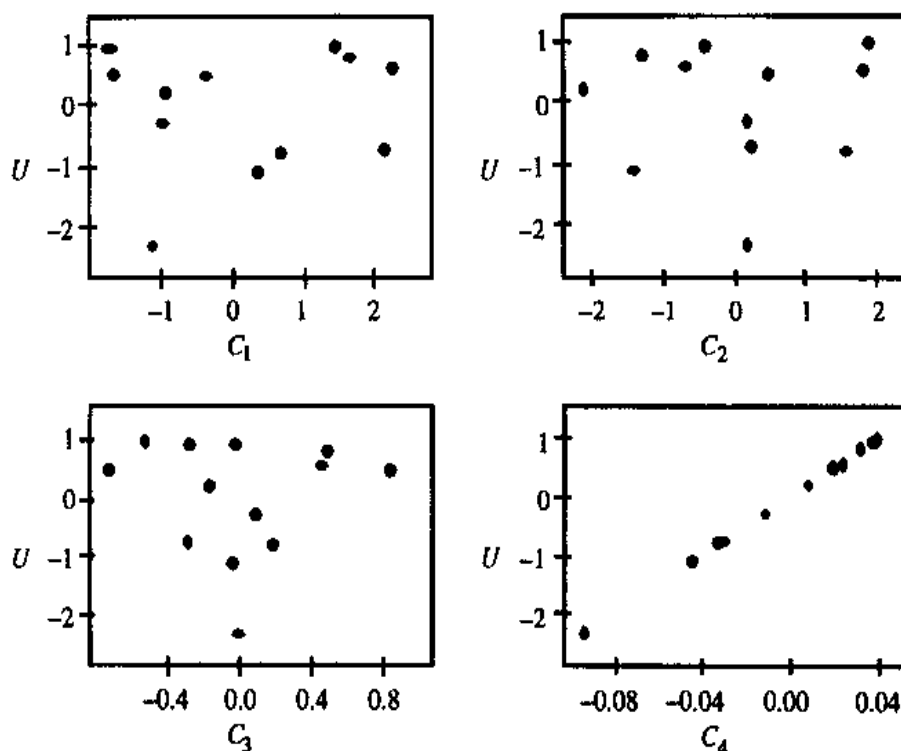
得到的回归结果列于表 10.5。最后一个主成分 C_4 的系数是高度显著的,而其他三个系数不显著。现在,如果我们丢掉方差最小的主成分 C_4 ,那我们得到表 10.6 中的结果。从表 10.5 和 10.6 的比较来看,很明显,所有四个主成分捕捉了 U 中几乎所有的波动,但前三个没有解释 U 的任何波动。因此,在舍弃任何主成分时都应当小心。

^① <http://www.ilr.cornell.edu/~hadi/RABE>

表 10.6 对 Hald 数据采用前三个主成分得到的回归结果

变量	系数	标准误	t -检验	p -值
C_1	-0.001	0.223	-0.01	0.9957
C_2	-0.000	0.266	-0.00	0.9996
C_3	0.002	0.772	0.00	0.9975
$n = 13$	$R^2 = 0.00$	$R_a^2 = -0.33$	$\hat{\sigma} = 1.155$	$d.f. = 9$

使用主成分回归的另一个问题是, 结果可能会过度地受到高杠杆点和异常点的影响 (参见第 4 章对异常和影响的具体讨论)。这是因为主成分是由相关矩阵算得的, 相关矩阵本身可能严重地受到数据中的异常点影响。如果数据中存在异常点的话, 那么响应变量与每个主成分的散点图以及两两主成分之间的散点图应该能指出异常点。 U 和每个主成分的散点图 (图 10.1) 显示, 数据中没有异常点, 而且 U 仅和 C_4 相关, 这与表 10.5、10.6 的结果是一致的。两两主成分的散点图 (这里没有罗列) 也显示数据中没有异常点。主成分回归其他一些可能的缺陷参见 Hadi and Ling (1998)。

图 10.1 Hald 数据中 U 关于每个主成分的散点图

10.6 岭回归

岭回归^①提供了另一种估计方法, 运用于预测变量高度共线性的场合是有益

^① Hoerl (1959) 因这方法与其早期在研究多变量二次响应曲面的工作中所用的岭脊分析方法的相似性, 把它命名为岭回归。

的。定义、计算岭估计（见本章附录）有多种办法，我们选择了与岭迹相结合的办法来介绍。这是一种图形方法，也可以视为一种探索性技术。当怀疑有多重共线性时，使用岭迹的岭分析是解决检测和估计问题的一种一体化的方法。由此导出的估计量是有偏的，但其均方误差往往比 OLS 估计小 (Hoerl and Kennard, 1970)。

回归系数的岭估计可通过解一个形式与正规方程组（参见第 3 章）略有不同的方程组获得。假定回归模型的标准化形式为

$$\tilde{Y} = \theta_1 \tilde{X}_1 + \theta_2 \tilde{X}_2 + \cdots + \theta_p \tilde{X}_p + \varepsilon'. \quad (10.15)$$

岭回归系数的估计方程组为

$$\begin{aligned} (1+k)\theta_1 + r_{12}\theta_2 + \cdots + r_{1p}\theta_p &= r_{1y}, \\ r_{21}\theta_1 + (1+k)\theta_2 + \cdots + r_{2p}\theta_p &= r_{2y}, \\ \vdots &\vdots \\ r_{p1}\theta_1 + r_{p2}\theta_2 + \cdots + (1+k)\theta_p &= r_{py}, \end{aligned} \quad (10.16)$$

其中 r_{ij} 为第 i 个与第 j 个预测变量之间的相关系数， r_{iy} 为第 i 个预测变量与响应变量 \tilde{Y} 之间的相关系数。(10.16) 的解 $\hat{\theta}_1, \dots, \hat{\theta}_p$ 就是岭回归系数的一组估计。岭估计可以看作是对数据稍略作改动后得到的。规范的处理请参阅本章的附录。

区别岭回归与 OLS 的关键参数是 k 。注意到，当 $k=0$ 时，各 $\hat{\theta}$ 就是 OLS 估计。参数 k 可称为偏倚参数。随着 k 从 0 开始逐渐增大，估计的偏倚也随之增大。另一方面，总方差（各回归系数估计的方差之和）

$$\text{总方差}(k) = \sum_{j=1}^p \text{Var}(\hat{\theta}_j(k)) = \sigma^2 \sum_{j=1}^p \frac{\lambda_j}{(\lambda_j + k)^2} \quad (10.17)$$

却是 k 的一个减函数。公式 (10.17) 显示了岭参数对于各回归系数岭估计的总方差的效应。将 $k=0$ 代入 (10.17)，我们得到

$$\text{总方差}(0) = \sigma^2 \sum_{j=1}^p \frac{1}{\lambda_j}. \quad (10.18)$$

这就显示了小的特征根对于回归系数 OLS 估计之总方差的影响。

若 k 无限制地增大下去，那么所有回归系数的估计都将趋于 0^①。岭回归的想法就是选一个 k 值，使得总方差的减少量不超过偏倚的增加量。

有人已经证明，存在一个正的 k 值，其相应的岭估计对于估计数据中的小幅变动将是稳定的 (Hoerl and Kennard, 1970)。在实践中，通常是这样来确定 k 值的：在 0, 1 之间挑一些数作为 k 值，计算相应的 $\hat{\theta}_1, \dots, \hat{\theta}_p$ ，并画出这些结果关于 k 的连线图。这张图就是所谓的岭迹，用于确定合适的 k 值。下例中给出了选择 k 的指导原则。

^① 因为岭回归方法是将各回归系数的估计向 0 压缩，因此，有时候将岭估计量通称为压缩估计量。

10.7 岭估计

源于岭分析的检测多重共线性的方法, 处理的是由于估计数据的微小变动而导致的系数估计的不稳定性。这种不稳定性可以从岭迹上观察到。岭迹是将各回归系数 $\hat{\theta}_1, \dots, \hat{\theta}_p$ 关于 k 的连线画在一起的一张图, k 取诸如 0.001, 0.002 等等各种不同的值。图 10.2 是关于进口数据的岭迹。该图根据表 10.7 中的数据绘制, 其中有 k 取从 0 至 1 共 29 个不同值时各系数的岭估计。典型地, k 值在区间下端点附近取得比较密集。假如 k 值较小时系数估计的波动较大, 那就显示出了不稳定性, 很可能有多重共线性在起作用。

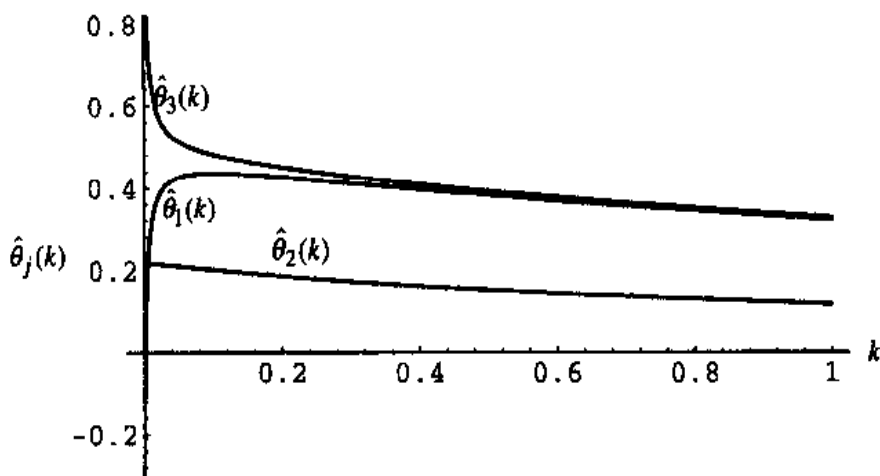


图 10.2 岭迹图: 进口数据 (1949—1959 年)

从岭迹或等价地从表 10.7 明显可见的是, k 值较小时, 系数 θ_1, θ_3 的估计值相当不稳定。 θ_1 的估计值从一个不可思议的负值 -0.339 迅速变化到一个较稳定的值约 0.43 。 θ_3 的估计值从 1.303 稳定到约 0.50 。 $\tilde{X}_2(\text{STOCK})$ 的系数 θ_2 不受多重共线性的影响, 一直稳定在 0.21 附近。

岭分析的下一步是确定 k 值并得到相应的回归系数的估计值。如果多重共线性很严重, 那么当 k 慢慢地从 0 开始增大时, 岭估计将会戏剧性地变化。在 k 继续增大时, 系数最终将稳定下来。因为 k 是偏倚参数, 它的大小直接与其导致的偏倚大小有关, 因此希望选一个最小的、能使系数估计稳定的 k 值。人们提出了几种选择 k 的方法, 包括:

1. 取固定值。Hoerl, Kernard and Baldwin (1975) 提出用

$$k = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^p [\hat{\theta}_j(0)]^2} \quad (10.19)$$

来估计 k , 其中 $\hat{\theta}_1(0), \dots, \hat{\theta}_p(0)$ 为对数据拟合模型 (10.15) 时获得的 $\theta_1, \dots, \theta_p$ 的最小二乘估计 (即取 $k=0$), $\hat{\sigma}^2(0)$ 为相应的残差均方。

表 10.7 岭估计 $\hat{\theta}_j(k)$ 与岭参数 k 之间的函数关系: 进口数据 (1949–1959 年)

k	$\hat{\theta}_1(k)$	$\hat{\theta}_2(k)$	$\hat{\theta}_3(k)$
0.000	-0.339	0.213	1.303
0.001	-0.117	0.215	1.080
0.003	0.092	0.217	0.870
0.005	0.192	0.217	0.768
0.007	0.251	0.217	0.709
0.009	0.290	0.217	0.669
0.010	0.304	0.217	0.654
0.012	0.328	0.217	0.630
0.014	0.345	0.217	0.611
0.016	0.359	0.217	0.597
0.018	0.370	0.216	0.585
0.020	0.379	0.216	0.575
0.022	0.386	0.216	0.567
0.024	0.392	0.215	0.560
0.026	0.398	0.215	0.553
0.028	0.402	0.215	0.548
0.030	0.406	0.214	0.543
0.040	0.420	0.213	0.525
0.050	0.427	0.211	0.513
0.060	0.432	0.209	0.504
0.070	0.434	0.207	0.497
0.080	0.436	0.206	0.491
0.090	0.436	0.204	0.486
0.100	0.436	0.202	0.481
0.200	0.426	0.186	0.450
0.300	0.411	0.173	0.427
0.400	0.396	0.161	0.408
0.500	0.381	0.151	0.391
0.600	0.367	0.142	0.376
0.700	0.354	0.135	0.361
0.800	0.342	0.128	0.348
0.900	0.330	0.121	0.336
1.000	0.319	0.115	0.325

2. 迭代法。Hoerl and Kennard (1976) 提出用下列迭代程序来选择 k : 用 (10.19) 算得 k 的初始值, 记之为 k_0 ; 接着, 计算

$$k_1 = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^p [\hat{\theta}_j(k_0)]^2}, \quad (10.20)$$

然后用 k_1 计算 k_2

$$k_2 = \frac{p\hat{\sigma}^2(0)}{\sum_{j=1}^p [\hat{\theta}_j(k_1)]^2}. \quad (10.21)$$

重复这一过程,直到 k 的前后两个估计值之间的差异可以忽略不计而停止。

3. 岭迹法。作为 k 的一个函数, $\hat{\theta}_j(k)$ 的表现很容易从岭迹观察到。 k 应选使所有系数 $\hat{\theta}_j(k)$ 都稳定的最小的一个值。另外,对于所选的 k 值,残差平方和应依然接近其最小值。各方差膨胀因子^① $VIF_j(k)$, 都应当降至 10 以下。(回忆一下,值 1 是一个正交系统的特征,一个小于 10 的值表明系统无共线性现象或较稳定。)
4. 其他方法。文献中还提出了许多其他估计 k 的方法。譬如,可参阅 Marquardt (1970), Mallows (1973), Goldstein and Smith (1974), McDonald and Galarneau (1975), Dempster et al. (1977), 以及 Wahba, Golub and Heath (1979)。然而,岭迹的诱人之处在于,它用图形来反映多重共线性对于系数估计的效应。

对于进口数据,公式 (10.19) 给出的固定值为

$$k = \frac{3 \times 0.0101}{(-0.339)^2 + (0.213)^2 + (1.303)^2} = 0.0164. \quad (10.22)$$

迭代法则给出了下列序列: $k_0 = 0.0164, k_1 = 0.0161, k_2 = 0.0161$ 。因此,两步迭代之后收敛于 $k = 0.0161$ 。图 10.2 中的岭迹(或看表 10.7)看上去在 k 为 0.04 附近时较稳定。因此我们有了 k 的三种估计值(0.0164, 0.0161 及 0.04)。

从表 10.7 我们看到,无论取这三个值中的哪一个, θ_1 之估计值为负的不正常现象消失了,且系数也基本稳定下来了($k = 0.016$ 时为 0.359, $k = 0.04$ 时为 0.42)。从表 10.8 我们看到,残差平方和 $SSE(k)$ 在 $k = 0$ 处为 0.081,在 $k = 0.016$ 处仅增大到 0.108,在 $k = 0.04$ 处也仅增大到 0.117。同时,方差膨胀因子 $VIF_1(k)$ 与 $VIF_3(k)$ 也从约 185 下降到 1 至 4 之间。明显地, k 取区间 $[0.016, 0.04]$ 中的值看来都是合适的。

以标准化变量表示的和以原始变量表示的模型中各系数的估计总结于表 10.9。原始系数 $\hat{\beta}_j$ 根据 (10.3) 由标准化的系数 $\hat{\theta}_j$ 算得。譬如, $\hat{\beta}_1$ 是这样算出来的

$$\hat{\beta}_1 = (s_y/s_1)\hat{\theta}_1 = (4.5437/29.9995)(0.4196) = 0.0635.$$

因此,用岭方法取 $k = 0.04$ 时拟合的模型,用原始变量的形式可表示为

$$\text{IMPORT} = -8.5537 + 0.0635 \cdot \text{DOPROD} + 0.5859 \cdot \text{STOCK} + 0.1156 \cdot \text{CONSUM}.$$

该方程对各预测变量与响应变量之间的关系给出了一种可能的解释。注意,该方程与用前两个主成分导出的主成分回归结果(见表 10.3)差别不大,尽管这两种计算方法看上去很不一样。

^① $VIF_j(k)$ 的公式在本章附录中给出。

表 10.8 残差平方和 $SSE(k)$ 及方差膨胀因子 $VIF_j(k)$ 与岭参数 k 之间的函数关系:
进口数据 (1949–1959 年)

k	$SSE(k)$	$VIF_1(k)$	$VIF_2(k)$	$VIF_3(k)$
0.000	0.0810	186.11	1.02	186.00
0.001	0.0837	99.04	1.01	98.98
0.003	0.0911	41.80	1.00	41.78
0.005	0.0964	23.00	0.99	22.99
0.007	0.1001	14.58	0.99	14.57
0.009	0.1027	10.09	0.98	10.09
0.010	0.1038	8.60	0.98	8.60
0.012	0.1056	6.48	0.98	6.48
0.014	0.1070	5.08	0.97	5.08
0.016	0.1082	4.10	0.97	4.10
0.018	0.1093	3.39	0.97	3.39
0.020	0.1102	2.86	0.96	2.86
0.022	0.1111	2.45	0.96	2.45
0.024	0.1118	2.13	0.95	2.13
0.026	0.1126	1.88	0.95	1.88
0.028	0.1132	1.67	0.95	1.67
0.030	0.1139	1.50	0.94	1.50
0.040	0.1170	0.98	0.93	0.98
0.050	0.1201	0.72	0.91	0.72
0.060	0.1234	0.58	0.89	0.58
0.070	0.1271	0.49	0.87	0.49
0.080	0.1310	0.43	0.86	0.43
0.090	0.1353	0.39	0.84	0.39
0.100	0.1400	0.35	0.83	0.35
0.200	0.2052	0.24	0.69	0.24
0.300	0.2981	0.20	0.59	0.20
0.400	0.4112	0.18	0.51	0.18
0.500	0.5385	0.17	0.44	0.17
0.600	0.6756	0.15	0.39	0.15
0.700	0.8191	0.14	0.35	0.14
0.800	0.9667	0.13	0.31	0.13
0.900	1.1163	0.12	0.28	0.12
1.000	1.2666	0.11	0.25	0.11

表 10.9 进口数据 (1949–1959 年) 回归系数的 OLS 估计及岭估计

变量	OLS($k=0$)		岭估计 ($k=0.04$)	
	标准化系数	原始系数	标准化系数	原始系数
常数	0	-10.1300	0	-8.5537
DOPROD	-0.3393	-0.0514	0.4196	0.0635
STOCK	0.2130	0.5869	0.2127	0.5859
CONSUM	1.3027	0.2868	0.5249	0.1156
$R^2=0.992$			$R^2=0.988$	

10.8 岭回归：几点说明

岭回归提供了一种工具，用以判断对给定的一批数据作最小二乘分析的稳定性。在高度共线性的情况下，正如已经指出的那样，数据中较小的变动（扰动）会导致回归系数估计的巨大改变。岭回归将揭示这一情况。在这种情况下最小二乘回归应当慎用。针对数据中微小扰动的影响，岭回归提供了比最小二乘估计更为稳健的估计。该方法将显示出，当数据有小变动时，最小二乘系数的敏感性（或稳定性）。

就不受估计数据中微小变动的影响这个意义而言，岭估计是稳定的。又因为其均方误差较小，故可以想像，各系数的岭估计值要比 OLS 估计更接近各回归系数的真实值。同时，对不在估计数据集中的预测变量值预报响应变量，也会更精确。

偏倚参数 k 的估计较为主观。估计 k 的方法有许多，但哪种方法更好，没有一致的意见。无论用哪种方法估计 k ，其估计值都会受到数据中异常点的影响。因此，不管用哪种方法估计 k ，都应当仔细地甄别异常点，以保证估计值不过度地受到数据中异常点的影响。

如同主成分方法那样，用于判断何时岭估计优于 OLS 估计的准则依赖于模型中回归系数的真值。尽管这些值不可能知道，但我们仍然建议，在怀疑多重共线性非常严重的场合，岭分析是有用的。岭回归系数蕴涵着对数据的另一种解释，这能使人对所研究的过程有更深入的理解。

关于岭回归的另一个实际问题是，它在一些统计软件包中尚未被实现。如果一个统计软件包没有岭回归功能，那可以对数据稍作改动，然后用标准的最小二乘软件包获得岭回归的估计值。具体地，回归系数的岭估计可由 Y^* 关于 X_1^*, \dots, X_p^* 的回归得到。新的响应变量 Y^* ，是在 \tilde{Y} 的基础上添加 p 个新的虚拟观测得到的。其中每一个观测都为 0。类似地，新的预测变量 X_j^* 也是在 \tilde{X}_j 的基础上，添加 p 个新的虚拟观测得到的，其中除了第 j 个位置上为 \sqrt{k} 外，其余都为 0，而 k 就是所选的岭参数。可以证明，岭估计 $\hat{\theta}_1(k), \dots, \hat{\theta}_p(k)$ 能由 Y^* 关于 X_1^*, \dots, X_p^* 的不含常数项模型的最小二乘回归获得。

10.9 小结

岭回归和主成分回归，这两种可供选择的估计方法，都对分析数据提供了额外的信息。我们已经看到，预测变量的相关矩阵之特征根在检测多重共线性及分析其效应时起着重要的作用。由这些方法得到的回归的估计值是有偏的，但在均方误差意义下可能比 OLS 估计更精确。对一个具体问题评价精度上的得益，这是不可能的，因为这两种方法与 OLS 之比较需要知道系数的真实值。但是，当怀疑有严重多重共线性时，我们建议除了 OLS 估计之外，还至少用其中的一种方法估计。这些估计值可能喻示着以前没有考虑过的一种对数据的解释。

使用主成分或者岭回归方法，理论上没有很强的依据。我们建议，在严重的

多重共线性出现时, 这些方法可用作形象的诊断工具, 来判断用最小二乘法分析这些数据是否合适。当主成分或岭回归分析揭示了一个具体的数据集的不稳定性之后, 分析人员应当首先考虑减少一些变量作最小二乘回归 (如第 9 章所示)。如果最小二乘回归仍然不令人满意 (各 VIF 较大, 系数的符号不对, 条件数很大), 此时才应当考虑用主成分或岭回归。

习 题

10.1 Longley(1967) 的数据集是多重共线性数据的一个经典的例子。数据 (见表 10.10) 由一个响应变量 S 和六个预测变量 X_1, \dots, X_6 组成。数据也可从本书的网站上找到。用原始变量表示的初始模型为

$$S = \beta_0 + \beta_1 X_1 + \dots + \beta_6 X_6 + \varepsilon, \quad (10.23)$$

可用标准化的变量表示为

$$\tilde{S} = \theta_1 \tilde{X}_1 + \dots + \theta_6 \tilde{X}_6 + \varepsilon'. \quad (10.24)$$

- (a) 采用最小二乘法对数据拟合模型 (10.24)。你能从这些数据中得出什么结论?
 - (b) 根据拟合模型 (10.24) 的结果, 计算模型 (10.23) 中各回归系数的最小二乘估计。
 - (c) 再采用最小二乘法对数据拟合模型 (10.23), 验证得到的结果与上面的结果是一致的。
 - (d) 计算六个预测变量的相关矩阵, 并画出相应的散点图矩阵。你看出共线性的迹象了吗?
 - (e) 计算相应的主成分、其样本方差, 以及条件数。数据中存在多少种不同的多重共线性关系? 每一种与哪些变量有关?
 - (f) 根据你所保留的主成分数目, 计算模型 (10.23)、(10.24) 中系数的主成分估计。
 - (g) 采用岭方法, 构造岭迹。你建议用哪个 k 值去估计模型 (10.23)、(10.24) 的参数? 根据你所选的 k 值计算 (10.23)、(10.24) 中回归系数的岭估计。
 - (h) 比较你用三种方法得到的估计值。你推荐用哪个? 请作解释。
- 10.2** 用 10.5 节中讨论过的 Hald 数据重复习题 10.1, 但原始响应变量为 Y , 四个预测变量为 X_1, \dots, X_4 。数据见表 10.11。
- 10.3** 根据你对 Longley 和 Hald 两个数据集的分析, 你有没有观察到 10.5 节中指出的那类问题?

表 10.10 Longley (1967 年) 数据

Y	X_1	X_2	X_3	X_4	X_5	X_6
60323	830	234289	2356	1590	107608	1947
61122	885	259426	2325	1456	108632	1948
60171	882	258054	3682	1616	109773	1949
61187	895	284599	3351	1650	110929	1950
63221	962	328975	2099	3099	112075	1951
63639	981	346999	1932	3594	113270	1952
64989	990	365385	1870	3547	115094	1953
63761	1000	363112	3578	3350	116219	1954
66019	1012	397469	2904	3048	117388	1955
67857	1046	419180	2822	2857	118734	1956
68169	1084	442769	2936	2798	120445	1957
66513	1108	444546	4681	2637	121950	1958
68655	1126	482704	3813	2552	123366	1959
69564	1142	502601	3931	2514	125368	1960
69331	1157	518173	4806	2572	127852	1961
70551	1169	554894	4007	2827	130081	1962

表 10.11 Hald 数据

Y	X_1	X_2	X_3	X_4
78.5	7	26	6	60
74.3	1	29	15	52
104.3	11	56	8	20
87.6	11	31	8	47
95.9	7	52	6	33
109.2	11	55	9	22
102.7	3	71	17	6
72.5	1	31	22	44
93.1	2	54	18	22
115.9	21	47	4	26
83.8	1	40	23	34
113.3	11	66	9	12
109.4	10	68	8	12

来源: Draper and Smith (1998), p.348.

附录：岭回归

本附录中，我们用矩阵记号来介绍岭回归方法。

A. 模型

回归模型可表示为

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\varepsilon}, \quad (\text{A.1})$$

其中 \mathbf{Y} 是响应变量的一个 $n \times 1$ 观测向量， $\mathbf{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_p)$ 是由 p 个预测变量的 n 个观测组成的一个 $n \times p$ 矩阵， $\boldsymbol{\theta}$ 是 $p \times 1$ 回归系数向量， $\boldsymbol{\varepsilon}$ 为 $n \times 1$ 随机误差向量。假定 $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2\mathbf{I}$ ，其中 \mathbf{I} 为 n 阶单位阵。另外，不失一般性，假定 \mathbf{Y} 、 \mathbf{Z} 都经过了中心化和尺度变换，因此， $\mathbf{Z}^T\mathbf{Z}$ 和 $\mathbf{Z}^T\mathbf{Y}$ 都是相关系数构成的矩阵^①。

$\boldsymbol{\theta}$ 之最小二乘估计为 $\hat{\boldsymbol{\theta}} = (\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{Y}$ 。可以证明

$$E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})] = \sigma^2 \sum_{j=1}^p \lambda_j^{-1}, \quad (\text{A.2})$$

其中 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$ 是 $\mathbf{Z}^T\mathbf{Z}$ 的特征根。(A.2) 的左边称为总的均方误差。它是回归系数的估计值与真实值之间的距离平方的一个综合量度。

B. 多重共线性的效应

在第 9 章以及第 9 章的附录中已经论证过，存在多重共线性与存在较小的特征根实际是同一回事。由方程 (A.2) 知，当一个或多个 λ 较小时， $\hat{\boldsymbol{\theta}}$ 的总的均方误差很大，也就蕴涵着最小二乘估计方法的不精确。岭回归方法试图构造另一种估计量，其总的均方误差较小。

C. 岭回归估计量

Hoerl and Kennard (1970) 提出了以 $k > 0$ 为参数的一族估计量。给定 k 值，估计量为

$$\hat{\boldsymbol{\theta}}(k) = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Y} = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}\hat{\boldsymbol{\theta}}. \quad (\text{A.3})$$

$\hat{\boldsymbol{\theta}}(k)$ 之期望值为

$$E[\hat{\boldsymbol{\theta}}(k)] = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}\boldsymbol{\theta}, \quad (\text{A.4})$$

方差-协方差矩阵为

$$\text{Var}[\hat{\boldsymbol{\theta}}(k)] = (\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\mathbf{Z}^T\mathbf{Z}(\mathbf{Z}^T\mathbf{Z} + k\mathbf{I})^{-1}\sigma^2. \quad (\text{A.5})$$

^① 注意， Z_j 是由原始预测变量 X_j 经变换 $z_{ij} = (x_{ij} - \bar{x}_j)/\sqrt{\sum (x_{ij} - \bar{x}_j)^2}$ 而得。因而， Z_j 经过了中心化和长度单位化变换，即 $\sum z_{ij}^2 = 1$ 。

方差膨胀因子 $VIF_j(k)$, 为 k 的一个函数, 就是矩阵 $(\mathbf{Z}^T \mathbf{Z} + k\mathbf{I})^{-1} \mathbf{Z}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z} + k\mathbf{I})^{-1}$ 的第 j 个对角元。

残差平方和可表为

$$\begin{aligned} SSE(k) &= (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}}(k))^T (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}}(k)) \\ &= (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}})^T (\mathbf{Y} - \mathbf{Z}\hat{\boldsymbol{\theta}}) + (\hat{\boldsymbol{\theta}}(k) - \hat{\boldsymbol{\theta}})^T \mathbf{Z}^T \mathbf{Z} (\hat{\boldsymbol{\theta}}(k) - \hat{\boldsymbol{\theta}}). \end{aligned} \quad (\text{A.6})$$

总的均方误差为

$$\begin{aligned} \text{TMSE}(k) &= E[(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})^T (\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})] \\ &= \sigma^2 \text{trace}[(\mathbf{Z}^T \mathbf{Z} + k\mathbf{I})^{-1} \mathbf{Z}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z} + k\mathbf{I})^{-1}] \\ &\quad + k^2 \boldsymbol{\theta}^T (\mathbf{Z}^T \mathbf{Z} + k\mathbf{I})^{-2} \boldsymbol{\theta} \\ &= \sigma^2 \sum_{j=1}^p \lambda_j (\lambda_j + k)^{-2} + k^2 \boldsymbol{\theta}^T (\mathbf{Z}^T \mathbf{Z} + k\mathbf{I})^{-2} \boldsymbol{\theta}. \end{aligned} \quad (\text{A.7})$$

注意, 方程 (A.7) 右边的首项是 $\hat{\boldsymbol{\theta}}(k)$ 各分量的方差之和 (总方差), 第二项是偏倚的平方。Hoerl and Kennard (1970) 证明了, 存在一个 $k > 0$, 使得

$$E[(\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})^T (\hat{\boldsymbol{\theta}}(k) - \boldsymbol{\theta})] < E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^T (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})].$$

也就是说, 岭估计量 $\hat{\boldsymbol{\theta}}(k)$ 的均方误差小于 OLS 估计量 $\hat{\boldsymbol{\theta}}$ 的均方误差。Hoerl and Kennard (1970) 指出, 可通过观察岭迹及诸如 $SSE(k)$ 、 $VIF_j(k)$ 等一些对 $\hat{\boldsymbol{\theta}}(k)$ 作补充的概述性统计量, 来选择合适的 k 值。所选 k 值应当是使 $\hat{\boldsymbol{\theta}}(k)$ 稳定的最小的一个值。另外, 在所选的 k 值点上, 残差平方和应接近其最小值, 且如同第 9 章所讨论的, 方差膨胀因子应小于 10。

人们已从多个方面对岭估计量作了推广。它们有时通称为压缩估计量, 因为这些方法都是将回归系数的估计值向 0 压缩。我们来看一种可能的推广, 考虑将回归模型改用主成分 $\mathbf{C} = (\mathbf{C}_1, \dots, \mathbf{C}_p)$ 表示, 这在第 9 章的附录中已讨论过。一般的模型具有形式

$$\mathbf{Y} = \mathbf{C}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \quad (\text{A.8})$$

其中

$$\begin{aligned} \mathbf{C} &= \mathbf{Z}\mathbf{V}, \quad \boldsymbol{\alpha} = \mathbf{V}^T \boldsymbol{\theta}, \\ \mathbf{V}^T \mathbf{Z}^T \mathbf{Z} \mathbf{V} &= \mathbf{\Lambda}, \quad \mathbf{V}^T \mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}, \end{aligned} \quad (\text{A.9})$$

而

$$\mathbf{\Lambda} = \begin{pmatrix} \lambda_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & 0 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_{p-1} & 0 \\ 0 & 0 & 0 & \cdots & 0 & \lambda_p \end{pmatrix}, \quad \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$$

为由 $\mathbf{Z}^T \mathbf{Z}$ 之特征根排序后构成的一个对角矩阵。(A.7) 中总的均方误差变为

$$\begin{aligned} \text{TMSE}(k) &= E[(\hat{\theta}(k) - \theta)^T (\hat{\theta}(k) - \theta)] \\ &= \sigma^2 \sum_{j=1}^p \frac{\lambda_j}{(\lambda_j + k)^2} + \sum_{j=1}^p \frac{k^2 \alpha_j^2}{(\lambda_j + k)^2}, \end{aligned} \quad (\text{A.10})$$

其中 $\alpha^T = (\alpha_1, \alpha_2, \dots, \alpha_p)$ 。我们可以考虑取多个不同的 k 值, 记为 k_1, k_2, \dots, k_p , 而不是单个 k 值, 也就是对每个回归系数单独地考虑岭参数 (即压缩因子)。现在, k 不再是一个标量而是一个向量了, 记之为 \mathbf{k} 。(A.10) 给出的总的均方误差现在变成了

$$\begin{aligned} \text{TMSE}(\mathbf{k}) &= E[(\hat{\theta}(\mathbf{k}) - \theta)^T (\hat{\theta}(\mathbf{k}) - \theta)] \\ &= \sigma^2 \sum_{j=1}^p \frac{\lambda_j}{(\lambda_j + k_j)^2} + \sum_{j=1}^p \frac{k_j^2 \alpha_j^2}{(\lambda_j + k_j)^2}. \end{aligned} \quad (\text{A.11})$$

取 $k_j = \sigma^2 / \alpha_j^2$ 便最小化 (A.11) 给出的总的均方误差。人们提出了一个迭代估计的步骤。第一步, 由普通的最小二乘法得到的 σ^2 与 α_j 的估计值, 计算 k_j 。然后用公式

$$\hat{\alpha}(\mathbf{k}) = (\mathbf{C}^T \mathbf{C} + \mathbf{K})^{-1} \mathbf{C}^T \mathbf{Y}$$

计算 $\hat{\alpha}(\mathbf{k})$, 其中 \mathbf{K} 是以第一步中获得的 k_1, \dots, k_p 为对角元的对角阵。重复该过程, 直至 $\hat{\alpha}(\mathbf{k})$ 各元素在相继两步迭代中的改变值可以忽略不计为止。再用方程 (A.9) 得到 θ 之估计为

$$\hat{\theta}(\mathbf{k}) = \mathbf{V} \hat{\alpha}(\mathbf{k}). \quad (\text{A.12})$$

Dempster et al. (1977) 讨论了前面定义的这两种岭型估计量 (单个 k 值、多个 k 值) 以及普通最小二乘估计的其他一些替代方法。文中采用 Monte Carlo 技术对不同的估计量进行了比较和评价。一般面言, 对于一个具体问题, 最佳估计方法的选择依赖于特定的模型与数据。Dempster et al. (1977) 提示, 在一项分析中, 这可能用来对一个给定的数据集确定最佳估计方法。目前, 我们还是偏向于最简单的岭方法, 即通过考察岭迹选取单个岭参数 k 。

11

变量选择的方法

11.1 引言

迄今为止，我们讨论回归问题都是假定进入方程的变量是事先选定的。我们的分析是围绕对方程的考察展开的，主要关心：函数形式的设定是否正确，关于误差项的假定是否合理。分析都事先假定，已经明确方程中包含哪些变量。然而，在回归分析的诸多应用中，回归模型中包含的变量集事先并没有确定下来，而且，选择变量常常是分析的第一步。有一些场合，出于理论或其他考虑，方程中包含的变量已被确定，此时就不存在变量选择的问题。但在没有清晰理论的场合，回归方程的变量选择问题便至关重要了。

变量选择问题与方程函数形式的设定是相互关联的。在表述一个回归模型时，需要回答的问题有：应包含哪些变量？应取这些变量的何种形式？也就是它们是该以一个原始变量的形式 X 进入方程呢，还是以诸如 $X^2, \log X$ 等一些变换形式进入方程，还是以两者组合的形式进入？尽管理想地，这两个问题应同时解决，但出于简化的目的，我们还是建议依次地处理它们。我们首先确定，哪些变量将包含于方程中，之后，再考察变量进入方程的准确形式。这是一种简化的做法，但它能使变量选择问题更容易处理。一旦选定了包含于方程的变量，我们便可以运用前几章介绍的方法得到方程的确切形式。

11.2 问题的归纳

设有一个响应变量 Y 和 q 个预测变量 X_1, X_2, \dots, X_q 。用 q 个变量描述 Y 的线性模型为

$$y_i = \beta_0 + \sum_{j=1}^q \beta_j x_{ij} + \varepsilon_i, \quad (11.1)$$

其中各 β_j 为参数， ε_i 表示随机干扰。有时候，尤其当 q 较大时，我们可能会舍弃一些变量，只用一个变量子集建立方程，而不直接处理变量的全集。本章关心的

是确定哪些变量该保留在方程中。我们将保留的变量记为 X_1, X_2, \dots, X_p , 将那些舍弃的变量记为 $X_{p+1}, X_{p+2}, \dots, X_q$. 让我们在如下两个一般的条件下考察剔除变量的效应:

1. Y 关于诸 X 的模型中, 所有 $\beta(\beta_0, \beta_1, \dots, \beta_q)$ 均非 0.
2. 模型中 $\beta_0, \beta_1, \dots, \beta_p$ 非 0, 但 $\beta_{p+1}, \beta_{p+2}, \dots, \beta_q$ 为 0.

假如我们拟合如下子集模型, 而非 (11.1),

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i. \quad (11.2)$$

我们将阐述在刚刚讲的两情形下, 拟合诸 X 之全集模型及子集模型的效果。简而言之, 就是把该舍弃的变量 (因其总体回归系数为 0) 纳入方程的效应, 以及把该包含的变量 (因其总体回归系数非 0) 排除在外的效应。我们将考察舍弃变量对于参数估计值的影响以及对于 Y 之预测值的影响。一旦弄明白了保留不重要的变量以及舍弃重要变量对于一个方程的影响, 那变量选择问题的解就会逐渐明了起来。

11.3 剔除变量的后果

在用变量 X_1, X_2, \dots, X_q 之全集拟合模型 (11.1) 时, 回归系数的估计记为 $\hat{\beta}_0^*, \hat{\beta}_1^*, \dots, \hat{\beta}_q^*$. 拟合模型 (11.2) 时, 回归系数的估计记为 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$. 记 \hat{y}_i^*, \hat{y}_i 分别为相应于观测 $(x_{i1}, x_{i2}, \dots, x_{iq})$ 之全集与子集的预测值。现在可将结果概述如下 (用矩阵记号表示的概述在本章附录中给出): $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 是 $\beta_0, \beta_1, \dots, \beta_p$ 的有偏估计, 除非其余模型参数 $\beta_{p+1}, \beta_{p+2}, \dots, \beta_q$ 为 0 或变量 X_1, X_2, \dots, X_p 与变量集 $(X_{p+1}, X_{p+2}, \dots, X_q)$ 正交。估计量 $\hat{\beta}_0^*, \hat{\beta}_1^*, \dots, \hat{\beta}_p^*$ 的精度相应地比 $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ 差, 即

$$\text{Var}(\hat{\beta}_j^*) \geq \text{Var}(\hat{\beta}_j), \quad j = 0, 1, \dots, p.$$

一个简约的方程中, 变量的回归系数估计量的方差不大于全模型中相应估计量的方差。变量的舍弃降低了, 或更准确地说, 没有增加其余回归系数估计量的方差。因为 $\hat{\beta}_j$ 有偏而 $\hat{\beta}_j^*$ 无偏, 故若要更好地比较估计量的精度, 应拿 $\hat{\beta}_j$ 的均方误差与 $\hat{\beta}_j^*$ 的方差比。仅当被舍弃变量的回归系数的数量级比这些系数相应的估计量的标准差小时, $\hat{\beta}_j$ 的均方误差 (MSE) 小于 $\hat{\beta}_j^*$ 的方差。根据子集模型获得的 σ^2 的估计一般偏大。

现在, 让我们再来看剔除变量对于预测的影响。预测值 \hat{y}_i 是有偏的, 除非剔除变量的回归系数为 0^①。由子集模型获得的预测值之方差不大于由全模型获得的预测值之方差, 即

$$\text{Var}(\hat{y}_i) \leq \text{Var}(\hat{y}_i^*).$$

^① 译注: 原文中此处还有一句“或者留下的变量集与删去的变量集正交”。这是 $\hat{\beta}_0, \dots, \hat{\beta}_p$ 无偏的充分条件, 并非 \hat{y}_i 无偏的充分条件。原文有误。参见本章附录的译注。

$MSE(\hat{y}_i)$ 小于 $Var(\hat{y}_i^*)$ 的条件与前面我们叙述的 $MSE(\hat{\beta}_j)$ 小于 $Var(\hat{\beta}_j^*)$ 的条件相同。更详细的讨论参阅 Chatterjee and Hadi (1988)。

变量选择的基本原理可概括如下：尽管舍弃变量的回归系数非 0，留下来的那些变量的回归系数，用子集模型去估计比用全模型去估计方差可能更小。对响应变量预测的方差结果亦然。剔除变量的代价是在估计量中引进了偏倚。然而，如我们上面所述，存在这样的条件，有偏估计的 MSE 会小于无偏估计的方差；也就是，精度上的得益没有被偏倚的平方抵消掉。另一方面，如果保留的变量中有一些是无关的或非本质的，即其系数为 0 或系数的数量级比估计量的标准差更小，那么将它们保留在方程中会降低估计和预测的精度。

读者可在 3.5 节、4.12 节及 4.13 节中查阅更多有关回归系数的解释以及变量在回归建模中的作用的细节。

11.4 回归方程的用途

一个回归方程有诸多用途，概括如下。

11.4.1 描述与建模

一个回归方程可用于描述一个给定的过程或用作一个复杂的关联系统的模型。方程的目的可以纯粹是描述性的，仅在于说明该复杂的关联性质。对于这种用途，有两种矛盾的需求：(1) 尽可能多地解释变异，这种需求倾向于将较多变量纳入模型；(2) 坚持吝啬原则，意味着为方便理解和解释，我们应力争用尽可能少的变量来描述过程。在主要目的是为描述的场所，我们尽量选择最少的、能解释响应变量最本质的那部分变异的预测变量。

11.4.2 估计与预测

有时为预测而构造回归方程。根据该回归方程，我们希望预测一个未来观测的值，或对一个给定的观测估计其响应的均值。当回归方程用于这个目的时，选择变量着眼于最小化预测的 MSE 。

11.4.3 控制

回归方程可作控制工具。构建方程的目的可能是：为使响应（目标）变量达到某个设定的值，确定某个预测变量须改变的幅度。这时，回归方程被视为一个以 Y 为响应变量的响应函数。出于控制的目的，希望准确地度量方程中变量的系数，即要求各回归系数的标准误小。

回归方程有广泛的用途。这些功能偶尔会重合在一起，此时一个方程是为若干个或所有这些目的构造的。我们要指出的主要一点是，构建回归方程的目的决定了其构建过程中要被优化的准则。紧接着会遇到的问题是，一个变量子集，从一个目的来看可能是最优的，但从另一个目的来看却可能不是最优的。“最优”变量子集这一概念，常常需要附加的先决条件。

在讨论实际的选择方法之前,我们作两个基本的注释。

第一,论及多元回归方程中的“最优变量集”并不总是有意义的。因为不存在一个唯一的“最优变量集”。一个回归方程能被用于多种目的。一个变量集对于一个目的而言可能是最优的,但对于另一个而言就不一定是最优的了。在变量选择过程中,应牢牢记住构建回归方程的目的。后面我们将说明,构建方程的目的决定了选择及评价变量贡献的准则。

第二,因为没有最优的变量集,也就可能有多个变量子集,它们都是适当的并都可能用来建立方程。一种好的变量选择方法应当能指出这多个集合,而不是只生成单个所谓的“最优”集合。多种适当的变量集合提供了数据结构的信息,有助于我们对潜在过程的理解。实际上,变量选择方法应被视为对预测变量间相关结构一种深层的分析,揭示它们是如何单独地或共同地影响我们所研究的响应变量。这两点影响着我们提出的与变量选择有关的方法论。

11.5 评价回归方程的准则

为评价所拟合的各方程的充分性,我们需要一个准则。统计文献中已提出过多种准则。我们介绍两个我们认为最有用的准则。详尽的准则清单可从 Hocking (1976) 中找到。

11.5.1 残差均方

用于判断方程拟合得是否充分的一种量度为残差均方 (RMS)。对于一个含 p 个项的方程, RMS 定义为

$$\text{RMS}_p = \frac{SSE_p}{n-p}, \quad (11.3)$$

其中 SSE_p 为含 p 个项的方程的残差平方和。两个方程相比, RMS 较小的那个较好,尤其当目标为预报时。

很明显, RMS_p 与复相关系数之平方 R_p^2 及修正的复相关系数之平方 R_{ap}^2 都有联系,这些都曾在第 3 章中介绍来作为判断一个方程的拟合效果的量度。这里,我们对 R^2, R_a^2 添加了一个下标,表示它们与方程中项数的相依关系。这些量之间的关系为

$$R_p^2 = 1 - (n-p) \frac{\text{RMS}_p}{SST}, \quad (11.4)$$

$$R_{ap}^2 = 1 - (n-1) \frac{\text{RMS}_p}{SST}, \quad (11.5)$$

其中

$$SST = \sum (y_i - \bar{y})^2.$$

注意,在对预测变量数不同的模型作比较时, R_{ap}^2 比 R_p^2 更合适,因为 R_{ap}^2 对模型中的预测变量数作了修正(惩罚)。

11.5.2 Mallows 的 C_p 准则

我们前面已经指出, 根据某个变量子集的回归方程得到的预测值一般是有偏的。为判断一个方程的表现, 我们应当考虑预测值的均方误差而不是方差。所有观测点上预测的标准化均方误差之和可用

$$J_p = \frac{1}{\sigma^2} \sum_{i=1}^n MSE(\hat{y}_i) \quad (11.6)$$

度量, 其中 $MSE(\hat{y}_i)$ 为根据含 p 个项的方程作的第 i 个预测值的均方误差, σ^2 是随机误差的方差。 $MSE(\hat{y}_i)$ 包括两个部分, 由估计产生的预测方差, 以及由剔除变量导致的偏倚。

Mallows (1973) 采用统计量

$$C_p = \frac{SSE_p}{\hat{\sigma}^2} + (2p - n) \quad (11.7)$$

估计 J_p , 其中 $\hat{\sigma}^2$ 是 σ^2 的估计, 通常由含全部 q 个变量的线性模型获得。可以证明, 当拟合的含 p 个项的方程无偏倚时, C_p 的期望值为 p 。于是, C_p 与 p 之间的差异就可作为偏倚的度量。因此, C_p 统计量用所有观测点上预测的标准化均方误差之和来衡量变量的表现, 而不管未知的真实模型是什么。它同时考虑了偏倚和方差。 C_p 接近于 p 的变量子集是理想的子集。“好”子集的选择可借助图形实施。对不同的子集, 画出 C_p 关于 p 的散点图, 并将直线 $C_p = p$ 也标于图中。图中的点越接近直线 $C_p = p$, 则其对应的变量集越好, 即用来构造方程越理想。我们将通过 11.10 节中的例子对 C_p 图的使用作更详细的说明和讨论。Daniel and Wood (1980) 中对 C_p 统计量作了非常全面的论述。

11.6 多重共线性和变量选择

在讨论变量选择方法时, 我们区分两种比较普遍的情况:

1. 预测变量非共线性的, 即没有有力的多重共线性的证据。
2. 预测变量是共线性的, 即数据有高度多重共线性关系。

根据预测变量的相关结构, 我们提出了不同的变量选择方法。如果分析的数据非共线性的, 则按一种方法进行, 反之, 则按另一种方法进行。

作为变量选择方法的第一步, 我们建议计算各方差膨胀因子 (VIF) 或预测变量相关矩阵的特征根。如果没有一个 VIF 大于 10, 那就没有共线性问题。此外, 就如我们在第 9 章中解释的, 出现小的特征根就意味着有共线性。若条件数^①大于 15, 则变量是共线性的。我们也可以看所有特征根的倒数之和。若任何单个的特征根小于 0.01 或所有特征根的倒数之和大于问题中预测变量数的 5 倍, 那我们可以认为变量是共线性的。若上述条件不成立, 那可认为变量是非共线性的。

^① 回忆一下, 第 9 章中条件数定义为 $\kappa = \sqrt{\lambda_{\max}/\lambda_{\min}}$, 其中 $\lambda_{\max}, \lambda_{\min}$ 分别为相关矩阵的最大、最小特征根。

11.7 评价所有可能的方程

我们将介绍的第一种方法非常直接,应用于共线性的和非共线性的数据一样地好。该方法就是对一批给定的数据拟合所有可能的子集方程。用 q 个变量总共可拟合 2^q 个方程,其中一个含所有变量,另一个不含任何变量。后者也就是 $\hat{y}_i = \bar{y}$, 由拟合模型 $Y = \beta_0 + \varepsilon$ 获得。这方法明确地提供给分析者关于 Y 与这组 X 间关系本质最多的信息。然而,方程的数目及必须得看的辅助信息大得惊人。甚至在仅六个预测变量时,已有 $64(2^6)$ 个方程要考虑,七个变量时方程数目增至 $128(2^7)$, 既不可行也不实用。利用所有可能方程的拟合结果的一个有效办法是,在变量数相同的方程中根据 R^2 , C_p 或 RMS 挑出三个“最好的”。接着,分析这个较小的方程子集,以获得最后的模型。在运用所有子集回归方法时,可采用 C_p 或者 RMS 来鉴别那些最有希望的方程。然后再仔细地分析这些方程以决定最后的模型,如通过残差去考察异常点、自相关,或作变换的必要性等。各种各样的变量子集蕴涵着对数据的多种解释,而这在限制较多的变量选择方法中往往会被忽视。

当变量数很大时,评价所有可能的方程实际上往往是不可行的。人们也提出了一些捷径 (Furnival and Wilson, 1974; La Motte and Hocking, 1970), 在搜索理想的子集时不计算所有的方程。但变量很多时,这些方法的计算量仍然相当可观。还有一些变量选择方法,它们不需要对所有可能的方程作评价。采用这些方法提供给分析者的信息没有拟合全部可能方程那么多,但能够大大地减少计算量,因而也许是唯一切实可行的办法。11.8 节讨论这些方法。它们对于非共线性的数据非常有效。但我们不建议将它们用于共线性数据。

11.8 若干变量选择方法

对于有大量潜在的预测变量的情况,人们提出了一系列不用计算全部可能方程的方法。这些方法的共同特点是,每次只引进一个变量或每次只从方程中剔除一个变量,并且只考察全部可能方程的一个子集。对于 q 个变量,这些方法最多只需评价 $q+1$ 个方程,这与考察全部可能方程时必须评价 2^q 个方程形成鲜明的对照。这些方法可以分为两大类:(1) 前向选择法(FS); (2) 后向剔除法(BE)。还有一类 FS 法的改版,称为逐步法。下面我们介绍并比较这三类方法。

11.8.1 前向选择法

前向选择法始于一个不含任何预测变量、只含一个常数项的方程。与响应变量 Y 的简单相关性最大的一个变量首先被引进方程。如果该变量的回归系数显著非 0, 那么将它保留在方程中, 然后进行第二个变量的搜索。第二个被引进方程的,是与经第一个变量修正后的 Y 相关性最大的那个变量,也就是与第一步中的残差有着最大的简单相关系数的那个变量。接着,检验第二个变量的回归系数的显著性。如果显著,则循着同样的线路搜索第三个变量。若最后引进方程的那个变量的回归系数不显著,或所有变量都已被引进方程,那么该过程终止。最后引

进的那个变量的回归系数的显著性,由根据最后那个方程计算的标准 t -检验来判断。绝大多数前向选择算法在检验新引进变量的系数时,采用一个较低的 t 截止值;因此,前向选择法搜遍了整个变量集,提供给我们最多是 $q+1$ 个可能的方程。

11.8.2 后向剔除法

后向剔除法始于全变量的方程,逐次从中剔除变量,每次剔除一个。剔除哪个则根据它们对于减少残差平方和的贡献来决定。对残差平方和的降低贡献最小的那个变量首先被剔除。这等价于删去方程中 t -值绝对值最小的那个变量。如果所有的 t -检验均显著,那么变量之全集保留在方程中。假定有一个或多个变量的 t -检验不显著,该方法则选择 t -检验最不显著的那个变量剔除。接着,用剩下的 $q-1$ 个变量拟合方程,对新的回归系数作 t -检验。当所有的 t -检验都显著、或所有变量都已被剔除时,过程终止。在绝大多数后向剔除法中, t -检验的截止值被设得很高,以使该方法能够搜遍整个变量集,也即从 q 变量的方程出发至仅含一个常数项的方程结束。后向剔除法最多拟合 $q+1$ 个方程。

11.8.3 逐步法

逐步法实质是一种前向选择法,但附加了一个条件,即在每一步骤中如同向后剔除法那样考虑剔除一个变量的可能性。在该方法中,一个在先前的步骤中引进的变量可能在后面的步骤中被剔除。为引进、剔除变量所作的计算和 FS、BE 法相同。通常,对引进变量和剔除变量所设定的显著性水平是不同的。

11.9 变量选择方法的一般说明

上面讨论的变量选择方法使用时应多加小心。我们不应机械地用这些方法来确定“最优的”变量集。变量选择过程中,变量进入或离开方程的次序不应该被解释为是变量相对重要性的体现。如果记住这些告诫的话,那么这些方法将是在非共线性场合变量选择的有力工具。对于非共线性数据,这三类方法给出几乎相同的变量选择。它们所需的计算量大大少于对全部可能方程的分析。

人们对变量选择方法提出了多种停止规则。据说非常有效的一种停止规则为:

1. 对 FS 法:当 t -检验的最小绝对值小于 1 时停止。
2. 对 BE 法:当 t -检验的最小绝对值大于 1 时停止。

在下面这个例子中,我们来说明变量选择中不同停止规则的效果。

与 FS 法相比,我们优先推荐 BE 法作变量选择。一个明显的理由是,BE 法计算全变量方程,尽管它不一定被用作最后的方程,但可用于校验。虽然我们并不主张这些变量选择方法在共线性场合使用,但与 FS 法相比,BE 法比较能处理多重共线性 (Mantel,1970)。

在应用变量选择方法时,会生成多个方程,每个方程包含的变量数不同。接着,可用诸如 C_p 或 RMS 等统计量来评价这些不同的方程。还应该考察各方程的

残差。残差图不理想的方程该被拒绝。只有全面、综合地分析,才能给出合适的变量选择以及有用的回归方程。我们通过下面的例子来说明变量选择的这一做法。

11.10 主管人员业绩的研究

为说明非共线性场合变量选择的方法,我们考虑 3.3 节中讨论过的主管人员业绩数据。我们需要一个回归方程,去研究是哪些素质使得其下属评其为优秀主管。建立方程是力图弄明白管理的过程以及各变量的相对重要性。根据回归方程的用途,这意味着我们希望准确估计回归系数,而这与仅用于预测的方程显然不同。问题中的变量列于表 3.2,数据在表 3.3 中,也可从本书的网站^①获得。

由 Y 关于 X_1, X_2, \dots, X_6 的回归得各方差膨胀因子为

$$\begin{aligned} \text{VIF}_1 &= 2.7, \quad \text{VIF}_2 = 1.6, \quad \text{VIF}_3 = 2.3, \\ \text{VIF}_4 &= 3.1, \quad \text{VIF}_5 = 1.2, \quad \text{VIF}_6 = 2.0. \end{aligned}$$

各 VIF 的变化范围(从 1.2 至 3.1)表明这些数据不存在共线性问题。如果我们考察数据的相关矩阵(表 11.1)之特征根,呈现的是同样的现象。相关矩阵的特征根为

$$\begin{aligned} \lambda_1 &= 3.169, \quad \lambda_2 = 1.006, \quad \lambda_3 = 0.763, \\ \lambda_4 &= 0.553, \quad \lambda_5 = 0.317, \quad \lambda_6 = 0.192. \end{aligned}$$

这些特征根的倒数之和为 12.8。因为没有特别小的特征根(条件数为 4.1),且特征根的倒数和仅约为变量数的两倍,所以我们判断本例中的数据不是严重共线性的,我们可以使用刚刚介绍的变量选择方法。

表 11.1 表 3.3 中主管人员业绩数据的相关矩阵

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.000					
X_2	0.558	1.000				
X_3	0.597	0.493	1.000			
X_4	0.669	0.445	0.640	1.000		
X_5	0.188	0.147	0.116	0.377	1.000	
X_6	0.225	0.343	0.532	0.574	0.283	1.000

由前向选择法得到的结果列于表 11.2。我们给出了每个方程所含的变量, RMS, 及 C_p 统计量的值。最后一列显示的是由 FS 法得到的子集在同样大小的子集中的名次(根据 RMS 排名)。 p 的值是方程中的预测变量数, 包括一个常数项。使用的两种停止规则为:

1. 当 t -检验的最小绝对值小于 $t_{0.05}(n-p)$ 时停止。

^① <http://www.ilr.cornell.edu/~hadi/RABE>

2. 当 t -检验的最小绝对值小于 1 时停止。

第一条规则比较严格, 在选进变量 X_1 、 X_3 后终止; 第二条规则不太严格, 在选进变量 X_1 、 X_3 及 X_6 后终止。

表 11.2 用前向选择法选择的变量

方程中的变量	$\min(t)$	RMS	C_p	p	名次
X_1	7.74	6.993	1.41	2	1
X_1X_3	1.57	6.817	1.11	3	1
$X_1X_3X_6$	1.29	6.734	1.60	4	1
$X_1X_3X_6X_2$	0.59	6.820	3.28	5	1
$X_1X_3X_6X_2X_4$	0.47	6.928	5.07	6	1
$X_1X_3X_6X_2X_4X_5$	0.26	7.068	7.00	7	—

表 11.3 用后向剔除法选择的变量

方程中的变量	$\min(t)$	RMS	C_p	p	名次
$X_1X_2X_3X_4X_5X_6$	0.26	7.068	7.00	7	—
$X_1X_2X_3X_4X_6$	0.47	6.928	5.07	6	1
$X_1X_2X_3X_6$	0.59	6.820	3.28	5	1
$X_1X_3X_6$	1.29	6.734	1.60	4	1
X_1X_3	1.57	6.817	1.11	3	1
X_1	7.74	6.993	1.41	2	1

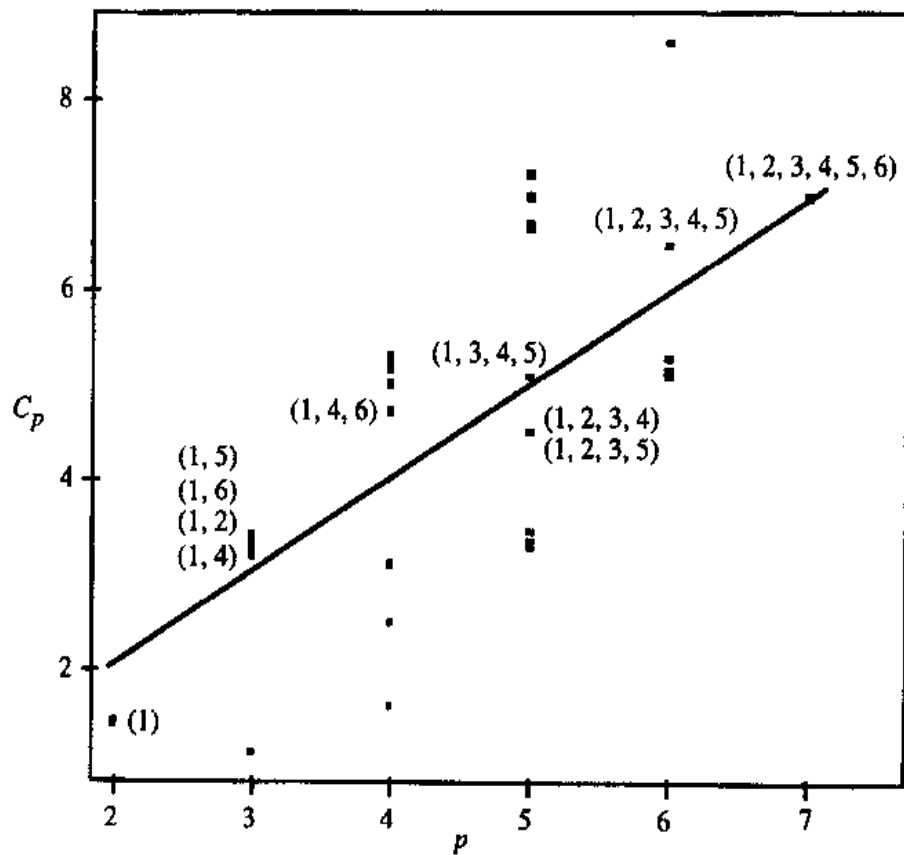
应用 BE 法得到的结果列于表 11.3。结构上它们与表 11.2 相同。对于 BE, 我们使用如下停止规则:

1. 当 t -检验的最小绝对值大于 $t_{0.05}(n-p)$ 时停止。
2. 当 t -检验的最小绝对值大于 1 时停止。

采用第一条停止规则选得的变量是 X_1 、 X_3 。采用第二条停止规则选得的变量是 X_1 、 X_3 和 X_6 。对于这个问题, FS 与 BE 给出了相同的方程, 但情况不总是这样的 (例子在 11.12 节中给出)。我们选择方程

$$Y = 13.58 + 0.62X_1 + 0.31X_3 - 0.19X_6$$

来描述主管人员的业绩。该方程的残差图 (未列出) 是令人满意的。因为现在的问题只有六个变量, 能拟合的、至少含一个变量的方程总共 63 个。所有 63 个方程的 C_p 值列于表 11.4。 C_p 值关于 p 的散点图见图 11.1。根据 C_p 值判断的最好的变量子集在表 11.5 中给出。

图 11.1 主管人员业绩数据: $C_p < 10$ 的子集其 C_p 值关于 p 的散点图表 11.4 C_p 统计量的值 (所有可能的方程)

变量	C_p	变量	C_p	变量	C_p	变量	C_p
1	1.41	1 5	3.41	1 6	3.33	1 5 6	5.32
2	44.40	2 5	45.62	2 6	46.39	2 5 6	47.91
1 2	3.26	1 2 5	5.26	1 2 6	5.22	1 2 5 6	7.22
3	26.56	3 5	27.94	3 6	24.82	3 5 6	25.02
1 3	1.11	1 3 5	3.11	1 3 6	1.60	1 3 5 6	3.46
2 3	26.96	2 3 5	28.53	2 3 6	24.62	2 3 5 6	25.11
1 2 3	2.51	1 2 3 5	4.51	1 2 3 6	3.28	1 2 3 5 6	5.14
4	30.06	4 5	31.62	4 6	27.73	4 5 6	29.50
1 4	3.19	1 4 5	5.16	1 4 6	4.70	1 4 5 6	6.69
2 4	29.20	2 4 5	30.82	2 4 6	25.91	2 4 5 6	27.74
1 2 4	4.99	1 2 4 5	6.97	1 2 4 6	6.63	1 2 4 5 6	8.61
3 4	23.25	3 4 5	25.23	3 4 6	16.50	3 4 5 6	18.42
1 3 4	3.09	1 3 4 5	5.09	1 3 4 6	3.35	1 3 4 5 6	5.29
2 3 4	24.56	2 3 4 5	26.53	2 3 4 6	17.57	2 3 4 5 6	19.51
1 2 3 4	4.49	1 2 3 4 5	6.48	1 2 3 4 6	5.07	1 2 3 4 5 6	19.51
5	57.91	6	57.95	5 6	58.76		

表 11.5 根据 C_p 统计量选择的变量

方程中的变量	$\min(t)$	RMS	C_p	p	名次
X_1	7.74	6.993	1.41	2	1
$X_1 X_4$	0.47	7.093	3.19	3	2
$X_1 X_4 X_6$	0.69	7.163	4.70	4	5
$X_1 X_3 X_4 X_5$	0.07	7.080	5.09	5	6
$X_1 X_2 X_3 X_4 X_5$	0.11	7.139	6.48	6	4
$X_1 X_2 X_3 X_4 X_5 X_6$	0.26	7.068	7.00	7	-

可见根据 C_p 选出的子集既不同于根据残差均方选出的, 也不同于由变量选择方法获得的。这反常现象提醒读者该记住关于 C_p 统计量的一个要点。在使用 C_p 统计量时, 需要 σ^2 的一个估计。通常, 该估计通过全模型的残差平方和获得。如果全模型包含了大量没有解释能力的变量 (即它们的总体回归系数为 0), 则由全模型之残差平方和获得的 σ^2 的估计将偏大。因为分母中损失的自由度与残差平方和的减小量不平衡。若 $\hat{\sigma}^2$ 偏大, 那 C_p 值就偏小。为使 C_p 正常地起作用, 必须提供 σ^2 的一个好的估计。如果没有好的 σ^2 的估计, 那 C_p 的作用有限。在目前这个例子中, 含六个变量的全模型的 RMS 大于含 X_1, X_3, X_6 三个变量的模型的 RMS。因此, C_p 值失真了, 对于本例的变量选择不是非常有用。我们描述的这类情形可通过审视 p 不同时的 RMS 来发现。RMS 起初随 p 增大而下降, 但后阶段则增大。该表现说明后来的变量对减少残差平方和没有显著贡献。有效运用 C_p 要求同时监视 RMS 以避免失真。

11.11 共线性数据的变量选择

在第 9 章中曾指出, 对于共线性数据, 标准分析会导致严重失真。因此, 在这种场合下我们建议一套不同的选择变量的方法。当相关矩阵有一个或多个较小特征根时, 表明存在共线性。变量较少时, 我们可评价全部可能的方程并用前面介绍的方法从中选取一个方程。但变量较多时, 这方法不可行。

人们提出了两种不同的方法来解决这个问题。第一种方法试图通过剔除一些变量来消解数据的共线性。变量间的共线性结构可被很小的特征根对应的特征向量揭示 (见第 9、10 章)。一旦共线性关系被确认, 那就可删去部分变量得到一个简约的、非共线性的数据集。然后我们再运用前面介绍的方法。第二种方法采用岭回归作为主要工具。我们假定读者已经熟悉岭回归的基本术语和概念 (第 10 章)。实践中几乎总使用第一种方法 (明智地剔除些相关变量)。

11.12 凶杀数据

在一项调查中, 研究手枪对底特律日益上升的凶杀率所起的作用, 收集了 1961-1973 年的数据。数据发表于 Gunst and Mason (1980), p360。响应变量 (凶

杀率) 及被认为对凶杀率的增长有影响或有关的预测变量的定义列于表 11.6, 数据列于表 11.7 和 11.8。数据也可从本书的网站上找到。

表 11.6 凶杀数据: 变量的描述

变量	符号	描述
1	FTP	每 100000 人口中全职警察数
2	UEMP	失业人口百分比
3	M	制造业工人数 (单位: 千人)
4	LIC	每 100000 人口中签发的持手枪许可证数
5	GR	每 100000 人口中签发的手枪执照数
6	CLEAR	凶杀案中因拘捕而结案的比例
7	W	人口中白种男性数
8	NMAN	非制造业工人数 (单位: 千人)
9	G	政府机构中的工人数 (单位: 千人)
10	HE	平均小时收入
11	WE	平均周收入
12	H	每 100000 人口中凶杀数

表 11.7 凶杀数据的第一部分

年份	FTP	UNEMP	M	LIC	GR	CLEAR
1961	260.35	11.0	455.5	178.15	215.98	93.4
1962	269.80	7.0	480.2	156.41	180.48	88.5
1963	272.04	5.2	506.1	198.02	209.57	94.4
1964	272.96	4.3	535.8	222.10	231.67	92.0
1965	272.51	3.5	576.0	301.92	297.65	91.0
1966	261.34	3.2	601.7	391.22	367.62	87.4
1967	268.89	4.1	577.3	665.56	616.54	88.3
1968	295.99	3.9	596.9	1131.21	1029.75	86.1
1969	319.87	3.6	613.5	837.80	786.23	79.0
1970	341.43	7.1	569.3	794.90	713.77	73.9
1971	356.59	8.4	548.8	817.74	750.43	63.4
1972	376.69	7.7	563.4	583.17	1027.38	62.5
1973	390.19	6.3	609.3	709.59	666.50	58.9

我们借这些数据来说明, 在共线性情况下机械地套用诸如 FS、BE 等变量选择方法的危险性。我们感兴趣的是拟合模型

$$H = \beta_0 + \beta_1 G + \beta_2 M + \beta_3 W + \varepsilon.$$

变量经中心化、尺度变化后, 模型改变为

$$\hat{H} = \theta_1 \tilde{G} + \theta_2 \tilde{M} + \theta_3 \tilde{W} + \varepsilon'. \quad (11.8)$$

表 11.8 凶杀数据的第二部分

年份	W	NMAN	G	HE	WE	H
1961	558724	538.1	133.9	2.98	117.18	8.60
1962	538584	547.6	137.6	3.09	134.02	8.90
1963	519171	562.8	143.6	3.23	141.68	8.52
1964	500457	591.0	150.3	3.33	147.98	8.89
1965	482418	626.1	164.3	3.46	159.85	13.07
1966	465029	659.8	179.5	3.60	157.19	14.57
1967	448267	686.2	187.5	3.73	155.29	21.36
1968	432109	699.6	195.4	2.91	131.75	28.03
1969	416533	729.9	210.3	4.25	178.74	31.49
1970	401518	757.8	223.8	4.47	178.30	37.39
1971	398046	755.3	227.7	5.04	209.54	46.26
1972	373095	787.0	230.9	5.47	240.05	47.24
1973	359647	819.8	230.2	5.76	258.05	52.33

OLS 的结果列于表 11.9。这个模型中预测变量数是否可能减少？假如标准假定成立，那么变量 G 的 t -检验绝对值 (0.68) 很小，也就说明相应的回归系数是不显著的， G 可从模型中剔除。现在，让我们用前向选择和后向剔除法来看哪些变量被选中。对标准化的变量实现这两种方法，我们所需要的回归结果总结在表 11.10 中。出于比较的目的，我们在这个表中给出了每个模型的系数估计及其 t -检验值，以及修正的复相关系数平方 R_a^2 。

表 11.9 凶杀数据：拟合模型 (11.8) 的 OLS 结果

变量	系数	标准误	t -检验	VIF
G	0.235	0.345	0.68	42
M	-0.405	0.090	-4.47	3
W	-1.025	0.378	-2.71	51
$n = 13$	$R^2 = 0.975$	$R_a^2 = 0.966$	$\hat{\sigma} = 0.0531$	$d.f. = 9$

被 FS 首先选入的变量是 G ，因为在三个含单个变量的模型中（表 11.10 中模型 (a) 至 (c)）它的 t -检验绝对值最大。在两个候选的二变量模型中（模型 (d) 至 (e)），模型 (d) 优于模型 (e)。因此，进入方程的第二个变量是 M 。第三个进入方程的变量为 W （模型 (f)），因为其 t -检验显著。但注意， G 的显著性在模型 (a),(d),(f) 中的戏剧性变化。它在模型 (a),(d) 中高度显著，但在模型 (f) 中变得不显著了。共线性是嫌疑对象！

BE 方法始于三变量模型 (f)。首先剔除的变量是 G （因为它的 t -检验绝对值最小），这就得到了模型 (g)。 M 与 W 在模型 (g) 中 t -检验均显著，因而 BE 过程终止。

注意，FS 首先选择的变量 G 正是 BE 首先剔除的变量。也就是说，被 FS 选

表 11.10 凶杀数据: 系数估计, t -检验值, 修正的复相关系数平方 R_a^2

变量	模型						
	(a)	(b)	(c)	(d)	(e)	(f)	(g)
G: 系数	0.96			1.15	0.87	0.24	
t -检验	11.10			11.90	1.62	0.68	
M: 系数		0.55		-0.27		-0.40	-0.43
t -检验		2.16		-2.79		-4.47	-5.35
W: 系数			0.95		-0.09	-1.02	-1.28
t -检验			-9.77		-0.17	-2.71	-15.90
R_a^2	0.91	0.24	0.89	0.95	0.90	0.97	0.97

为三个变量中最重要变量 G , 却被 BE 认为最不重要! 和其他情况一样, 这种反常结果的原因是共线性。相关矩阵的特征根为 $\lambda_1 = 2.65, \lambda_2 = 0.34, \lambda_3 = 0.011$, 产生了很大的条件数 ($\kappa = 15.6$)。三个变量中的两个 (G 和 W) 的 VIF 很大 (42 和 51)。特征根的倒数和也非常大 (96)。除了共线性, 因为各观测按时间顺序取得 (1961 至 1973 年), 我们这里处理的是时间序列数据。因此, 误差项可能是自相关的 (参阅第 8 章)。考察各对变量的散点图将揭示数据中的另外一些问题。

本例明确地显示了, 对多重共线性数据运用变量选择方法自动地选择变量, 会导致选出一个错误的模型。在 11.13 节与 11.14 节中, 我们在多重共线性场合使用岭回归来处理变量选择问题。

11.13 运用岭回归选择变量

岭回归的一个目标是生成一个系数稳定的回归方程。系数稳定指的是它们不受估计数据中微小变动的影响。一个好的变量选择方法的目标是: (1) 选择一组变量, 对研究的过程提供清晰的理解; (2) 构建一个方程, 对研究中未涉及到的预测变量值, 给出相应的响应变量的精确预报。看来, 好的变量选择方法与岭回归有着非常相似的目标, 因此, 岭回归可用来实现变量选择。

变量选择是通过考察岭迹, 即岭回归系数关于岭参数 k 的图形进行的。一个共线性系统的岭迹的典型模式已在第 10 章中介绍过了。岭迹可用于从方程中剔除变量。剔除变量的指导性准则为:

1. 剔除系数稳定但绝对值很小的变量。因为岭回归处理的是标准化的数据, 故不同系数的数值大小可直接比较。
2. 剔除系数不稳定而无预测能力的变量, 即趋于 0 的不稳定系数。
3. 剔除一个或多个系数不稳定的变量。用剩下来的 p 个变量建立回归方程。

在上述步骤中, 每做一步, 我们就用剩下的变量重新拟合模型, 然后再进行下一步。

对剔除后剩下的变量子集,应该检查其中是否不再有共线性现象。我们用-一个例子来说明这方法。

11.14 大气污染研究中的变量选择

McDonald and Schwing (1970) 进行了一项总死亡率与气候、社会经济及污染变量之间关系的研究。15 个预测变量被选来供研究,列于表 11.11。响应变量是由各种原因导致的经年龄修正后的总死亡率。我们将不从流行病学的角度作评论,只是拿这批数据作为变量选择的一个示范例子。McDonald 和 Schwing 在他们的文章中对这个问题作了非常详细的讨论,有兴趣的读者可以参考这篇文章获得更多信息。

表 11.11 变量的描述、均值和标准差 ($n=60$)

变量	描述	均值	标准差
X_1	年均降水量(英寸)	37.37	9.98
X_2	一月份平均气温(华氏度)	33.98	10.17
X_3	七月份平均气温(华氏度)	74.58	4.76
X_4	65 岁以上人口百分比	8.80	1.46
X_5	每个家庭人口	3.26	0.14
X_6	接受学校教育年数的中位数	10.97	0.85
X_7	健康住宅单元的百分比	80.92	5.15
X_8	每平方英里人口数	3876.05	1454.10
X_9	非白种人口百分比	11.87	8.92
X_{10}	白领职业的百分比	46.08	4.61
X_{11}	收入低于 3000 美元的家庭百分比	14.37	4.16
X_{12}	碳氢化合物的相对潜在污染	37.85	91.98
X_{13}	氮氧化化合物的相对潜在污染	22.65	46.33
X_{14}	二氧化硫的相对潜在污染	53.77	63.39
X_{15}	相对湿度	57.67	5.37
Y	各种原因导致的经年龄修正后的总死亡率	940.36	62.21

原始数据我们没有,但表 11.12 给出了响应与 15 个预测变量的相关矩阵。仅仅根据相关矩阵作分析并不是一项好的实践,因为没有原始数据,我们就不能进行诊断,而这对于任何全面的数据分析来说都是必须的。我们仍是在线性回归模型的标准假定成立的假设下开始分析。从变量的性质可以想像,某些变量之间是高度相关的。如果我们考察相关矩阵的特征根的话,共线性的证据是明显的。这

些特征根为

$$\begin{aligned}\lambda_1 &= 4.5272, & \lambda_6 &= 0.9605, & \lambda_{11} &= 0.1665, \\ \lambda_2 &= 2.7547, & \lambda_7 &= 0.6124, & \lambda_{12} &= 0.1275, \\ \lambda_3 &= 2.0545, & \lambda_8 &= 0.4729, & \lambda_{13} &= 0.1142, \\ \lambda_4 &= 1.3487, & \lambda_9 &= 0.3708, & \lambda_{14} &= 0.0460, \\ \lambda_5 &= 1.2227, & \lambda_{10} &= 0.2163, & \lambda_{15} &= 0.0049.\end{aligned}$$

有两个非常小的特征根；最大的特征根比最小特征根大近 1000 倍。特征根的倒数之和为 263，是变量数的近 17 倍。数据显示出有力的共线性证据。

表 11.12 表 11.11 中变量的相关矩阵

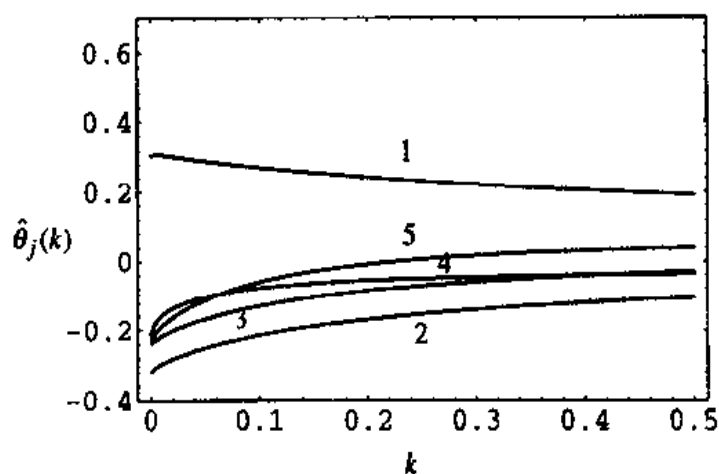
	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1.0000	.0922	.5033	.1011	.2634	-.4904	-.4903	-.0035
X_2		1.0000	.3463	-.3981	-.2092	.1163	.0139	-.1001
X_3			1.0000	-.4340	.2623	-.2385	-.4155	-.0610
X_4				1.0000	-.5091	-.1389	.0649	.1620
X_5					1.0000	-.3951	-.4095	-.1843
X_6						1.0000	.5515	-.2439
X_7							1.0000	.1806
X_8								1.0000
	X_9	X_{10}	X_{11}	X_{12}	X_{13}	X_{14}	X_{15}	Y
X_1	.4132	-.2973	.5066	-.5318	-.4873	-.1069	-.0773	.5095
X_2	.4538	.2380	.5653	.3508	.3210	-.1078	.0679	-.0300
X_3	.5753	-.0214	.6193	-.3565	-.3377	-.0993	-.4528	.2770
X_4	-.6378	-.1177	-.3098	-.0205	-.0021	.0172	.1124	-.1746
X_5	.4194	-.4257	.2599	-.3882	-.3584	-.0041	-.1357	.3573
X_6	-.2088	.7032	-.4033	.2868	.2244	-.2343	.1765	-.5110
X_7	-.4091	.3376	-.6806	.3859	.3476	.1180	.1224	-.4248
X_8	-.0057	-.0318	-.1629	.1203	.1653	.4321	-.1250	.2655
X_9	1.0000	-.0044	.7049	-.0259	.0184	.1593	-.1180	.6437
X_{10}		1.0000	-.1852	.2037	.1600	-.0685	.0607	-.2848
X_{11}			1.0000	-.1298	-.1025	-.0965	-.1522	.4105
X_{12}				1.0000	.9838	.2823	-.0202	-.1772
X_{13}					1.0000	.4094	-.0459	-.0774
X_{14}						1.0000	-.1026	.4259
X_{15}							1.0000	-.0885
Y								1.0000

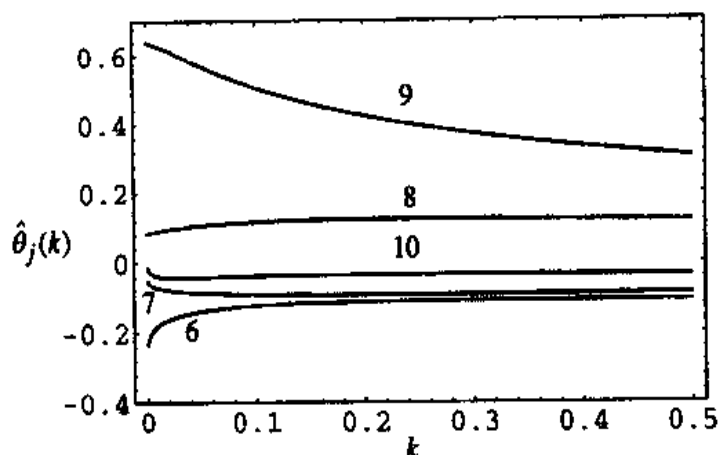
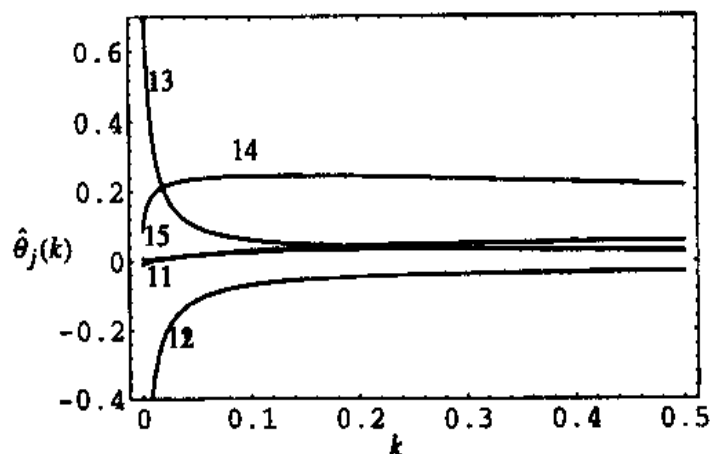
表 11.13 给出了对经中心化与尺度变换的数据拟合线性模型的初始的 OLS 结果。尽管该模型的 R^2 很大，但某些系数估计的 t -检验绝对值较小。在多重共线性存在时，小的 t -检验绝对值并不一定意味着相应的变量不重要，可能由于方差膨胀导致。从表 11.13 可见， VIF_{12} 和 VIF_{13} 非常大。

表 11.13 大气污染数据的 OLS 回归输出 (15 个预测变量)

变量	系数	标准误	t-检验	VIF
X_1	0.306	0.148	2.063	4.11
X_2	-0.318	0.181	-1.755	6.13
X_3	-0.237	0.146	-1.627	3.97
X_4	-0.213	0.200	-1.064	7.46
X_5	-0.232	0.152	-1.527	4.31
X_6	-0.233	0.161	-1.448	4.85
X_7	-0.052	0.146	-0.356	3.97
X_8	0.084	0.094	0.890	1.66
X_9	0.640	0.190	3.359	6.78
X_{10}	-0.014	0.123	-0.112	2.84
X_{11}	-0.010	0.216	-0.042	8.72
X_{12}	-0.979	0.724	-1.353	97.92
X_{13}	0.983	0.747	1.316	104.22
X_{14}	0.090	0.150	0.599	4.21
X_{15}	0.009	0.101	0.093	1.91
$n = 60$	$R^2 = 0.764$	$R_a^2 = 0.648$	$\hat{\sigma} = 0.073$	$d.f. = 44$

15 个回归系数的岭迹作于图 11.2 至 11.4。每张图显示五条曲线。如果我们把所有 15 条曲线画在一张图上, 那会过于混杂而无法看清每条曲线。为使这三张图可比, 我们使各图的刻度保持一致。从岭迹可见, 有些系数很不稳定, 有些系数不论岭参数 k 取何值总是很小。

图 11.2 大气污染数据: $\hat{\theta}_1, \dots, \hat{\theta}_5$ 的岭迹 (15 个变量的模型)

图 11.3 大气污染数据: $\hat{\theta}_6, \dots, \hat{\theta}_{10}$ 的岭迹 (15 个变量的模型)图 11.4 大气污染数据: $\hat{\theta}_{11}, \dots, \hat{\theta}_{15}$ 的岭迹 (15 个变量的模型)

现在, 我们根据多重共线性数据变量选择的指导性准则来操作。根据第一条准则, 我们剔除变量 7, 8, 10, 11 及 15。它们的平坦的岭迹表明, 这些变量的系数都比较稳定, 但都非常小。尽管变量 14 在 $k=0$ 时的系数很小 (见表 11.13), 但当 k 增大时, 其值也迅速增大。因此, 从这一点看, 它不该被剔除。

我们再对余下的 10 个变量 1, 2, 3, 4, 5, 6, 9, 12, 13 及 14 重复同样的分析。相应的 OLS 结果在表 11.14 中给出。仍有多重共线性的迹象。最大特征根 $\lambda_1 = 3.377$ 大约是最小特征根 $\lambda_{10} = 0.005$ 的 600 倍。变量 12, 13 的 VIF 仍然很大。相应的岭迹作于图 11.5 与 11.6 中。变量 14 的系数在 $k=0$ 处仍很小, 但仍是随 k 增大而增大。因此, 在这阶段, 它仍应被保留在模型中。另外九个变量也都不符合第一条准则。

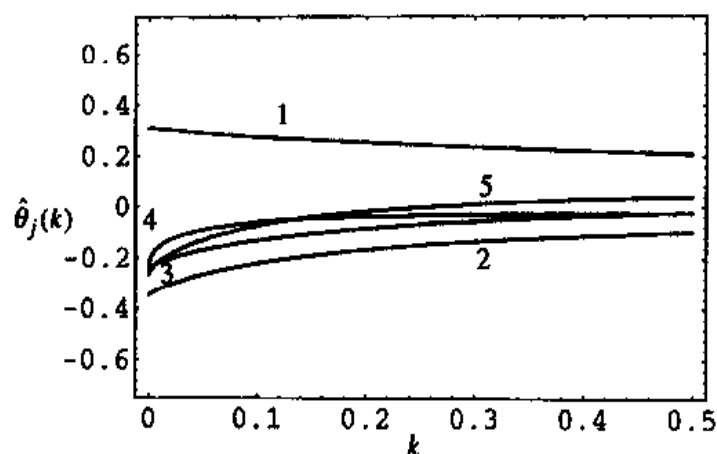
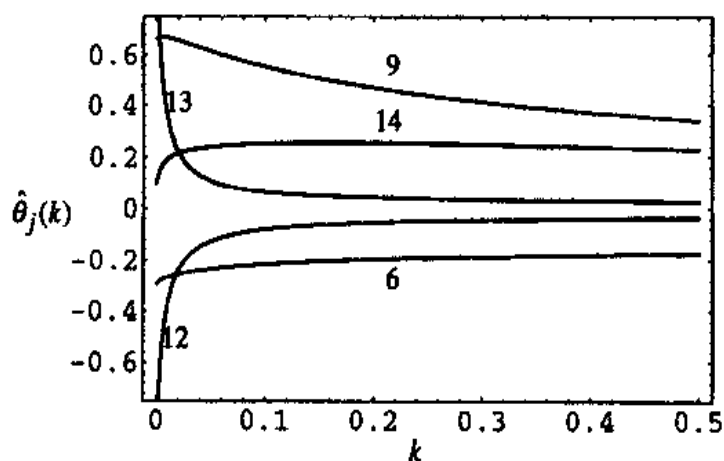
图 11.5 大气污染数据: $\hat{\theta}_1, \dots, \hat{\theta}_5$ 的龄迹 (10 个变量的模型)图 11.6 大气污染数据: $\hat{\theta}_6, \hat{\theta}_9, \hat{\theta}_{12}, \hat{\theta}_{13}, \hat{\theta}_{14}$ 的龄迹 (10 个变量的模型)

表 11.14 大气污染数据的 OLS 回归输出 (10 个预测变量)

变量	系数	标准误	t-检验	VIF
X_1	0.306	0.135	2.260	3.75
X_2	-0.345	0.119	-2.907	2.88
X_3	-0.244	0.108	-2.256	2.39
X_4	-0.222	0.175	-1.274	6.22
X_5	-0.268	0.137	-1.959	3.81
X_6	-0.292	0.103	-2.842	2.15
X_9	0.664	0.140	4.748	3.99
X_{12}	-1.001	0.658	-1.522	88.30
X_{13}	1.001	0.673	1.488	92.40
X_{14}	0.098	0.127	0.775	3.29
$n = 60$	$R^2 = 0.760$	$R_a^2 = 0.711$	$\hat{\sigma} = 0.070$	$d.f. = 49$

第二条准则建议, 剔除系数不稳定的、且趋向于 0 的变量。考察图 11.5 与

图 11.6 中的岭迹, 发现变量 12, 13 属于这一类。

关于剩下 8 个变量的 OLS 结果列于表 11.15。共线性现象消失了。现在, 最大、最小的特征根分别为 2.886 与 0.094, 相应的条件数 $\kappa = 5.5$ 较小。特征根倒数之和为 23.5, 约是变量数的三倍。所有的 VIF 值都小于 10。因为余下的变量是非共线性的, 因而接下来, 我们就可运用 11.7 及 11.8 节中讨论的非共线性数据的变量选择方法了。这留给读者作为练习。

表 11.15 大气污染数据的 OLS 回归输出 (8 个预测变量)

变量	系数	标准误	t-检验	VIF
X_1	0.331	0.120	2.765	2.911
X_2	-0.351	0.106	-3.313	2.279
X_3	-0.217	0.104	-2.087	2.191
X_4	-0.155	0.163	-0.946	5.419
X_5	-0.221	0.134	-1.656	3.621
X_6	-0.270	0.102	-2.654	2.097
X_9	0.692	0.133	5.219	3.567
X_{14}	0.230	0.083	2.767	1.405
$n = 60$	$R^2 = 0.749$	$R_a^2 = 0.709$	$\hat{\sigma} = 0.070$	$d.f. = 51$

分析这些大气污染数据的另一条途径是: 原始的 15 个变量中的共线性现象实际是非常简单的一种情况, 仅与两个变量 (12 及 13) 密切相关。所以, 分析可在剔除其中任何一个后展开。读者可以验证, 余下的 14 个变量是非共线性的。于是, 就可采用非共线性数据变量选择的标准程序了。我们把这也留给读者作为练习。

我们在分析大气污染数据时, 没有用到第三条准则, 但有些场合下这条准则是需要的。我们应当注意到, 在这个例子中, 岭回归被成功地用作了变量选择的工具。在中间阶段, 因为我们发现所选的变量是非共线性的, 所以也利用了标准的 OLS。

Henderson and Wellenam (1981) 对这批数据给出了非岭回归的分析。他们全面地分析了这批数据, 详情请读者参阅他们的文章。

我们希望, 我们的讨论能使大家清晰地认识到, 变量选择体现着艺术与科学的结合, 应当仔细并慎重。我们概括了一系列方法和指导性准则, 而没有规定一套正式的程序。总之, 我们必须强调前面已经叙述过的这一观点, 变量选择不应该被机械地、完全地作为一种目标来追求, 而应当作为对所研究数据的结构的一种探索。正如所有真正的探索一样, 研究者应以理论、直觉及常识作指导。

11.15 拟合回归模型的可能策略

在本章总结性的这一节中, 我们概括了一套可能的实施步骤, 可用来圆满地拟合回归模型。我们首先强调, 没有唯一正确的做法。读者可能更习惯于某套与

此不同的步骤,那就尽管放心地去按那套步骤做。几乎所有的情况下,这里描述的分析总能导出有意义的、可解释的模型,这些模型在实际应用中往往也有用。

假定响应变量为 Y , 一组变量为 X_1, X_2, \dots, X_p , 我们欲考察 Y 与其中部分或全部变量之间的关系。该变量集, X_1, X_2, \dots, X_p , 往往出于对外在主题的考虑而形成的。这个集合常常过大, 我们希望获得一个可以接受的简约的集合。我们的目标是构建一个合理的且可行的回归模型。一套可能的实施步骤为:

1. 逐个地考察变量 (Y, X_1, X_2, \dots, X_p)。这可以通过计算概述性统计量来实施, 也可通过看直方图、点图或箱线图 (参阅第 4 章) 来完成。各变量的分布不当太偏斜, 变化范围也不该过大。查看异常值 (检查抄写错误)。作变换以获得对称性、消除偏斜。此时, 常用对数变换 (参见第 6 章)。
2. 画成对变量的散点图。当预测变量数 p 较大时, 这可能不可行。成对变量的散点图对两两变量间的关系提供的信息非常丰富。审视相关矩阵, 指出明显的共线性问题。剔除冗余变量。计算相关矩阵的条件数, 对共线性的严重程度有个大致的了解 (第 9、10 章)。
3. 拟合完全的线性回归模型。删除无显著解释能力的变量 (t -检验不显著)。对于简化的模型, 考察残差:
 - (a) 检查线性性。若非线性, 对变量作变换 (见第 6 章)。
 - (b) 检查异方差与自相关 (对时间序列数据) 现象。如果存在, 采取合适的措施 (见第 7、8 章)。
 - (c) 查看异常点、高杠杆点及强影响点。若存在, 则采取适当的措施 (见第 4 章)。
4. 考察是否有多余的变量可剔除而无损模型完整性。考察是否有新变量可引进模型 (添加变量图, 残差加分量图) (见第 4、11 章)。重复步骤 3。
5. 对最后拟合的模型, 检查方差膨胀因子。确保有令人满意的残差图, 以及没有任何不利的诊断信息 (见第 3、5、6、9 章)。如若必要, 重复步骤 4。
6. 接着, 应当尽力对拟合的模型作验证。当数据量很大时, 可用部分数据拟合模型, 用剩下的数据作验证。诸如自助法、刀切法、交叉核实等重抽样方法也是可能的选择, 尤其当数据量不太大时 (参见 Efron(1982) 及 Diaconis and Efron (1983))。

实践中, 我们所述的这些步骤常常不是依次实施的, 而是同时进行的。我们所述的这一过程是一个循环反复的过程, 为获得一个满意的模型, 可能有必要重复上面概括的步骤多次。这些步骤列举了在建立一个满意的模型时必须考虑的种种因素。

在我们概括的步骤中, 还没有包括另一个重要的方面, 即分析者对于需建模的那个领域的专业知识。在建模过程中总是应该利用这方面的知识, 这常常会加快获得满意模型的进程, 因为这将大大地有助于合适地选择变量及相应的变换。该说的说了, 该做的做了, 之后, 统计建模便是一项艺术。我们讲的全部技术, 就是有条不紊地开展这项任务的一些工具。

11.16 文献

有大量关于变量选择的文献散见于各种统计杂志。在 Hocking (1976) 中可找到一个非常全面的评论, 还附带非常广泛的参考文献。在 Daniel and Wood (1980) 一书中, 对变量选择作了详细讨论, 其中特别强调 C_p 统计量。Mallows (1973) 对 C_p 统计量的应用作了改进。Draper and Smith (1998) 一书中讨论了变量选择方法。Hoerl and Kennard (1970) 及 McDonald and Schwing (1973) 讨论了岭回归在变量选择方面的运用。

习 题

- 11.1 如我们在 11.14 节中所见, 存在下面三个非共线性的预测变量子集。对每个子集采用一种或多种方法作变量选择, 并比较最终导出的模型:
- (a) 由八个变量构成的子集: 1, 2, 3, 4, 5, 6, 9 和 14。
 - (b) 由剔除变量 12 后的 14 个变量构成的子集。
 - (c) 由剔除变量 13 后的 14 个变量构成的子集。
- 11.2 表 11.13 中是各变量标准化形式的回归系数估计值, 因为它们由响应与预测变量的相关矩阵算得。用表 11.11 中各变量的均值与标准差, 写出用原始变量表示的回归方程 (未作中心化与尺度变换的回归方程)。
- 11.3 在 11.12 节讨论的凶杀数据中, 我们看到, 在拟合模型 (11.8) 时, FS 和 BE 方法给出了相互矛盾的结果。事实上, 这批数据中还存在其他若干个子集 (不一定含三个预测变量), 用 FS 和 BE 方法会给出矛盾的结果。请找出一个或多个这样的子集。
- 11.4 采用变量选择方法, 找出表 11.7 和 11.8 中预测变量的一个或多个、能最好地解释响应变量 H 之变异的子集。这里姑且认为采用变量选择方法是合适的。
- 11.5 房产估价: 科学全面估价法 (Scientific mass appraisal) 是将线性回归方法应用于房产估价问题的一项技术。其目的是, 根据建筑物的某些物理特性及购房税 (地方税、教育税、县税) 去预测住宅的销售价格。我们从杂志 Multiple Listing (Vol.87) 中获得了宾夕法尼亚州伊利市的 24 个观测, 该市在目录中被归为第 12 区 (Area 12)。这些数据 (见表 11.17) 最初是由 Narula and Wellington (1977) 提供的。变量清单在表 11.16 中给出。
- 回答下列问题, 在每一项中, 都用合适的分析说明你的理由。
- (a) 若要拟合一个反映销售价格与各项税收、各建筑特性之间关系的回归模型, 你会将所有变量包含在内吗?
 - (b) 一个经验老到的房地产代理商建议, 地方税收、房间数、房龄就能充分刻画售价了。你同意吗?
 - (c) 该项目请来的一位房地产专家作了如下推理: 一套住宅的销售价格由其吸引力决定, 当然这是该建筑的物理特性的一个函数。综合估价已反映于房东所付的地方税中了; 因此, 销售价格最佳的预测因子是地方税。所

以, 在一个包含了地方税的回归方程中, 各建筑特性是冗余的。销售价格单单关于地方税的方程可能就足够了。请考察若干个模型, 来检验该论断。你同意该论断吗? 提出你认为最为充分的一个或多个、用于预测宾夕法尼亚州伊利市的住宅销售价格的模型。

表 11.16 表 11.17 中数据的变量清单

变量	定义
Y	房屋的售价(千美元)
X_1	税款(地方税、教育税、县税)(千美元)
X_2	盥洗室间数
X_3	大小(千平方英尺)
X_4	起居空间(千平方英尺)
X_5	车库间数
X_6	房间数
X_7	卧室数
X_8	房龄(年)
X_9	壁炉数

11.6 请查阅表 9.18、9.19 中的汽油消耗数据。

(a) 你会用所有变量去预测各类汽车的油耗吗? 请作解释。

(b) 有人提出了六个模型供选择:

- i. Y 关于 X_1 的回归;
- ii. Y 关于 X_{10} 的回归;
- iii. Y 关于 X_1 与 X_{10} 的回归;
- iv. Y 关于 X_2 与 X_{10} 的回归;
- v. Y 关于 X_8 与 X_{10} 的回归;
- vi. Y 关于 X_8 、 X_5 与 X_{10} 的回归。

在这些回归模型中, 你会选择哪个去预测汽车的油耗? 你能否提出一个更好的模型?

(c) 逐个地画 Y 关于 X_1 、 X_2 、 X_8 及 X_{10} 的散点图。这些散点图是否表明 Y 与那 11 个预测变量之间的关系可能不是线性的?

(d) 油耗是由驾驶各辆载重相同的汽车、行驶同样的道路(约 123 英里长的道路)测得的。有人提议考虑一个新变量 $W = 100/Y$ (加仑/百英里)来替代 Y (英里/加仑)。画 W 关于 X_1 、 X_2 、 X_8 及 X_{10} 的散点图。 W 与那 11 个预测变量的关系, 和 Y 与那 11 个预测变量的关系相比是否更接近线性?

(e) 用 W 代替 Y 再做习题 11.6b。你的结论怎样?

(f) 作 Y 关于 X_{13} 的回归, 其中 $X_{13} = X_8/X_{10}$ 。

(g) 写一份简要的报告描述你的发现。推荐一个用于预测汽车油耗的模型。

表 11.17 各项建筑特性与销售价格

行	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	Y
1	4.918	1.000	3.472	0.998	1.0	7	4	42	0	25.90
2	5.021	1.000	3.531	1.500	2.0	7	4	62	0	29.50
3	4.543	1.000	2.275	1.175	1.0	6	3	40	0	27.90
4	4.557	1.000	4.050	1.232	1.0	6	3	54	0	25.90
5	5.060	1.000	4.455	1.121	1.0	6	3	42	0	29.90
6	3.891	1.000	4.455	0.988	1.0	6	3	56	0	29.90
7	5.898	1.000	5.850	1.240	1.0	7	3	51	1	30.90
8	5.604	1.000	9.520	1.501	0.0	6	3	32	0	28.90
9	5.828	1.000	6.435	1.225	2.0	6	3	32	0	35.90
10	5.300	1.000	4.988	1.552	1.0	6	3	30	0	31.50
11	6.271	1.000	5.520	0.975	1.0	5	2	30	0	31.00
12	5.959	1.000	6.666	1.121	2.0	6	3	32	0	30.90
13	5.050	1.000	5.000	1.020	0.0	5	2	46	1	30.00
14	8.246	1.500	5.150	1.664	2.0	8	4	50	0	36.90
15	6.697	1.500	6.902	1.488	1.5	7	3	22	1	41.90
16	7.784	1.500	7.102	1.376	1.0	6	3	17	0	40.50
17	9.038	1.000	7.800	1.500	1.5	7	3	23	0	43.90
18	5.989	1.000	5.520	1.256	2.0	6	3	40	1	37.90
19	7.542	1.500	5.000	1.690	1.0	6	3	22	0	37.90
20	8.795	1.500	9.890	1.820	2.0	8	4	50	1	44.50
21	6.083	1.500	6.727	1.652	1.0	6	3	44	0	37.90
22	8.361	1.500	9.150	1.777	2.0	8	4	48	1	38.90
23	8.140	1.000	8.000	1.504	2.0	7	3	3	0	36.90
24	9.142	1.500	7.326	1.831	1.5	8	4	31	0	45.80

11.7 查阅表 5.17 中总统选举数据, 如同习题 9.3, 考虑拟合一个 V 关于所有变量的模型 (包括表示选举年份的时间趋势), 加上尽可能多的二变量或三变量交互作用项。

(a) 从习题 9.3a 中的模型出发。运用不少于两种变量选择的方法, 选出在预测未来的总统选举中表现最好的一个或若干个模型。

(b) 从习题 9.3d 中的模型出发, 再做上面的习题。

(c) 在前面得到的那些模型中, 你认为哪一个最合适?

(d) 用你选出的模型去预测, 2000、2004 及 2008 年美国总统选举中某个总统候选人的预期得票率。

(e) 在上面的三个预测中, 你认为哪个预测比另两个更精确? 请作解释。

(f) 在本书这一版付印之际, 2000 年总统选举的结果尚未揭晓。如果你碰巧在 2000 年选举之后读到了本书, 那你习题中的预测是否正确呢?

11.8 香烟消费数据: 考虑在习题 3.14 中描述的、表 3.17 中给出的香烟消费数据。该组织希望建立一个回归方程, 去刻画全州范围内香烟消费 (人均) 与各社会经济变量、人口统计变量之间的关系, 并确定这些变量在预测香烟消费时是否有用。

- (a) 建立一个解释该州人均香烟销售额的线性回归模型。在分析中, 特别注意异常点。看剔除一个异常点是否影响你的发现。在决定最后的模型前请审视残差图。除非分析表明需要全部变量, 否则模型不必包含全部变量。你的目标是找到最少的变量, 去有意义地、合理地描述该州的香烟销售额。
- (b) 写一篇报告描述你的发现。

附录: 误设模型的影响

在本附录中, 我们采用矩阵记号讨论不正确的模型设定对于回归系数的估计及预测值的影响。定义下列矩阵、向量:

$$\mathbf{X} = \begin{bmatrix} x_{10} & x_{11} & \cdots & x_{1p} & x_{1(p+1)} & \cdots & x_{1q} \\ x_{20} & x_{21} & \cdots & x_{2p} & x_{2(p+1)} & \cdots & x_{2q} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n0} & x_{n1} & \cdots & x_{np} & x_{n(p+1)} & \cdots & x_{nq} \end{bmatrix}, \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \\ \text{-----} \\ \beta_{p+1} \\ \vdots \\ \beta_q \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix},$$

其中 $x_{i0} = 1, i = 1, \dots, n$ 。将 n 行、 $q+1$ 列的矩阵 \mathbf{X} 分成 \mathbf{X}_p 和 \mathbf{X}_r 两个子矩阵, 阶数分别为 $n \times (p+1)$ 和 $n \times r$, 其中 $r = q - p$ 。向量 $\boldsymbol{\beta}$ 类似地分为 $\boldsymbol{\beta}_p$ 、 $\boldsymbol{\beta}_r$, 分别含 $p+1$ 、 r 个元素。

包含所有 q 个变量的完全线性模型为

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbf{X}_p\boldsymbol{\beta}_p + \mathbf{X}_r\boldsymbol{\beta}_r + \boldsymbol{\varepsilon}, \quad (\text{A.1})$$

其中, 各 ε_i 是独立的、0 均值、同方差、正态分布的误差。

仅含 p 个变量的线性模型 (即含 $p+1$ 个项的方程) 为

$$\mathbf{Y} = \mathbf{X}_p\boldsymbol{\beta}_p + \boldsymbol{\varepsilon}. \quad (\text{A.2})$$

我们将由全模型 (A.1) 获得的 β 之最小二乘估计记为 $\hat{\beta}^*$, 其中

$$\hat{\beta}^* = \begin{pmatrix} \hat{\beta}_p^* \\ \hat{\beta}_r^* \end{pmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}.$$

由子集模型 (A.2) 获得的 β_p 之估计 $\hat{\beta}_p$ 为

$$\hat{\beta}_p = (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T \mathbf{Y}.$$

由 (A.1)、(A.2) 获得的 σ^2 的估计分别记为 $\hat{\sigma}_q^2$ 和 $\hat{\sigma}_p^2$, 即有

$$\hat{\sigma}_q^2 = \frac{\mathbf{Y}^T \mathbf{Y} - \hat{\beta}^{*T} \mathbf{X}^T \mathbf{Y}}{n - q - 1}$$

和

$$\hat{\sigma}_p^2 = \frac{\mathbf{Y}^T \mathbf{Y} - \hat{\beta}_p^T \mathbf{X}_p^T \mathbf{Y}}{n - p - 1}.$$

根据标准的理论可知, $\hat{\beta}^*$ 与 $\hat{\sigma}_q^2$ 分别是 β 与 σ^2 的无偏估计。也可证明

$$E(\hat{\beta}_p) = \beta_p + \mathbf{A}\beta_r,$$

其中

$$\mathbf{A} = (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \mathbf{X}_p^T \mathbf{X}_r.$$

另外还有

$$\begin{aligned} \text{Var}(\hat{\beta}_p) &= (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \sigma^2, \\ \text{Var}(\hat{\beta}^*) &= (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2, \end{aligned}$$

及

$$MSE(\hat{\beta}_p) = (\mathbf{X}_p^T \mathbf{X}_p)^{-1} \sigma^2 + \mathbf{A}\beta_r\beta_r^T \mathbf{A}^T.$$

我们将 $\hat{\beta}_p$ 与 $\hat{\beta}_p^*$ 的性质总结如下:

1. $\hat{\beta}_p$ 是 β_p 的有偏估计, 除非 (1) $\beta_r = 0$ 或 (2) $\mathbf{X}_p^T \mathbf{X}_r = 0$.
2. 矩阵 $\text{Var}(\hat{\beta}^*) - \text{Var}(\hat{\beta}_p)$ 是半正定的, 也就是说, 由全模型获得的回归系数之最小二乘估计的方差比由子集模型获得的相应估计的方差要大。换句话说, 删除变量常常导致余下变量的回归系数估计的方差减小。
3. 若矩阵 $\text{Var}(\hat{\beta}_r^*) - \beta_r\beta_r^T$ 是半正定的, 那么矩阵 $\text{Var}(\hat{\beta}_p^*) - MSE(\hat{\beta}_p)$ 也是半正定的。这意味着, 当被剔除变量的回归系数比全模型中这些系数估计的标准差还小时, 由子集模型获得的回归系数之最小二乘估计的均方误差比由全模型获得的小。
4. 作为 σ^2 的估计, $\hat{\sigma}_p^2$ 通常偏大些。

为查看模型误设对于预测的影响,我们来考察相应于一个观测的预测值,譬如说观测 $\mathbf{x}^T = (\mathbf{x}_p^T : \mathbf{x}_r^T)$ 。用变量全集给出的相应于 \mathbf{x}^T 的预测值记为 \hat{y}^* 。则 $\hat{y}^* = \mathbf{x}^T \hat{\beta}^*$, 均值为 $\mathbf{x}^T \beta$, 预测的方差 $\text{Var}(\hat{y}^*)$ 为:

$$\text{Var}(\hat{y}^*) = \sigma^2(1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}).$$

另一方面,若采用子集模型 (A.2), 预测值为 $\hat{y} = \mathbf{x}_p^T \hat{\beta}_p$, 其均值为

$$E(\hat{y}) = \mathbf{x}_p^T \beta_p + \mathbf{x}_p^T \mathbf{A} \beta_r,$$

预测的方差为

$$\text{Var}(\hat{y}) = \sigma^2(1 + \mathbf{x}_p^T(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{x}_p),$$

预测的均方误差为

$$MSE(\hat{y}) = \sigma^2(1 + \mathbf{x}_p^T(\mathbf{X}_p^T\mathbf{X}_p)^{-1}\mathbf{x}_p) + (\mathbf{x}_p^T \mathbf{A} \beta_r - \mathbf{x}_p^T \beta_r)^2.$$

\hat{y}^* 与 \hat{y} 的性质可以总结如下:

1. \hat{y} 是有偏的, 除非 $\beta_r = 0$ ^①。
 2. $\text{Var}(\hat{y}^*) \geq \text{Var}(\hat{y})$ 。
 3. 若矩阵 $\text{Var}(\hat{\beta}_r^*) - \beta_r \beta_r^T$ 是半正定的, 那么 $\text{Var}(\hat{y}^*) \geq MSE(\hat{y})$ 。
- 在变量选择背景下这些结果的意义及解释已在本章的主体部分给出。

^① 译注: 原文此处为 $\mathbf{X}_p^T \mathbf{X}_r \beta_r = 0$, 那是 $E(\hat{y}) = \mathbf{X}_p^T \beta_p$ 的条件, 并不能保证 \hat{y} 无偏, 显然有误。

12

Logistic 回归

12.1 引言

迄今为止，在我们讨论的回归分析中，一直将响应变量 Y 视为定量变量。而预测变量却既有定量的，也有定性的。先前所述的示性变量就属于后者。然而，确实存在响应变量是定性变量的情况。本章我们就来介绍处理这类问题的一些方法。这些方法与前文讨论的最小二乘法迥然不同。

先来看根据一系列测试的得分选拔人才的例子。经过五年的考察，可将候选人分为“好”、“差”两类。我们关心的是这些测试对于候选人工作业绩的预测能力如何。这里，业绩是一个二值响应变量。我们不妨可以将“好”编码为 1，将“差”编码为 0。预测变量是各项测试的得分。

在致癌因素的研究中，我们收集了若干人的健康记录，包括年龄、性别、抽烟史、日常饮食及家庭病史等变量的数据。响应变量是，一个人得了癌症 ($Y = 1$)，还是没得癌症 ($Y = 0$)。

在金融界，最为关心的是企业的“健康”状况。响应变量是公司的偿付能力 (破产 = 0, 有偿付能力 = 1)，预测变量是公司的各项财务特征。二值响应变量的情形非常普遍，广泛出现于各类统计应用中。

12.2 定性数据的建模

二值响应变量处理的定性数据往往可用 0 或 1 两个数值编码。我们想解决的是对响应取二值之一的概率建模，而不是直接预测其取值。因而前文所述的标准线性模型的局限性是显而易见的。

我们用仅含一个预测变量的简单回归问题来说明这一点。对多元回归的讨论类似。记 π 为在 $X = x$ 时 $Y = 1$ 的概率。假如我们用标准线性模型去描述 π ，那么模型为

$$\pi = Pr(Y = 1|X = x) = \beta_0 + \beta_1 x + \varepsilon. \quad (12.1)$$

由于 π 是一个概率, 故它必定介于 0 和 1 之间。而 (12.1) 给出的线性函数是无界的, 因此不能用于对概率建模。普通最小二乘法的不适用还有一个原因。响应变量 Y 是一个二项分布的随机变量, 其方差是 π 的函数、依赖于 X , 因而方差齐性假定不再成立。尽管我们能采用加权最小二乘法, 但该方法也有问题。因 π 的值未知, 故使用加权最小二乘法时, 我们必须从某个猜测的初始值出发进行迭代。下面, 我们讨论另一种对概率建模的方法, 而不是这种复杂的方法。

12.3 Logit 模型

概率 π 与 X 的关系通常能由 logistic 响应函数来描述。它类似于一条 S-型曲线, 示意图见图 12.1。随着 X 的增大, 一开始, 概率 π 增大得很缓慢, 然后加速, 最后又趋平缓, 但始终不超过 1。这符合直观认识。譬如, 调查表的回收概率与所给的现金报酬之间的函数关系, 或者, 通过一项考试的概率与所花的学习时间之间的函数关系, 都大致如此。

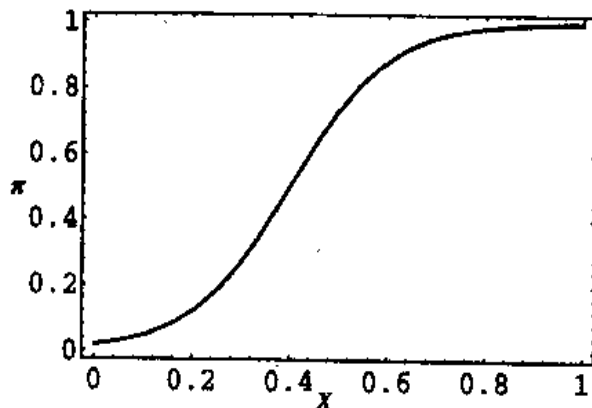


图 12.1 Logistic 响应函数

图 12.1 中的 S-型曲线可由下列模型生成:

$$\pi = Pr(Y = 1|X = x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}, \quad (12.2)$$

其中 e 为自然对数的底。这实际是用 logistic 分布的分布函数来对概率建模。用其他分布函数对概率建模的方法同样也能生成 S-型曲线。譬如, 也可用正态分布的分布函数, 它建立的是 probit 模型。这里我们不讨论 probit 模型, 因为我们觉得 logistic 模型更为简单而且优于 probit 模型。

logistic 模型可以直接推广到多个预测变量的场合。相应地, 概率 π 的模型为

$$\pi = Pr(Y = 1|X_1 = x_1, \dots, X_p = x_p) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}. \quad (12.3)$$

方程 (12.3) 称为 logistic 回归函数。它关于参数 $\beta_0, \beta_1, \dots, \beta_p$ 是非线性的。然而,

通过logit变换可以将它线性化^①。我们不直接分析 π ，而是对它变换后的值进行分析。如果 π 是某事件发生的概率，那么比率 $\pi/(1-\pi)$ 称为该事件的优势比。由于

$$1 - \pi = Pr(Y = 0 | X_1 = x_1, \dots, X_p = x_p) = \frac{1}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}},$$

故

$$\frac{\pi}{1 - \pi} = e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}. \quad (12.4)$$

对(12.4)两边同时取自然对数，可得

$$\begin{aligned} g(x_1, \dots, x_p) &= \log\left(\frac{\pi}{1 - \pi}\right) \\ &= \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p. \end{aligned} \quad (12.5)$$

优势比的对数称为logit。由(12.5)可见，logit变换产生了参数 $\beta_0, \beta_1, \dots, \beta_p$ 的一个线性函数。另外注意，当(12.3)中的 π 在0与1之间取值时， $\log(\pi/(1-\pi))$ 则在 $-\infty$ 与 $+\infty$ 之间取值，因而用线性回归去拟合各logit的值（即对数优势比值）更为合适。

拟合logistic回归也就是采用logistic分布对响应概率建模、并估计(12.3)中模型的参数。在logistic回归中，拟合是对各logit值进行的。logit变化导出了一个关于参数为线性的模型。通常采用极大似然法来估计参数。可采用迭代法获得极大似然估计的数值解。与最小二乘拟合不同的是，这里参数估计不存在精确的表达式。我们不准备深入讨论有关的计算问题，对此，读者可参阅McCullagh and Nelder(1983), Seber(1984), 以及Hosmer and Lemeshow(1989)等文献。

在实际拟合logistic回归时，电脑程序是关键。大多数回归软件包含logistic回归功能。拟合之后，接着面对的就是一组在通常的线性回归中同样要考虑的问题，包括：模型是否合适，哪些变量该保留，拟合的效果如何等等。最小二乘回归中常用的 R^2 , t 和 F 检验等工具不再适用，我们得采用另外的工具去解决同样的问题。做假设检验的方法也不同，因为估计方法采用的是极大似然法而不是最小二乘法。

12.4 例子：破产概率的估计



诊察发现运营不良的金融商业机构，是审计核查的一项重要功能。审计核查的分类失败会导致灾难性的后果，比如，美国1980年代的储蓄-贷款的惨败事件。表12.1列出了66家公司的一些运营的财务比率，其中33家在2年后破产，另33家在同期保持偿付能力。数据能在本书的网页上找到^②。用变量 X_1, X_2, X_3 拟合一个多元logistic回归模型，结果见表12.2。

^① 参阅第6章变量的变换。

^② <http://www.ilr.cornell.edu/~hadi/RABE>

表 12.1 有偿付能力及破产公司的财务比率

行	Y	X ₁	X ₂	X ₃	行	Y	X ₁	X ₂	X ₃
1	0	-62.8	-89.5	1.7	34	1	43.0	16.4	1.3
2	0	3.3	-3.5	1.1	35	1	47.0	16.0	1.9
3	0	-120.8	-103.2	2.5	36	1	-3.3	4.0	2.7
4	0	-18.1	-28.8	1.1	37	1	35.0	20.8	1.9
5	0	-3.8	-50.6	0.9	38	1	46.7	12.6	0.9
6	0	-61.2	-56.2	1.7	39	1	20.8	12.5	2.4
7	0	-20.3	-17.4	1.0	40	1	33.0	23.6	1.5
8	0	-194.5	-25.8	0.5	41	1	26.1	10.4	2.1
9	0	20.8	-4.3	1.0	42	1	68.6	13.8	1.6
10	0	-106.1	-22.9	1.5	43	1	37.3	33.4	3.5
11	0	-39.4	-35.7	1.2	44	1	59.0	23.1	5.5
12	0	-164.1	-17.7	1.3	45	1	49.6	23.8	1.9
13	0	-308.9	-65.8	0.8	46	1	12.5	7.0	1.8
14	0	7.2	-22.6	2.0	47	1	37.3	34.1	1.5
15	0	-118.3	-34.2	1.5	48	1	35.3	4.2	0.9
16	0	-185.9	-280.0	6.7	49	1	49.5	25.1	2.6
17	0	-34.6	-19.4	3.4	50	1	18.1	13.5	4.0
18	0	-27.9	6.3	1.3	51	1	31.4	15.7	1.9
19	0	-48.2	6.8	1.6	52	1	21.5	-14.4	1.0
20	0	-49.2	-17.2	0.3	53	1	8.5	5.8	1.5
21	0	-19.2	-36.7	0.8	54	1	40.6	5.8	1.8
22	0	-18.1	-6.5	0.9	55	1	34.6	26.4	1.8
23	0	-98.0	-20.8	1.7	56	1	19.9	26.7	2.3
24	0	-129.0	-14.2	1.3	57	1	17.4	12.6	1.3
25	0	-4.0	-15.8	2.1	58	1	54.7	14.6	1.7
26	0	-8.7	-36.3	2.8	59	1	53.5	20.6	1.1
27	0	-59.2	-12.8	2.1	60	1	35.9	26.4	2.0
28	0	-13.1	-17.6	0.9	61	1	39.4	30.5	1.9
29	0	-38.0	1.6	1.2	62	1	53.1	7.1	1.9
30	0	-57.9	0.7	0.8	63	1	39.8	13.8	1.2
31	0	-8.8	-9.1	0.9	64	1	59.5	7.0	2.0
32	0	-64.7	-4.0	0.1	65	1	16.3	20.4	1.0
33	0	-11.4	4.8	0.9	66	1	21.7	-7.8	1.6

这三个财务比率的含义为：

$$\begin{aligned} X_1 &= \frac{\text{未分配利润}}{\text{总资产}}, \\ X_2 &= \frac{\text{支付利息税金前的利润}}{\text{总资产}}, \\ X_3 &= \frac{\text{销售额}}{\text{总资产}}. \end{aligned}$$

响应变量定义为

$$Y = \begin{cases} 0, & \text{若 2 年后破产,} \\ 1, & \text{若 2 年后仍有偿付能力.} \end{cases}$$

表 12.2 用 X_1, X_2, X_3 作 Logistic 回归的输出结果

变量	系数	标准误	Z-检验	p-值	优势比	95% 置信区间	
						下限	上限
常数	-10.15	10.84	-0.94	0.349			
X_1	0.33	0.30	1.10	0.27	1.39	0.77	2.51
X_2	0.18	0.11	1.69	0.09	1.20	0.97	1.48
X_3	5.09	5.08	1.00	0.32	161.98	0.01	3.43×10^6
对数似然 = -2.906		$G = 85.683$		$d.f. = 3$		$p\text{-值} < 0.000$	

表 12.2 与标准回归的输出有些类似，有些输出的作用也相似。现在我们来解释拟合 logistic 回归得到的输出结果。记 π 为一个公司 2 年后仍有偿付能力的概率，其 logit 的拟合值为

$$\hat{g}(x_1, \dots, x_p) = -10.15 + 0.33x_1 + 0.18x_2 + 5.09x_3. \quad (12.6)$$

这相当于标准分析中的回归方程。这里，我们得到的是一个预测 logit 值 $\log(\pi/(1-\pi))$ 的模型，而不是预测 Y 的模型。对 logit 值作变换，便可以得到概率的预测值。常数项及各项系数可直接从表中的第二栏读出。各系数的标准误 (s.e.) 列于第三栏。题头标着“Z-检验”的第四栏为系数与其标准误的比值，比如相应于 X_2 之系数的 Z 值就是由 0.181 除以 0.107 得到的。这相当于标准回归中的 t -检验值。logistic 回归中该比值近似服从正态分布，而不是线性回归中的 t -分布。第五栏给出了相应于 Z 值的 p -值，解释与第 2、3 章中的 p -值类似，用来判断系数的显著性。若 p -值小于 0.05，那我们可以下结论：在 5% 的显著性水平下，相应的系数显著地不为 0。从表 12.2 中的 p -值来看，单个变量对于预测各观测的 logit 值来说都不显著。

在标准回归的输出中，回归系数有很简单的解释，第 j 个预测变量 X_j 的系数就是，其他变量固定不动、 X_j 改变一个单位所导致的 Y 之期望的变化量。而

(12.6) 中 X_2 的系数则是, 其他变量固定不动、 X_2 变化一个单位导致的 logit 值期望的变化量。logistic 回归的系数还有另一个解释, 很实用。在固定 X_1 、 X_3 时, X_2 增加一个单位, 则相对优势

$$\frac{\text{Pr(公司 2 年后仍有偿付能力)}}{\text{Pr(公司破产)}}$$

是原来的 $e^{\hat{\beta}_2} = e^{0.181} = 1.198$ 倍, 即增大了约 20%。这些值列于题头为“优势比”的第六列中, 表示在其他变量恒定、某一个变量变化一个单位所导致的优势比的变化量。譬如, 变量 X_j 改变一个单位、其他变量不变, 则优势比改变 $e^{\hat{\beta}_j}$ 倍。若 x_j 是二值变量, 取 1 或 0, 则 $e^{\hat{\beta}_j}$ 是 x_j 两种状态下的实际优势比, 而不是优势比的变化量。

表中最后两列给出了优势比的 95% 置信区间。如果置信区间不包含 1, 那么对应的变量对优势比有显著效应。若区间位于 1 之下, 则该变量增大会显著降低相对优势。另一方面, 若区间位于 1 之上, 则该变量增大会显著地提高相对优势。

为考察变量合起来是否对解释 logit 值有作用, 须检验系数 β_1, \dots, β_p 是否同时为 0。这就相当于在多元回归情形下检验是否所有的回归系数同时为 0。表 12.2 底部的 G 统计量就起这个作用。统计量 G 近似地服从 χ^2 分布。 p -值远小于 0.05, 表明这些变量合起来对 logit 值有显著影响。

12.5 Logistic 回归诊断

拟合 logistic 回归之后, 也可以通过一些诊断量度来检测异常点、高杠杆点、强影响观测, 以及其他模型缺陷。第 4 章中针对标准线性模型建立起来的一些诊断量度也适用于 logistic 回归模型。含 logistic 回归功能的回归软件包通常提供多种诊断量度, 包括:

1. 概率的估计值 $\hat{\pi}_i, i = 1, \dots, n$ 。
2. 一种或多种残差, 比如, 标准化偏离残差 DR_i , 和标准化皮尔逊残差 $PR_i, i = 1, \dots, n$ 。
3. 加权杠杆值 p_{ii}^* , 度量预测变量的观测值对于所得 logistic 回归结果的潜在效应。
4. 删去第 i 个观测之后回归系数的标准化差异 $DBETA_i, i = 1, \dots, n$ 。
5. 删去第 i 个观测之后 χ^2 统计量 G 的变化 $DFG_i, i = 1, \dots, n$ 。

这些量度的公式及其推导超出了本书讨论的范围。感兴趣的读者请参阅 Pregibon (1981), Landwehr, Pregibon and Shoemaker (1984), Hosmer and Lemeshow (1989) 等文献以及其中提到的文献。但上面这些量度与线性拟合得到的对应量度(第 4 章)用法相同。比如, 我们同样可以考察下列图形:

1. DR_i 关于 $\hat{\pi}_i$ 的散点图。
2. PR_i 关于 $\hat{\pi}_i$ 的散点图。
3. $DR_i, DBETA_i, DFG_i$ 及 p_{ii}^* 的序列图。

以破产数据为例, 由拟合 logistic 回归 (12.6) 得到的 DR_i , $DBETA_i$, DG_i 之序列图分别显示于图 12.2, 图 12.3 以及图 12.4。从这几张图容易看出, 第 9, 52, 36 号观测不太正常, 它们对 logistic 回归的结果可能有过度的影响。那么, 若把它们删去, 是否会显著地改变分析结果、结论呢? 我们把这个作为问题作为一个练习留给读者考察。

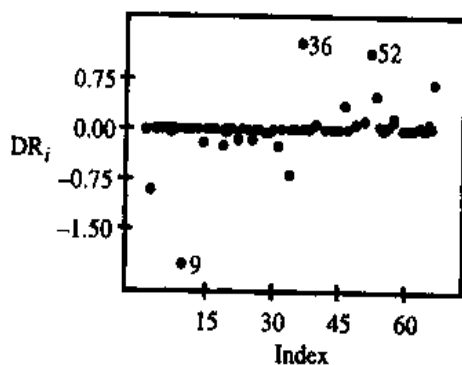


图 12.2 破产数据: 标准化偏离残差 DR_i 序列图

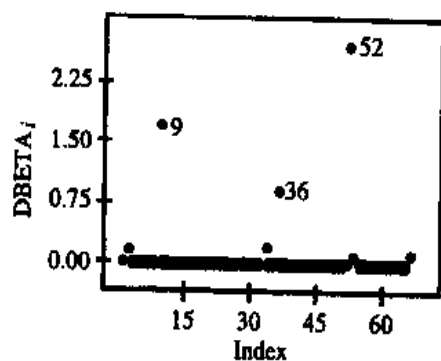


图 12.3 破产数据: $DBETA_i$ 序列图, 删去第 i 个观测之后回归系数的标准化差异

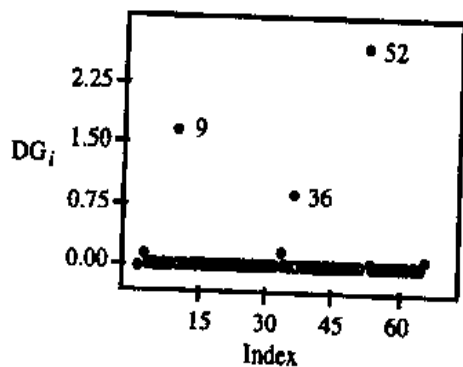


图 12.4 破产数据: DG_i 序列图, 删去第 i 个观测之后 χ^2 统计量 G 的变化 DFG_i

12.6 变量选择

对于破产数据的分析,我们现已确定,变量 X_1 、 X_2 、 X_3 合起来具有解释能力。我们是否需要所有这三个变量?这类似于第11章讨论的多元回归中的变量选择问题。这里,我们看的不再是残差平方和的减少量,而是看拟合的两个模型之间似然函数值的变化(精确地说,对数似然)。原因在于拟合 logistic 回归的准则是似然函数值,而最小二乘法的准则是平方和。记 $L(p)$ 为一个含 p 个变量和一个常数项的模型的对数似然值。类似地, $L(p+q)$ 为一个含 $p+q$ 个变量和一个常数项的模型的对数似然值。我们根据 $2(L(p+q) - L(p))$ 来判断添加 q 个变量的作用是否显著。这个量是两个模型的对数似然值之差的两倍,在添加的 q 个变量不显著时,它近似地服从自由度为 q 的 χ^2 分布(参阅附表 A.3)。

这个量的大小用来确定检验的显著性。若这个量较小,那相应的结论是,添加的 q 个变量对预测 logit 值无显著的改善,因而在模型中是不必要的。若这个量较大,那就要求在模型中保留这 q 个变量。临界值由检验的显著性水平决定。当用来拟合模型的观测数量 n 较大时,该检验方法是合理的。

我们来看,在破产数据的分析中,去掉变量 X_3 是否对模型影响不大。我们希望回答的问题是:变量 X_3 是否该保留在模型中?我们用 X_1 、 X_2 拟合了一个 logistic 回归,结果列于表 12.3。含 X_1 、 X_2 和 X_3 三个变量的模型的对数似然值为 -2.906 ,而仅含 X_1 、 X_2 两个变量的模型的对数似然值为 -4.736 。这里 $p=2$, $q=1$, $2(L(3) - L(2)) = 3.66$ 。应拿它与自由度为 1 的 χ^2 分布相比。从附表 A.3 查得 1 个自由度的 χ^2 分布的 5% 临界值为 3.84。因此我们可以在 5% 的显著性水平下下结论:去掉变量 X_3 ,将不影响模型的效果。

表 12.3 用 X_1, X_2 作 Logistic 回归的输出结果

变量	系数	标准误	Z-检验	p-值	优势比	95% 置信区间	
						下限	上限
常数	-0.550	0.951	-0.58	0.563			
X_1	0.157	0.075	2.10	0.036	1.17	1.01	1.36
X_2	0.195	0.122	1.59	0.112	1.21	0.96	1.54
对数似然 = -4.736		$G = 82.024$		$d.f. = 2$		$p\text{-值} < 0.000$	

再来看,我们是否能再去掉 X_2 。 Y 关于 X_1 的回归结果列于表 12.4,对数似然值为 -7.902 。我们前面所述的检验统计量的取值为 6.332,自由度为 1。和前面一样,5% 的临界值为 3.84。分析显示,我们不应该把 X_2 从模型中去掉。该检验的 p -值为 0.019。因此,若用我们的数据去预测公司的破产概率,模型应当包括 X_1 和 X_2 。

上述方法可以用于检验任何嵌套模型。一组模型若能看作某个较大模型的特例,那么我们称之为是嵌套的。这里的方法类似于多元回归中对嵌套模型的分析方法,差别仅在于这里的检验统计量是根据对数似然构造的,而不是平方和。

表 12.4 用 X_1 作 Logistic 回归的输出结果

变量	系数	标准误	Z-检验	p-值	优势比	95% 置信区间	
						下限	上限
常数	-1.167	0.816	-1.43	0.153			
X_1	0.177	0.057	3.09	0.002	1.19	1.07	1.33
对数似然 = -7.902		$G = 75.692$		$d.f. = 1$		$p\text{-值} < 0.000$	

12.7 Logistic 回归拟合程度的判断

一个多元回归模型整体的拟合程度可由 R^2 等量度来判断。但对 logistic 回归, 不存在这样一个简单而令人满意的量度。人们也提出了一些特定的、基于似然比的量度, 其中绝大多数量度是在二项分布假定下模型似然与数据似然之比的函数。这些量度提供的信息不是特别充分, 我们将考虑一种不同的方法。

logistic 回归方程意在对 Y 取 0 或 1 两个值的概率建模。我们用模型对观测分类, 然后根据分类正确的观测数去判断模型的表现。具体做法是: 对数据拟合 logistic 模型, 计算 logit 的拟合值, 并由此算得每个观测的拟合概率。如果一个观测的拟合概率大于 0.5, 那么将它判归 1 组 ($Y = 1$); 若小于 0.5, 则判归 0 组 ($Y = 0$)。然后看分类的正确率。正确率越高, 表明 logistic 模型的表现越好; 反之, 则越差。

文献中还提出了不少不同于 0.5 的分割点。在大多数应用场合中, 如果没有诸如误分类的相对损失或总体中两类个体的相对频数等辅助信息, 一般建议取 0.5 为分割点。

一个稍有疑问的问题是, 分类正确率该多高才能认为 logistic 回归是有效的。设在一个容量为 n 的样本中, 有 n_1 个观测来自 1 组, n_2 个观测来自 0 组。假如我们把所有观测判归其中的某一组, 那么分类的正确率为 n_1/n 或 n_2/n 。我们可以拿 $\max(n_1/n, n_2/n)$ 作为分类正确率的基准水平。如果 logistic 模型是有用的, 那么由 logistic 回归得到的分类正确率应显著高于该基准水平。

对于我们正在分析的破产数据, logistic 回归表现很好。我们发现, 包含 X_1, X_2 两个变量的模型, 将一个仍有偿付能力的观测 (第 36 号观测) 和一个破产的观测 (第 9 号观测) 分错了类。总的分类正确率为 $64/66 = 0.97$, 这比基准水平 0.5 大多了。

看待分类正确率还应谨慎。实践中, 若用这个 logistic 回归对来自该总体的一组新的观测分类, 其表现很可能不会同样地好。分类正确率一般偏大, 原因在于用来判断模型表现的与用来拟合模型的是同一批数据。对给定的一批数据拟合的模型, 在同一批数据上的表现理应是好的。正确衡量 logistic 回归在分类方面表现的量度, 应当是对一个新观测而不是一个样本观测正确分类的概率。对分类正确率估计的偏性可用诸如刀切法、自助法等重抽样方法来减小。这里将不再讨论了。读者可参阅 Efron (1982) 及 Diaconis and Efron (1983)。

12.8 分类问题：另一种方法

我们前面讨论了, 根据给定的若干特征的测量值, 用 logistic 回归对一个观测隶属于某一类别的概率建模的方法。接着, 我们还讨论了如何用 logit 的拟合值将一个观测判归于两类中的一类。如果我们的主要兴趣是分类, 那么还有另外一类常用的统计方法, 即判别分析。这里我们不讨论判别分析, 仅仅指出它可用一种简单的回归方法实现。读者可从 Mclachlan (1992), Rencher (1995) 和 Johnson (1998) 中找到关于判别分析的讨论。

判别分析的基本想法是, 要找到预测变量 X_1, \dots, X_p 的一种线性组合, 使得该线性组合的得分值尽可能地将观测分成两类。实现这样的分割的一条途径是, 对数据拟合一个多元回归模型。响应变量是 Y , 取 0 或 1 两个值, 预测变量为 X_1, \dots, X_p 。前面已经指出过, 有些拟合值将会落在 0, 1 范围之外。但这不要紧, 因为我们不是试图对概率建模, 而仅仅是预测类别。我们计算所有观测的预测值之平均。若一给定观测的预测值大于那个平均预测值, 我们将此观测判归 $Y = 1$ 那类; 若预测值小于平均预测值, 那我们将之判归 $Y = 0$ 那类。对这种分类, 我们也计算样本中分类正确的观测数。也可用多元回归中变量选择的方法来确定哪些变量该用于分类。

我们拿先前曾用于说明 logistic 回归的破产数据来演示这个方法。表 12.5 给出了用三个预测变量 X_1, X_2, X_3 得到的普通最小二乘回归的结果。这三个变量都具有显著的回归系数, 都应留在分类方程中。

表 12.5 Y 关于 X_1, X_2, X_3 的 OLS 回归结果

变量	系数	标准误	t -检验	p -值
常数	0.322	0.087	3.68	0.0005
X_1	0.003	0.001	3.76	0.0004
X_2	0.004	0.001	2.96	0.0044
X_3	0.149	0.045	3.28	0.0017
$n = 66$	$R^2 = 0.57$	$R_a^2 = 0.55$	$\hat{\sigma} = 0.3383$	$d.f. = 62$

表 12.6 列出了破产数据中 Y 的观测值、预测值, 以及判归的类别。 Y 的预测值之平均为 0.5。所有预测值小于 0.5 的观测判归 $Y = 0$, 预测值大于 0.5 的判归 $Y = 1$ 。误判的观测标以 *。从结果可见, 有 5 个破产公司被误判为有偿付能力的, 1 个有偿付能力的公司被误判为破产的。应当注意到, logistic 回归仅误判了两个观测: 一个有偿付能力的公司和一个破产公司被误判。就表 12.2 中的破产数据而言, 在对这些样本数据分类时, logistic 回归比多元回归有更好的表现。这一般是对的。logistic 回归不必对预测变量作多元正态性这样的限制性假定。对于分类问题, 我们建议使用 logistic 回归。如果没有 logistic 回归软件, 那么可以试试多元回归方法。

表 12.6 由拟合值对观测作的分类

行	Y	拟合值	分类结果	行	Y	拟合值	分类结果
1	0	-0.00	0	34	1	0.72	1
2	0	0.48	0	35	1	0.82	1
3	0	-0.12	0	36	1	0.73	1
4	0	0.31	0	37	1	0.80	1
5	0	0.23	0	38	1	0.65	1
6	0	0.14	0	39	1	0.80	1
7	0	0.33	0	40	1	0.75	1
8	0	-0.32	0	41	1	0.76	1
9	0	0.52	1*	42	1	0.83	1
10	0	0.12	0	43	1	1.10	1
11	0	0.23	0	44	1	1.42	1
12	0	-0.07	0	45	1	0.86	1
13	0	-0.80	0	46	1	0.66	1
14	0	0.55	1*	47	1	0.81	1
15	0	0.03	0	48	1	0.58	1
16	0	-0.45	0	49	1	0.97	1
17	0	0.64	1*	50	1	1.03	1
18	0	0.45	0	51	1	0.77	1
19	0	0.44	0	52	1	0.48	0*
20	0	0.14	0	53	1	0.60	1
21	0	0.22	0	54	1	0.74	1
22	0	0.37	0	55	1	0.81	1
23	0	0.18	0	56	1	0.84	1
24	0	0.05	0	57	1	0.62	1
25	0	0.55	1*	58	1	0.81	1
26	0	0.56	1*	59	1	0.74	1
27	0	0.39	0	60	1	0.84	1
28	0	0.34	0	61	1	0.86	1
29	0	0.39	0	62	1	0.80	1
30	0	0.26	0	63	1	0.68	1
31	0	0.39	0	64	1	0.83	1
32	0	0.12	0	65	1	0.61	1
33	0	0.44	0	66	1	0.59	1

习 题

- 12.1 诊断图 12.2, 12.3, 12.4 显示出破产数据中的三个不正常的观测。去掉这三个观测, 用余下的 63 个观测拟合 logistic 回归, 并将你的结果与 12.5 节的结果作比较。去掉这三个点的是否导致了 logistic 回归结果实质性的改变?
- 12.2 对 Y 关于 X_1, X_2 的 logistic 回归 (表 12.3) 作各种 logistic 回归诊断, 判断数据中是否含不正常观测。
- 12.3 用于空间发射的助推火箭中的 O 型环对火箭防爆起着重要作用。通常认为 O 型环的失效概率与温度有关。有关问题背景的详细讨论可参见 Chatterjee, Handcock, and Simonoff (1995) 中 The Flight of the Space Shuttle Challenger (pp. 33-35) 一文。每次飞行中有六个 O 型环可损坏。来自 23 次飞行的数据列于表 12.7, 也可从本书的网页^①上找到。对每次飞行, 我们记录有 O 型环的损坏数以及发射温度。
- (a) 拟合一个关于 O 型环的失效概率与温度之间关系的 logistic 回归。解释系数。
- (b) 第 18 次飞行的发射温度 75 度被认为是有疑问的。去掉这一点, 用剩下的数据拟合一个 logistic 回归。解释系数。
- (c) 根据拟合的模型, 估计发射温度为 31 度时 O 型环的失效概率。这是 1986 年 1 月 20 日发射不幸的挑战者号飞行那天的温度预报。
- (d) 你对那天的发射有何建议?

表 12.7 挑战者号航天飞机 23 次飞行中 O 型环的损坏数以及发射时的温度 (华氏度)

航次	损坏数	温度	航次	损坏数	温度
1	2	53	13	1	70
2	1	57	14	1	70
3	1	58	15	0	72
4	1	63	16	0	73
5	0	66	17	0	75
6	0	67	18	2	75
7	0	67	19	0	76
8	0	67	20	0	78
9	0	68	21	0	79
10	0	69	22	0	81
11	0	70	23	0	76
12	0	70			

- 12.4 泛美橄榄球联合会 (AFL) 和国家橄榄球联合会 (NFL) 1969 年赛季的定位射门数据列于表 12.8, 也可从本书网页上找到。记 $\pi(X)$ 为从 X 码处踢进球门

① <http://www.ilr.cornell.edu/~hadi/RABF>

的概率。

(a) 对这两个联合会，分别拟合模型

$$\pi(X) = \frac{e^{\beta_0 + \beta_1 X + \beta_2 X^2}}{1 + e^{\beta_0 + \beta_1 X + \beta_2 X^2}}.$$

(b) Z 表示哪个协会，即

$$Z = \begin{cases} 1, & \text{表示 AFL,} \\ 0, & \text{表示 NFL.} \end{cases}$$

用两个协会的数据拟合单个模型，其中包括示性变量 Z 。即拟合

$$\pi(X, Z) = \frac{e^{\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z}}{1 + e^{\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 Z}}.$$

(c) 模型中二次项的作用显著吗？

(d) 在同样的地方踢，两队的得分概率是否相同？

表 12.8 泛美橄榄球联合会 (AFL) 和国家橄榄球联合会 (NFL)
1969 年赛季的定位射门表现。示性变量 Z 表示球队。

球队	距离	成功次数	尝试次数	Z
NFL	14.5	68	77	0
NFL	24.5	74	95	0
NFL	34.5	61	113	0
NFL	44.5	38	138	0
NFL	52.0	2	38	0
AFL	14.5	62	67	1
AFL	24.5	49	70	1
AFL	34.5	43	79	1
AFL	44.5	25	82	1
AFL	52.0	7	24	1

附录：统计用表

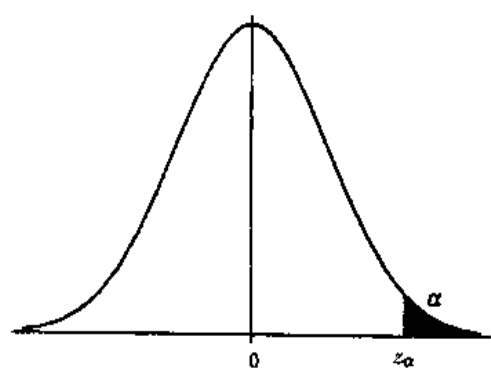
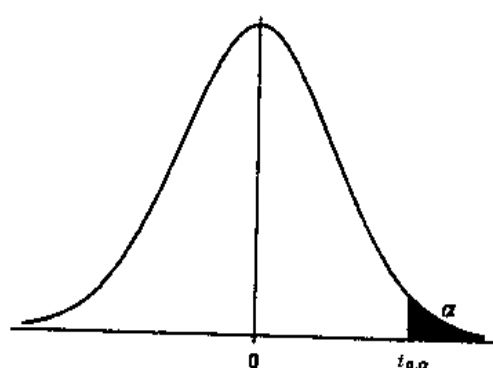


图 A.1 标准正态分布的概率密度函数

表 A.1 标准正态分布 Z 的临界值 z_α , 其中 $Pr(Z \geq z_\alpha) = \alpha$

α	z_α	α	z_α	α	z_α	α	z_α	α	z_α
.50	0.00	.050	1.64	.030	1.88	.020	2.05	.010	2.33
.45	0.13	.048	1.66	.029	1.90	.019	2.07	.009	2.37
.40	0.25	.046	1.68	.028	1.91	.018	2.10	.008	2.41
.35	0.39	.044	1.71	.027	1.93	.017	2.12	.007	2.46
.30	0.52	.042	1.73	.026	1.94	.016	2.14	.006	2.51
.25	0.67	.040	1.75	.025	1.96	.015	2.17	.005	2.58
.20	0.84	.038	1.77	.024	1.98	.014	2.20	.004	2.65
.15	1.04	.036	1.80	.023	2.00	.013	2.23	.003	2.75
.10	1.28	.034	1.83	.022	2.01	.012	2.26	.002	2.88
.05	1.64	.032	1.85	.021	2.03	.011	2.29	.001	3.09

来源：摘自 Lindley and Miller (1958), *Cambridge Elementary Statistical Tables*, published by Cambridge University Press. Table 2. 承蒙作者和发行者慷慨应允转载。

图 A.2 自由度 (*d.f.*) 为 *n* 的学生氏 *t*-分布的概率密度函数表 A.2 自由度 (*d.f.*) 为 *n* 的学生氏 *t*-分布 T_n 的临界值 $t_{n,\alpha}$, 其中 $Pr(T_n \geq t_{n,\alpha}) = \alpha$

<i>n</i> (<i>d.f.</i>)	α				
	0.10	0.05	0.025	0.010	0.005
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.97	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.42	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
12	1.36	1.78	2.18	2.68	3.06
14	1.34	1.76	2.14	2.62	2.98
16	1.34	1.75	2.12	2.58	2.92
18	1.33	1.73	2.10	2.55	2.88
20	1.32	1.72	2.09	2.53	2.84
30	1.31	1.70	2.04	2.46	2.75
40	1.30	1.68	2.02	2.42	2.70
60	1.30	1.67	2.00	2.39	2.66
120	1.29	1.66	1.98	2.36	2.62
∞	1.28	1.64	1.96	2.33	2.58

来源：摘自 Fisher and Yates (1963), *Statistical Tables for Biological, Agricultural and Medical Research*, 6th Ed., published by Oliver and Boyd, Edinburgh. Table III.
承蒙作者和发行者慷慨应允转载。

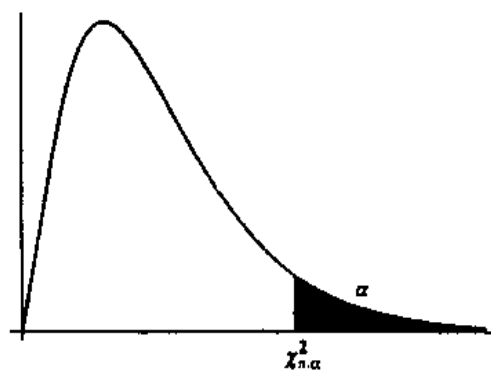


图 A.3 自由度 (d.f.) 为 n 的 χ^2 分布的概率密度函数

表 A.3 自由度 ($d.f.$) 为 n 的 χ^2 分布 χ_n^2 的临界值 $\chi_{n,\alpha}^2$, 其中 $Pr(\chi_n^2 \geq \chi_{n,\alpha}^2) = \alpha$

n ($d.f.$)	α				
	0.10	0.05	0.025	0.010	0.005
1	2.71	3.84	5.02	6.63	7.88
2	4.61	5.99	7.38	9.21	10.60
3	6.25	7.81	9.35	11.34	12.84
4	7.78	9.49	11.14	13.28	14.86
5	9.24	11.07	12.83	15.09	16.75
6	10.65	12.59	14.45	16.81	18.55
7	12.02	14.07	16.01	18.48	20.28
8	13.36	15.51	17.53	20.09	21.96
9	14.68	16.92	19.02	21.67	23.59
10	15.99	18.31	20.48	23.21	25.19
11	17.28	19.68	21.92	24.72	26.76
12	18.55	21.03	23.34	26.22	28.30
13	19.81	22.36	24.74	27.69	29.82
14	21.06	23.68	26.12	29.14	31.32
15	22.31	25.00	27.49	30.58	32.80
16	23.54	26.30	28.85	32.00	34.27
17	24.77	27.59	30.19	33.41	35.72
18	25.99	28.87	31.53	34.81	37.16
19	27.20	30.14	32.85	36.19	38.58
20	28.41	31.41	34.17	37.57	40.00
21	29.62	32.67	35.48	38.93	41.40
22	30.81	33.92	36.78	40.29	42.80
23	32.01	35.17	38.08	41.64	44.18
24	33.20	36.42	39.36	42.98	45.56
25	34.38	37.65	40.65	44.31	46.93
26	35.56	38.89	41.92	45.64	48.29
27	36.74	40.11	43.19	46.96	49.65
28	37.92	41.34	44.46	48.28	50.99
29	39.09	42.56	45.72	49.59	52.34
30	40.26	43.77	46.98	50.89	53.67
40	51.81	55.76	59.34	63.69	66.77
50	63.17	67.50	71.42	76.15	79.49
60	74.40	79.08	83.30	88.38	91.95
70	85.53	90.53	95.02	100.42	104.22
80	96.58	101.88	106.63	112.33	116.32
90	107.57	113.14	118.14	124.12	128.30
100	118.50	124.34	129.56	135.81	140.17

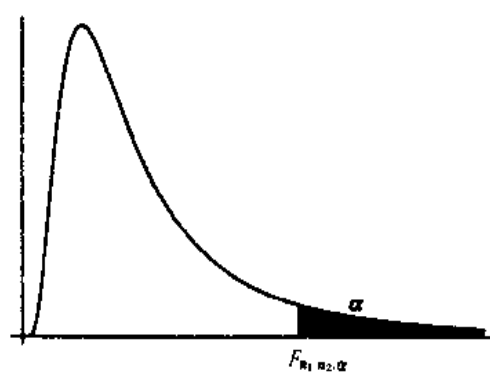


图 A.4 自由度 (*d.f.*) 为 n_1 (分子)、 n_2 (分母) 的 *F*-分布的概率密度函数

表 A.4 自由度 ($d.f.$) 为 n_1 (分子)、 n_2 (分母) 的 F -分布 F_{n_1, n_2} 的 5% 临界值
 $f_{n_1, n_2; 0.05}$, 其中 $\Pr(F_{n_1, n_2} \geq f_{n_1, n_2; 0.05}) = 0.05$

n_2	n_1								
	1	2	4	6	8	10	12	24	∞
1	161.4	199.5	224.6	234.0	238.9	241.9	243.9	249.1	254.3
2	18.51	19.00	19.25	19.33	19.37	19.40	19.41	19.45	19.50
3	10.13	9.55	9.12	8.94	8.85	8.79	8.74	8.64	8.53
4	7.71	6.94	6.39	6.16	6.04	5.96	5.91	5.77	5.63
5	6.61	5.79	5.19	4.95	4.82	4.74	4.68	4.53	4.36
6	5.99	5.14	4.53	4.28	4.15	4.06	4.00	3.84	3.67
7	5.59	4.74	4.12	3.87	3.73	3.64	3.57	3.41	3.23
8	5.32	4.46	3.84	3.58	3.44	3.35	3.28	3.12	2.93
9	5.12	4.26	3.63	3.37	3.23	3.14	3.07	2.90	2.71
10	4.96	4.10	3.48	3.22	3.07	2.98	2.91	2.74	2.54
11	4.84	3.98	3.36	3.09	2.95	2.85	2.79	2.61	2.40
12	4.75	3.89	3.26	3.00	2.85	2.75	2.69	2.51	2.30
13	4.67	3.81	3.18	2.92	2.77	2.67	2.60	2.42	2.21
14	4.60	3.74	3.11	2.85	2.70	2.60	2.53	2.35	2.13
15	4.54	3.68	3.06	2.79	2.64	2.54	2.48	2.29	2.07
20	4.35	3.49	2.87	2.60	2.45	2.35	2.28	2.08	1.84
25	4.24	3.39	2.76	2.49	2.34	2.24	2.16	1.96	1.71
30	4.17	3.32	2.69	2.42	2.27	2.16	2.09	1.89	1.62
40	4.08	3.23	2.61	2.34	2.18	2.08	2.00	1.79	1.51
60	4.00	3.15	2.53	2.25	2.10	1.99	1.92	1.70	1.39
120	3.92	3.07	2.45	2.17	2.02	1.91	1.83	1.61	1.25
∞	3.84	3.00	2.37	2.10	1.94	1.83	1.75	1.52	1.00

来源：节选自 Pearson and Hartley (1954), *Biometrika Tables for Statisticians, Volume I*, published at the Cambridge University Press for the *Biometrika* Trustees. Table 18. 承蒙作者和发行者慷慨应允转载。

表 A.5 自由度 ($d.f.$) 为 n_1 (分子)、 n_2 (分母) 的 F -分布 F_{n_1, n_2} 的 1% 临界值
 $f_{n_1, n_2; 0.01}$, 其中 $Pr(F_{n_1, n_2} \geq f_{n_1, n_2; 0.01}) = 0.01$

n_2	n_1								
	1	2	4	6	8	10	12	24	∞
1	4052	5000	5625	5859	5981	6056	6106	6235	6366
2	98.50	99.00	99.25	99.33	99.37	99.40	99.42	99.46	99.50
3	34.12	30.82	28.71	27.91	27.49	27.23	27.05	26.60	26.13
4	21.20	18.00	15.98	15.21	14.80	14.55	14.37	13.93	13.46
5	16.26	13.27	11.39	10.67	10.29	10.05	9.89	9.47	9.02
6	13.75	10.92	9.15	8.47	8.10	7.87	7.72	7.31	6.88
7	12.25	9.55	7.85	7.19	6.84	6.62	6.47	6.07	5.65
8	11.26	8.65	7.01	6.37	6.03	5.81	5.67	5.28	4.86
9	10.56	8.02	6.42	5.80	5.47	5.26	5.11	4.73	4.31
10	10.04	7.56	5.99	5.39	5.06	4.85	4.71	4.33	3.91
11	9.65	7.21	5.67	5.07	4.74	4.54	4.40	4.02	3.60
12	9.33	6.93	5.41	4.82	4.50	4.30	4.16	3.78	3.36
13	9.07	6.70	5.21	4.62	4.30	4.10	3.96	3.59	3.17
14	8.86	6.51	5.04	4.46	4.14	3.94	3.80	3.43	3.00
15	8.68	6.36	4.89	4.32	4.00	3.80	3.67	3.29	2.87
20	8.10	5.85	4.43	3.87	3.56	3.37	3.23	2.86	2.42
25	7.77	5.57	4.18	3.63	3.32	3.13	2.99	2.62	2.17
30	7.56	5.39	4.02	3.47	3.17	2.98	2.84	2.47	2.01
40	7.31	5.18	3.83	3.29	2.99	2.80	2.66	2.29	1.80
60	7.08	4.98	3.65	3.12	2.82	2.63	2.50	2.12	1.60
120	6.85	4.79	3.48	2.96	2.66	2.47	2.34	1.95	1.38
∞	6.63	4.61	3.32	2.80	2.51	2.32	2.18	1.79	1.00

来源: 节选自 Pearson and Hartley (1954), *Biometrika Tables for Statisticians, Volume 1*, published at the Cambridge University Press for the *Biometrika* Trustees. Table 18.
 承蒙作者和发行者慷慨应允转载。

表 A.6 Durbin-Watson 统计量 d 的分布: d_L 与 d_U 的 5% 显著点 (p 为预测变量个数)

n	$p = 1$		$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	1.08	1.36	0.95	1.54	0.82	1.75	0.69	1.97	0.56	2.21
16	1.10	1.37	0.98	1.54	0.86	1.73	0.74	1.93	0.62	2.15
17	1.13	1.38	1.02	1.54	0.90	1.71	0.78	1.90	0.67	2.10
18	1.16	1.39	1.05	1.53	0.93	1.69	0.82	1.87	0.71	2.06
19	1.18	1.40	1.08	1.53	0.97	1.68	0.86	1.85	0.75	2.02
20	1.20	1.41	1.10	1.54	1.00	1.68	0.90	1.83	0.79	1.99
21	1.22	1.42	1.13	1.54	1.03	1.67	0.93	1.81	0.83	1.96
22	1.24	1.43	1.15	1.54	1.05	1.66	0.96	1.80	0.86	1.94
23	1.26	1.44	1.17	1.54	1.08	1.66	0.99	1.79	0.90	1.92
24	1.27	1.45	1.19	1.55	1.10	1.66	1.01	1.78	0.93	1.90
25	1.29	1.45	1.21	1.55	1.12	1.66	1.04	1.77	0.95	1.89
26	1.30	1.46	1.22	1.55	1.14	1.65	1.06	1.76	0.98	1.88
27	1.32	1.47	1.24	1.56	1.16	1.65	1.08	1.76	1.01	1.86
28	1.33	1.48	1.26	1.56	1.18	1.65	1.10	1.75	1.03	1.85
29	1.34	1.48	1.27	1.56	1.20	1.65	1.12	1.74	1.05	1.84
30	1.35	1.49	1.28	1.57	1.21	1.65	1.14	1.74	1.07	1.83
31	1.36	1.50	1.30	1.57	1.23	1.65	1.16	1.74	1.09	1.83
32	1.37	1.50	1.31	1.57	1.24	1.65	1.18	1.73	1.11	1.82
33	1.38	1.51	1.32	1.58	1.26	1.65	1.19	1.73	1.13	1.81
34	1.39	1.51	1.33	1.58	1.27	1.65	1.21	1.73	1.15	1.81
35	1.40	1.52	1.34	1.58	1.28	1.65	1.22	1.73	1.16	1.80
36	1.41	1.52	1.35	1.59	1.29	1.65	1.24	1.73	1.18	1.80
37	1.42	1.53	1.36	1.59	1.31	1.66	1.25	1.72	1.19	1.80
38	1.43	1.54	1.37	1.59	1.32	1.66	1.26	1.72	1.21	1.79
39	1.43	1.54	1.38	1.60	1.33	1.66	1.27	1.72	1.22	1.79
40	1.44	1.54	1.39	1.60	1.34	1.66	1.29	1.72	1.23	1.79
45	1.48	1.57	1.43	1.62	1.38	1.67	1.34	1.72	1.29	1.78
50	1.50	1.59	1.46	1.63	1.42	1.67	1.38	1.72	1.34	1.77
55	1.53	1.60	1.49	1.64	1.45	1.68	1.41	1.72	1.38	1.77
60	1.55	1.62	1.51	1.65	1.48	1.69	1.44	1.73	1.41	1.77
65	1.57	1.63	1.54	1.66	1.50	1.70	1.47	1.73	1.44	1.77
70	1.58	1.64	1.55	1.67	1.52	1.70	1.49	1.74	1.46	1.77
75	1.60	1.65	1.57	1.68	1.54	1.71	1.51	1.74	1.49	1.77
80	1.61	1.66	1.59	1.69	1.56	1.72	1.53	1.74	1.51	1.77
85	1.62	1.67	1.60	1.70	1.57	1.72	1.55	1.75	1.52	1.77
90	1.63	1.68	1.61	1.70	1.59	1.73	1.57	1.75	1.54	1.78
95	1.64	1.69	1.62	1.71	1.60	1.73	1.58	1.75	1.56	1.78
100	1.65	1.69	1.63	1.72	1.61	1.74	1.59	1.76	1.57	1.78

来源: Durbin and Watson (1951)。

表 A.7 Durbin-Watson 统计量 d 的分布: d_L 与 d_U 的 1% 显著点 (p 为预测变量个数)

n	$p = 1$		$p = 2$		$p = 3$		$p = 4$		$p = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
15	0.81	1.07	0.70	1.25	0.59	1.46	0.49	1.70	0.39	1.96
16	0.84	1.09	0.74	1.25	0.63	1.44	0.53	1.66	0.40	1.90
17	0.87	1.10	0.77	1.25	0.67	1.43	0.57	1.63	0.48	1.85
18	0.90	1.12	0.80	1.26	0.71	1.42	0.61	1.60	0.52	1.80
19	0.93	1.13	0.83	1.26	0.74	1.41	0.65	1.58	0.56	1.77
20	0.95	1.15	0.86	1.27	0.77	1.41	0.68	1.57	0.60	1.74
21	0.97	1.16	0.89	1.27	0.80	1.41	0.72	1.55	0.63	1.71
22	1.00	1.17	0.91	1.28	0.83	1.40	0.75	1.54	0.66	1.69
23	1.02	1.19	0.94	1.29	0.86	1.40	0.77	1.53	0.70	1.67
24	1.04	1.20	0.96	1.30	0.88	1.41	0.80	1.53	0.72	1.66
25	1.05	1.21	0.98	1.30	0.90	1.41	0.83	1.52	0.75	1.65
26	1.07	1.22	1.00	1.31	0.93	1.41	0.85	1.52	0.78	1.64
27	1.09	1.23	1.02	1.32	0.95	1.41	0.88	1.51	0.81	1.63
28	1.10	1.24	1.04	1.32	0.97	1.41	0.90	1.51	0.83	1.62
29	1.12	1.25	1.05	1.33	0.99	1.42	0.92	1.51	0.85	1.61
30	1.13	1.26	1.07	1.34	1.01	1.42	0.94	1.51	0.88	1.61
31	1.15	1.27	1.08	1.34	1.02	1.42	0.96	1.51	0.90	1.60
32	1.16	1.28	1.10	1.35	1.04	1.43	0.98	1.51	0.92	1.60
33	1.17	1.29	1.11	1.36	1.05	1.43	1.00	1.51	0.94	1.59
34	1.18	1.30	1.13	1.36	1.07	1.43	1.01	1.51	0.95	1.59
35	1.19	1.31	1.14	1.37	1.08	1.44	1.03	1.51	0.97	1.59
36	1.21	1.32	1.15	1.38	1.10	1.44	1.04	1.51	0.99	1.59
37	1.22	1.32	1.16	1.38	1.11	1.45	1.06	1.51	1.00	1.59
38	1.23	1.33	1.18	1.39	1.12	1.45	1.07	1.52	1.02	1.58
39	1.24	1.34	1.19	1.39	1.14	1.45	1.09	1.52	1.03	1.58
40	1.25	1.34	1.20	1.40	1.15	1.46	1.10	1.52	1.05	1.58
45	1.29	1.38	1.24	1.42	1.20	1.48	1.16	1.53	1.11	1.58
50	1.32	1.40	1.28	1.45	1.24	1.49	1.20	1.54	1.16	1.59
55	1.36	1.43	1.32	1.47	1.28	1.51	1.25	1.55	1.21	1.59
60	1.38	1.45	1.35	1.48	1.32	1.52	1.28	1.56	1.25	1.60
65	1.41	1.47	1.38	1.50	1.35	1.53	1.31	1.57	1.28	1.61
70	1.43	1.49	1.40	1.52	1.37	1.55	1.34	1.58	1.31	1.61
75	1.45	1.50	1.42	1.53	1.39	1.56	1.37	1.59	1.34	1.62
80	1.47	1.52	1.44	1.54	1.42	1.57	1.39	1.60	1.36	1.62
85	1.48	1.53	1.46	1.55	1.43	1.58	1.41	1.60	1.39	1.63
90	1.50	1.54	1.47	1.56	1.45	1.59	1.43	1.61	1.41	1.64
95	1.51	1.55	1.49	1.57	1.47	1.60	1.45	1.62	1.42	1.64
100	1.52	1.56	1.50	1.58	1.48	1.60	1.46	1.63	1.44	1.65

来源: Durbin and Watson (1951)。

参考文献

- Anscombe, F. J. (1960), "Rejection of Outliers," *Technometrics*, 2, 123-167.
- Anscombe, F. J. (1973), "Graphs in Statistical Analysis," *The American Statistician*, 27, 17-21.
- Atkinson, A. C. (1985), *Plots, Transformations, and Regression: An Introduction to Graphical Methods of Diagnostic Regression Analysis*, Oxford: Clarendon Press.
- Barnett, V. and Lewis, T. (1994), *Outliers in Statistical Data*, 3rd ed., New York: John Wiley & Sons.
- Bartlett, G., Stewart, J., and Abrahamowicz, M. (1998), "Quantitative Sensory Testing of Peripheral Nerves," *Student: A Statistical Journal for Graduate Students*, 2, 289-301.
- Bates, D. M. and Watts, D. G. (1988), *Nonlinear Regression Analysis and Its Applications*, New York: John Wiley & Sons.
- Becker, R. A., Cleveland, W. S., and Wilks, A. R. (1987), "Dynamic Graphics for Data Analysis," *Statistical Science*, 2, 4, 355-395.
- Belsley, D. A. (1991), *Conditioning Diagnostics: Collinearity and Weak Data in Regression*, New York: John Wiley & Sons.
- Belsley, D. A., Kuh, E., and Welsch, R. E. (1980), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, New York: John Wiley & Sons.
- Billor, N., Chatterjee, S., and Hadi, A. S. (1999), "A Re-Weighted Least Squares Method for Robust Regression Estimation and Outlier Detection," *Technical Report #99-002*, Department of Social Statistics, Cornell University.

- Birkes, D. and Dodge, Y. (1993), *Alternative Methods of Regression*, New York: John Wiley & Sons.
- Box, G. E. P. and Pierce, D. A. (1970), "Distribution of Residual Autocorrelation in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, 64, 1509-1526.
- Carroll, R. J. and Ruppert, D. (1988), *Transformation and Weighting in Regression*, London: Chapman and Hall.
- Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*, Boston: Duxbury Press.
- Chatterjee, S. and Hadi, A. S. (1988), *Sensitivity Analysis in Linear Regression*, New York: John Wiley & Sons.
- Chatterjee, S., Handcock, M. S., and Simonoff, J. S. (1995), *A Casebook for a First Course in Statistics and Data Analysis*, New York: John Wiley & Sons.
- Chatterjee, S. and Mächler, M. (1997), "Robust Regression: A Weighted Least Squares Approach," *Communications in Statistics, Theory and Methods*, 26, 1381-1394.
- Chi-Lu, C. and Van Ness, J. W. (1999), *Statistical Regression With Measurement Error*, London: Arnold.
- Coakley, C. W. and Hettmansperger, T. P. (1993), "A Bounded Influence, High Breakdown, Efficient Regression Estimator," *Journal of the American Statistical Association*, 88, 872-880.
- Christensen, R. (1996), *Analysis of Variance, Design and Regression: Applied Statistical Methods*, New York: Chapman and Hall.
- Cochrane, D. and Orcutt, G. H. (1949), "Application of Least Squares Regression to Relationships Containing Autocorrelated Error Terms," *Journal of the American Statistical Association*, 44, 32-61.
- Coleman, J. S., Cambell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., and York, R. L. (1966), *Equality of Educational Opportunity*, U.S. Government Printing Office, Washington, D.C.
- Conover, W. J. (1980), *Practical Nonparametric Statistics*, New York: John Wiley & Sons.
- Cook, R. D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15-18.
- Cook, R. D. and Weisberg, S. (1982), *Residuals and Influence in Regression*, London: Chapman and Hall.
- Cox, D. R. (1989), *The Analysis of Binary Data*, 2nd ed., London: Methuen.
- Daniel, C. and Wood, F. S. (1980), *Fitting Equations to Data: Computer Analysis of Multifactor Data*, 2nd ed., New York: John Wiley & Sons.
- Dempster, A. P., Schatzoff, M., and Wermuth, N. (1977), "A Simulation Study of Alternatives to Ordinary Least Squares," *Journal of the American Statistical Association*, 72, 77-106.
- Diaconis, P. and Efron, B. (1983), "Computer Intensive Methods in Statistics," *Scientific American*, 248, 116-130.

- Dodge, Y. and Hadi, A. S. (1999), "Simple Graphs and Bounds for the Elements of the Hat Matrix" *Journal of Applied Statistics*, 26, 817-823.
- Draper, N. R. and Smith, H. (1998), *Applied Regression Analysis*, 3rd ed., New York: John Wiley & Sons.
- Durbin, J. and Watson, G. S. (1950), "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, 37, 409-428.
- Durbin, J. and Watson, G. S. (1951), "Testing for Serial Correlation in Least Squares Regression, II," *Biometrika*, 38, 159-178.
- Efron, B. (1982), "The Jackknife, the Bootstrap and Other Resampling Plans," *CBMS- National Science Monograph 38*, Society of Industrial and Applied Mathematics.
- Ezekiel, M. (1924), "A Method for Handling Curvilinear Correlation for Any Number of Variables," *Journal of the American Statistical Association*, 19, 431-453.
- Finney, D. J. (1964), *Probit Analysis*, London: Cambridge University Press.
- Fox, J. (1984), *Linear Statistical Models and Related Methods*, New York: John Wiley & Sons.
- Friedman, M. and Meiselman, D. (1963), "The Relative Stability of Monetary Velocity and the Investment Multiplier in the United States, 1897-1958," in *Commission on Money and Credit, Stabilization Policies*, Englewood Cliffs, N.J.: Prentice-Hall.
- Fuller, W. A. (1987), *Measurement Error Models*, New York: John Wiley & Sons.
- Furnival, G. M. and Wilson, R. W., Jr. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 499-512.
- Gibbons, J. D. (1993), *Nonparametric Statistics: An Introduction*, Newbury Park, CA: Sage Publications.
- Goldstein, M. and Smith, A. F. M. (1974), "Ridge-Type Estimates for Regression Analysis," *Journal of the Royal Statistical Society (B)*, 36, 284-291.
- Gray, J. B. (1986), "A Simple Graphic for Assessing Influence in Regression," *Journal of Statistical Computation and Simulation*, 24, 121-134.
- Gray, J. B. and Ling, R. F. (1984), "K-Clustering as a Detection Tool for Influential Subsets in Regression (with Discussion)," *Technometrics*, 26, 305-330.
- Graybill, F. A. (1976), *Theory and Application of the Linear Model*, Belmont, CA: Duxbury Press.
- Graybill, F. A. and Iyer, H. K. (1994), *Regression Analysis: Concepts and Applications*, Belmont, CA: Duxbury Press.
- Green, W. H. (1993), *Econometric Analysis*, 2nd ed., Saddle River, NJ: Prentice-Hall.
- Gunst, R. F. and Mason, R. L. (1980), *Regression Analysis and Its Application: A Data-Oriented Approach*, New York: Marcel Dekker.
- Hadi, A. S. (1988), "Diagnosing Collinearity-Influential Observations," *Computational Statistics and Data Analysis*, 7, 143-159.

- Hadi, A. S. (1993), "Graphical Methods for Linear Models," Chapter 23 in *Handbook of Statistics: Computational Statistics*, (C. R. Rao, Ed.), Vol. 9, New York: North-Holland Publishing Company, 775-802.
- Hadi, A. S. (1996), *Matrix Algebra As a Tool*, Belmont, CA: Duxbury Press.
- Hadi, A. S. and Ling, R. F. (1998), "Some Cautionary Notes on the Use of Principal Components Regression," *The American Statistician*, 52, 15-19.
- Hadi, A. S. and Simonoff, J. S. (1993), "Procedures for the Identification of Multiple Outliers in Linear Models," *Journal of the American Statistical Association*, 88, 1264-1272.
- Hadi, A. S. and Son, M. S. (1997), "Detection of Unusual Observations in Regression and Multivariate Data," Chapter 13 in *Handbook of Applied Economic Statistics*, (A. Ullah and D. E. A. Giles, Eds.), New York: Marcel Dekker, 441-463.
- Hadi, A. S. and Velleman, P. F. (1997), "Computationally Efficient Adaptive Methods for the Identification of Outliers and Homogeneous Groups in Large Data Sets," *Proceedings of the Statistical Computing Section, American Statistical Association*, 124-129.
- Haith, D. A. (1976), "Land Use and Water Quality in New York Rivers," *Journal of the Environmental Engineering Division, ASCE* 102 (No. EE1. Proc. Paper 11902, Feb. 1976), 1-15.
- Hamilton, D. J. (1987), "Sometimes $R^2 > r_{y \cdot x_1}^2 + r_{y \cdot x_2}^2$, Correlated Variables Are Not Always Redundant," *The American Statistician*, 41, 2, 129-132.
- Hamilton, D. J. (1994), *Time Series Analysis*, Princeton, NJ: Princeton University Press.
- Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986), *Robust Statistics: The Approach Based on Influence Functions*, New York: John Wiley & Sons.
- Hand, D. J., Daly, F., Lunn, A. D., McConway, K. J., and Ostrowski, E. (1994), *A Handbook of Small Data Sets*, New York: Chapman and Hall.
- Hawkins, D. M. (1980), *Identification of Outliers*, London: Chapman and Hall.
- Henderson, H. V. and Velleman, P. F. (1981), "Building Multiple Regression Models Interactively," *Biometrics*, 37, 391-411.
- Hildreth, C. and Lu, J. (1960), "Demand Relations With Autocorrelated Disturbances," *Technical Bulletin No. 276*, Michigan State University, Agricultural Experiment Station.
- Hoaglin, D. C. and Welsch, R. E. (1978), "The Hat Matrix in Regression and ANOVA," *The American Statistician*, 32, 17-22.
- Hocking, R. R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1-49.
- Hoerl, A. E. (1959), "Optimum Solution of Many Variables," *Chemical Engineering Quart. Progr.*, 55, 69-78.
- Hoerl, A. E. and Kennard, R. W. (1970), "Ridge Regression: Biased Estimation for Nonorthogonal Problems," *Technometrics*, 12, 69-82.

- Hoerl, A. E. and Kennard, R. W. (1976), "Ridge Regression: Iterative Estimation of the Biasing Parameter," *Communications in Statistics, Theory and Methods*, A5, 77-88.
- Hoerl, A. E., Kennard, R. W., and Baldwin, K. F. (1975), "Ridge Regression: Some Simulations," *Communications in Statistics, Theory and Methods*, 4, 105-123.
- Hollander, M. and Wollfe, D. A. (1999), *Nonparametric Statistical Methods*, New York: John Wiley & Sons.
- Hosmer, D. W. and Lemeshow, S. (1989), *Applied Logistic Regression*, New York: John Wiley & Sons.
- Huber, P. J. (1981), *Robust Statistics*, New York: John Wiley & Sons.
- Huber, P. J. (1991), "Between Robustness and Diagnostics," in *Directions in Robust Statistics and Diagnostics*, (W. Stahel and S. Weisberg, Eds.), New York: Springer-Verlag.
- Iversen, G. R. (1976), *Analysis of Variance*, Beverly Hills, CA: Sage Publications.
- Iversen, G. R. and Norpoth, H. (1987), *Analysis of Variance*, Beverly Hills, CA: Sage Publications.
- Jerison, H. J. (1973), *Evolution of the Brain and Intelligence*, New York: Academic Press.
- Johnson, D. E. (1998), *Applied Multivariate Methods for Data Analysts*, Belmont, CA: Duxbury Press.
- Johnson, R. A. and Wichern, D. W. (1992), *Applied Multivariate Statistical Analysis*, 3rd ed., Englewood Cliffs, N.J.: Prentice-Hall.
- Johnston, J. (1984), *Econometric Methods*, 2nd ed., New York: McGraw-Hill.
- Kmenta, J. (1986), *Elements of Econometrics*, New York: Macmillan.
- Krasker, W. S. and Welsch, R. E. (1982), "Efficient Bounded-Influence Regression Estimation," *Journal of the American Statistical Association*, 77, 595-604.
- Krishnaiah, P. R. (Ed.) (1980), *Analysis of Variance*, New York: North-Holland Publishing Co.
- La Motte, L. R. and Hocking, R. R. (1970), "Computational Efficiency in the Selection of Regression Variables," *Technometrics*, 12, 83-93.
- Landwehr, J., Pregibon, D., and Shoemaker, A. (1984), "Graphical Methods for Assessing Logistic Regression Models," *Journal of the American Statistical Association*, 79, 61-83.
- Larsen, W. A., and McCleary, S. J. (1972), "The Use of Partial Residual Plots in Regression Analysis," *Technometrics*, 14, 781-790.
- Lawless, J. F. and Wang, P. (1976), "A Simulation of Ridge and Other Regression Estimators," *Communications in Statistics, Theory and Methods*, A5, 307-323.
- Lehmann, E. L. (1975), *Nonparametric Statistical Methods Based on Ranks*, New York: McGraw-Hill.
- Lindman, H. R. (1992), *Analysis of Variance in Experimental Design*, New York: Springer-Verlag.

- Malinvaud, E. (1968), *Statistical Methods of Econometrics*, Chicago: Rand McNally.
- Mallows, C. L. (1973), "Some Comments on C_p ," *Technometrics*, 15, 661-675.
- Manly, B. F. J. (1986), *Multivariate Statistical Methods*, New York: Chapman and Hall.
- Mantel, N. (1970), "Why Stepdown Procedures in Variable Selection," *Technometrics*, 12, 591-612.
- Manly, B. F. J. (1986), *Multivariate Statistical Methods*, New York: Chapman and Hall.
- Marquardt, D. W. (1970), "Generalized Inverses, Ridge Regression, Biased Linear Estimation and Nonlinear Estimation," *Technometrics*, 12, 591-612.
- McCallum, B. T. (1970), "Artificial Orthogonalization in Regression Analysis," *Review of Economics and Statistics*, 52, 110-113.
- McCullagh, P. and Nelder, J. A. (1983), *Generalized Linear Models*, London: Chapman and Hall.
- McCulloch, C. E. and Meeter, D. (1983), Discussion of "Outliers," by R. J. Beckman and R. D. Cook, *Technometrics*, 25, 119-163.
- McDonald, G. C. and Galarneau, D. I. (1975), "A Monte Carlo Evaluation of Some Ridge Type Estimators," *Journal of the American Statistical Association*, 70, 407-416.
- McDonald, G. C. and Schwing, R. C. (1973), "Instabilities of Regression Estimates Relating Air Pollution to Mortality," *Technometrics*, 15, 463-481.
- McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: John Wiley & Sons.
- Moore, D. S. and McCabe, G. P. (1993), *Introduction to the Practice of Statistics*, New York: W. H. Freeman and Company.
- Morris, C. N. and Rolph, J. E. (1981), *Introduction to Data Analysis and Statistical Inference*, Englewood Cliffs, NJ: Prentice-Hall.
- Mosteller, F. and Moynihan, D. F. (Eds.) (1972), *On Equality of Educational Opportunity*, New York: Random House.
- Mosteller, F. and Tukey, J. W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley.
- Myers, R. H. (1990), *Classical and Modern Regression with Applications*, 2nd ed., Boston: PWS-KENT Publishing Company.
- Narula, S. C. and Wellington, J. F. (1977), "Prediction, Linear Regression, and the Minimum Sum of Relative Errors," *Technometrics*, 19, 2, 185-190.
- Obenchain, R. L. (1975), "Ridge Analysis Following a Preliminary Test of the Shrunk Hypothesis," *Technometrics*, 17, 431-441.
- Pregibon, D. (1981), "Logistic Regression Diagnostics," *The Annals of Statistics*, 9, 705-724.
- Rao, C. R. (1973), *Linear Statistical Inference and Its Applications*, New York: John Wiley & Sons.
- Retkowsky, D. A. (1983), *Nonlinear Regression Modeling: A Unified Practical Approach*, New York: Marcel Dekker.

- Ratkowsky, D. A. (1990), *Handbook of Nonlinear Regression Models*, New York: Marcel Dekker.
- Rencher, A. C. (1995), *Methods of Multivariate Analysis*, New York: John Wiley & Sons.
- Rousseeuw, P. J. and Leroy, A. M. (1987), *Robust Regression and Outlier Detection*, New York: John Wiley & Sons.
- Scheffé, H. (1959), *The Analysis of Variance*, New York: John Wiley & Sons.
- Searle, S. R. (1971), *Linear Models*, New York: John Wiley & Sons.
- Seber, G. A. F. (1977), *Linear Regression Analysis*, New York: John Wiley & Sons.
- Seber, G. A. F. (1984), *Multivariate Observations*, New York: John Wiley & Sons.
- Seber, G. A. F. and Wild, C. J. (1989), *Nonlinear Regression*, New York: John Wiley & Sons.
- Sen, A. and Srivastava, M. (1990), *Regression Analysis: Theory, Methods, and Applications*, New York: Springer-Verlag.
- Shumway, R. H. (1988), *Applied Statistical Time Series Analysis*, Englewood Cliffs, NJ: Prentice-Hall.
- Silvey, S. D. (1969), "Multicollinearity and Imprecise Estimation," *Journal of the Royal Statistical Society, (B)*, 31, 539-552.
- Snedecor, G. W. and Cochran, W. G. (1980), *Statistical Methods*, 7th ed., Ames, IA: Iowa State University Press.
- Staudte, R. G. and Sheather, S. J. (1990), *Robust Estimation and Testing*, New York: John Wiley & Sons.
- Strang, G. (1988), *Linear Algebra and Its Applications*, 3rd ed., San Diego: Harcourt Brace Jovanovich.
- Thomson, A. and Randall-Maciver, R. (1905), *Ancient Races of the Thebaid*, Oxford: Oxford University Press.
- Velleman, P. F. (1999), *Data Desk*, Ithaca, NY: Data Description.
- Velleman, P. F. and Welsch, R. E. (1981), "Efficient Computing of Regression Diagnostics," *The American Statistician*, 35, 234-243.
- Vinod, H. D. and Ullah, A. (1981), *Recent Advances in Regression Methods*, New York: Marcel Dekker.
- Wahba, G., Golub, G. H., and Heath, C. G. (1979), "Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter," *Technometrics*, 21, 215-223.
- Welsch, R. E. and Kuh, E. (1977), "Linear Regression Diagnostics," *Technical Report 923-77*, Sloan School of Management, Cambridge, MA.
- Wildt, A. R. and Ahtola, O. (1978), *Analysis of Covariance*, Beverly Hills, CA: Sage Publications.
- Wood, F. S. (1973), "The Use of Individual Effects and Residuals in Fitting Equations to Data," *Technometrics*, 15, 677-695.

索引

- Added-variable plot, 95, 110
- Adjusted multiple correlation coefficient, 289
- Analysis of covariance, 12, 15
- Analysis of variance, 12, 15, 66, 141
- ANOVA, 141
 - multiple regression, 66
 - simple regression, 70
- Assumption
 - constant variance, 86
 - Homogeneity, 86
 - homoscedasticity, 86
 - independent-errors, 86
 - linearity, 32, 86
 - normality, 37, 86
- Autocorrelation, 18, 86, 181, 201
 - first-order, 205
- Autoregressive structure
 - first-order, 205
- Backward elimination, 292
- Best linear unbiased estimator, 59, 83
- Beta coefficients, 253
- Binary
 - response data, 198
 - response variable, 320, 325
 - variable, 12
- BLUE, 59, 83
- Book's Web site, 3
- 添加变量图, 75, 86
- 修正的复相关系数, 48
- 协方差分析, 10, 12
- 方差分析, 10, 12, 52
- 方差分析
 - 多元回归的 ~, 52
 - 简单回归的 ~, 55
- 假定
 - 等方差 ~, 68
 - 方差齐性 ~, 68
 - 方差齐性 ~, 68
 - 误差独立 ~, 68
 - 线性 ~, 67
 - 正态性 ~, 68
- 自相关, 14, 68
 - 一阶 ~
- 自回归结构
 - 阶 ~
- 后向剔除法, 235
- 最佳线性无偏估计量, 47, 65
- Beta 系数, 204
- 二值
 - ~ 响应数据
 - ~ 响应变量, 257
 - ~ 变量, 10
- 最佳线性无偏估计量, 47, 65
- 书的网站, 2

- Carriers, 2
- Centering, 241
- Chi-square distribution, 325
- Classification, 330
- Closed Shop Contract, 4
- Coefficient
 - constant, 29
 - of determination, 42, 60
 - slope, 29
- Coefficients table, 36
- Collinearity, 18, 226
 - problem, 88
- Collinearity-influential observations, 249
- Computer repair data, 26, 31-32, 36, 43
- Condition number, 247, 280, 291
- Confidence
 - interval for β_j , 62
 - interval, 39
 - region, 38
- Constant
 - coefficient, 55-56
 - variance assumption, 86
- Control group, 141
- Cook's distance, 104
- Corrected sums of squares and products matrix, 59
- Correlated errors, 142
- Correlation coefficient, 21, 24
 - between Y and \hat{Y} , 40, 60
 - population, 37
 - sample, 37, 40
- Correlation
 - coefficient, 243
 - matrix, 118, 244, 246, 253, 256
- $Cov(\hat{\beta}_i, \hat{\beta}_j)$, 83
- Covariance, 21, 23
- Covariates, 2
- 承载变量, 1
- 中心化, 195
- χ^2 分布, 262
- 分类, 266
- 只雇佣工会会员的合同, 3
- 系数
 - 常数 \sim , 23
 - 决定 \sim , 34, 47
 - 斜率 \sim , 23
- 系数表, 29
- 共线性, 14, 180
 - \sim 问题, 68
- 共线性影响观测, 200
- 计算机修理数据, 20, 24, 28, 34
- 条件数, 199, 217, 225
- 置信
 - β_j 的 \sim 区间, 49
 - \sim 区间, 29
 - \sim 域, 30
- 常数
 - \sim 系数, 45
 - 等方差假定, 68
- 控制组
- Cook 距离, 82
- 校正的平方与乘积和矩阵, 47
- 相关的误差
- 相关系数, 16, 18
 - Y 与 \hat{Y} 的 \sim , 32, 47
 - 总体 \sim , 29
 - 样本 \sim , 29
- 相关
 - \sim 系数
 - \sim 矩阵
- $\hat{\beta}_i$ 与 $\hat{\beta}_j$ 的协方差
- 协方差
- 协变量

Cross-sectional data, 132, 219–220
 Cumulative distribution function, 320
 Cyclical process, 16
 DASL, 2
 Data set
 advertising, 237
 air pollution, 303–304
 Anscombe, 25
 Anscombe's quartet, 25
 bacteria deaths, 157
 bankruptcy, 322–323
 brain, 174–175
 cigarette consumption, 1, 79, 315
 college expense, 184
 computer repair (expanded), 116
 computer repair, 26–27
 consumer expenditure, 202, 221
 corn yields, 148
 cost of education, 184
 domestic immigration, 6
 education expenditure, 142, 144, 187–188
 Egyptian skulls, 6
 equal educational opportunity, 227
 exam scores, 75
 examination data, 76
 exponential growth, 108
 field goal kicking, 333
 for Exercise 4.10, 119
 for exercises 4.12–4.14, 120
 fuel consumption, 258
 Hald, 269, 280
 Hamilton, 95
 hard disk prices, 179
 heights of husbands and wives, 47, 77
 homicide, 300
 housing starts, 210

截面数据
 累积分布函数
 循环过程
 Data and story library, 2
 数据集
 广告 ~, 190
 大气污染 ~
 Anscombe~
 Anscombe 的四组数据
 细菌死亡 ~
 企业破产 ~
 大脑 ~
 香烟消费 ~
 大学开支 ~
 计算机修理 ~ (增补)
 计算机修理 ~
 消费者开支 ~
 玉米产量 ~
 教育费用 ~
 国内移民 ~
 教育开支 ~
 埃及人颅骨 ~
 教育机会均等 ~
 考试成绩 ~
 考试数据
 指数生长 ~
 定位球 ~
 习题 4.10 用到的 ~
 习题 4.12–4.14 用到的 ~
 燃料消耗 ~
 Hald~
 Hamilton~
 硬盘价格 ~
 丈夫与妻子的身高 ~
 凶杀 ~
 住房启动 ~

- import, 233, 264
- injury incidents, 164
- labor force, 47
- Longley, 279-280
- magazine advertising, 177
- milk production, 3
- New York rivers, 7, 99, 102, 106
- newspapers, 48
- nonlinear, 24
- oil production, 179, 221
- preemployment testing, 133
- presidential election, 149, 151, 178, 260
- publicly available, 2
- quantitative sensory testing, 14
- real estate, 312
- right-to-work laws, 4
- salary survey, 124
- Scottish Hills Races, 111
- ski sales, 141-142, 215-216
- Space Shuttle, 331
- students information, 148-149
- supervisor performance, 52-53, 62, 295
- supervisors, 166, 182
- wind chill factor, 177
- Data
 - collection, 11
 - transformation, 13
- Degrees of freedom, 33, 56, 89
- DFITS, 105
- Diagonal elements, 244
- Discriminant analysis, 330
- Distance
 - orthogonal, 30
 - perpendicular, 30
 - vertical, 30
- DJIA, 221
- Dose-response curve, 182, 197
- 进口 ~
- 伤害事件 ~
- 劳动力 ~
- Longley ~
- 杂志广告 ~
- 牛奶产量 ~
- 纽约州河流 ~
- 报纸 ~
- 非线性 ~
- 石油产量 ~
- 雇佣前测试 ~
- 总统选举 ~
- 公用 ~
- 定量的感觉测试 ~
- 房地产 ~
- 工作权利法 ~
- 薪水调查 ~
- 苏格兰山地赛马 ~
- 雪橇销售 ~
- 航天飞机 ~
- 学生信息 ~
- 主管人员业绩 ~
- 主管人员 ~
- 风寒因素 ~
- 数据
 - 收集 ~
 - ~ 变换, 11
- 自由度
- DFITS, 82
- 对角元, 196
- 判别分析, 266
- 距离
 - 正交 ~
 - 垂直 ~
 - 纵向 ~
- 道琼斯工业平均指数
- 剂量 - 响应曲线

- Dow Jones Industrial Average, 221
- Draftsman's plot, 94
- Dummy variable, 123
- Durbin-Watson statistic, 204
- Eigenvalues, 246
- Eigenvector, 246
- Estimators
 - shrinkage, 272
- Externally studentized residual, 90
- Factors, 2
- Fitted
 - regression equation, 38
 - value, 15, 31, 56
- Fitting
 - method of, 14
 - models to data, 14
- Forecast interval, 39
- Forecasted value, 16
- Forward selection, 292
- F -test, 64
- Full model, 63
- Function
 - cumulative distribution, 320
 - intrinsically nonlinear, 13
 - linear, 13
 - logistic, 198
 - nonlinear, 13, 197
- Goodness-of-fit index, 42
- Hadi's influence measure, 105
- Hat matrix, 82, 89
- Heterogeneity
 - problem, 86
- Heteroscedasticity, 162, 181
 - problem, 86
- High leverage, 100, 269
- Homogeneity, 86
- Homoscedasticity, 161
- Independent-errors assumption, 86
- 道琼斯工业平均指数
- 打样图, 73
- 哑变量
- Durbin-Watson 统计量
- 特征根, 198
- 特征向量, 198
- 估计量
 - 压缩 ~, 219
- 外学生化残差, 71
- 因子, 1
- 拟合的
 - ~ 回归方程,
 - ~ 值
- 拟合
 - ~ 方法, 12
 - 对数据 ~ 模型, 12
- 预报(测)区间
- 预报值, 13
- 前向选择法, 235
- F -检验
- 全模型, 50
- 函数
 - (累积) 分布 ~, 258
 - 本质非线性 ~, 11
 - 线性 ~, 11
 - logistic~
 - 非线性 ~, 11
- 拟合效果指标, 33
- Hadi 的影响量度
- 帽子矩阵, 65, 69
- 异方差(性)
 - ~ 问题, 68
- 异方差(性)
 - ~ 问题, 68
- 高杠杆
- 方差齐性, 68
- 方差齐性
- 误差独立假定, 68

- Independently and identically distributed, 86 独立同分布, 68
- Indicator variable, 119, 123, 216 示性变量
- Industrial psychology, 52 工业心理学, 42
- Influential 强影响
 - observations, 91 ~ 观测
 - point, 99 ~ 点
- Interaction effects, 128 交互效应
- Intercept, 29, 55 截距, 23, 44
- Internally studentized residual, 90 内学生化残差, 71
- Interpretation 解释
 - of regression coefficients, 57 回归系数的 ~
- Intrinsically nonlinear functions, 13 本质非线性函数, 11
- L-R plot, 107 杠杆 - 残差图,
- Ladder of transformation, 174 阶梯变换
- Lagged variables, 220 滞后变量
- Least squares, 14 最小二乘,
 - estimates, 30 ~ 估计
 - line, 30 ~ 直线
 - method, 30, 55 ~ 法
 - properties, 59 ~ 性质
 - weighted, 18, 70, 181 加权 ~ 法
- Leverage-values, 89 杠杆值
- Leverage-residual plot, 107 杠杆 - 残差图,
- Leverages 杠杆
 - weighted, 325 加权 ~ 值
- Linear regression, 15 线性回归
- Linearity 线性
 - assumption, 32, 86 ~ 假定
- Linearizable, 13 可线性化的
- Logistic Logistic
 - distribution, 321 ~ 分布
 - function, 179, 198 ~ 函数
 - models, 182, 198 ~ 模型
 - regression diagnostics, 325 ~ 回归诊断
 - regression, 12, 15, 18, 141, 319, 321 ~ 回归
 - response function, 320 ~ 响应函数
- Logit, 322 Logit, 258
 - model, 198 ~ 模型

- transformation, 321
- Masking problem, 101
- Matrix, 243
 - corrected sums of squares and products, 59
 - correlation, 244, 246, 253, 256
 - Draftsman's, 94
 - hat, 89
 - plot, 94, 118
 - projection, 89
 - variance-covariance, 243-244, 261
- Maximum likelihood, 14, 322
- Mean
 - sample, 22
- Model
 - fitting, 14
 - full, 63
 - non-intercept, 185, 190
 - probit, 198, 321
 - random walk, 224
 - reduced, 63
 - through the origin, 43
- Multicollinearity, 67, 226
- Multiple correlation coefficient, 60, 66, 69, 87, 289
- Multiple regression, 15, 52, 60
 - ANOVA table, 66
 - assumptions, 59
- Multiplicative effect, 128
- Multivariate regression, 14-15
- Nested models, 63, 328
- No-intercept model, 43, 185, 190
- Nonlinear
 - function, 197
 - regression, 15
- Nonparametric statistics, 204
- Normal
 - equations, 55, 82
 - scores, 97
 - ~ 变换, 259
- 伪装问题
- 矩阵
 - 校正的平方与乘积和 ~
 - 相关 ~
 - 打样图
 - 帽子 ~
 - 图 ~
 - 投影 ~
 - 方差 - 协方差 ~
- 最大似然
- 均值
 - 样本 ~
- 模型
 - 拟合 ~
 - 全 ~
 - 无截距 ~
 - probit~
 - 随机游动 ~
 - 简化 ~, 50
 - 通过原点的 ~, 34
- 多重共线性
- 复相关系数
- 多元回归
 - ~ 方差分析表
 - ~ 假定
- 乘积效应
- 多变量回归, 11
- 嵌套模型
- 无截距模型, 34
- 非线性
 - ~ 函数
 - ~ 回归, 11-12
- 非参数统计
- 正态
 - ~ 方程组
 - ~ 得分

- Normality assumption, 34, 86
 Normalizing transformations, 154
 Odds ratio, 321
 One-sample t -test, 44
 Orthogonal
 predictors, 225, 242
 regression, 30
 Outliers, 100, 269
 P-R plot, 107
 Parsimony, 68
 Partial
 regression coefficients, 51, 57
 regression plot, 110
 residual plot, 111
 Plot
 added-variable, 95, 110
 L-R, 107
 matrix, 94, 118
 of Y versus X, 32
 P-R, 107
 partial residual, 111
 partial regression, 110
 potential-residual, 107
 residual plus component, 95
 scatter, 32
 sequence, 203
 Potential-residual plot, 107
 Predicted value, 15
 Prediction
 errors, 224
 interval, 39
 Principal components, 14, 242, 246, 260-261
 Probit
 model, 198, 321
 Projection matrix, 82, 89
 Pure error, 187
 P -value
 for F -test, 65
 正态性假定
 正态化变换
 优势比, 259
 单样本 t -检验, 35
 正交
 ~ 预测变量, 180
 ~ 回归, 23
 异常点
 位势 - 残差图
 吝啬, 53
 偏
 ~ 回归系数
 ~ 回归图
 ~ 残差图
 图
 添加变量 ~
 杠杆 - 残差 ~
 ~ 矩阵
 Y 关于 X 的散点 ~
 位势 - 残差 ~
 偏残差 ~
 偏回归 ~
 位势 - 残差 ~
 残差加分量 ~
 散点 ~
 序列 ~
 位势 - 残差图
 预测值
 预测
 ~ 误差
 ~ 区间
 主成分
 Probit
 ~ 模型
 投影矩阵
 纯误差
 P -值
 F -检验的 ~

- for t -test, 34, 61
- R^2 , 42, 60
- Random walk model, 224
- Reduced model, 63
- Regression
 - assumptions, 85
 - coefficients, 2, 29, 51
 - definition, 1
 - elements of, 7
 - examples of, 3
 - line, 30
 - linear, 15
 - logistic, 12, 15, 18
 - model through the origin, 43
 - multiple, 13, 15, 17, 52, 60
 - multivariate, 14–15
 - nonlinear, 15
 - package, 17
 - parameters, 2, 29
 - partial coefficients, 51
 - ridge, 271
 - robust, 116
 - simple, 13, 15, 17, 52, 62, 70, 74
 - sum of squares, 41
 - trivial models, 44
 - univariate, 14–15
- Regressors, 2
- Relationship between simple and multiple regression coefficients, 75
- Residual
 - internally studentized, 90
 - sum of squares, 33, 41
 - mean square, 43, 289
 - ordinary least squares, 31, 56
 - plots, 90
 - plus component plot, 95, 110–111
 - standardized, 89
- Ridge
 - bias of estimators, 283
- t -检验的 ~
- R^2
- 随机游动模型
- 简化模型
- 回归
 - ~ 假定
 - ~ 系数
 - ~ 定义
 - ~ 的要素
 - ~ 的例子
 - ~ 直线
 - 线性 ~
 - logistic~
 - 通过原点的 ~ 模型
 - 多元 ~
 - 多变量 ~
 - 非线性 ~
 - ~ 软件包
 - ~ 参数
 - 偏 ~ 系数
 - 岭 ~
 - 稳健 ~
 - 简单 ~
 - ~ 平方和
 - 平凡的 ~ 模型
 - 单变量 ~
- 回归变量, 1
- 简单回归与多元回归系数之间的关系
- 残差
 - 内学生化 ~
 - ~ 平方和
 - ~ 均方
 - 普通最小二乘 ~
 - ~ 图
 - ~ 加分量图
 - 标准化 ~
- 岭
 - ~ 估计量的偏倚

- method, 14
- parameter, 272
- regression, 271
- trace, 271, 273
- variance of estimators, 283
- RMS, 289
- Robust regression, 116
- Sample
 - standard deviation, 44
 - variance, 44
- Sampling distribution
 - of $\hat{\beta}_0$, 37
 - of $\hat{\beta}_1$, 33, 37
- Scaling, 241
 - unit length, 242
- Scatter plot
 - of Y versus X, 32
- sequence plot, 203
- Shrinkage estimators, 272
- Simple regression, 15, 52, 62, 70, 74
 - ANOVA table, 70
 - ANOVA, 70
- Simultaneous confidence region, 38
- Slope, 29
- SSE, 41
- SSR, 41
- SST, 41
- Standard deviation, 34
 - sample, 24, 44
- Standard error
 - of $\hat{\beta}$, 83
 - of estimate, 34
- Standard normal distribution, 90
- Standardized
 - deviance residuals, 325
 - Pearsonian residuals, 325
 - residual, 89
 - variables, 243
- Standardizing, 242
 - ~ 方法
 - ~ 参数
 - ~ 回归
 - ~ 迹
 - ~ 估计量的方差
- 残差均方
- 稳健回归
- 样本
 - ~ 标准差
 - ~ 方差
- 抽样分布
 - $\hat{\beta}_0$ 的 ~
 - $\hat{\beta}_1$ 的 ~
- 尺度变换
 - 长度单位化 ~, 195
- 散点图
 - Y 关于 X 的 ~
- 序列图
- 压缩估计量
- 简单回归
 - ~ 的方差分析表
 - ~ 的方差分析
- 同时置信域
- 斜率
- 残差平方和
- 回归平方和
- 总的平方和
- 标准差
 - 样本 ~
- 标准误
 - $\hat{\beta}$ 的 ~
 - 估计的 ~
- 标准正态分布
- 标准化
 - ~ 的偏离残差
 - ~ Pearson 残差
 - ~ 残差
 - ~ 变量
- 标准化

- Stepwise method, 293
- Stimulus-response relationships, 197
- Sum of squared residuals, 56
- Supervisor performance data, 53, 62
- Swamping problem, 101
- Test
 - Durbin-Watson, 204
 - F -, 64-67, 69
 - one-sample, 44
 - runs, 204
 - t -, 34-35, 61-62, 66, 71
 - two-sample, 44
 - Time Series data, 132, 203, 219
 - Total
 - mean square error, 282-283
 - sum of squares, 41
 - variance of OLS estimators, 272
 - variance of ridge estimators, 272
 - variance, 272
 - Transformation, 13, 18, 32
 - ladder of, 174
 - to achieve linearity, 155
 - to achieve normality, 154
 - variance-stabilizing, 154
 - Trivial regression models, 44
 - t -test
 - one-sample, 44
 - two-sample, 44
 - Two-sample t -test, 44
 - Uniqueness of LS solution, 31, 55
 - Univariate regression, 14-15
 - Variable
 - binary, 12, 18, 320, 325
 - dependent, 1
 - dummy, 12, 123
 - explanatory, 1
 - independent, 2
 - indicator, 12, 119, 123, 216
 - lagged, 220
 - 逐步回归方法
 - 刺激-反应关系
 - 残差平方和
 - 主管人员业绩数据
 - 淹没问题
 - 检验
 - Durbin-Watson~
 - F -~
 - 单样本~
 - 游程~
 - t -~
 - 两样本~
 - 时间序列数据
 - 全部(总体)
 - ~均方误差
 - ~平方和
 - ~普通最小二乘估计量的方差
 - ~岭估计量的方差
 - ~方差
 - 变换
 - ~阶梯
 - 线性化~
 - 正态化~
 - 方差稳定~
 - 平凡的回归模型
 - t -检验
 - 单样本~
 - 两样本~
 - 两样本 t -检验
 - 最小二乘解的唯一性
 - 单变量回归
 - 变量
 - 二值~
 - 因~
 - 哑~
 - 解释~
 - 独立~
 - 示性~
 - 滞后~

- predictor, 1
 - qualitative, 12, 18
 - quantitative, 12
 - regressor, 2
 - response, 1
 - role of, 316
 - selection procedures, 285
 - selection, 18
 - standardized, 243
- Variance
 - inflation factor, 240, 282, 291
 - sample, 44
- Variance-covariance matrix, 243-244, 261
- Variance-stabilizing transformations, 154
- VIF, 240
- Web site
 - book's, 3, xiv
 - Case book
 - DASL, 2
 - Electronic Dataset Service, 3
- Weighted
 - least squares, 18, 170, 181
 - leverages, 325
- Welsch and Kuh Measure, 105
- 预测 ~
- 定性 ~
- 定量 ~
- 回归 ~
- 响应 ~
- ~ 的作用
- ~ 选择方法
- ~ 选择
- ~ 标准化
- 方差
 - ~ 膨胀因子, 193
 - 样本 ~,
- 方差 - 协方差矩阵, 196-198, 210
- 方差稳定变换
- 方差膨胀因子, 193
- 网站
 - 本书的 ~, 2
 - Case book~, 2
 - DASL~, 2
 - 电子数据服务系统 ~, 2
- 加权
 - ~ 最小二乘, 12
 - ~ 杠杆值, 262
- Welsch-Kuh 量度, 82