状态风险分析

及其在生物医学中的应用

上册

定常协变量问题

刘韵源 著

科学出版社

内 容 简 介

《状态风险分析及其在生物医学中的应用》一书分两册出版，分别介绍定常和时序协变量问题的风险分析理论与应用. 上册包括相关风险分析、列联风险分析、模型风险分析及风险研究设计四篇. 内容涉及相关研究、回顾和前瞻性研究资料的经典与模型统计分析策略和方法，并附有大量实际应用例子. 作者引用信息量寻优准则和状态变量技术，能较好地分析多因素资料，揭示变量间的交互作用效应，提供高风险状态的重要信息. 本书中的全部算法已由"状态风险分析程序"实现，备有软件盘，以便读者应用.

本书可供生物学、医学、应用统计学及其他有关学科，特别是预防肿瘤学的科研人员、大专院校有关专业师生参考.

状态风险分析
及其在生物医学中的应用
上 册
定常协变量问题
刘韵源 著
责任编辑 马素卿 王爱琳
科学出版社出版
北京东黄城根北街16号
邮政编码：100707
中国科学院印刷厂印刷
新华书店北京发行所发行 各地新华书店经售
*
1990年7月第一版 开本：787×1092 1/16
1990年7月第一次印刷 印张：34 插页：1
印数：001—750 字数：797 000
ISBN 7-03-001626-2/R·72
定价：34.20元

序

现代肿瘤学的研究，需要进行大量复杂的风险分析，内容涉及生物统计学的各个方面，同时也促进了统计理论的发展．作者长期从事癌生物统计学的研究，有机会经常与生物医学专家商讨问题，并深入癌症高发现场调查实况，迫切感到应有一本系统介绍风险分析理论和应用的著作，供有关科研人员参考，并建立一套风险分析统计软件，以便推广应用．在国家科委和我所领导的支持下，我们总结了近年来的研究成果，同时吸收了国内外文献中的先进技术，编写成此书及其支持软件系统，期望能对读者有所裨益．

在本书中，我们提出了广义分量分析与 Medimax 正交旋转理论、组群回归模型和多因变量逐步聚类算法；借助交叉积差和统计量，发展了非参数统计中的一系列算法和定性资料的数量化方法；在引入危险状态概念和信息量寻优准则的基础上，建立了广为适用的变量选择技巧及风险分析策略，提出了广义风险分析模型，能较好处理包括多项目广类目定性资料在内的多因素数据的统计分析，揭示变量间的高次交互作用效应，提供高风险状态的重要信息．相应的统计软件实现了在微机上分析较大资料的要求，一轮分析能给出多项系列结果，具有较高效率．本书分两册出版，分别介绍定常和时序协变量问题的风险分析理论与应用．上册包括相关风险分析、列联风险分析、模型风险分析和风险研究设计四篇，内容涉及相关研究、回顾性与前瞻性研究，以及高风险人群干预试验研究资料的经典与模型统计分析策略和方法，并附有大量实际应用例子，以便于读者理解和参考．由于任何真实事件都是作为过程而存在的，故下册中将讨论时序协变量问题的分析技术，慢性病发病与死亡率的长期预测，肿瘤高发区建立癌症监测-防治网的战略与方法，最佳普查间隔的确定及效益-代价比分析；危险因素有效累积暴露剂量的计算、疾病潜伏期的估计；肿瘤发病多阶段模型和多状态转移理论的时序风险分析等复杂问题．尽管由于作者的工作关系，书中例子多半取自癌症研究领域，但方法本身具有普遍意义，适宜于生物学、医学、生物统计学、人口学、心理学、市场和能源分析，以及应用统计软件的科研人员、大专院校有关专业师生参考．

陈元立同志参加编写了第 9 章；周家丽同志协助编写了第 10 章，并负责完成了上册中全部图表的设计与绘制．美国 Memphis 大学生物统计学家 Wai-yuan Tan（谭外元）教授曾与作者系统讨论过状态风险分析法的原理和应用，给予颇多启示；美国国立癌症研究所的生物统计学家 W. J. Blot 博士、D. Byar 博士、C. C. Brown 博士、Hai-tung King（金海童）教授、M. Gail 博士和 J. Lubin 博士，美国 Cornell 大学生物统计学家 W. Federer 教授，曾对此方法提出过许多宝贵建议；高级软件工程师 D.Pee（皮殿武）博士受美国国立癌症研究所专家组委托，在 IBM 370 计算机上对状态风险分析法的部分算法功能进行过考核，作者谨借此机会表示衷心感谢．

由于作者水平所限，编写过程中肯定有错误和不妥之处，恳切希望同行专家和广大读者不吝赐教，以便得到改进．

作者于中国医学科学院肿瘤研究所

1987 年 10 月

目　　录

第1篇　相关风险分析

第2篇　列联风险分析

第3篇 模型风险分析

第 4 篇　风险研究设计

第1篇　相关风险分析

第1章　多元统计与相关风险分析

相关风险分析主要通过变量间的统计相关来揭示事物或事件之间的连带风险，为探究因果关系提供线索．由于相关研究设计多姿多彩，相应的分析方法也层出不穷，多元统计在相关风险分析中有着广泛应用．本章以美国癌症死亡率与环境及人种因素的分析为例，介绍一些多元统计方法的实际应用．

Wellington 和 Macdonald 等人（1979）在《恶性肿瘤死亡率——环境与人种因素》一书中，介绍了美国 1950—1969 年间人口死亡资料，包括 23 个人口统计学变量，6 个经济收入变量，11 个气候、气象变量，37 个空气污染变量，3 个辐射、20 个消费及 74 个人种变量，目的在于通过多元分析寻找与肿瘤发病有关的重要因素组合．在分析的第 1 阶段，借助多元回归、主成分分析和判别分析技术，从上述诸多因素中选出四组变量，用来导出各种癌症死亡率模型，以确保不同模型间因素效应的可比性．表 1.1 列出了这些变量组合，表 1.2 说明了各变量的含义．在第 2 阶段分析中，采用逐步回归、岭回归（ridge regression）和 Mallows Cp 统计量寻优法，建立各种癌症死亡率的“最佳”子集模型，并绘出包含 12 个变量之模型的岭迹（ridge trace）图，确定每种因素净效应的稳定性．从四组变量选出的“最佳”子集模型中，具有最高复相关系数者被推荐为正式癌症死亡率模型．对“最佳”子集模型的两类离群点（outlier）进行了分析．

表 1.1　用以建立癌症死亡率模型的变量组合

6 变量组合	8 变量组合	10 变量组合	12 变量组合
TEMP	TEMP	TEMP	TEMP
PRELEV	PRELEV	PRELEV	PRELEV
POLLUT	POLLUT	POLLUT	POLLUT
INCM	INCM	INCM	INCM
CIGS	CIGS	CIGM	CIGM
ALCO	MALT, WINE, DISP	ALCO	MALT, WINE, DISP
		FRAN	FRAN
		BRIT	BRIT
		SCAN	SCAN
		OTHEUR	OTHEUR

尽管数据有一定缺陷，尤其是可能观测的因素受到限制，但仍然建立了多数癌症的死亡率模型，不仅用一些因素解释了癌症死亡率的变差，揭示出不同癌症死亡率水平与因素

表 1.2 癌症死亡率模型中各因素效应的解释

变 量	含 义	正效应	负效应
TEMP	1947 年 1 月和 7 月平均气温的第一主成分得分	1 月和 7 月气温高	1 月和 7 月气温低
PRELEV	1921—1950 年间平均年降雨量和平均海拔高度的第一主成分得分	高降雨量低海拔高度	低降雨量高海拔高度
POLLUT	下述变量组的第一主成分得分：用 1970 年人口加权的每平方公里地面上的悬浮微粒、二氧化硫、一氧化碳、碳氢化合物、氮氧化合物含量，以及用 1969 年人口加权的每立方米空气中的苯并芘含量	高污染水平	低污染水平
INCM	1950 年按人头计算每人的经济收入	高收入水平	低收入水平
CIGS	1955 年按人头计算每人的烟草消费量	高烟草消费	低烟草消费量
MALT	1950 年每个成年人的啤酒消费量	高啤酒消费	低啤酒消费
WINE	1966 年每个成年人的葡萄酒消费量	高葡萄酒消费	低葡萄酒消费
DISP	1966 年每个成年人的蒸馏酒消费量	高蒸馏酒消费	低蒸馏酒消费
ALCO	MALT，WINE 和 DISP 的第一主成分得分	高酒精消费	低酒精消费
FRAN	1950 年出生于法国的白人百分比	法国人比例高	法国人比例低
BRIT	1950 年出生于英格兰、威尔斯、苏格兰、北爱尔兰和爱尔兰的白人百分比	英国人比例高	英国人比例低
SCAN	1950 年出生于挪威、瑞典、丹麦的白人百分比	斯堪的纳维亚半岛人比例高	斯堪的纳维亚半岛人比例低
OTHEUR	下述变量组的第一主成分得分：1950年出生于意大利、联邦德国、俄罗斯、波兰、奥地利、捷克斯洛伐克、匈牙利、尼得兰、比利时、立陶宛的白人百分比，以及其中犹太人所占的百分比	其他欧洲人种比例高	其他欧洲人种比例低

组合间的关联，还描述了各种癌症死亡率模型间的联系．尤其令人瞩目的两大因素是都市化程度与人口密度．Wellington 等人早就发现，都市化程度与许多癌症及循环系统疾病的高死亡水平密切相关，而农村田园生活与其他几类主要疾病(包括意外伤亡和暴病)死亡率间的关联较强．这次研究的目的之一是要确定，哪些原始因素造成了慢性病死亡模型中的都市化效应，哪些导致了人口密度效应．结果表明，诸如经济收入、酒精和烟草消费水平，以及人种等社会因素与都市化程度关系更为密切；而气温、雨量、海拔高度、污染及本底辐射等环境因素，与人口密度效应的关联性更强．都市化程度中的经济收入水平被证明主要为一修正因素，而酒精消费因素在消化系统的癌症死亡率模型中显示出重要影响．所有的消费变量都与人种因素高度相关；烟草消费的净效应，由于回归方程中引入了人种变量而被减弱．在各类癌症死亡率模型中，显示出不同人种效应与癌症死亡率风险间关联的惊人一致性．研究结果表明，环境因素中的气候变量(包括雨量和海拔高度)，是所有变量中影响最甚者，并且在一些模型中，此变量还代表了若干未选入因素，例如本底辐射或墨西哥人种；因此可以推测，尚有其他这样的因素漏选．在 6 个污染因素中，每个的加权人群暴露综合变量，显示出相当强的作用，但无疑都不如更为精确的一项测度指标所能有的影响大．气温效应虽然不强，却证明是区分呼吸系统与其他几种主要癌症死亡率的一个有意义的判别因素．对离群点的分析，明确指出了对入选因素不足以解释具有过高或过低癌症死亡水平的个别地区，可能存在其他重要致癌因素．在导出癌症死亡率模型时，显示出特殊影响的一些地区也被找出，并考察了其影响的类型和程度，从中能发现一些入选因素对它们的癌症死亡率水平有着特别显著的影响．

限于篇幅，不宜作详细描述，有兴趣的读者可参阅原著，可以较全面了解多元统计在癌症死亡率资料分析中的实际应用．下面介绍统计分析中的一些重要概念．

§1.1 变量的复共线性

当自变量之间有近似线性关系时，会导致统计分析的困难，例如对回归系数的最小二乘估计有如下影响：

(a) 当一个自变量被引入或剔除时，其余变量的回归系数有较大变化.

(b) 当新数据参加计算或去掉一个数据时，回归系数变化较大.

(c) 回归系数的数值和符号与经验不一致，结果难以解释.

在文献中，常常把这种近似线性关系称为复共线关系 (multicollinearity). 表 1.3 列出了 12 变量组合的单相关矩阵，在各人种变量之间，人种变量与收入水平之间，以及人种与消费变量之间的单相关系数很高，存在复共线关系，直接作回归分析效果未必理想，因此，原作者首先对此 12 个变量作主成分分析，提取它们的主成分来进行下步分析. 发现第 1 主成分解释了联合变差的 56%，能够认为代表了一个综合的都市化因素，其中在收入水平，法国、英国和其他欧洲人种百分比，以及啤酒与烟草消费水平上，有较高的因子载荷. 第 2 主成分解释了总变差的 18%，似乎代表了人口密度因素，在污染水平和气候因素上有较高的因子载荷. 取全部 12 个主成分的样本得分为"自变量"，与各种癌症死亡率作多元回归时，发现在男、女性胃癌、大肠癌、直肠癌、淋巴肉瘤、何杰金氏病，以及男性肾癌、膀胱癌、食管癌和女性胰腺癌、乳癌、卵巢癌的死亡率模型中，第 1 主成分，即综合都市化因素，具有最大的影响；而第 2 主成分，即污染/气候因素，在男、女口腔癌，咽癌，喉癌，主气管、支气管和肺癌，以及男性肝癌、胰腺癌和女性膀胱癌、宫颈癌死亡率模型中的作用最为显著. 用原始变量作直接回归，也能看出此种现象.

复共线资料的统计分析是难度较大的，采用主成分分析作预处理，然后进行主成分回

表 1.3 各因素的相关矩阵

	环境因素			消费、因素						人种因素			
	气候						酒精饮料						
	TEMP	PRELEV	POLLUT	INCM	CIGS	ALCO	MALT	WINE	DISP	FRAN	BRIT	SCAN	OTHEUR
TEMP	1.000	0.354	0.171	−0.406	−0.321	−0.318	−0.515	−0.064	−0.274	−0.246	−0.475	−0.692	−0.429
PRELEV		1.000	0.538	−0.158	0.005	−0.053	−0.177	−0.034	0.057	−0.086	−0.092	−0.443	−0.004
POLLUT			1.000	0.407	0.228	0.395	0.326	0.368	0.330	0.302	0.289	−0.126	0.487
INCM				1.000	0.698	0.806	0.831	0.622	0.656	0.806	0.908	0.660	0.842
CIGS					1.000	0.756	0.631	0.583	0.746	0.742	0.686	0.342	0.566
ALCO						1.000	0.790	0.896	0.903	0.842	0.759	0.500	0.723
MALT							1.000	0.537	0.553	0.712	0.765	0.659	0.834
WINE								1.000	0.763	0.776	0.615	0.301	0.548
DISP									1.000	0.697	0.604	0.367	0.523
FRAN										1.000	0.855	0.515	0.717
BRIT											1.000	0.675	0.821
SCAN												1.000	0.661
OTHEUR													1.000

归分析，是常用的方法之一．须注意的是，由自变量所提取的主成分，只是按照它们对自变量总方差的贡献率来排队的，而非依据与因变量的关联强度选取的；因而可能出现下述情况：第1或第2主成分与因变量的相关系数，反而要小于后面一些主成分（常常在分析中舍弃）与因变量的相关系数．另外，在结果解释时，主成分回归也只能提供粗线条印象，有时不能满足要求．岭回归是分析复共线资料的另一常用技术，将在后面详细介绍．

§1.2 线性模型中的变量选择方法

实际应用时，经常要求从众多的因素中筛选出与因变量有显著关联的变量和变量组合，常常称为“最佳”变量子集．所谓“最佳”变量子集，其实并无通常理解下的最优含义，实际上可能并不存在变量的唯一“最佳”子集．一个回归方程能用于各种目的，目的不同，“最佳”子集的内含可以各异．能有若干子集都适宜组成回归方程，一种好的变量选择方法应该指出全部这样的子集，而不是仅仅产生单一的“最佳”子集．适当变量所构成的各种子集有助于说明数据的结构，帮助我们较好理解客观的基本规律．变量选择过程应被视为旨在探究自变量相关结构，以及它们是怎样单独和联合地对因变量起作用的一细致分析．已经提出了产生全部可能回归子集的程序，然而当有m个自变量时，所需拟合的方程总数达到2^m个．计算量的庞大，使得高速电子计算机也如牛负重，难以胜任．因此，当m较大时，这些方法就无法实现．这样，一些简便的替代算法便应运而生，广为流行，虽然它们只需要有限的计算，却能得到某些有效的实用解．通常所说的“最佳”变量子集，真实含义不过如此，只是多个较为适宜的子集之一，既非绝无仅有的，也未必是真正最佳的．在生物医学研究中更是这样，“最佳”子集还必须有合理的生物医学解释．

除了目的不同，选择和评价各变量贡献的标准也随之有别外，一般说来，采用简易法选择变量有如下特点：

(a) 对因变量确无显著关联的变量，不会选入“最佳”子集．

(b) 对因变量有显著关联，而与其他自变量又不相关的变量，将作为重要变量入选．

(c) 当两个或多个变量对因变量有着显著的综合作用时，虽然它们单独对因变量的作用并不显著，却能结合在一起，构成“最佳”子集，对因变量的变差作出重要贡献．这是多因素分析的明显优点之一，也是研究人员所期望的．后向剔除法有时优于前向选择法和逐步剔选法，主要原因在于存在这样的变量组合：它们单独对因变量的作用极其微弱，一旦结合便产生很强的综合效应．这样的变量子集，采用后向法更易于找到；前向法和逐步法可能漏选．

(d) 与因变量高度相关，彼此间关联也很强的变量，由于它们对因变量所提供的统计信息有相当大的部分是重叠的，它们对因变量的作用，基本上可由少数变量代表，因而只有其中的个别变量作为代表入选，其余的被排除在“最佳”子集之外，总的统计信息却无明显损失，即不会严重影响对因变量取值的估计．至于哪些个别变量能“有幸”作为代表入选，与数据的相关结构及具体算法有关，并无绝对的客观标准．入选的代表变量不同，便构成多个“最佳”变量子集．对于预测目的来说，采用易于观测的子集已经足够．但在风险分析中，由于不同入选变量有着特定的生物医学解释，在此情况下，可在一定的研究假说下，首选符合研究目的的变量子集，通过相应的统计检验最后决定取舍．我们要强调指

出，由于此种原因而未被选入“最佳”子集的变量，仍然是显著的重要变量，不可轻率舍弃，否则会丢失大量有用信息．这点很容易证实：只要将相应代表变量删除，不参加变量选择过程，就会有新的代表来替代，预测效果大体相当．

(e) 当数据中自变量具有复共线结构时，回归系数的估计对于数据的微小变化会非

表 1.4 各种最佳子集癌症死亡率模型中具有显著效应的变量

模型效应	TEMP		PRELEV		POLLUT	
	男	女	男	女	男	女
+***	黑素瘤	黑素瘤	口腔和咽癌，喉癌 主气管、支气管和肺癌 支气管和肺癌 b 食管癌 大肠癌 胰腺癌 膀胱癌 其他皮肤癌			口腔和咽癌 脑瘤
+**	主气管、支气管和肺癌 a	主气管、支气管和肺癌 a	直肠癌	膀胱癌 乳癌 大肠癌	口腔和咽癌 脑瘤 何杰金氏病	食管癌 直肠癌 何杰金氏病 白血病
+*	主气管、支气管和肺癌 其他皮肤癌		肝癌 肾癌 白血病	食管癌 肾癌 卵巢癌 其他皮肤癌	大肠癌 直肠癌 喉癌 主气管、支气管和肺癌 a	淋巴肉瘤 乳癌 大肠癌
−*	肾癌	食管癌 胰腺癌 何杰金氏病 膀胱癌 大肠癌		肝癌	支气管和肺癌 b	
−**	前列腺癌	宫颈癌 卵巢癌			其他皮肤癌	
−***	胃癌 直肠癌	胃癌 直肠癌 肝癌 肾癌 乳癌 子宫体癌				
模型效应	INCM		CIGS		MALT	
	男	女	男	女	男	女
+***				食管癌		肝癌
+**		主气管、支气管和肺癌 a	直肠癌 何杰金氏病	直肠癌 宫颈癌 大肠癌	肝癌	支气管和肺癌 b
+*			大肠癌	口腔和咽癌 卵巢癌 何杰金氏病	大肠癌	大肠癌 膀胱癌
−*	胰腺癌 前列腺癌	喉癌 大肠癌 支气管和肺癌 b 食管癌	肝癌			其他皮肤癌
−*	胰腺癌					

续表 1.4

模型效应	INCM		CIGS		MALT	
	男	女	男	女	男	女
−**	胃癌 何杰金氏病	胃癌 何杰金氏病			脑瘤	脑瘤 食管癌
−***	直肠癌	直肠癌			黑素瘤	黑素瘤 口腔和咽癌
模型效应	WINE		DISP		FRAN	
	男	女	男	女	男	女
+***	胰腺癌	食管癌	食管癌 肾癌	膀胱癌	口腔和咽癌 喉癌 主气管、支气管和肺癌 a 支气管和肺癌 b	
+**	胃癌	胰腺癌	喉癌 前列腺癌		黑素瘤 其他皮肤癌	喉癌 支气管和肺癌 b
+*		主气管、支气管和肺癌 胃癌	口腔和咽癌 膀胱癌	喉癌 支气管和肺癌 b	肝癌 膀胱癌	主气管、支气管和肺癌 其他皮肤癌
−*	前列腺癌	大肠癌	胃癌 何杰金氏病	胃癌 其他皮肤癌	淋巴肉瘤	主气管、支气管和肺癌 a 肝癌 脑瘤 淋巴肉瘤
−**			其他皮肤癌			卵巢癌
−***		子宫体癌				
模型效应	BRIT		SCAN		OTHEUR	
	男	女	男	女	男	女
+***	直肠癌		脑瘤 白血病	脑瘤 白血病 淋巴肉瘤 何杰金氏病	食管癌 胃癌	肝癌 卵巢癌
+**	大肠癌	直肠癌 膀胱癌 子宫体癌	何杰金氏病		直肠癌 肝癌	主气管 支气管和肺癌 胃癌 子宫体癌
+*		大肠癌 乳癌	肾癌 淋巴肉瘤	肾癌 卵巢癌	大肠癌 肾癌	直肠癌 淋巴肉瘤 乳癌
−*	肝癌	口腔和咽癌 肾癌	其他皮肤癌	其他皮肤癌		
−**	其他皮肤癌	白血病	喉癌 主气管、支气管和肺癌			
−***	白血病		支气管和肺癌 b 肝癌	主气管、支气管和肺癌 支气管和肺癌 b 肝癌 膀胱癌 宫颈癌 子宫体癌		

说明：正、负效应的偏相关 t 测验显著性水平：*$0.01<p<0.05$；** $0.001<p<0.01$；*** $p<0.001$.

常敏感，有关变量的效应表现极不稳定，难以作出确切推断．这种现象在调查项目很多，各项指标又彼此密切关联的预防医学研究资料中，经常出现．用原始变量的高阶相乘项作为衍生变量参加分析时，如果高阶相乘效应不存在，相应衍生变量就不会提供显著的附加统计信息，而与原始变量高度相关，导致复共线现象．除了已经提到的技术外，还可以通过删除回归系数不稳定的变量来破坏数据的复共线结构，用约化的非共线数据进行分析．具体方法是引入各变量的入选优先级，在电脑程序中设置可控变量选择功能．

仍采用前面的例子，表 1.4 列出了各种“最佳”癌症死亡率模型中具有显著效应的变

表 1.5　各种岭回归癌症死亡率模型中具有稳定效应的变量

效　应	TEMP	PRELEV	POLLUT	INCM	CIGS	MALT
正效应(+)	男、女性: 口腔和咽癌 主气管、支气管和肺癌 主气管、支气管和肺癌 a 黑素瘤 其他皮肤癌	男、女性: 口腔和咽癌 喉癌 主气管、支气管和肺癌 支气管和肺癌 b 食管癌 大肠癌 直肠癌 膀胱癌 何杰金氏病 黑素瘤 其他皮肤癌	男、女性: 口腔和咽癌 喉癌 主气管、支气管和肺癌 主气管、支气管和肺癌 a 食管癌 大肠癌 直肠癌 肝癌 脑瘤 淋巴肉瘤	男、女性: 主气管、支气管和肺癌 a 白血病	男、女性: 口腔和咽癌 喉癌 主气管、支气管和肺癌 支气管和肺癌 b 食管癌 大肠癌 直肠癌 膀胱癌	男、女性: 胃癌 大肠癌 直肠癌 肝癌 肾癌 膀胱癌 淋巴肉瘤
	男性: 喉癌	男性: 主气管、支气管和肺癌 a 肝癌 胰腺癌 肾癌 脑瘤 淋巴肉瘤		男性: 大肠癌 肾癌	男性: 主气管、支气管和肺癌 a 肾癌 何杰金氏病 黑素瘤	男性: →食管癌 何杰金氏病
		女性: 乳癌 卵巢癌 宫颈癌 子宫体癌	女性: 胰腺癌 膀胱癌 何杰金氏病 白血病 乳癌 卵巢癌 宫颈癌 子宫体癌	女性: 淋巴肉瘤 乳癌 卵巢癌	女性: 乳癌 卵巢癌 宫颈癌	女性: 主气管、支气管和肺癌 支气管和肺癌 b 乳癌
负效应(−)	男、女性: 胃癌 直肠癌 大肠癌 肾癌 膀胱癌 淋巴肉瘤 何杰金氏病				男、女性: 脑瘤 白血病	男、女性: 黑素瘤 其他皮肤癌
	男性: 前列腺癌			男性: 胰腺癌 何杰金氏病 前列腺癌		男性: 脑瘤
	女性: 肝癌			女性: 喉癌	女性: 胃癌	女性: 口腔和咽癌
负效应(−)	女性: 胰腺癌 乳癌 卵巢癌 子宫体癌			女性: 食管癌 宫颈癌 其他皮肤癌		女性: 喉癌 主气管、支气管和肺癌 a →食管癌

续表 1.5

效 应	WINE	DISP	FRAN	BRIT	SCAN	OTHEUR
正效应(+)	男、女性: 口腔和咽癌 喉癌 主气管、支气管和肺癌 主气管、支气管和肺癌 a 支气管和肺癌 b 食管癌 胰腺癌 膀胱癌	男、女性: 喉癌 支气管和肺癌 b 食管癌 胰腺癌 膀胱癌	男、女性: 口腔癌和咽癌 喉癌 主气管、支气管和肺癌 支气管和肺癌 b	男、女性: 食管癌 大肠癌 直肠癌 膀胱癌	男、女性: 胃癌 肾癌 脑瘤 淋巴肉瘤 何杰金氏病 白血病	男、女性: 胃癌 大肠癌 直肠癌 肝癌 肾癌 膀胱癌 淋巴肉瘤 白血病
	男性: 黑素瘤	男性: 口腔和咽癌 肾癌 前列腺癌	男性: 主气管、支气管和肺癌 a 胰腺癌 膀胱癌 黑素瘤 肝癌	男性: 口腔和咽癌 肾癌 淋巴肉瘤 何杰金氏病 前列腺癌	男性: 前列腺癌	男性: →喉癌 食管癌
		女性: 主气管、支气管和肺癌 乳癌 卵巢癌		女性: 喉癌 乳癌 卵巢癌 子宫体癌	女性: 乳癌 卵巢癌	女性: 主气管、支气管和肺癌 胰腺癌 乳癌 卵巢癌 子宫体癌
负效应(−)		男、女性: 其他皮肤癌	男、女性: 脑瘤	男、女性: 白血病 黑素瘤 其他皮肤癌	男、女性: 口腔和咽癌 喉癌 主气管、支气管和肺癌 支气管和肺癌 b 其他皮肤癌	男、女性: 黑素瘤
	男性: 前列腺癌		男性: 前列腺癌	男性: 肝癌	男性: 肝癌	男性: 其他皮肤癌 前列腺癌
	女性: 大肠癌 肾癌 子宫体癌	女性: 肾癌 肝癌	女性: 肾癌 淋巴肉瘤 白血病		女性: 食管癌 膀胱癌 宫颈癌 子宫体癌	女性: 口腔和咽癌 →喉癌 宫颈癌

量. 对于男性原发性主气管、支气管和肺癌的死亡率，就存在两个不同的“最佳”子集模型. 在表 1.4 中，确定为原发性癌者用符号“a”表示，未确定为原发性癌的用“b”表示.

由于存在复共线性，需要检验所得癌症死亡模型中各因素效应的稳定性，为此作岭回归分析，获得更为明确的结果(见表 1.5)，可与表 1.4 中“最佳”子集模型进行比较. 例如，在岭回归模型中，气候因素 TEMP 更确定地把呼吸系统的癌症与肠道及某性别特有的肿瘤区分开来，其对前者有正效应，而对后者却是负效应，正效应中还包括所有的皮肤癌. PRELEV 具有更突出的作用，且似乎无负效应. POLLUT 因素的作用也明显加强，特别在呼吸系统、肝和肠癌死亡率模型中是如此，而且它不作为一负效应因素进入方程. 读者不难对其他因素的效应作出类似比较. 对于在男、女性同一种癌症的死亡率模型中具有相反效应的因素，表中用箭头标出. 有关岭回归的原理和方法，将在后面介绍.

§1.3 线性模型中离群点的分析

离群点分析是回归模型拟合中的重要一环，主要包括对剩余离群点（residual outliers）和影响空间（influence space）离群点的研究．通常的线性回归模型为

$$y_t = \beta_0 + \sum_{l=1}^{m} \beta_l x_{tl} + \varepsilon_t \quad (t = 1, 2, \cdots, N) \tag{1.1}$$

其中 ε_t 表示第 t 次观测误差，并假定 ε_t 相互独立地遵从正态分布 $N(0,\sigma^2)$，σ^2 为误差 ε 的公共方差．若回归系数 $\beta_0,\{\beta_l\}$ 的估计为 $\hat{\beta}_0,\{\hat{\beta}_l\}$，则 y_t 的估计值可写成：

$$\hat{y}_t = \hat{\beta}_0 + \sum_{l \in L} \hat{\beta}_l x_{tl} \quad (t = 1, 2, \cdots, N) \tag{1.2}$$

ε_t 的估计值 e_t 称为残差或剩余，L 为入选子集．

$$e_t = y_t - \hat{y}_t \tag{1.3}$$

在本章例子中，y_t 为某一时期内，地区 t 的某种癌症的年龄调整死亡率．(1.2) 是最小二乘法意义下的“最佳”子集模型．e_t 的绝对值愈大，说明估计值 $\hat{y}_t$ 与实测值 y_t 偏离愈远，这或许是数据固有的可变性所致，也能够因为观测或抽样错误而造成，例如，死因诊断或记录有误，或者包含了不属于研究人群的样本．如果某一地区的癌症死亡率由一个非常突出的高危险因素所表征，而其他地区并无这样的特征，回归方程中也未包括此因素，则形如 (1.2) 的模型就不能真实地描述该地区癌症死亡率与流行因素间的关系．倘若将此因素引入回归方程，模型又可能歪曲其他地区癌症死亡率与各流行因素间的总括关系．当剩余极值是由于观测或抽样错误引起的时，从计算中删除相应样本是适宜的．然而，在产生剩余极值的观测中，有些却属于研究人群的“合法”成员，并反映了数据的固有特性．因此，对剩余绝对值特别大的样本，即剩余离群点，应该仔细加以研究，考察它在模型中的地位和作用，有可能得出有关小概率事件的重要认识．

确定剩余离群点的方法颇多，最简单的是利用标准剩余，即剩余值除以剩余标准差：

$$e_{ts} = e_t / s_y \quad (t = 1, 2, \cdots, N) \tag{1.4}$$

e_{ts} 遵从近似独立的标准正态分布 $N(0,1)$．e_{ts} 的绝对值大于 2 者被确定为一离群点．更严格的标准是 e_{ts} 的绝对值大于 3，甚至 4 时，才定为离群点．在本章例子中，剩余离群点分析的兴趣在于，如果某一地区癌症死亡率的估计值偏低，是否说明该地区有关流行因素的值太低？或者某些能够表征其高死亡率水平的重要危险因素未包括在模型中？倘若癌症死亡率的估计值偏高，是否意味着有关流行因素值过高？对于该地区所有各种癌症死亡率水平的估计是否一致地都偏高？

检测离群点的另一方法是采用距离统计量，即在自变量所构成的影响空间中，计算每个样本点与其他观测点之间的距离．最小二乘法的回归系数估计，难以反映小概率事件，即不能很好揭示个别地区流行因素与癌症死亡率间的特殊关系，因而剩余离群点的分析对此也无能为力．在此情况下，可以采用加权平方标准距离 (WSSD)：

$$\mathrm{WSSD}_t = \frac{1}{s_y^2} \sum_{l \in L} [\hat{\beta}_l (x_{tl} - \bar{x}_l)]^2 \quad (t = 1, 2, \cdots, N) \tag{1.5}$$

其中 $\bar{x}_l = \frac{1}{N}\sum_{i=1}^{N} x_{il}$，为 x_l 的N次观测的平均值，s_y^2 为癌症死亡率模型的剩余均方估计值．此距离测度了每个样本点对模型整个影响空间的贡献大小．对于每个“最佳”子集模型，算出N个样本点的加权平方标准距离 WSSD_i $(i = 1,2,\cdots,N)$，若此值远大于其他样本点的相应值，则判其为影响空间中的离群点，应作进一步分析．

§1.4 相关风险分析中的多元统计方法

从本章的例子不难看出，多元统计方法在相关风险分析中有着广阔的应用前景．回归分析、判别分析、主成分分析和因子分析、典型相关分析、对应分析、聚类分析、非线性映射、数量化理论、通径分析，以及训练迭代法技术，都能够在相关风险分析中发挥积极作用．有关这些统计方法与电脑程序的文献很多，限于篇幅，本书将不作重复介绍，有兴趣的读者可参阅有关文献．我们将集中讨论如下问题的分析方法：设对某种现象进行了N次试验，每次试验 i 都观测了 p 个指标 $\mathbf{y} = (y_1, y_2, \cdots, y_p)'$，并记录了可能影响这 p 个指标的m个因素的状态 $\mathbf{x} = (x_1, x_2, \cdots, x_m)'$，$N$次试验的数据用矩阵表示为 $\mathbf{X} = (x_{il})_{N\times m}$，$\mathbf{Y} = (y_{ti})_{N\times p}$，$x$ 和 y 皆可以是定量或定性资料．分析目标是，在考察两组变量 $\mathbf{x}$ 和 $\mathbf{y}$ 相关或关联的同时，依据观测数据中所含的信息，从中挑选出某些重要指标，以及对这些指标有显著影响的某些重要因素来，对N个样本进行分类，评价事件出现的条件与风险，为探究因果关系提供线索．