信度分析

@文彤老师

信度分析

- 探索研究事物间的相似性
- 用于评价问卷这种测量工具的稳定性或可靠性，具体来说就是用问卷对同一事物进行重复测量时，所得结果的一致性程度。
  - 内在信度：调查表中的一组问题（或整个调查表）是否测量的是同一个概念，也就是这些问题之间的内在一致性如何。
  - 外在信度：在不同时间进行测量时调查表结果的一致性程度。最常用的外在信度指标是重测信度，即用同一问卷在不同时间对同一对象进行重复测量，然后计算一致程度。

信度分析

- 系数应当到多大才能认为该问卷信度较高没有统一的标准，但根据多数学者的观点，任何测验或量表的信度系数如果在0.9以上，则该测验或量表的信度甚佳；信度系数在0.8以上都是可接受的；如果在0.7以上，则该量表应进行较大修订，但仍不失其价值；如果低于0.7，问卷的测量效果非常差！
  - 在大型量表中，往往一组问题用来集中测量某一方面的信息。此时信度分析应当按问题组来进行，即测量同一信息的一组问题间信度如何，而不是直接测量整个量表的信度。

案例：问卷信度分析

- 数据item.sav记录了某问卷测量的数据，该问卷共有10道题目，均为9分量表，高分数代表同意此观点，共测量了100人。现希望考察该问卷的信度究竟如何。
  - Scale if item deleted

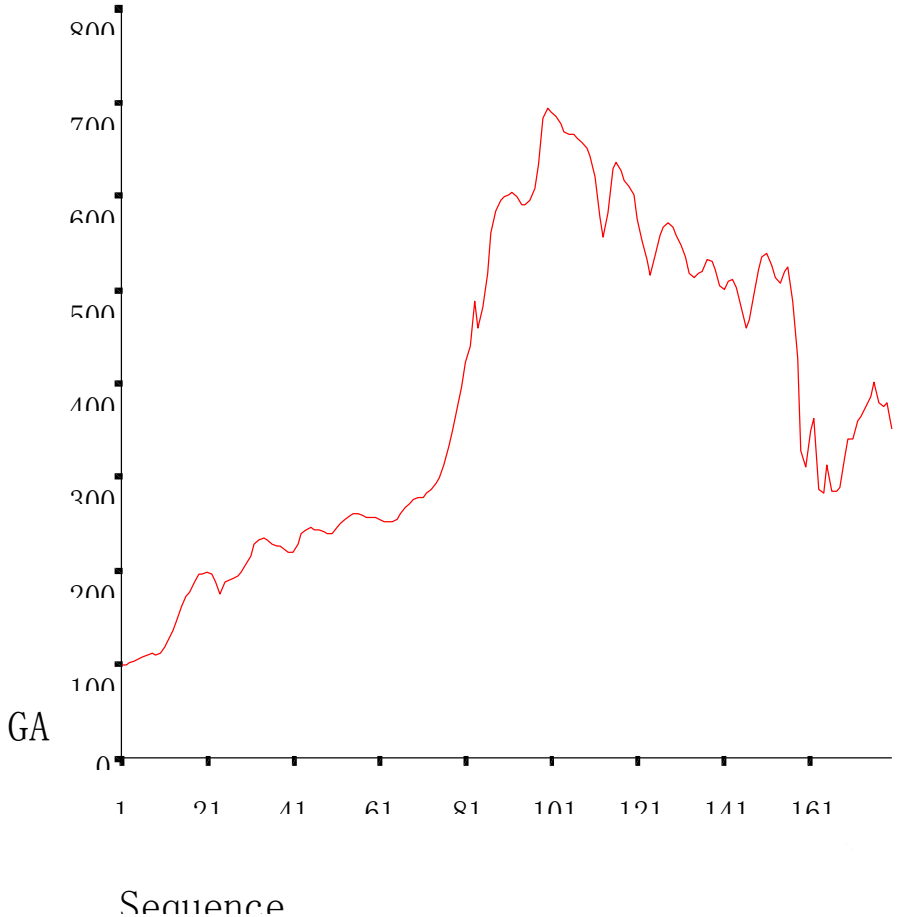
时间序列模型

@文彤老师



时间序列模型简介

- 依时间顺序排列起来的一系列观测值
  - 观测值之间不独立
- 它考虑的不是变量间的因果关系，而是重点考察变量在时间方面的发展变化规律，并为之建立数学模型
- 使用时间序列模型的前提
  - 有足够长的数据序列
  - 数据序列的变动是稳定而规律的



方法分类

- Time domain：将时间序列看成是过去一些点的函数，或者认为序列具有随时间系统变化的趋势，它可以用不多的参数来加以描述，或者说可以通过差分、周期等还原成随机序列。
- Frequency domain：认为时间序列是由数个正弦波成份叠加而成，当序列的确来自一些周期函数集合时，该方法特别有用。

时间序列的构成

- 长期趋势
  - 指一种长期的变化趋势。它采取一种全局的视角，不考虑序列局部的波动
  - 中国的GDP呈现一种上升的长期趋势
- 季节变化(Season)
  - 反映一种周期性的变化一般在一年中完成
  - 虽然称作“季节”，但是周期并不一定是季度，也可以是月、周等其它能在一年内完成的周期。因为，大多数的周期都以季节的形式出现，所以称作季节变化。
    - 圣诞节效应
    - 冷饮的销售情况

时间序列的构成

- 循环变化(Cyclic)
  - 指一种较长时间的周期变化。一般来说循环时间为2—15年。循环变化一般会出现波峰和波谷，呈现一种循环往复的现象。
  - 经济危机
- 不规则变化(error)
  - 指时间序列中无法预计的部分
  - 序列的随机波动
  - 突发事件引起的
  - 它是无法预测的
    - 在分析中往往又将其称为白噪声

时间序列的构成

- 并不是每个序列都包含所有4种成分。比如以年为时间单位的序列就不会有季节变化；
- 一些较短的序列也看不出循环变化。所以在分析的时候得具体情况具体分析。

时间序列分析面临的问题

- 基础统计数据
  - 报表系统——质量
  - 统计口径——不统一
  - 范围变化、区域划分——变化
- 理论和技术培训

时间序列分析对长度的要求

- 不同的序列分析方法
  - 差分
- 历史数据越多，对预测的帮助越大
- 如果存在周期，一般需要4个周期以上
  - 有人认为应当在20个周期以上

SPSS中涉及的功能

- 预处理模块
  - 用于填充序列缺失值的Replace Missing Values过程
  - 建立时间变量的Define dates过程
  - 将序列平稳化的Create Time Series过程。
- 图形化观察 / 分析模块
  - 时间序列在分析中高度依赖图形
  - 序列图(Sequence Chart)
  - 自相关 / 偏相关图(Autocorrelation Function, ACF & Partial Autocorrelation Function, PACF)
  - 交叉相关图（Cross- correlation Function, CCF）、周期图(Periodogram)和谱密度图（Spectral Chart）

SPSS中涉及的功能

- 分析模块
  - 指数平滑法（Exponential Smoothing）
  - 自回归线性模型（Autoregressive model）
  - ARIMA模型
  - 季节解构（Seasonal Decomposition）
- 软件的基本趋势是越来越高度自动化：当然是以足够多的数据量为前提
  - 从14版之后开始引入自动建模功能

时间序列问题的基本分析思路

- 平滑与季节分解
  - 用移动平均的方式消除波动，反映出主要的趋势
- 回归模型
  - 根据历史数据建立相应的回归模型用于预测
- ARIMA
  - 真正的时间序列模型建模分析方法

平滑与季节分解

- 是利用时间序列资料进行短期预测的一种方法。它的基本思想是：除去一些不规则变化后，时间序列将剩下一些基本的变化模式，而这种变化模式将延续到将来。
- 平滑的主要目的就是除去时间序列的不规则变化，把时间序列的基本变化模式突现出来，作为短期预测的基础。
  - 因此也有人把平滑称作“修匀”

平滑与季节分解

- 平滑的方法很多。广义上说，可以认为回归也是一种平滑。因为拟合回归曲线可以把杂乱的观测数据修匀得到连续而光滑的曲线。
- 移动平均和移动中位数等这些不同的平滑技术又被称为平滑器。
  - 平滑处理后，可以得到一些新的序列。不同的平滑器得到的新序列是不同的。选择合适的平滑器，寻求最佳的预测效果是平滑处理的关键。

案例：NRC数据的建模预测

- 美国1947年1月到1969年12月住宅建筑的数据，分析目的是希望能过通过历史数据来预测1970年全年的情况。 Nrc.sav

时间序列模型简介

- 预处理过程
  - 缺失值的填补
  - 时间变量的定义
  - 时间序列的平稳化
    - 一次差分
    - 季节差分

时间序列模型简介

- 时间序列趋势的图形化观察
  - Sequence Chart：序列图
    - 实际上就是一种特殊的线图
  - Autocorrelation Chart：做单个序列，任意滞后（包括负的滞后，也就是超前）的自相关和偏相关图
    - 对序列图的初步观察结果作进一步确认
    - 重点关心主要的相关趋势，然后再对模型进一步修正

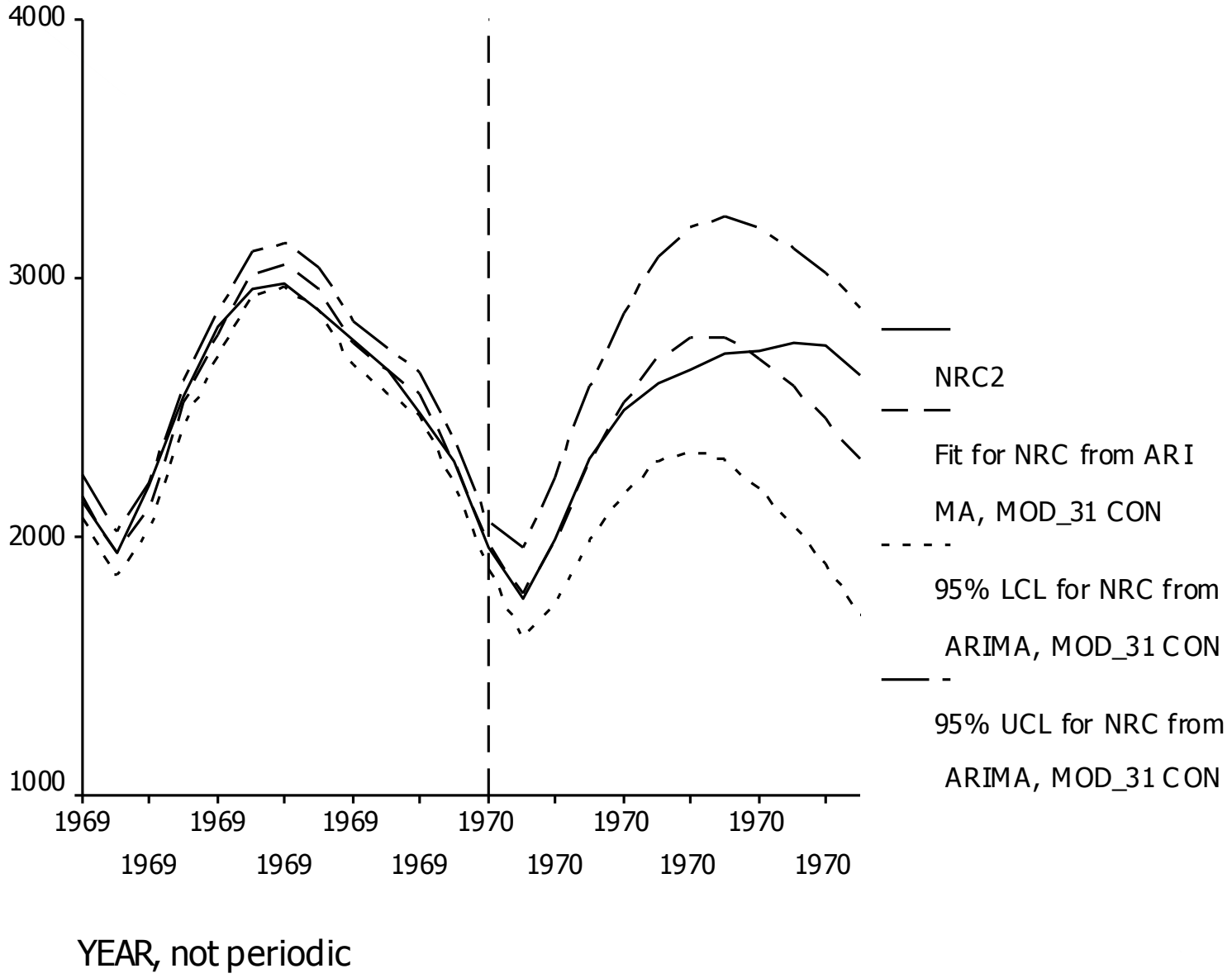
时间序列模型简介

- 时间序列趋势的图形化观察
  - Autocorrelation Chart
    - 自相关图：自回归系数的变化
    - 偏相关图：偏回归系数的变化

模型	自回归系数	偏回归系数
AR(p)	拖尾	lag=p 后骤减
MA(q)	lag=q 后骤减	拖尾
ARMA(p,q)	拖尾	拖尾

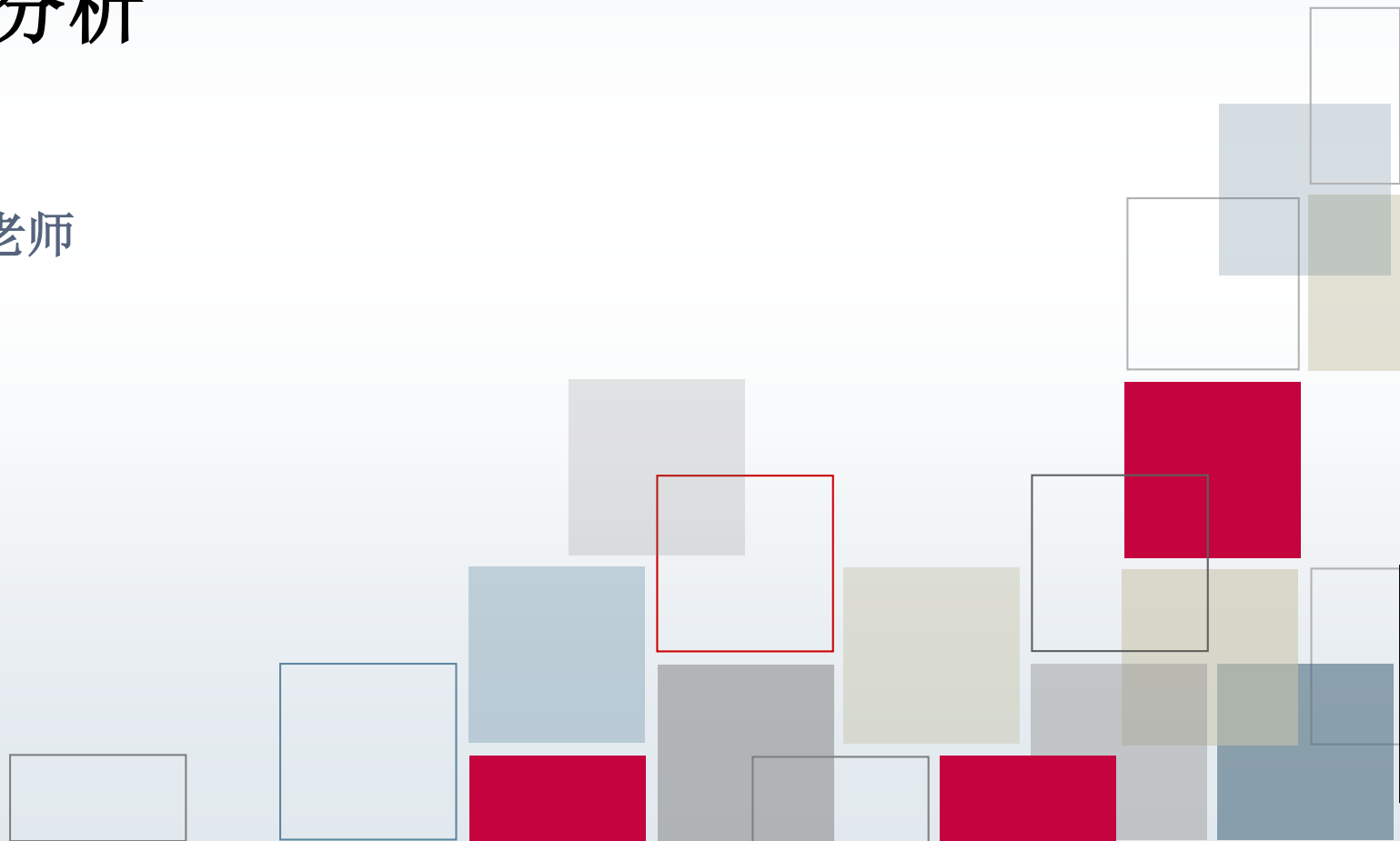
时间序列模型简介

- 模型拟合
  - 几乎均可包含在ARIMA模型族中
  - ARIMA（1,1,1）*（0,1,1）12
  - 寻找适当的参数是一个反复尝试的过程
- 预测
  - 近期预测效果较好
    - 一般预测范围不要超过2个周期
  - 放大局部便于观察


4000
3000
2000
1000
1969
1969
1969
1970
1970
1970
1969
1969
1969
1970
1970
1970
NRC2
Fit for NRC from ARI
MA, MOD_31 CON
95% LCL for NRC from
ARIMA, MOD_31 CON
95% UCL for NRC from
ARIMA, MOD_31 CON
YEAR, not periodic

生存分析

@文彤老师



生存分析的用途

- 职员在公司任职的时间
- 产品的寿命
- 晋升所需时间
- 客户忠诚度研究

传统方法在分析随访资料时的困难

- 时间和生存结局都成为了要关心的因素
  - 除了生存结局作为判定标准以外，只要能让病人存活时间延长，这种药物也应当是被认为有效的。即时间延长也认为有效
  - 如果将两者均作为因变量拟和多元模型，则时间分布不明（肯定不呈正态分布，在不同情况下的分布规律也不同），拟和多元模型极为困难

传统方法在分析随访资料时的困难

- 存在大量失访
  - 失去联系（病人搬走，电话号码改变）
  - 无法观察到结局（死于其他原因）
  - 研究截止
- 显然，将失访数据无论是算作死亡还是存活似乎都不大合理

泼些冷水

- 在针对总体生存时间都比较短的疾病的研究中，生存分析比只考虑结局，不考虑时间因素的Logistic模型等方法的效率要高得多
- 相反，如果相应疾病的结局都非常好，比如几乎人人都可痊愈的感冒，生存分析的利用价值也许并不大，因为除非是特殊情况，感冒晚好两天对医生或病人来说又能有多大影响呢？

生存分析的主要研究内容

- 描述生存过程
  - 研究人群生存状态的规律
  - 研究生存率曲线的变动趋势
  - 是人寿保险业的基础
- 生存过程影响因素分析及结局预测
  - 了解哪些因素会影响生存过程
  - 对生存结局加以预测
    - 在临床中应用的非常广泛

常用术语

- 失效事件(Failure event)
  - 也被称为称“死亡”事件或失败事件，表示观察到随访对象出现了我们所规定的结局。失效事件的认定是生存分析的基石，必须绝对准确。
  - 失效事件应当由研究目的而决定，并非一定是死亡（如研究灯泡寿命），而死亡也被并非一定是发生了失效事件（如肺癌患者死于其他疾病）。

常用术语

- 截尾值（Censored value）
  - 终止随访不是由于失效事件发生，而是无法继续随访下去，常用符号“＋”表示。
  - 生存但中途失访：包括拒绝访问、失去联系或中途退出试验。
  - 死于其它与研究无关的原因：如肺癌患者死于心机梗塞、自杀或因车祸死亡，终止随访时间为死亡时间。
  - 随访截止：随访研究结束时观察对象仍存活。

常用术语

- 生存时间（Survival time）
  - 随访观察持续的时间，按失效事件发生或失访前最后一次的随访时间记录，常用符号*t*表示。
  - 根据失效事件的定义不同，生存时间可以是各种类型的指标，甚至于根本就不是“时间”
    - 从症状缓解到再次恶化
    - 设备从开始使用到失效
    - 汽车的累积行驶里程

常用术语

- 生存率（Survival rate）
  - 实际上应当是生存概率，指某个观察对象活过*t*时刻的概率，常用*p(x>t)*表示。
  - 根据不同随访资料的失效事件，生存率可以是缓解率、有效率等。

生存分析方法分类

- 参数法
  - 首先要求观察的生存时间*t*服从某一特定的分布，采用估计分布中参数的方法获得生存率*p(X>t)*的估计值。
  - 生存时间的分布可能为指数分布、Weibull分布、对数正态分布等，这些分布曲线都有相应的生存率函数形式。只需求得相应参数的估计值，即可获得*p(X>t)*的估计值和曲线。

生存分析方法分类

- 非参数法
  - 实际工作中，多数生存时间的分布不符合上述所指的分布，就不宜用参数法进行分析，应当用非参数法。
  - 这类方法的检验假设与以往所学的非参数法一样，假设两组或多组的总体生存率曲线分布相同，而不论总体的分布形式和参数如何。
  - 非参数法是随访资料的常用分析方法。

生存分析方法分类

- 半参数法
  - 只规定了影响因素和生存状况间的关系，但是没有对时间（和风险函数）的分布情况加以限定
  - 这种方法主要用于分析生存率的影响因素，属多因素分析方法，其典型方法是Cox比例风险模型

SPSS中的相应模块

- Nonliner过程
  - 可以针对任何种类的时间分布加以拟和
- Life tables过程
  - 分析分组生存资料，主要用于计算寿命表
- Kaplan-Meier过程
  - 用于未分组生存资料
- Cox Regression过程
  - 最重要的一个分析方法
- Cox w/Time-Dep Cov过程
  - 对比例风险模型的扩展，允许影响因素的影响程度虽时间而变化。

Kaplan-Meier过程

- 是最基本的一种生存分析方法
- 案例：Prednisolone新药对慢性肝炎疗效的研究

表 25.1　44 名慢性活动性肝炎患者的生存时间（月）

组别	生存时间（月）
新药	2 6 12 54 56† 68 89 96 96 125^{+} 128^{+} 131^{+} 140^{+} 141^{+} 143 145^{+} 146^{+} 148^{+} 162^{+} 168 173^{+} 181^{+}
对照	2 3 4 7 10 22 28 29 32 37 40 41 54 61 63 71 127^{+} 140^{+} 146^{+} 158^{+} 167^{+} 182^{+}

Cox比例风险模型入门

- 属于半参数模型
- 与参数模型相比，该模型不能给出各时点的风险率，但对生存时间分布无要求，可估计出各研究因素对风险率的影响，因而应用范围更广。为了纪念Cox的贡献，统计学家把它称为Cox比例风险模型。

Cox比例风险模型入门

- 在比例风险模型中，假设在时点t个体出现观察结局的风险大小可以分解为两个部分。
- 有一个基本（本底）风险量$h_0(t)$，代表没有任何自变量影响下的生存状况。
- 第i个影响因素使得在任意一个时点t的死亡风险从$h_0(t)$增加$e^{(bi\,X\,i)}$倍而成为$h_0(t)*e^{(bi\,X\,i)}$。

Cox比例风险模型入门

- 因此，如果在k个因素同时影响生存过程的情况下，在时点t的风险量（常称为风险率hazard rate或风险函数（hazard function）就为：
  - $h(t)=h_0(t)e^{(b1 X 1)}e^{(b2 X 2)}\ldots e^{(bi X i)}$
  - $h(t,X)= h_0(t)e^{(b1 X 1+b2 X 2+\ldots+bk X k)}$
- 上式取对数，移项得
  - $Log[Rh(t)]=Log[h(t,X)/h_0(t)]= b_1 X_1+\ldots+b_k X_k$

Cox比例风险模型入门

- $Log[Rh(t)]=Log[h(t,X)/h_0(t)]= b_1 X_1+\ldots+b_k X_k$
- $h_0(t)$
  - 表示个体在时点t的基准死亡情况（基础风险函数，为发病密度或死亡密度）。
- $Beta_i$
  - 与以前相同，可直接理解为Xi的回归系数
  - β的实际含义是：当变量X改变一个单位时，引起的死亡风险改变倍数的自然对数值。

Cox比例风险模型入门

- E^{beta}
  - 相对危险度（RR，Relative Risk）
  - 表示两种情况下发病密度或者说发病概率之比
  - 显然，如果RR>1则说明相应的自变量取值增加，会导致个体的发病/死亡风险增加若干倍。
  - 因此RR在医学中得到了极为广泛的应用，其使用价值也高于OR。