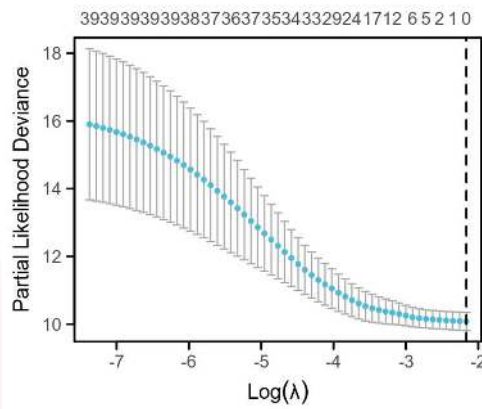


临床意义 — 预后 Lasso 系数筛选[云]



网址: <https://www.xiantao.love>



更新时间: 2023.03.06

目录

基本概念	3
应用场景	3
分析流程	4
结果解读	6
数据格式	8
参数说明	9
预后参数	9
ID 列表	10
方法	11
点	13
误差线	14
标题文本	15
风格	16
图片	16
结果说明	17
主要结果	17
补充结果	18
方法学	19
如何引用	20
常见问题	21

基本概念

- **Lasso 回归**：在线性回归的基础上，通过增加**惩罚项**（ $\lambda \times \text{斜率的绝对值}$ ），减少模型的过拟合，提高模型的泛化能力。另外一种也是通过增加惩罚项来减少模型的过拟合的方法是岭回归，对应的惩罚项是（ $\lambda \times \text{斜率的平方}$ ）。惩罚项在机器学习领域也叫做正则化，其中，Lasso 回归的惩罚项是**L1 正则化**（曼哈顿距离（参数绝对值求和）），而岭回归的惩罚项是**L2 正则化**（欧氏距离（参数平方值求和））
- Lasso 可用于 logistics、Cox 其中，此模块就是 Lasso 在预后中的应用。预后 Lasso 常常出现在构建预后模型或者筛选变量上，最常出现两种图，一种是 系数(λ)筛选的图，另外一种为变量轨迹图。Lasso 的 λ 筛选一般会采用**交叉验证**的手段进行筛选，常见的会有五折和十折交叉验证。

应用场景

将预后 Lasso 系数筛选过程中各个 λ 值（惩罚项）对应的统计量(似然偏差值或 C 指数)进行可视化，以**构建预后模型或者筛选变量**。当样本较少或者变量较多（少于样本数一半的变量）时，可以用 Lasso 直接构建预后模型或者筛选变量。

分析流程

云端数据 → lasso 预后分析 → lasso 系数筛选可视化

➤ 数据格式：

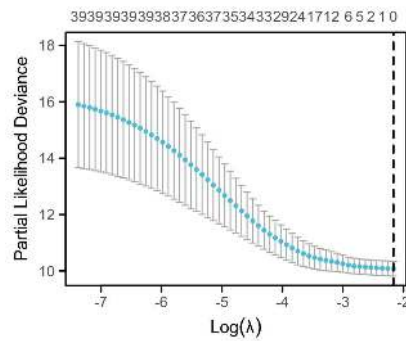
- 云端数据：提供预清洗好的云端数据，不同平台的云端数据集对应的分子有可能不同。
- 在主要参数[预后参数]选择预后类型，不同的数据集之间的预后类型可能不一样！可以选择
 - ◆ OS[Overall Survival]：总体生存期
 - ◆ DSS[Disease Specific Survival]：无病生存期
 - ◆ PFI[Progress Free Interval]：无进展间隔
- 在主要参数[ID 列表]选择样本数据中需要进行分析的分子列表

➤ Lasso 预后分析：

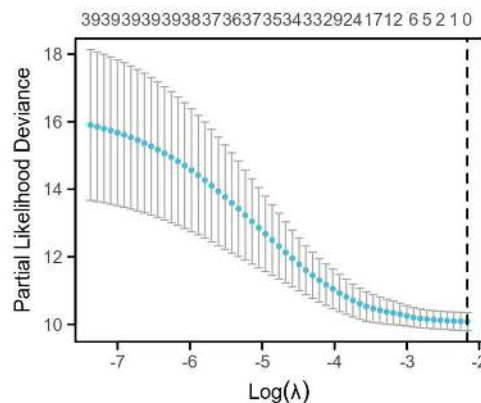
- 构建 lasso 预后模型
- 计算模型的 lambda 值
- 通过 lambda 值计算变量的系数值
- 筛选掉 lambda 值对应系数为 0 的变量(系数为 0 表示变量之间不存在相关关系，在预后模型中没有实质上的意义)

➤ Lasso 系数筛选可视化

- Lasso 预后分析得到的 λ 值取对数, 对应到 lasso 系数筛选可视化结果的横坐标值
- Lasso 预后分析得到的不同指标下的似然偏差值 (deviance) (默认) 或 C 指数 (c-index) 对应到 lasso 系数筛选可视化结果的纵坐标
- 进行可视化, 结果如下:



结果解读



- 下方 x 轴：表示 Lasso 回归中惩罚项 λ 值取对数 ($\log(\lambda)$)
- 上方 x 轴的数字：表示每个 λ 值对应的非 0 系数的变量个数
 - 这些数字对应的值是说：不同 λ 值计算得到模型中所有变量系数不为 0 的变量的个数，而不是所有的变量(要是数值与变量个数对应不上，则是因为缺少的那些变量间不存在相关关系(系数为 0)被筛选掉了)
 - 由于可视化结果是 ggplot2 格式，故不能展示全部的数值
- y 轴：表示在不同指标下的似然偏差值 (deviance) (默认) 或 C 指数 (c-index)
- 每个点：表示数据在进行交叉验证过程中，每个 λ 对应的似然偏差值 (默认) 或 C 指数的均值
- 每条竖线 (误差线)：表示数据在进行交叉验证过程中，每个 λ 对应的似然偏差值的标准误
- 左边虚线：表示评价指标最佳的 λ 值 (λ_{\min})
- 右边虚线：表示评价指标在最佳值 1 个标准误范围的模型的 λ 值 (λ_{1se})
- 当选择的指标为 deviance 时，y 值 (似然偏差值) 越小对应的模型越好

- 当选择的指标为 c-index 时, y 值 (C 指数) 越大对应的模型越好
- 当 λ_{\min} 和 λ_{1se} 一样(图中只有 1 根虚线并且在最右侧), 说明模型没有筛选出来任何一个非 0 系数的变量; λ_{\min} 可能对模型过于严格, λ_{1se} 对应的变量越少, 模型会更加简洁; 两个都可以选, 比较常用的是 λ_{\min} , 如果 λ 对应的变量较多, 也会用 λ_{1se}



数据格式

提供预清洗好的云端数据，不同平台的云端数据集对应的分子有可能不同。选择了云端数据后，只需要把分子列表输入到右侧「分子列表」选项卡中即可进行分析和可视化

(此样本数据使用的是食管癌数据，如下：)

数据参数

云端数据 ⓘ

食管癌 / TCGA / TCGA-ESCA / RNAseq / STAR / TPM @过滤:去除正常+去除无临床信息 @处理:log2(v...



参数说明

(说明：标注了颜色的为常用参数。)

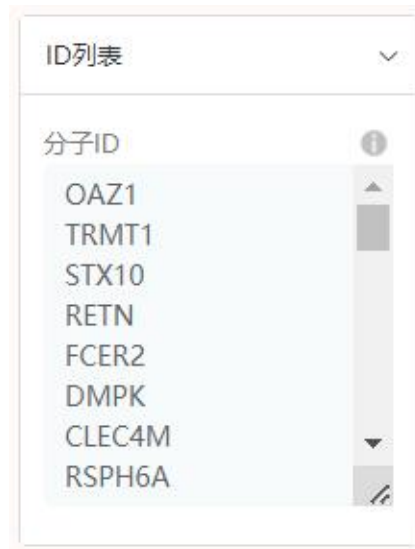
预后参数



- 预后类型：可选不同的预后类型。不同的数据集之间的预后类型可能不一样!
可以选择：

- OS[Overall Survival] (默认)：总体生存期
- DSS[Disease Specific Survival]：无病生存期
- PFI[Progress Free Interval]：无进展间隔

ID 列表



- 分子 ID: 这里输入的是进入到 lasso 筛选的分子列表。这里的格式可以是分子/基因名
 - 注意: 输入的分子有可能会匹配不上对应的表达谱数据, 输入的分子最终的匹配的情况需要在说明文本中进行查看
 - 一般不会输入很多分子的, 一般是在几十到一两百 (最多支持 200 个), 太多的分子未必就能筛选出好的结果, 太少的分子也是一样的。如果分子列表的分子很多, 建议是和“筛选预后分子”模块中进行了预后分析有意义的分子先进行一次交集分析 (利用韦恩图), 得到预后有意义的分子 (缩小分子数量) 后再进行 lasso, 这个也是可以作为一种尝试

方法

方法

lambda指标

deviance

验证方法

十折交叉验证

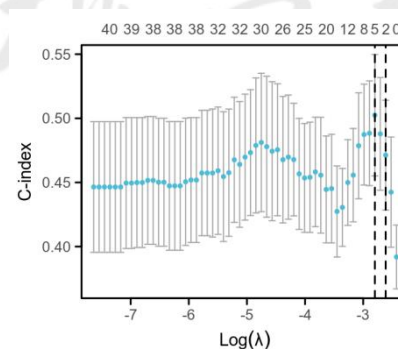
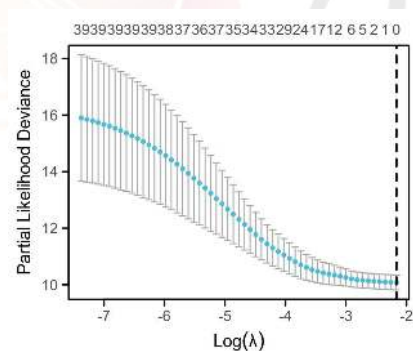
种子号

2022

➤ lambda 指标：可以选择 lasso 系数筛选的指标：deviance（似然偏差值）（默认）或 c-index（C 指数），如下：左侧为 deviance，右侧为 c-index

■ 当选择的指标为 deviance 时，y 值（似然偏差值）越小对应的模型越好

■ 当选择的指标为 c-index 时，y 值（C 指数）越大对应的模型越好



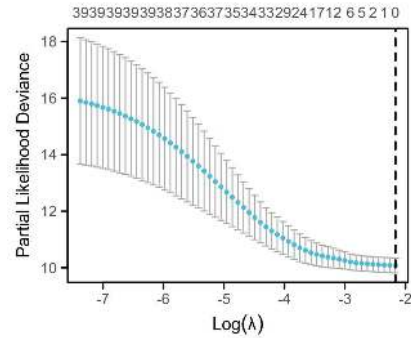
➤ 验证方法（交叉验证的倍数）：可选三折、五折、七折、十折交叉验证。例如十折交叉验证，就是把数据分成 10 份，轮流把 9 份数据作为训练集训练模型，另外 1 份作为验证集验证模型。默认是选择十折交叉验证，如下：

方法

筛选指标 deviance

交叉验证 十折交叉验证

种子号 2022

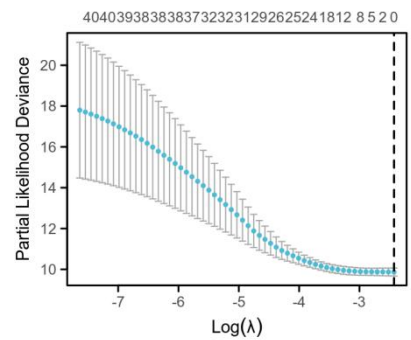


方法

筛选指标 deviance

交叉验证 三折交叉验证

种子号 2022



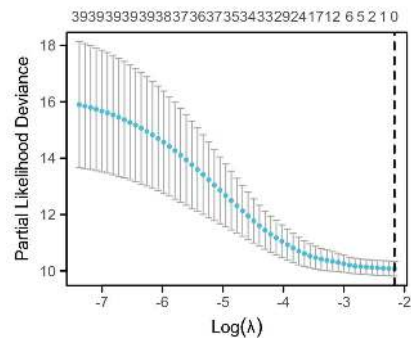
- 种子号：可填入其他的数字，默认为 2022。由于在进行交叉验证的过程中会涉及到对数据的抽样和分训练集和验证集，故不同的种子号对应的结果都会有不同，但是只要是同一份数据同一个种子号，对应的结果都是一样的，如下：

方法

筛选指标 deviance

交叉验证 十折交叉验证

种子号 2022

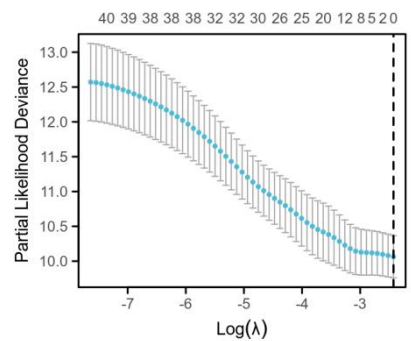


方法

筛选指标 deviance

交叉验证 十折交叉验证

种子号 2023



点



点

描边色

填充色

样式 圆形

大小 0.5

不透明度 1

- 描边色：可以修改图中点的描边色
- 填充色：可以修改图中点的填充色
- 样式：可以选择图中点样式类型，可选择圆形、正方形、菱形、三角形、倒三角
- 大小：可修改点的大小
- 不透明度：可修改点的透明度。0 为完全透明，1 为完全不透

误差线

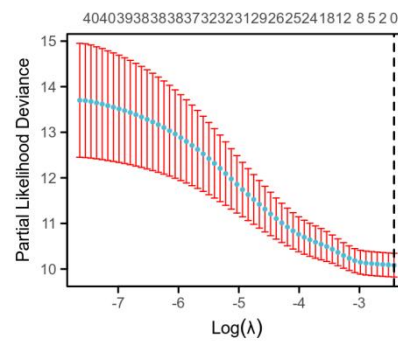
误差线

颜色
 粗细 0.50pt
 不透明度 1

- 颜色：可以修改图中误差线(竖线)的颜色，如下：

误差线

颜色
 粗细 0.50pt
 不透明度 1



- 粗细：可以修改图中误差线的线条粗细
- 不透明度：可以修改图中误差线的不透明度，1 表示完全不透明，0 表示完全透明

标题文本

标题	
大标题	大标题内容
x轴标题	x轴标题内容
y轴标题	y轴标题内容

- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

补充：在要换行的中间插入\n。如果需要上标，可以用两个英文输入法下的大括号括住，比如 {{2}}；如果需要下标，可以用两个英文输入法下的中括号括住，比如 [[2]]

风格



风格

边框 ☒

网格 ☐

文字大小 7pt

- 外框：是否添加外框，默认添加
- 网格：是否添加网格
- 文字大小：控制整体文字大小，默认为 7pt



图片



图片

宽度 (cm) 6

高度 (cm) 5

字体 Arial

- 宽度：图片横向长度，单位为 cm
- 高度：图片纵向长度，单位为 cm
- 字体：可以选择图片中文字的字体

结果说明

主要结果

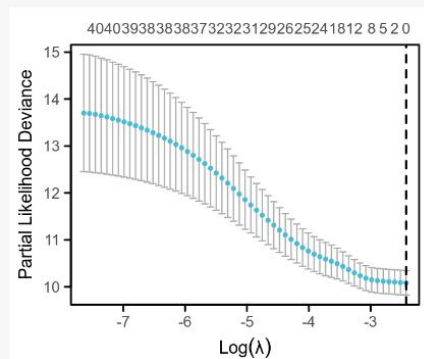
预后lasso系数筛选-云

预后lasso系数筛选: 可视化lasso系数筛选过程中各个lambda值 (惩罚值) 对应的统计量(似然偏差值或C指数)

作用: 预后lasso常用于构建预后模型或者筛选变量

交叉验证: 十折交叉验证 || 种子号: 2022

预后类型: OS[Overall Survival]



预后lasso系数筛选.pdf

预后lasso系数筛选.tiff

预后lasso系数筛选.pptx

变量-Lasso系数.xlsx

Lasso-RiskScore.xlsx

对于主要结果的解释:

- 变量-系数的 excel 表, 文件内一共有两个 sheet, 其中一个是 lambda.min 对应的变量和系数情况, 另外一个为 lambda.1se 对应的变量和系数的情况
 - 一般 lasso 是看非 0 系数的, 系数为 0 的变量为剔除的变量。
- 包含有预后资料、系数非 0 的变量以及对应的 RiskScore 的 excel 表, 分别 2 个表, 分别对应 lambda.min 和 lambda.1se 对应的情况。可用于: 临床意义-预后-风险因子图的绘制 (只需要修改 RiskScore 的列名); 基础绘图-时间依赖 ROC 曲线 (提取时间资料和 RiskScore 列) 基础绘图-生存曲线 (提取时间资料和 RiskScore 列)

补充结果

Lasso-交叉验证

交叉验证：十折交叉验证

种子号：2022

评价指标(统计量/图中y值)：deviance

	lambda值	Index	统计量	标准误(SE)	系数非0的个数
lambda.min	0.088315	1	10.084	0.2598	0
lambda.1se	0.088315	1	10.084	0.2598	0

说明：

· lambda.min表示：统计量(似然偏差值最小或C指数最大)对应的lambda

· lambda.1se表示：统计量(似然偏差值最小或C指数最大)且在1倍标准误以内对应的lambda

· Index表示：lambda.min与lambda.1se在所有的lambda值中的位置

补充：

· lambda.min与lambda.1se均可作为cutoff，但lambda.min相对严格，lambda.1se对应的变量个数更少，模型相对更简洁

· 模型的变量(表格中的系数非0的个数列)尽量控制在10个左右

· lasso可作为筛选变量的方法，如果筛选出来的变量仍很多，可以对这些筛选出来的变量进一步多因素Cox回归，构建Cox模型

这里提供 lasso-交叉验证表格：可以查看种子号为 2022、指标为 deviance 且使用十折交叉验证方法进行 lasso 系数筛选时的数据信息

- lambda.min 代表：统计量(似然偏差值最小或 C 指数最大)对应的 lambda
- lambda.1se 代表：统计量(似然偏差值最小或 C 指数最大)且在 1 倍标准误以内对应的 lambda
- Index 代表：lambda.min 与 lambda.1se 在所有的 lambda 值中的位置
- lambda.min 与 lambda.1se 均可作为 cutoff，但 lambda.min 相对严格
- lambda.1se 对应的变量个数更少，模型相对更简洁
- 模型的变量(表格中的系数非 0 的个数列)尽量控制在 10 个左右
- lasso 可作为筛选变量的方法，如果筛选出来的变量仍很多，可以对这些筛选出来的变量进一步多因素 Cox 回归，构建 Cox 模型

方法学

统计分析和可视化均在 R 4.2.1 版本中进行

涉及的 R 包：survival 包（用于构建模型）、glmnet（用于分析）

处理过程：

- (1) 使用 survival 包对云端数据进行模型构建,使用 glmnet 包进行分析得到 lambda 值、最大似然数或 C 指数等
- (2) 对数据进行可视化



如何引用

生信工具分析和可视化用的是 R 语言，可以直接写自己用 R 来进行分析和可视化即可，可以无需引用仙桃，如果想要引用仙桃，可以在致谢部分 (Acknowledge) 致谢仙桃学术 (www.xiantao love)。

方法学部分可以参考对应说明文本中的内容以及一些文献中的描述。



常见问题

1. 图里面看不到两根竖（虚）线，只有 1 根？

答：

这个情况说明 lasso 筛选得到的 cutoff 重合了，如果两根竖线都在图的最右侧，非 0 系数的变量个数为 0 或者 1（上方 x 轴），说明 lasso 无法筛选出来变量或无法构建模型。

2. lambda.min 和 lambda.1se 对应的系数非 0 的变量个数为 0，如何才能让结果能好？

答：

由于 lasso 系数筛选过程中涉及到交叉验证和样本抽样的过程，所以不同的种子号是可能会对应不同的一个情况，如果是想要结果更加“好看”一些，可以手动修改种子号。

3. 为什么图上方非 0 系数变量的个数与数据中的变量个数对应不上？为什么看不到所有的数字，只是一小部分？

答：

①图上方的这些数字对应的值是说：不同 lambda 值计算得到模型中所有变量系数不为 0 的变量的个数，而不是所有的变量（要是数值与变量个数对应不上，则

是因为缺少的那些变量间不存在相关关系(系数为 0)被筛选掉了，或者变量在数据处理过程中就被筛选掉了)

②由于可视化结果是 ggplot2 格式，故不能展示全部的数值

