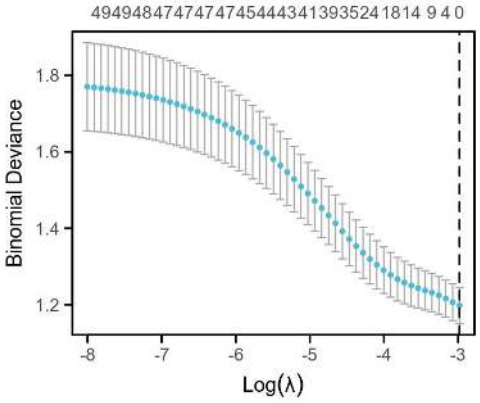
临床意义 – 诊断 Lasso 系数筛选



网址：https://www.xiantao.love

更新时间：2023.05.29

目录

基本概念

- Lasso 回归：在线性回归的基础上，通过增加惩罚项（lambda × 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 × 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 在诊断中的应用。诊断 Lasso 常常出现在构建诊断模型或者筛选变量上，最常出现两种图，一种是 系数(lambda)筛选的图，另外一种是变量轨迹图。Lasso 的 lambda 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

将诊断 Lasso 系数筛选过程中各个 lambda 值（惩罚项）对应的统计量(似然偏差值或分类错误率)进行可视化，以构建诊断模型或者筛选变量。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 直接构建诊断模型或者筛选变量。

分析流程

上传数据 ⟹ 数据处理(清洗) ⟹ lasso 诊断分析 ⟹ lasso 系数筛选可视化

- 数据格式：xlsx ／ csv ／ txt 文件格式：
  - 第 1 列数据作为结局变量(事件发生情况)，可以是数值类型也可以是分类类型数据，需要是二分类类型，可以用（0 和 1，0 表示未发生事件，1 表示发生了事件），默认会把先出现的组作为参考组。注：第 1 列不能都是删失

	A	B	C	D	E	F	G	H	I
1	event	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8
2	1	-0.022280617	-2.770435561	-0.467054253	0.659710026	-0.515117091	0.026623033	0.964868218	-1.064910322
3	1	-1.183217086	-0.316118003	0.60603766	-0.303598493	0.148727871	-1.21013808	1.213376601	-0.195115978
4	0	-0.622974392	1.845886201	1.517668885	-0.839217309	-0.273699539	-1.914794488	0.986931949	-0.185920375
5	1	-0.961432206	-0.136129034	0.707027039	-2.240777738	-0.115896498	-1.678393183	0.581343256	1.248839562
6	1	-2.009057914	0.754472479	-1.360111214	0.743456609	1.242012981	0.373015672	0.65579689	-0.654581201
7	0	0.79356585	-0.236620914	-0.501233774	0.938055954	-1.219659013	-1.625081979	0.328081467	1.04612915
8	1	-0.291946728	-0.194660575	0.208889948	-0.744460884	-1.593752647	-0.118096617	-1.29451323	1.312739415
9	1	0.709801941	-0.255890999	1.345437412	-1.040223718	-0.040457512	0.702152223	-0.592764399	2.245095696
10	1	0.257086806	0.372722112	-0.013511554	-1.046213702	0.86385945	0.767475738	0.847607122	-1.561368255
11	1	2.504925108	1.371715847	-0.114156296	-1.722567123	-0.052316113	-1.274151487	1.361514751	-0.678908796
12	1	0.44265243	-0.456238875	-0.015031461	1.584885619	0.05882662	-1.299684282	-1.501592696	-1.009258186

  - 第 2 列开始直至后面每一列都代表一个样本/变量/分子，必须是数值类型数据

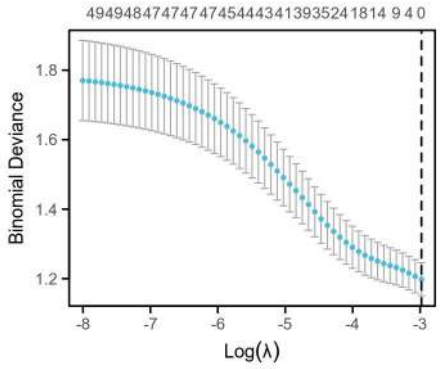
- 数据处理：分别对第 1 列（事件）、第 2 列开始后的所有变量进行清洗（去除掉数据中的非数值或者不符合条件的数据）

- Lasso 诊断分析：
  - 构建 lasso 诊断模型
  - 计算模型的 lambda 值
  - 通过 lambda 值计算变量的系数值

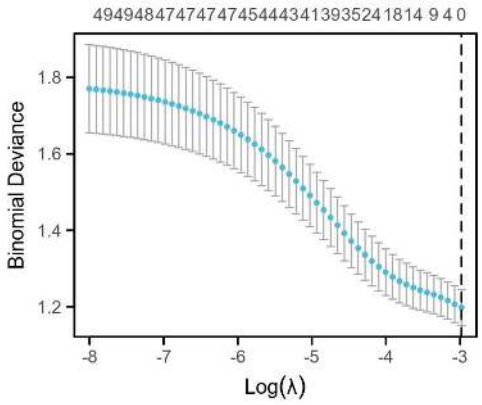
- 筛选掉 lambda 值对应系数为 0 的变量(系数为 0 表示变量之间不存在相关关系，在诊断模型中没有实质上的意义)

➢ Lasso 系数筛选可视化

- Lasso 诊断分析得到的 lambda 值取对数，对应到 lasso 系数筛选可视化结果的横坐标值
- Lasso 诊断分析得到的不同指标下的似然偏差值（deviance）（默认）或分类错误率（class）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 lambda 值取对数（log(λ)）
- 上方 x 轴的数字：表示每个 lambda 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 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)
  - 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 (−2)倍的对数似然函数值（deviance）（默认）或 模型的分类错误率（class）
- 每个点：表示数据在进行交叉验证过程中，每个 Lambda 对应的 (−2)倍的对数似然函数值（默认）或 模型的分类错误率 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 lambda 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 lambda 值（lambda.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 lambda 值（lambda.1se）

- 不管选择的指标为 (-2)倍的对数似然函数值 还是 模型的分类错误率，y 值越小对应的模型越好
- 当 lambda.min 和 lambda.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;lambda.min 可能对模型过于严格，lambda.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 lambda.min，如果 lambda 对应的变量较多，也会用 lambda.1se

数据格式

	A	B	C	D	E	F	G	H	I
1	event	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6	Gene 7	Gene 8
2	1	-0.022280617	-2.770435561	-0.467054253	0.659710026	-0.515117091	0.026623033	0.964868218	-1.064910322
3	1	-1.183217086	-0.316118003	0.60603766	-0.303598493	0.148727871	-1.21013808	1.213376601	-0.195115978
4	0	-0.622974392	1.845886201	1.517668885	-0.839217309	-0.273699539	-1.914794488	0.986931949	-0.185920375
5	1	-0.961432206	-0.136129034	0.707027039	-2.240777738	-0.115896498	-1.678393183	0.581343256	1.248839562
6	1	-2.009057914	0.754472479	-1.360111214	0.743456609	1.242012981	0.373015672	0.65579689	-0.654581201
7	0	0.79356585	-0.236620914	-0.501233774	0.938055954	-1.219659013	-1.625081979	0.328081467	1.04612915
8	1	-0.291946728	-0.194660575	0.208889948	-0.744460884	-1.593752647	-0.118096617	-1.29451323	1.312739415
9	1	0.709801941	-0.255890999	1.345437412	-1.040223718	-0.040457512	0.702152223	-0.592764399	2.245095696
10	1	0.257086806	0.372722112	-0.013511554	-1.046213702	0.86385945	0.767475738	0.847607122	-1.561368255
11	1	2.504925108	1.371715847	-0.114156296	-1.722567123	-0.052316113	-1.274151487	1.361514751	-0.678908796
12	1	0.44265243	-0.456238875	-0.015031461	1.584885619	0.05882662	-1.299684282	-1.501592696	-1.009258186

数据要求：

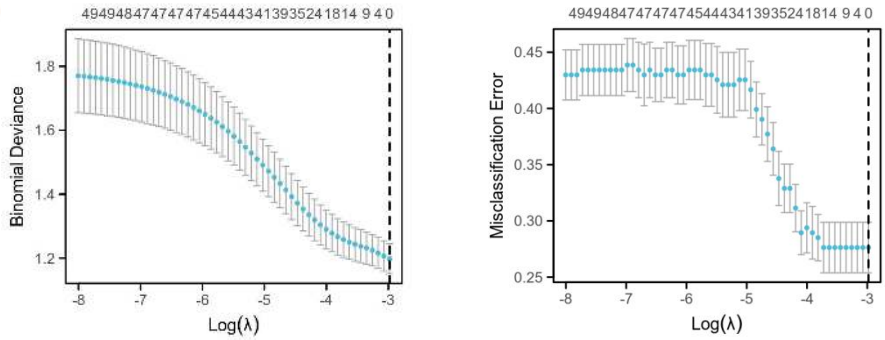
- 列数：至少需要 3 列以上的数据，最多 300 列（299 个变量）的数据，行数：至少需要 20 个以上的样本(20 行)，暂时支持最多 3000 个以上的样本
  - 第 1 列表示事件发生的情况（或结局），二分类类型
    - 可以是数值类型也可以是分类类型数据
    - 不能含有无法识别的特殊字符或者非法字符
  - 第 2 列及以后每一列数据都需要是数值类型
    - 不能含有非数值类型数据，或者混合数值与非数值类型数据
- 列名(样本名)不能重复

参数说明

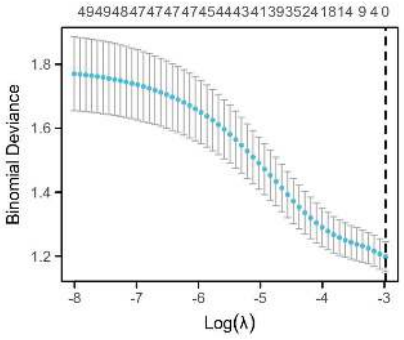
（说明：标注了颜色的为常用参数。）

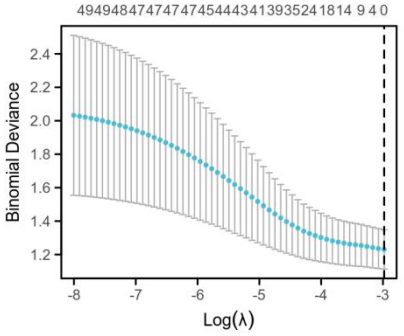
方法

- 筛选指标：可以选择 lasso 系数筛选的指标：
  - deviance（(-2)倍的对数似然函数值）（默认）
  - class （模型的分类错误率），如下：

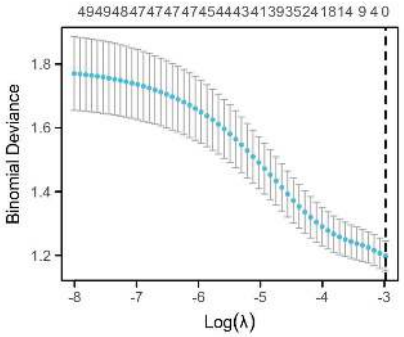


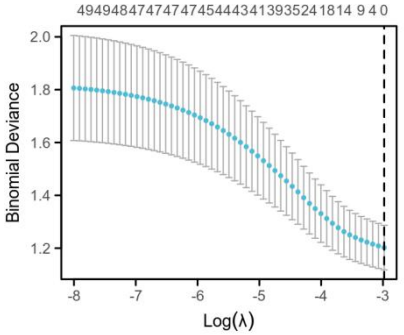
- 交叉验证（交叉验证(方法)的倍数）：可选三折、五折、七折、十折交叉验证。例如十折交叉验证，就是把数据分成 10 份，轮流把 9 份数据作为训练集训练模型，另外 1 份作为验证集验证模型。默认是选择十折交叉验证，如下：





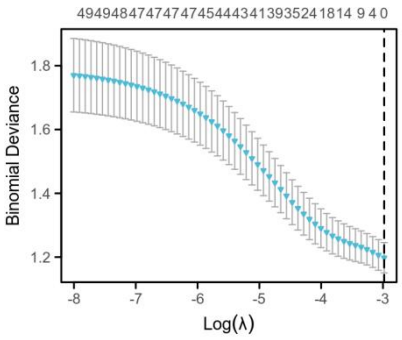
- 种子号：可填入其他的数字，默认为 2022。由于在进行交叉验证的过程中会涉及到对数据的抽样和分训练集和验证集，故不同的种子号对应的结果都会有不同，但是只要是同一份数据同一个种子号，对应的结果都是一样的，如下：





点

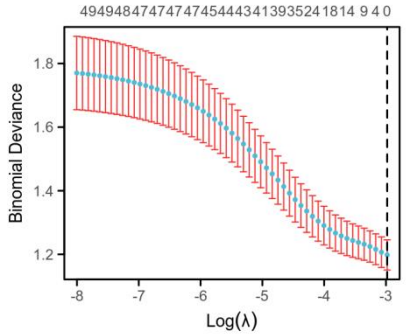
- 描边色：可以修改图中点的描边色
- 填充色：可以修改图中点的填充色
- 样式：可以选择图中点样式类型，可选择圆形、正方形、菱形、三角形、倒三角，如下：



- 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

- 颜色：可以修改图中误差线(竖线)的颜色，如下：



- 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

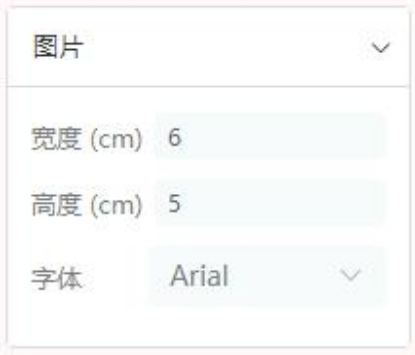
- 大标题：大标题文本
- x 轴标题：x 轴标题文本
- y 轴标题：y 轴标题文本

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

- 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

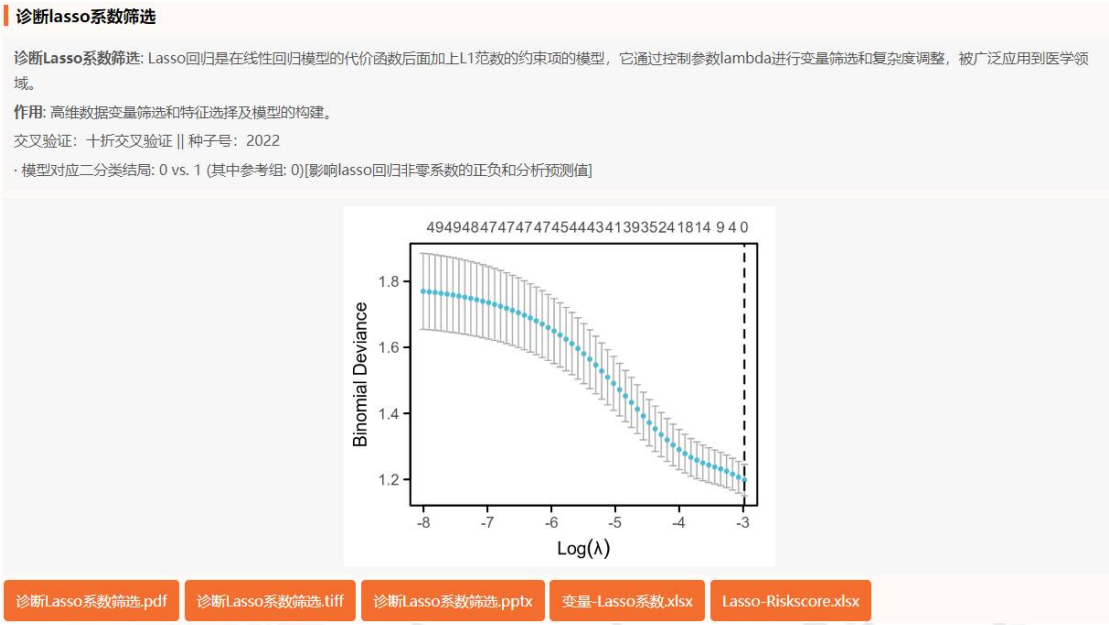
图片



- 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



- 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，另外一个是 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。
- 包含有诊断资料、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

补充结果

Lasso-交叉验证

交叉验证(十折交叉验证折方法):

种子号：2022

	lambda值	Index	统计量	标准误(SE)	系数非0的个数
lambda.min	0.05079	1	1.1982	0.047501	0
lambda.1se	0.05079	1	1.1982	0.047501	0

说明:

· lambda.min表示平均误差（目标参量均值）最小时对应的lambda。

· lambda.1se表示平均误差（目标参量均值）在1个标准差以内的最大的lambda。

· index表示lambda.min和lambda.1se在所有lambda值中所处的位置。

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Logistic回归，构建Logistic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，可筛选变量进行后面 logical 回归分析

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：glmnet（用于分析可视化）

处理过程：

(1) 使用 glmnet 包对清洗过后的数据进行分析得到变量 lambda 值、似然数值或分类错误率等

(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会可能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