R生存分析|关心的变量KM曲线不显著，还有救吗？

发布于2021-09-09 11:06:48 阅读 497

如果想查看某些因素，如年龄，性别，分期，肿瘤数目，大小，实验室指标 或者 通过生信手（tao）段（lu）构建的模型和评分是否对预后有影响时候，经常会把连续变量变为分类变量，然后绘制KM曲线或者列线图等。

这时候会有一些常用的方法：

（1）实验室指标：根据正常范围进行分类

（2）临床指标：根据临床意义进行分类

（3）生信模型评分：根据中位数，平均值等进行分类

（4）生信模型评分：根据统计上的最优cutoff来分类

本次主要介绍基于统计上的最优cutoff分类的方法，并与常见的中位数进行简单的比较。

一 载入数据，R包

为了复现方便，使用内置myeloma数据集

```
#载入所需的R包
library("survival")
library("survminer")
#查看myeloma数据集
data(myeloma)
head(myeloma)
```

```
> head(myeloma)
     molecular_group chr1q21_status treatment event  time   CCND1 CRIM1 DEPDC1    IRF4   TP53   WHSC1
GSM50986      Cyclin D-1       3 copies       TT2     0 69.24  9908.4 420.9  523.5 16156.5   10.0   261.9
GSM50988      Cyclin D-2       2 copies       TT2     0 66.43 16698.8  52.0   21.1 16946.2 1056.9   363.8
GSM50989           MMSET       2 copies       TT2     0 66.50   294.5 617.9  192.9  8903.9 1762.8 10042.9
GSM50990           MMSET       3 copies       TT2     1 42.67   241.9  11.9  184.7 11894.7  946.8  4931.0
GSM50991             MAF           <NA>       TT2     0 65.00   472.6  38.8  212.0  7563.1  361.4   165.0
GSM50992    Hyperdiploid       2 copies       TT2     0 65.20   664.1  16.9  341.6 16023.4 2096.3   569.2
>
```

二 KM-中位数分类

以TP53基因为例，按照常用的中位数分为表达量高，低两组

```
#按照中位数进行分类
myeloma <- myeloma %>%
  mutate(TP53_cat = ifelse(TP53 > median(TP53) ,"High" , "Low"))

head(myeloma)
```

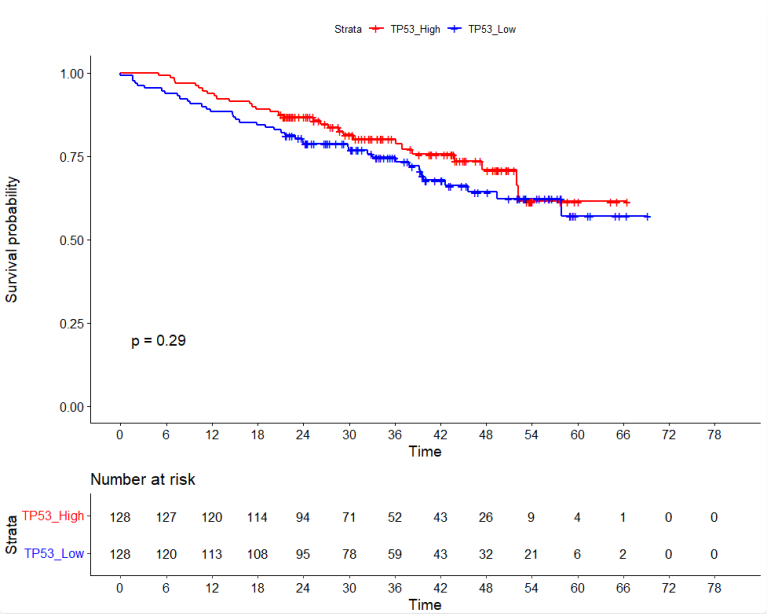
构建模型，并绘制KM曲线

```
#构建模型
fit <- survfit(Surv(time, event) ~ TP53_cat, data = myeloma)

#绘制生存曲线并显示P值
ggsurvplot(fit,
           data = myeloma,
           palette=c("red", "blue"), #自定义颜色
           legend.labs=c("TP53_High","TP53_Low"), #自定义标签
```


目录

```
            risk.table = TRUE,
            break.x.by = 6, #横坐标刻度间隔
            pval = T) #是否显示P值
```



如图显示P值不显著，这时候可以试一下最优cutoff。

更多调整可参考R|生存分析 - KM曲线，必须拥有姓名和颜值

三 KM-最优cutoff分类

3.1 计算最优cutoff

使用 `surv_cutpoint` 函数找到最优cutoff

```
res.cut <- surv_cutpoint(myeloma,
                         time = "time",
                         event = "event",
                         variables = c("TP53", "WHSC1")) #可以
summary(res.cut)#查看最佳cutoff
      cutpoint statisticTP53     748.3  2.928906WHSC1    3205.6
```

可以看到TP53 和 WHSC1 基因统计得到的最优cutoff。

3.2 根据最优cutoff分类

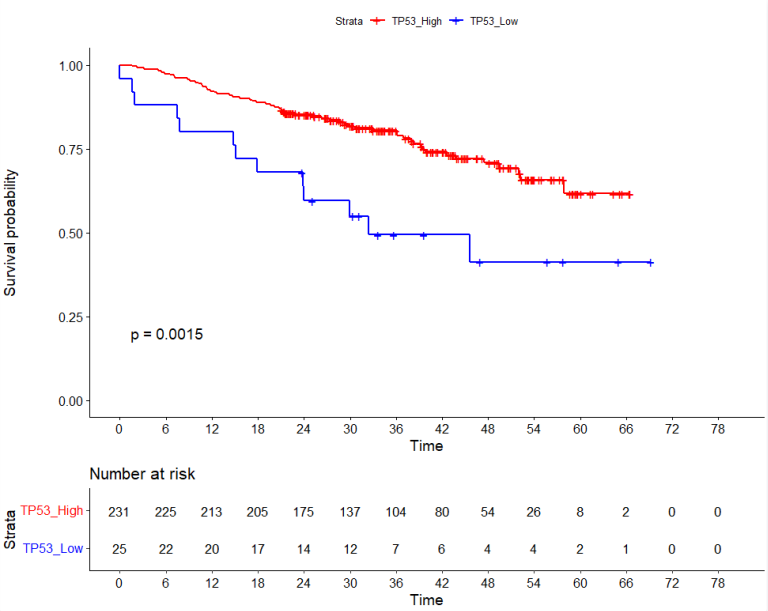
A：根据得到的最优cutoff 自行分类

```
myeloma <- myeloma %>%
  mutate(TP53_cutoff = ifelse(TP53 > 748.3 ,"High" , "Low"))

head(myeloma)
```

构建模型，并绘制KM曲线

```
#构建模型
fit <- survfit(Surv(time, event) ~ TP53_cutoff, data = myeloma)
#绘制生存曲线
ggsurvplot(fit,
           data = myeloma,
           palette=c("red", "blue"), #自定义颜色
           legend.labs=c("TP53_High","TP53_Low"), #自定义标签
           risk.table = TRUE,
           break.x.by = 6,##横坐标间隔
           pval = T) #是否展示P值
```



可以看到P值 < 0.05了，变化还是很明显的。

B：根据 `surv_categorize` 函数获取重新构建的矩阵

此处推荐这种方法，能比较简单的获取重新构建的矩阵

```
##重新构建的矩阵
res.cat <- surv_categorize(res.cut)
head(res.cat)
          time event TP53 WHSC1GSM50986 69.24     0  low   low(
```

构建模型，并绘制KM曲线

```
fit <- survfit(Surv(time, event) ~TP53, data = res.cat)
#绘制生存曲线
ggsurvplot(fit,
           data = res.cat,
           palette=c("red", "blue"),
           legend.labs=c("TP53_High","TP53_Low"), #标签
           risk.table = TRUE,
           break.x.by = 6, ##横坐标间隔
           pval = T)
```

结果和自行根据最优cutoff，使用ifelse进行分类得到的结果一致，此处不展示了。

参考资料：https://www.rdocumentation.org/packages/survminer/versions/0.4.9

https://www.rdocumentation.org/packages/survminer/versions/0.4.9/topics/surv_cutpoint

1 条评论 最新 高赞

用户9396076 2022-01-18

这样子就是上帝视角了，因为只要你想，就能做出显著的结果来

0 回复

相关文章

R|生存分析（1）

生存分析：研究各个因素与生存时间有无关系以及关联程度大小。可拓展到疾病复发时间，机器的故障时间等。 起始事件：反应研究对象开始生存过程的起始特征事件。 …

西游东行

生存分析|知道这些又没有坏处

生存分析：研究各个因素与生存时间有无关系以及关联程度大小。可拓展到疾病复发时间，机器的故障时间等。

西游东行

受限平均生存时间（Restricted mean surviva…

前些天我的学徒写了教程：人人都可以学会生存分析（学徒数据挖掘） 吸引到了读者：武汉大学金文意，他希望可以…

生信技能树

这篇3+分教你筛选拿出来几个基因应该如何分析

Construction of a novel gene-based model for prognosis prediction of clear cell ...

科研菌

R基于TCGA数据画生存曲线

常见如对于同一种癌症类型使用放疗的患者与使用化疗的患者之间的生存是否存在显著差异，从而判断使用哪种治疗方法更有利于患者的生存。

生信交流平台

纯生信文章补几张免疫组化真的很重要!

今天和大家分享的是2020年发表在Journal of cancer（IF：3.565)上的一篇文章，“Genome-wide Analysis of the ...

科研菌

TCGA的28篇教程- 对TCGA数据库的任意癌…

生存分析一般来说是针对RNA表达数据，可以说mRNA-seq

的转录组数据，也可以说miRNA-seq数据，或者基因表达…

生信技能树

KM生存曲线经logRNA检验后也可以计算HR值

可以很明显看到，根据基因表达量把病人分成高表达组合低表达组，经过log rank 检验，可以看到两组病人的生存是…

生信技能树

最权威的生存分析神器，你值得拥有！

Kaplan-meier Plotter数据库基于GEO、EGA以及TCGA等公共数据库的基因芯片和RNA-seq数据构建而成。

作图丫

生存分析凭什么不需要矫正P值

虽然生存分析如此重要而且如此常见，但是仍然有一些未解之谜，不同数据库来源，病人的不同时期的记录信息，以及不同的阈值分组，拿到的结果居然是可以不一样的！…

生信技能树

多元统计分析 期末总结

Logistic 模型 常用 极大似然估计法 估计参数，用 Newton - Raphson 迭代求解

yiyun

纯生信免疫微环境末班车

今天和大家分享的是2020年2月发表在Aging（IF：4.831)上的一篇文章，“Profiles of immune cell infiltration a...

科研菌

二十年前做科研你只需要检测一些基因在一…

检测的基因是 CUL-5： 以及 other CUL family members (CUL-1, -2, -3, -4A, and -4B)。实际上哺乳动...

生信技能树

精心整理（含图PLUS版）|R语言生信分析，…

为了能更方便的查看，检索，对文章进行了精心的整理（PLUS）。建议收藏，各取所需，当前没用也许以后就用…

西游东行

Oh my god！不做实验也能发3分SCI！

大家好，本期给大家推荐的文献是Differentially Expressed lncRNAs in Gastric Cancer Patients: A Po...

百味科研芝士

生存分析——KM生存曲线、hazard比例、P…

与完全数据相反，如果在研究结束的时候，研究对象发生了研究之外的其他事件或生存结局，无法明确的观察记录到…

悟乙己

欧洲裔和非裔美国乳腺癌患者差异可以TCGA数据库验证

差异分析相信大家都不陌生了，基本上看我六年前的表达芯片的公共数据库挖掘系列推文即可；

生信技能树

更多文章