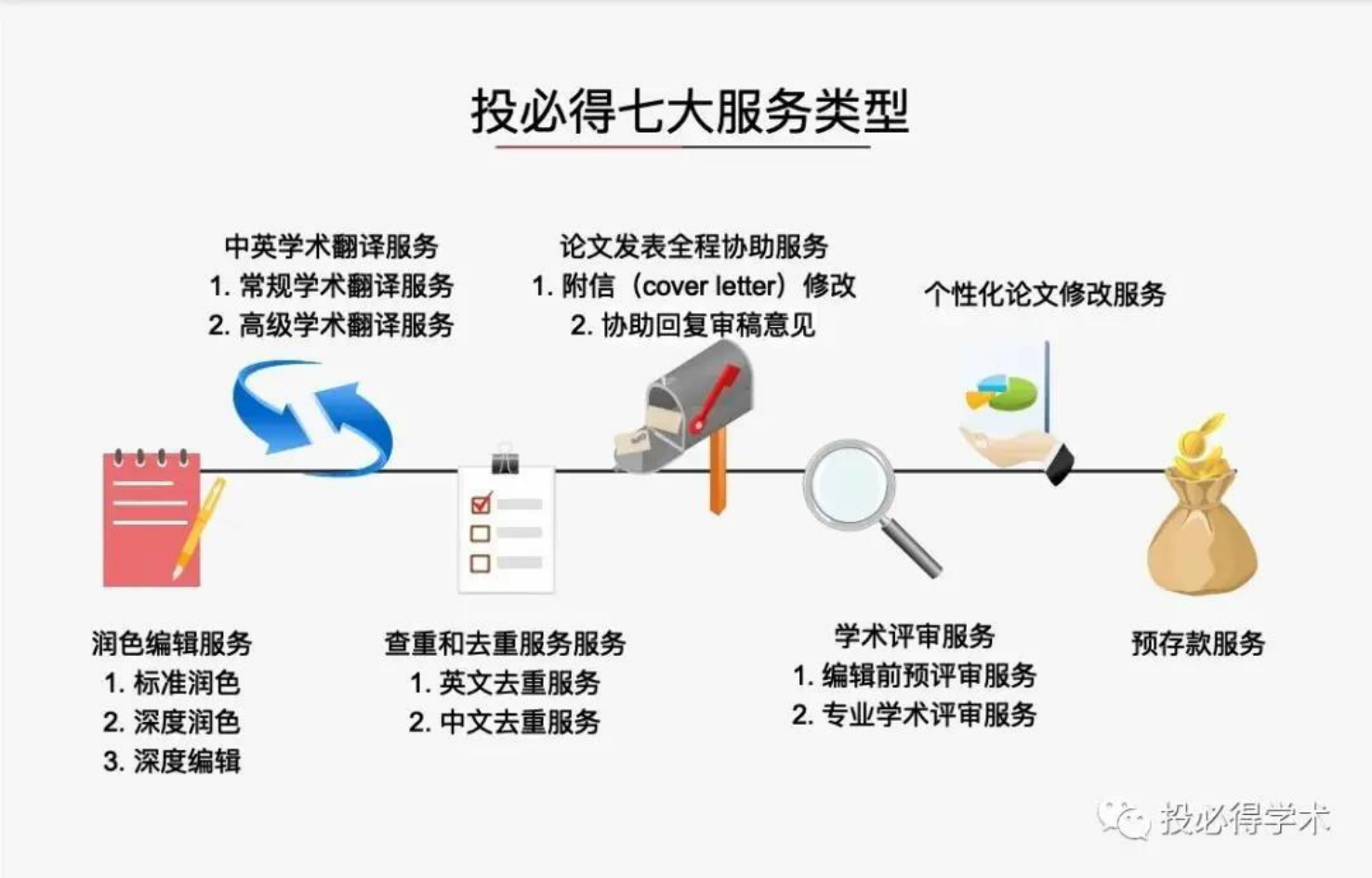
知乎 首发于 投必得统计分析大讲堂



第二十八讲 R语言-Cox比例风险模型1

投必得论...
已认证的官方帐号

在第二十五到二十七讲中，我们介绍了生存分析的基本概念，KM生存曲线及绘图，以及比较多组生存曲线是否存在差异。KM生存曲线和Log-rank检验是单变量多分析方法，只能通过分层的方式，考虑一个水平（因子 / 因素）的作用，而忽略其他因素多影响。但是当数据存在多个因素需要考虑，或因素不是分类变量，而是连续型变量时，KM曲线和Log-rank检验就无法应对了。这时，该怎么办呢？Cox比例风险模型应运而生。

Cox比例风险模型(proportional hazards model，简称Cox模型)，是由英国统计学家D.R.Cox(1972) 年提出的一种半参数回归模型。该模型以生存结局和生存时间为应变量，可同时分析众多因素对生存期的影响，能分析带有截尾生存时间的资料，且不要求估计资料的生存分布类型。

知乎 首发于 投必得统计分析大讲堂

在实际临床研究中，影响事件发生的因素往往不止一个，它是多个因素综合作用的结果。Cox比例风险模型，它既适用于连续型变量也适用于类别变量。此外，Cox回归模型扩展了生存分析方法，以同时评估几种风险因素对生存时间的影响，并且给每一个因素提供了统计量的大小以反映因素对事件发生的影响大小。

该模型的目的是同时评估多种因素对生存的影响。换句话说，它使我们可以探索特定因素是如何影响特定时间点的特定事件（例如，感染，死亡）发生率低。该比率通常称为风险率。预测变量（或因子）在生存分析中通常称为协变量（covariates）。

Cox模型由h(t)表示的危险函数表示。简而言之，风险函数可以解释为在时间t死亡的风险。可以估计如下：

$$h(t)=h0(t)\times\exp(b1x1+b2x2+...+bpxp)$$

其中，

- t表示生存时间
- h(t)是由一组数量为p的协变量 (x1,x2,...,xpx1,x2,...,xp)
- 系数 (b1,b2,...,bpb1,b2,...,bp)可以衡量协变量的影响（即影响大小）。
- h0被称为基准风险，即当所有xi等于零（exp（0）数值等于1）时的事件风险。H(t)中的“t”表示该风险是随时间变化的。

Cox模型可以写成多元线性模型的形式:

$$Ln(h(t))= (b1x1+b2x2+...+bpxp) + ln(h0(t))$$

Exp(bi)称为风险比（Hazard ratios，HR）。值bi大于零，即风险比大于1，表示随着协变量ith的增加，事件风险也增加，于是生存期减少。

换句话说，风险比大于1表示协变量与事件概率正相关，与生存时间负相关 。

综上所述，

- HR = 1：无效
- HR < 1：减少风险
- HR > 1：增加风险

- 风险比 > 1（即：b > 0）的协变量称为不良预后因素
- 风险比 < 1（即：b < 0）的协变量被称为良好预后因素

Cox模型的一个关键假设条件是观察组（或患者）的生存曲线应成比例，并且不能交叉。

比如，考虑两个x值不同的患者k和k'。其相应的危险函数可以简单地写成如下

- 对患者k的风险函数：

- 对患者k'的风险函数：

- 这两名患者的危险比

应该要与时间t无关。

因此，Cox模型是比例风险模型：即在任何组中，事件的风险都是在协变量的影响下成比例变化的。所以，在Cox比例风险模型中，各组的生存曲线也应成比例，并且不能交叉。

换句话说，如果一个人在某个初始时间点的死亡风险是另一个人的两倍，那么在以后的所有时间，死亡风险仍然要是另一个人的两倍。

我们将在第三十讲中，为大家介绍，如何评估Cox比例风险模型的假设条件

知乎 首发于 投必得统计分析大讲堂

2.1安装并加载所需的R包

我们将使用两个R包：

- survival用于计算生存分析
- survminer用于总结和可视化生存分析结果
- 安装软件包

```
install.packages(c("survival", "survminer"))
```

- 加载软件包

```
library("survival")
library("survminer")
```

2.2 用于计算Cox模型的R函数：coxph（）

生存包中的函数coxph（）可用于计算R中的Cox比例风险回归模型。

简化格式如下：

```
coxph(formula, data, method)
```

•formula：是将生存对象作为响应变量的线性模型。 使用功能Surv（）创建生存对象：Surv（time，event）。

•data：包含变量的数据框

•method：用于指定如何处理领带。 默认值为“efron”。 其他选项是“breslow”和“exact”。 通常，默认的“efron”优于曾经流行的“breslow”方法。 “exact”方法的计算量会很大。

2.3 示例数据集

```
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
4    5  210      2  57   1       1       90        60     1150      11
5    1  883      2  60   1       0      100        90       NA       0
6   12 1022      1  74   1       1       50        80      513       0
```

- inst：机构代码
- time：以天为单位的生存时间
- status：删失状态1 = 删失，2 = 出现失效事件
- age：岁
- sex：性别，男= 1女= 2
- ph.ecog：ECOG评分（0 =好，5 =死）
- ph.karno：医师进行的Karnofsky评分（0 = 差，100 = 好）
- pat.karno：患者自行进行的Karnofsky评分（0 = 差，100 = 好）
- meal.cal：用餐时消耗的卡路里
- wt.loss：最近六个月的体重减轻

2.4 计算Cox模型

我们将使用以下协变量来拟合Cox回归：age，sex，ph.ecog和wt.loss。

我们首先计算所有这些变量的单变量Cox分析。然后我们将使用两个变量来拟合多变量Cox分析，以描述这些因素如何共同影响生存。

2.4.1 单变量Cox模型

单变量Cox分析可以如下计算：

```
res.cox <- coxph(Surv(time, status) ~ sex, data = lung)
res.cox
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)
       coef exp(coef) se(coef)     z      p
sex -0.531     0.588    0.167 -3.18 0.0015
```

```
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)
  n= 228, number of events= 165
       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816
Concordance= 0.579  (se = 0.022 )
Rsquare= 0.046   (max possible= 0.999 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001111
Wald test            = 10.09  on 1 df,   p=0.001491
Score (logrank) test = 10.33  on 1 df,   p=0.001312
```

Cox回归结果可以解释为：

1. **统计意义：**标记为“z”的列给出了Wald的统计量。它对应于每个回归系数与其标准误的比率（z = coef / se（coef））。Wald统计量会评估beta（β，coef）给定变量的系数在统计上是否与0有显著差异。从上面的输出中，我们可以得出结论，变量sex的系数（coef）与0具有统计学差异（P = 0.00149）。
2. **回归系数。**Cox模型结果中要注意的第二个特征是回归系数（coef）的符号。正号表示该变量值为不良风险因素，与受试者高事件发生率有关，因此预后较差。变量sex被编码为数字向量。1：男性； 2：女性。Cox模型的summary()提供了第二组相对于第一组（即女性与男性）的风险比（HR，exp（coef））。在这些数据中，性别的β系数（coef）= -0.53，表示女性比男性具有更低的死亡风险。
3. **风险比。**指数系数（exp（coef）= exp（-0.53）= 0.59），也称为风险比，给出了协变量的影响大小。例如，女性（性别= 2）可将危险降低0.59或41％。女性是良好预后的因素。
4. **风险比的置信区间。**Summary()输出还给出了风险比（exp（coef））的 95％置信区间，下限95％= 0.4237，上限95％= 0.816。
5. **模型的全局统计意义。**最后，结果中输出了模型的总体重要性检验结果，病提供了三个备选检验的p值：似然比检验（likelihood-ratio test），Wald检验（Wald test）和得分对数秩统计（score logrank statistics）。这三种方法基本等效。对于足够大的N，它们将给出相似的结果。对于较小的N，它们可能会有所不同。对于小样本量，似然比检验结果更准确，因此通常是首选。

```
 function(x) as.formula(paste('Surv(time, status)~', x)))

univ_models <- lapply( univ_formulas, function(x){coxph(x, data = lung)})
# 提取数据，并制作数据表格
univ_results <- lapply(univ_models,
 function(x){
                          x <- summary(x)
                          p.value<-signif(x$wald["pvalue"], digits=2)
                          wald.test<-signif(x$wald["test"], digits=2)
                          beta<-signif(x$coef[1], digits=2);#coeficient beta
                          HR <-signif(x$coef[2], digits=2);#exp(beta)
                          HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)
                          HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)
                          HR <- paste0(HR, " (",
                                       HR.confint.lower, "-", HR.confint.upper, ")")
                          res<-c(beta, HR, wald.test, p.value)
                          names(res)<-c("beta", "HR (95% CI for HR)", "wald.test",
                                        "p.value")
 return(res)
                          #return(exp(cbind(coef(x),confint(x))))
                         })
res <- t(as.data.frame(univ_results, check.names = FALSE))
as.data.frame(res)
           beta HR (95% CI for HR) wald.test p.value
age       0.019            1 (1-1)       4.1   0.042
sex       -0.53   0.59 (0.42-0.82)        10  0.0015
ph.karno -0.016      0.98 (0.97-1)       7.9   0.005
ph.ecog    0.48        1.6 (1.3-2)        18 2.7e-05
wt.loss  0.0013         1 (0.99-1)      0.05    0.83
```

上面的输出显示了每个变量相对于总生存率的回归beta系数，效应大小（以风险比给出）和统计显着性。通过单独的单变量Cox回归评估每个因素。

从上面的输出中，

- 性别，年龄和ph.ecog变量具有极高的统计学意义，而ph.karno的系数则不显着。
- 年龄和ph.ecog具有正β系数，而性别具有负系数。因此，年龄较大和phogog较高与事件发生

素（性别，年龄和ph.ecog）纳入多元模型。

下一讲中， 我们将介绍，在经过单变量Cox分析后选取的多个因素，如何进行多元Cox回归分析，如何解释回归分析的结果，以及如何通过作图来展示结果。

2.4.2 多元Cox回归分析

多元cox回归分析的R代码如下，其中3个因素（性别，年龄和ph.ecog）纳入多元模型：

```
res.cox <- coxph(Surv(time, status) ~ age + sex + ph.ecog, data =  lung)
summary(res.cox)
Call:
coxph(formula = Surv(time, status) ~ age + sex + ph.ecog, data = lung)
  n= 227, number of events= 164
   (1 observation deleted due to missingness)
             coef exp(coef)  se(coef)      z Pr(>|z|)
age      0.011067  1.011128  0.009267  1.194 0.232416
sex     -0.552612  0.575445  0.167739 -3.294 0.000986 ***
ph.ecog  0.463728  1.589991  0.113577  4.083 4.45e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
        exp(coef) exp(-coef) lower .95 upper .95
age        1.0111     0.9890    0.9929    1.0297
sex        0.5754     1.7378    0.4142    0.7994
ph.ecog    1.5900     0.6289    1.2727    1.9864
Concordance= 0.637  (se = 0.026 )
Rsquare= 0.126   (max possible= 0.999 )
Likelihood ratio test= 30.5  on 3 df,   p=1.083e-06
Wald test            = 29.93  on 3 df,   p=1.428e-06
Score (logrank) test = 30.5  on 3 df,   p=1.083e-06
```

所有三个总体检验（似然性，Wald和得分）的p值均显着，表明该模型具有显着性意义。这些检验评估了总体beta的综合原假设（β）为0。在上面的示例中，检验统计数据非常一致，并且完全拒绝了综合原假设。即由3个因素（性别，年龄和ph.ecog）组成的模型对风险比的影响系数不为0。

在元Cox分析中，协变量性别和ph.ecog保持显着性（p <0.05）。但是，协变量年龄不显着（p = 0.23，大于0.05）

件下，女性（性别= 2）相比于男性，死亡风险低0.58或42%。我们得出的结论是，成为女性与良好的预后相关。

同样，ph.ecog的p值为4.45e-05，风险比HR = 1.59，表明ph.ecog值与死亡风险增加之间有很强的关系。如果保持其他协变量不变，则ph.ecog的值越高，存活率越低，即在其他协变量都一致的人群中，ph.ecog每增高一个单位，死亡风险增高59%。

相比之下，年龄的p值现在为p = 0.23。风险比HR = exp（coef）= 1.01，95%置信区间为0.99至1.03。由于HR的置信区间为1，因此该结果表明，在调整了ph.ecog值和患者的性别后，年龄对HR差异的贡献较小，并且不显着。例如，在其他协变量保持不变的情况下，每老一岁会引起每日死亡危险， 1%，但这并没有统计学意义，也不是重大贡献。

3. 可视化生存时间的估计分布情况

将Cox模型拟合到数据后，就可以可视化特定风险组在任何给定时间点的预测的生存率。函数survfit（）估计生存比例，默认情况下输出的为协变量的平均值。

与KM生存曲线不同的是，Cox模型拟合曲线输出的是在矫正了其他协变量因素以后的预测的生存率，而不是实际观察到的生存率情况。

```
# 绘出基线的生存函数
ggsurvplot(survfit(res.cox), color = "#2E9FDF",
           ggtheme = theme_minimal())
```

我们可能希望显示估算的生存率如何被特定协变量影响的。

考虑到这一点，我们想评估性别对估计生存率的影响。在这种情况下，我们先创建一个有两行的新数据表，每一行代表一种性别。其他协变量则固定为其平均值（如果是连续变量）或最低水平（如果它们是离散变量）。对于协变量为哑变量的，平均值为数据集中编码为1的比例。该数据表通过newdata参数传递给survfit（）：

```
# 创建新的数据表
sex_df <- with(lung,
               data.frame(sex = c(1, 2),
                          age = rep(mean(age, na.rm = TRUE), 2),
                          ph.ecog = c(1, 1)
                          )
               )
sex_df
  sex      age ph.ecog
```

```
ggsurvplot(fit, conf.int = TRUE, legend.labs=c("Sex=1", "Sex=2"),
           ggtheme = theme_minimal())
```

参考内容：sthda.com

发布于 07-27

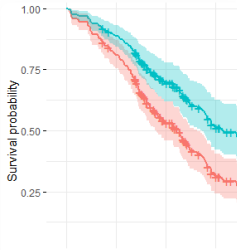
生存分析 风险模型 R（编程语言）

知乎 首发于 投必得统计分析大讲堂

推荐阅读

第二十九讲 R语言-Cox比例风险模型2

投必得论文... 发表于投必得统计...



COX比例风险模

经济小海狸

还没有评论

写下你的评论...