生存分析作业 1

罗震林

1 准备

- 包

```
rm(list=ls())   # 清楚当前环境对象
library(survival)
library(survminer)  # 生存曲线图包
library(muhaz)    # 风险函数包
```

这里共需要 survival, survminer, muhaz 三个包, 其中

1. survival 是生存分析常用包，包含创建生存对象等生存分析常用函数
2. survminer 是基于 ggplot 的画图包，包含对 survival 中一些函数返回结果进行画图的函数，比直接作图更加美观
3. muhaz 包中则含有 **kphaz()** 函数，可以用来计算 Kaplan-Meier 类型的风险函数

- 读取数据

```
actg <- read.table("ACTG175(speff2trial).txt",header = TRUE,sep=",")
```

ACTG175 数据集共有 2139 个样本，27 个变量，在本报告中，会使用到 cens, days, arm 和 age 四个变量

变量	含义
days	事件第一次发生的天数
cens	1，事件发生；0，数据有删失
age	研究时的受访者年龄
arm	根据使用药物分组

- 创建生存对象

```
obj_surv <-Surv(actg$days, actg$cens==1)       # 创建生存对象
obj_surv[1:10]
```

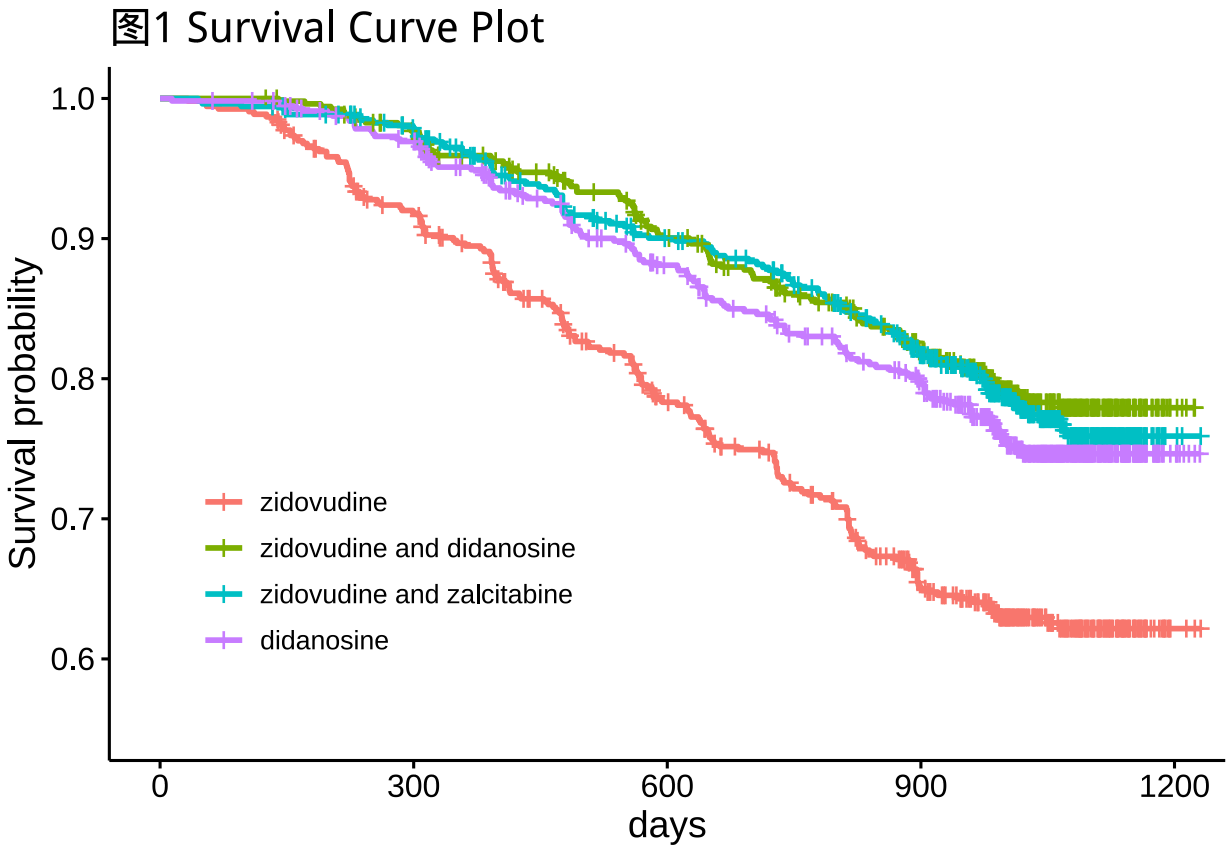
```
##  [1]  948+ 1002   961+ 1166+ 1090+ 1181+  794   957+  198   188
```

创建的生存对象 obj_surv，后面会用于做为模型中的响应变量。以前 10 个数据为例，带 + 号表示数据有删失。

2 Plot

2.1 Survival Curve Plot

```
fit_surv <- survfit(obj_surv  ~ actg$arms, ctype = 1) # 生存曲线数据
# 生存函数曲线
ggsurvplot(fit_surv,
           data = actg,
           title = " 图 1 Survival Curve Plot",
           xlab = "days",
           ylim = c(0.55,1),
           legend.title = "",      # 图例标题
           legend.labs = c("zidovudine", "zidovudine and didanosine",
                           "zidovudine and zalcitabine", "didanosine"),
           legend = c(0.25,0.3),   # 图例位置
           )
```

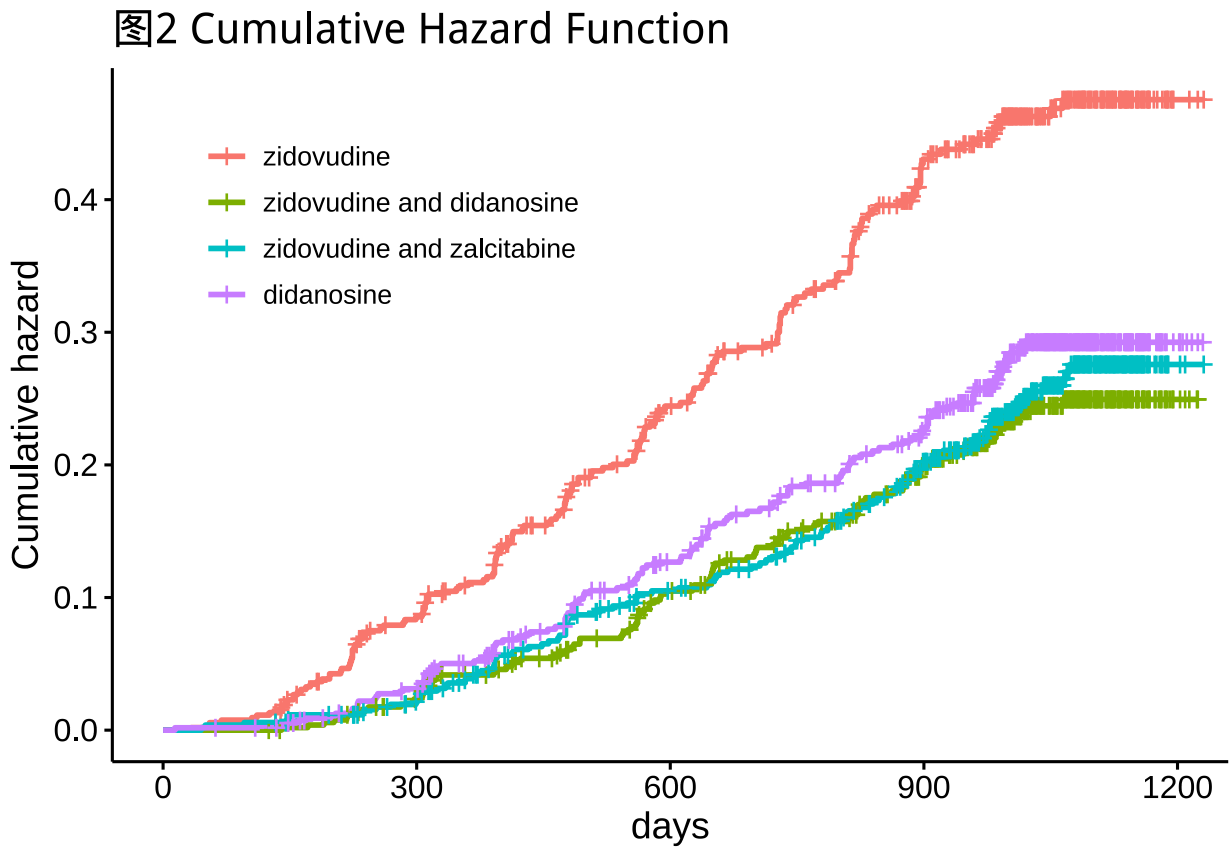


从图 1 可以看出：

1. 其它 3 组都比单独使用 zidovudine 的组生存概率高，说明其它 3 种疗法更为有效
2. 单独使用 didanosine 组的生存概率虽较单独使用 zidovudine 组高，但仍比另外两组混合治疗方法的生存概率低
3. 在前期，zidovudine+didanosine 组与 zidovudine+zalcitabine 组的生存概率十分接近，但是在后期，则是 zidovudine+didanosine 组的生存概率更高

2.2 Cumulative Hazard Function

```
ggsurvplot(fit_surv,
           data = actg,
           fun = "cumhaz",
           title = " 图 2 Cumulative Hazard Function",
           legend.title = "",
           legend.labs = c("zidovudine", "zidovudine and didanosine",
                           "zidovudine and zalcitabine", "didanosine"),
           legend=c(0.25, 0.8),
           xlab = "days") # 绘制累计风险曲线
```



这里绘制的累积风险函数采用 Nelson-Aalen 估计，由在 survfit() 中的 ctype = 1 参数来确定。

从图 2 可以看出：

1. 其它 3 组的累积风险都比单独使用 zidovudine 组低，说明其它 3 种疗法更为有效
2. 单独使用 didanosine 组的累积风险虽较单独使用 zidovudine 组低，但仍比另外两组混合治疗方法的累积风险高
3. 在前期，zidovudine+didanosine 组与 zidovudine+zalcitabine 组的累积风险十分接近，但是在后期，则是 zidovudine+didanosine 组的累积风险更低

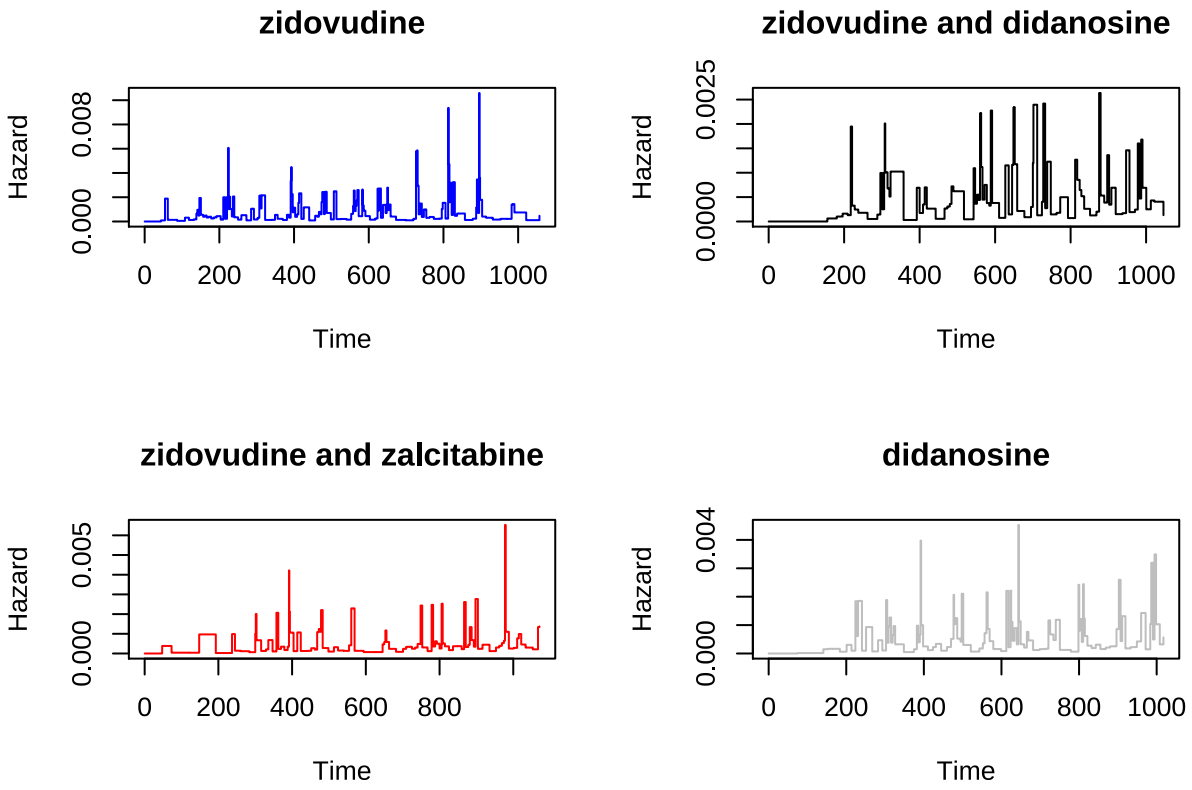
综上，图 2 和图 1 得出的结论相同，基本上可以说明 zidovudine+didanosine 组是四种方法中疗效最好的

2.3 Hazard Function

```
# 划分数据集
zio <- subset(actg,arms==0)
zio_did <- subset(actg, arms==1)
zio_zal <- subset(actg, arms==2)
did <- subset(actg, arms==3)

# 拟合
haz1 <- kphaz.fit(zio$days, zio$cens, method = "product-limit")
haz2 <- kphaz.fit(zio_did$days, zio_did$cens,method = "product-limit" )
haz3 <- kphaz.fit(zio_zal$days, zio_zal$cens, method = "product-limit")
haz4 <- kphaz.fit(did$days, did$cens, method = "product-limit")

# 作图
par(mfrow=c(2,2))
kphaz.plot(haz1, col = "blue", main = "zidovudine")
kphaz.plot(haz2, main ="zidovudine and didanosine")
kphaz.plot(haz3, main ="zidovudine and zalcitabine", col = "red")
kphaz.plot(haz4, main ="didanosine", col = "grey")
```



这里使用 muhaz 包中的函数 kphaz.fit() 和 kphaz.plot() 进行拟合和画图，采用的非参数方法是 Kaplan-Meier 估计。

可以看出：

1. zidovudine 组的风险是最高的，且明显高于其它 3 组
2. didanosine 组的风险比 zidovudine 组低，平均风险比其它两组要稍高
3. zidovudine+didanosine 组与 zidovudine+zalcitabine 组的风险十分接近，但是 zidovudine+didanosine 组的风险存在几处偏高的极端值，所以 zidovudine+didanosine 组更加不稳定

结合生存函数和累积风险函数，可以进一步判断 zidovudine+didanosine 组治疗效果是最好的，zidovudine+zalcitabine 组次之，didanosine 组也会好于 zidovudine 组。但是，这些数据并没有控制其它变量，可能会导致对某一特定人群治疗效果与结论存在差异，所以需进一步控制想研究的变量来分析

3 Log-rank Test

3.1 未分层

- 假设

我们想检验假设

$$H_0：所有生存曲线相同 \quad vs \quad H_1：至少有两组生存曲线不同$$

* 计算统计量和 p 值

```
#log-rank test
lrt <- survdiff(obj_surv  ~ actg$arms)
lrt
```

```
## Call:
## survdiff(formula = obj_surv ~ actg$arms)
##
##                N Observed Expected (O-E)^2/E (O-E)^2/V
## actg$arms=0 532      181      116    37.030     47.67
## actg$arms=1 522      103      134     6.988      9.40
## actg$arms=2 524      109      132     4.158      5.58
## actg$arms=3 561      128      139     0.933      1.27
##
##  Chisq= 49.2  on 3 degrees of freedom, p= 1e-10
```

```
#log-rank test age
actg$age_d <- 0    # 创建新列
actg[which(actg$age <= 25),]$age_d=1            # 数据分层
actg[which(actg$age > 25 & actg$age <= 55),]$age_d = 2
actg[which(actg$age > 55),]$age_d = 3
```

统计量 $X^2 = 49.2$ 在原假设下服从自由度为 3 的卡方分布，并且可以得到 $p = 1e - 10 < 0.01$

- 结论

在显著性水平大于 0.01 的情况下，都可以拒绝原假设，认为四组的生存曲线是不同的

3.2 根据 age 分层

- 数据预处理

将变量 age 的数据以 <=25，25~55，>55 的标准分为三组，分别标记为 1，2，3，储存在 stra_age 的新变量中

```
actg$stra_age <- 0    # 创建新列
actg[which(actg$age <= 25),]$stra_age=1
actg[which(actg$age > 25 & actg$age <= 55),]$stra_age = 2
actg[which(actg$age > 55),]$stra_age = 3
```

- 假设

这里想检验的假设同上：

$$H_0：所有生存曲线相同 \quad vs \quad H_1：至少有两组生存曲线不同$$

- 计算统计量和 p 值

```
lrt_st <- survdiff(obj_surv ~ arms + strata(stra_age), data = actg)
lrt_st
```

```
## Call:
## survdiff(formula = obj_surv ~ arms + strata(stra_age), data = actg)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## arms=0 532      181      115     37.29     48.01
## arms=1 522      103      134      7.08      9.53
## arms=2 524      109      133      4.23      5.69
## arms=3 561      128      139      0.89      1.22
##
##  Chisq= 49.6  on 3 degrees of freedom, p= 1e-10
```

统计量 $X^2 = 49.6$ 在原假设下服从自由度为 3 的卡方分布，并且可以得到 $p = 1e - 10 < 0.01$。

- 结论

在显著性水平大于 0.01 的情况下，都可以拒绝原假设，认为四组的生存曲线是不同的。注意这里统计量比未根据年龄分层时的统计量更大，从而计算的 p 值会更小，所以更能拒绝原假设。总而言之，分层和未分层的检验结果几乎一样，相差不大