生存分析 2

罗震林 17306071

2021 年 4 月 18 日

1

准备工作：加载相关包，载入数据以及创建生存对象；其中 nnet 包只用到其中生成哑变量的函数 class.ind()

```
rm(list=ls())
library(survival)
library(survminer)
library(nnet)
actg <- read.table("ACTG175(speff2trial).txt",header = TRUE,sep=",")
y_surv <-Surv(actg$days, actg$cens==1)
```

1.1 (a)

在这里，我们建立 Cox proportional hazards model (Model 1), 其中解释变量（风险因素）选取 arms, cd40, cd80, age, wtkg, hemo, homo, drug, karnof, race, gender and symptom. 在 Cox model 中, arms 需要作为因子变量处理，在 R 中可以通过两种方法实现，一是直接使用 factor()，二是将 arms 作为哑变量处理，实现代码如下：

```
# 方法一：直接使用factor()
model1 <- coxph(y_surv~factor(arms)+cd40+cd80+age+wtkg+hemo+homo+drugs+karnof+race+gender+
                symptom,data=actg)
summary(model1)

# 方法二：  先创建哑变量
dumv <- class.ind(actg$arms)  # 创建哑变量
colnames(dumv) <- c("arms0","arms1","arms2","arms3")
actg <- cbind(actg,dumv) # 合并到数据框中
model2 <- coxph(y_surv~arms1+arms2+arms3+cd40+cd80+age+wtkg+hemo+homo+
                drugs+karnof+race+gender+symptom,data=actg)
summary(model2)
```

在创建哑变量的方法中，以 arms=0 的组别为参照组，所以在 coxph() 中代入除 arms=0 的其它三个变量进行 cox 回归，两种方法的结果分别见图 1和图 2

```
Call:
coxph(formula = y_surv ~ factor(arms) + cd40 + cd80 + age + wtkg +
    hemo + homo + drugs + karnof + race + gender + symptom, data = actg)

  n= 2139, number of events= 521

                    coef  exp(coef)   se(coef)      z Pr(>|z|)
factor(arms)1 -7.796e-01  4.586e-01  1.242e-01 -6.279 3.42e-10 ***
factor(arms)2 -6.591e-01  5.173e-01  1.218e-01 -5.411 6.28e-08 ***
factor(arms)3 -5.672e-01  5.671e-01  1.160e-01 -4.888 1.02e-06 ***
cd40          -4.259e-03  9.957e-01  4.446e-04 -9.581  < 2e-16 ***
cd80           4.454e-04  1.000e+00  8.413e-05  5.294 1.19e-07 ***
age            8.632e-03  1.009e+00  5.222e-03  1.653 0.098314 .
wtkg           7.009e-05  1.000e+00  3.525e-03  0.020 0.984137
hemo           1.046e-01  1.110e+00  2.136e-01  0.490 0.624200
homo           2.068e-02  1.021e+00  1.581e-01  0.131 0.895922
drugs         -3.109e-01  7.328e-01  1.522e-01 -2.043 0.041006 *
karnof        -2.577e-02  9.746e-01  6.985e-03 -3.690 0.000225 ***
race          -1.155e-01  8.910e-01  1.110e-01 -1.040 0.298323
gender        -3.082e-02  9.697e-01  1.835e-01 -0.168 0.866608
symptom        3.961e-01  1.486e+00  1.035e-01  3.825 0.000131 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

图 1: 方法一结果

```
Call:
coxph(formula = y_surv ~ arms1 + arms2 + arms3 + cd40 + cd80 +
    age + wtkg + hemo + homo + drugs + karnof + race + gender +
    symptom, data = actg)

  n= 2139, number of events= 521

               coef  exp(coef)   se(coef)      z Pr(>|z|)
arms1   -7.796e-01  4.586e-01  1.242e-01 -6.279 3.42e-10 ***
arms2   -6.591e-01  5.173e-01  1.218e-01 -5.411 6.28e-08 ***
arms3   -5.672e-01  5.671e-01  1.160e-01 -4.888 1.02e-06 ***
cd40    -4.259e-03  9.957e-01  4.446e-04 -9.581  < 2e-16 ***
cd80     4.454e-04  1.000e+00  8.413e-05  5.294 1.19e-07 ***
age      8.632e-03  1.009e+00  5.222e-03  1.653 0.098314 .
wtkg     7.009e-05  1.000e+00  3.525e-03  0.020 0.984137
hemo     1.046e-01  1.110e+00  2.136e-01  0.490 0.624200
homo     2.068e-02  1.021e+00  1.581e-01  0.131 0.895922
drugs   -3.109e-01  7.328e-01  1.522e-01 -2.043 0.041006 *
karnof  -2.577e-02  9.746e-01  6.985e-03 -3.690 0.000225 ***
race    -1.155e-01  8.910e-01  1.110e-01 -1.040 0.298323
gender  -3.082e-02  9.697e-01  1.835e-01 -0.168 0.866608
symptom  3.961e-01  1.486e+00  1.035e-01  3.825 0.000131 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

图 2: 方法二结果

可以看到，两种方法得到的结果是一致的，后面都以方法一的结果来分析。其中风险因子 arms, cd40, cd80, age, drugs, karnof, symptom 的 p 值是小于 0.05，所以在显著性水平 $\alpha = 0.05$ 的情况下，这些因子在统计上是显著的。

1.2 (b)

已知接受 zidovudine 和 zalcitabine 治疗的组别其 arms=2，只接受 zidovudine 治疗的组其 arms=0（即参照组），model 1 得到的 hazard ration 可以使用 summary() 函数得到，结果承接图 1, 具体见图 3

	exp(coef)	exp(-coef)	lower .95	upper .95
factor(arms)1	0.4586	2.1806	0.3595	0.5850
factor(arms)2	0.5173	1.9330	0.4075	0.6568
factor(arms)3	0.5671	1.7633	0.4517	0.7119
cd40	0.9957	1.0043	0.9949	0.9966
cd80	1.0004	0.9996	1.0003	1.0006
age	1.0087	0.9914	0.9984	1.0190
wtkg	1.0001	0.9999	0.9932	1.0070
hemo	1.1103	0.9007	0.7305	1.6875
homo	1.0209	0.9795	0.7489	1.3918
drugs	0.7328	1.3647	0.5438	0.9874
karnof	0.9746	1.0261	0.9613	0.9880
race	0.8910	1.1224	0.7167	1.1075
gender	0.9697	1.0313	0.6768	1.3893
symptom	1.4860	0.6730	1.2131	1.8203

图 3: Model 1 hazard ration

我们可以看到 arms=2 的 exp(coef)=0.5173，其中由文档知，该计算结果是控制其它因子取平均值后计算得到，即

$$\hat{HR} = \exp(\beta_{arms}) = 0.5173$$

所以 arms=2 与 arms=0 的风险比为 0.5173，其 95% 置信区间为 $(0.4075, 0.6568)$. 从而说明接受 zidovudine 和 zalcitabine 治疗的组风险是只接受 zidovudine 治疗的组的 51.73%，因此，zidovudine 和 zalcitabine 的组合治疗效果更好，有利于提高病人的生存时间

1.3 (c)

1.3.1 log-log plot

这里绘制 log-log plot 的方法基于 KM 估计，没有假设潜在的 Cox 模型

- cd40
  在数据集中 cd40 是连续变量，为了进行 PH 假设检验，需先将 cd40 分类，这里我们分成三个水平：low, medium, high，分组依据由其四分位数（见图 4）确定，代码如下

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	263.5	340.0	350.5	423.0	1199.0

图 4: Summary of cd40

```r
# cd40分组
summary(actg$cd40)
actg$s_cd40<- 0
actg[which(actg$cd40 <= 263.5),]$s_cd40=1
actg[which(actg$cd40 > 263.5 & actg$cd40 <= 423.0),]$s_cd40 = 2
actg[which(actg$cd40 > 423.0),]$s_cd40 = 3

km1 <- survfit(y_surv~actg$s_cd40)
km2 <- summary(km1)
km3 <- data.frame(km2$strata,km2$time,km2$surv)
names(km3)=c("cd40","time","survival")
l_cd40 <- km3[km3$cd40=="actg$s_cd40=1",]
m_cd40 <- km3[km3$cd40=="actg$s_cd40=2",]
h_cd40 <- km3[km3$cd40=="actg$s_cd40=3",]
gr_cd40 <- rbind(l_cd40, m_cd40, h_cd40)

p2 <- ggplot(gr_cd40, aes(x=time,y=survival,color=cd40)) + geom_line()
p2 + xlab("survival time in days") +
  ylab("log-log survival") +
  scale_colour_manual(breaks=c("actg$s_cd40=1","actg$s_cd40=2","actg$s_cd40=3"),
                      labels=c("low","medium","high"),
                      values= c('#000000', '#56B4E9', '#E69F00'))
```

结果见图 5，可以看出，随着时间的增加，三条曲线的差距在逐渐增大，所以可以初步判断因子 cd40 不满足 PH 假设

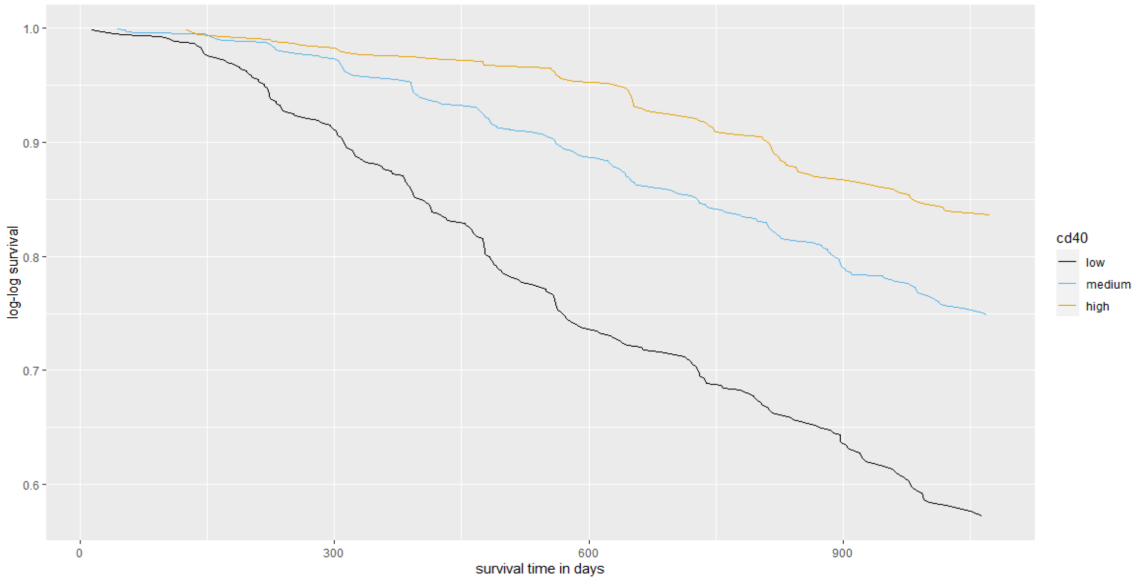

图 5: log-log plot by cd40

- gender
  因为性别是个 01 变量，所以直接绘制根据 gender 分组的 log-log 图，代码如下，结果见图 6

```r
# gender
kmfit1 <- survfit(y_surv~actg$gender)
kmfit2 <- summary(kmfit1)
kmfit3 <- data.frame(kmfit2$strata,kmfit2$time,kmfit2$surv)
names(kmfit3)=c("gender","time","survival")
gender0 <- kmfit3[kmfit3$gender=="actg$gender=0",] #分组
gender1 <- kmfit3[kmfit3$gender=="actg$gender=1",]
gr_gender <- rbind(gender0,gender1) # 整合数据

p1 <- ggplot(gr_gender, aes(x=time,y=survival,color=gender))+geom_line()
p1 + xlab("survival time in days") +
  ylab("log-log survival") +
  scale_color_discrete(breaks=c("actg$gender=0","actg$gender=1"),
                       labels=c("gender=0","gender=1")) # 修改图例名称
```

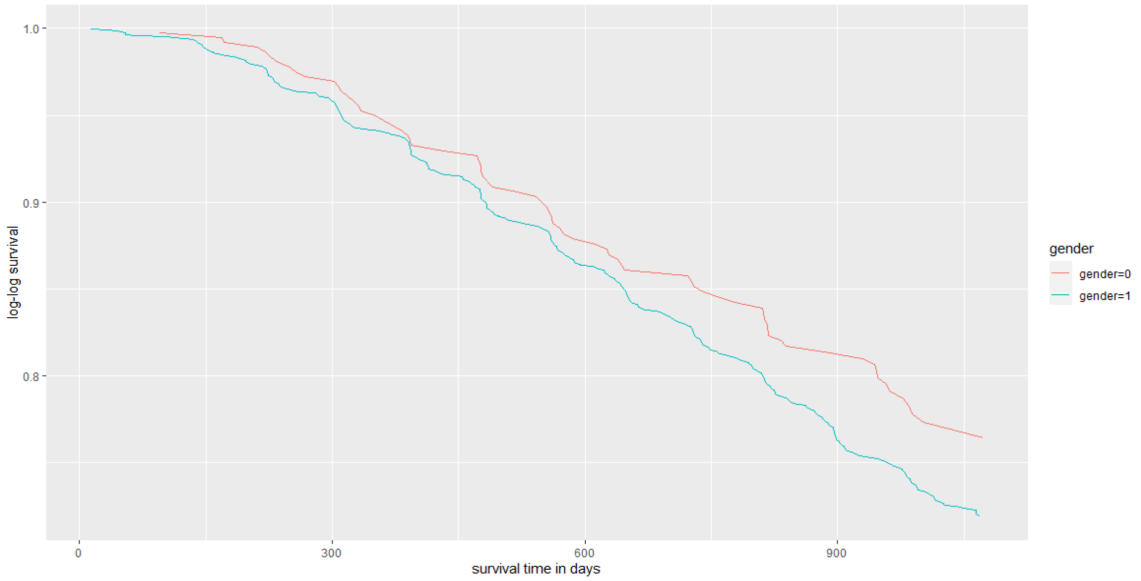

图 6: log-log plot by gender

从图中可以看出，两条曲线没有相交，且差距的变化不明显，在最后的部分，两条曲线的差距才逐渐来开，所以不能确定因子 gender 是否满足 PH 假设，需要进一步通过检验确定

1.3.2 GOF 和 SChoenfeld residual plot

这里拟合优度检验采用教材上基于 Schoenfeld 残差的检验，代码如下

```r
model3 <- coxph(y_surv~gender+cd40,data=actg)
phtest <- cox.zph(model3,transform = rank)
phtest
ggcoxzph(phtest,point.size = 1,ggtheme =  scale_y_continuous())
```

值得注意的是 cox.zph() 中第二个参数选择 rank，则会检验 ranked survival times 相较于 Schoenfeld residuals 而不是默认的 survival times

ggcoxzph() 函数针对每个协变量生成标准化的 Schoenfeld 残差相对于时间的相关性图

最后结果见图 7，Shoenfeld residuals plot 见图 8

```
        chisq df      p
gender  0.012  1 0.9129
cd40   10.756  1 0.0010
GLOBAL 10.757  2 0.0046
```

图 7: result of GOF

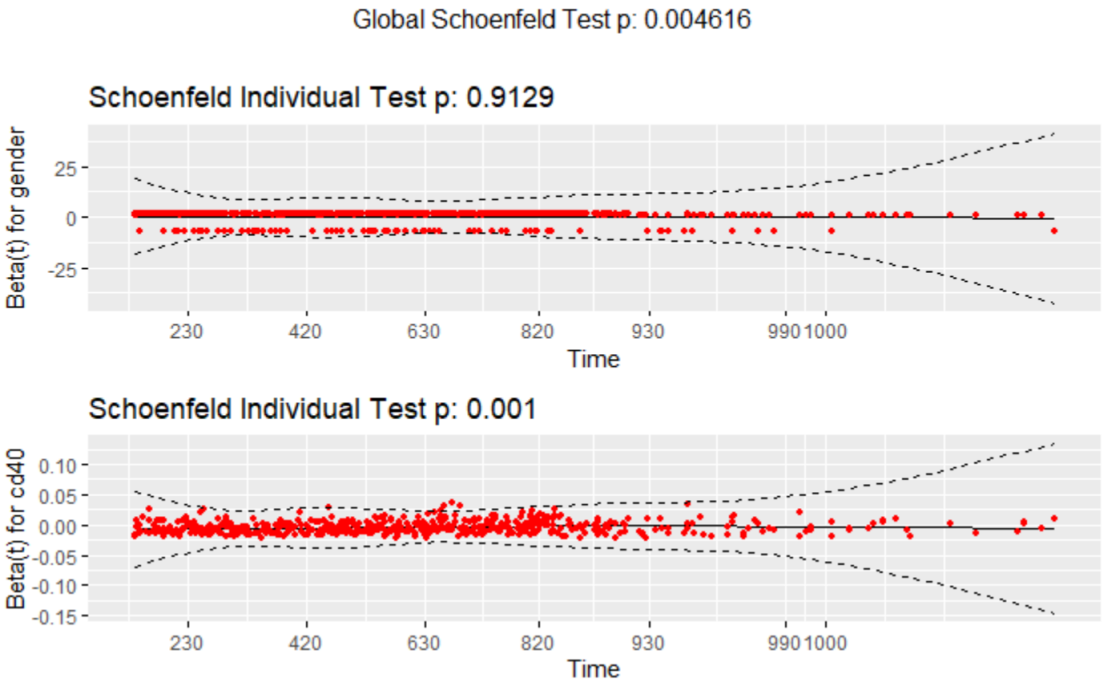

图 8: Shoenfeld residuals plot

从 GOF 检验结果来看，gender 的 p 值 >0.05，所以可以接受原假设，认为 gender 满足 PH 假设；而 cd40 的 p 值 <0.05，所以拒绝原假设，认为 cd40 不满足 PH 假设.

图 8中，实线是拟合的样条平滑曲线，虚线表示拟合曲线上下 2 个单位的标准差。如果曲线偏离 2 个单位的标准差则表示不满足比例风险假定。如果仅从图判断，各协变量满足 PH 风险假设，但实际上 cd40 的图中残差方差较大，散点分布不均匀，导致拟合出现较为平滑的效果，实际上通过 p 值，其并不满足 PH 假设

1.4 (d)

加入 age^2,age^3 的代码如下，得到的结果见图 9

```
model4 <- coxph(y_surv~factor(arms)+cd40+cd80+age+I(age^2)+I(age^3)+
                wtkg+hemo+homo+drugs+karnof+race+gender+symptom,data=actg)
summary(model4)
```

```
                     coef  exp(coef)   se(coef)      z Pr(>|z|)
factor(arms)1 -7.811e-01  4.579e-01  1.243e-01 -6.282 3.35e-10 ***
factor(arms)2 -6.529e-01  5.205e-01  1.219e-01 -5.355 8.54e-08 ***
factor(arms)3 -5.648e-01  5.685e-01  1.161e-01 -4.866 1.14e-06 ***
cd40          -4.327e-03  9.957e-01  4.452e-04 -9.719  < 2e-16 ***
cd80           4.542e-04  1.000e+00  8.437e-05  5.383 7.31e-08 ***
age           -1.161e-01  8.904e-01  1.063e-01 -1.092 0.274806
I(age^2)       2.332e-03  1.002e+00  2.689e-03  0.867 0.385703
I(age^3)      -1.215e-05  1.000e+00  2.173e-05 -0.559 0.576053
wtkg           9.872e-04  1.001e+00  3.508e-03  0.281 0.778413
hemo          -2.006e-02  9.801e-01  2.263e-01 -0.089 0.929386
homo           2.613e-02  1.026e+00  1.584e-01  0.165 0.868967
drugs         -2.771e-01  7.580e-01  1.528e-01 -1.814 0.069743 .
karnof        -2.600e-02  9.743e-01  6.999e-03 -3.715 0.000203 ***
race          -1.227e-01  8.845e-01  1.109e-01 -1.107 0.268460
gender        -2.291e-02  9.774e-01  1.836e-01 -0.125 0.900720
symptom        4.119e-01  1.510e+00  1.039e-01  3.964 7.37e-05 ***
```

图 9: Summary of model 4

可以看到 age^2 和 age^3 的 p 值都大于 0.05，所以在显著性水平 0.05 的情况下，不能拒绝原假设，认为 age^2 和 age^3 的系数不为 0，应当添加进模型中，age 对生存时间的影响是非线性的

1.5 (e)

将交叉项 gender:factor(arms) 添加进 model 1 中，代码如下

```
model5 <- coxph(y_surv~factor(arms)+cd40+cd80+age+wtkg+hemo+homo+drugs+
                karnof+race+gender+symptom+gender:factor(arms),data=actg)
summary(model5)
```

结果见图 10，gender 与 arms 的交叉项 p 值都大于 0.05，说明不能拒绝原假设，结果不显著，因此我们认为交叉项应不添加进模型中

```
                              coef  exp(coef)   se(coef)      z Pr(>|z|)
factor(arms)1         -9.754e-01  3.770e-01  3.548e-01 -2.749 0.005974 **
factor(arms)2         -6.242e-01  5.357e-01  3.279e-01 -1.903 0.056983 .
factor(arms)3         -3.152e-01  7.297e-01  2.943e-01 -1.071 0.284229
cd40                  -4.255e-03  9.958e-01  4.458e-04 -9.546  < 2e-16 ***
cd80                   4.430e-04  1.000e+00  8.428e-05  5.256 1.47e-07 ***
age                    8.500e-03  1.009e+00  5.244e-03  1.621 0.105022
wtkg                   5.367e-05  1.000e+00  3.550e-03  0.015 0.987937
hemo                   1.062e-01  1.112e+00  2.135e-01  0.497 0.618924
homo                   2.244e-02  1.023e+00  1.580e-01  0.142 0.887054
drugs                 -3.078e-01  7.351e-01  1.524e-01 -2.020 0.043346 *
karnof                -2.607e-02  9.743e-01  6.994e-03 -3.728 0.000193 ***
race                  -1.203e-01  8.867e-01  1.113e-01 -1.080 0.279970
gender                 1.199e-02  1.012e+00  2.532e-01  0.047 0.962213
symptom                3.945e-01  1.484e+00  1.036e-01  3.808 0.000140 ***
factor(arms)1:gender   2.253e-01  1.253e+00  3.790e-01  0.594 0.552206
factor(arms)2:gender  -4.055e-02  9.603e-01  3.529e-01 -0.115 0.908505
factor(arms)3:gender  -2.976e-01  7.426e-01  3.204e-01 -0.929 0.353028
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

图 10: Summary of model 5

2

2.1 (a)

将新的数据代入 model 1 的代码如下

```
patern0 <- data.frame(arms=c(0,1,2,3),cd40=400, cd80=500, age=25, wtkg=70,
hemo=0,homo=0,drugs=1,karnof=90,race=1,gender=1,symptom=0)

new_model <- survfit(model1,newdata=patern0)
ggsurvplot(new_model,data=new_model,
           xlab = "days",
           ylim = c(0.7,1),
           legend.title = "",      #图例标题
           legend.labs = c("zidovudine", "zidovudine and didanosine",
                           "zidovudine and zalcitabine", "didanosine"),
           legend = c(0.25,0.3))
```

作出的生存曲线如图 11, 可以看出 zidovudine 和 didanosine 组合治疗组的生存率下降最慢，说明该治疗方法对这个新病人来说效果是最好的，能够延长其生存时间

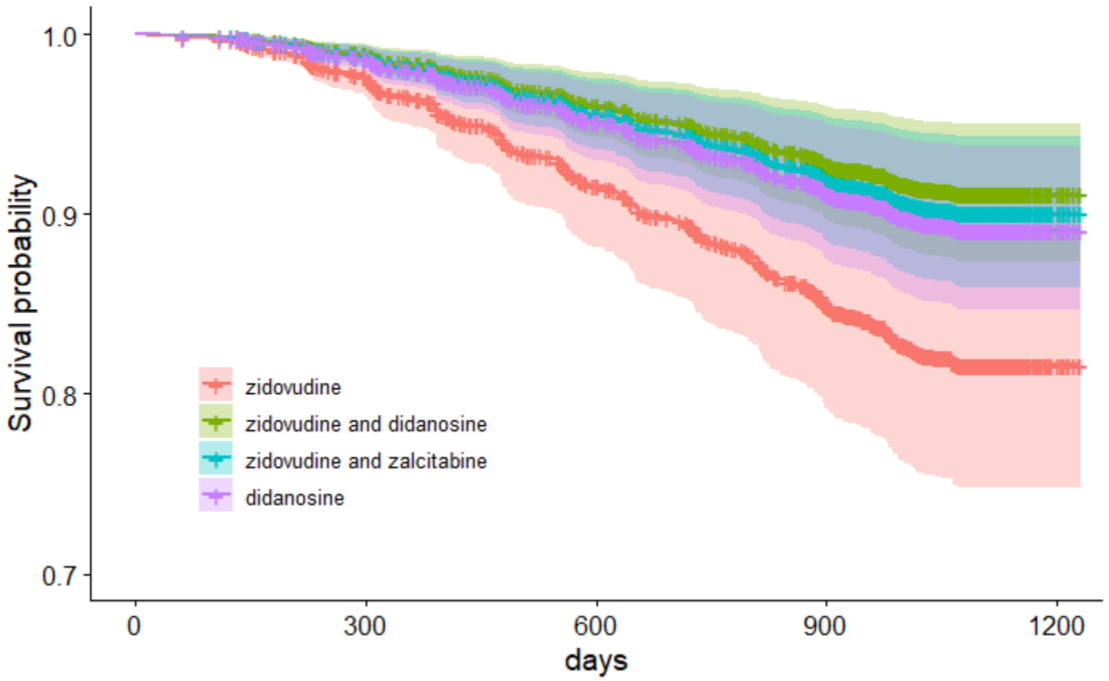

图 11: Survival curve of 4 treatment

2.2 (b)

```
summary(new_model,times=365*2)
```

```
Call: survfit(formula = model1, newdata = patern0)

 time n.risk n.event survival1 survival2 survival3 survival4
  730   1595     349      0.89     0.948     0.941     0.936
```

图 12: Probability of surviving more than two years

代码及结果见上,可以得到四种方案 (arms=0,1,2,3) 的存活时间超过两年的生存率分别为 0.89, 0.948, 0.941, 0.936. 所以 zidovudine 和 didanosine 组合治疗的方案对该病人来说有最大的生存率

2.3 (c)

代码见下，计算数据集中四种方案得到的 risk score 中位数和平均数，与新病人的日 risk score 进行比较

```
outcome <- predict(model1,newdata=patern0,type="risk")
names(outcome)=c("0","1","2","3")

s_mod1 <- summary(model1)
p_mod1 <- predict(model1,type="risk")
arm0 <- p_mod1[actg$arms==0]
arm1 <- p_mod1[actg$arms==1]
arm2 <- p_mod1[actg$arms==2]
arm3 <- p_mod1[actg$arms==3]
mean_four <- c(mean(arm0), mean(arm1), mean(arm2), mean(arm3))
median_four <- c(median(arm0), median(arm1), median(arm2), median(arm3))
cbind(outcome, mean_four, median_four)
```

结果见图 13，可以看出无论是 ACTG175 研究中的数据的中位数还是平均数，新病人的 risk score 都低于这两个水平，说明 4 种方案相比较而言，新病人的 risk score 处于较低水平，风险较低

```
   outcome mean_four median_four
0 0.7231630 1.9578929   1.6451296
1 0.3316369 0.9233367   0.7697687
2 0.3741224 1.0076599   0.8207676
3 0.4101078 1.1126701   0.9589386
```

图 13: Comparsion between new patient and ACTG175 Study