Assignment 1
(Due: 2020/06/03, 11:59pm)

Note:

- **No late assignment accepted**;
- Submit your assignment in a single PDF file ($name + ID.pdf$) with all R code and outputs; submit it to
  `statistics_sysu@163.com`
  by the deadline;
- Write your assignment in Chinese or English.

Analysis of ACTG175 Data

```
Data Source: ACTG175(speff2trial).txt
Reference: Hammer et al. (1996), New England Journal of Medicine
```

ACTG 175 was a randomized clinical trial to compare monotherapy with zidovudine or didanosine with combination therapy with zidovudine and didanosine or zidovudine and zalcitabine in adults infected with the human immunodeficiency virus type I whose CD4 T cell counts were between 200 and 500 per cubic millimeter.

Explanation of the dataset: A data frame with 2139 observations on the following 27 variables.

```
pidnum: patients ID number
age: age in years at baseline
wtkg: weight in kg at baseline
hemo: hemophilia (0=no, 1=yes)
homo: homosexual activity (0=no, 1=yes)
drugs: history of intravenous drug use (0=no, 1=yes)
karnof: Karnofsky score (on a scale of 0-100)
oprior: non-zidovudine antiretroviral therapy prior to initiation of study treatment
 (0=no, 1=yes)
```

z30: zidovudine use in the 30 days prior to treatment initiation (0=no, 1=yes)
zprior: zidovudine use prior to treatment initiation (0=no, 1=yes)
preanti: number of days of previously received antiretroviral therapy
race: race (0=white, 1=non-white)
gender: gender (0=female, 1=male)
str2: antiretroviral history (0=naive, 1=experienced)
strat: antiretroviral history stratification (1=antiretroviral naive, 2=> 1 but
less than 52 weeks of prior
antiretroviral therapy, 3=> 52 weeks)
symptom: symptomatic indicator (0=asymptomatic, 1=symptomatic)
treat: treatment indicator (0=zidovudine only, 1=other therapies)
offtrt: indicator of off-treatment before 96 plus/minus 5 weeks (0=no,1=yes)
modSearch 3
cd40: CD4 T cell count at baseline
cd420: CD4 T cell count at 20 plus/minus 5 weeks
cd496: CD4 T cell count at 96 plus/minus 5 weeks (=NA if missing)
r: missing CD4 T cell count at 96 plus/minus 5 weeks (0=missing, 1=observed)
cd80: CD8 T cell count at baseline
cd820: CD8 T cell count at 20 plus/minus 5 weeks
cens: indicator of observing the event in days
days: number of days until the first occurrence of: (i) a decline in CD4 T cell count of
at least 50
(ii) an event indicating progression to AIDS, or (iii) death.
arms treatment arm (0=zidovudine, 1=zidovudine and didanosine,
2=zidovudine and zalcitabine, 3=didanosine).

Reading the data:

```
#Read the data file
ACTG175<-read.csv("E:/ACTG175(speff2trial).txt",   header=TRUE,
   sep=",")
#Obtain the number of rows and columns of the dataset.
dim(ACTG175)
[1] 2139   27
#Display the first 3 row of data
ACTG175[1:3,]
#The output is:
  pidnum age    wtkg hemo homo drugs karnof oprior z30 zprior preanti race
```

```
1  10056  48 89.8128    0    0     0    100      0   0      1       0    0
2  10059  61 49.4424    0    0     0     90      0   1      1     895    0
3  10089  45 88.4520    0    1     1     90      0   1      1     707    0
  gender str2 strat symptom treat offtrt cd40 cd420 cd496 r cd80 cd820 cens
1      0    0     1       0     1      0  422   477   660 1  566   324    0
2      0    1     3       0     1      0  162   218    NA 0  392   564    1
3      1    1     3       0     1      1  326   274   122 1 2063  1893    0
#Display  the cd40 values for first 100 rows:
ACTG175$cd40[1:100]
#The output is:
 [1] 422 162 326 287 504 235 244 401 214 221 471 340 540 212 120 150 350 330
 [19] 180 233 320 470 230 400 344 421 227 357 486 238 236 407 257 342 444 496
 [37] 370 186 386 332 422 393 266 454 416 293 224 331 253 307 364 340 293 227
 [55] 601 483 470 256 389 421 204 251 211 199 158 209 245 499 505 260 210 360
 [73] 250 410 430 400 420 310 510 540 770 430 350 470 300 490 210 290 260 420
 [91] 320 360 280 300 240 270 360 530 168 272
```

In particular, CD4 cell count is an important biomarker for HIV/AIDS disease, lower CD4 cell count means worse situation of HIV/AIDS disease.

Note: `cens` is binary variable δ_i: 1 indicates event occurrence; 0 indicates censoring; `days` is the observed survival time T_i. We set significance level $\alpha = 0.05$.

Questions:

1. Draw a survival curve plot using the Kaplan-Meier approach for the four treatment groups (i.e., `arms`). Compare the four survival curves in this plot. From the plot, which treatment is most effective in prolonging the patients' survival time?

2. Does the median survival time in each of the four treatment groups exist? Why?

3. What is the KM estimate of the survival function $S(t)$ in each of the four treatment groups for $t = 365$, and the corresponding 95% confidence interval? What is the KM estimate of the survival function $S(t)$ in each of the four treatment groups for $t = 730$, and the corresponding 95% confidence interval?

4. Assume that the survival time T^* is continuous and it satisfies $E(T^{*2}) < \infty$. Show that $E(T^*) = \int_0^\infty S(t)dt$.