MTH6157 Survival Models – Assessed Coursework 1

This is the first of two assessed courseworks, each of which count towards 15% of your total module mark. The deadline for submission is 5pm UK time Wednesday 18 October (week 4). Late submissions will not be accepted and if you fail to submit by the deadline and do not have an accepted EC claim your mark will be zero. You should submit your solution in a MS Word document. Please paste all R code and R output into the document along with any typed answers. You should ensure that you submit your own work and that your submission complies with QMUL policies on plagiarism and collusion.

Question

Medical researchers are looking to assess a new pharmaceutical treatment for a certain virus. They conduct a trial with 120 patients, half taking an existing medication and the other half taking the new drug. The researchers measure the number of days before all virus symptoms are gone.

The file `Coursework1Data.csv` available on QM Plus contains the results of this trial with the following data recorded:

`patient`	a patient number from 1 to 120
`treatment`	1 if the patient was given the existing medication, 2 if the patient was given the new drug
`delta`	1 if all virus symptoms were recorded as gone, 0 if the patient left the trial before all symptoms were gone
`time`	the number of days before either all symptoms were recorded as gone or the patient left the trial

(a) Load the `Coursework1Data.csv` file into R and install the `survival` package. [1]

```
> Coursework1Data <- read.csv("~/6157/Coursework1Data.csv")
> treatment <- Coursework1Data$treatment
> time <- Coursework1Data$time
> delta <- Coursework1Data$delta
> library(survival)
```

(b) Give an example of how informative right censoring might be present in this trial. [2]

A patient assigned the new drug in the trial experiencing side effects may stop taking this drug and swap to the existing medication.

There are a wide range of potential correct answers here – the key elements are that the censoring is non-random and causes an early exit from the trial.

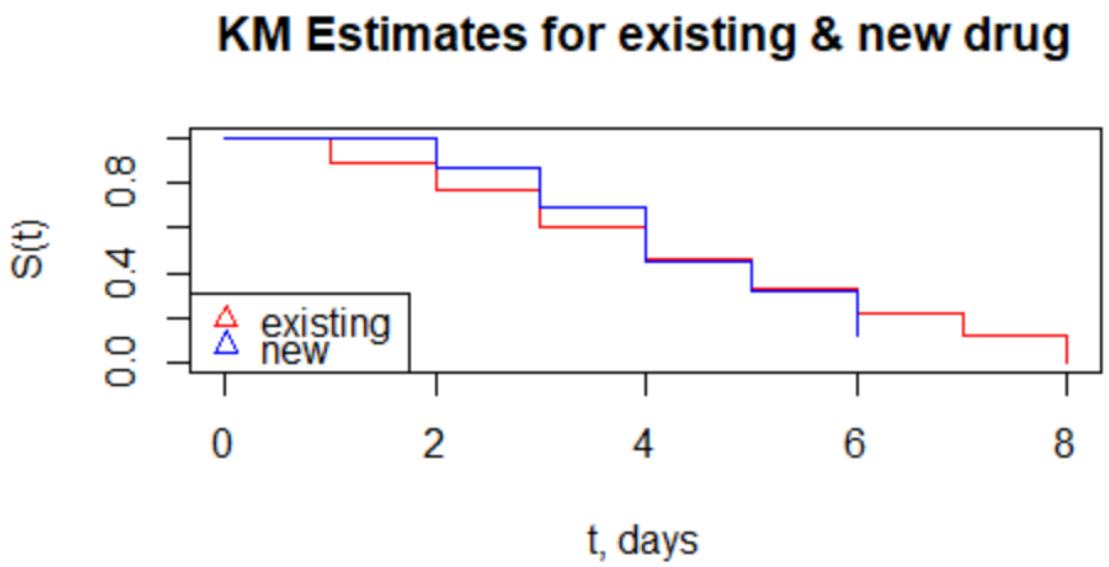
(c) Construct R code to calculate the Kaplan Meier Estimate of the survival function separately for patients on the existing medication and those given the new drug. [7]

```
> KM <- survfit( Surv(time, delta) ~ treatment, conf.int=0.95, conf.type="plain")
```

Note that it is also possible to construct two KM estimates separately for each treatment by splitting the dataset into two by treatment. Full marks will be awarded for this approach too although the single KM function code above is both more elegant and more scalable to >2 treatments. Where the separate KM estimates are derived here in (c) code will be needed to combine them into a single plot in (d) below.

(d) Plot the two survival functions on one graph with a different colour for each using R, making sure you add suitable title, axis labels and legend. [6]

```
> plot(KM, xlab = "t, days", ylab = "S(t)", main = "KM Estimates for existing & new drug", col = c("red", "blue"))
> legend("bottomleft", legend = c("existing", "new"), col = c("red", "blue"), pch=2)
```



(e) What does your plot in (d) say about which treatment is better? [5]

- We seek to minimize the survival function here
- Both treatments are effective in eliminating virus symptoms
- The shape of the two survival functions is very similar
- The existing treatment appears more effective between 1 and 4 days
- The new treatment removes symptoms for all patients faster (t=6) than the existing treatment (t=8)
- If the goal is to maximise early recovery then the existing treatment is slightly better
- If the goal is to eliminate symptoms in all patients then the new drug is better
- This trial therefore cannot state definitively that one treatment is better
- It can be noted that the sample sizes are quite large for a new drug trial

(f) What concerns might the researchers have about the trial results? [4]

- The effect of censoring is not the same in the two treatments
- An investigation of the dataset shows
    - 3 censored patients for the existing treatment (5% of the sample)
    - 15 censored patients for the new drug (1/4 of the sample)
    - Therefore there is 5 times as much censoring with the new drug
- We do not know the cause of the censoring
- This should be investigated
- To determine whether this censoring is informative (which would seem likely given the size of the difference)
- For example does the new drug have significant side effects leading to right censoring?

To gain full marks in this section you need to refer to censoring, explore the effect of censoring in the dataset, and explain why this might be concerning.

[Total 25]

Your mark out of 25 will be converted to a % to be comparable with the second coursework.