Proposal

JEFFREY CHAN

10/26/2020

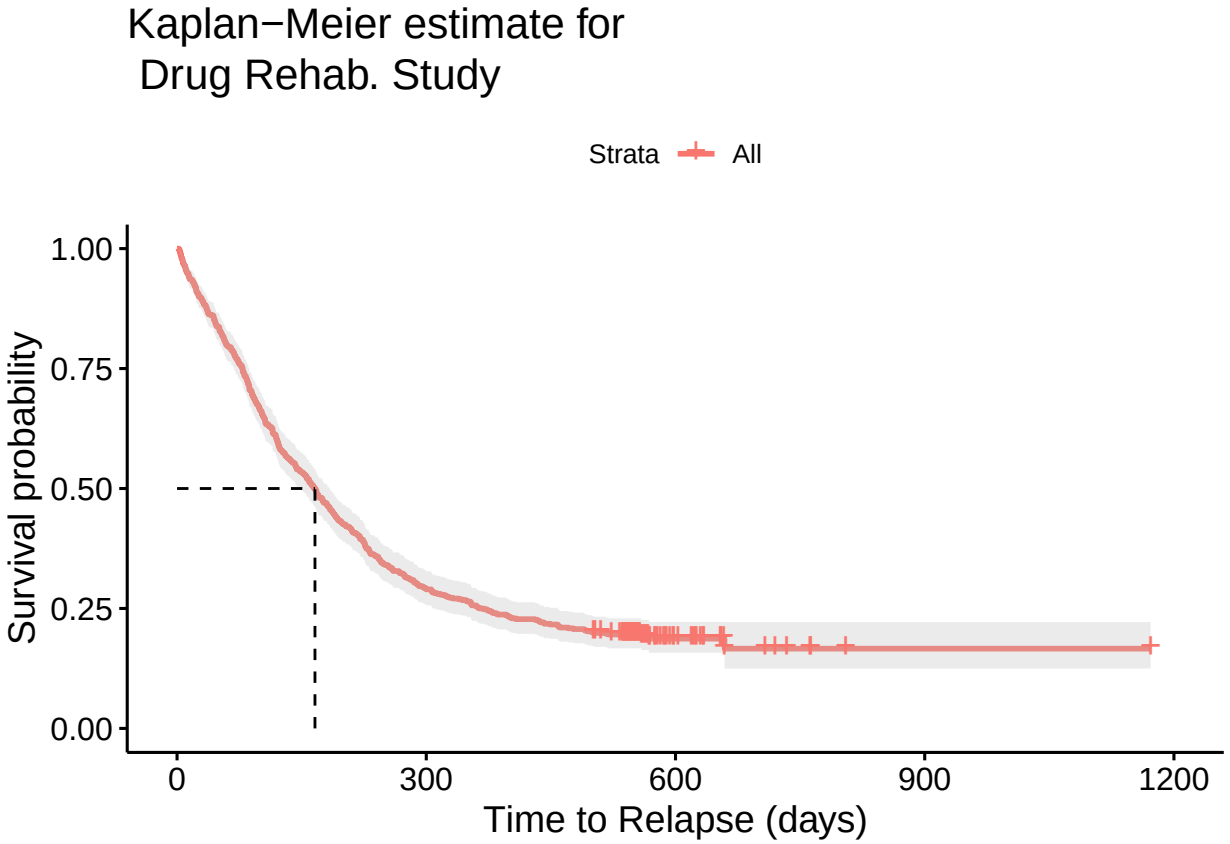
In this project I will analyze the UIS UMARU Impact study data. The data are measurements from a drug treatment study. The data take the form of survival data with the quantity of interest being the time until Return to Drug Use (Time data). The covariates recorded are the age of the subject; their Beck depression score at admission; whether they used heroin, cocaine, neither or both in the 3 months prior to admission; their history of intraveneous drug use at admission; the number of prior treatments a subject had received; the race (recorded only as White or other); a treatment randomization assignment (Long or Short); their treatment site (A or B); the length of treatment; and whether they returned to drug use (event of interest) or were censored from the study. These data are recorded for 628 patients. The dataset:

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

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
library("ggplot2")
drugs <- read.delim("/Users/jeffreychan/uis.txt",
                    header = F,
                    stringsAsFactors =F,
                    sep =""
                    )
colnames(drugs)<-c("id","age","beck","hercoc",
                   "ivhx","ndrugtx","race","treat",
                   "site","los","time","censor"
                   )
surv.Obj <- Surv(drugs$time, drugs$censor)
surv.fit <- survfit(surv.Obj~1, data = drugs)
ggsurvplot( surv.fit,data=drugs,
            conf.int=T,
            title="Kaplan-Meier estimate for \n Drug Rehab. Study",
            xlab = "Time to Relapse (days)",
            surv.median.line ="hv"
            )
```



Some initial scientific questions we might ask of this data set are: What is the estimated median time to relapse? What is a 95% confidence interval for this median time and can we conclude at the 5% significance level that this median time is less than one year? These are some superficial questions that can be quickly determined from the Kaplan-Meier curve.

More interesting questions would involve the covariates in the data set. We will for example apply the log-rank test to test whether the Long and Short treatment assignments follow the same survival curve, and whether treatment sites A and B follow the same survival curve. Applying the log-rank test for several groups would let us ask questions like whether the survival curves differ for the various groups of previous drug exposure (cocaine use, heroin use, both and neither).

We expect age, previous drug exposure, history of IV drug use, and the Beck depression score to all be interesting covariates. The most likely confounding covariate would be the treatment assignment (Long or Short), since this will correlate with the length of treatment covariate and probably also the relapse time. It seems that the number of previous drug treatments could perhaps also be a confounding covariate, since this seems likely to simultaneously correlate with their prior exposure to heroin/cocaine, with their history of IV drug use, and also their relapse time. We will test for confounding covariates.

The greatest difficulty I see at the moment is how to deal with the length of treatment variable. It does not seem appropriate to use any of our tests for comparing survival curves of groups to this covariate. It may be interesting to try to analyze its effect using linear regression, perhaps multivariate linear regression together with other relevant covariates.