Homework5

Sherine George

04 December, 2023

The Publication data in the ISLR2 R package contains information about the time to publication for the results of 244 clinical trials funded by the National Heart, Lung, and Blood Institute. Take some time to read more about this dataset in chapter 11.5.4 of ISL. You can also type ?Publication in R for more information after loading the dataset.

Question 1 - 5 points

Load the Publication dataset in R.

Solution:

```
data("Publication")
```

Question 2 - 10 points

Calculate for how many clinical trials the time associated with the event of interest (i.e., the time of publication) is observed and calculate for how many clinical trials the time of the event of interest is censored.

Solution:

```
table(Publication$status)
```

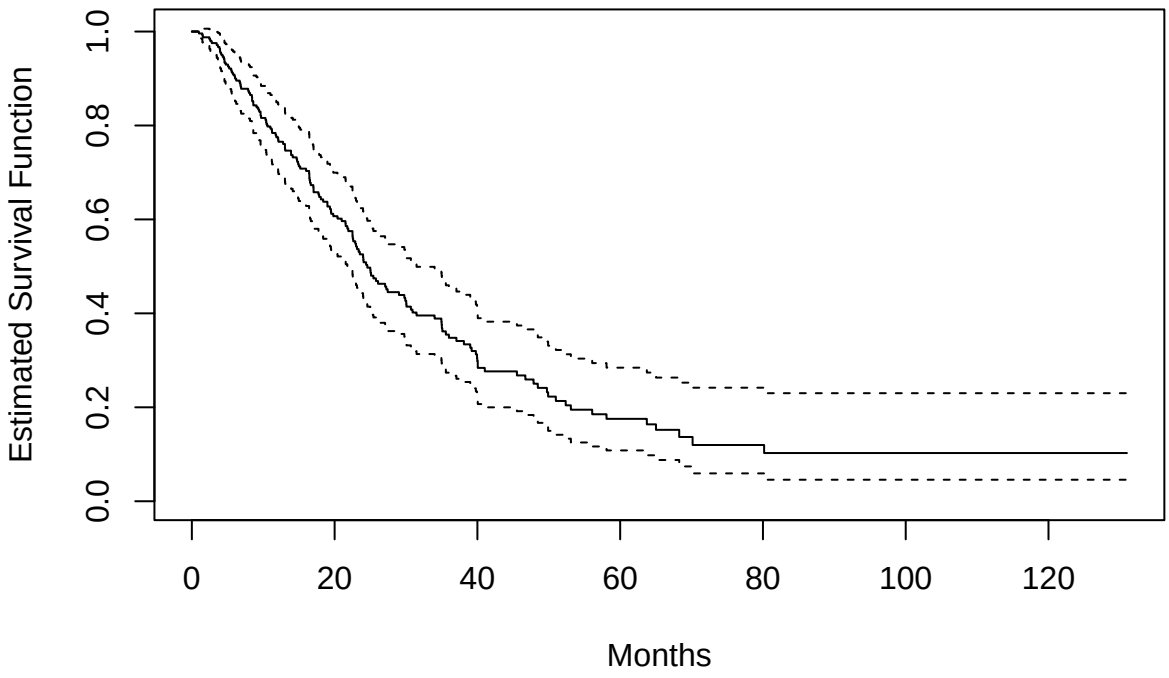
```
##
##   0   1
##  88 156
```

Question 3 - 10 points

Produce and plot the Kaplan-Meier estimator for the time to publication of all the clinical trials in the dataset. Include 99% pointwise confidence bands in the plot.

Solution:

```
# Creating the Kaplan-Meier survival function estimate.
kapmei <- survfit(Surv(time, status) ~ 1, data = Publication)
plot(
    kapmei,
    xlab = "Months",
    ylab = "Estimated Survival Function",
    conf.int=0.99
)
```

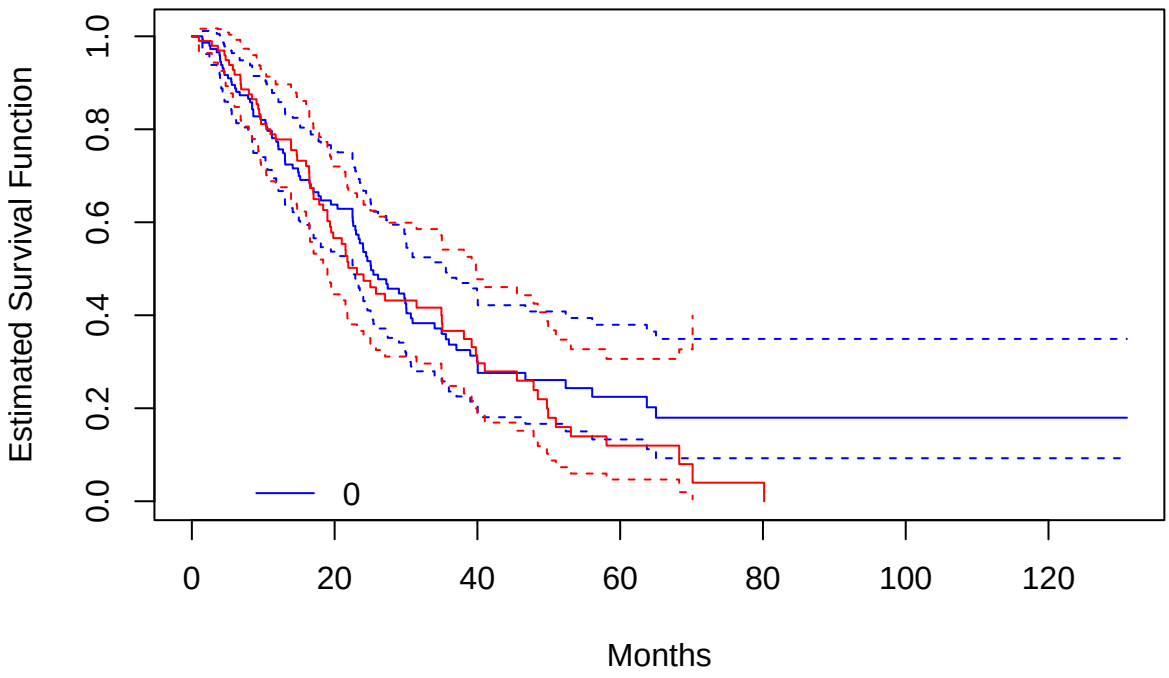


Question 4 - 20 points

Produce and plot the Kaplan-Meier estimator for the time to publication of the clinical trials for the two subgroups corresponding to the posres variable (i.e., for the group of clinical trials that resulted in positive findings and for the group of clinical trials that did not result in positive findings). Then, use the log-rank test to test the null hypothesis that the time to publica- tion is not associated with whether or not the clinical trial resulted in a positive finding (posres). State in English the result of the log-rank test.

Solution:

```
km_p <- survfit(Surv(time, status) ~ posres, data = Publication)
plot(
    km_p,
    xlab = "Months",
    ylab = "Estimated Survival Function",
    conf.int = 0.99,
    col = c("blue", "red")
)
legend(
    5, 0.1,
    c(0,1),
    col = c("blue", "red"),
    lty = 1,
    box.lty = 0
)
```




```
p_logrank <- survdiff(Surv(time, status) ~ posres, data = Publication)
p_logrank
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ posres, data = Publication)
##
##             N Observed Expected (O-E)^2/E (O-E)^2/V
## posres=0 146       87     92.6     0.341     0.844
## posres=1  98       69     63.4     0.498     0.844
##
##  Chisq= 0.8  on 1 degrees of freedom, p= 0.4
```

Given the p-value is 0.4, it suggests insufficient evidence to reject the null hypothesis, which posits no association between the time to publication and the outcome of a clinical trial being positive (posres).

Question 5 - 10 points

Fit a Cox proportional hazards model to these data using the following predic- tors: • posres • multi • clinend • budget. Also produce the model summary with the summary function.

Solution:

```
cox_model <- coxph(Surv(time, status) ~ posres+multi+clinend+budget, data = Publication)
summary(cox_model)
```

```
## Call:
```

```
## coxph(formula = Surv(time, status) ~ posres + multi + clinend +
##     budget, data = Publication)
##
##   n= 244, number of events= 156
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## posres  0.533728  1.705278 0.178275 2.994  0.00275 **
## multi   0.633555  1.884298 0.227922 2.780  0.00544 **
## clinend 1.641604  5.163447 0.241385 6.801 1.04e-11 ***
## budget  0.002282  1.002285 0.001800 1.268  0.20477
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## posres      1.705     0.5864    1.2024     2.418
## multi       1.884     0.5307    1.2054     2.945
## clinend     5.163     0.1937    3.2172     8.287
## budget      1.002     0.9977    0.9988     1.006
##
## Concordance= 0.731  (se = 0.021 )
## Likelihood ratio test= 81.39  on 4 df,   p=<2e-16
## Wald test            = 97.91  on 4 df,   p=<2e-16
## Score (logrank) test = 125.1  on 4 df,   p=<2e-16
```

Question 6 - 10 points

Do the global likelihood ratio, Wald, and score test suggest that the model is better than a model that does not use any predictor?

Solution:

Null Hypothesis: Our model exhibits no improvement over a baseline model that lacks any predictors.

Considering that the p-value is less than or equal to 2e-16, significantly lower than our chosen significance level of 0.01, we have substantial grounds to reject the null hypothesis with approximately 99% confidence. Consequently, we can assert that our model displays enhanced performance compared to the baseline model.

Question 7 - 20 points In English, interpret the estimated effect of each predictor on the hazard function corresponding to the time to publication.

Solution:

βposres = 0.533728; ebβposres = 1.705278. The analysis indicates that for every one-unit increase in the posres predictor, there is a 1.705278-fold multiplicative change in the hazard rates, holding all other variables constant. In simpler terms, a single-unit elevation in the posres predictor is associated with an approximately 70.5278% increase in the hazard rates.

βmulti = 0.633555; ebβmulti = 1.884298. The examination suggests that with each one-unit increase in the multi predictor, there is a 1.884298-fold change in hazard rates, assuming all other factors remain constant. In simpler terms, a unit increase in the multi predictor is associated with an estimated 88.4298% rise in the hazard rates.

βclinend = 1.641604; ebβclinend = 5.163447. The analysis indicates that an elevation of one unit in the clinend predictor is linked to a 5.163447-fold rise in the hazard rates, holding all other variables constant. Put more plainly, each incremental unit increase in the clinend predictor corresponds to a surge of 416.3447% in the hazard rates.

βbudget = 0.002282; ebβbudget = 1.002285.The examination suggests that a one-unit increase in the budget predictor is associated with a 1.002285-fold alteration in hazard rates, with all other variables held constant.

This implies that for each unit rise in the budget predictor, there is an estimated uptick of 0.2285% in the hazard rates.

Question 8 - 15 points the cumulative hazard function

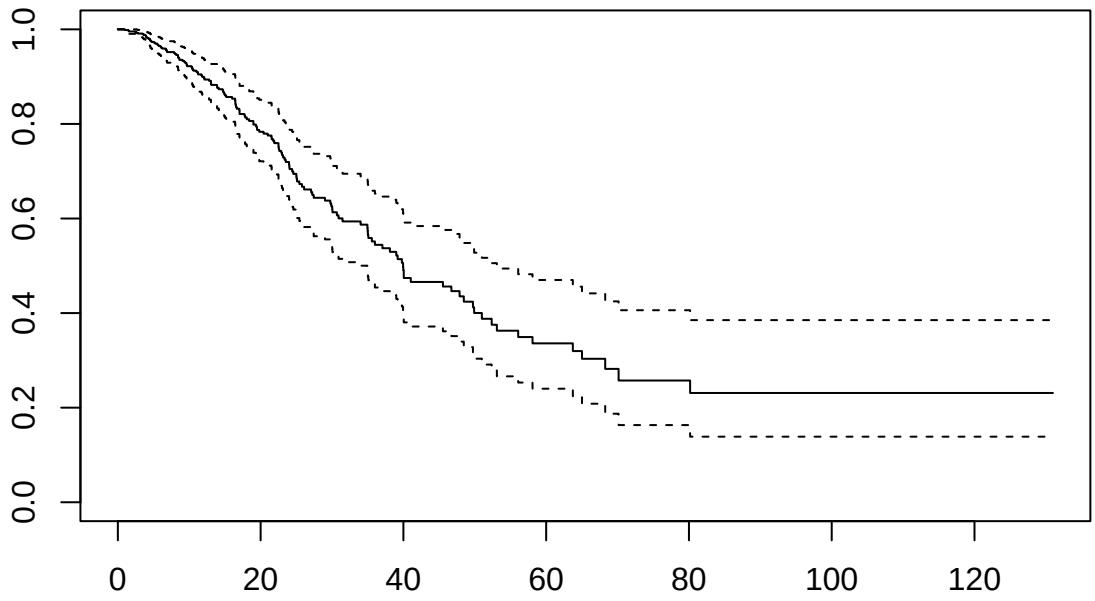
In R, you can compute the estimated survival function Sb from a Cox propor- tional hazards model and plot it as follows:

estimated_survival_function <- survfit(, newdata =) plot(estimated_survival_function)

First, use the above commands to plot the estimated baseline survival function S0 for the time to publication (this is obtained by setting all predictors to 0 in newdata). Then, use the above commands to plot. the estimated survival function for the time to publication of a new clinical trial with predictor values – posres: 0 – multi: 0 – clinend: 1 – budget: 8.5 the estimated survival function for the time to publication of a new clinical trial with predictor values – posres: 0 – multi: 0 – clinend: 0 – budget: 1.3 Based on the two estimated survival functions, which of these two clinical trials do you think is more likely to be published sooner? Explain.

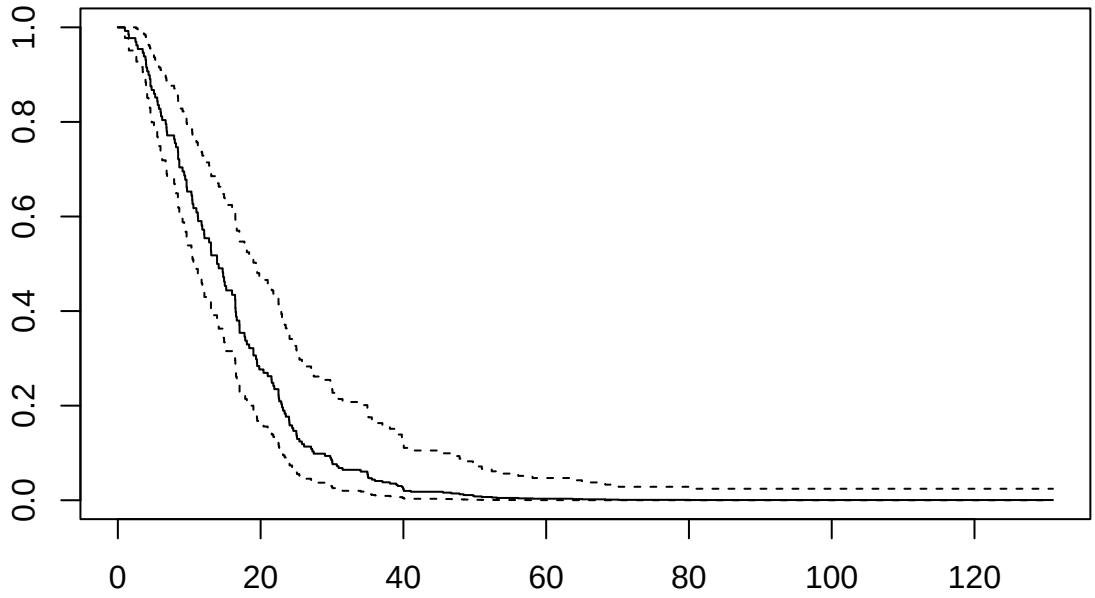
Solution:

```
#Baseline
newdata <- data.frame(
  posres = 0,
  multi = 0,
  clinend = 0,
  budget = 0
)
estimated_survival_function <- survfit(cox_model,newdata)
plot(estimated_survival_function)
```



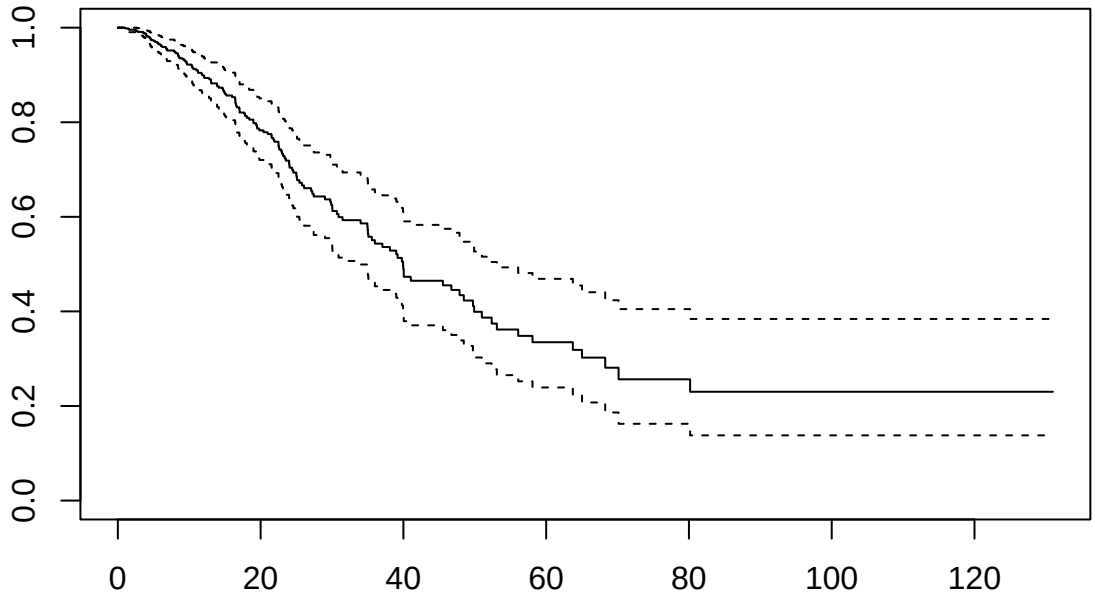
```r
#The estimated survival function for the time to publication of a new clinical trial with predictor val
#- posres: 0
#- multi: 0
#- clinend: 1
#- budget: 8.5

newdata <- data.frame(
  posres = 0,
  multi = 0,
  clinend = 1,
  budget = 8.5
)
estimated_survival_function <- survfit(cox_model,newdata)
plot(estimated_survival_function)
```




```r
#The estimated survival function for the time to publication of a new clinical
#trial with predictor values
#- posres: 0
#- multi: 0
#- clinend: 0
#- budget: 1.3
newdata <- data.frame(
  posres = 0,
  multi = 0,
  clinend = 0,
```

```
  budget = 1.3
)
estimated_survival_function <- survfit(cox_model,newdata)
plot(estimated_survival_function)
```



Given the steeper curve observed in the first estimated survival function, it is anticipated that this particular survival function will be published sooner than the others.