BST 222 - Homework 3

Xiaowei Zeng

2023-10-26

The addicts data set is from a study by Caplehorn et al. ("Methadone Dosage and Retention of Patients in Maintenance Treatment," Med. J. Aust., 1991)in Sydney, Australia and contains a cohort of 238 heroin addicts who entered maintenance programs between February 1986 and August 1987. The time in the clinic was determined by dates of entry into and exit or transfer from the methadone maintenance program to which they had been assigned. There are two further covariates, namely, prison record and methadone dose, believed to affect the survival times.

The data set and R input code are on the website. The variables are as follows:

- id: Subject ID
- clinic: Clinic (1 or 2)
- status: Survival status (1 = left the clinic, 0 = censored, i.e. we do not know when they left)
- time: Survival time in days
- prison: Prison record (0 = none, 1 = any)
- methadone: Methadone dose (mg/day)

Load the R packages and the dataset.

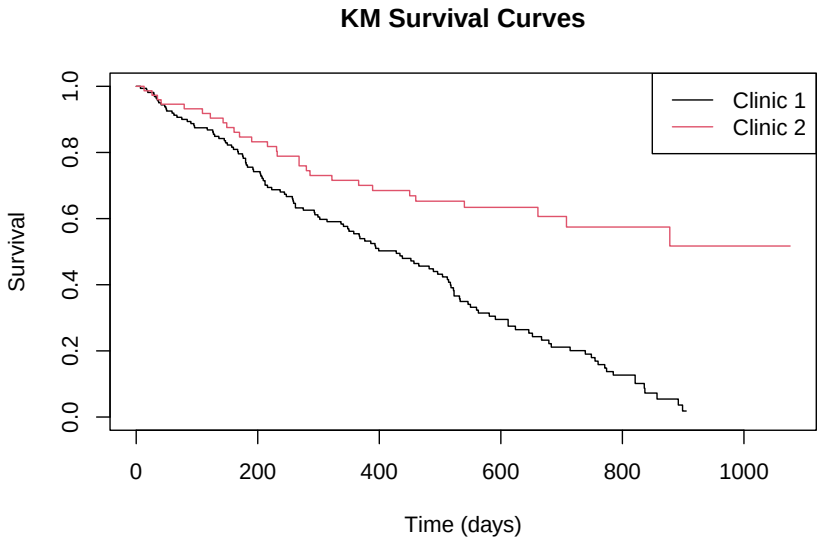
```
library(survival)
library(KMsurv)
library(MASS)

vars <- c("id","clinic","status","time","prison","methadone")
addicts <- read.table("https://dmrocke.ucdavis.edu/Class/BST222.2023.Fall/addicts.txt",
                      header=F, col.names=vars)

#change variables to factors to be used in Cox PH
addicts$clinic <- factor(addicts$clinic,labels=c("Clinic1","Clinic2"))
addicts$prison <- factor(addicts$prison,labels=c("No","Yes"))

head(addicts)
```

```
##   id  clinic status time prison methadone
## 1  1 Clinic1      1  428     No        50
## 2  2 Clinic1      1  275    Yes        55
## 3  3 Clinic1      1  262     No        55
## 4  4 Clinic1      1  183     No        30
## 5  5 Clinic1      1  259    Yes        65
## 6  6 Clinic1      1  714     No        55
```

```
# Create a survival object
dfsurv <- Surv(addicts$time, addicts$status)
```

1. Plot the Kaplan-Meier survival curves for the two clinics.

```
KMcurves <- survfit(dfsurv ~ clinic, data = addicts)
plot(KMcurves, col = 1:2, xlab = 'Time (days)', ylab = 'Survival')
legend('topright', c('Clinic 1', 'Clinic 2'), col = 1:2, lwd = 1)
title('KM Survival Curves')
```



Clinic 1 has a lower probability of staying at the clinic (namely a higher risk for leaving) than Clinic 2 for nearly all time points. Subjects are leaving Clinic 1 quicker and do not stay as long as those in Clinic 2.

2. Test whether the two survival curves could have come from the same process.

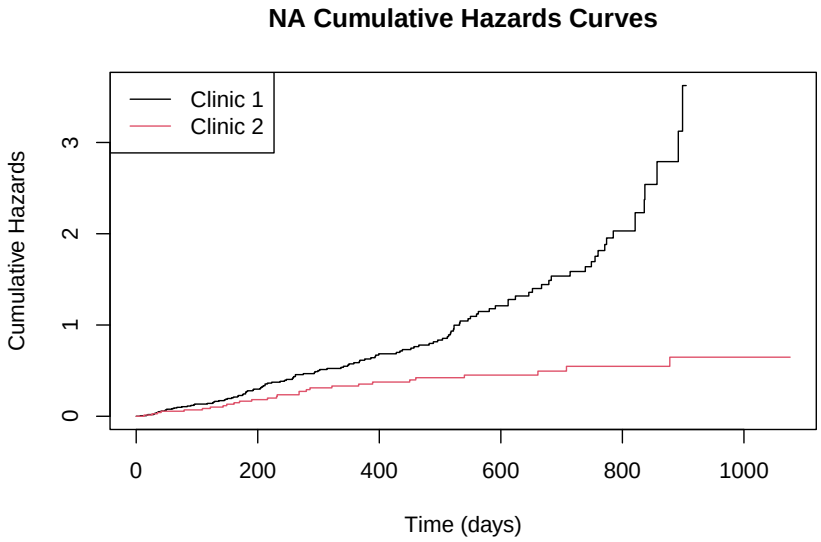
```
survdiff(dfsurv ~ clinic, data = addicts)
```

```
## Call:
## survdiff(formula = dfsurv ~ clinic, data = addicts)
##
##                   N Observed Expected (O-E)^2/E (O-E)^2/V
## clinic=Clinic1 163      122     90.9      10.6      27.9
## clinic=Clinic2  75       28     59.1      16.4      27.9
##
##  Chisq= 27.9  on 1 degrees of freedom, p= 1e-07
```

Suppose that the significance level α is 0.05. The p-value of the test is 1e-07 < 0.05, so we will reject the null hypothesis that the two true survival curves come from the same process.

3. Plot the cumulative hazards from the Nelson-Aalen estimator.

```
NAcurves = survfit(dfsurv ~ clinic, data = addicts, type = 'fleming-harrington')
plot(NAcurves, col = 1:2, fun = 'cumhaz', xlab = 'Time (days)', ylab = 'Cumulative Hazards')
legend('topleft', c('Clinic 1', 'Clinic 2'), col = 1:2, lwd = 1)
title('NA Cumulative Hazards Curves')
```



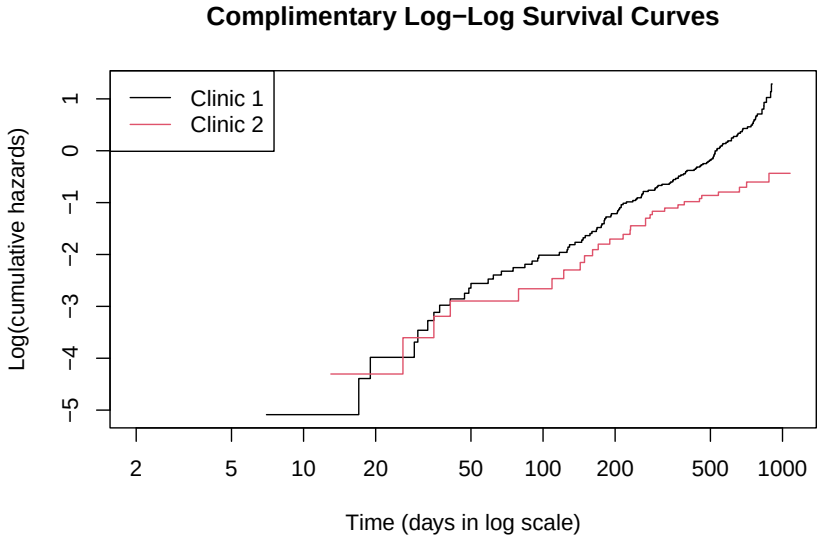
The cumulative hazards of Clinic 1 are larger than those of Clinic 2 for nearly all time points (except the first several days). The cumulative hazards of Clinic 1 surge after 600 days, which means the subjects in Clinic 1 are more likely to leave after 600 days.

4. Complementary log-log survival plot.

A common comparison plot for proportional hazards is the complimentary log-log survival plot which plots $\ln(-\ln[\hat{S}(x)]) = \ln \hat{H}(t)$ against $\ln(t)$. This helps us check the proportional hazards assumption and it can also help us decide if the Weibull distribution would be appropriate for the failure times.

- If the hazards are proportional, then so are the cumulative hazards, and after taking logs, the curves should be parallel.
- If the lines are straight, then the Weibull model may be appropriate. Make this plot for the two clinics using the NelsonAalen estimator and comment on the results.

```
plot(NAcurves, col=1:2, fun='cloglog', xlab='Time (days in log scale)',
     ylab='Log(cumulative hazards)')
legend('topleft', c('Clinic 1', 'Clinic 2'), col = 1:2, lwd = 1)
title('Complimentary Log-Log Survival Curves')
```



Except the initial fluctuations before 50 days, the two lines are roughly parallel. Since the survival time ranges from 0 to 1076 and 50 is quite short compared to this scale, we can assume that the hazards are proportional.

5. Construct a Cox model using only the clinic variable.

Using Clinic 1 as the reference group (baseline), fit a Cox model labeled as Model 1.

```
cox1 <- coxph(dfsurv ~ clinic, data = addicts)
summary(cox1)
```

```
## Call:
## coxph(formula = dfsurv ~ clinic, data = addicts)
##
##   n= 238, number of events= 150
##
##                   coef exp(coef) se(coef)      z Pr(>|z|)
## clinicClinic2 -1.0754    0.3412   0.2127 -5.057 4.26e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## clinicClinic2    0.3412      2.931    0.2249    0.5176
##
## Concordance= 0.574  (se = 0.022 )
## Likelihood ratio test= 30.99  on 1 df,   p=3e-08
## Wald test            = 25.57  on 1 df,   p=4e-07
## Score (logrank) test = 27.92  on 1 df,   p=1e-07
```

Note that when the model only includes one independent variable, the local test (t-test) and the global test are equivalent. According to the results, the p-value of the clinic variable is 4.26e-07, far smaller than 0.05 (the significance level), indicating that the coefficient is significantly different from 0. The p-values of the global tests, namely Likelihood ratio test, Wald test and logrank test, are also far smaller than 0.05. Therefore, we can conclude that the "survival" is significantly different at the two clinics.

The hazard ratio for Clinic 2 vs Clinic 1 is $e^{-1.0754} = 0.3412$, which means Clinic 2 has 0.3412 times the risk for leaving the clinic compared to Clinic 1. The 95% confidence interval for $e^{\beta_{\text{clinic2}}}$ is $(0.2249, 0.5176)$, which does not include 1 (the entire interval is less than 1), so we can say that the risk for leaving in Clinic 2 is significantly lower than in Clinic 1.

6. Pick one test for the null hypothesis that the clinics do not differ.

I would like to choose the likelihood ratio test because in most cases it has the best performance.

Recall that three global tests are automatically conducted in `coxph`: the likelihood ratio test based on the maximized likelihood itself, the Wald test based on the estimators standardized by use of the information matrix (asymptotic normality), and the score test (or the log rank test when there are no ties) based on the first derivatives of the log likelihood. All of these tests are asymptotically $\chi^2(p)$ where p is the number of coefficients in the vector.

Empirical studies have shown that the convergence rate of the likelihood ratio and Wald tests are similar, but the LR test slightly outperforms the Wald test for small sample size, the same as the conclusion shown in the lecture notes that the LR test has faster convergence than the Wald test in many cases. The score

test converges less rapidly to the limiting chi-squared distribution and thus is less accurate (particularly for small sample size). Therefore, given the same sample size, the LR test has the largest power of test, which is consistent to the result that it has the smallest p-value of 3e-08, indicating that the two clinics do differ.

7. Consider adding the prison and methodone variables.

Without taking the interaction term into consideration, we fix a second Cox model, labeled as Model 2, with clinic, prison and methadone as the covariates.

```
cox2 <- coxph(dfsurv ~ clinic + methadone + prison, data = addicts)
summary(cox2)
```

```
## Call:
## coxph(formula = dfsurv ~ clinic + methadone + prison, data = addicts)
##
##   n= 238, number of events= 150
##
##                     coef exp(coef)  se(coef)      z Pr(>|z|)
## clinicClinic2 -1.009896  0.364257  0.214889 -4.700 2.61e-06 ***
## methadone     -0.035369  0.965249  0.006379 -5.545 2.94e-08 ***
## prisonYes      0.326555  1.386184  0.167225  1.953   0.0508 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##               exp(coef) exp(-coef) lower .95 upper .95
## clinicClinic2    0.3643     2.7453    0.2391    0.5550
## methadone        0.9652     1.0360    0.9533    0.9774
## prisonYes        1.3862     0.7214    0.9988    1.9238
##
## Concordance= 0.665  (se = 0.025 )
## Likelihood ratio test= 64.56  on 3 df,   p=6e-14
## Wald test            = 54.12  on 3 df,   p=1e-11
## Score (logrank) test = 56.32  on 3 df,   p=4e-12
```

It turns out that the p-values of the clinic and methadone variables are all far smaller than 0.05, indicating significant coefficients, while the p-value for the prison variable is larger than 0.05. So we wonder if the prison variable will improve the model.

There are usually two common ways to figure out this question:

- Analysis of Variance (ANOVA).

```
anova(cox2)
```

```
## Analysis of Deviance Table
##  Cox model: response is dfsurv
## Terms added sequentially (first to last)
##
##            loglik   Chisq Df Pr(>|Chi|)
## NULL      -705.54
## clinic    -690.04 30.9896  1  2.594e-08 ***
## methadone -675.15 29.7980  1  4.795e-08 ***
```

```
## prison    -673.26  3.7727  1    0.05209 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The negative log likelihood decreases when each variable is added to the model, so all the variables do improve the model.

- AIC / Likelihood Ratio Test.

```
drop1(cox2, test='Chisq')
```

```
## Single term deletions
##
## Model:
## dfsurv ~ clinic + methadone + prison
##           Df    AIC     LRT  Pr(>Chi)
## <none>       1352.5
## clinic     1 1376.9 26.3506 2.847e-07 ***
## methadone  1 1381.3 30.7820 2.887e-08 ***
## prison     1 1354.3  3.7727   0.05209 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results show that the model with all three variables has the lowest AIC (which is the best), while the likelihood ratio test gives the prison variable a p-value of $0.05209 > 0.05$. Since it doesn't exceed the significance level much, we can assume that the prison variable positively contribute to the model.

Therefore, both covariates seem to improve the model and Model 2 should be better than Model 1.

8. Compare the survival curves from Cox model and KM estimates.

Since we've already fit Cox Model 1 (the Cox PH model with only the clinic variable) and Cox Model 2 (the Cox PH model with clinic, methadone and prison variables), I will plot the survival curves from both models and KM estimates.

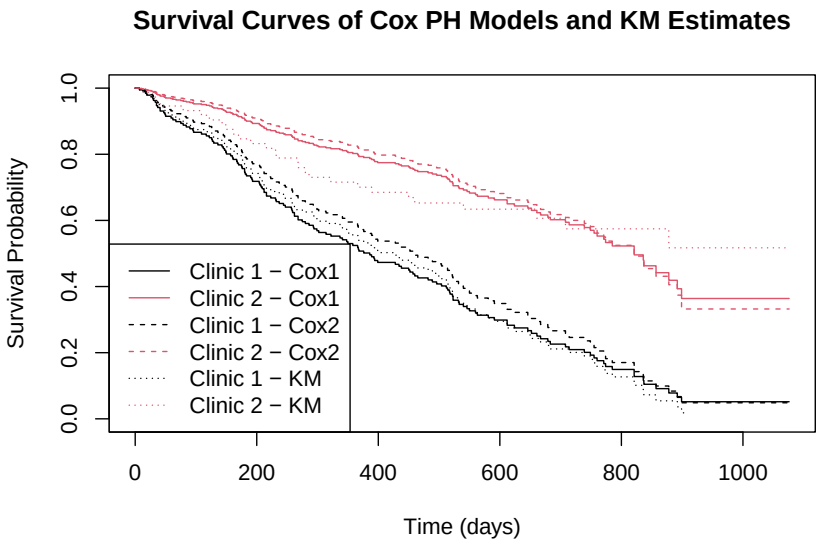
```
# Settings for Model 1
covariates <- data.frame(clinic = c('Clinic1', 'Clinic2'))
cox3 <- survfit(cox1, covariates, conf.int = F)
```

In Model 2, we set the covariates levels as the mean of numeric variable (methadone) and the mode of categorical variable (prison).

```
# Settings for Model 2
m = mean(addicts$methadone) # mean of methadone
p = names(which.max(table(addicts$prison))) # mode of prison
covariates <- data.frame(clinic = c('Clinic1', 'Clinic2'),
                         methadone = c(m, m),
                         prison = c(p, p))
cox4 <- survfit(cox2, covariates, conf.int = F)
```

Now plot all the survival curves.

```
plot(cox3, col=1:2, ylab='Survival Probability', xlab='Time (days)')
lines(cox4, col=1:2, lty = 2)
lines(KMcurves, col = 1:2, lty = 3)
legend('bottomleft', c('Clinic 1 - Cox1', 'Clinic 2 - Cox1',
                       'Clinic 1 - Cox2', 'Clinic 2 - Cox2',
                       'Clinic 1 - KM', 'Clinic 2 - KM'),
       col = rep(1:2, 3), lty = c(1, 1, 2, 2, 3, 3), lwd = 1)
title('Survival Curves of Cox PH Models and KM Estimates')
```



According to the plot, for Clinic 1 the survival curves of Cox PH models and those of KM estimates are very close. But for Clinic 2, before 700 days the survival probability of Cox model exceeds that of KM estimates, while after 700 days it's the reverse. Note that the KM estimator is unrestricted and suitable for all dataset; the Cox PH model allows additional covariates to be included and thus is more flexible, but it requires the Proportional Hazard (PH) assumption that KM estimates do not need. Therefore, the slight difference between Cox and KM may be attributed to the PH assumption.