Computer Project

Hortencia J. Hernandez[*] University of Texas at El Paso (UTEP)

7 May, 2024

Contents

[*] hjhernandez4@miners.utep.edu

1 EDA

The dataset uis is from the University of Massachusetts AIDS Research Unit IMPACT Study. It was a 5 year project comprised of two concurrent randomized trails of residential treatment for drug abuse. The goal was to compare treatment programs of different planned durations designed to reduce drug abuse and to prevent high-risk HIV behavior. The dataset contains 628 participants and 12 variables that was collected.

1.1

Missing Values

	Missing Information	
Variable	Percentage Missing(%)	Total Missing
id	0.0000	0
age	0.7962	5
beck	5.2548	33
hercoc	2.8662	18
ivhx	2.8662	18
ndrugtx	2.7070	17
race	0.9554	6
treat	0.0000	0
site	0.0000	0
los	0.0000	0
time	0.0000	0
status	0.0000	0

From the table we can see that the variables: age, beck, hercoc, ivhx, ndrugtx, and race contain missing values. Since the percentage of missing is relatively small ($\leq$ 10%) we will proceed with listwise deletion.

Listwise deletion will leave us with 575 observations (53 observations were removed).

1.2

Censoring rate on the observed event times?

The censoring rate on the observed event times is approximately 19.3043

1.3

Among the covariates, how many of them are continuous and how many are categorical?

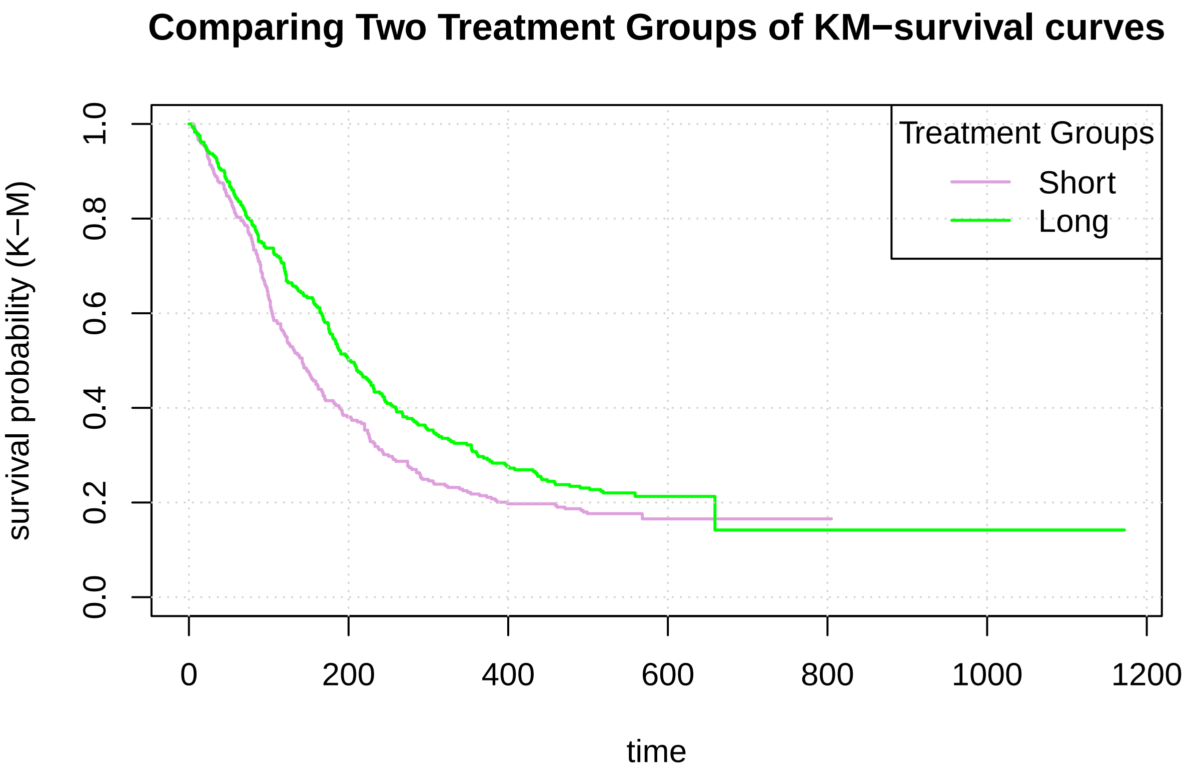
Covariates	**Codes/Values**	**Type**
age	Years	*Continuous*
beck	0.000-54.000	*Continuous*
heroc	Heroin & Cocaine/Heroin Only/Cocaine Only/Neither	*Categorical*
ivhx	Never/Previous/Recent	*Categorical*
ndrugtx	0-40	*Continuous*
race	White/Non-white	*Categorical*
treat	Short/Long	*Categorical*
site	A/B	*Categorical*
los	Days	*Continuous*
time	Days	*Continuous*

We can see from the table, there are 4 continuous variables and 5 categorical variables

2 Treatment Effect Assessment

2.1

Plot the Kaplan-Meier survival curves in the two treatment groups and compare. Does the proportional hazards (PH) seem to hold?



Notice that the two curves cross, this would indicate that the hazards are not proportional and thus that the groups vary over time. We can also see that the survival rate decreases over time but declines faster in the short treatment group.

2.2

Use logrank test to asses the effect of "treat". Fit a CoxPH with "treat" only. Among the three tests (LRT, score, and Wald) available in the output of Cox PH model, to which one is the logrank test closest? Interpret the results in terms of the hazard ratio or relative risk between the two treatment groups.

The logrank test statistic is 6.528 with $p = 0.01$ suggests that there is significant evidence (since $0.01 < \alpha = 0.05$) that there is a difference in survival between the short and long treatment.

Fitting a CoxPH with `treat` it appears that the three tests are close to the logrank test (6.528). In particular the closest appears to be the score test (6.537). Since the `coef` is -0.24 and is a negative value it suggests the hazard risk is lower in the Long treatment. The hazard ratio (0.79) suggests that the Long treatment has a lower hazard compared to the Short treatment and their difference is statistically significant (since $\alpha = 0.0107 < 0.05$).

2.3

It is often of interest to examine treatment-by-site interaction in a multi-center trial. Fit a Cox PH model with the interaction term `treat x site` and determine whether site is an effect-moderator.

Examining the treatment-by-site interaction, the coefficient for site is approximately -0.2122 with hazard ratio of 0.8088, but it is not statistically significant (p = 0.1329) which indicates that the site alone may not have a significant influence in survival outcome. For the interaction term `treat x site`, its coefficient is approximately 0.193 with hazard ratio of 1.2129, but it is not statistically significant (p = 0.3426) which suggests that the interaction between the two may not significantly influence the survival outcome.

3 Cox PH Modeling

3.1

For variable screening purpose, fit simple Cox PH model by including variables one at a time. Output the p-value of LRT associated with each variable and tabulate the results.

Variable	LRT	P-Value
age	3.159065	0.0755064
beck	4.487811	0.0341374
hercoc	3.782576	0.0517889
ivhx	13.641799	0.0002212
ndrugtx	12.694466	0.0003667
race	7.754668	0.0053574
treat	6.507069	0.0107446
site	1.111861	0.2916777
los	132.539400	0.0000000

The variables with significant p-values include beck, ivhx, ndrugtx, race, treat, los

3.2

Build your 'best' predictive Cox PH model with a variable selection procedure of your choice; call it bfit.ph. Again, remove all rows that contain a missing value. Optionally, you may also consider including first-order interaction terms.

```
## Start:  AIC=5154.38
## Surv(time, status) ~ id + age + beck + hercoc + ivhx + ndrugtx +
##     race + treat + site + los
## 
##           Df    AIC
## - id       1 5152.4
## - hercoc   1 5152.4
## - beck     1 5153.4
## <none>       5154.4
## - treat    1 5154.6
## - age      1 5158.9
## - site     1 5159.8
## - race     1 5160.5
## - ndrugtx  1 5160.6
## - ivhx     1 5162.6
## - los      1 5297.5
## 
## Step:  AIC=5152.42
## Surv(time, status) ~ age + beck + hercoc + ivhx + ndrugtx + race + 
##     treat + site + los
## 
##           Df    AIC
## - hercoc   1 5150.5
## - beck     1 5151.5
## <none>       5152.4
## - treat    1 5152.7
## + id       1 5154.4
## - age      1 5157.0
## - race     1 5158.5
## - ndrugtx  1 5158.6
## - ivhx     1 5160.7
## - site     1 5165.3
## - los      1 5295.6
## 
## Step:  AIC=5150.47
## Surv(time, status) ~ age + beck + ivhx + ndrugtx + race + treat + 
##     site + los
## 
##           Df    AIC
## - beck     1 5149.5
## <none>       5150.5
## - treat    1 5150.7
## + hercoc   1 5152.4
## + id       1 5152.4
## - age      1 5155.0
## - race     1 5156.5
## - ndrugtx  1 5156.9
## - ivhx     1 5162.3
## - site     1 5163.4
## - los      1 5293.7
```

```
##
## Step:  AIC=5149.53
## Surv(time, status) ~ age + ivhx + ndrugtx + race + treat + site +
##     los
##
##           Df    AIC
## <none>       5149.5
## - treat    1 5149.9
## + beck     1 5150.5
## + hercoc   1 5151.5
## + id       1 5151.5
## - age      1 5154.3
## - race     1 5155.2
## - ndrugtx  1 5156.0
## - site     1 5162.3
## - ivhx     1 5162.5
## - los      1 5294.2
```

```
## Call:
## fitfunc(formula = as.formula(x), data = data)
##
##   n= 575, number of events= 464
##
##               coef  exp(coef)   se(coef)       z Pr(>|z|)
## age     -0.0216841  0.9785493  0.0081494  -2.661 0.007795 **
## ivhx     0.2256559  1.2531443  0.0601880   3.749 0.000177 ***
## ndrugtx  0.0267754  1.0271371  0.0084818   3.157 0.001595 **
## race    -0.2919416  0.7468121  0.1143344  -2.553 0.010668 *
## site     0.4065997  1.5017029  0.1082646   3.756 0.000173 ***
## los     -0.0091383  0.9909033  0.0008061 -11.336  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##         exp(coef) exp(-coef) lower .95 upper .95
## age        0.9785     1.0219    0.9630    0.9943
## ivhx       1.2531     0.7980    1.1137    1.4100
## ndrugtx    1.0271     0.9736    1.0102    1.0444
## race       0.7468     1.3390    0.5969    0.9344
## site       1.5017     0.6659    1.2146    1.8567
## los        0.9909     1.0092    0.9893    0.9925
##
## Concordance= 0.741  (se = 0.011 )
## Likelihood ratio test= 188.4  on 6 df,   p=<2e-16
## Wald test            = 164.5  on 6 df,   p=<2e-16
## Score (logrank) test = 168.5  on 6 df,   p=<2e-16
```

3.3

Interpret your 'best' Coxmodel bfit.ph. Which variables are highly predictive of drug relapse? Are there any variables that are significant in the simple Cox model but not selected by the multiple Cox model? Are there any variables that are insignificant in the simple Cox model but becomes significant in the multiple Cox model?

From the previous section the variables that are highly predictive of drug relapse include: age, ivhx, ndrugtx, race, site, and los. That is, the variables that have negative impact or a decrease in the risk of drug relapse are: age with a hazard rate of 0.9785, and los (length of stay in treatment) with a hazard rate of 0.9909. That is the longer a patient is in treatment the less of a chance of relapse, as well as the older they get. Note that race also has a negative impact that is to say the if a person is non-white they have a higher chance of going into relapse at a risk of 0.7468. IVHX (IV Drug Use History), ndrugtx (number of prior drug treatments), and dependent on the treatment site increases the risk of drug relapse. (with hazard rates of 1.2531, 1.0271, 1.5017, respectively). That is if they have had a history of intravenous drug use, prior drug treatment or treated at site B they have a higher risk of relapse.

Variable(s) that are significant in the simple Cox model but *NOT* in the multiple model is beck (Beck Depression Score) and treat (treatment). Based on the two models age is insignificant in the simple Cox model and it becomes significant in the multiple Cox Model.

4 Appendix:

4.1 Code Results

```
## [1] "2.2: Logrank"
```

```
## Call:
## survdiff(formula = Surv(time, status) ~ treat, data = uis_clean)
##
##           N Observed Expected (O-E)^2/E (O-E)^2/V
## treat=0 289      239      212      3.52      6.53
## treat=1 286      225      252      2.96      6.53
##
##  Chisq= 6.5  on 1 degrees of freedom, p= 0.01
```

```
## [1] "2.2: CoxPH"
```

```
## Call:
## coxph(formula = Surv(time, status) ~ treat, data = uis_clean,
##     ties = "efron")
##
##   n= 575, number of events= 464
##
##           coef exp(coef) se(coef)      z Pr(>|z|)
## treat -0.23744   0.78864  0.09308 -2.551   0.0107 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## treat    0.7886      1.268    0.6571    0.9465
##
## Concordance= 0.538  (se = 0.012 )
## Likelihood ratio test= 6.51  on 1 df,   p=0.01
## Wald test            = 6.51  on 1 df,   p=0.01
## Score (logrank) test = 6.54  on 1 df,   p=0.01
```

```
## [1] "2.3: CoxPH_Multi-Center"
```

```
## Call:
## coxph(formula = Surv(time, status) ~ treat * site, data = uis_clean)
##
##   n= 575, number of events= 464
##
##               coef exp(coef) se(coef)      z Pr(>|z|)
## treat      -0.3010    0.7401   0.1111 -2.710  0.00674 **
## site       -0.2122    0.8088   0.1412 -1.503  0.13293
## treat:site  0.1930    1.2129   0.2034  0.949  0.34262
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
##             exp(coef) exp(-coef) lower .95 upper .95
## treat          0.7401     1.3512    0.5953    0.9201
## site           0.8088     1.2363    0.6133    1.0667
## treat:site     1.2129     0.8245    0.8142    1.8069
##
## Concordance= 0.545  (se = 0.014 )
## Likelihood ratio test= 8.84  on 3 df,   p=0.03
## Wald test            = 9.05  on 3 df,   p=0.03
## Score (logrank) test = 9.11  on 3 df,   p=0.03
```

```
## [1] "3.1: Variable Screening Purpose"
```

```
##       Variable        LRT      P_Value
## test        age   3.159065 7.550640e-02
## test1      beck   4.487811 3.413737e-02
## test2    hercoc   3.782576 5.178886e-02
## test3      ivhx  13.641799 2.212053e-04
## test4   ndrugtx  12.694466 3.667393e-04
## test5      race   7.754668 5.357388e-03
## test6     treat   6.507069 1.074464e-02
## test7      site   1.111861 2.916777e-01
## test8       los 132.539400 1.140212e-30
```

4.2 All Code Used in This Report

```r
knitr::opts_chunk$set(echo = TRUE)
library(dplyr)
library(kableExtra)
library(knitr)
library(tidyr)
library(survival)
library(tidyverse)
library(MASS)
library(glmulti)

uis <- read.table(file="uissurv.txt", sep="", header=F,
                  na.strings = ".",
                  col.names=c("id", "age", "beck", "hercoc", "ivhx", "ndrugtx",
                              "race", "treat", "site", "los", "time", "status"))
head(uis)
# Calculate percentage missing and total missing
missing_summary <- uis %>%
  summarise_all(~ mean(is.na(.)) * 100) %>%
  pivot_longer(everything(), names_to = "Variable",
               values_to = "Percentage_Missing") %>%
  mutate(Total_Missing = colSums(is.na(uis)))

kable(missing_summary, format = "markdown",
      col.names = c("Variable", "Percentage Missing(%)", "Total Missing"),
      digits = 4) %>%
  # Add some formatting using kableExtra
  kable_styling(full_width = FALSE) %>%
  add_header_above(c(" " = 1, "Missing Information" = 2))

#Listwise Deletion

uis_clean <- uis[complete.cases(uis), ]

# Find censoring rate of the observed event times
#   0 is censored; the following code gives the
#   proportion of cases for both status = 0 and status = 1

# censoring_rate[1] represents the censoring rate

censoring_rate <- prop.table(table(uis_clean$status))


#2.1

fitKM <- survfit(Surv(time,status)~treat, data = uis_clean,
                 type='kaplan', conf.type='log-log')
#fitKM
#summary(fitKM)
plot(fitKM, xlab='time', ylab='survival probability (K-M)',
     lty  = 1, col = c("plum", "green"), lwd = 1.5,
     main = "Comparing Two Treatment Groups of KM-survival curves")
grid()
legend("topright", legend = c("Short", "Long"), col = c("plum", "green"),
       lty = 1, lwd = 1.5, title = "Treatment Groups")


# Logrank test

logrank <- survdiff(Surv(time,status)~treat, data=uis_clean)
#logrank

# Fit CoxPH with 'treat' only

fit.cox <- coxph(Surv(time,status)~treat, data = uis_clean, ties="efron")
```

```
#summary(fit.cox)


# treatment by site

fit.cox_multi <- coxph(Surv(time, status) ~ treat * site, data = uis_clean)
#summary(fit.cox_multi)


#3.1 CoxPH Modeling

results <- data.frame(Variable = character(),
                      LRT = numeric(),
                      p_value = numeric(),
                      stringsAsFactors = FALSE)

for (var in names(uis_clean)) {
  # Skip 'time' and 'status' variables as they are part of the survival formula
  if (var %in% c("id", "time", "status")) {
    next
  }

  # Fit a simple Cox PH model for each variable
  fit_cox <- coxph(Surv(time, status) ~ uis_clean[[var]], data = uis_clean)
  #Extract LRT
  lrt <- summary(fit_cox)$logtest["test"]
  # Extract the p-value of the likelihood ratio test
  p_value <- summary(fit_cox)$logtest["pvalue"]

  results <- rbind(results, data.frame(Variable = var,
                                       LRT = lrt, P_Value = p_value))
}

sig_vars <- results$Variable[which(results$P_Value < 0.05)]
knitr::kable(results, format = "markdown",
             col.names = c("Variable", "LRT", "P-Value"), row.names = FALSE)

# 3.2

# Perform stepwise selection using AIC
selected_model <- stepAIC(coxph(Surv(time, status) ~ ., data = uis_clean),
                          direction = "both")

# Extract the subset of variables selected by stepwise selection
selected_variables <- names(selected_model$coefficients)

# Perform variable selection using glmulti with the selected variables
```

```r
result.bfit.ph <- glmulti(
  Surv(time, status) ~ .,
  data = uis_clean[, c("time", "status", selected_variables)],
  level = 1,
  method = "h",
  crit = "bic",
  confsetsize = 1,
  plotty = FALSE,
  report = FALSE,
  fitfunction = "coxph"
)

# Extract the best subset model from glmulti
bfit.ph <- result.bfit.ph@objects[[1]]

# Summarize the best subset model
summary(bfit.ph)




#Appendix: Code Results - Code

print("2.2: Logrank")
logrank

print("2.2: CoxPH")
summary(fit.cox)

print("2.3: CoxPH_Multi-Center")
summary(fit.cox_multi)

print("3.1: Variable Screening Purpose")
results
```