Survival Analysis part 2

Adam J Sullivan, PhD

03/21/2018

Cox-PH Regression

Physicians Health Study and Aspirin

Recall the Colorectal Cancer component of the Physicians Health Study

NAME	DESCRIPTION
age	Age in years at time of Randomization
asa	0 - placebo, 1 - aspirin
bmi	Body Mass Index (kg/m^2)
hypert	1 - Hypertensive at baseline, 0 - Not
alcohol	0 - less than monthly, 1 - monthly to less than daily, 2 - daily consumption

NAME	DESCRIPTION
dm	0 = No diabetes Mellitus, 1 - diabetes Mellitus
sbp	Systolic BP (mmHg)
exer	0 - No regular, 1 - Sweat at least once per week
csmoke	0 - Not currently, 1 - < 1 pack per day, 2 - $\geq$ 1 pack per day
psmoke	0 - never smoked, 1 - former < 1 pack per day, 2 - former $\geq$ 1 pack per day
pkyrs	Total lifetime packs of cigarettes smoked
crc	0 - No colorectal Cancer, 1 - Colorectal cancer
cayrs	Years to colorectal cancer, or death, or end of follow-up.

For this study each participant contributed 2 pieces of information during follow-up:

1. Information on whether of not they had a Colorectal Cancer(CRC) during follow-up
2. Follow-up time in years, specified as time from randomization until first of
    - end of Study
    - death
    - Colorectal Cancer
    - Loss to follow-up

We can load this data into R.

```
library(tidyverse)
library(haven)
phscrc <- read_dta("phscrc.dta")
phscrc <- phscrc %>% mutate(age.cat = cut(age, c(40, 50, 60,
    70, 90), right = FALSE)) %>% mutate(alcohol.use = factor(alcohol >
    1, labels = c("no", "yes"))) %>% mutate(obese = factor(bmi >
    30, labels = c("Not Obese", "Obese")))
```

Proportional Hazards Model

The general is

$$h(t|X_1, \ldots, X_p) = h_0(t)\exp(\beta_1 x_1 + \cdots + \beta_p x_p)$$

$$\text{or}$$

$$\log[h(t|X_1, \ldots, X_p)] = \log[h_0(t)] + \beta_1 x_1 + \cdots + \beta_p x_p$$

where $h_0(t)$ is the baseline hazard function and the "intercept" is $\log[h_0(t)]$.

Semi-Parametric Regression

- Weibull and Exponential are examples of parametric proportional hazards models, where $h_0(t)$ is a specified function.
- In 1972, Cox generalized these types of models so that we can make inferences on the $\beta_1, \ldots, \beta_p$ without specifying $h_0(t)$.
- We call Cox a semi-parametric regression model
- We fit this using something called
- Once again we use an algorithm to maximize the partial likelihood.

Interpeting the Model

Let

- $X = 0$ be the control group
- $X = 1$ be the treatment group

Then

$$\begin{aligned} h(t|X=x) &= h_0(t)\exp(\beta x) \\ h(t|X=0) &= h_0(t) \\ &= \text{ baseline hazard for control group} \\ h(t|X=1) &= h_0(t)\exp(\beta) \\ &= \text{ hazard for treated group} \\ \exp(\beta) &= \frac{h(t|X=1)}{h(t|X=0)} \end{aligned}$$

What Does This Mean?

- This means that the hazard ratio is constant over time (**Proportional Hazards**)
- β is the log hazard ratio or log-relative risk
- According to the Cox model

$$\log[h]h(t|X=0)] = \log[h_0(t)]$$
$$\log[h]h(t|X=1)] = \log[h_0(t)] + \beta$$

- This means the log of the hazard functions are parallel over time.
- We make no assumptions about $h_0(t)$.

Verifying Proportional Hazards Assumption

Recall

$$S(t) = \exp(-\Lambda(t))$$

with a binary X we have that

$$\begin{aligned} \Lambda_1(t) &= \Lambda_0(t)\exp(\beta) \\ S_0(t) &= \exp(\Lambda_0(t)) \\ -\log(S_0(t)) &= \Lambda_0(t) \\ \log(-\log(S_0(t))) &= \log(\Lambda_0(t)) \end{aligned}$$

$$\begin{aligned} S_1(t) = exp(-\Lambda_1(t)) &= \exp[\Lambda_0(t)\exp(\beta)] \\ -\log(S_1(t)) &= \Lambda_0(t)\exp(\beta) \\ \log(-\log(S_1(t))) &= \log(\Lambda_0(t)) + \beta \end{aligned}$$

Verifying Proportional Hazards Assumption

- Thus we can see that under the assumption of
  - $\log(-\log(K - M))$ should be parallel over time.
  - We typically verify this graphically.
  - Recall the CRC study:

Example: Kaplan-Meier Survival

```
library(survival)

model <- survfit(Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc,
    cayrs > 0))
model
```

Example: Kaplan-Meier Survival

```
library(survival)

model <- survfit(Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc,
    cayrs > 0))
model
```

Plotting the Kaplan-Meier

```
library(survminer)
ggsurvplot(model, legend.labs = c("Non-Drinker", "Drinker"),
    break.time.by = 2, fun = "cloglog")
```

Plotting the Kaplan-Meier

```
## Error in ggsurvplot(model, legend.labs = c("Non-Drinker", "Drinker"), : object 'model' not found
```

Cox PH in R

- We can run the Cox PH in R:

```
cox.crc <- coxph(Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc,
    cayrs > 0))
summary(cox.crc)
```

Cox PH in R

```
## Call:
## coxph(formula = Surv(cayrs, crc) ~ alcohol.use, data = subset(phscrc,
##     cayrs > 0))
##
##   n= 16018, number of events= 254
##
##                  coef exp(coef) se(coef)    z Pr(>|z|)
## alcohol.useyes 0.414     1.514    0.135 3.08   0.0021 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## alcohol.useyes      1.51      0.661      1.16      1.97
##
## Concordance= 0.541  (se = 0.013 )
## Rsquare= 0.001   (max possible= 0.262 )
## Likelihood ratio test= 8.97  on 1 df,   p=0.00275
## Wald test            = 9.48  on 1 df,   p=0.00208
## Score (logrank) test = 9.61  on 1 df,   p=0.00193
```

Interpretation

- This would suggest that the hazard of Colorectal Cancer for those who drink daily is 51% higher than those who drink less than daily.

Continuous Example of Cox PH

- Let's consider `age` and `smoking`:

```
library(survival)
crc.cox <- coxph(Surv(cayrs, crc) ~ csmok + age, data = subset(phscrc,
    cayrs > 0))
summary(crc.cox)
```

Continuous Example of Cox PH

```
## Call:
## coxph(formula = Surv(cayrs, crc) ~ csmok + age, data = subset(phscrc,
##     cayrs > 0))
##
##   n= 16018, number of events= 254
##
##          coef exp(coef) se(coef)     z Pr(>|z|)
## csmok 0.31715   1.37320  0.09767  3.25   0.0012 **
## age   0.07904   1.08224  0.00628 12.58   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## csmok      1.37      0.728      1.13      1.66
## age        1.08      0.924      1.07      1.10
##
## Concordance= 0.724  (se = 0.018 )
## Rsquare= 0.01   (max possible= 0.262 )
## Likelihood ratio test= 160  on 2 df,   p=0
## Wald test            = 163  on 2 df,   p=0
## Score (logrank) test = 177  on 2 df,   p=0
```

Interpretation

- Then we could say that for two people with the same smoking status a one year increase in age would lead to an 8.2% increase in the hazard of colorectal cancer with a 95% CI of 6.9% to 9.6%.
- We would also be able to say that for 2 people the same age, a person who is a current smoker would have a 37% increase in hazard of colorectal cancer than a non smoker.

Cox-PH Regression

Assessing Diagnostics and Model Fit with Cox-PH

- With the Cox PH model we will consider 2 things
  - Checking Proportional Hazards Assumption
  - Checking for Influential Observations

The Data

- We will consider Recidivism of 432 male patients.
- They all were observed for 1 year prior to release from prison.
- The following slides will contain the variables.

Variables

VARIABLE	DESCRIPTION
week	week of first arrest after release, or censoring time.
arrest	the event indicator, equal to 1 for those arrested during the period of the study and 0 for those who were not arrested.
fin	a factor, with levels yes if the individual received financial aid after release from prison, and no if he did not; financial aid was a randomly assigned factor manipulated by the researchers.

Variables

VARIABLE	DESCRIPTION
age	in years at the time of release.
wexp	a factor with levels yes if the individual had full-time work experience prior to incarceration and no if he did not.
mar	a factor with levels married if the individual was married at the time of release and not married if he was not.
paro	a factor coded yes if the individual was released on parole and no if he was not.
prior	number of prior convictions.

Variables

VARIABLE	DESCRIPTION
educ	education, a categorical variable coded numerically, with codes 2 (grade 6 or less), 3 (grades 6 through 9), 4 (grades 10 and 11), 5 (grade 12), or 6 (some post-secondary).6
emp1 – emp52	factors coded yes if the individual was employed in the corresponding week of the study and no otherwise.
race	a factor with levels black and other.

Reading Data in

```
url <- "http://socserv.mcmaster.ca/jfox/Books/Companion/data/Rossi.txt"
Rossi <- read.table(url, header = TRUE)
```

Our Model

```
library(survival)
mod1 <- coxph(Surv(week, arrest) ~ fin + age + race + wexp +
    mar + paro + prio, data = Rossi)
tidy1 <- tidy(mod1, exponentiate = T)
knitr::kable(tidy1[-c(3, 4)])
```

Our Model

```
## Error in tidy(mod1, exponentiate = T): could not find function "tidy"
```

```
## Error in inherits(x, "list"): object 'tidy1' not found
```

Plotting Regression

```
library(ggplot2)
library(ggfortify)
autoplot(survfit(mod1), surv.linetype = "dashed", surv.colour = "blue",
    conf.int.fill = "dodgerblue3", conf.int.alpha = 0.5, censor = FALSE)
```

Plotting Regression

```
## Error in library(ggfortify): there is no package called 'ggfortify'
```

```
## Error: Objects of type survfit.cox/survfit not supported by autoplot.
```

Checking Proportional Hazards

- We have tested these before with Schoenfeld Residuals
- We can do this with the `cox.zph()` function.

Checking Proportional Hazards

```
cox.zph(mod1)
```

```
##                       rho    chisq        p
## finyes            0.00646  0.00502 0.943519
## age              -0.26455 11.27897 0.000784
## raceother         0.11224  1.41652 0.233977
## wexpyes           0.22976  7.14021 0.007537
## marnot married   -0.07295  0.68627 0.407435
## paroyes          -0.03618  0.15496 0.693841
## prio             -0.01366  0.02304 0.879353
## GLOBAL                 NA 17.65862 0.013609
```

Conclusion

- We see there is an issue
  - `age` is an issue
  - `wexp` is an issue as well.
- What do we do???

Enter Stratification

- We can adjust for a variable that does not meet the proportional hazards assumption by stratification.
- Assume we have Z which does not allow for PH

$$h(t|X, Z = j) = h_j(t)exp(X\beta)$$

- $j = 1, ldots, C$ levels of Z.

Create Age Categories

```
Rossi$age.cat <- cut(Rossi$age, c(0, 19, 25, 30, Inf))
xtabs(~age.cat, data = Rossi)
```

```
## age.cat
##   (0,19]  (19,25]  (25,30] (30,Inf]
##       66      236       66       64
```

Re Run the Model

```
mod2 <- coxph(Surv(week, arrest) ~ fin + race + mar + paro +
    prio + strata(wexp, age.cat), data = Rossi)
tidy2 <- tidy(mod2, exponentiate = T)
```

```
## Error in tidy(mod2, exponentiate = T): could not find function "tidy"
```

```
knitr::kable(tidy2[-c(3, 4)])
```

```
## Error in knitr::kable(tidy2[-c(3, 4)]): object 'tidy2' not found
```

PH Assumption

```
cox.zph(mod2)
```

```
##                    rho  chisq     p
## finyes         -0.0147 0.0252 0.874
## raceother       0.1086 1.3066 0.253
## marnot married -0.0794 0.8033 0.370
## paroyes        -0.0112 0.0141 0.906
## prio           -0.0174 0.0326 0.857
## GLOBAL              NA 2.3420 0.800
```

Influential Observations

- to test this let's build a smaller model

```
mod3 <- coxph(Surv(week, arrest) ~ fin + prio + strata(age.cat),
    data = Rossi)
```

Influential Observations

- With Cox PH we will use DFBETA to tell.
- DFBETA's measure how much an observation has effected the estimated coefficient.
- We look for values to be under $\frac{2}{\sqrt{n}}$.

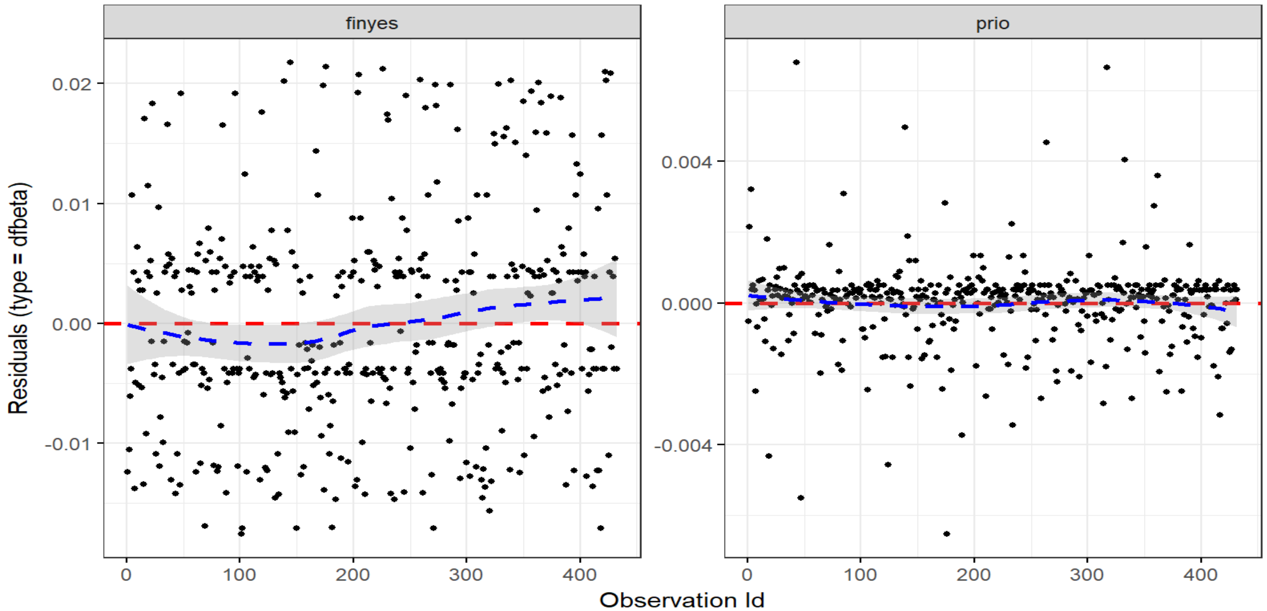
DFBETA in R

```
library(survminer)
2/sqrt(dim(Rossi)[1])
```

```
## [1] 0.0962
```

```
ggcoxdiagnostics(mod3, type = "dfbeta", linear.predictions = FALSE,
    ggtheme = theme_bw())
```

DFBETA in R


finyes
prio
Residuals (type = dfbeta)
Observation Id
0.02
0.01
0.00
-0.01
0.004
0.000
-0.004
0
100
200
300
400