University of Illinois
Fall 2001

Department of Economics
Roger Koenker

# Economics 478
# Lecture 6
# Time Dependent Covariates in the Cox Proportional Hazard Model

We have seen how fitting of the Cox model works in simple cases with "baseline covariates," i.e., with covariates that are fixed over the study period. But there are many applications for which the covariates change over the duration of the study period. These cases are handled naturally within the counting process framework. I will try to illustrate practical aspects of this approach following closely the analysis of the Crowley and Hu (1977) Stanford Heart Transplant Data provided by Therneau and Grambsch.

In Table 1 we illustrate the format of the original data from Crowley and Hu (1977). There are 103 patients the table gives the data for the first 10 of them. There are two types of patients: those who have received a transplant, and those who were enrolled in the study and therefore expected to receive a transplant, but never did. In this case treatment itself is a time dependent covariate. Survival times are measured from the point of enrollment and response to treatment is expected to be influenced by how promptly the transplant occurred.

A crucial aspect of the analysis is the transformation of the original data into a format consistent with the counting process formulation. In Table 2 we illustrate the transformed data for the first 10 patients. The first thing to note is that the 10 original lines has become 14; each of the patients receiving a transplant, i.e., patients $\{3, 4, 7, 10\}$ now have 2 lines rather than one. Instead of the conventional data structure $(y_i, \delta_i, x_i)$ giving event times, censoring indicator and covariate vector we now have $((t_{ij-1} - t_{ij}], \delta_{ij}, x_{ij})$ where $(t_{ij-1} - t_{ij}]$ denotes an interval over which the covariate vector $x_{ij}$ prevailed for patient $i$. In the present example, for transplant patients we have two such records for each patient; one corresponding to the time interval between enrollment and the transplant, the other to the interval between the transplant and either death or the point at which the patient is lost to follow up. In the terminology of Therneau's survival software we have the data format

(start, stop], status, covariates

as indicated in Figure 2. In more complicated settings with more frequent changes in the covariates we would have more multiple lines for each patient. For patients who did not receive a transplant we have only a single line. In

| Patient | Date of birth | Date of acceptance | Date of transplant | Date last seen | Dead=1 Alive=0 | Previous surgery |
|---|---|---|---|---|---|---|
| 1 | 1/10/37 | 11/15/67 | | 1/3/68 | 1 | 0 |
| 2 | 3/2/16 | 1/2/68 | | 1/7/68 | 1 | 0 |
| 3 | 9/19/13 | 1/6/68 | 1/6/68 | 1/21/68 | 1 | 0 |
| 4 | 12/23/27 | 3/28/68 | 5/2/68 | 5/5/68 | 1 | 0 |
| 5 | 7/28/47 | 5/10/68 | | 5/27/68 | 1 | 0 |
| 6 | 11/8/13 | 6/13/58 | | 6/15/68 | 1 | 0 |
| 7 | 8/29/17 | 7/12/68 | 8/31/68 | 5/17/70 | 1 | 0 |
| 8 | 3/27/23 | 8/1/68 | | 9/9/68 | 1 | 0 |
| 9 | 6/11/21 | 8/9/68 | | 11/1/68 | 1 | 0 |
| 10 | 2/9/26 | 8/11/68 | 8/22/68 | 10/7/68 | 1 | 0 |

TABLE 1. Original Form of the Stanford Heart Transplant Data: First 10 patients

| | start | stop | event | age | year | surgery | transplant | id |
|---|---|---|---|---|---|---|---|---|
| 1 | 0.0 | 50.0 | 1 | −17.15537303 | 0.12320329 | 0 | 0 | 1 |
| 2 | 0.0 | 6.0 | 1 | 3.83572895 | 0.25462012 | 0 | 0 | 2 |
| 3 | 0.0 | 1.0 | 0 | 6.29705681 | 0.26557153 | 0 | 0 | 3 |
| 4 | 1.0 | 16.0 | 1 | 6.29705681 | 0.26557153 | 0 | 1 | 3 |
| 5 | 0.0 | 36.0 | 0 | −7.73716632 | 0.49007529 | 0 | 0 | 4 |
| 6 | 36.0 | 39.0 | 1 | −7.73716632 | 0.49007529 | 0 | 1 | 4 |
| 7 | 0.0 | 18.0 | 1 | −27.21423682 | 0.60780287 | 0 | 0 | 5 |
| 8 | 0.0 | 3.0 | 1 | 6.59548255 | 0.70088980 | 0 | 0 | 6 |
| 9 | 0.0 | 51.0 | 0 | 2.86926762 | 0.78028747 | 0 | 0 | 7 |
| 10 | 51.0 | 675.0 | 1 | 2.86926762 | 0.78028747 | 0 | 1 | 7 |
| 11 | 0.0 | 40.0 | 1 | −2.65023956 | 0.83504449 | 0 | 0 | 8 |
| 12 | 0.0 | 85.0 | 1 | −0.83778234 | 0.85694730 | 0 | 0 | 9 |
| 13 | 0.0 | 12.0 | 0 | −5.49760438 | 0.86242300 | 0 | 0 | 10 |
| 14 | 12.0 | 58.0 | 1 | −5.49760438 | 0.86242300 | 0 | 1 | 10 |

TABLE 2. Transformed Form of the Stanford Heart Transplant Data: First 10 patients. Note that age is in decimal years -40 and the event times are in days.

the entire data set there are 69 transplants of 103 total patients so the transformed data has $2.69 + 44 = 172$ lines.

Given the new form of the data we have effectively broken the event times for each patient into several subintervals over which the model may wish to assign different hazards. Obviously, if the time varying covariate has a zero effect, then the estimated model is entitled to say so. The new data structure

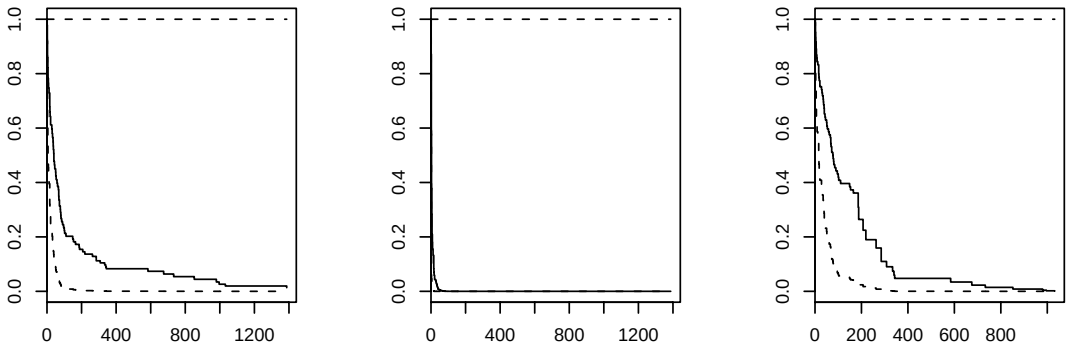


FIGURE 1. Three hypothetical survival functions for the Stanford Heart Transplant Analysis.

is accommodated by the model fitting software by expanding the usual call to

Surv (start, stop, status).

So for example we can illustrate the fitting with the following fragment of code and output from *R*.

```
#Extended Example  based on Stanford Heart Transplant data
library(survival)
data(heart)
fit1_coxph( Surv(start,stop,event)~(age + surgery)*transplant,
data=heart, method='breslow')
#fit1
#summary(fit1)
#Now try to plot survival curves first for patients who never get a transplant
#Next for patients who don't have prior surgery and get a transplant immediately
#Note that there are no such patients in the sample so the latter is suspect
#Note that this syntax conflicts slightly with advice of p 51 of Therneau(1999)
fit1.1_survfit(fit1,data.frame(age=50,surgery=0,transplant=0))
fit1.2_survfit(fit1,data.frame(age=50,surgery=0,transplant=1))
#Now consider a more complicated case
#Joe is a hypothetical patient age 50 with prior surgery who gets a transplant
#after 6 months (183 days).  Note that the event argument is ignored and the
#result of survfit estimates the  survival curve for Joe under above scenario
Joe_data.frame(start=c(0,183),stop=c(183,3*365),event=c(1,1),age=c(50,50),
surgery=c(1,1),transplant=c(0,1))
```

```
fit1.3_survfit(fit1,Joe,individual=T)
#Now plot all three fits
postscript("fig1.ps", horizontal=F,width=7,height=3)
par(mfrow=c(1,3))
plot(fit1.1)
plot(fit1.2)
plot(fit1.3)
```

A point that is emphasized in Therneau and Grambsch and also in the useful technical report that Therneau has written to accompany his software is the delicate nature of the interpretation of "fitting survival functions" under time varying covariate schemes. Now the survival function is no longer conditioned simply on a vector of baseline covariates, but an entire time path of these covariates. So in order to plot survival curves it is necessary to specify such time paths, and as usual, it is important to restrict such hypothetical exercises to regions of covariate space that are within the realm of possibility given the data. This is particularly problematic in cases where there is more frequent monitoring of the time varying covariate. This is all illustrated in the fragment of $R$ code for the Stanford data. The figure illustrate three survival curves for hypothetical patients. Note in the last fit1.3 construction we have a survival curve for a more complicated case.

*Postscript*

After a heated discussion at coffee on the morning following this lecture, Xuming He made a strong case for questioning the interpretation of the time-varying covariate version of the Cox model. In particular, when we estimate a model with only baseline covariate we have a clear prognostic view of the survival prospects viewed from time zero. In the Cox model with time varying covariate we have as of time zero, to compute various survival curves for various sample paths of the covariates and then combine these in some way to obtain an unconditional result , i.e., conditional only on baseline covariates, in order to get to a valid prognostic object. From this perspective I would like to raise the question: Suppose we consider a quantile regression model – of the sort used in the medfly paper – for the Stanford Heart Transplant Data. This model will have two pieces that are completely disconnected (this immediately sounds fishy ...). One piece would use all the data for those before transplant, obviously the transplant patient would appear in this data set as censored at the time of transplant. The other piece would analyze the transplant patients as if time zero was the time of transplant. My questions are: first, how does this model correspond to the Cox model discussed earlier, and second, how does it correspond to an extended version of the Cox model that we didn't discuss, but probably should have, in which separate baseline hazards are estimated for treatment and won treatment observations? What about using the time to transplant as a covariate in the second model.