PHST 683 – Survival Analysis
Course Project – Option 2

OPTION 2 – Construction and Analysis of a Time-Dependent Covariate Data Set

The department of Child Protective Services (CPS) in a certain metropolitan area is pilot testing a new program for victims of physical child abuse. The goal of the program is to reduce the rate at which future instances of physical child abuse occurs. After an "index event", i.e. a call to CPS for a suspected occurrence of abuse or neglect, children experiencing the index event and their families were randomized to receive the pilot intervention or to receive a standard intervention. The pilot intervention involves regular, expanded family counseling and access to a 24-hour telephone hotline which provides point-of-care counseling.

After the index event, investigators tracked children enrolled in the study for up to 1 year and noted the occurrence of secondary instances of physical child abuse, which serves as the response variable for this analysis. Children not experiencing physical child abuse within 1 year or who moved away were right censored. Investigators also measured a number of demographic characteristics of each child enrolled and certain household risk factors for physical abuse. Investigators also measured whether or not each child was removed from the home and placed in a foster home and the time at which the removal occurred for those children. For children removed from the home at any time during the year of observation, investigators also measured whether or not the child was returned from home within the year and the time of return to home.

The data are available in the data set **dat** within the file **AbuseStudy.Rdata** posted to Blackboard. A listing and description of the included variables is provided below:

Variable	**Description**
id	Child identifier
time	Time to secondary occurrence of abuse (days)
event	Indicator of secondary occurrence of abuse
program	Program of enrollment (Pilot, Standard)
age	Age of child at enrollment
sex	Biological sex of child (Male, Female)
minority	Indicator of child being of minority race
poverty	Indicator of family living below poverty threshold
subst.abuse	Indicator of substance abuse in the household
crim.hist	Indicator of criminal history in the household
first	Indicator that index event was first ever CPS contact for child
region	Region of metropolitan area (North, South)
agency	Agency contact for index event (Local, State)
removal	Indicator of child being removed from home to foster care
time.removal	Time of removal from home to foster care (days)
return	Indicator of child being returned to home from foster care
time.return	Time of return to home from foster care (days)

The investigators are primarily interested in comparing the risk of secondary occurrences of physical child abuse, but want to control for other factors measured in the data set. As the study statistician, your job is to conduct such an analysis.

Your first task is to transform this data set to a time-dependent data set. Specifically, the variables measuring removal from and return to the home define a time-dependent covariate that can be called "outofhome". This variable needs to be an indicator taking value 1 when a child is outside of the home. Note that all children begin the study in the home and were only removed from/returned to the home once in the year of observation (there were no instances of multiple removals/returns). Here is an example of such a translation (not real data):

Original Data

id	time	event	removal	time.removal	return	time.return
1	292	1	1	35	0	292
2	365	0	1	172	1	215
3	32	1	0	32	0	32
4	97	1	1	15	1	62

New (Time-Dependent) Data

id	start	stop	outofhome	event
1	0	35	0	0
1	35	292	1	1
2	0	172	0	0
2	172	215	1	0
2	215	365	0	0
3	0	32	0	1
4	0	15	0	0
4	15	62	1	0
4	62	97	0	1

Some notes on the translation:

- Child 1 was in the home until day 35, was removed from the home, and stayed out of the home until day 292, when a secondary instance of abuse occurred out of the home
- Child 2 was in the home until day 172, was removed from the home, returned to the home on day 215, and did not experience abuse before day 365
- Child 3 experienced a secondary instance of abuse in the home on day 32, with no removal from/return to the home
- Child 4 was in the home until day 15, was removed from the home, returned on day 62, and then experienced the event in the home on day 97

Here's a useful snippet of code for building the time-dependent data set, which creates the time intervals and time-dependent covariate for the first row of the data frame:

```
start <- 0
stop <- numeric(length=0)
outofhome <- 0
event <- numeric(length=0)
if (dat$removal[1]==1) {
   start <- c(start, dat$time.removal[1])
   stop <- c(stop, dat$time.removal[1])
   outofhome <- c(outofhome,1)
   event <- c(event, 0)
}
if (dat$return[1]==1) {
   start <- c(start, dat$time.return[1])
   stop <- c(stop, dat$time.return[1])
   outofhome <- c(outofhome,0)
   event <- c(event, 0)
}
stop <- c(stop, dat$time[1])
event <- c(event, dat$event[1])
temp.frame <- data.frame(id=dat$id[1],start,stop,event,outofhome)
```

Play with this code to make sure you understand how it works. You can use this code to build the time-dependent data set. For reference, examine the details of the construction of the time-dependent BMT data set from Week 10 to see how the above code can be used to create the time dependent data set.

After you've built the time-dependent version of the data set, you'll need to conduct analyses to address the following on behalf of the investigators:

- Were the intervention groups (pilot and standard) balanced with respect to the other measured covariates? (*NOTE: This will require standard (i.e. non-survival) methods*.)
- Was the risk of secondary occurrences of abuse lower for children enrolled in the pilot program?
- What other measured covariates were associated with the risk of secondary occurrences of abuse?
- Did the intervention factor significantly interact with any covariates found to be related to the risk of secondary occurrences of abuse?

One important note on the analysis is that the region of the metropolitan area and the agency to which the index event were reported are not of interest as covariates to the investigators and should be included only as stratifying variables.

Please follow these important instructions while building your models (to keep things simpler):

1. You may use any automated variable selection process in R that you like (if it helps you).

2. First, build an initial model containing only main effects, i.e. no interactions. Once you've identified covariates related to the outcome, then start considering interactions.
3. Consider only interactions between the intervention factor and other included covariates.

Produce a report of your findings. **DO NOT SUBMIT AN R MARKDOWN FILE AS YOUR REPORT!!** You should write the report as though it is to be submitted to the investigators on the study (not to your instructor), whom you should consider a lay audience. There are no length requirements for the report, but it should be double-spaced and provide clear and full explanations of the results of your analyses, along with any tables and graphs of your results as you see fit, to highlight key findings from your analysis. Include your R code by which you built your model as an appendix to your report.