Exploratory Analysis

Alex Brown,David Geyfman, Jason Freeberg, Faraz Farooq

November 4, 2016

Introduction

Our dataset comes from a longitudinal study conducted in the United States. Researchers observed the time until divorce of 3371 couples, and tracked three covariates listed below. The data's time variable is measured in years with up to three decimals of precision. The event indicator is labeled 0 for censorship, and 1 for divorce.

- The husband's education level, coded as...
  - 0 -> less than 12 years (only high school)
  - 1 -> 12 to 15 years (only bachelors or equivalent)
  - 2 -> 16 or more years (some form of graduate studies)
- The husband's race, coded as...
  - 1 if the husband is black
  - 0 otherwise
- Whether or not both partners are black, coded as...
  - 1 if both partners are **not** black
  - 0 if one parter is not black, and the other is

Research question

Our team wanted to see how education affected time until divorce while controlling for the races of the couples.

Methodology

Our team decided to build our model bottom-up, meaning we fit a single covariate initially and added additional significant covariates.

Exploratory Analysis of Covariates

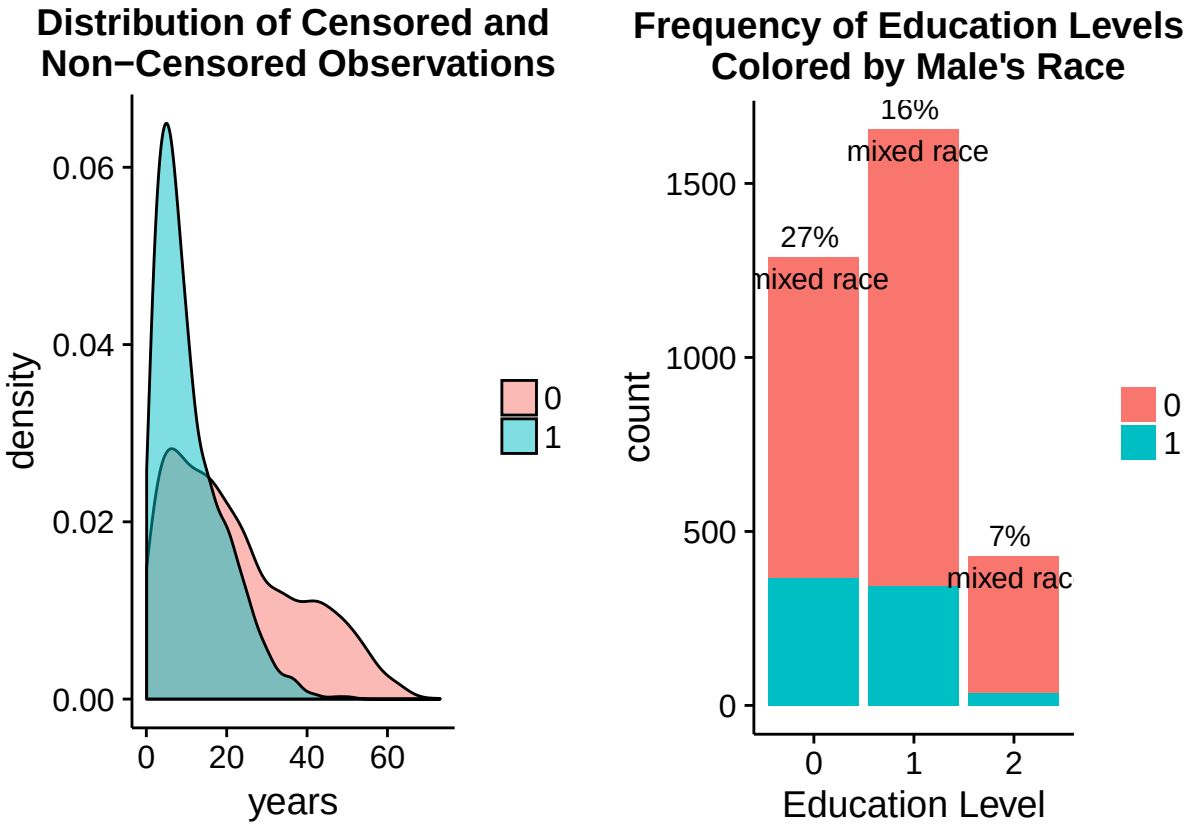
The following plots explore the distributions of our covariates.

```
#loading necessary libraries
library(ggplot2)
library(survival)
library(cowplot)
library(devtools)
library(ggkm)

# Uncomment and
# devtools::install_github("sachsmc/ggkm")
# https://github.com/sachsmc/ggkm

# Load data
colNames <- c("id", "edu", "hblack", "mixed", "years", "div")
```

```
divorce <- read.table(file = "divorce.txt", header = F, col.names = colNames)
head(divorce)
```

```
##   id edu hblack mixed  years div
## 1  9   1      0     0 10.546   0
## 2 11   0      0     0 34.943   0
## 3 13   0      0     0  2.834   1
## 4 15   0      0     0 17.532   1
## 5 33   1      0     0  1.418   0
## 6 36   0      0     0 48.033   0
```

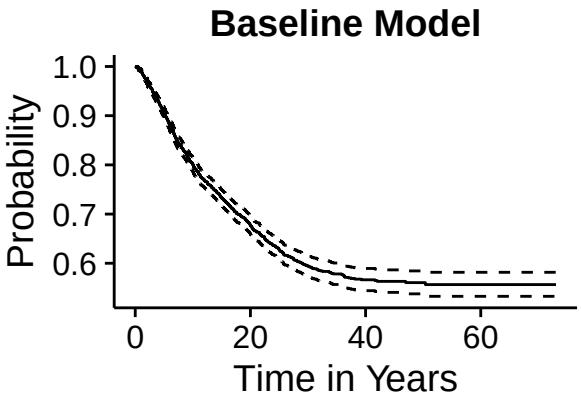
We turn the covariates into factors so we can group the covariate values as categories. In order to focus on education we wanted to treat race as a controlled variable, instead of having two covariates involving race, we made a new column that fully encodes the couple's racial makeup.

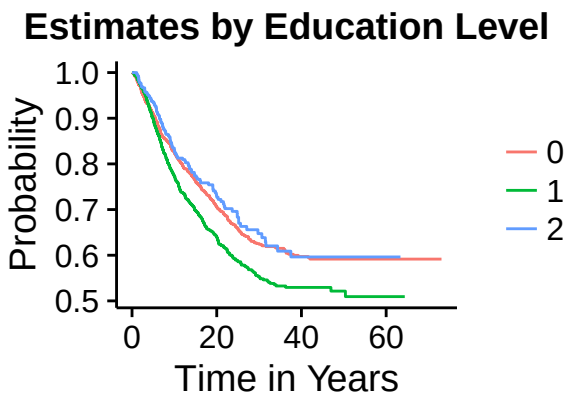


Left We can see that the bulk of our observed divorces, colored in blue, occur around the ten year mark. Our censored observations, colored in orange, drop steeply at 30 years into the study.

Right This plot shows the frequency of different education levels in our data. The bars are colored by their proportion of black husbands within the education level. So we can see that our data is largely composed of couples with husbands that are not African American. Lastly, the percentages over each bar report the percent of mixed race coupled within their respective education bracket. For example, 27% of couples in education bracket **0** are mixed.

Exploratory Survival Curves





Plot analysis clockwise from top left.

Top Left Thanks to the large number of observations, our KM estimate is nearly a smooth line.

Top Right This plot is especially interesting, as one would expect there to be a linear relationship between the education levels and the respective hazard rates... for example, $S_1(X) < S_2(X) < S_3(X)$ (where S_i denotes the survival function of the ith education level). Interestingly, we do not see this trend. The survival rates of couples with a college-educated husband are visibly lower than that of couples with either a high-school educated or graduate-educated husband.

Bottom Right

Bottom Left Non-mixed couples are likely to survive longer, shown by their higher survival function and that their observations carried on longer into the study.

Modeling

Here we fit our base model using only the covariate encoding the husband's education level.

```
## Call:
## coxph(formula = Surv(years, div) ~ edu, data = divorce)
##
##   n= 3371, number of events= 1032
##
##          coef exp(coef) se(coef)      z Pr(>|z|)
## edu1  0.23880   1.26973  0.06692  3.568 0.000359 ***
```

```
## edu2 -0.07784   0.92512  0.10799 -0.721 0.471063
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## edu1    1.2697     0.7876    1.1136     1.448
## edu2    0.9251     1.0809    0.7486     1.143
##
## Concordance= 0.535  (se = 0.009 )
## Rsquare= 0.005   (max possible= 0.99 )
## Likelihood ratio test= 17.36  on 2 df,   p=0.0001699
## Wald test            = 17.29  on 2 df,   p=0.000176
## Score (logrank) test = 17.39  on 2 df,   p=0.0001674
```

Our base model is significant ..... **YAY, ELABORATE ON OUTPUT HERE**

Controling for Racial Makeup

```
edu_race_coxph <- coxph(Surv(years, div) ~ edu + couple, data = divorce)

summary(edu_race_coxph)
```

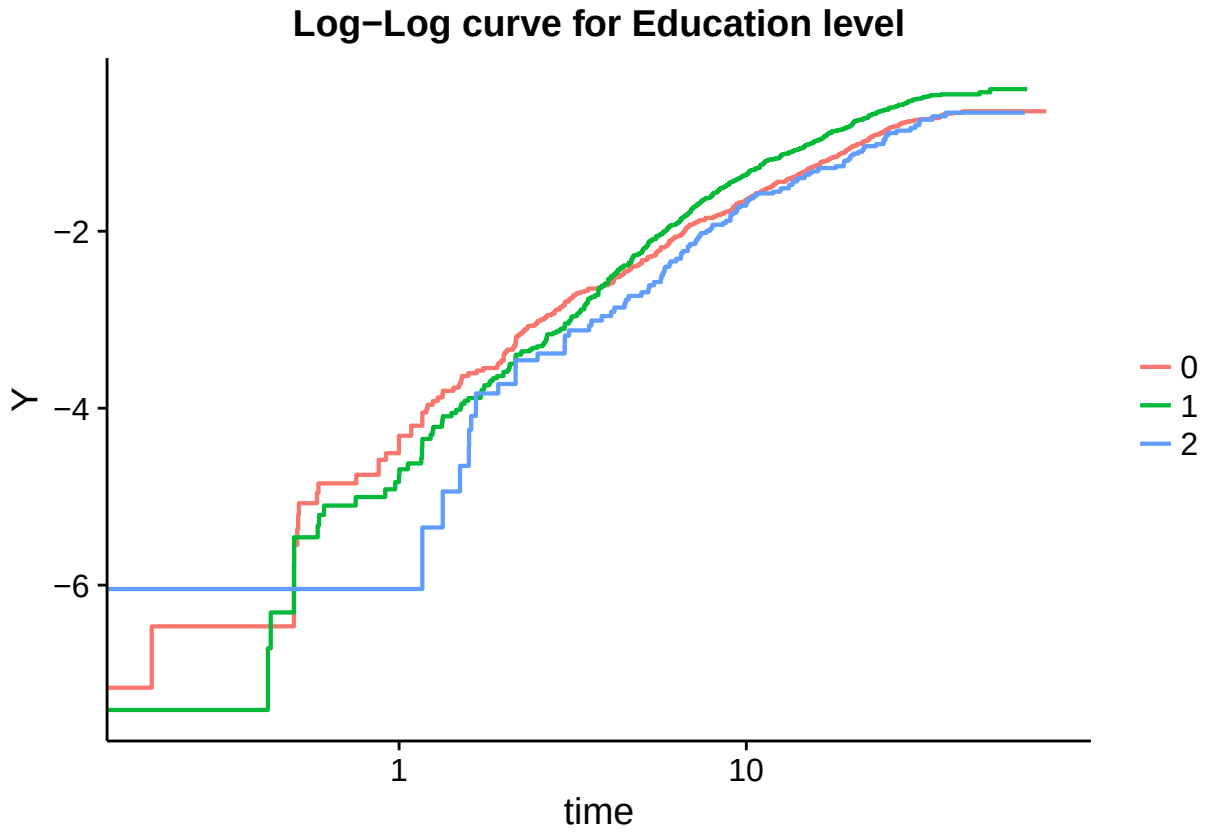
```
## Call:
## coxph(formula = Surv(years, div) ~ edu + couple, data = divorce)
##
##   n= 3371, number of events= 1032
##
##              coef exp(coef) se(coef)      z Pr(>|z|)
## edu1      0.29358   1.34122  0.06834  4.296 1.74e-05 ***
## edu2      0.02362   1.02390  0.11120  0.212   0.8318
## coupleBO  0.21434   1.23905  0.13231  1.620   0.1052
## coupleOB  0.05093   1.05225  0.12645  0.403   0.6871
## coupleOO -0.19430   0.82341  0.09975 -1.948   0.0514 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## edu1        1.3412     0.7456    1.1731     1.533
## edu2        1.0239     0.9767    0.8234     1.273
## coupleBO    1.2390     0.8071    0.9560     1.606
## coupleOB    1.0522     0.9503    0.8213     1.348
## coupleOO    0.8234     1.2145    0.6772     1.001
##
## Concordance= 0.547  (se = 0.009 )
## Rsquare= 0.011   (max possible= 0.99 )
## Likelihood ratio test= 35.77  on 5 df,   p=1.055e-06
## Wald test            = 36.51  on 5 df,   p=7.495e-07
## Score (logrank) test = 36.73  on 5 df,   p=6.794e-07
```

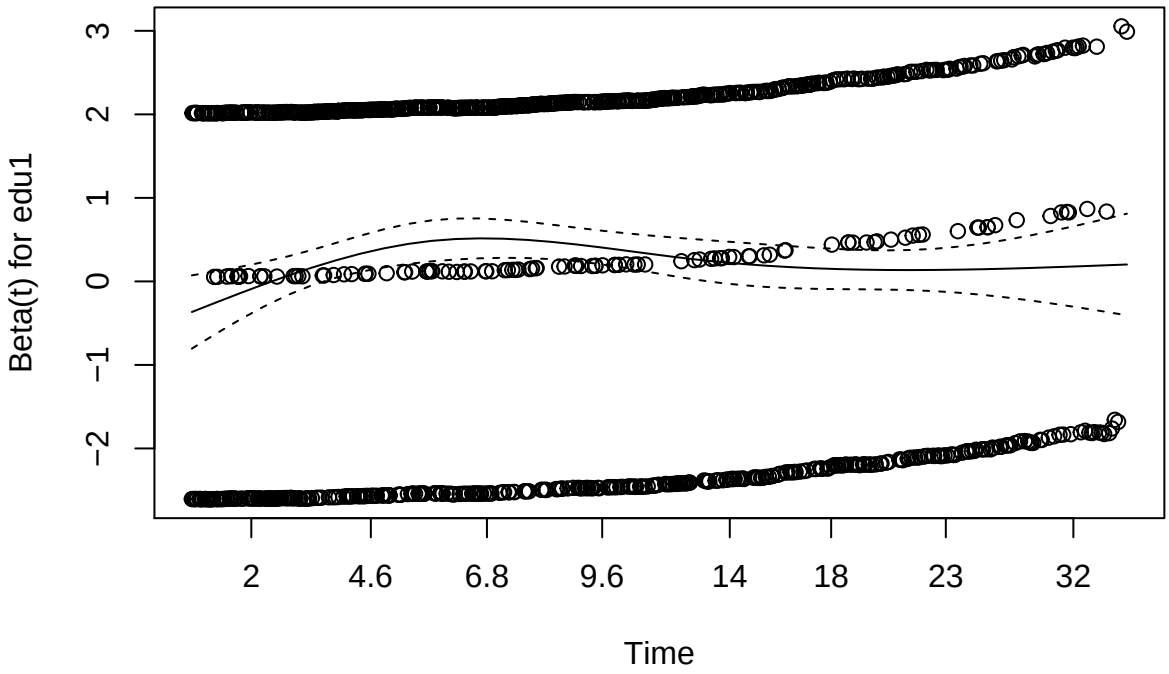
Elaborate on the P-values of the education factors, note that the whole model is still significant.

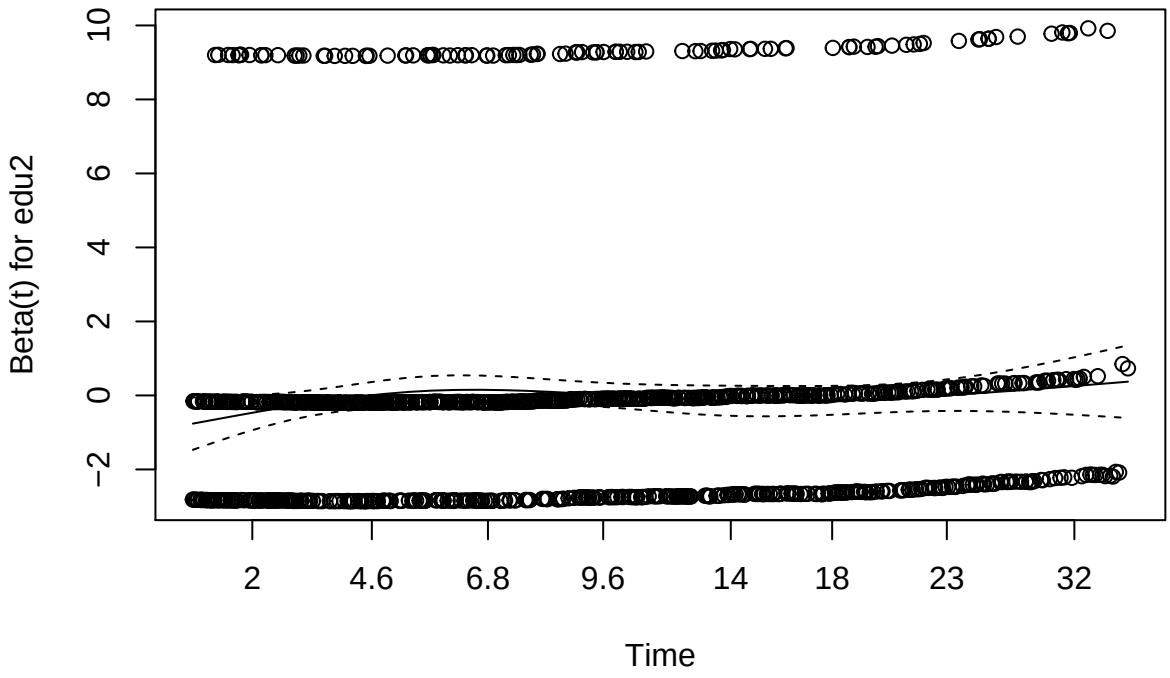
Checking Proportional Hazard Assumptions

Graphical Approach

```
## Warning: Transformation introduced infinite values in continuous x-axis
```




Beta(t) for edu1
3
2
1
0
−1
−2
2
4.6
6.8
9.6
14
18
23
32
Time



The plot **is** concerning because some of the lines intersect.

.... ANALYSE THE SCATTERPLOT OF SCHOENFELD RESIDUALScix

Numerical Approach

```
summary(edu_zph)
```

```
##           Length Class  Mode
## table        9   -none- numeric
## x         1032   -none- numeric
## y         2064   -none- numeric
## var          4   -none- numeric
## call         2   -none- call
## transform    1   -none- character
```