

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 partner is not black, and the other is

We also encoded two new variables using the above data, *wblack* and *couple*.

- The wife's race, coded as...
 - 1 if the wife is black
 - 0 otherwise
- The couple's racial makeup, encoded as...
 - BB if both are black
 - BO if the husband is black, and the wife is not
 - OB if the wife is black, and the husband is not
 - OO if both are not black

Research question

Our team was primarily interested in how the husband's highest education level affects time until divorce, using the various racial covariates as controlling variables. Our secondary goal was to investigate the influence of racial makeups as a stratified variable.

Methodology

After an exploratory analysis, our team began modeling the data with a base Cox-PH model including only the husband's education level. We then constructed various control models using the racial covariates. After establishing that the base model was still significant under the controls, we turned our interest to the possible influence of these racial predictors as stratified variables. Finally, our team investigated the necessity of a time-transform to compensate for the near non-proportionality between two levels of the education predictor. Sadly, the time-transformation offered no additional insight.

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)

# ggkm is not on CRAN, using devtools we can install it from GitHub
# 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)

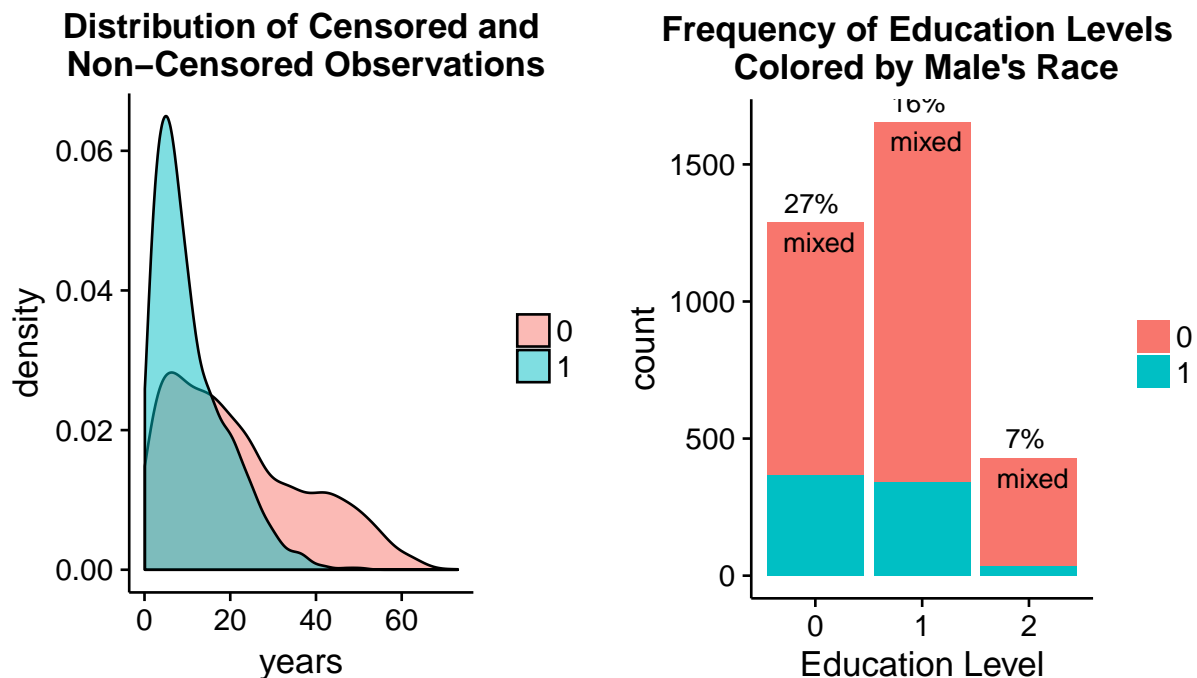
# Coerce numeric variables
divorce$edu <- as.factor(divorce$edu)
divorce$hblack <- as.factor(divorce$hblack)
divorce$mixed <- as.factor(divorce$mixed)

# Encode race variables
divorce <- femalecol(divorce)
divorce <- couple_column(divorce)

head(divorce)
```

```
##   id edu hblack mixed  years div wblack couple
## 1  9   1      0      0 10.546  0      0      00
## 2 11   0      0      0 34.943  0      0      00
## 3 13   0      0      0  2.834  1      0      00
## 4 15   0      0      0 17.532  1      0      00
## 5 33   1      0      0  1.418  0      0      00
## 6 36   0      0      0 48.033  0      0      00
```

The functions `femalecol()` and `couple_column()` encode the wife's race and the couple's racial makeup respectively. Their levels are explained in the Introduction above.

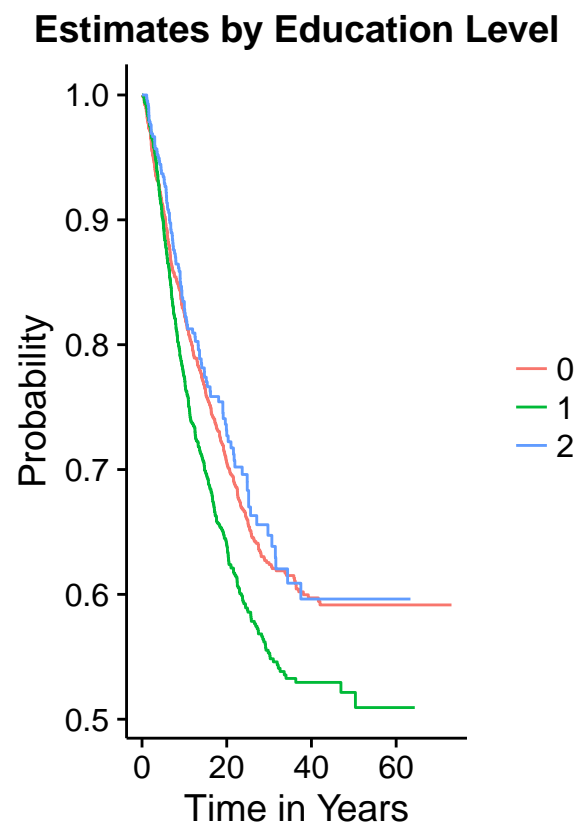
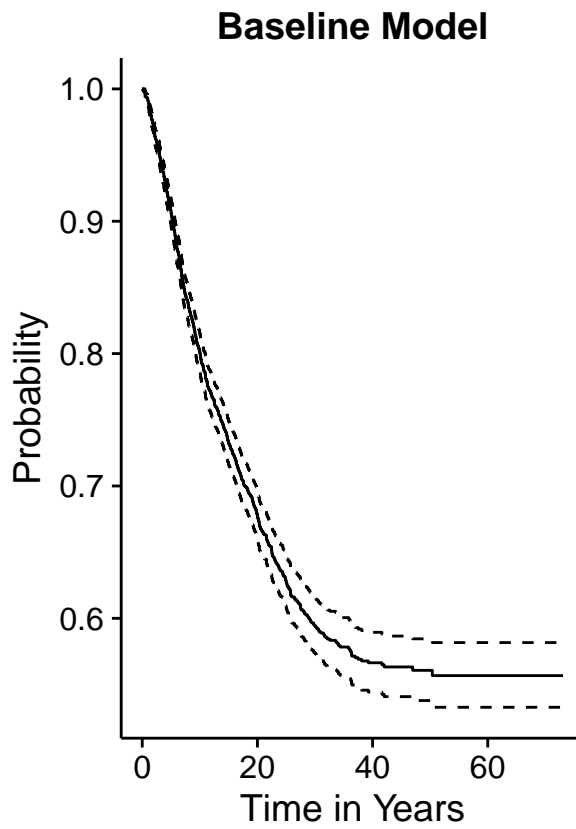


Plot analysis from left to right

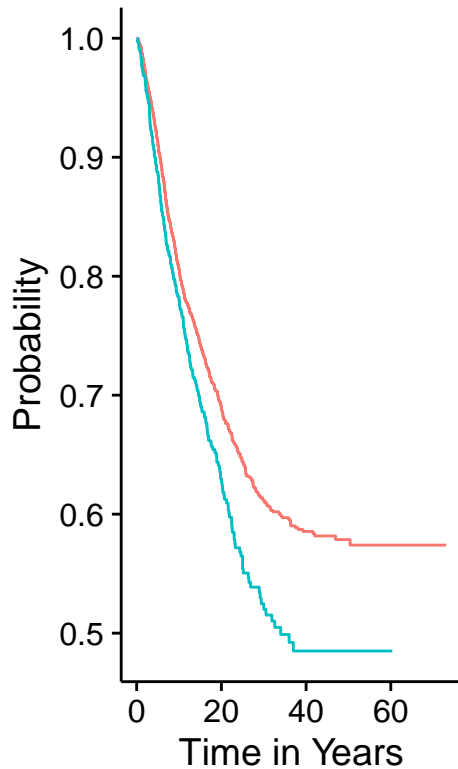
Fig. 1 We can see that the bulk of our observed divorces, colored in blue, occur around the ten year mark. Our censored observations, colored in orange, drop steeply at 30 years into the study.

Fig. 2 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.

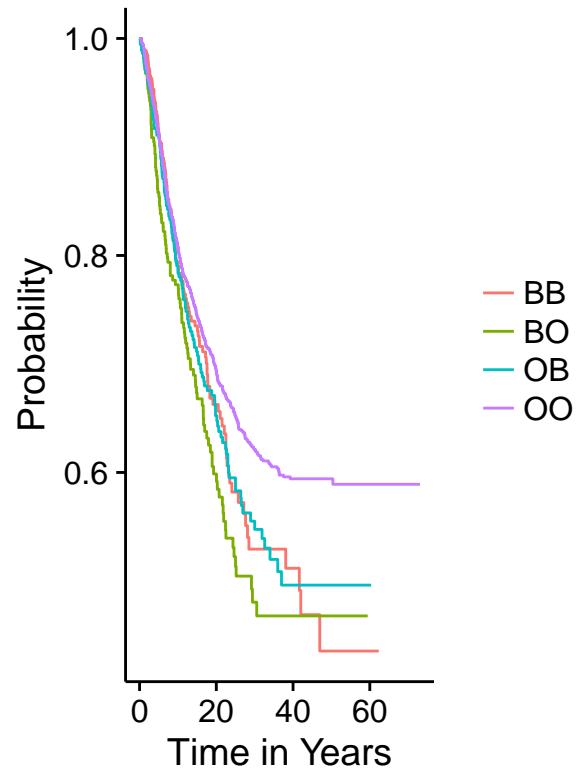
Exploratory Survival Curves



Estimates of Mixed Race Couples



Estimates by Racial Makeup



Plot analysis clockwise from top left.

Fig. 3 Thanks to the large number of observations, our KM estimate is nearly a smooth line.

Fig. 4 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 i th 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.

Fig. 5 The *couple* predictor gives us a greater insight than *mixed*. Interestingly, we can see that couples with neither partner being black have a higher survival function.

Fig. 6 Non-mixed couples are likely to survive longer, shown by their higher survival curve and that their marriages carried on longer into the study.

Modeling

```
# Base model using only education
edu_coxph <- coxph(Surv(years, div) ~ edu, data = divorce)
edu_coxph
```

```
## Call:
## coxph(formula = Surv(years, div) ~ edu, data = divorce)
##
##           coef exp(coef) se(coef)      z      p
```

```
## edu1 0.2388 1.2697 0.0669 3.57 0.00036
## edu2 -0.0778 0.9251 0.1080 -0.72 0.47106
##
## Likelihood ratio test=17.4 on 2 df, p=0.00017
## n= 3371, number of events= 1032
```

Although the P-value for the ratio between education levels 0 and 2 (highschool and graduate) is not significant, the model as a whole attains a significant P-value of 0.017%. The high P-value for education levels 0 and 2 is not surprising, however, given how close the two curves are in **Figure 4**.

Control for Racial Makeup

Here we add the racial covariates to previous. Our goal is to ensure that *education* is still a significant predictor in all models.

```
eh_coxph <- coxph(Surv(years,div) ~ edu + hblack, data=divorce)
em_coxph <- coxph(Surv(years,div) ~ edu + mixed, data=divorce)
emb_coxph <- coxph(Surv(years,div) ~ edu + mixed + hblack, data=divorce)
embw_coxph <- coxph(Surv(years,div) ~ edu + mixed + hblack + wblack, data=divorce)
```

```
## Call:
## coxph(formula = Surv(years, div) ~ edu + hblack, data = divorce)
##
##           coef exp(coef) se(coef)      z      p
## edu1      0.2683    1.3077  0.0677   3.97 7.3e-05
## edu2     -0.0210    0.9792  0.1097  -0.19 0.8481
## hblack1   0.2460    1.2789  0.0765   3.22 0.0013
##
## Likelihood ratio test=27.3 on 3 df, p=5.17e-06
## n= 3371, number of events= 1032
```

```
## Call:
## coxph(formula = Surv(years, div) ~ edu + mixed, data = divorce)
##
##           coef exp(coef) se(coef)      z      p
## edu1      0.27822    1.32077  0.06789   4.10 4.2e-05
## edu2     -0.00771    0.99232  0.10993  -0.07 0.94405
## mixed1   0.28333    1.32754  0.07603   3.73 0.00019
##
## Likelihood ratio test=30.6 on 3 df, p=1.03e-06
## n= 3371, number of events= 1032
```

```
## Call:
## coxph(formula = Surv(years, div) ~ edu + mixed + hblack, data = divorce)
##
##           coef exp(coef) se(coef)      z      p
## edu1      0.2928    1.3401  0.0682   4.29 1.8e-05
## edu2      0.0217    1.0220  0.1107   0.20 0.8444
## mixed1    0.2342    1.2640  0.0791   2.96 0.0031
## hblack1   0.1829    1.2008  0.0796   2.30 0.0216
##
## Likelihood ratio test=35.7 on 4 df, p=3.28e-07
## n= 3371, number of events= 1032
```

```
## Call:
## coxph(formula = Surv(years, div) ~ edu + mixed + hblack + wblack,
##       data = divorce)
##
##              coef exp(coef) se(coef)      z      p
## edu1         0.2936    1.3412  0.0683  4.30 1.7e-05
## edu2         0.0236    1.0239  0.1112  0.21  0.8318
## mixed1       0.2298    1.2583  0.0825  2.78  0.0054
## hblack1      0.1789    1.1959  0.0824  2.17  0.0300
## wblack1      0.0154    1.0156  0.0822  0.19  0.8511
##
## Likelihood ratio test=35.8 on 5 df, p=1.05e-06
## n= 3371, number of events= 1032
```

The R output above confirms that *education* is a significant predictor in all control situations. Now that we have established that *education* alone is a stable covariate, our team decided to investigate the above models further. Our motivation is that one (or more) of the models may in fact be a better choice than *education* alone.

Effect of Racial Makeup

```
edu_BIC <- BIC(edu_coxph)
eh_BIC <- BIC(eh_coxph)
em_BIC <- BIC(em_coxph)
emb_BIC <- BIC(emb_coxph)
embw_BIC <- BIC(embw_coxph)

# A new model using the new "couple" covariate
edu_couple_coxph <- coxph(Surv(years, div) ~ edu + couple, data = divorce)
edu_couple_BIC <- BIC(edu_couple_coxph)

## Base model = 15685.14
## Edu + Husband = 15682.16
## Edu + Mixed = 15678.84
## Edu + Mixed + Husband = 15680.64
## Edu + Mixed + Husband + Wife = 15687.54
## Edu + Couple = 15687.54
```

Although the BIC on the model using *education* and *mixed* was the lowest, we decided to continue our analysis using a new variable, *couple* which encodes all the racial makeups into one predictor. Our reasoning is that *couple* encodes far more information than *mixed* and may provide a more insightful model.

To justify our decision, we used a likelihood ratio test to ensure there is no significant difference between the model using the newly encoded predictor, and the model which attained the lowest BIC (*em_coxph*).

```
chi_test <- 2*(edu_couple_coxph$loglik[[2]] - em_coxph$loglik[[2]])

# 5 parameters in the bigger model (lots of factor levels)
# 3 parameters in the restricted model
1 - pchisq(chi_test, df = 5-3)

## [1] 0.07526716
```

The P-value is 7.5%, which is above any common significance level. Therefore, there is no significant difference between the two models using...

- $base \sim Education + Couple$
- $base \sim Education + Mixed$

This allows us to use a more condensed predictor for later analysis. Lastly, it is interesting to note that the model using Education + Mixed + Hblack + Wblack is in fact the same model as Education + Couple. We expected to see this since they encode the same information, but it is encouraging nonetheless.

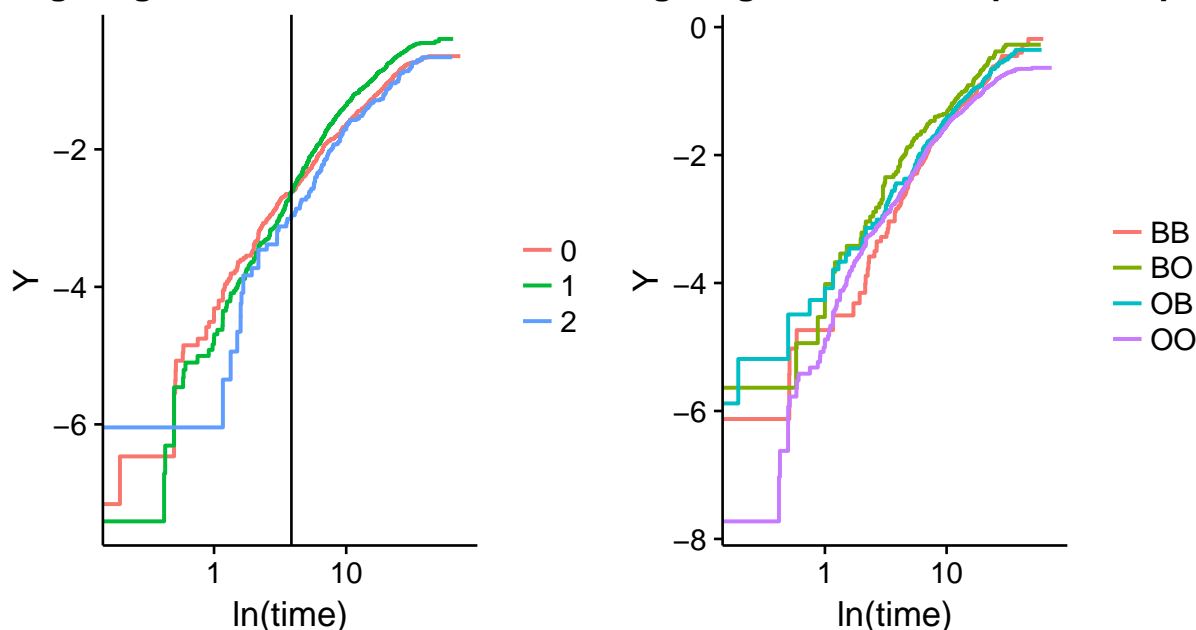
Checking Proportional Hazard Assumptions

Graphical Approach

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

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

Log-Log curve for Education level Log-Log curve for Couple Makeup



Plot analysis from left to right.

Fig. 7 The intersecting lines are concerning because they indicate a possible violation of the proportionality assumption of the Cox PH model. Our team investigated the necessity of a time-transformation on education to account for the issue. The vertical line to indicate where the lines cross; it is placed at $\ln(47) = 3.85$.

Fig. 8 Again, the intersecting lines are concerning. In this case, however, we will investigate the necessity of stratifying the racial makeup predictor.

Numerical Approach


```

edu_zph <- cox.zph(edu_coxph)
edu_couple_zph <- cox.zph(edu_couple_coxph)

edu_couple_zph

```

```

##           rho  chisq      p
## edu1      0.0307  0.995 0.31854
## edu2      0.0516  2.818 0.09321
## coupleB0 -0.0513  2.720 0.09913
## couple0B -0.0346  1.239 0.26564
## couple00 -0.0922  9.031 0.00265
## GLOBAL          NA 11.411 0.04382

```

While our numerical test indicates that the education covariate does **not** violate the proportional hazards assumption, we will still investigate the improvement made by including a time-transformation.

The numerical test for racial makeup of the couple confirms our suspicions from the c-log-log plot. The predictor **does** violate the proportional hazards assumption and we will use it as a stratified variable.

Stratification

Here we use *couple* as a stratified covariate because it violated the proportionality assumption and we are still interested in gaining insight from its inclusion.

```

# Construct Cox-PH model with the stratified variable
edu_strata_coxph <- coxph(Surv(years, div) ~ edu + strata(couple), data = divorce)
edu_strata_zph <- cox.zph(edu_strata_coxph)
strata_BIC <- round(BIC(edu_strata_coxph), r)

```

```

## Call:
## coxph(formula = Surv(years, div) ~ edu + strata(couple), data = divorce)
##
##           coef exp(coef) se(coef)      z      p
## edu1 0.3035      1.3546   0.0687 4.42 9.9e-06
## edu2 0.0304      1.0309   0.1114 0.27  0.78
##
## Likelihood ratio test=22 on 2 df, p=1.65e-05
## n= 3371, number of events= 1032

```

```

##           rho chisq      p
## edu1      0.0332  1.16 0.2804
## edu2      0.0521  2.87 0.0904
## GLOBAL          NA  3.07 0.2151

```

```

## Base model BIC = 15685.14
## Strata model BIC = 13543

```

Confidence Intervals

```
conf_int <- summary(educ_strata_coxph)$conf.int[,3:4]
conf_int
```

```
##      lower .95 upper .95
## edu1 1.1839834  1.549700
## edu2 0.8286285  1.282513
```

Numerical test analysis

Cox Model Stratifying on *couple* highlights the differences between the levels of *education*, as the P-values for both coefficients are even more disparate than the original model. As before, the ratio between divorce rates of highschool-educated and graduate-educated marriages does not achieve a significant P-value. Stratifying on *couple* shows that the difference between the divorce rates of highschool-educated and graduate-educated marriages are even more pronounced. The total model is significant.

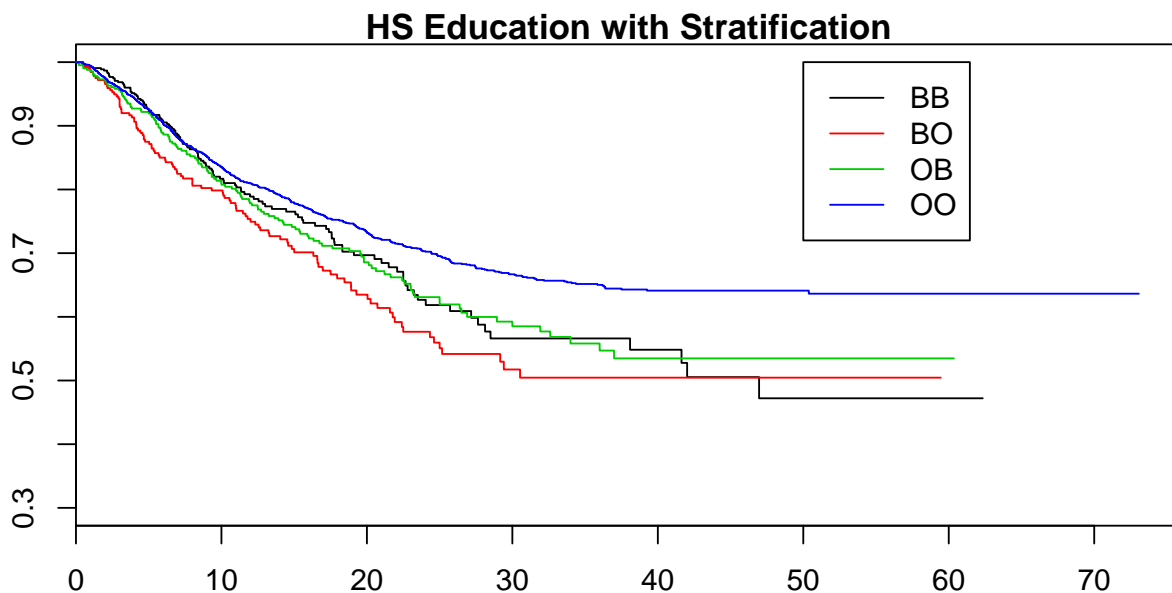
Proportionality Test As expected, stratifying *couple* ensures the model does not violate the proportionality assumption. Neither the factor levels of *education* nor the total model achieve a significant P-value.

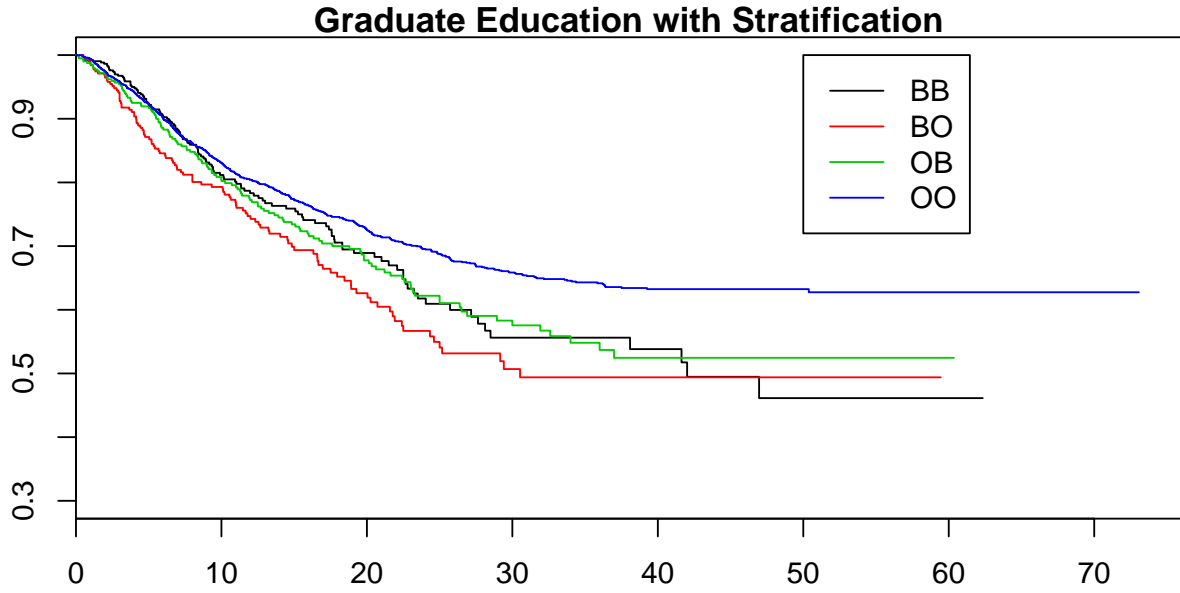
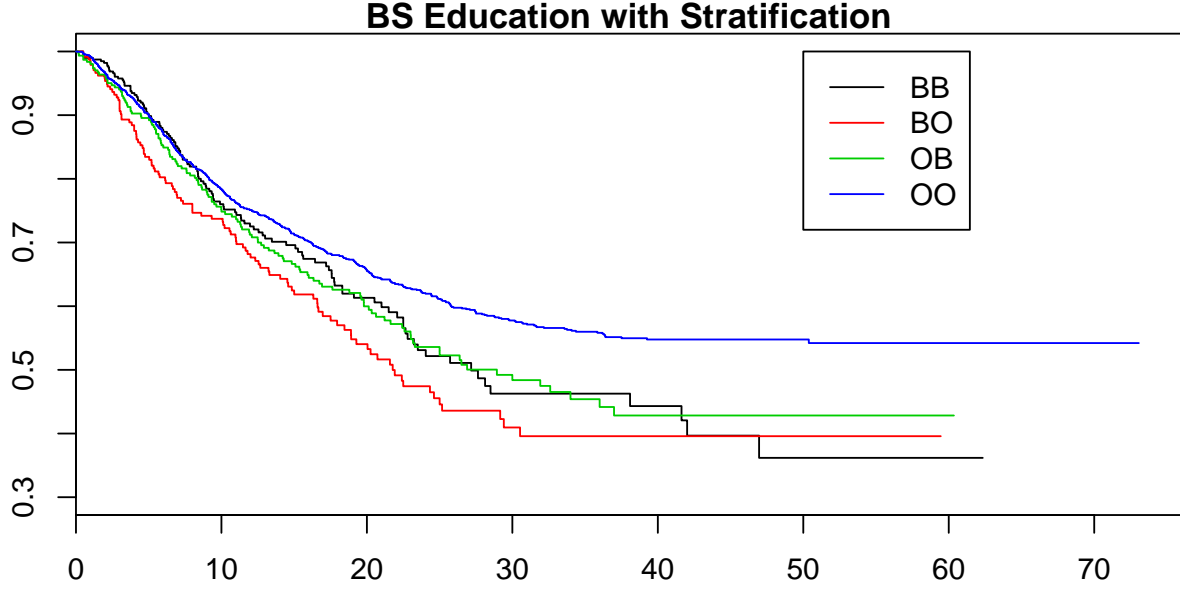
BIC The BIC of the stratified model is quite lower than the original model, indicating that it is likely a better fit for our data.

Confidence Intervals

Bachelors (edu1) : We are 95% sure that our true ratio for a bachelor's degree is between 1.18 and 1.55

Master's or Above (edu2) : We are 95% sure that our true ratio for a master's degree or above is between .83 and 1.28 which verifies that it is not significantly different from having a high school education.





Plot interpretation from top to bottom. Please note that the y-axis minimum is set to 0.30

Fig. 9

Fig. 10

Fig. 11

Time Transform

As we learned in **Figure 7**, the time in question occurred at 47 years. Our team decided to split each observation at this time. We opted for this approach over introducing a new term, $\beta * g(X)$, because the four factor levels in the stratified variable make model interpretation difficult.

```

# Split on t = exp(3.85) = 47
edu_split_df <- survSplit(Surv(years, div) ~ edu, data = divorce, cut = 47, episode = "Ep")
edu_surv_split <- coxph(Surv(years, div) ~ edu:Ep, data = edu_split_df)

## Warning in fitter(X, Y, strats, offset, init, control, weights = weights, :
## Loglik converged before variable 1,2,3 ; beta may be infinite.

# edu_couple_surv_split <- coxph(Surv(years, div) ~ edu + couple, data = edu_split_df) ... necessary?
edu_surv_split

## Call:
## coxph(formula = Surv(years, div) ~ edu:Ep, data = edu_split_df)
##
##              coef exp(coef)  se(coef)      z      p
## edu0:Ep -1.73e+01  3.14e-08  4.88e+02 -0.04  0.97
## edu1:Ep -1.70e+01  4.00e-08  4.88e+02 -0.03  0.97
## edu2:Ep -1.74e+01  2.91e-08  4.88e+02 -0.04  0.97
##
## Likelihood ratio test=251  on 3 df, p=0
## n= 3589, number of events= 1032

```

Sadly, there are very few observed instances of divorce past 47 years, as shown by **Figure 1**. This class imbalance is at the root of the strange output of the Cox model. Although the numerical output is uninterpretable, the class imbalance past 47 years is also the reason for the crossing lines in the clog-log plot in **Figure 7**.

Final Results/Summary

For this project we decided to explore if and how a husband's education affects divorce rates over time while taking into account the race of the couple. Our goal was to find the most significant model using various statistical methods that we learned in class as well as from our textbook. We determined that education definitely has an impact on divorce rate and then tested how other variables contributed to our model. For our final model we chose to use the education and stratified couple variables as it gives us our most significant p-value and lowest BIC. Regarding the education coefficients we see that with a bachelor's education the marriage roughly fails 1.4 times faster than with a high school education and having a Master's degree or above results in the same failure rate as with a high school education. Although we cannot know with certainty why we ended up with the above results, we think that there is less complications between spouse's when the husband has a lower education because it is likely that his partner has a similar education where as the variance between the spouse's education is higher when the husband has a bachelor's degree. When the husband has a master's degree or above then the same reason as for the highschool education applies. To have more evidence to determine whether our claim is accurate we would need further exploration with the wife's education and income for both the husband and wife as additional variables. In conclusion, according to our results we would recommend pursuing a Master's degree to ensure a more successful marriage.