Marriage Dissolution in the United States
PSTAT 175 Final Project (Revised):

By: Simranjit Kaur, Joanna Kim, and Brandon Lee

12/20/2020

Data Description

For our project, we will be analyzing the "Marriage Dissolution in the United States" data set. The data provides data for 3371 couples in the United States. The unit of observation is the couple and the event of interest is divorce, with interview and widowhood treated as censoring events. The data contains three covariates: the education of the husband, whether the husband is African-American, and whether the couple is mixed. The variables in the data set are:

1) id: a couple identification number.
2) heduc: the education of the husband, coded
   - 0 = less than 12 years of education
   - 1 = 12 to 15 years of education
   - 2 = 16 or more years of education
3) heblack: coded 1 if the husband is black and 0 otherwise.
4) mixed: coded 1 if the husband and wife have different ethnicity (defined as black or other), 0 otherwise.
5) years: duration of the marriage in years, from the date of the wedding to divorce or censoring (due to widowhood or interview).
6) div: the failure indicator, coded 1 for divorce and 0 for censoring.

This data set was retrieved from https://data.princeton.edu/wws509/datasets/#divorce.

Key Scientific Question

For this project, our group is interested in seeing how the different covariates affect the survival probability of the couple's marriages, and we aim to examine the interactions between the covariates to see if any of these interactions are significant. Furthermore, we are also interested in seeing whether or not the covariates have differing effects on the probability of divorce depending on how long a couple has been married. We initially hypothesize that the hazard rate is high at the beginning of a marriage, and that, after a certain point in time, the hazard rate will decrease significantly. Our ultimate goal for this project is to test whether or not this hypothesis is true by using various statistical tools to successfully manipulate and analyze our data.

```
library(survival)
library(survminer)
```

```
## Loading required package: ggplot2
```

```
## Loading required package: ggpubr
```

```
divorce <- read.table("divorce.txt")
names(divorce)[names(divorce) == "V1"] <- "id"
names(divorce)[names(divorce) == "V2"] <- "heduc"
names(divorce)[names(divorce) == "V3"] <- "heblack"
names(divorce)[names(divorce) == "V4"] <- "mixed"
names(divorce)[names(divorce) == "V5"] <- "years"
names(divorce)[names(divorce) == "V6"] <- "div"

years <- as.vector(divorce$years)
div <- as.vector(divorce$div)
divorce.surv <- Surv(years, div)
divorce.fit <- survfit(divorce.surv ~ 1)

divorce[1:10,]
```

```
##    id heduc heblack 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
## 7  43     2       0     0 16.706   0
## 8  47     0       0     0 24.999   0
## 9  50     0       0     0 24.999   0
## 10 56     0       1     0  3.869   0
```

```
# first 10 rows of Marriage Dissolution Data
```

Data Analysis

Before we begin plotting our data, we can use some simple functions to learn more about the data.

```
#View(divorce)
sum(divorce$heduc[divorce$heduc == 1])
sum(divorce$heduc[divorce$heduc == 2])
3371 - sum(divorce$heduc[divorce$heduc == 1])-
  sum(divorce$heduc[divorce$heduc == 2])

3371-sum(divorce$heblack[divorce$heblack == 1])
3371-sum(divorce$mixed[divorce$mixed == 1 ])
```

```
summary(divorce$years)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.079   6.485  14.497  18.410  26.141  73.068
```

From our data exploration, we find that 860 out of the 3371 couples went to school for less than 12 years, 1655 went to school for 12-15 years, and 856 went to school for more than 16 years. Furthermore, we find that 2626 out of 3371 husbands are non-black, while 745 were black. Lastly, we find that 2730 out of 3371 couples have the same ethnicity as their partner while 641 of the couples were of mixed ethnicities. Moreover, we use the summary() function on the years variable and find that, on average, marriages lasted for 18 years before the couple got divorced. To further understand and visualize our data for the different covariates, we proceed by plotting some KM plots.

Kaplan-Meier Curves for Marriage Dissolution Data

We begin analyzing our data by using KM curves. First, we plot the KM with all of the covariates taken into consideration. Then, we plot the KM curves to further analyze the separate impact of each covariate on the survival probability.

```
plot(divorce.fit,
     xlab="Time (in Years)",
     ylab="Survival Probability",
     main = "Kaplan-Meier Curves \n for Marriage Dissolution Data",
     col = c("red", "black","black"),
     conf.int = T)
```

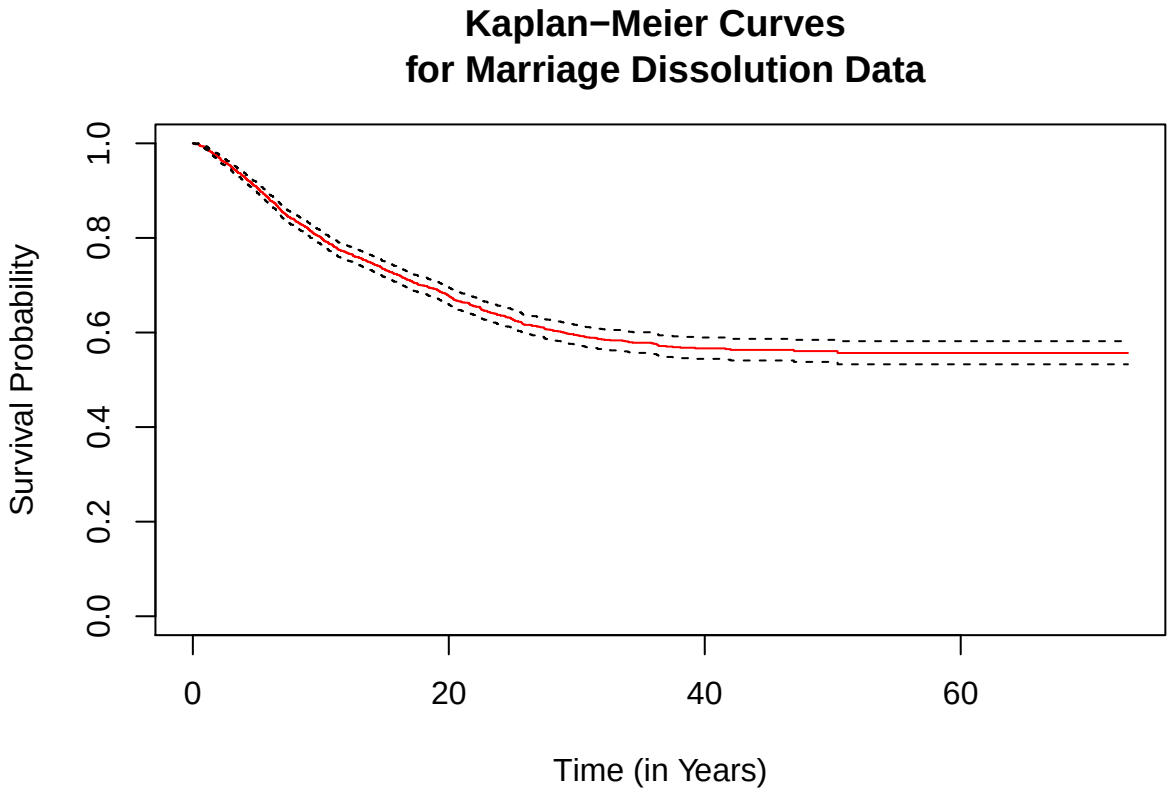


Figure 1: Kaplan-Meier Curve of Marriage Dissolution Data with all covariates.

From Figure 1, we can see that as time goes by, the survival probability of a couple's marriage decreases and then becomes steady after 40 years.

```
divorce.fit.2 <- survfit(Surv(divorce$years,
            divorce$div)~divorce$heduc, data = divorce)

plot(divorce.fit.2, col = c("red", "blue", "green"),
     conf.int = F,
     xlim=c(0,70),
     main = "Kaplan-Meier Curves for Marriage Dissolution Data
     with Respect to Husband's Education Level",
     ylab= "Survival Probability",
     xlab = "Time (in Years)")
legend("topright", legend = c("less than 12 years of education",
                              "12-15 years of education",
                              "16 or more years of education"),
       col = c("red", "blue", "green"), lty = 1, cex = 1)
```

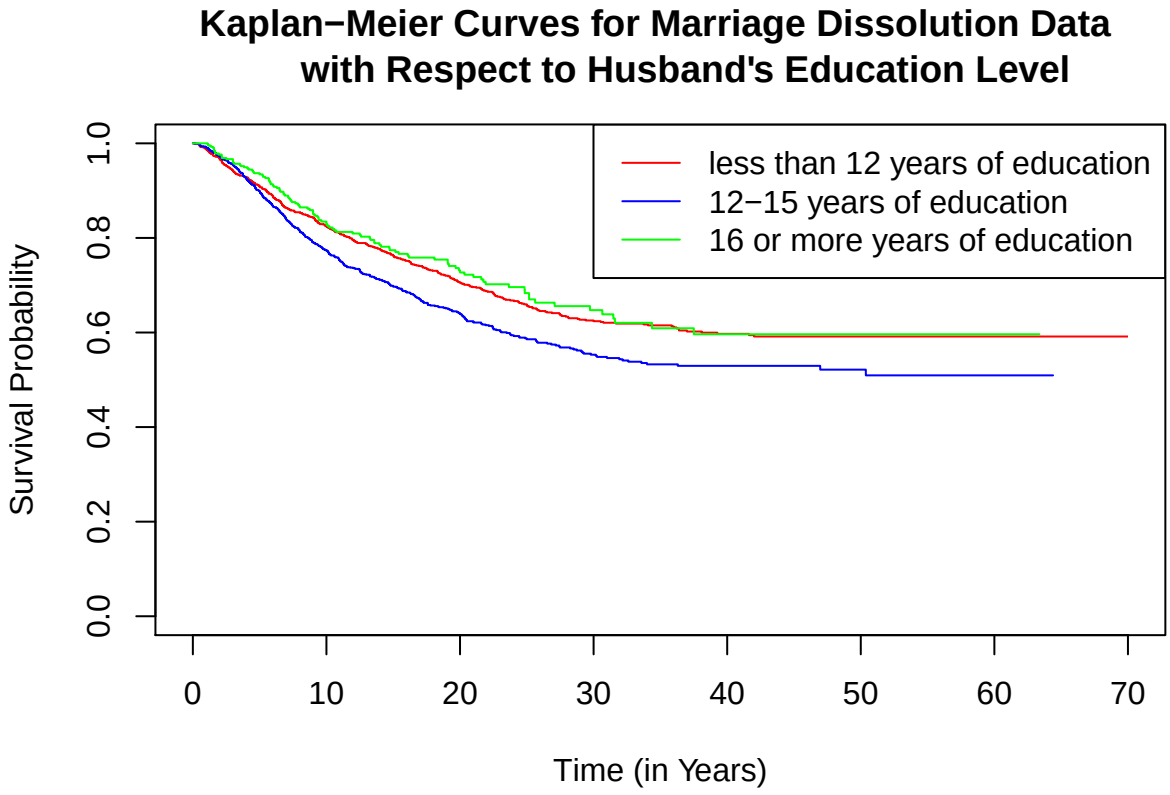


Figure 2: Kaplan-Meier Curves comparing the survival probabilities of couples with varying education levels of the husbands.

From Figure 2, we can conclude that the education level of the husband impacts the survival probability of the couple's marriage. It seems that couples with a husband who has been in the education system for 12 to 15 years have the lowest survival probability in comparison to the couples with husbands who have either been in the education system for less than 12 years or more than 16 years. Nonetheless, from these KM curves, we can conclude that the education level of the husband does affect the survival probability.

```
divorce.fit.3 <- survfit(Surv(divorce$years,
            divorce$div)~divorce$heblack, data = divorce)

plot(divorce.fit.3, col = c("red", "blue"),
     conf.int = F,
     xlim=c(0,70),
     main = "Kaplan-Meier Curves for Marriage Dissolution Data
     with Respect to Husband Being Black",
     ylab= "Survival Probability",
     xlab = "Time (in Years)")
legend("topright", legend = c("Non-Black Husband","Black Husband"),
       col = c("red", "blue"), lty = 1, cex = 1)
```

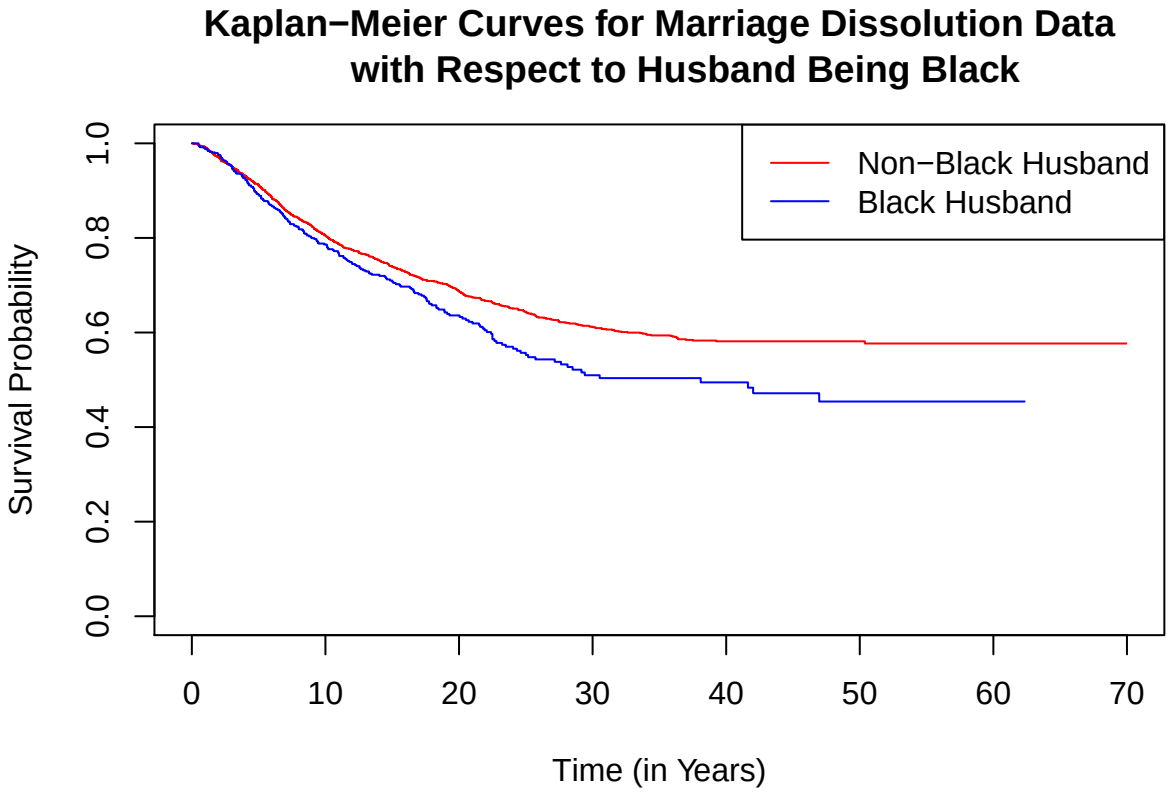


Figure 3: Kaplan-Meier Curves comparing the survival probabilities of couples with varying ethnicities (black or non-black) of the husbands.

From Figure 3, we can conclude that the ethnicity (black or non-black) of the husband impacts the survival probability of the couple's marriage. It seems that when the husband is black, the couple's marriage does not survive as well or as long as the couples where the husband is not black.

```
divorce.fit.4 <- survfit(Surv(divorce$years,
            divorce$div)~divorce$mixed, data = divorce)

plot(divorce.fit.4, col = c("red", "blue"),
     conf.int = F,
     xlim=c(0,70),
     main = "Kaplan-Meier Curves for Marriage Dissolution Data
     with Respect to Couple's Ethnicity",
     ylab= "Survival Probability",
     xlab = "Time (in Years)")
legend("topright", legend = c("Same Ethnicity","Different Ethnicity"),
       col = c("red", "blue"), lty = 1, cex = 1)
```

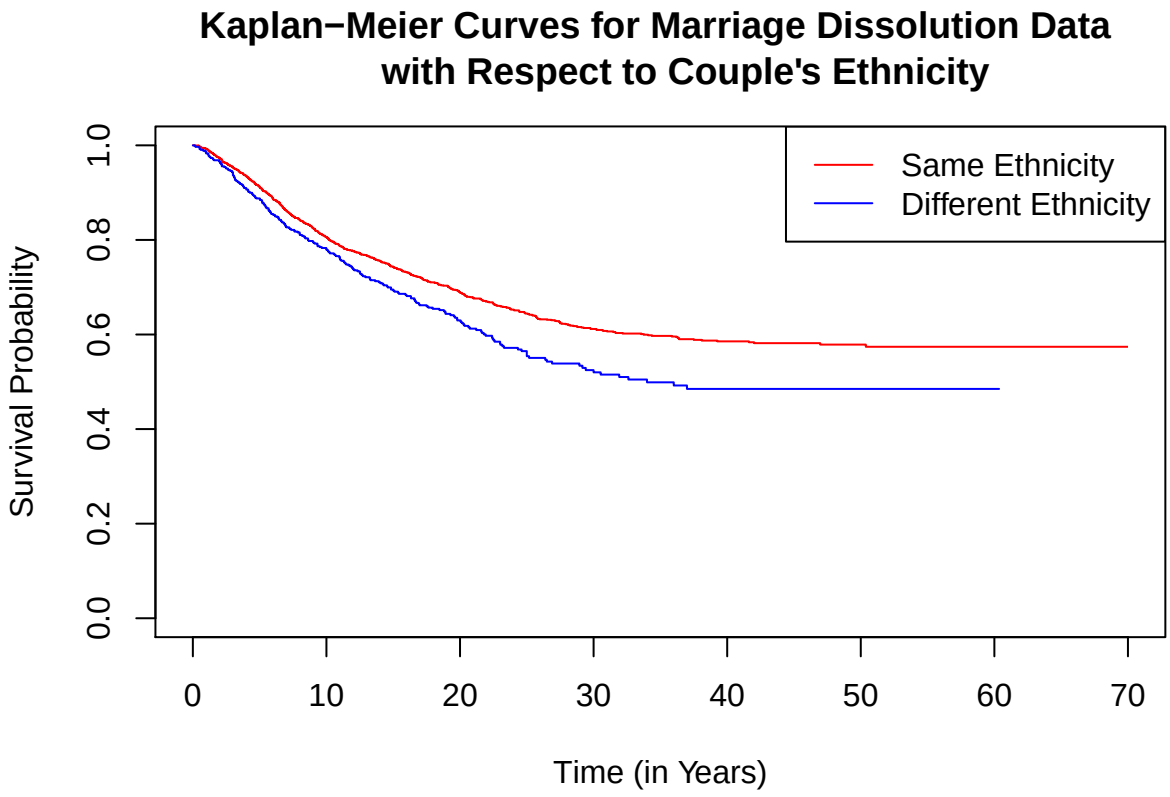


Figure 4: Kaplan-Meier Curves comparing the survival probabilities of couples with varying ethnicities (mixed or non-mixed) of the couple.

From Figure 4, we can conclude that the ethnicity of the couples impacts the survival probability of the couple's marriage. It seems that when the couples have the same ethnicity, their marriage has a better probability of surviving in comparison to the couples who have mixed ethnicities.

In conclusion, from the Kaplan-Meier plots, we can claim that all of the covariates impact how quickly the couple gets divorced. To further check our claims, we conduct a Log Rank Test.

Log Rank Test

```
(divorce.lrt <- survdiff(Surv(years,div) ~ divorce$heduc, data = divorce))
```

```
## Call:
## survdiff(formula = Surv(years, div) ~ divorce$heduc, data = divorce)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## divorce$heduc=0 1288      393      436      4.30      7.51
## divorce$heduc=1 1655      529      463      9.26     16.93
## divorce$heduc=2  428      110      132      3.73      4.28
##
##  Chisq= 17.4  on 2 degrees of freedom, p= 2e-04
```

From this LRT we see that the p-value is significant since it is smaller than our significance value of 0.05.

Therefore, we can claim that there is a statistically significant difference between the survival rates of couples where the education level of the husband varies.

```
(divorce.lrt <- survdiff(Surv(years,div) ~ divorce$heblack, data = divorce))
```

```
## Call:
## survdiff(formula = Surv(years, div) ~ divorce$heblack, data = divorce)
##
##                         N Observed Expected (O-E)^2/E (O-E)^2/V
## divorce$heblack=0 2626      802      839      1.62      8.69
## divorce$heblack=1  745      230      193      7.03      8.69
##
##  Chisq= 8.7  on 1 degrees of freedom, p= 0.003
```

From this LRT we can see that the p-value is significant since it is smaller than our significance value of 0.05. Therefore, we can claim that there is a statistically significant difference between the survival probabilities of couples with African-American husbands and couples with non-African-American husbands.

```
(divorce.lrt <- survdiff(Surv(years,div) ~ divorce$mixed, data = divorce))
```

```
## Call:
## survdiff(formula = Surv(years, div) ~ divorce$mixed, data = divorce)
##
##                       N Observed Expected (O-E)^2/E (O-E)^2/V
## divorce$mixed=0 2730      797      839      2.12      11.3
## divorce$mixed=1  641      235      193      9.22      11.3
##
##  Chisq= 11.3  on 1 degrees of freedom, p= 8e-04
```

From this LRT we can see that the p-value is significant since it is smaller than our significance value of 0.05. Therefore, we can claim that there is a statistically significant difference between the survival rates of interracial couples and non-interracial couples.

After running a Log Rank Test on all three of our covariates, we see that each p-value is smaller than our significance value of 0.05. Therefore, it is safe to assume that our prior claim still follows. We conclude that the covariates have a statistically significant impact on the survival probability of the couple's marriage.

Now that we have verified our assumptions with both the KM curves and the Log Rank Test, we proceed to build a Cox Proportional Hazard Model.

Cox PH Hazard Model

To begin building our model, we start by using Forward Selection to fit each covariate separately, find the covariate that minimizes the AIC, and add the AIC minimizing covariate to the model, repeating these steps until we are unable find any more significant covariates.

We begin by creating the cox variables for each of our covariates, and then we analyze the summaries and find the following:

```r
divorce$heduc <- as.factor(divorce$heduc)
heduc.ph <- coxph(Surv(years, div) ~ heduc, data = divorce)
divorce$heblack <- as.factor(divorce$heblack)
heblack.ph <- coxph(Surv(years, div) ~ heblack, data = divorce)
divorce$mixed <- as.factor(divorce$mixed)
mixed.ph <- coxph(Surv(divorce$years, divorce$div) ~ mixed, data = divorce)
```

```r
summary(heduc.ph)
```

```
## Call:
## coxph(formula = Surv(years, div) ~ heduc, data = divorce)
##
##   n= 3371, number of events= 1032
##
##            coef exp(coef) se(coef)      z Pr(>|z|)
## heduc1  0.23880   1.26973  0.06692  3.568 0.000359 ***
## heduc2 -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
## heduc1    1.2697     0.7876    1.1136     1.448
## heduc2    0.9251     1.0809    0.7486     1.143
##
## Concordance= 0.535  (se = 0.008 )
## Likelihood ratio test= 17.36  on 2 df,   p=2e-04
## Wald test            = 17.29  on 2 df,   p=2e-04
## Score (logrank) test = 17.39  on 2 df,   p=2e-04
```

The p-value is 2e-04, which is less than 0.05. We can conclude that the difference in education does affect the divorce rates between married couples.

The coefficient for education amount of 12 to 15 years (heduc1) is positive, meaning that compared to married men who had less than 12 years of education, the hazard rate is larger, meaning the divorce rate is higher for married men with education amounts of 12 to 15 years. The confidence interval is (1.1136, 1.448). Because the CI only contains values greater than 1, it further proves that the divorce rate is higher. Because the hazard ratio is 1.26973, the education amount of 12-15 years has a 27% increase of hazard rate than married men with less than 12 years of education.

The coefficient for education amount of 16+ years (heduc2) is negative, meaning that compared to married men who had less than 12 years of education, the hazard rate is smaller. However, the confidence interval is (0.7486, 1.143). Because the CI includes 1, this shows that for married men with 16+ years of education, the hazard rate is almost the same as our baseline, married men with less than 12 years of education. The hazard ratio is 0.92512, which is very close to 1, meaning that for married men with 16+ years of education, they have a very similar chance of marriage survival, if not just a slightly higher chance, than men with 12 years or less of education.

```r
summary(heblack.ph)
```

```
## Call:
## coxph(formula = Surv(years, div) ~ heblack, data = divorce)
##
```

```
##   n= 3371, number of events= 1032
##
##             coef exp(coef) se(coef)     z Pr(>|z|)
## heblack1 0.22066   1.24690  0.07501 2.942  0.00326 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##          exp(coef) exp(-coef) lower .95 upper .95
## heblack1     1.247      0.802     1.076     1.444
##
## Concordance= 0.514  (se = 0.007 )
## Likelihood ratio test= 8.3  on 1 df,   p=0.004
## Wald test            = 8.65  on 1 df,   p=0.003
## Score (logrank) test = 8.69  on 1 df,   p=0.003
```

The p-value is 0.004, which is less than 0.05. We can conclude that whether the male in the relationship is African-American or not does affect the divorce rates between married couples.

The coefficient for married men in the marriage being African-American (heblack1) is positive, meaning that compared to marriages with men who are not African-American (heblack0), the hazard rate is larger, meaning that the divorce rate is higher for relationships where the male is African-American. The confidence interval is (1.076, 1.444). Since the CI only contains values greater than 1, that further proves that marriages where the male is African-American has a higher hazard rate, which also means a higher divorce rate, than in marriages where the male is not African-American. Because the hazard ratio is 1.2469, marriages with men who are African-American have almost a 25% increase in rate of divorce than marriages with men who are not African-American.

```
summary(mixed.ph)
```

```
## Call:
## coxph(formula = Surv(divorce$years, divorce$div) ~ mixed, data = divorce)
##
##   n= 3371, number of events= 1032
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## mixed1 0.24945   1.28332  0.07426 3.359 0.000782 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## mixed1     1.283     0.7792     1.109     1.484
##
## Concordance= 0.519  (se = 0.007 )
## Likelihood ratio test= 10.77  on 1 df,   p=0.001
## Wald test            = 11.28  on 1 df,   p=8e-04
## Score (logrank) test = 11.34  on 1 df,   p=8e-04
```

The p-value is 0.001, which is less than 0.05. We can conclude that whether the married couples are of mixed ethnicity does affect the divorce rates between married couples.

The coefficient for mixed ethnicity (mixed1) is positive, meaning that compared to non-mixed ethnicity marriages (mixed0), the hazard rate is larger, meaning that the divorce rate is higher for mixed-ethnicity marriages. The confidence interval is (1.109, 1.484). Since the CI only contains values greater than 1, that further proves that mixed-ethnicity marriages have a higher hazard and divorce rate than non-mixed

marriages. Because the hazard ratio is 1.28332, mixed-ethnicity marriages have about a 28% increase in rate of divorce than non-mixed ethnicity marriages.

After this, we begin forward selection and use the AIC() function to determine which of covariates would help build the best model.

```
AIC(heduc.ph, heblack.ph, mixed.ph)
```

```
##            df      AIC
## heduc.ph    2 15675.26
## heblack.ph  1 15682.31
## mixed.ph    1 15679.85
```

From this AIC() output, we see that the "heduc" model minimizes the AIC, and therefore it is the best model at predicting survival for now. From here, we fit "heduc" with the "mixed" covariate and we fit "heduc" with the "heblack" covariate.

```
heduc.ph.mix <- coxph(Surv(years, div) ~ heduc + mixed, data = divorce)
heduc.ph.black <- coxph(Surv(years, div) ~ heduc + heblack, data = divorce)
AIC(heduc.ph.mix, heduc.ph.black)
```

```
##                df      AIC
## heduc.ph.mix    3 15664.02
## heduc.ph.black  3 15667.35
```

From this, we see that the covariate "heduc" added with the covariate "mixed" minimizes the AIC, and therefore it is a better model at predicting survival. From here, we fit the full model with all 3 covariates.

```
heduc.ph.full <- coxph(Surv(years, div) ~ heduc + mixed + heblack , data = divorce)
anova(heduc.ph.full)
```

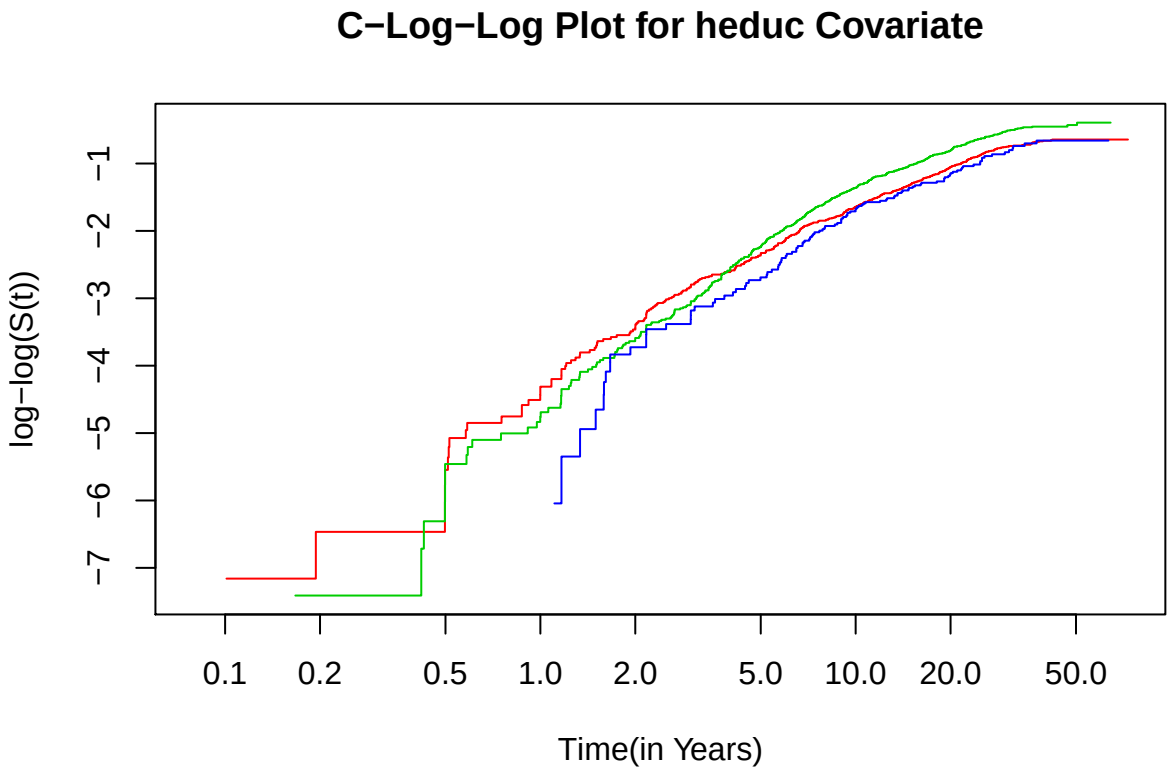
```
## Analysis of Deviance Table
##  Cox model: response is Surv(years, div)
## Terms added sequentially (first to last)
##
##          loglik   Chisq Df Pr(>|Chi|)
## NULL    -7844.3
## heduc   -7835.6 17.3608  2  0.0001699 ***
## mixed   -7829.0 13.2379  1  0.0002743 ***
## heblack -7826.4  5.1382  1  0.0234051 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the ANOVA test output we see that all three of the covariates are indeed significant in our model. After using forward selection to construct our model, we run some tests to make sure that these covariates satisfy the Proportional Hazards (PH) Assumption.

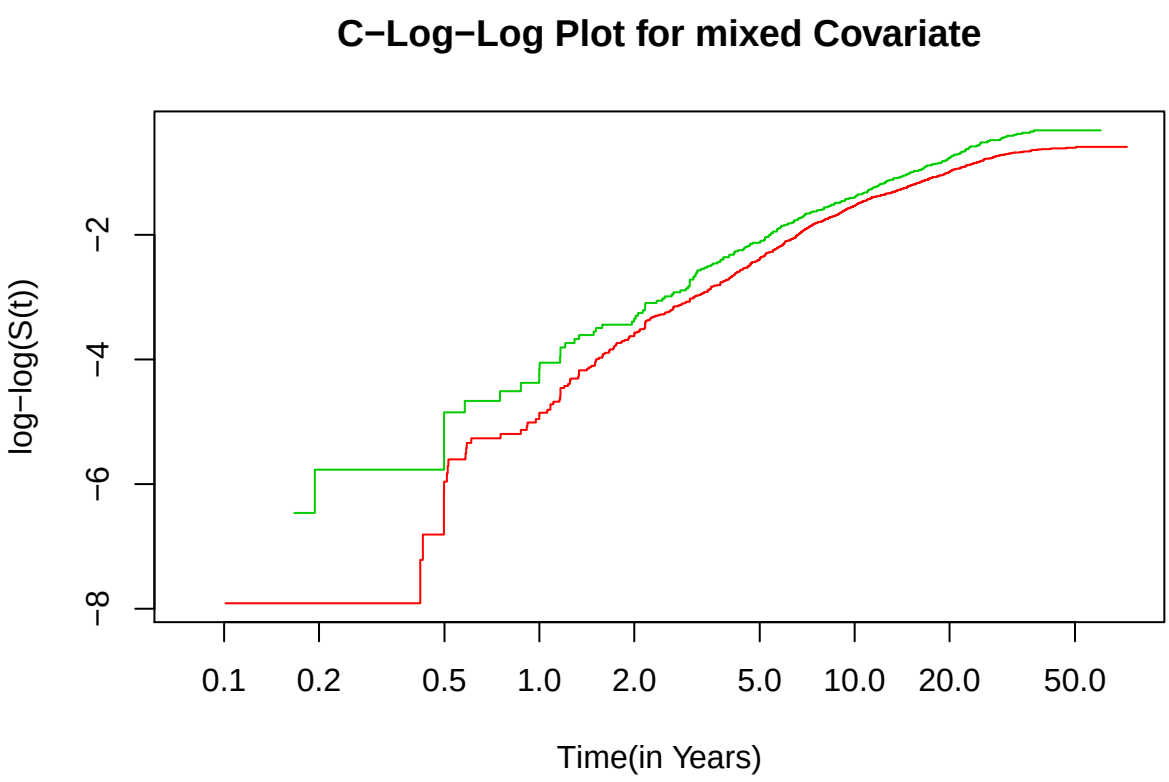
Cox PH Hazard Model Assumption Check

To check our model assumptions, we use C-Log-Log plots to check that the PH assumption is valid for each covariate. Furthermore, to confirm our findings from the plots, we use the cox.zph() function in R.

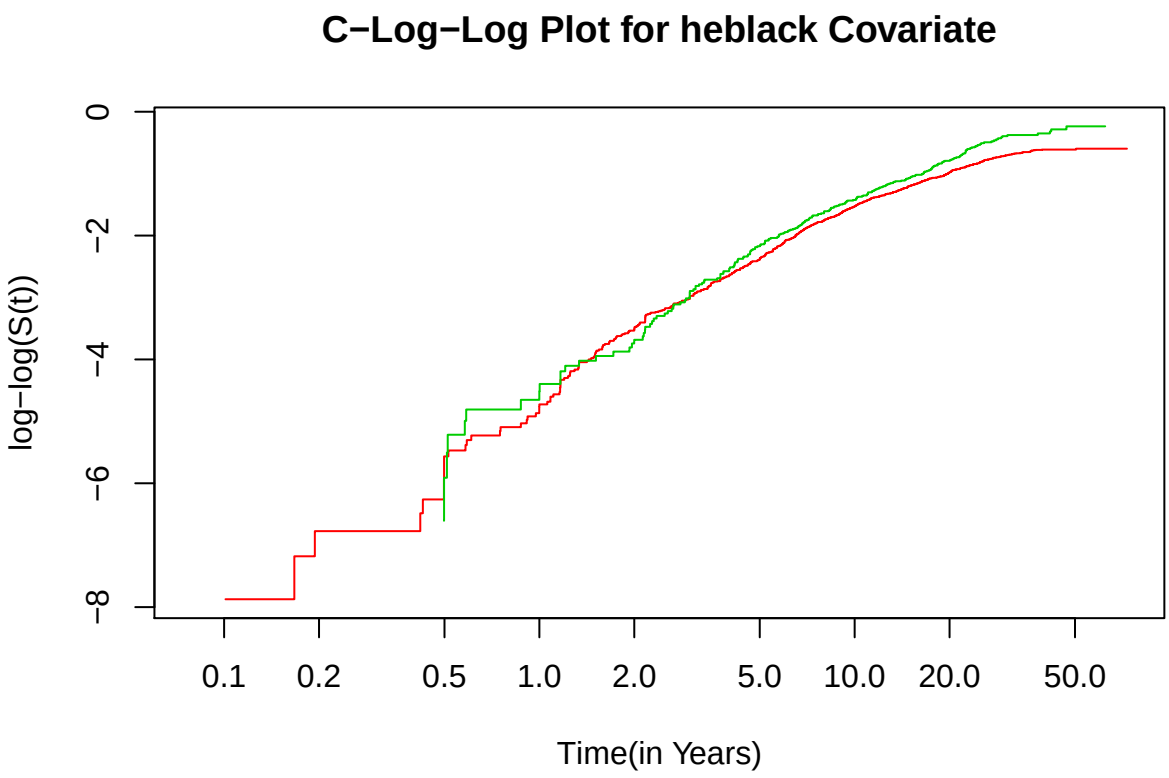
```
plot(survfit(divorce.surv ~ divorce$heduc), fun = "cloglog",
     col = c(2,3,4),
     ylab = "log-log(S(t))",
     xlab = "Time(in Years)",
     main = "C-Log-Log Plot for heduc Covariate")
```




```
plot(survfit(divorce.surv ~ divorce$mixed), fun = "cloglog",
     col = c(2,3),
     ylab = "log-log(S(t))",
     xlab = "Time(in Years)",
     main = "C-Log-Log Plot for mixed Covariate")
```




```
plot(survfit(divorce.surv ~ divorce$heblack), fun = "cloglog",
     col = c(2,3),
     ylab = "log-log(S(t))",
     xlab = "Time(in Years)",
     main = "C-Log-Log Plot for heblack Covariate")
```



From the C-Log-Log plots, we see that the heblack covariate might be the most concerning since there are multiple instances where the lines cross each other. The heduc covariate does have some instances where the lines cross, but it seems parallel for the most part. The mixed covariate seems to be pretty parallel and there isn't an instance where the lines cross each other. We predict that mixed might be a better covariate to include in our analysis in comparison to the heblack covariate, which might be a bit more concerning. To check our predictions, we use the cox.zph() function in R.

```
cox.zph(heduc.ph.full)
```

```
##          chisq df     p
## heduc    1.204  2 0.548
## mixed    0.489  1 0.485
## heblack 4.150  1 0.042
## GLOBAL   6.510  4 0.164
```

H0: The Proportional Hazards Assumption is reasonable to use
HA: The Proportional Hazards Assumption is NOT reasonable to use

Since our p-value for the mixed covariate is greater than the significance level (0.05), we fail to reject the null hypothesis and claim that there is insufficient evidence for us to abandon the PH assumption. Therefore, we are justified in using the PH assumption in our modeling of the effect of heduc. We can see that the only concerning covariate is the heblack covariate, since it's p-value is smaller than our significance level of 0.05. Since this matches our claim from the C-Log-Log plots, we reject the null hypothesis and claim that there is sufficient evidence for us to be concerned about including the heblack covariate in our analysis and that the PH assumption is not reasonable to use. As a result, our next step is to stratify the heblack covariate.

Stratification

```
heduc.ph.full.strata <- coxph(Surv(years, div) ~ heduc + mixed + strata(heblack) , data = divorce)
summary(heduc.ph.full.strata)
```

```
## Call:
## coxph(formula = Surv(years, div) ~ heduc + mixed + strata(heblack),
##     data = divorce)
##
##   n= 3371, number of events= 1032
##
##           coef exp(coef) se(coef)     z Pr(>|z|)
## heduc1 0.30106   1.35129  0.06847 4.397  1.1e-05 ***
## heduc2 0.02725   1.02762  0.11090 0.246   0.8059
## mixed1 0.23064   1.25940  0.07923 2.911   0.0036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##        exp(coef) exp(-coef) lower .95 upper .95
## heduc1     1.351     0.7400    1.1816     1.545
## heduc2     1.028     0.9731    0.8269     1.277
## mixed1     1.259     0.7940    1.0783     1.471
##
## Concordance= 0.535  (se = 0.01 )
## Likelihood ratio test= 28.09  on 3 df,   p=3e-06
## Wald test            = 28.27  on 3 df,   p=3e-06
## Score (logrank) test = 28.41  on 3 df,   p=3e-06
```

From this model, we can see that our LRT gives us a significant value. Therefore, a model with stratification is worth using. After stratifying our model, we begin to check interactions and determine whether or not any of them are significant.

Likelihood Ratio Test for Interactions

```
heduc.interaction <- coxph(Surv(years, div) ~ strata(heblack) * heduc , data = divorce)
heduc.additive <- coxph(Surv(years, div) ~ strata(heblack) + heduc , data = divorce)
LRT.1 <- 2*(logLik(heduc.interaction)- logLik(heduc.additive))
round(pchisq(as.numeric(LRT.1), 2, lower.tail = FALSE),3)
```

```
## [1] 0.064
```

```
anova(heduc.interaction) #double check LRT
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(years, div)
## Terms added sequentially (first to last)
##
##                       loglik   Chisq Df Pr(>|Chi|)
```

```
## NULL                  -7291.0
## heduc                 -7281.0 19.8963  2  4.782e-05 ***
## strata(heblack):heduc -7278.3  5.5107  2    0.06359 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
mixed.interaction <- coxph(Surv(years, div) ~ strata(heblack) * mixed , data = divorce)
mixed.additive <- coxph(Surv(years, div) ~ strata(heblack) + mixed , data = divorce)
LRT.2 <- 2*(logLik(mixed.interaction)- logLik(mixed.additive))
round(pchisq(as.numeric(LRT.2), 1, lower.tail = FALSE),3)
```

```
## [1] 0.949
```

```
anova(mixed.interaction)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(years, div)
## Terms added sequentially (first to last)
##
##                        loglik  Chisq Df Pr(>|Chi|)
## NULL                  -7291.0
## mixed                 -7287.9 6.2767  1    0.01223 *
## strata(heblack):mixed -7287.9 0.0041  1    0.94909
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
heduc.mixed.interaction <- coxph(Surv(years,div) ~ heduc * mixed, data = divorce)
heduc.mixed.additive <- coxph(Surv(years,div) ~ heduc + mixed, data = divorce)
LRT.3 <- 2*(logLik(heduc.mixed.interaction)-
              logLik(heduc.mixed.additive))
round(pchisq(as.numeric(LRT.3), 2, lower.tail = FALSE),3)
```

```
## [1] 0.371
```

```
anova(heduc.mixed.interaction)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(years, div)
## Terms added sequentially (first to last)
##
##               loglik   Chisq Df Pr(>|Chi|)
## NULL         -7844.3
## heduc        -7835.6 17.3608  2  0.0001699 ***
## mixed        -7829.0 13.2379  1  0.0002743 ***
## heduc:mixed -7828.0  1.9815  2  0.3713043
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

By doing these different LRT tests, we see that none of the interactions are significant in our model, since they all have p-values that are larger than our significance value of 0.05. Therefore, we conclude that the covariates that we should include to create the best model are heduc + heblack + strata(heblack).

To take our project a step further, we proceed to split our data and use time-varying covariates to analyze the married couples at different stages of their marriage. We aim to explore the impact that the different covariates have on the survival probability depending on the different time periods of the marriages.

Time Varying Covariates

Since our data focuses on the length of marriages in the US, we start by hypothesizing that the hazard rate is high at the beginning of a marriage, and that, after a certain point (or points) in time, the hazard rate decreases. We will begin testing this hypothesis by first finding the "best" years to split the data with the goal of examining any changes over time in mind. To accomplish this, we use the method of trial and error and split our data set into two or three sections of time repeatedly until we determine the "best possible" fit. After performing trial and error, we find that it is best to split the data set at the 3^{rd} and/or 16^{th} years. In the data that was split at the 3^{rd} year alone, however, only the heduc covariate seemed to be significantly affected by time. On the other hand, in the data that was split at the 16^{th} year alone, only the heblack covariate seemed to be significantly affected by time. We then try splitting the data at both the 3^{rd} and 16^{th} years and find that, in this data, both the heduc and heblack covariates seemed to be significantly affected by time, rather than only one of the two being significantly affected by time. Thus, we find that it is best to split the data set at the 3^{rd} and 16^{th} years and use this information to split the data using the survSplit() function.

```
#create time split data
divorce$years2 <- divorce$years
divorce$years2[divorce$years==0] <- 0.25
divorce.split <- survSplit(Surv(years2,div)~heduc + heblack + mixed,
data=divorce, cut=c(3,16), id="Subject",
episode="Episode")
```

After splitting the data, we proceed by creating two different models: a time-independent model that is our starting point and a time-dependent model that is in the start-stop data form.

```
#time-independent model (starting point)
coxph(Surv(years2,div)~heduc + heblack + mixed,data=divorce)
```

```
## Call:
## coxph(formula = Surv(years2, div) ~ heduc + heblack + mixed,
##     data = divorce)
##
##              coef exp(coef) se(coef)     z        p
## heduc1   0.29275   1.34011  0.06819 4.293 1.76e-05
## heduc2   0.02173   1.02197  0.11073 0.196  0.84442
## heblack1 0.18295   1.20075  0.07962 2.298  0.02158
## mixed1   0.23425   1.26395  0.07914 2.960  0.00308
##
## Likelihood ratio test=35.74  on 4 df, p=3.278e-07
## n= 3371, number of events= 1032
```

```
#time-dependent model (start-stop data form)
coxph(Surv(tstart,years2,div)~heduc + heblack + mixed,data=divorce.split)
```

```
## Call:
## coxph(formula = Surv(tstart, years2, div) ~ heduc + heblack +
##     mixed, data = divorce.split)
##
##              coef exp(coef) se(coef)     z        p
## heduc1   0.29275   1.34011  0.06819 4.293 1.76e-05
## heduc2   0.02173   1.02197  0.11073 0.196  0.84442
```

```
## heblack1 0.18295   1.20075  0.07962 2.298  0.02158
## mixed1   0.23425   1.26395  0.07914 2.960  0.00308
##
## Likelihood ratio test=35.74  on 4 df, p=3.278e-07
## n= 7918, number of events= 1032
```

We observe that, excluding the number of observations (due to splitting the data), the coefficients, the values, and the likelihood-ratio test of the two models are exactly the same, since the risk pools still look the same at any given time. As a result, we decide to implement a time change effect (we call this variable "Episode", and it represents the sections of time created by the splits) to divide the risk pool between the varying sections of time. To properly implement this time change effect to the model, we treat Episode as a stratified variable and create interaction terms between the covariates and the Episode variable. We then perform an ANOVA test to determine whether or not the hypothesis that these interactions exist is true.

```
#create time change effect (Episode variable)
divorce.split$Episode <- as.factor(divorce.split$Episode)

#model with interaction with stratification
divorce.split.fit <- coxph(Surv(tstart,years2,div)~(heduc + heblack + mixed)*strata(Episode),
data=divorce.split)
divorce.split.fit
```

```
## Call:
## coxph(formula = Surv(tstart, years2, div) ~ (heduc + heblack +
##     mixed) * strata(Episode), data = divorce.split)
##
##                               coef exp(coef) se(coef)      z        p
## heduc1                    -0.21880   0.80348  0.16864 -1.297 0.194470
## heduc2                    -0.43570   0.64681  0.28885 -1.508 0.131448
## heblack1                  -0.03789   0.96282  0.19572 -0.194 0.846514
## mixed1                     0.26631   1.30514  0.19361  1.376 0.168973
## heduc1:strata(Episode)2    0.66063   1.93602  0.19075  3.463 0.000533
## heduc2:strata(Episode)2    0.53215   1.70258  0.32243  1.650 0.098860
## heduc1:strata(Episode)3    0.51051   1.66614  0.21974  2.323 0.020169
## heduc2:strata(Episode)3    0.62013   1.85917  0.36422  1.703 0.088634
## heblack1:strata(Episode)2  0.16685   1.18157  0.22087  0.755 0.450007
## heblack1:strata(Episode)3  0.56674   1.76251  0.25739  2.202 0.027673
## mixed1:strata(Episode)2   -0.06833   0.93395  0.21907 -0.312 0.755093
## mixed1:strata(Episode)3    0.02548   1.02581  0.25367  0.100 0.919980
##
## Likelihood ratio test=55.7  on 12 df, p=1.357e-07
## n= 7918, number of events= 1032
```

```
anova(divorce.split.fit)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(tstart, years2, div)
## Terms added sequentially (first to last)
##
##                          loglik   Chisq Df Pr(>|Chi|)
## NULL                    -7844.3
## heduc                   -7835.6 17.3608  2  0.0001699 ***
```

```
## heblack                 -7830.7  9.9102  1  0.0016436 **
## mixed                   -7826.4  8.4659  1  0.0036187 **
## heduc:strata(Episode)   -7820.0 12.9580  4  0.0114827 *
## heblack:strata(Episode) -7816.6  6.7282  2  0.0345932 *
## mixed:strata(Episode)   -7816.5  0.2739  2  0.8719992
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

H0: There is no interaction between Episode and a given covariate, and there are no significant differences between the given covariate's parameters in the time periods of before 3 years, between 3 and 16 years, and longer than 16 years of marriage.

HA: There is an interaction between Episode and a given covariate, and there are significant differences between the given covariate's parameters in the time periods of before 3 years, between 3 and 16 years, and longer than 16 years of marriage.

From the ANOVA test we see that the p-values for heduc and heblack interactions have p-values of 0.0114827 and 0.0345932, respectively. Since these p-values are both less than $\alpha = 0.05$, we reject the null hypothesis and conclude that there is sufficient evidence to suggest that there is an interaction between Episode and both heduc and heblack and that there are significant differences in the heduc and heblack parameters before 3 years, between 3 and 16 years, and longer than 16 years of marriage. However, we also see that the p-value for the mixed interaction is 0.8719992, which is much greater than $\alpha = 0.05$. As such, we fail to reject the null hypothesis and conclude that there is insufficient evidence to suggest that there is an interaction between Episode and mixed or that there are significant differences in the mixed parameters before 3 years, between 3 and 16 years, and longer than 16 years of marriage. Thus we conclude that two of the three covariates, heduc and heblack, have differing impacts on the survival probability dependent on time.

Next, we work towards determining whether or not our hypothesis that the hazard rate for a marriage is high at the beginning and becomes lower after certain points in time is true. We do this by tracking the changes in the hazard ratios for each covariate between each Episode of time for said covariate in the coxph() output:

- Starting with the heduc covariate, we see that heduc1 begins with a hazard ratio of 0.80348 in the first 3 years of marriage (Episode 1). However, between the 3^{rd} and 16^{th} years (Episode 2), heduc1 has a hazard ratio of 1.93602, which indicates a very large increase in hazard rate. After the 16^{th} year (Episode 3), the hazard ratio of heduc1 is 1.66614, which is smaller than the hazard ratio between the 3^{rd} and 16^{th} years but still much larger than the hazard ratio of heduc1 in the first 3 years.
- We move on to analyzing heduc2 in a similar fashion and observe that heduc2 has a hazard ratio of 0.64681 in Episode 1, a hazard ratio of 1.70258 in Episode 2, and a hazard ratio of 1.85917 in Episode 3. In this case the hazard ratio of heduc2 increased from both Episode 1 to Episode 2 and from Episode 2 to Episode 3.
- Next we analyze the hazard ratios of the covariate heblack1, which has a hazard ratio of 0.96282 in Episode 1, a hazard ratio of 1.18157 in Episode, and a hazard ratio of 1.76251 in Episode 3. For heblack1, once again, we see a covariate's hazard ratios increasing from both Episode 1 to Episode 2 and from Episode 2 to Episode 3.
- Lastly, we look at the mixed1 covariate's hazard ratios: 1.30514 in Episode 1, 0.93395 in Episode 2, and 1.02581 in Episode 3. We observe that the mixed covariate's hazard ratios are the only ones that seem to be higher in the first 3 years of marriage. However, from the ANOVA test above, we concluded that the mixed covariate is fixed and does not depend on time. As such, the observed decrease in the mixed covariate's hazard ratios over time is not significant and cannot be used to support the hypothesis that the hazard rate for divorces is high at the beginning of a marriage and becomes lower after certain points in time.

Based on the hazard ratio analyses of each covariate above, we see that every covariate that is actually dependent on time has hazard ratios (and thus hazard rates) that are at their smallest within the first 3 years and become larger after 3 years. Thus, we conclude that our hypothesis that the hazard rate is high at the beginning of a marriage and becomes lower after certain points in time is false, and that, instead, the opposite seems to be true with the hazard rate increasing after certain points in time.

Lastly, to validate that the model with the time splits we used to test our hypothesis does in fact resemble a "correct" time-varying model, we create a model with a continuous, increasing time effect by creating time splits at each event with the goal of comparing the results of the ANOVA test output between the two models.

```
#model with increasing time effect
all.times <- divorce$years[divorce$div == 1]
length(all.times)
```

```
## [1] 1032
```

```
divorce.bigsplit <- survSplit(Surv(years2,div)~heduc + heblack + mixed,data=divorce,
                              cut=all.times,id="Subject",episode="Episode")
dim(divorce.bigsplit)
```

```
## [1] 1880697       8
```

```
divorce.bigsplit$heduc.lt <- (as.numeric(divorce.bigsplit$heduc))*
  log(divorce.bigsplit$years2)
divorce.bigsplit$heblack.lt <- (as.numeric(divorce.bigsplit$heblack))*
  log(divorce.bigsplit$years2)
divorce.bigsplit$mixed.lt <- (as.numeric(divorce.bigsplit$mixed))*
  log(divorce.bigsplit$years2)
logt.divorce.fit <- coxph(Surv(tstart,years2,div) ~ heduc + heblack + mixed + heduc.lt +
                            heblack.lt + mixed.lt,data=divorce.bigsplit)
logt.divorce.fit
```

```
## Call:
## coxph(formula = Surv(tstart, years2, div) ~ heduc + heblack +
##     mixed + heduc.lt + heblack.lt + mixed.lt, data = divorce.bigsplit)
##
##                coef exp(coef) se(coef)      z      p
## heduc1      0.06068   1.06256  0.12793  0.474 0.6352
## heduc2     -0.45385   0.63518  0.24893 -1.823 0.0683
## heblack1   -0.19394   0.82371  0.19044 -1.018 0.3085
## mixed1      0.25483   1.29024  0.18567  1.372 0.1699
## heduc.lt    0.11726   1.12441  0.05381  2.179 0.0293
## heblack.lt  0.19551   1.21593  0.08756  2.233 0.0256
## mixed.lt   -0.01246   0.98762  0.08479 -0.147 0.8832
##
## Likelihood ratio test=44.46  on 7 df, p=1.744e-07
## n= 1880697, number of events= 1032
```

```
#ANOVA test for continuous increase model
anova(logt.divorce.fit)
```

```
## Analysis of Deviance Table
##  Cox model: response is Surv(tstart, years2, div)
## Terms added sequentially (first to last)
##
##             loglik   Chisq Df Pr(>|Chi|)
## NULL       -7844.3
## heduc      -7835.6 17.3608  2  0.0001699 ***
## heblack    -7830.7  9.9102  1  0.0016436 **
## mixed      -7826.4  8.4659  1  0.0036187 **
## heduc.lt   -7824.7  3.3814  1  0.0659345 .
## heblack.lt -7822.1  5.3153  1  0.0211391 *
## mixed.lt   -7822.1  0.0216  1  0.8832434
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the two ANOVA test outputs, we see that the conclusions that there is sufficient evidence to suggest that there are significant differences in the heblack parameters over time and that there is insufficient evidence to suggest that there are significant differences in the mixed parameters over time hold true for both models. The only difference between the ANOVA tests of the two models is that while there is sufficient evidence to suggest that there are significant differences in the heduc parameters over time in the model we used to test our hypothesis, there is insufficient evidence to suggest the same conclusion in the continuous increase model. However, since the heduc interaction covariate's p-value of 0.0659345 only barely fails the ANOVA test at $\alpha = 0.05$ and passes the test at $\alpha = 0.10$, we conclude that the model we used to test our hypothesis is close enough in resemblance to the continuous time effect model to be used for hypothesis testing.

Conclusion

To conclude this project, we will summarize what we have learned by answering our Key Scientific Questions. We used Kaplan-Meier estimators and the Log Rank Tests for each of the covariates and we learned that all three covariates played a significant role in affecting the couple's survival probability. The KM estimators as well as the CoxPH showed the same conclusions. We found when building our model that the survival rate of a marriage does not linearly increase as the education amount increases. Instead, there is a dip in the survival rate of marriage when the husband has 12-15 years of education. Husbands with 16+ years of education have the highest marriage survival rates. Husbands with less than 12 years of education have the second-highest marriage survival rate. Lastly, husbands with 12-15 years of education have the lowest marriage survival rate. In regards to the husband in the marriage being African-American and the marriage being of mixed ethnicity, the hazard coefficient for both covariates is positive, meaning that the marriage survival rate decreases when the husband is African-American and if the marriage is of mixed ethnicity. We learned that the interactions between the covariates were not significant.

Furthermore, we learned that two of the three covariates, heduc and heblack, have differing impacts on the survival probability dependent on time. Moreover, we found that every covariate that is dependent on time has hazard ratios (and thus hazard rates) that are at their smallest within the first 3 years and become larger after 3 years. Thus, we found that our hypothesis that the hazard rate is high at the beginning of a marriage and becomes lower after certain points in time is false, and that, instead, the opposite seems to be true with the hazard rate increasing after certain points in time.

References

Carter, Aaron. Class lectures, PSTAT 175 Survival Analysis, University of California Santa Barbara, Santa Barbara, 2020.

"GR's Website." Princeton University, The Trustees of Princeton University, 2000, data.princeton.edu/wws509/datasets/.

Lillard and Panis (2000), aML Multilevel Multiprocess Statistical Software, Release 1.0, EconWare, LA, California. (original source extracted by Princeton University).