University of California, Santa Barbara

Department of Probability and Statistics

PSTAT 175 Final Project

# Marriage Dissolution in the U.S.

Report by: Sam Zhang, Zian He, Jordan Jang

Instructor: Drew Carter

Nov 30th, 2018

## Abstract

*In the modern United States, divorce has become a common process for married couple. In this project, we build a Cox Proportional Hazards model to see how the race and education level of a couple affects the rate of divorce over time, using the right-censored dataset of 3000 married couples.*

## Data source and Background information

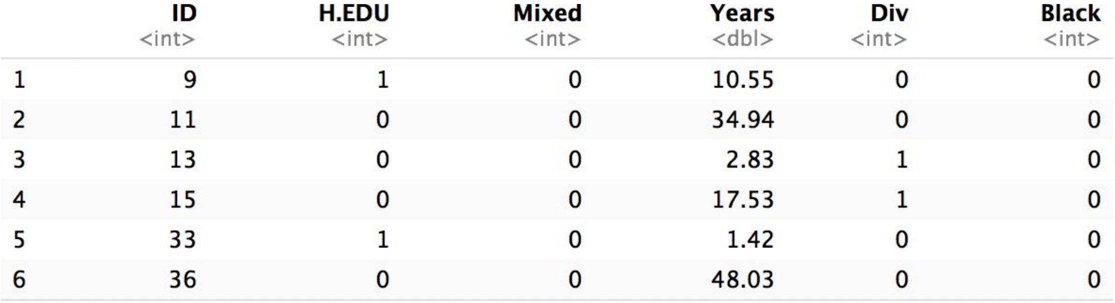
Our dataset is titled marriage dissolution in the US. It is adapted from an example in the software package aML and is based on a longitudinal survey conducted in the U.S. We obtained our data from a course website, WWS 509 of Princeton University. The dataset includes the marriage data of 3371 couples in the US. The unit of observation is the couple and the event of interest is divorce, with interview and widowhood treated as censoring events.

We have three fixed covariates: education of the husband, indicator of the couple's ethnicity: whether the couple is mixed and indicator of the husband's race: whether the husband is black.

The variables are:

- `H.EDU` (education of husband), coded as 0 = less than 12 years, 1= 12 to 15 years, 2= 16 or more years.
- `Mixed` (Whether the couple is in mixed ethnicity), coded 1 if the husband and wife have different ethnicity 0 otherwise.
- `Black` (Whether the husband is African American), coded 1 if the husband is black and 0 otherwise

- `Div` (Whether they are divorced), 1 for divorced and 0 for censored.
- `Years` (Duration of marriage), from the date of wedding to divorce or censoring.

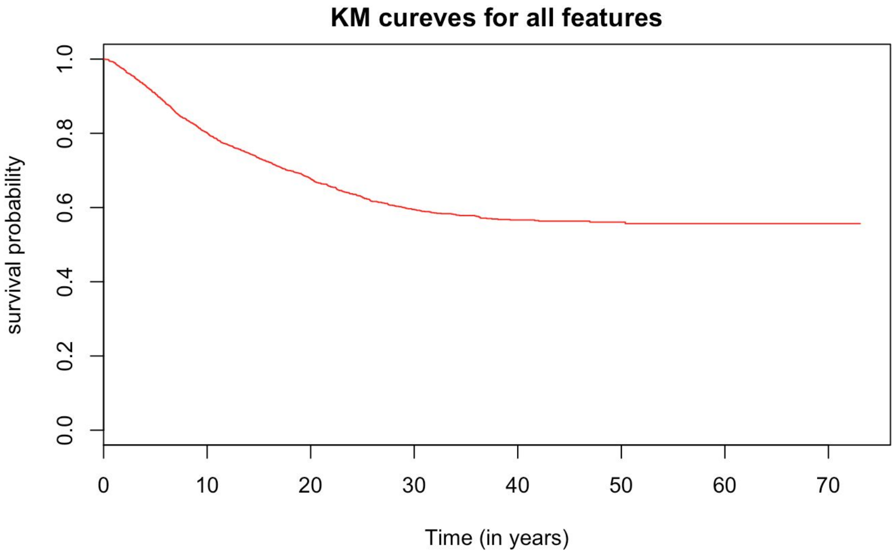


| | ID | H.EDU | Mixed | Years | Div | Black |
|---|---|---|---|---|---|---|
| | <int> | <int> | <int> | <dbl> | <int> | <int> |
| 1 | 9 | 1 | 0 | 10.55 | 0 | 0 |
| 2 | 11 | 0 | 0 | 34.94 | 0 | 0 |
| 3 | 13 | 0 | 0 | 2.83 | 1 | 0 |
| 4 | 15 | 0 | 0 | 17.53 | 1 | 0 |
| 5 | 33 | 1 | 0 | 1.42 | 0 | 0 |
| 6 | 36 | 0 | 0 | 48.03 | 0 | 0 |

## Research Question

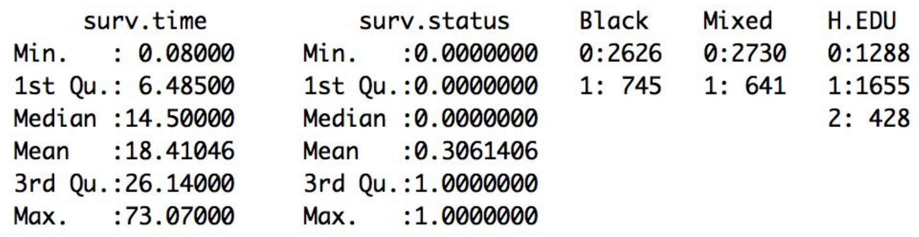
We are interested in whether ethnicity combination, husband education level, or black indicator would lead to divorce and how do they affect the marriage dissolution. In addition, we want to see if there is any interaction between any of these covariates: `Mixed`, `H.EDU` and `Black`.

Moreover, we are interested in whether there is a relationship between marriage length and divorce, then we predict the marriage length given their ethnicity combination, husband education level, and an indicator of whether husband is black.

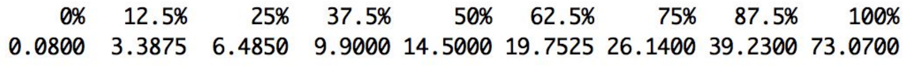


## Data Exploration

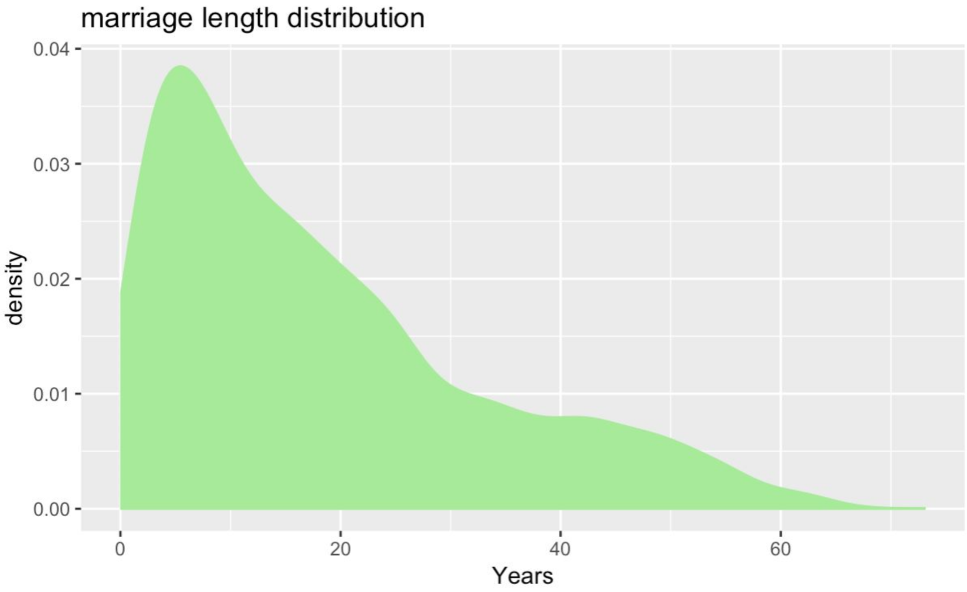
Prior to building the Cox PH model, we try to understand the dataset. By using the data summary, we find that 2626 out of 3371 husbands are non-black, 2730 out of 3371 couple have same race as his/her couple, and 50% of people have been educated between 12 and 15 years. By using the function "quantile" function in R, we plot the marriage length distribution. From the summary, we see that 62.5% of the observations end their marriage before 20 years.



```
     surv.time          surv.status        Black    Mixed    H.EDU
 Min.   : 0.08000    Min.   :0.0000000   0:2626   0:2730   0:1288
 1st Qu.: 6.48500    1st Qu.:0.0000000   1: 745   1: 641   1:1655
 Median :14.50000    Median :0.0000000                     2: 428
 Mean   :18.41046    Mean   :0.3061406
 3rd Qu.:26.14000    3rd Qu.:1.0000000
 Max.   :73.07000    Max.   :1.0000000
```

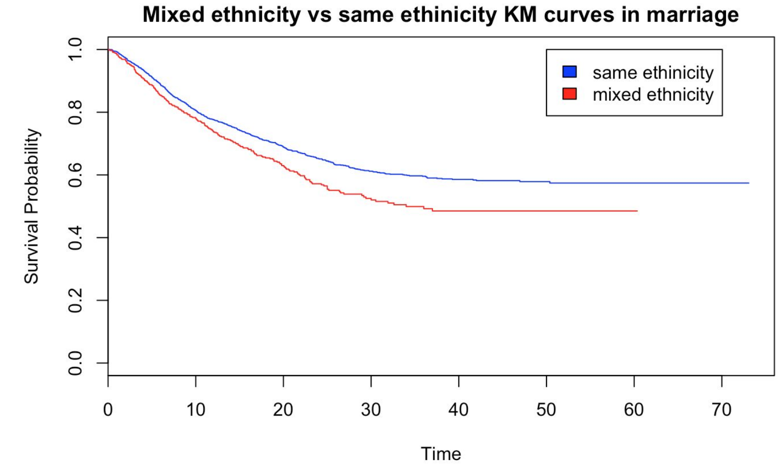


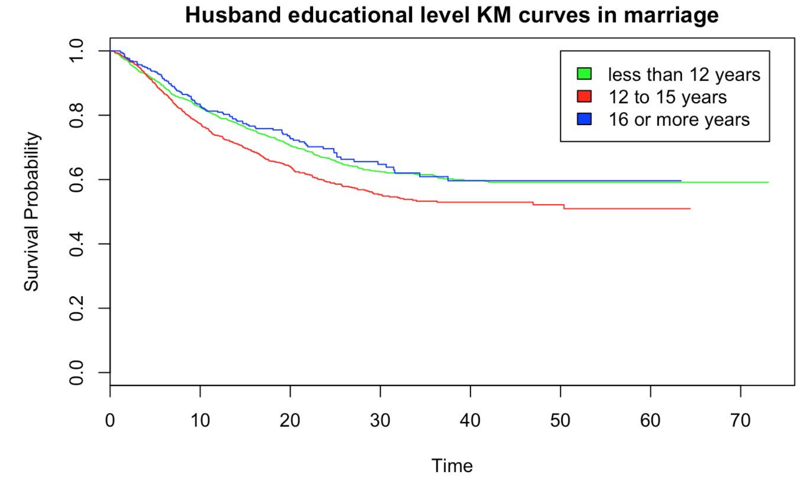
```
     0%   12.5%     25%   37.5%     50%   62.5%     75%   87.5%    100%
 0.0800  3.3875  6.4850  9.9000 14.5000 19.7525 26.1400 39.2300 73.0700
```

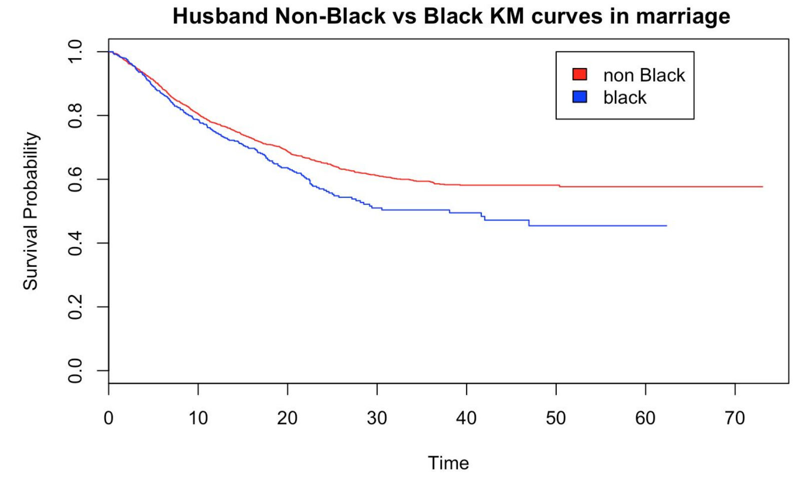


## Kaplan-Meier estimation curves

Next, we plot the Kaplan-Meier survival curves to visually analyze the effects of each covariate `Black`, `Mixed` and `H.EDU` on marriage length. Looking at the plots below, we can conclude that all three covariates do affect marriage length. Couples with non African American husband tend to have longer marriage. Couples with same ethnicity are less likely to get divorced than couples with different race. As to the husband educational, an interesting finding is that husband with middle education level are more likely to get divorced than the other groups, which may be caused by different sample sizes of each group.

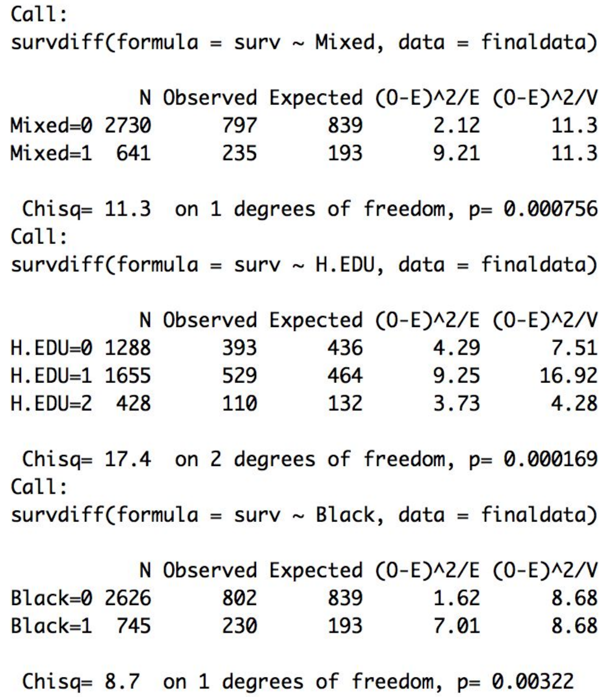

Mixed ethnicity vs same ethinicity KM curves in marriage
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Time
🟦 same ethinicity
🟥 mixed ethnicity


Husband educational level KM curves in marriage
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Time
🟩 less than 12 years
🟥 12 to 15 years
🟦 16 or more years


Husband Non-Black vs Black KM curves in marriage
Survival Probability
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70
Time
🟥 non Black
🟦 black

## Log rank test

After plotting Kaplan-Meier curves, we also conduct log rank test on each variable. All of the p values are smaller than 0.05, which indicates all of these variables have significant effect on the marriage dissolution.

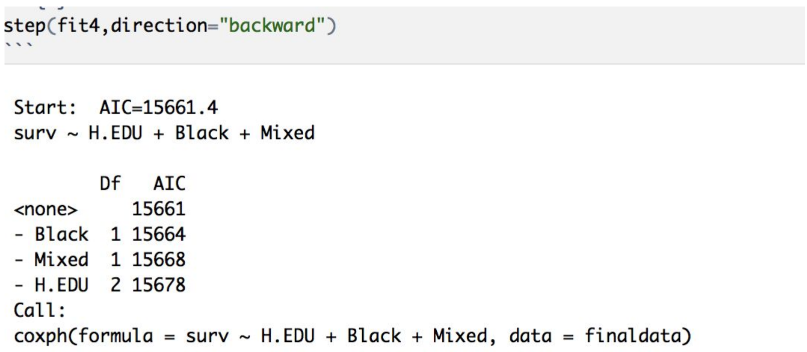


```
Call:
survdiff(formula = surv ~ Mixed, data = finaldata)

           N Observed Expected (O-E)^2/E (O-E)^2/V
Mixed=0 2730      797      839      2.12      11.3
Mixed=1  641      235      193      9.21      11.3

 Chisq= 11.3  on 1 degrees of freedom, p= 0.000756
Call:
survdiff(formula = surv ~ H.EDU, data = finaldata)

           N Observed Expected (O-E)^2/E (O-E)^2/V
H.EDU=0 1288      393      436      4.29      7.51
H.EDU=1 1655      529      464      9.25     16.92
H.EDU=2  428      110      132      3.73      4.28

 Chisq= 17.4  on 2 degrees of freedom, p= 0.000169
Call:
survdiff(formula = surv ~ Black, data = finaldata)

           N Observed Expected (O-E)^2/E (O-E)^2/V
Black=0 2626      802      839      1.62      8.68
Black=1  745      230      193      7.01      8.68

 Chisq= 8.7  on 1 degrees of freedom, p= 0.00322
```

## Model Building

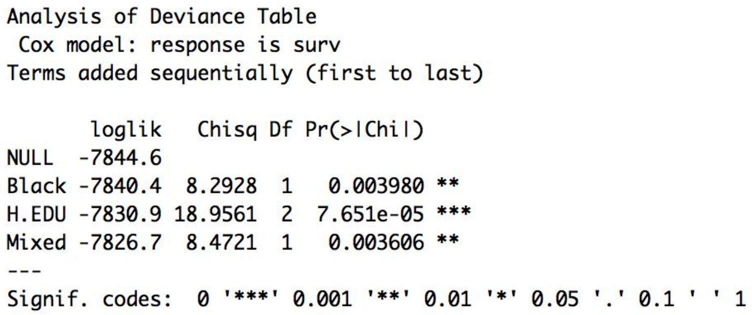
Now, we start to build our Cox PH model. We are using both backward elimination method and forward stepwise selection method to pick the right set of covariates.

First we build a full model with all three covariates. Then we use function “step” in R to apply backward elimination method. The step function stops at the full model  indicates that we should include all three covariates in our model.



```
step(fit4,direction="backward")
```
```
Start:  AIC=15661.4
surv ~ H.EDU + Black + Mixed

        Df   AIC
<none>     15661
- Black  1 15664
- Mixed  1 15668
- H.EDU  2 15678
Call:
coxph(formula = surv ~ H.EDU + Black + Mixed, data = finaldata)
```

Second, we use the  likelihood tests to select covariates. First we check the anova table for the full model with all corivates.



```
Analysis of Deviance Table
 Cox model: response is surv
Terms added sequentially (first to last)

       loglik   Chisq Df Pr(>|Chi|)
NULL  -7844.6
Black -7840.4  8.2928  1   0.003980 **
H.EDU -7830.9 18.9561  2  7.651e-05 ***
Mixed -7826.7  8.4721  1   0.003606 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

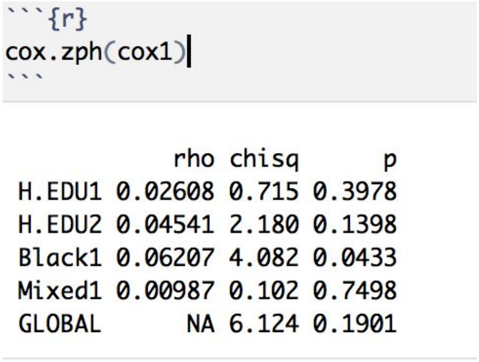
All three variables are significant. We choose the most significant variable H.EDU and then add Black to build a compared model, after running likelihood test, we should include the black covariate in our model. Then similarly we also add Mixed and conduct another likelihood test.  In the end, we get the same model as using backward elimination method.

## Model Checking

Since we are building the Cox PH model, we need to make sure all three covariates meet the Cox PH assumption. We use both residual tests and C-log-log plots to check Cox PH assumption.

## Residual tests

The function cox.zph() performs statistical tests on the PH assumption based on Schoenfeld residuals, to test for independence between residuals and time.We find that the p value of variable Black is 0.0433 which is less than 0.05. It indicates the covariate Black fail on the PH assumption check. As a result, we decide to stratify the variable Black.

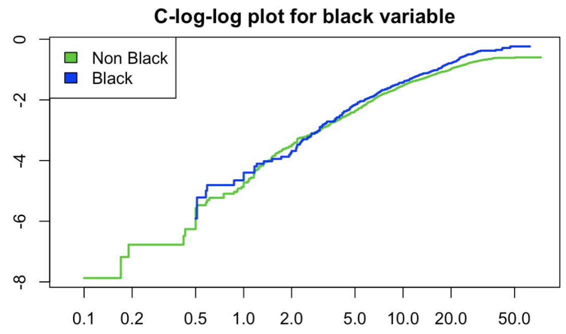


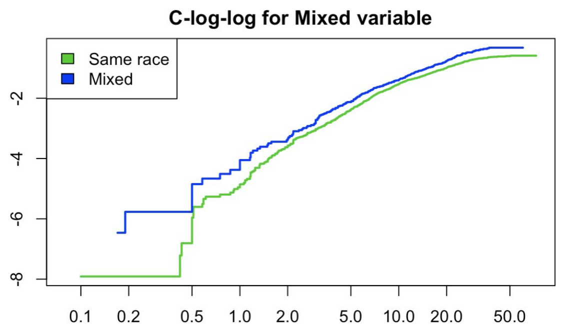
~~~
```{r}
cox.zph(cox1)
```
~~~

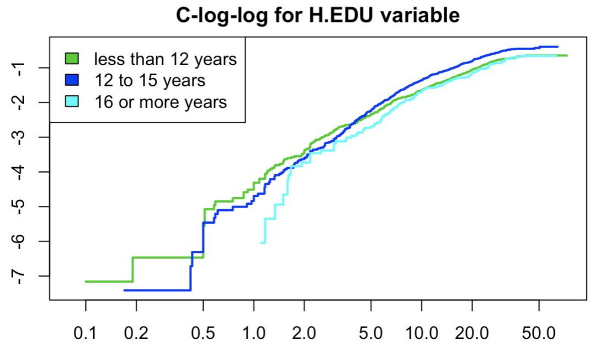
```
             rho chisq      p
H.EDU1 0.02608 0.715 0.3978
H.EDU2 0.04541 2.180 0.1398
Black1 0.06207 4.082 0.0433
Mixed1 0.00987 0.102 0.7498
GLOBAL      NA 6.124 0.1901
```

## C-log-log plot

Next we check the C-log-log plot. From the C-log-log plot, we can see in the `Black` plot the two lines get across which indicates that `Black` doesn't meet the Cox-PH assumption. What we are concerned is that `H.EDU` plot also has the crossing problem which is not shown in the `cox.zph` test. By thinking over on this problem we decide to follow `cox.zph` test since it takes all covariates into consideration when deciding which covariate doesn't meet the Cox-PH assumption, while the C-log-log plot only considers the single covariate. But in our model, we are considering the general effect of all these covariates when they are working together. Therefore, we decide to stratify the covariate `Black` only.

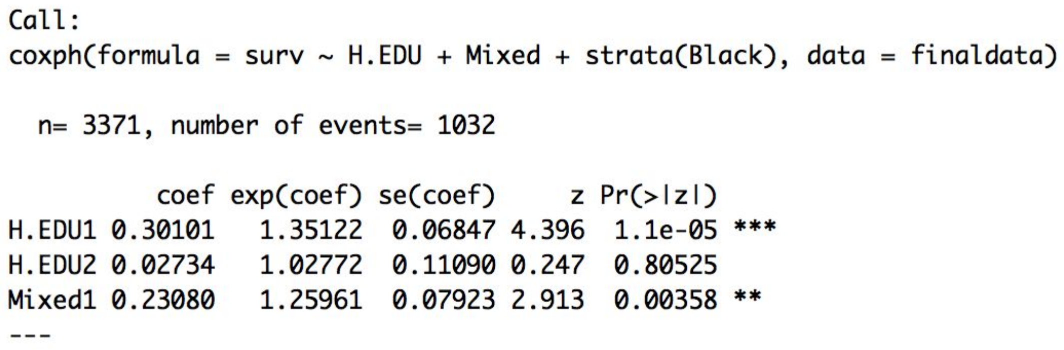






## Interaction term

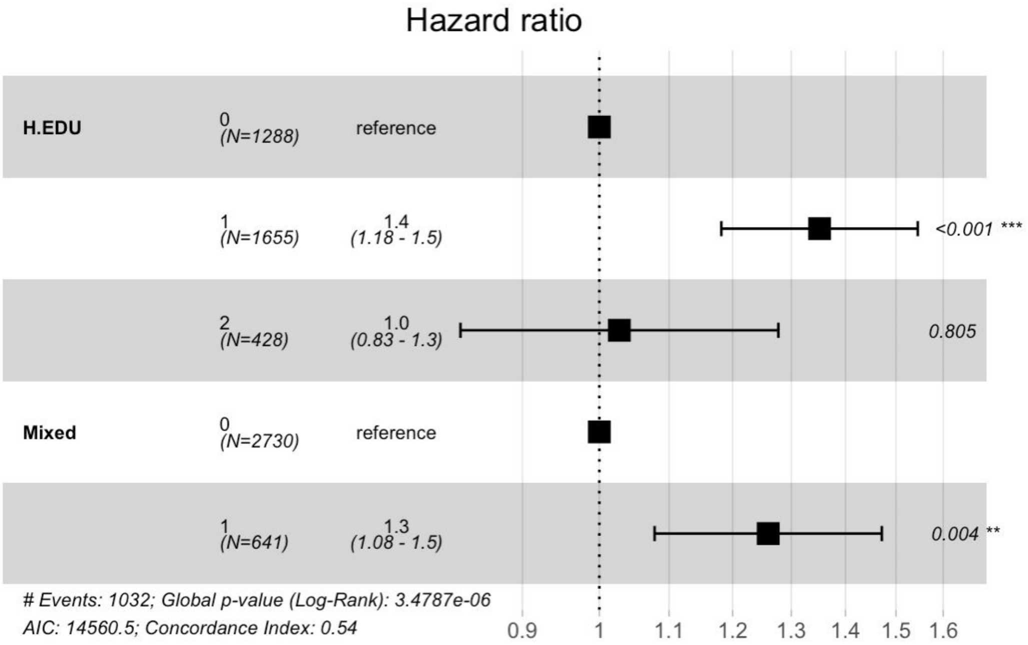
Now, we consider interaction terms in our model. There are three potential interaction terms : `Mixed*strata(Black)`, `Mixed*H.EDU` and `strata(Black)*H.EDU`. After running the likelihood ratio test on each potential interaction term, we conclude that none of them is significant since all three p values are greater than 0.05. Therefore, our final model is `Surv~H.EDU+Mixed+strata(Black)`.



```
Call:
coxph(formula = surv ~ H.EDU + Mixed + strata(Black), data = finaldata)

  n= 3371, number of events= 1032

          coef exp(coef) se(coef)     z Pr(>|z|)
H.EDU1 0.30101   1.35122  0.06847 4.396  1.1e-05 ***
H.EDU2 0.02734   1.02772  0.11090 0.247  0.80525
Mixed1 0.23080   1.25961  0.07923 2.913  0.00358 **
---
```

## Hazard Ratios and C.I.

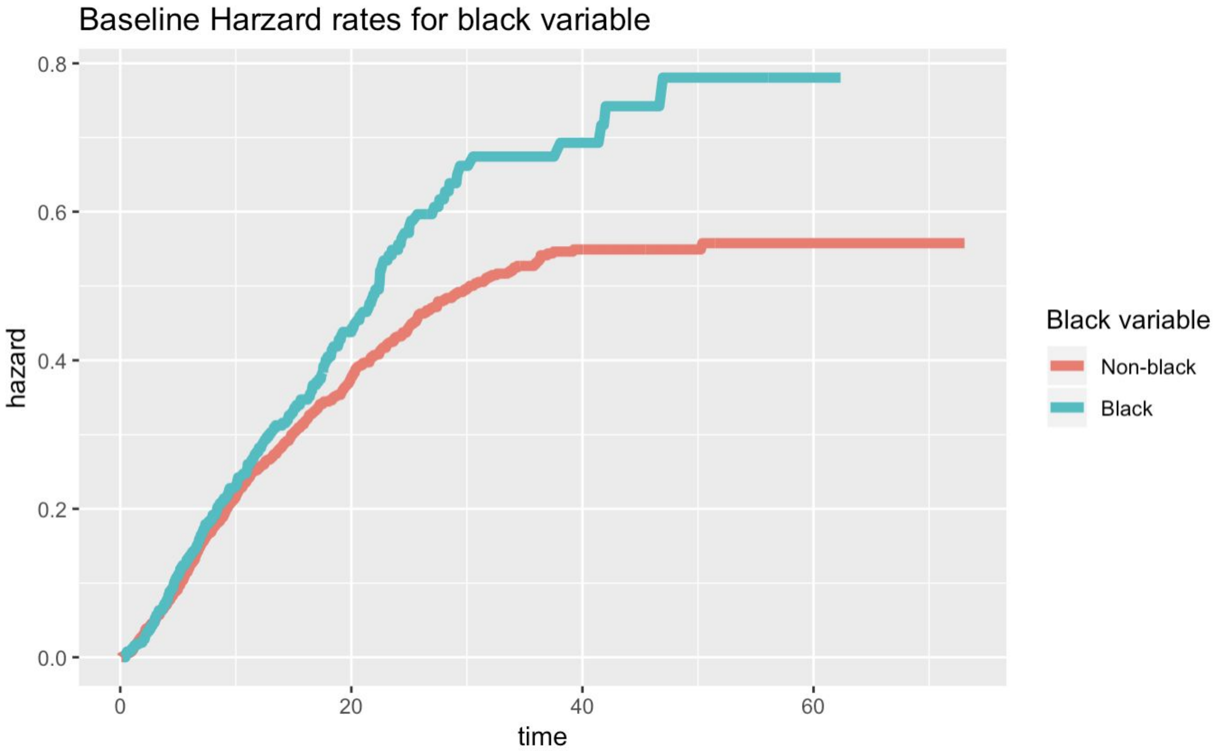
We are using `ggforest()` to visually showing the Hazard Ratios and Confidence Intervals for each covariate of different groups.



Looking at the hazard ratio chart above, we know that hazard ratio of mixed couple is centered at 1.3 and its 95% confidence interval is between 1.08 and 1.5. It indicates that mixed couples have 30% more likelihood to get divorced than couples with the same race.The hazard ratio of Husband education level 1 (12-15 years) is 1.4 (95% CI is [1.18-1.5]) compared to Husband education level 0 (less than 12 years). It indicates that husbands with education level 1 have 40% more likelihood to get divorce than husbands with education level 0. Moreover, sine the the hazard ratio of Husbands with education level 2(more than 16 years) is 1.0 (95% CI is [0.83,1.3]) compared to Husbands with education level 0, which indicates those two groups have same chance to get divorce regardless of other factors.

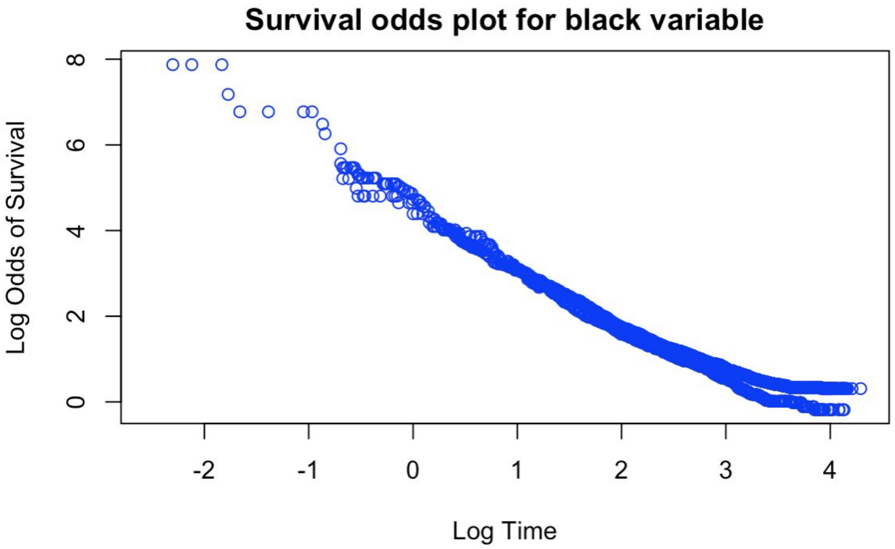
## Baseline Hazard Rates

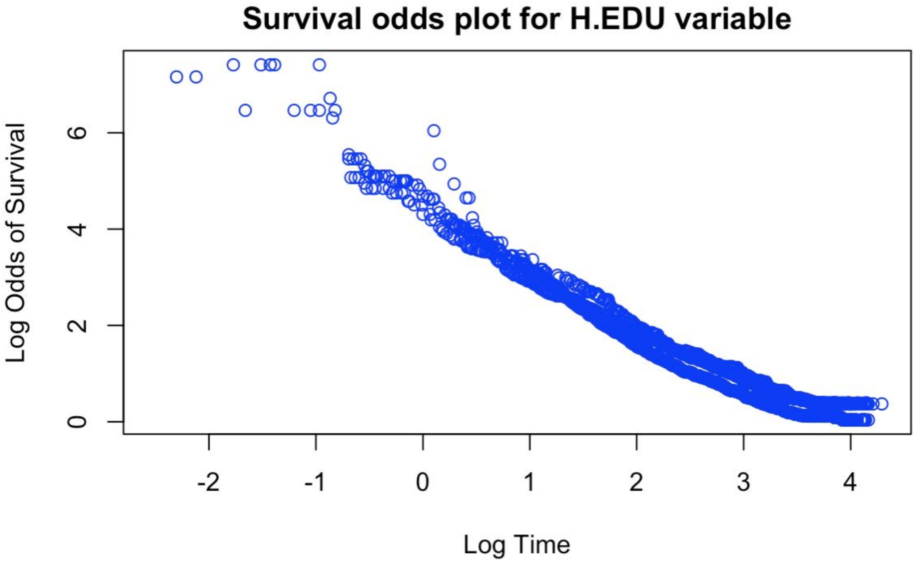
We also draw the baseline hazard plot for each strata (0 indicates non black, 1 indicates black). We can get the baseline hazards ($h_0(t)$) at different marriage length for each group (Black vs Non Black). From the plot below, we clearly see that black group has higher baseline hazard rates than non black group. As a result, the black group are more likely to get divorce than non-black group.

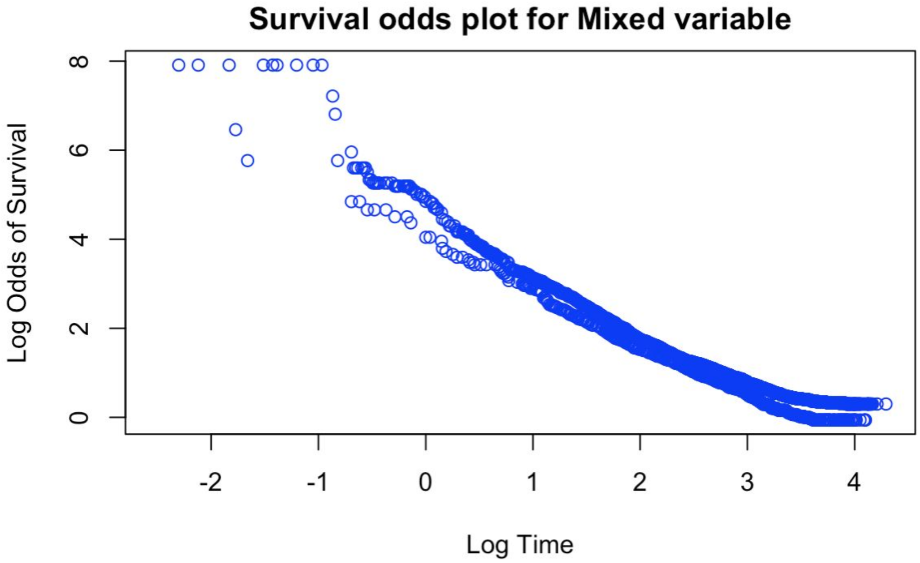


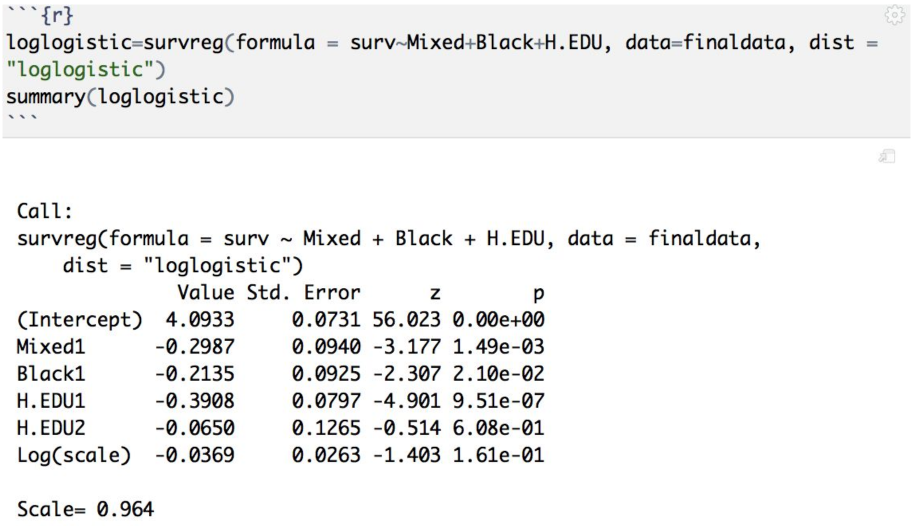
## Extension 1 - AFT model

For the first part of extension section, we build an AFT model. We try to build AFT model with log-logistic distribution so we check our assumption by drawing the survival odds plot for variables `Black`, `Mixed` and `H.EDU`. From the survival odds plot, the straight lines indicate log-logistic distribution. Therefore, we can use log-logistic distribution to model AFT model.









~~~
```{r}
loglogistic=survreg(formula = surv~Mixed+Black+H.EDU, data=finaldata, dist =
"loglogistic")
summary(loglogistic)
```
~~~

```
Call:
survreg(formula = surv ~ Mixed + Black + H.EDU, data = finaldata,
    dist = "loglogistic")
              Value Std. Error      z        p
(Intercept)  4.0933     0.0731 56.023 0.00e+00
Mixed1      -0.2987     0.0940 -3.177 1.49e-03
Black1      -0.2135     0.0925 -2.307 2.10e-02
H.EDU1      -0.3908     0.0797 -4.901 9.51e-07
H.EDU2      -0.0650     0.1265 -0.514 6.08e-01
Log(scale)  -0.0369     0.0263 -1.403 1.61e-01

Scale= 0.964
```

The estimated acceleration factor γˆ comparing black and nonblack is 0.81 (e^-0.2135). This leads to S(nonblack)=S(0.81*black). Therefore, the marriage time for black people is "accelerated" by a factor of 0.81 compared to the nonblack people based on AFT model with log-logistic distribution. As a result, the probability of non-black people lasting t years of marriage equals the probability of black people lasting 0.81 times t years of marriage.

Moreover, the estimated acceleration factor γ^ comparing mixed race and same race is 0.75 (e^-0.2987). This leads to S(same race)=S(0.75*mixed race). As a result, the probability of people with same race lasting t years of marriage equals the probability of mixed race people lasting 0.75 times t years of marriage.

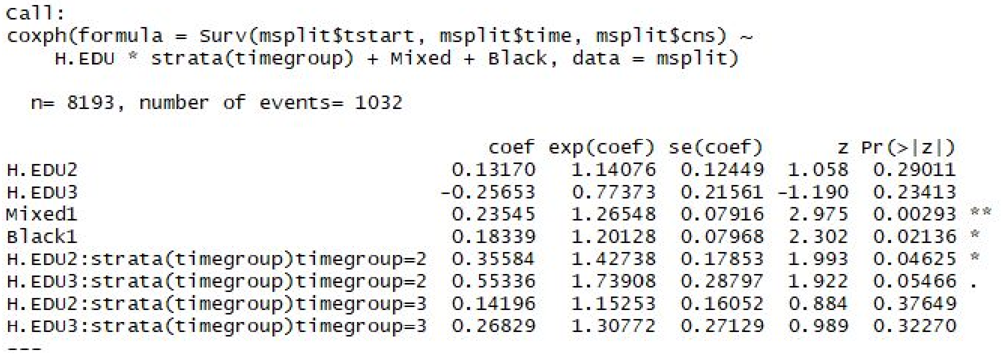
Finally, the estimated acceleration factor γ^ comparing education1(12-15 years education) and education0 (less than 12 years education) is 0.68 (e^-0.3908). This leads to S(education0)=S(0.68*education1). As a result, the probability of people with education0 lasting t years of marriage equals the probability of people with education1 lasting 0.68 times t years of marriage.

All in all, under the AFT model with log-logistic distribution, the covariate H.EDU (husband's education) has the highest effect on marriage length.

## Extension 2 - Time Varying parameters

For the second part of extension, we now split each married couple into different time groups and evaluate. By dividing the time periods, we intend to explore the possibility of racial and educational factors affecting chance of divorce differently depending on different time phases of marriage. The `Cluster` function was not applied because doing the analysis using robust variances did not make much of difference in this case. Then, we also use `anova` function to see if any of the variables show significant differences according to their time groups.

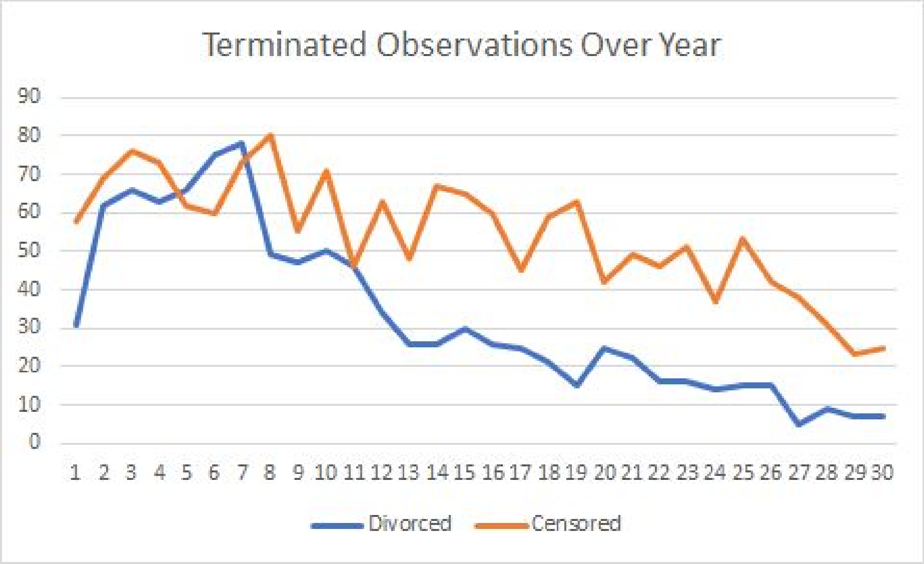
First, we arbitrarily set up the cutoffs at 5 and 10 years, since these two time periods are common for time periods used in different studies. Out of 3371 couples at total, 2088 couples had time variable of bigger than 10 years, 746 have between 5 and 10, and 636 have less than 5. Therefore, we conclude that the sample size in each time group is significant enough to perform individual Cox-PH regressions.



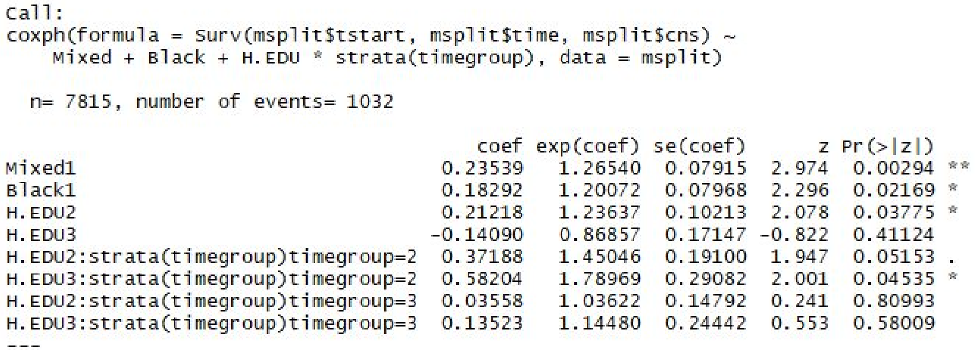
```
Call:
coxph(formula = Surv(msplit$tstart, msplit$time, msplit$cns) ~ 
    H.EDU * strata(timegroup) + Mixed + Black, data = msplit)

  n= 8193, number of events= 1032 

                                    coef exp(coef) se(coef)      z Pr(>|z|)   
H.EDU2                           0.13170   1.14076  0.12449  1.058  0.29011   
H.EDU3                          -0.25653   0.77373  0.21561 -1.190  0.23413   
Mixed1                           0.23545   1.26548  0.07916  2.975  0.00293 **
Black1                           0.18339   1.20128  0.07968  2.302  0.02136 * 
H.EDU2:strata(timegroup)timegroup=2  0.35584   1.42738  0.17853  1.993  0.04625 * 
H.EDU3:strata(timegroup)timegroup=2  0.55336   1.73908  0.28797  1.922  0.05466 . 
H.EDU2:strata(timegroup)timegroup=3  0.14196   1.15253  0.16052  0.884  0.37649   
H.EDU3:strata(timegroup)timegroup=3  0.26829   1.30772  0.27129  0.989  0.32270   
---
```

After using `survSplit` on the original data, we stratify the time groups, then apply them to each of `Black`, `Divorce`, and `H.EDU` to find if an interaction term is applicable to any of these variables. From the results, we see that `H.EDU` somewhat affects the coefficient when the husband has done between 12 to 16 years of education (i.e. when the variable equals 1), and the marriage belongs to the time group of between 5 to 10 years (i.e. when `timegroup` variable equals 2). However, the p-value of 0.0463 is very close to the criteria of 0.05, which means the effect is not highly significant. Any of the other variables did not show a significant deviation when applied an interaction term with stratified time groups, therefore we conclude that any of the covariates we discussed does not vary in its effect if the cutoff is set up at 5 and 10 years.

Because the arbitrary cutoffs of time periods did not result in any significant difference from covariates, we now look at the number of divorces and censored observations year by year to determine which points are appropriate for cutoffs.

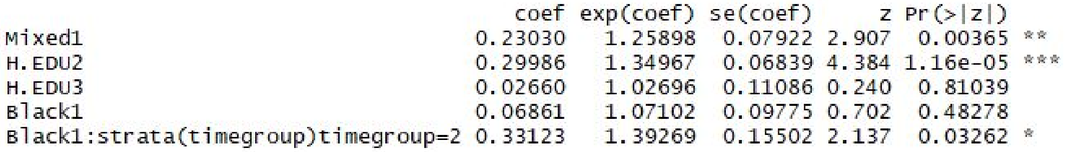


Overall, we can see that the number of divorce declines faster over time than number of censored observations. A point we are looking for should show a significant decline compared to its previous point, because a separation on divorce rate can signify a different stage of marriage based on its higher stability. Indeed, there are certain years where the number of divorce displays a steep decline: between year 7 and 8, and between year 11 and 12. We can consider these two points as the "tipping points" of the marriage. Out of 3372 total couples, 471 censored observations and 441 divorces occurred by year 7. Between year 8 and 11, 252 censored observations and 192 divorces occurred. Again, the sample size of each time group are significantly big. We perform multiple Cox-PH regressions, starting from cutoffs at 7 and 11 years. `anova` function will be utilized in order to check which cutoffs work the best for individual variables.



```
Call:
coxph(formula = Surv(msplit$tstart, msplit$time, msplit$cns) ~
    Mixed + Black + H.EDU * strata(timegroup), data = msplit)

  n= 7815, number of events= 1032

                                      coef exp(coef) se(coef)      z Pr(>|z|)
Mixed1                             0.23539   1.26540  0.07915  2.974  0.00294 **
Black1                             0.18292   1.20072  0.07968  2.296  0.02169 *
H.EDU2                             0.21218   1.23637  0.10213  2.078  0.03775 *
H.EDU3                            -0.14090   0.86857  0.17147 -0.822  0.41124
H.EDU2:strata(timegroup)timegroup=2 0.37188   1.45046  0.19100  1.947  0.05153 .
H.EDU3:strata(timegroup)timegroup=2 0.58204   1.78969  0.29082  2.001  0.04535 *
H.EDU2:strata(timegroup)timegroup=3 0.03558   1.03622  0.14792  0.241  0.80993
H.EDU3:strata(timegroup)timegroup=3 0.13523   1.14480  0.24442  0.553  0.58009
---
```

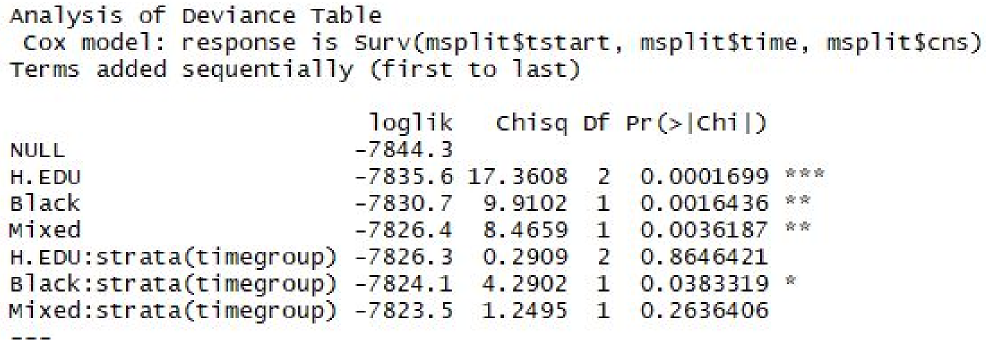
Once again, out of all 3 covariates, H.EDU was the only covariate that showed a significant coefficient based on a different timegroup strata, and its p-value indicated a barely significant as well. From the result, we conclude that splitting up the variables with cutoffs at 7 and 11 does not change the much. However, we notice that the overall p-values in other covariates showed a decrease, and would like to continue by dropping 7-year cutoff, since the small chunk between 7 and 11 may not be evenly splitting the data, especially because there are relatively low number of black and mixed couples.



```
                                    coef exp(coef) se(coef)     z Pr(>|z|)
Mixed1                           0.23030   1.25898  0.07922 2.907  0.00365 **
H.EDU2                           0.29986   1.34967  0.06839 4.384 1.16e-05 ***
H.EDU3                           0.02660   1.02696  0.11086 0.240  0.81039
Black1                           0.06861   1.07102  0.09775 0.702  0.48278
Black1:strata(timegroup)timegroup=2 0.33123   1.39269  0.15502 2.137  0.03262 *
```

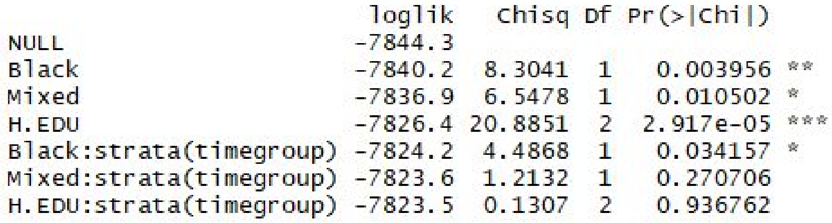
We now build another Cox-PH model with only 11 year mark remaining as the cutoff for the data. While H.EDU and Mixed covariates did not show any significance on their coefficients, we observe a fairly significant p-value for Black covariate after 11 years of marriage. The 95% interval for coefficient is (1.0278, 1.887). This indicates that after 11 years of marriage, black couples are significantly more likely to divorce compared to the non-black couples. To confirm this model, we perform a log-likelihood test using anova function.

Since anova test is sensitive to order, we first run the function with all covariates stratified by timegroup to determine the order of terms. The terms with lowest p-values should be included first, and we remove the covariate that comes last until all log-likelihood test terms suggest significance.



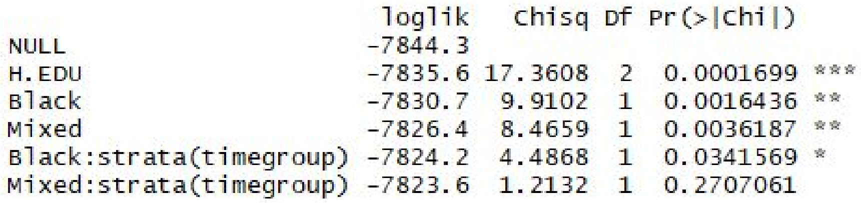
```
Analysis of Deviance Table
 Cox model: response is Surv(msplit$tstart, msplit$time, msplit$cns)
Terms added sequentially (first to last)

                          loglik   Chisq Df Pr(>|Chi|)
NULL                     -7844.3
H.EDU                    -7835.6 17.3608  2  0.0001699 ***
Black                    -7830.7  9.9102  1  0.0016436 **
Mixed                    -7826.4  8.4659  1  0.0036187 **
H.EDU:strata(timegroup) -7826.3  0.2909  2  0.8646421
Black:strata(timegroup) -7824.1  4.2902  1  0.0383319 *
Mixed:strata(timegroup) -7823.5  1.2495  1  0.2636406
---
```

This result suggests that the stratified covariates should go in order of Black, Mixed, and H.EDU. After ordering the stratified terms correctly, we use anova function once again:

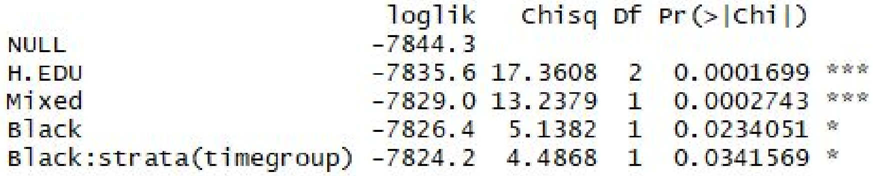


```
                          loglik   Chisq Df Pr(>|Chi|)
NULL                     -7844.3
Black                    -7840.2  8.3041  1   0.003956 **
Mixed                    -7836.9  6.5478  1   0.010502 *
H.EDU                    -7826.4 20.8851  2  2.917e-05 ***
Black:strata(timegroup) -7824.2  4.4868  1   0.034157 *
Mixed:strata(timegroup) -7823.6  1.2132  1   0.270706
H.EDU:strata(timegroup) -7823.5  0.1307  2   0.936762
```

We drop H.EDU covariate with a whopping 0.937 p-value, and run the function again. If Mixed does not show a significant p-value, we repeat the process and test Black's significance. In order for the model to agree with the prior conclusion of only variable Black significant, the model should drop Mixed and show Black as a significant covariate.



```
                           loglik   Chisq Df Pr(>|Chi|)
 NULL                      -7844.3
 H.EDU                     -7835.6 17.3608  2  0.0001699 ***
 Black                     -7830.7  9.9102  1  0.0016436 **
 Mixed                     -7826.4  8.4659  1  0.0036187 **
 Black:strata(timegroup)  -7824.2  4.4868  1  0.0341569 *
 Mixed:strata(timegroup)  -7823.6  1.2132  1  0.2707061
```



```
                           loglik   Chisq Df Pr(>|Chi|)
NULL                      -7844.3
H.EDU                     -7835.6 17.3608  2  0.0001699 ***
Mixed                     -7829.0 13.2379  1  0.0002743 ***
Black                     -7826.4  5.1382  1  0.0234051 *
Black:strata(timegroup)  -7824.2  4.4868  1  0.0341569 *
```

We conclude that log-likelihood tests agree with my prior conclusion based on Cox-PH model coefficients, and confirm that the model should only stratify `Black` variable by `timegroup`.

## Conclusion/Discussion

For this project, we were given a right-censored data of more than 3,000 married couples along with the covariates that indicate their race and education levels. We plotted Kaplan-Meier curve to check if any of the covariates affects divorce rate over time. With the curves indicating that all 3 covariates have significant effect on divorce rates, we then performed log-rank test to confirm the significance. With the result, we confirm that the dataset satisfies the model's assumption through residual test and C-log-log plot. However, some of C-log-log plots indicate that the model does not meet the proportional hazards assumption, which resulted in stratification of a variable (`Black`) in our model. We then tested for possible interaction terms, and concluded that none of them should be included in the model. Finally, we estimated the coefficients and confidence intervals for the two non-stratified variables, and baseline hazard rates for the stratified variable. The results indicated that divorce happens more often if the husband is college-educated (but not to postgraduate level), if the couple is black, and if the black husband marries non-black wife.

The extension was divided into two parts: AFT and time varying parameters. In the first part, we built an AFT model, estimating acceleration factor $\gamma\hat{}$ for the `Black` covariate. The result suggested that black couples have an accelerated "speed" to divorce compared to non-black couples. For the second part, we split up each couple's time period to see if the covariates have different magnitude of effect over different time periods after marriage. From the resulting cox-PH model, we conclude that black couples are more likely to divorce compared to non-black couples once they have spent 11 years of marriage, and confirmed the model with a stepwise log-likelihood test.