Remi HARDY – S18 DSTI – Survival analysis

1. Dataset description

The data is based on a survey conducted in the US.
Source http://data.princeton.edu/wws509/datasets/#divorce

The event of interest is the divorce among couples in the US.
The dataset comprises 3371 couples with the following information:

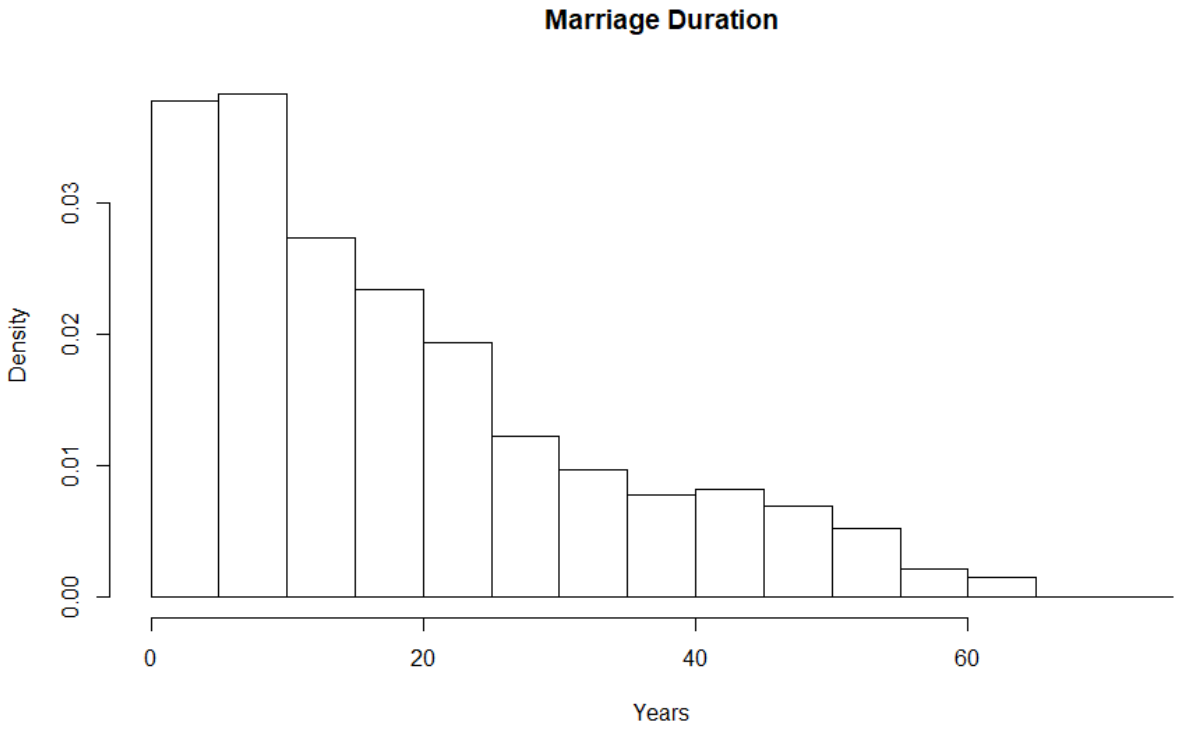
- **M_educ**: education level of the husband, coded
  0 = less than 12 years (count 1288/3371)
  1= 12 to 15 years (count 1655/3371)
  2 = 16 or more years (count 428/3371)
- **M_black**: if the husband is black or not
  1 = black (count 745/3371)
  0 = not black (count 2626/3371)
- **Mixed**: if the couple has mixed ethnicity
  1 = mixed couple (count 641/3371)
  0 = otherwise (count 2730/3371)
- **Years**: duration of the marriage, from wedding to divorce
- **Div**: event (ie divorce) indicator
  1 = divorced (count 1032/3371)
  0 for censoring (count 2339/3371)

2. Descriptive statistics

The dataset comprises the data of 3371 couples

First, we can explore the distribution of divorced and non-divorced couples, we get:
Divorced = 1032 (~30%)
Non-divorced = 2339 (~70%)

We can also check the distribution of the marriage duration, we get:



With:

Statistics	Complete Dataset
min	0.08
max	73.06
mean	18.4
median	14.49

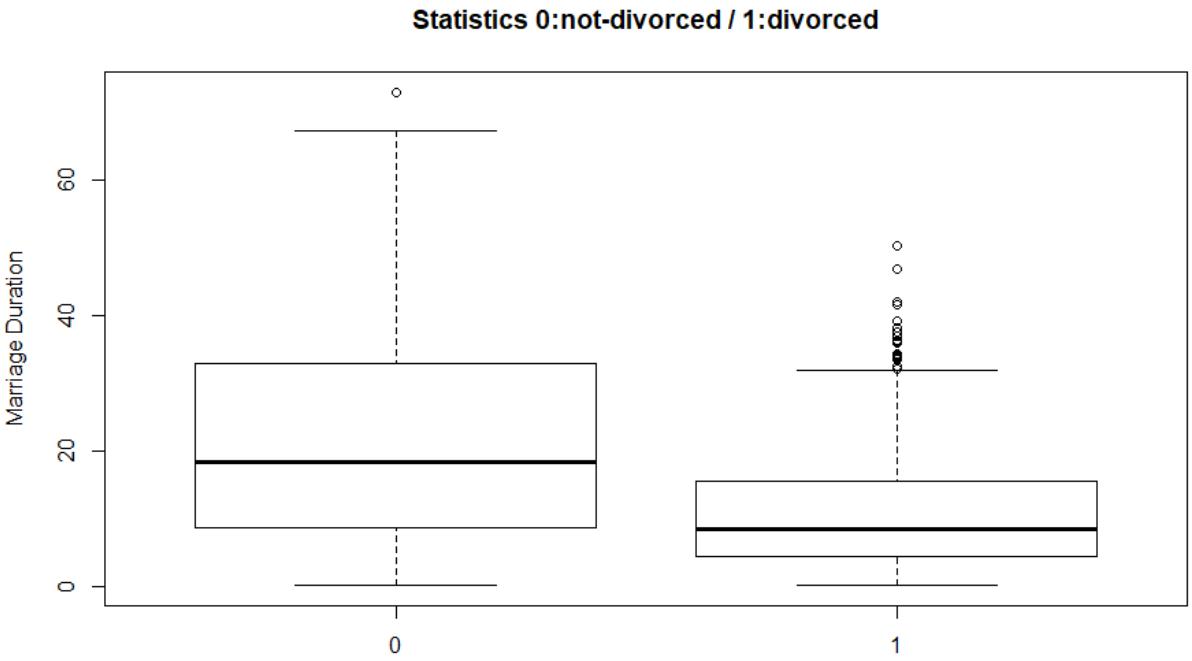
Notice: these values correspond only to the marriage duration, that the people - who answered the survey -reported.
In principle, they are not related to the divorce event: all the couples who reported a marriage duration lower than 14.489 years (50% of the sample size) may or may not have divorced.
To figure this out, we can get the number of divorces (div=1) among these first 50%:
out of 1685 couples, we get 744 divorced and 941 not-divorced.
This – however – brings an interesting information: out of the 1032 divorced couples of the dataset, 744 (~70%) divorced before 14.489 years, which is a significant proportion.

To go further into the analysis and confirm the above observation, we may generate 2 subsets out of the original one: one with divorced couples only, and one with not-divorced couples only, bringing the following statistics together:

Statistics	Divorced subset	Not-divorced subset
min	0.10	0.07
max	50.37	73.06
mean	10.75	21.78
median	8.34	18.23

As we can see - confirming the first observation - the couples who reported a divorce have mean and median marriage duration that is much lower than the other couples (who have not reported a divorce yet).

An easier method to visualize this, is to use the boxplot function from R (same 2 subsets d_div and d_ndiv in R code):



Next, we can explore education and ethnicity variables:

Education and ethnicity have roughly the same distribution in the 2 groups composed of divorced couples and not-divorced couples. Although it does not help us to analyze the influence of these parameter on the divorce event for the moment, at least it does not introduce a bias due to an over-representation of one or the other variable.

EDUCATION	Dataset	Divorced subset	Not-divorced subset
12y	38%	38%	38%
12y-15y	49%	51%	48%
15+ y	13%	11%	14%

BLACK	Dataset	Divorced subset	Not-divorced subset
0	78%	78%	78%
1	22%	22%	22%

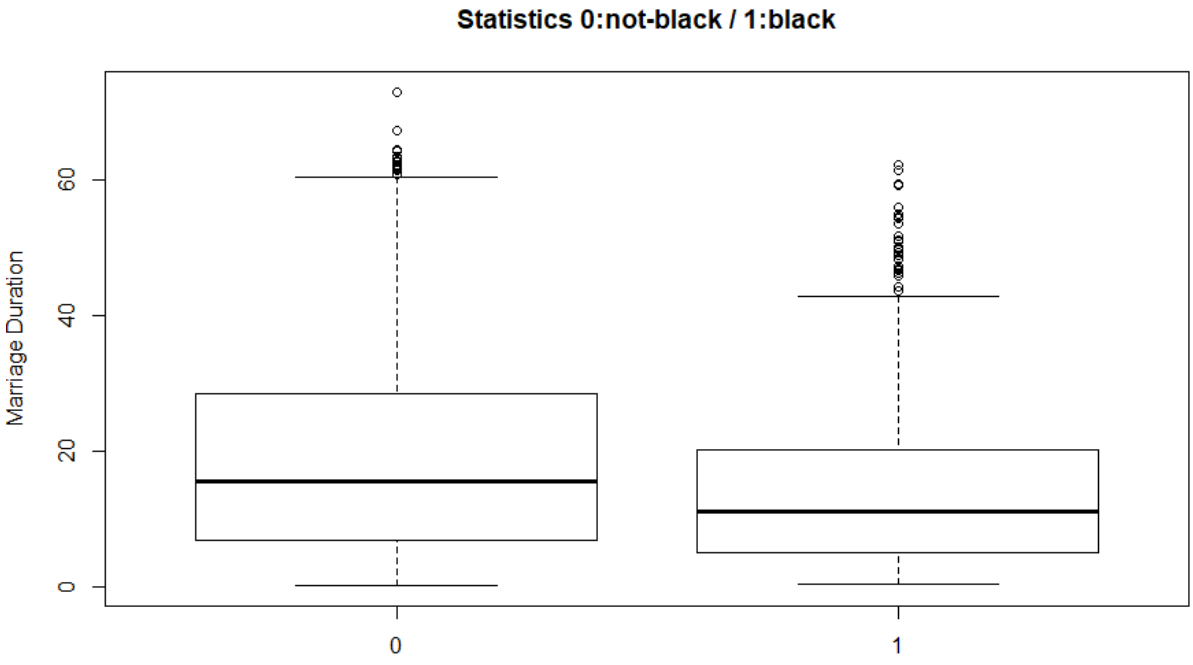
MIXED	Dataset	Divorced subset	Not-divorced subset
0	81%	77%	83%
1	19%	23%	17%

Some other observations:
Black ethnicity males in the couples represent 22%. This is quite low, we will see in the survival analysis if this variable has some influence despite a relatively small proportion.

Black ethnicity males in the couples received a shorter education as depicted in the table below:

EDUCATION	Black	Not Black
12y	49%	35%
12y-15y	46%	50%
15+ y	5%	15%

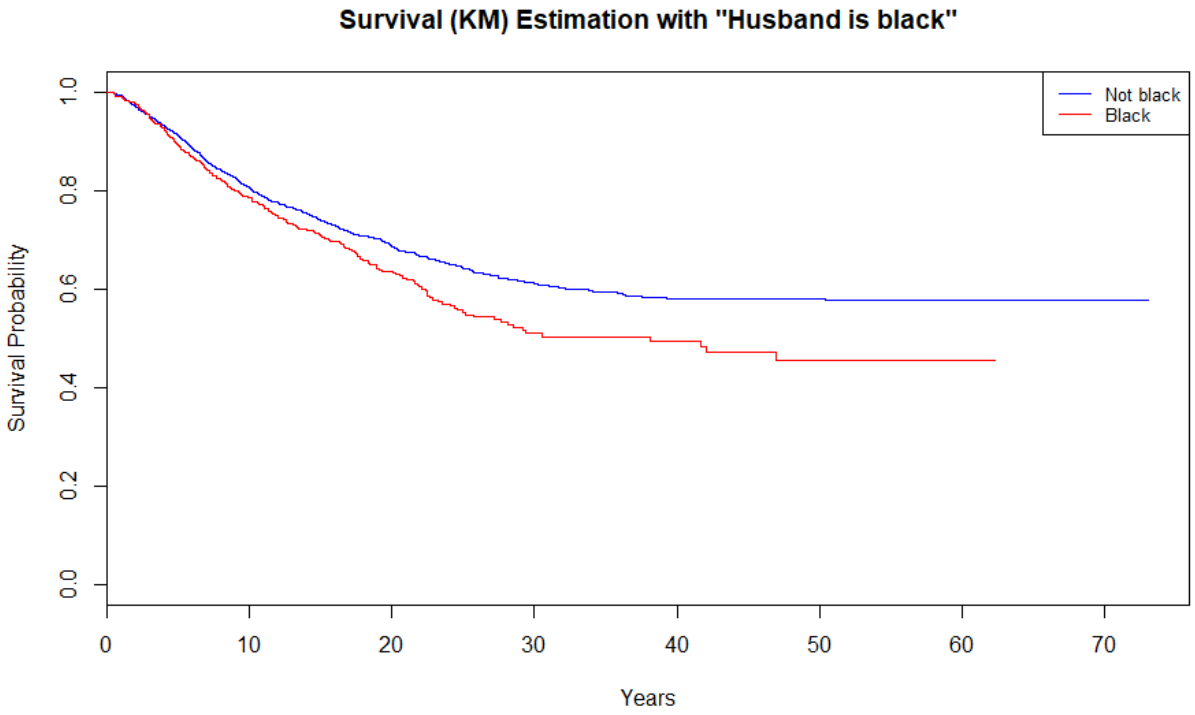
For the couples whose husband is black, we observe a marriage duration that seems to be slightly lower than for the other couples:



3. Survival analysis

The idea is to explore the time to divorce event depending on the education and ethnicity variables. To do that, we will use the Kaplan-Meier estimator and the Logrank Test.

a) "Husband is black" variable



From the scheme above, we can notice that the survival probability if "husband is black" in a couple is lower than that "husband is not black"

Another indication is the median survival probability that is 38.1 years for "husband is black" while it is NA for "husband is not black". NA means that the median value of survival probability of 50% is NEVER reach. So, the case where "husband is not black" is more favorable wrt time to divorce event.

```
                n events median 0.95LCL 0.95UCL
d$m_black=0 2626    802     NA      NA      NA
d$m_black=1  745    230   38.1    25.2      NA
```

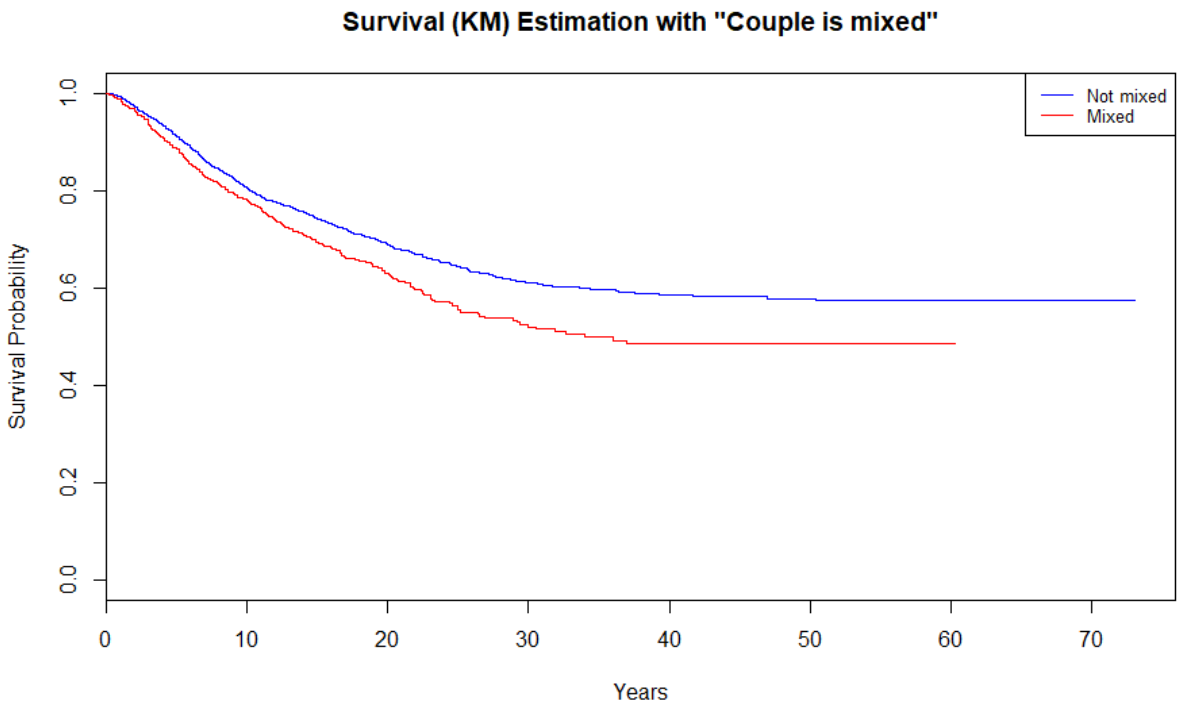
The p-value of the Logrank test (H0: groups are similar wrt survival) is small, below significance level of 0.05, meaning the 2 groups are significantly different wrt the time to divorce.

```
Call:
survdiff(formula = Surv(d$years, d$div) ~ d$m_black, data = d)

                N Observed Expected (O-E)^2/E (O-E)^2/V
d$m_black=0 2626      802      839      1.62      8.69
d$m_black=1  745      230      193      7.03      8.69

 Chisq= 8.7  on 1 degrees of freedom, p= 0.003
```

b) “Couple is mixed ethnicity” variable



From the scheme above, we can notice that the survival probability if “couple is mixed ethnicity” is lower than that “the couple is not of mixed ethnicity”.

```
              n events median 0.95LCL 0.95UCL
d$mixed=0 2730    797     NA      NA      NA
d$mixed=1  641    235     34    26.5      NA
```

Same remark regarding the median survival probability as previously: the median value for survival probability is 34 years, if the couple is of mixed ethnicity, while S=50% is never reached for a couple of the same ethnicity.

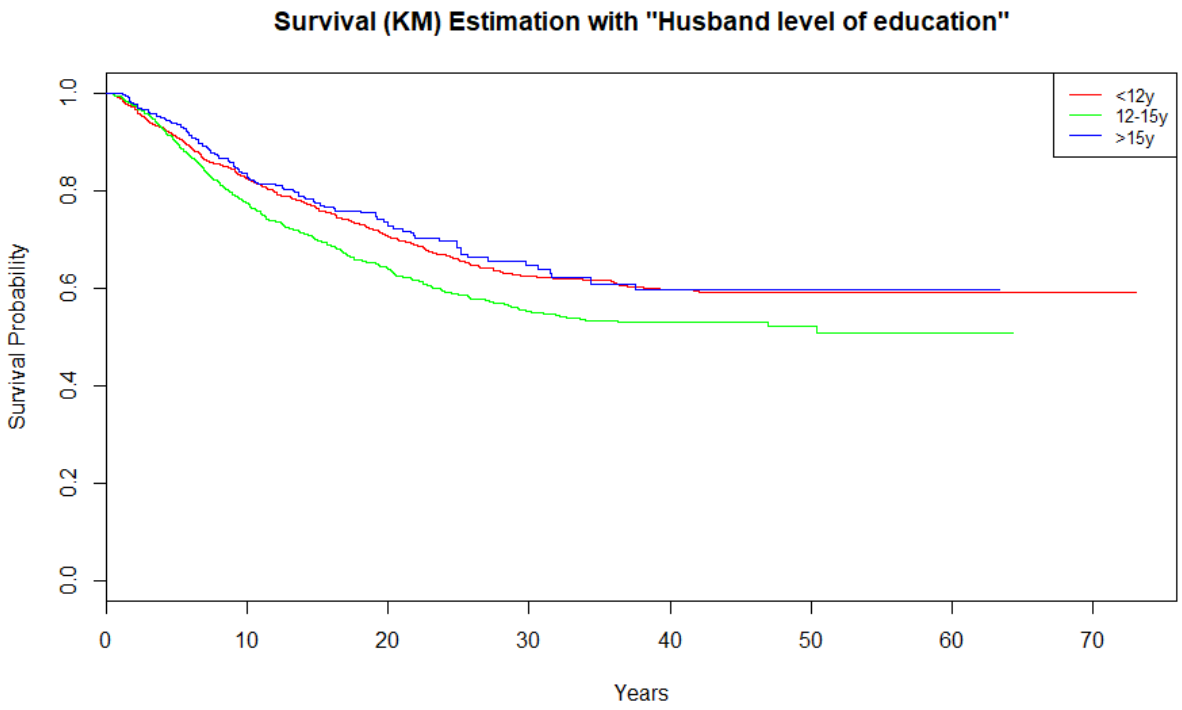
The p-value of the Logrank test (H0: groups are similar wrt survival) is small, below significance level of 0.05, meaning the 2 groups are significantly different wrt the time to divorce.

```
Call:
survdiff(formula = Surv(d$years, d$div) ~ d$mixed, data = d)

              N Observed Expected (O-E)^2/E (O-E)^2/V
d$mixed=0 2730      797      839      2.12      11.3
d$mixed=1  641      235      193      9.22      11.3

 Chisq= 11.3  on 1 degrees of freedom, p= 8e-04
```

c) Husband education level



Regarding the influence of the level of education of the husband in the couple, the class corresponding to “12-15y” has a lower survival probability.
This may come from the fact that this group has the largest proportion in the dataset.
This does not bring really any valuable information.
So, a more refined analysis is required here.

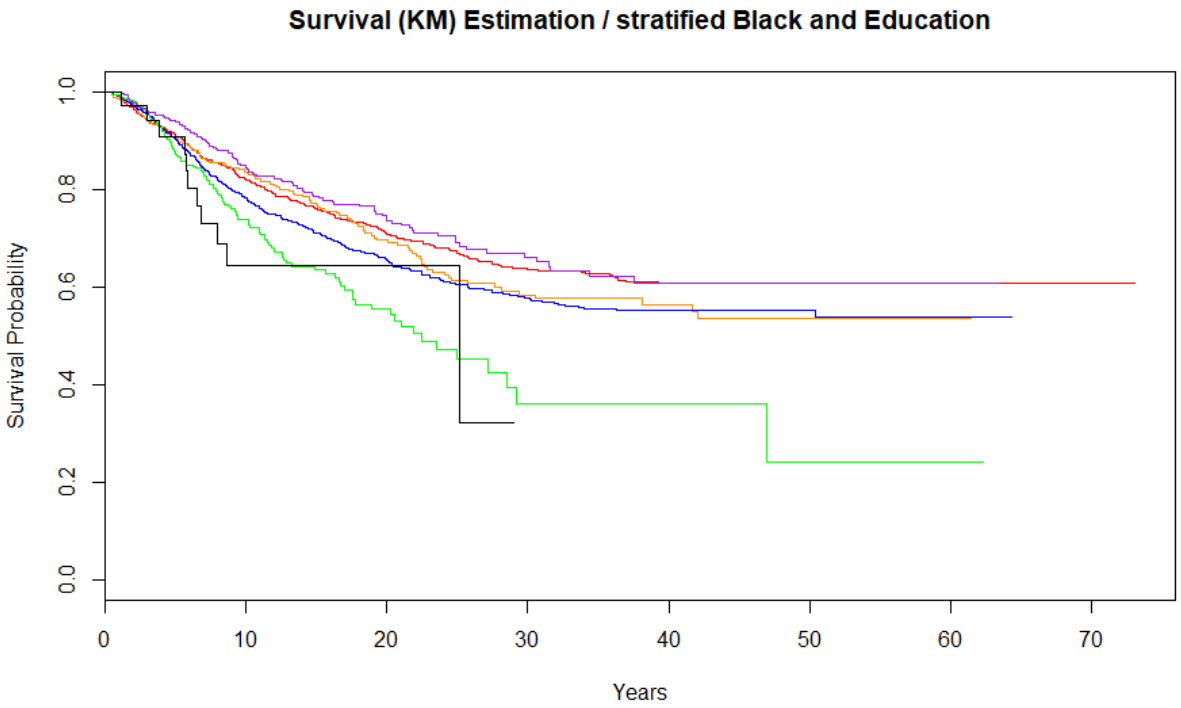
The median value of 50% is never reached for any of the education level classes (NAs).

```
               n events median 0.95LCL 0.95UCL
d$m_educ=0 1288    393     NA      NA      NA
d$m_educ=1 1655    529     NA      34      NA
d$m_educ=2  428    110     NA      NA      NA
> survdiff(Surv(d$years,d$div)~ d$m_educ,data=d)
Call:
survdiff(formula = Surv(d$years, d$div) ~ d$m_educ, data = d)

               N Observed Expected (O-E)^2/E (O-E)^2/V
d$m_educ=0 1288      393      436      4.30      7.51
d$m_educ=1 1655      529      463      9.26     16.93
d$m_educ=2  428      110      132      3.73      4.28
```

d) Stratified test of husband education level and ethnicity (husband is black)

As, the only information of education level does not bring any relevant information, it may be interesting to combine its analysis with the ethnicity variable. We can use a stratified Logrank test for that purpose.



As adding a legend on the scheme is not readable, I preferred to explicit it here:
The 3 classes of education levels are combined with ethnicity variable "husband is black"
We can consider 3 pairs of lines:

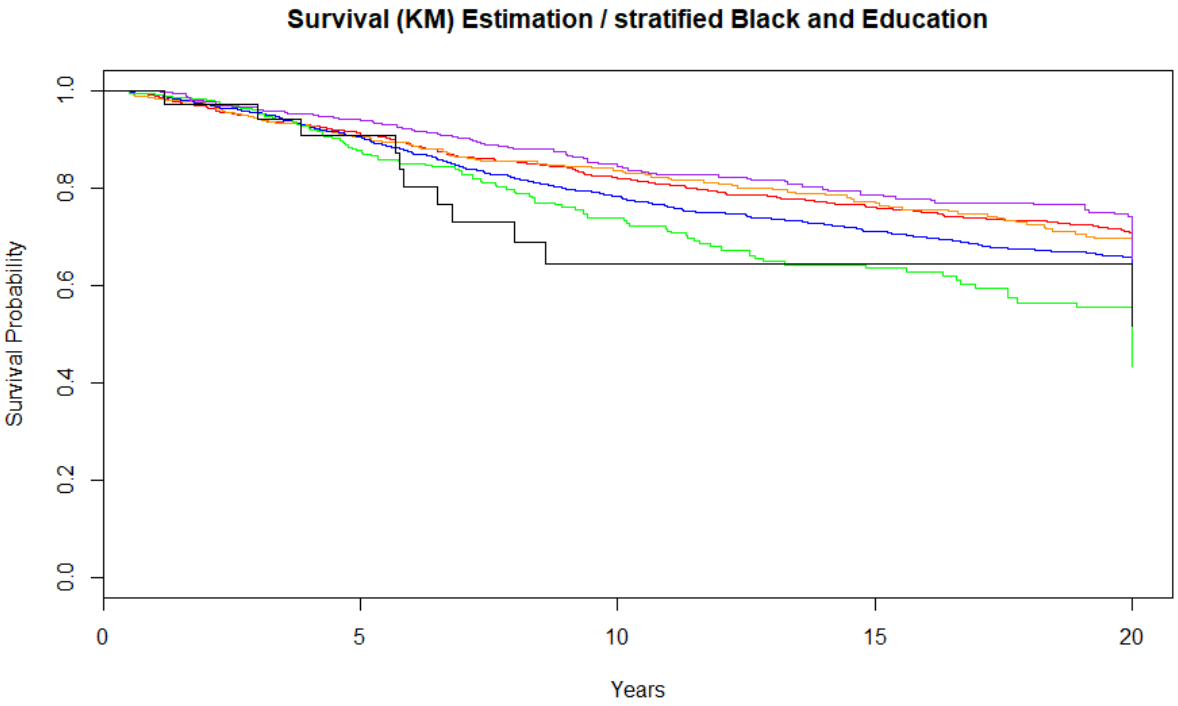
- Education level: 0 (<12y) ; "not black" in red ; "black" in orange
- Education level: 1 (12-15y) ; "not black" in blue ; "black" in green
- Education level: 2 (>15y) ; "not black" in purple ; "black" in black

We can notice 2 things:

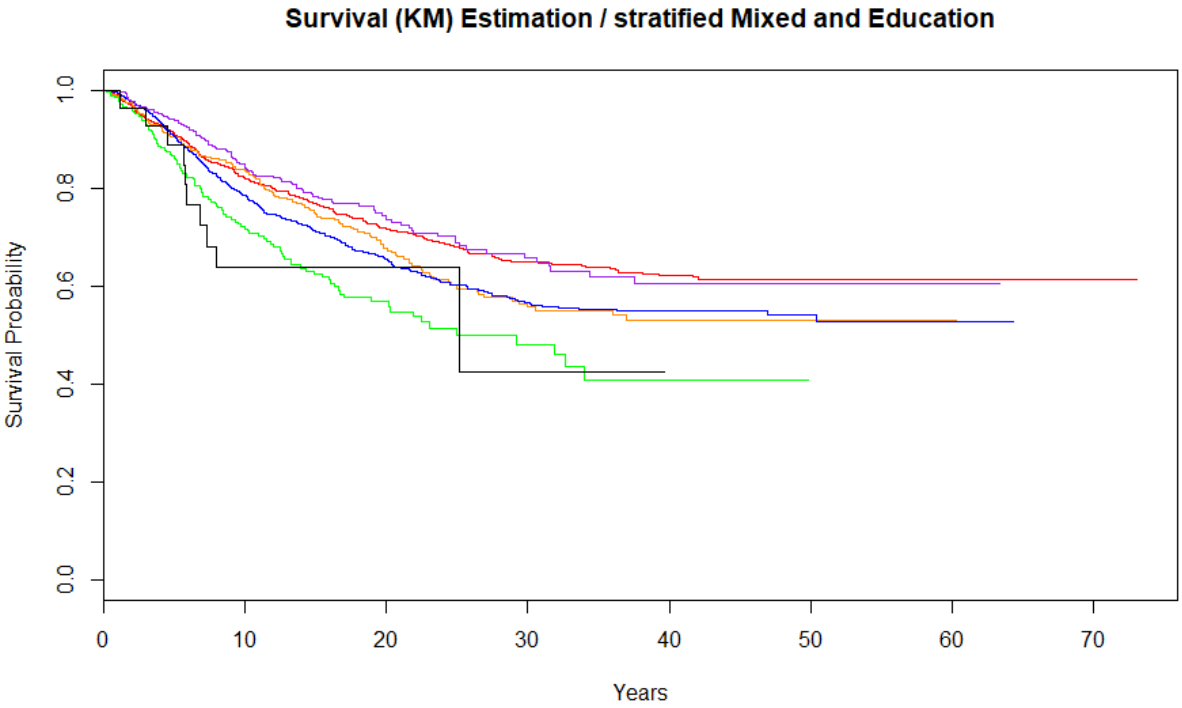
- The couples, whose husband is black, have always a lower survival probability than the others, whatever the level of education is.
  Notice: "husband is black" covers both cases where husband and wife are black (couple is NOT mixed ethnicity) or only the husband is black (couple is mixed ethnicity)
- The higher the education level, the bigger the difference in survival probability. Median value of survival probability for the 2 highest level of education is only 22.5 and 25.2 years.

	n	events	median	0.95LCL	0.95UCL
d$m_educ=0, strata(d$m_black)=d$m_black=0	922	282	NA	NA	NA
d$m_educ=0, strata(d$m_black)=d$m_black=1	366	111	NA	38.09	NA
d$m_educ=1, strata(d$m_black)=d$m_black=0	1312	421	NA	50.38	NA
d$m_educ=1, strata(d$m_black)=d$m_black=1	343	108	22.5	17.78	NA
d$m_educ=2, strata(d$m_black)=d$m_black=0	392	99	NA	NA	NA
d$m_educ=2, strata(d$m_black)=d$m_black=1	36	11	25.2	8.59	NA

A truncation at 20 years shows the different profiles a bit more clearly (and leads to the same observations of course).



e) <u>Stratified test of husband education level and ethnicity (couple is mixed ethnicity)</u>



```
Call: survfit(formula = Surv(d$years, d$div) ~ d$m_educ + strata(d$mixed),
    data = d)

                                          n events median 0.95LCL 0.95UCL
d$m_educ=0, strata(d$mixed)=d$mixed=0  944    270     NA      NA      NA
d$m_educ=0, strata(d$mixed)=d$mixed=1  344    123     NA    30.0      NA
d$m_educ=1, strata(d$mixed)=d$mixed=0 1387    427     NA    50.4      NA
d$m_educ=1, strata(d$mixed)=d$mixed=1  268    102   29.2    20.1      NA
d$m_educ=2, strata(d$mixed)=d$mixed=0  399    100     NA      NA      NA
d$m_educ=2, strata(d$mixed)=d$mixed=1   29     10   25.2     8.0      NA
```

The analysis of the "mixed ethnicity couple" variable brings a similar observation as for the "black husband ethnicity" variable: the mixed ethnicity couples have always a lower survival probability than the others, whatever the level of education is.

Notice: in both cases d) and e), the number of samples where education level =2 and ethnicity variable = true (black or mixed) is statistically low. This may lead to a bias in the results.

5. COX regression

Thanks to the COX regression, we may add some risk information on the time to divorce probability, from the education level and ethnicity variables.

Questions:
- What is the risk brought on the time to divorce probability by the 3 variables: education level of the husband, husband is black, couple is of mixed ethnicity?
- Can we confirm the higher risk brought by the 2 ethnicity variables as demonstrated by the previous tests?

From the COX regression results, we can notice the following:

- The 3 p-values are below a 0.05 significance level for *H0: beta=0*. This means that the 3 variables have a significant impact on the survival probability
- However, the "mixed couple" factor is the most significant by an order of magnitude vs the 2 others (0.00382 vs 0.02 and 0.04)
- It also has the highest risk (1.258) leading to the lowest survival probability (lowest time to divorce)
- Higher to lower risk: mixed > black > education level

The quantities exp(coeff=beta) are called hazard ratios (HR). A value of beta greater than zero, or equivalently a hazard ratio greater than 1, indicates that, as the value of the covariate increases, the event hazard increases and thus the length of survival decreases.

The p-value comes from testing the null hypothesis that this hazard ratio is 1.

```
Call:
coxph(formula = Surv(d$years, d$div) ~ d$m_educ + d$m_black +
    d$mixed, data = d)

  n= 3371, number of events= 1032

              coef exp(coef) se(coef)     z Pr(>|z|)
d$m_educ   0.09426   1.09885  0.04718 1.998  0.04571 *
d$m_black  0.18367   1.20162  0.07974 2.303  0.02126 *
d$mixed    0.22936   1.25779  0.07929 2.893  0.00382 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

           exp(coef) exp(-coef) lower .95 upper .95
d$m_educ       1.099     0.9100     1.002     1.205
d$m_black      1.202     0.8322     1.028     1.405
d$mixed        1.258     0.7950     1.077     1.469

Concordance= 0.524  (se = 0.009 )
Rsquare= 0.006   (max possible= 0.99 )
Likelihood ratio test= 18.81  on 3 df,   p=3e-04
Wald test            = 19.6  on 3 df,   p=2e-04
Score (logrank) test = 19.68  on 3 df,   p=2e-04
```

We can also split the dataset per education level to confirm the risk associated to ethnicity per education level:

Education level	variable	p-value	Exp(coeff)	comment
<12y	black	0.55	1.07	p-value not significant
	mixed	0.09	1.20	p-value close to significant, risk is higher
12-15y	black	0.02	1.29	p-value is significant, risk is higher
	mixed	0.02	1.28	p-value is significant, risk is higher
>15y	black	0.12	1.73	p-value not significant (?), but highest risk
	mixed	0.22	1.58	p-value not significant (?), but highest risk

These results confirm quantitatively the observations from the schemes from the previous section:

- Ethnicity variables ("husband is black" or "couple is mixed") bring a higher risk to the time to divorce event
- The higher the education level, the higher the risk

Notice: for education level > 15y, I would have expected a smallish p-value. When you correct this exam, I would be interested in having your explanation on this point please at remi.hardy@edu.dsti.institute, thanks in advance

APPENDIX 1: R Code

```
library(survival)

d=read.table("d:\\SURVIVAL\\divorce.raw",header=FALSE)
names(d)=c('id','m_educ','m_black','mixed','years','div')

#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 and two indicators of the couple's ethnicity: whether the husband is black and whether the couple is mixed. The variables are:
#id: a couple number.
#heduc: education of the husband, coded
#0 = less than 12 years,
#1 = 12 to 15 years, and
#2 = 16 or more years.
#heblack: coded 1 if the husband is black and 0 otherwise
#mixed: coded 1 if the husband and wife have different ethnicity (defined as black or other), 0 otherwise.
#years: duration of marriage, from the date of wedding to divorce or censoring (due to widowhood or interview).
#div: the failure indicator, coded 1 for divorce and 0 for censoring.

#---------------------------------------------
#STATISTICS / RAW ANALYSIS
#---------------------------------------------

#distribution of the categorical variables
table(d$mixed)
table(d$m_black)
table(d$m_educ)


#DIVORCED / NOT-DIVORCED variable
#get the distribution of divorced vs censored
table(d$div)
prop.table(table(d$div))
#get the distribution of marriage duration
h=hist(d$years,freq=FALSE, main='Marriage Duration',xlab='Years',ylab='Density')
min(d$years)
max(d$years)
mean(d$years)
median(d$years)

#how many divorces do we get from the first 50%
count=0
count_div=0
```

```
med=median(d$years)
for (i in (1:length(d[,5])))
{
 if (d[i,5]<med)
 {
  count=count+1
  if (d[i,6]==1) {count_div=count_div+1}
 }
}

#subset original dataset with divorced only (div=1)
d_div=subset(d,d$div==1)
min(d_div$years)
max(d_div$years)
mean(d_div$years)
median(d_div$years)

#subset original dataset with non divorced only (div=0)
d_ndiv=subset(d,d$div==0)
min(d_ndiv$years)
max(d_ndiv$years)
mean(d_ndiv$years)
median(d_ndiv$years)

#related boxplot
b=boxplot(d$years~d$div, main='Statistics 0:not-divorced / 1:divorced', ylab='Marriage Duration')


#EDUCATION / ETHNICITY variables
#education and ethnicity have the same distribution in divorced and not-divroced groups
prop.table(table(d$m_educ))
prop.table(table(d_div$m_educ))
prop.table(table(d_ndiv$m_educ))

prop.table(table(d$m_black))
prop.table(table(d_div$m_black))
prop.table(table(d_ndiv$m_black))

prop.table(table(d$mixed))
prop.table(table(d_div$mixed))
prop.table(table(d_ndiv$mixed))


#relation black/ education
table(d$m_black)
prop.table(table(d$m_black))
prop.table(table(d$m_educ,d$m_black),2)
```

```
#relation black / marriage duration
boxplot(d$years~d$m_black, main='Statistics 0:not-black / 1:black', ylab='Marriage Duration')


#-------------------------------------------
#SURVIVAL ANALYSIS / LOGRANK + KM estim
#-------------------------------------------


#BLACK
fit.km=survfit(Surv(d$years,d$div)~ d$m_black,data=d)
plot(fit.km, col=c('blue','red'),main='Survival (KM) Estimation with "Husband is black"',xlab='Years', ylab='Survival Probability')
legend("topright", legend=c("Not black", "Black"),col=c("blue", "red"), lty=1, cex=0.8)
fit.km
survdiff(Surv(d$years,d$div)~ d$m_black,data=d)

#MIXED
fit.km=survfit(Surv(d$years,d$div)~ d$mixed,data=d)
plot(fit.km, col=c('blue','red'),main='Survival (KM) Estimation with "Couple is mixed"',xlab='Years',ylab='Survival Probability')
legend("topright", legend=c("Not mixed", "Mixed"),col=c("blue", "red"), lty=1, cex=0.8)
fit.km
survdiff(Surv(d$years,d$div)~ d$mixed,data=d)

#EDUC
fit.km=survfit(Surv(d$years,d$div)~ d$m_educ,data=d)
plot(fit.km, col=c('red','green','blue'),main='Survival (KM) Estimation with "Husband level of education"',xlab='Years',ylab='Survival Probability')
legend("topright", legend=c("<12y", "12-15y",">15y"),col=c("red", "green","blue"), lty=1, cex=0.8)
fit.km
survdiff(Surv(d$years,d$div)~ d$m_educ,data=d)

#Stratified test on level of education and black
fit.km=survfit(Surv(d$years,d$div)~ d$m_educ + strata(d$m_black),data=d)
plot(fit.km, col=c('red','darkorange','blue','green','purple','black'),main='Survival (KM) Estimation / stratified Black and Education',xlab='Years',ylab='Survival Probability')
fit.km


#truncation on 20 years
d_trunc=d
trunc_th=20

for (i in (1:length(d_trunc[,5])))
{
if (d_trunc[i,5]>trunc_th)
   {
```

```
   d_trunc[i,5]=trunc_th
   d_trunc[i,6]=0
   }
}

fit.km=survfit(Surv(d_trunc$years,d$div)~ d_trunc$m_educ + strata(d_trunc$m_black),data=d_trunc)
plot(fit.km,  col=c('red','darkorange','blue','green','purple','black'),main='Survival (KM) Estimation / stratified Black and Education',xlab='Years',ylab='Survival Probability')
fit.km


#Stratified test on level of education and mixed
fit.km=survfit(Surv(d$years,d$div)~ d$m_educ + strata(d$mixed),data=d)
plot(fit.km,  col=c('red','darkorange','blue','green','purple','black'),main='Survival (KM) Estimation / stratified Mixed and Education',xlab='Years',ylab='Survival Probability')
fit.km


#---------------------------------------------
#SURVIVAL ANALYSIS / COX
#---------------------------------------------

#1- multi variate COX regression analysis
fit<-coxph(Surv(d$years,d$div)~d$m_educ+d$m_black+d$mixed,data=d)
summary(fit)

#2- split the dataset per education level class ...
d_low=subset(d,d$m_educ==0)
d_mid=subset(d,d$m_educ==1)
d_high=subset(d,d$m_educ==2)

#... to confirm ethnicity risk per education level class
fit<-coxph(Surv(d_low$years,d_low$div)~d_low$m_black+d_low$mixed,data=d_low)
summary(fit)
fit<-coxph(Surv(d_mid$years,d_mid$div)~d_mid$m_black+d_mid$mixed,data=d_mid)
summary(fit)
fit<-coxph(Surv(d_high$years,d_high$div)~d_high$m_black+d_high$mixed,data=d_high)
summary(fit)
```

APPENDIX 2: Dataset Extract (divorce.dat, not used, but more explicit)

id	heduc	heblack	mixed	years	div
9	12-15 years	No	No	10.546	No
11	< 12 years	No	No	34.943	No
13	< 12 years	No	No	2.834	Yes
15	< 12 years	No	No	17.532	Yes
33	12-15 years	No	No	1.418	No
36	< 12 years	No	No	48.033	No
43	16+ years	No	No	16.706	No
47	< 12 years	No	No	24.999	No
50	< 12 years	No	No	24.999	No
56	< 12 years	Yes	No	3.869	No
63	12-15 years	Yes	No	7.732	No
66	12-15 years	No	No	5.2105	Yes
70	12-15 years	No	No	15.444	No
77	16+ years	No	No	4.085	No
80	< 12 years	No	No	17.333	No
84	12-15 years	No	No	16.331	No
87	12-15 years	No	No	35.335	No
90	< 12 years	No	No	37.67	No
91	12-15 years	No	No	19.1865	Yes
94	12-15 years	No	No	22.697	No
109	< 12 years	No	No	50.776	No
111	< 12 years	No	No	37.495	No
114	< 12 years	No	No	30.738	No
116	< 12 years	No	No	48.654	No
118	< 12 years	No	No	33.024	No
120	< 12 years	No	No	36.668	No
122	16+ years	No	No	27.992	No
124	16+ years	No	No	27.789	No
128	12-15 years	No	No	3.2525	Yes
129	12-15 years	No	No	5.67	Yes
137	< 12 years	No	No	17.999	No
139	< 12 years	No	No	7.5975	Yes
141	< 12 years	No	No	1.791	No
145	< 12 years	No	No	2.475	No
148	12-15 years	No	No	35.959	No
153	12-15 years	No	No	13.136	No
160	12-15 years	No	No	5.717	No
162	12-15 years	No	No	5.717	No
164	12-15 years	No	No	26.278	No
165	12-15 years	No	No	4.331	Yes
166	< 12 years	No	No	8.893	No
170	12-15 years	No	No	15.743	No
183	< 12 years	No	No	4.671	Yes
194	< 12 years	No	No	48.356	No

APPENDIX 3: Dataset Extract (divorce.raw, loaded by R code)

9	1	0	0	10.5460	0
11	0	0	0	34.9430	0
13	0	0	0	2.8340	1
15	0	0	0	17.5320	1
33	1	0	0	1.4180	0
36	0	0	0	48.0330	0
43	2	0	0	16.7060	0
47	0	0	0	24.9990	0
50	0	0	0	24.9990	0
56	0	1	0	3.8690	0
63	1	1	0	7.7320	0
66	1	0	0	5.2105	1
70	1	0	0	15.4440	0
77	2	0	0	4.0850	0
80	0	0	0	17.3330	0
84	1	0	0	16.3310	0
87	1	0	0	35.3350	0
90	0	0	0	37.6700	0
91	1	0	0	19.1865	1
94	1	0	0	22.6970	0
109	0	0	0	50.7760	0
111	0	0	0	37.4950	0
114	0	0	0	30.7380	0
116	0	0	0	48.6540	0
118	0	0	0	33.0240	0
120	0	0	0	36.6680	0
122	2	0	0	27.9920	0
124	2	0	0	27.7890	0
128	1	0	0	3.2525	1
129	1	0	0	5.6700	1
137	0	0	0	17.9990	0
139	0	0	0	7.5975	1
141	0	0	0	1.7910	0
145	0	0	0	2.4750	0
148	1	0	0	35.9590	0
153	1	0	0	13.1360	0
160	1	0	0	5.7170	0
162	1	0	0	5.7170	0
164	1	0	0	26.2780	0
165	1	0	0	4.3310	1
166	0	0	0	8.8930	0
170	1	0	0	15.7430	0
183	0	0	0	4.6710	1
194	0	0	0	48.3560	0
196	0	0	0	39.3320	0
199	0	0	0	39.3320	0