A Statistical Report of a Study of US Divorces and Factors Related to Divorce

Introduction

This study is based on a longitudinal survey conducted in the US of divorces and factors related to divorce. The event of interest was divorce and time to event was measured in years. The main aim of this study was to find the best model to examine the risk factors for divorce, view, characterize and discuss overall survival rates and trends, differences in the survival distribution of couples with different education levels, ethnicity, age, and income.

Censoring

The type of censoring used in this study is random. For different reasons, some of the couples were randomly lost from the study and information about what happened after was not available. Some of the couples were censored as a result of widowhood while others who did not show up for the interview were also censored. About 69.39% were censored.

Descriptive Statistics

This study examined the risk factors related to divorce in the US using data collected from 3771 couples who were followed over time. The variables included were HEDUC (Education of the husband coded as <12, 12 to 15, and 16+), HEBLACK (coded as yes if the husband is black and 0 otherwise), MIXED (coded as 'yes' if the couple have different ethnicity and 0 otherwise), AGE (This is the average age of the couple at the date of the wedding), INCOME (annual household income for the married couple in units of thousands of dollars), YEARS (duration of marriage from the date of wedding to divorce or censoring) and DIVORCE (coded as 1 for divorce and 0 for censoring).

Table 1. summarizes the demographic characteristics of the couples in the study. One of the observations in the INCOME variable had a negative value. This was the only negative value in the dataset and after a

thorough assessment of the data, it was determined to be an error. It was changed to a positive value and kept in the dataset.

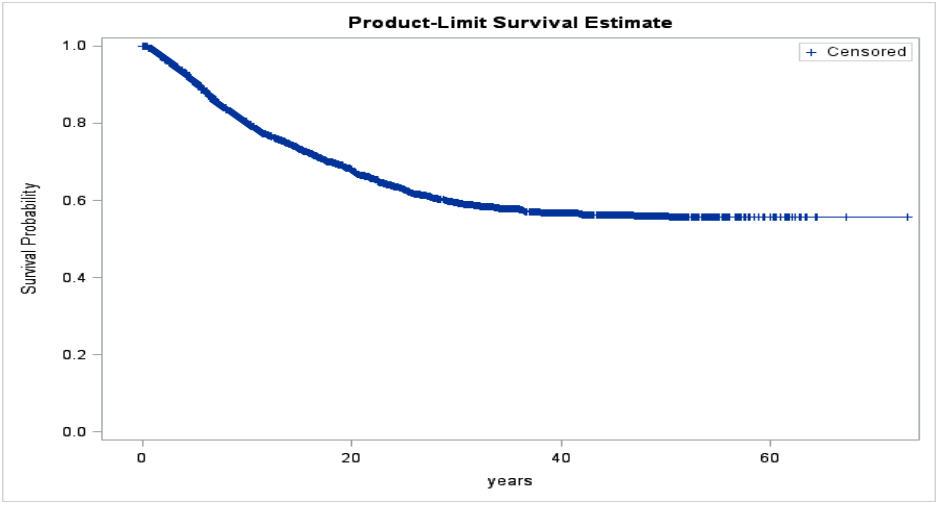
Characteristics	Frequency (N=3771)	Percent/Mean ± SD
Education(Husband)		
<12	1288	38.21
12 to 15	1655	49.10
16+	428	12.70
Black(Husband0		
Yes	745	22.10
No	2626	77.90
Mixed ethnicity		
Yes	641	19.02
No	2730	80.98
Age	3771	25.53 ± 2.21
Income	3771	66.59 ± 16.15
Divorced		
Yes	1032	30.61
Censored	2339	69.39

Table 1. Descriptive statistics

Overall Survival Rates and Trends

Kaplan Meier survival plots were used to characterize raw survival rates and trends. From the plot in figure 1, a gradual decline can be seen in the first 20 years. Subsequently the curve flattens out with every additional year. Approximately 67.71% of couples take longer than 20 years to experience a divorce. At 13.67 years, 25% of couples already have experienced a divorce. Censoring was spread all through the time in the study. It was not possible to obtain the median time to divorce in the study because more than half of the couples were censored.

Figure 1:Overall Survival plot for the study


Product-Limit Survival Estimate
Censored
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
years

Life tables were used to present survival information in 5-year intervals. From the table the number of couples that got divorced was more in the first 20 years compared to the rest of the period. Also, the number of censored observations was also more in the first 20 years of the study.

Interval		Number Failed	Number Censored	Conditional Probability of Failure	Survival	PDF	Hazard
[Lower,	Upper)						
0	**5**	294	343	0.0919	1.0000	0.0184	0.019263
5	**10**	303	343	0.1182	0.9081	0.0215	0.025135
10	**15**	163	297	0.0840	0.8007	0.0135	0.017546
15	**20**	113	272	0.0757	0.7334	0.0111	0.015744
20	**25**	85	251	0.0761	0.6779	0.0103	0.015814
25	**30**	43	164	0.0521	0.6263	0.00653	0.010703

Table 2. Life Table Showing the Survival, Pdf and Hazard rates for intervals up to 30.

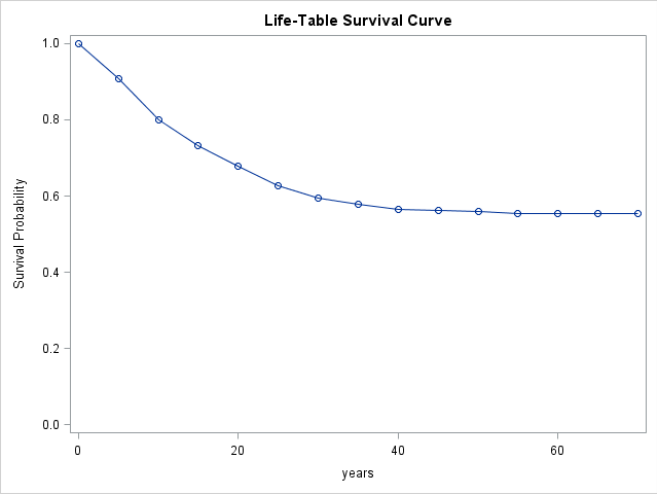


Figure 2: Life Table Survival Curves at 5-year intervals

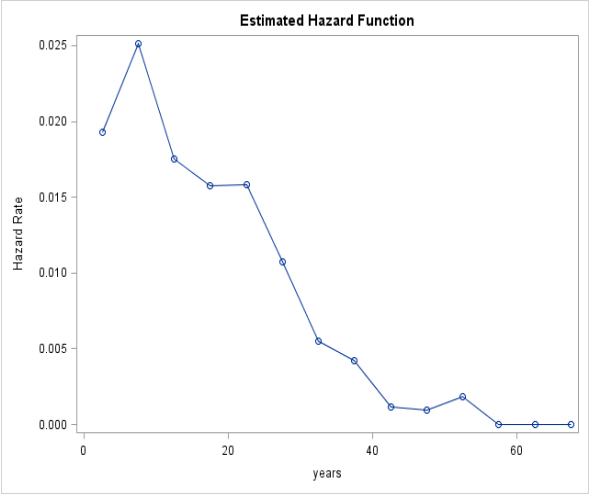


Figure 3: Hazard plot

The hazard of getting a divorce is high during the first five years of marriage and peaks between 5 and 10 years. From that point on, the hazard decreases with each additional year with a slight increase between 50 and 55 years.

Survival distributions of Different Groups

Kaplan-Meier survival plots and log-ranks test were used to compare couples with different levels of education, race, age, and income. The survival plot below shows the distribution of couples with different levels of education. Those with 16+ years of education did not appear to be significantly different from those with <12 years of education. The differences between the 3 groups were examined using a log-ranks test and it showed that at least 2 of the groups were significantly different from each other (χ^2 =17.3887; p = 0.0002).

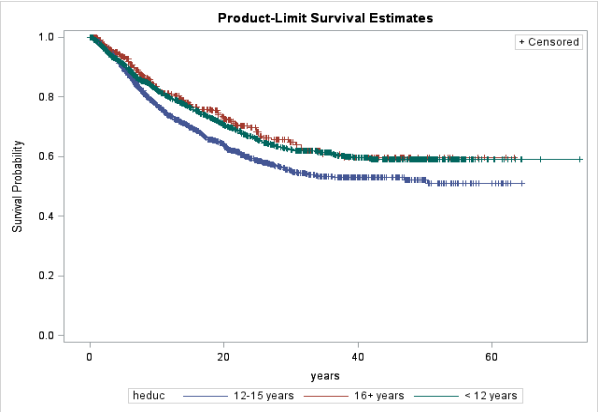


Figure 4. Survival plot for HEDUC

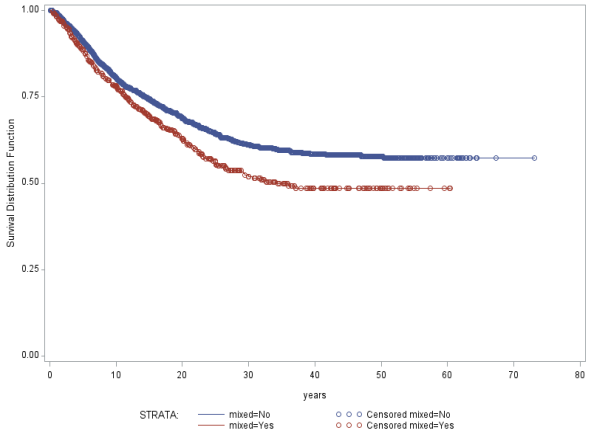


Figure 5. Survival plot for MIXED

Further analysis was done to determine the particular groups that were significantly different from each other, and it was found that couples whose husbands had <12 years of education were significantly different from couples whose husbands had 12 to 15 years of education (P<.0001). Therefore, HEDUC was recoded as EDUCAT (**0** as less than 12 years or 16+ and **1** as 12 to 15 years).

Figure 5. shows the distribution of couples with different races. Both curves appear to be different from each other. A log-ranks test showed that they were significantly different from one another (χ^2 =11.3427; p = 0.0008). In order to view differences in the survival distribution based on age and income (continuous variables), they were recoded as categorical variables only for the purpose of this comparison.

AGE which ranged from 18 to 33 was recoded as AGECAT (1=less than 25, 2 = 25+) and INCOME which ranged from 10 to 121 was recoded as INCOMECAT (1 = <60, 2 = 60+). A significant difference was seen between both groups in age (χ^2 =8.8825; p = 0.0029) and between both groups in income ((χ^2 =6.4886; p = 0.0109).

Cox Proportional Hazards (PH) Model

As a first step, a univariate proportional hazards model was carried out for each of the variables in the study and statistical significance was measured at .05. The following results obtained are shown in the table below.

Parameters	HR	P value
Educat	1.292	<.0001
Mixed	1.283	.0008
Heblack	1.247	.0033
Age	.954	.0005
Income	.995	.0137

After this initial analysis, a multivariate model including all 5 variables was done. All 10 (two-way) interaction effects among all these factors were added one at a time and none was significant at the.05 significance level.

The different models considered were:

MODEL	-2LogL	DESCRIPTION OF MODEL
Model 1	15636.106	EDUCAT (with 2 groups) + MIXED + HEBLACK + AGE + INCOME
Model 2	15634.854	HEDUC (with 3 groups) + MIXED + HEBLACK + AGE + INCOME
Model 3	15639.583	EDUCAT (with 2 groups) + MIXED + HEBLACK + AGECAT + INCOME

Between Model 2 (6 parameters) and Model 3 (5 parameters), using log likelihood as a criterion (critical value=3.84), Model 2 is better. Between Model 1 (5 parameters) and Model 2 (6 parameters), model 1 is better. Therefore, the chosen model is model 1.

Cox Proportional Hazards Model Assumptions

Model 1 was tested for the Cox model assumptions. For the nominal variables, graphical comparisons between groups were done via complementary log-log survivor plot versus time. If the hazards are proportional, there will be a constant difference in the log-log survivor function for any value of time.

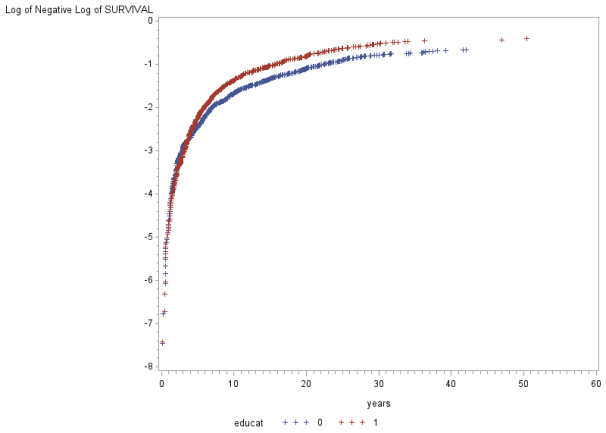


Figure 6: Log-log survivor plot versus time for Educat

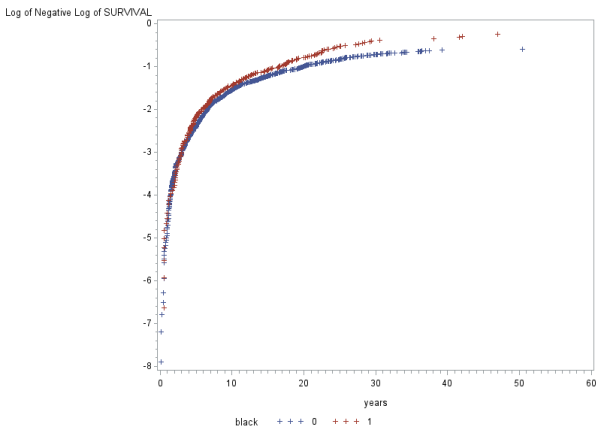


Figure 7: Log-Log Survivor Plot for Heblack

The log-log survivor plot for EDUCAT and MIXED appear parallel and there seems to be a constant difference at all times. For the HEBLACK log-log survivor plot, at the earlier times, both curves appear to be joined and no difference exists while at later times, there seems to be a difference between them.

Schoenfeld residuals were used to assess the assumptions for the continuous variables. Figures 8 and 9 show the plots of the Schoenfeld residuals against time for age and income respectively. Both show more points at earlier times and fewer points at later times. Even though the plots may appear to be randomly scattered about 0, they also look as though they have a funneling pattern. Further statistical testing using correlation is necessary to really confirm these results. Further testing of the correlation of the Schoenfeld residuals with time (years), log of time (log years) and square of time ($years^2$) was carried out. The Schoenfeld residual for age was not found to be significantly correlated with years (r= -0.035, p=0.2612), log of years (r=-0.01389, p=0.6558), and $years^2$ (r=-0.03571, p=0.2517).

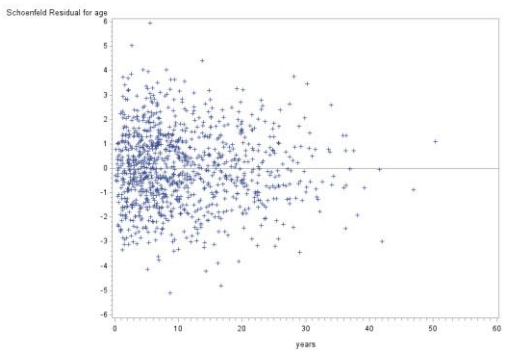


Figure 8: Schoenfeld Residuals for Age

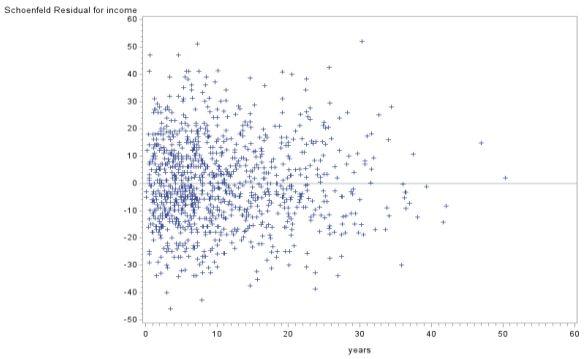


Figure 9: Schoenfeld residuals for Income

Also, the Schoenfeld residual for income was not found to be significantly correlated with years (r= -0.01459, p=0.6397), log of years (r=-0.01625, p=0.6022), and years^2 (r=-0.01211, p=0.6976). This therefore supports the Cox proportional hazards assumption.

Interactions of the variables with time were also done. Only the interaction between Heblack and Years was found to be significant (HR=1.020, p=0.0291) at the .05 level. This explained the log-log survivor plot in figure 7 where there was no difference between the curves at earlier times. Therefore, the interaction term, Black*Years was added to the final model.

$$\text{Log } h_i(t) = \alpha(t) + 0.30725X_{EDUCAT} + 0.21004X_{MIXED} - 0.04358X_{HEBLACK} - 0.\ 04272X_{AGE} - 0.00516X_{INCOME} + 0.01968Xt_{BLACKYEARS}$$

Parameters	Estimate	SE	P value	HR
Educat	0.30725	0.06357	<0.0001	1.360
Mixed	0.21004	0.07858	0.0075	1.234
Heblack	-0.04358	0.12327	0.7237	0.957
Age	-0.04272	0.01368	0.0018	0.958
Income	-0.00516	0.00193	0.0076	0.995
BlackYears	0.01968	0.00902	0.0291	1.020

Cox Model Interpretation

On average, if all other covariates are held constant, we expect couples who have husbands that have 12 to 15 years of education to have a hazard of divorce that is 36% higher than couples whose husbands have less than 12 years of education or those that have 16 or more years of education (p<.0001). With all the

covariates held constant, on average, couples that have different ethnicity will expect to have a 23.4% higher hazard of divorce than couples who have the same ethnicity (p=.0075). Also, holding all other factors constant, on average, after 10 years of marriage, couples whose husbands are black are expected to have a 16.56% higher hazard of divorce than those whose husbands are not black (HR=1.1656).

With all the covariates the same between two couples, on average, a couple whose average age at the date of their wedding is one year greater will expect to have a hazard of divorce that is 4.2% lower than one whose average age at the date of their wedding is one year lower (p=0.0018). Also, with all the covariates the same between two couples, on average, the couple with one-unit higher annual household will expect to have a hazard of divorce that is 0.5% lower than couples who have a one-unit lower annual household income (p=0.0076).

Parametric Model

Four parametric models were fitted to describe the relationship between a set of covariates, Educat, Mixed, Heblack, Age and Income and time to divorce measured in years. The four models include the Weibull, Gamma, Lognormal and Exponential model.

The Exponential model (Log likelihood = -3136.095): $\text{Log}(T_i) = 2.9626 - 0.3921X_{EDUCAT} - 0.2476X_{MIXED} - 0.2303X_{BLACK} + 0.0426X_{AGE} + 0.0051X_{INCOME}$

If the distribution is exponential, the plot of negative log survival function versus time should yield a straight line with an origin at 0. The plot in figure 10 which shows the plot of negative log survival function starts as a straight line from 0 but turns into a curve between 30 and 40 years. This does not support the exponential distribution. Also using the Lagrange Multiplier statistics(χ^2 =17.5103, p<.0001), it was concluded that the exponential model was not a good fit for the distribution.

Weibull Model (Log Likelihood = -3125.893): $\text{Log}(T_i) = 2.8858 - 0.4068X_{EDUCAT} - 0.2727X_{MIXED} - 0.2239X_{BLACK} + 0.0485X_{AGE} + 0.0058X_{INCOME}$

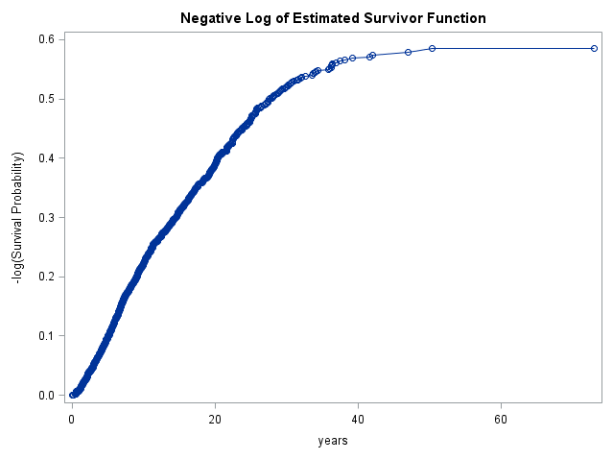


Figure 10: Plot of Negative Log Survival Function

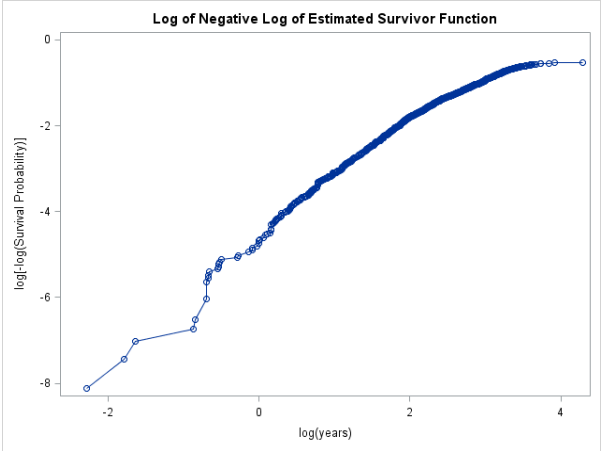


Figure 11: Plot Of Log Negative Log Survival Function

The plot in Figure 11 shows an initial mild zigzag pattern and then a straight line before gradually turning into a curve. As a result, the Weibull distribution is not supported.

Gamma Model (Log Likelihood = -3060.636): $\text{Log}\,(T_i) = 1.6234 - 0.2935X_{EDUCAT} - 0.2906X_{MIXED} - 0.0859X_{BLACK} + 0.0719X_{AGE} + 0.0056X_{INCOME}$

Log-Normal Model (Log Likelihood = -3067.958): $\text{Log}\,(T_i) = 2.0854 - 0.3469X_{EDUCAT} - 0.2829X_{MIXED} - 0.1380X_{BLACK} + 0.0647X_{AGE} + 0.006X_{INCOME}$

Best Parametric Model

Using the log likelihood as a criterion, the best parametric model was decided below. The null hypothesis for this test states that the model with fewer parameters is adequate while the alternative states that the more complex model is needed.

Model	2(L1-L2)	Critical Value	Better Model
Gamma vs Log-Normal	14.644	3.84	Gamma
Gamma vs Weibull	130.514	3.84	Gamma
Gamma vs Exponential	150.918	5.99	Gamma
Weibull vs Exponential	20.404	3.84	Weibull

The best parametric model is the gamma model.

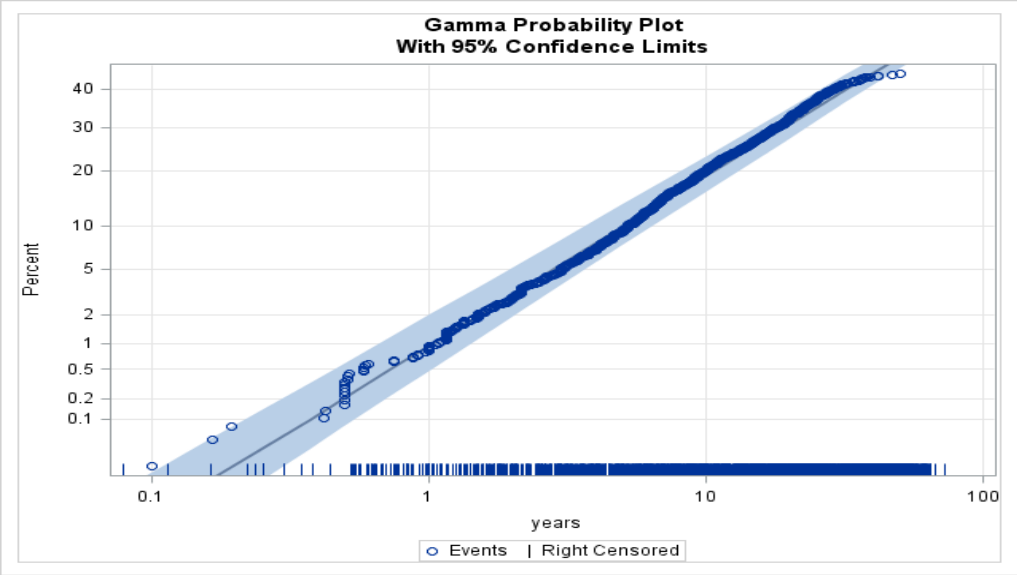


Figure 12. The Probability plot of the Gamma Model

To evaluate if the gamma model is the right fit for the model, the model is compared directly to the data using the probability plot shown above. It is clear that the model fits the data. Almost all points fell into the straight line and 95% confidence interval. This shows that the model meets the distribution assumption.

Parameters	Estimate	95% Confidence Limits		Chi-Square	P-value
Intercept	1.6234	0.5381	2.7086	8.60	0.0034
Education	-0.2935	-0.4502	-0.1367	13.47	0.0002
Mixed	-0.2906	-0.4880	-0.0932	8.33	0.0039
Black	-0.0859	-0.2809	0.1091	0.75	0.3879
Age	0.0719	0.0326	0.1113	12.83	0.0003
Income	0.0056	0.0008	0.0104	5.25	0.0219

Model Interpretation

On average, we predict an expected time to divorce of 5.07 years if all covariates are set to 0 (p=0.0034). With all the covariates held constant, on average, couples who have husbands that have 12 to 15 years of education are expected to have a 25.44% shorter time to divorce than couples whose husbands have less than 12 years or 16 or more years of education (p=0.0002). With all the covariates held constant, on average, couples that have different ethnicity are expected to have a 25.22% shorter time to divorce than couples who have the same ethnicity (p=0.0039). Also, holding all other factors constant, on average, couples whose

husbands are black are expected to have an 8.23% shorter time to divorce than those whose husbands are not black. This was found not to be significant (p=0.3879).

With all the covariates the same between two couples, on average, a couple whose average age at the date of their wedding is one year greater will expect to have a 7.45% longer time to divorce than one whose average age at the date of their wedding is one year lower (p=0.0003). Also, with all the covariates the same between two couples, on average, the couple with one-unit higher annual household will expect to have a 0.56% longer time to divorce than couples who have a one-unit lower annual household income (p=0.0219).

Final Model

After fitting both the parametric model and the cox proportional hazards model, it was found that the best model for parametric models was the Gamma model and the best model for the Cox PH model was a six-parameter model containing EDUCAT, MIXED, HEBLACK, AGE, INCOME, and the Interaction of BLACK with YEARS. Using the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC), the gamma model and the Cox PH model were compared.

MODEL	AIC	BIC
Gamma model	6137.273	6186.256
Cox PH model	15643.530	15673.165

Examining the estimates of the gamma and cox models, both appear to have similar interpretations for the variables explaining education, mixed ethnicity, black husbands, age, and income. Both models differ as regards to the black variable. In the Cox PH model, it was found to be significant but dependent on time while in the gamma model it was not significant. This may seem as though these two models are trying to explain the same thing.

Considering the similarities in both models and their AIC and BIC values (smaller is better), the gamma model appears to be the best fit for the data.