Andrew Heneghan

6/30/2023

Survival Analysis on Marriage Factors and Divorce

A survival analysis study was done to see what kinds of factors in a marriage can play big parts in divorce. In survival analysis, survival time is the time starting from a point defined by the researcher to the time of the event of interest. For this study, survival time, in years, would start on the day a couple got married and end at the time they either got divorced or the study ended. Censoring can also occur in survival analysis when survival time is not exactly known for a subject or the event of interest does not happen before the study's end. Censored observations in this study included if either couples were still together by the end of the study or if one partner became a widow or widower. These couples will not be included in the resulting analysis. This censoring follows a random, Type I pattern since the study ends at a fixed point of time and couples are censored for reasons beyond the researcher's control. In this case, it would be if they become widows or are not divorced by then. These couples would also be right-censored since either the study ends before they're divorced or they leave the study due to widowhood.

Life Table Survival Estimates

Interval												Evaluated at the Midpoint of the Interval			
[Lower,	Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	10	597	686	3028.0	0.1972	0.00723	1.0000	0	0	.	.	0.0197	0.000723	0.021872	0.00089
10	20	276	569	1803.5	0.1530	0.00848	0.8028	0.1972	0.00723	.	.	0.0123	0.000690	0.016572	0.000994
20	30	128	415	1035.5	0.1236	0.0102	0.6800	0.3200	0.00916	.	.	0.00841	0.000705	0.013176	0.001162
30	40	27	267	566.5	0.0477	0.00895	0.5959	0.4041	0.0106	.	.	0.00284	0.000536	0.004882	0.000939
40	50	3	252	280.0	0.0107	0.00615	0.5675	0.4325	0.0114	.	.	0.000608	0.000349	0.001077	0.000622
50	60	1	123	89.5	0.0112	0.0111	0.5614	0.4386	0.0118	.	.	0.000627	0.000624	0.001124	0.001124
60	70	0	26	14.0	0	0	0.5552	0.4448	0.0133	.	.	0	.	0	.
70	80	0	1	0.5	0	0	0.5552	0.4448	0.0133	.	.	0	.	0	.
80	.	0	0	0.0	0	0	0.5552	0.4448	0.0133	.	.	.	.	.	.

Table 1

For the different times shown in Table 1, the raw survival rate of this data, highlighted in Table 1 above, shows the probability that a couple will remain married longer than that time. The survival rate starts at 100% (1.0) since all couples start out married here. The rate then decreases as time from marriage increases and more couples get divorced. In the first two or three decades, the survival rate steadily decreases as many marriages seem to fall through in early years. After about 35 to 40 years into a marriage, survival rate levels out around 55.52% (0.5552). From here on, it seems the chances of staying married will remain around this value. These same decreasing survival rates are visualized in the survival probability curve in Figure 1. Unfortunately, this overall survival rate cannot go any lower. The censored couples count as part of the data used for the analysis, but not for the analysis itself. They are not shown in the survival curve or life table, but they are affecting the survival rate since they are a part of it more or less. That is why the rate levels out. If more observations were uncensored, survival rate would have decreased further.

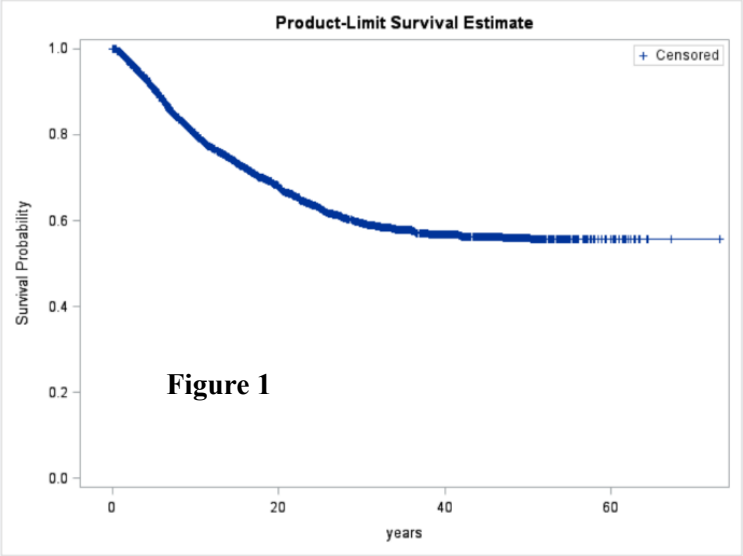

Figure 1

Table 2

Life Table Survival Estimates

Interval [Lower,	Interval Upper)	Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival	Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	Evaluated at the Midpoint of the Interval: PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	5	294	343	3199.5	0.0919	0.00511	1.0000	0	0	.	.	0.0184	0.00102	0.019263	0.001122
5	10	303	343	2562.5	0.1182	0.00638	0.9081	0.0919	0.00511	.	.	0.0215	0.00116	0.025135	0.001441
10	15	163	297	1939.5	0.0840	0.00630	0.8007	0.1993	0.00734	.	.	0.0135	0.00102	0.017546	0.001373
15	20	113	272	1492.0	0.0757	0.00685	0.7334	0.2666	0.00840	.	.	0.0111	0.00101	0.015744	0.00148
20	25	85	251	1117.5	0.0761	0.00793	0.6779	0.3221	0.00925	.	.	0.0103	0.00108	0.015814	0.001714
25	30	43	164	825.0	0.0521	0.00774	0.6263	0.3737	0.0101	.	.	0.00653	0.000975	0.010703	0.001632
30	35	17	146	627.0	0.0271	0.00649	0.5937	0.4063	0.0107	.	.	0.00322	0.000772	0.005497	0.001333
35	40	10	121	476.5	0.0210	0.00657	0.5776	0.4224	0.0111	.	.	0.00242	0.000760	0.004242	0.001341
40	45	2	136	338.0	0.00592	0.00417	0.5655	0.4345	0.0115	.	.	0.000669	0.000472	0.001187	0.000839
45	50	1	116	210.0	0.00476	0.00475	0.5621	0.4379	0.0117	.	.	0.000535	0.000534	0.000955	0.000955
50	55	1	87	107.5	0.00930	0.00926	0.5594	0.4406	0.0120	.	.	0.00104	0.00104	0.001869	0.001869
55	60	0	36	45.0	0	0	0.5542	0.4458	0.0129	.	.	0	.	0	.

This other survival rate, highlighted in Table 2, shows probability that a couple will remain married longer than a certain time, this time in 5-year intervals. The survival rate also starts at 100% (1.0) since all couples start out married here. The rate then decreases as time from marriage increases and more couples get divorced. In the first two intervals, the survival rate steadily decreases by roughly 10%. Then the survival rate in the next two intervals decreases by around 7%. This rate of decrease continues to get smaller until the survival rate starts leveling out around the 40-45 years interval. After about 55 to 60 years into a marriage, survival rate fully levels out around 55.42% (0.5542). From here on, it seems the chances of staying married will remain around this value. Again, this overall survival rate cannot go any lower due to the effects of over half of the couples used were censored. They are still not shown in the life table, but they are still affecting survival rate since they are a part of it. The survival rate would have decreased further with less censored observations. This survival distribution is relatively similar to that of the previous distribution above.

Table 3

divorce	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	2339	69.39	2339	69.39
1	1032	30.61	3371	100.00

Table 4

Summary Statistics for Time Variable years

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	.	LOGLOG	.	.
25	16.2190	LOGLOG	14.0400	18.5585

Since 69.39% of the couples are censored, according to Table 3, and cannot be used in the survival distribution, statistics like the median and upper quartile survival times cannot be found. These statistics show the number of years it takes for 50% and 75%, respectively, of the couples to divorce. Since only 31.61% are uncensored, it is not possible. However, the lower quartile survival time, which is 16.219 according to Table 4, can still be found since more than

25% of the couples are being used in the analysis. Therefore, it takes 16.219 years for 25% of the couples to divorce.

Next, log-rank and Wilcoxon tests were done to see if there was a difference in survival distributions between couples for differing education levels of the husbands. The null hypothesis, H_0, of these tests says the survival distributions are the same among the 3 education levels. Since the p-values of both tests, highlighted in Table 5 below, are less than alpha = 0.05, we will reject H_0. This indicates that at least 2 of the education levels of husbands between couples have survival distributions significantly different from each other. This difference in the distributions is displayed in Figure 2. Since the survival distribution curve for couples with the husbands having 12-15 years of education is lower than the others, this means that these couples are more likely to divorce as the marriage goes on compared to the other levels.

Table 5

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	17.3887	2	0.0002
Wilcoxon	16.0046	2	0.0003
-2Log(LR)	29.8217	2	<.0001

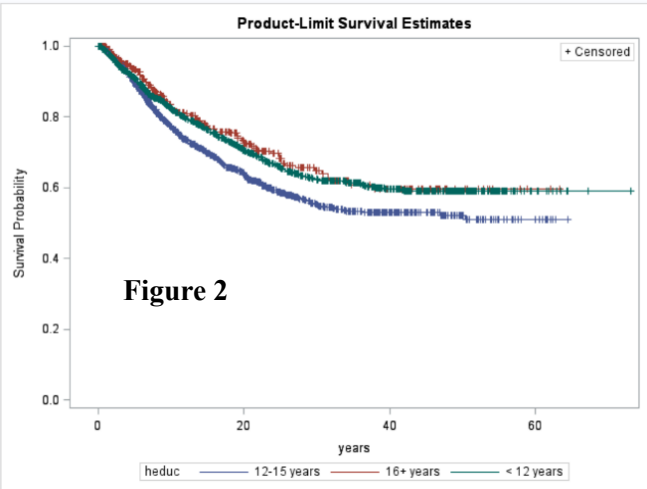

Figure 2

Both tests were also used to see if there was a difference in survival distributions between couples of differing ethnicities/races. The null hypothesis, H_0, of these tests says the survival distributions for both levels, whether they are different ethnicities/races or not, are the same. Since the p-values of both tests, highlighted in Table 6 below, are less than alpha = 0.05, we will

reject H_0. This indicates the survival distributions of both levels are significantly different from each other. This difference in the distributions is displayed in Figure 3. The survival distribution curve for couples of different ethnicities/races is lower than the other, which means that these couples are more likely to divorce as the marriage goes on compared to those of the same ethnicity/race.

Table 6

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	11.3427	1	0.0008
Wilcoxon	8.7862	1	0.0030
-2Log(LR)	15.0092	1	0.0001

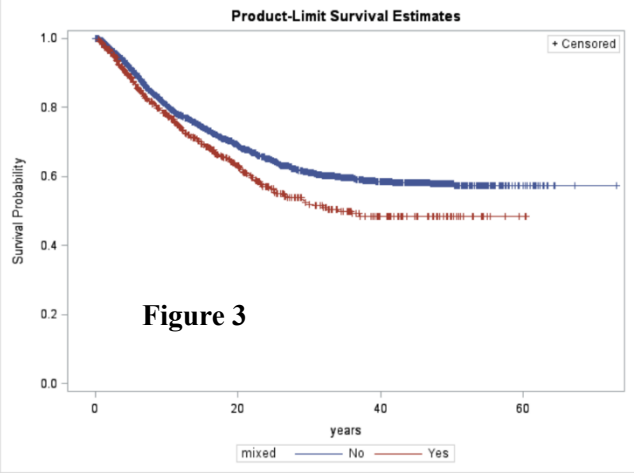

Figure 3

Differences in survival distributions were also looked at based upon age and income as well. The log-rank and Wilcoxon tests were used again, for each variable, with the forward stepwise sequence of chi-squares. The null hypothesis, H_0, of these tests says the survival distributions are the same. The p-values of both age tests, highlighted in Tables 7 and 8 below, are less than alpha = 0.05, so we will reject H_0. This indicates the survival distributions of age are significantly different from each other. As for income, the p-values of these tests, highlighted in Tables 9 and 10 below, are less than alpha = 0.05, so we will reject H_0. This indicates the survival distributions of income are also significantly different from each other.

Table 7

Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
age	1	12.1753	0.0005	12.1753	0.0005

Table 8

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
age	1	12.5790	0.0004	12.5790	0.0004

Table 9

Forward Stepwise Sequence of Chi-Squares for the Log-Rank Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
income	1	5.9539	0.0147	5.9539	0.0147

Table 10

Forward Stepwise Sequence of Chi-Squares for the Wilcoxon Test					
Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
income	1	5.5296	0.0187	5.5296	0.0187

From here on, I starting creating Cox-proportional hazard and parametric models to examine risk factors for divorce based on the variables in the study. I first split up the levels of the husband education variable and created a model, outlined in Table 11, with these variables and the pre-existing ones. However, the education 16+ level variable continued to act insignificant. I also looked at interactions between these variables and the duration of marriage (years), yet none of them were significant either.

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
heduc12to15		1	0.33250	0.06985	22.6613	<.0001	1.394	
heducgt16		1	0.13103	0.11734	1.2470	0.2641	1.140	
heblack	No	1	-0.16568	0.07987	4.3032	0.0380	0.847	heblack No
mixed	No	1	-0.22655	0.07924	8.1750	0.0042	0.797	mixed No
age		1	-0.04336	0.01368	10.0430	0.0015	0.958	
income		1	-0.00588	0.00203	8.3567	0.0038	0.994	

Table 11

Before getting to the final model, I checked the assumption of the Cox Model that all hazards were proportional, meaning relative hazard remains constant over time with different covariate levels. The P-values for all the Martingale Residuals, highlighted in Table 12, appeared to be greater than the 0.05 significance value, indicating no violation for the proportional hazards assumptions for all the variables.

Table 12

Supremum Test for Proportionals Hazards Assumption				
Variable	Maximum Absolute Value	Replications	Seed	Pr > MaxAbsVal
heduc12to15	1.3034	1000	390778001	0.1180
heducgt16	0.8975	1000	390778001	0.5220
heblackNo	1.1894	1000	390778001	0.1430
mixedNo	0.9762	1000	390778001	0.3200
age	0.6445	1000	390778001	0.2770
income	0.5864	1000	390778001	0.8720

After checking the assumption and taking out interactions, I used forward, backward, and stepwise selection to look at final models. They all had the same parameter estimates and p-values, and they also excluded the 16+ education level variable since it was insignificant. I went with the forward selection since it appeared to be the cleanest and never mentioned the insignificant variable. Therefore, I believe the best model for examining risk factors for divorce is **log[h(t)] = 0.30047*(heduc12to15) – 0.15557*(heblack) – 0.21535*(mixed) - 0.0433*(age) - 0.00515*(income).**

Table 13

Analysis of Maximum Likelihood Estimates

Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
heduc12to15		1	0.30047	0.06340	22.4608	<.0001	1.350	
heblack	No	1	-0.15557	0.07924	3.8548	0.0496	0.856	heblack No
mixed	No	1	-0.21535	0.07848	7.5290	0.0061	0.806	mixed No
age		1	-0.04330	0.01367	10.0315	0.0015	0.958	
income		1	-0.00515	0.00192	7.1616	0.0074	0.995	

Summary of Forward Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	heduc12to15	1	1	16.9285	<.0001
2	mixed	1	2	14.5860	0.0001
3	age	1	3	10.4711	0.0012
4	income	1	4	7.6884	0.0056
5	heblack	1	5	3.8578	0.0495

Table 13 above shows parameter estimates and hazard ratios for this model. The first coefficient, 0.30047, means that on average, with all covariates held constant, couples where the husband had at least 12 to 15 years of education are expected to have a 35.049% (exp(0.30047)-1) longer time to divorce than couples where the husband had less than 12 years of education. The accompanying hazard ratio of 1.35 means that, after holding all other covariates constant, hazard of divorce is 35% higher for couples where the husbands have 12 to 15 years of education compared to couples with husbands having less than 12 years. The next coefficient, -0.15557, means that on average, with all covariates held constant, couples where the husband wasn't black are expected to have a 14.4% (exp(-0.15557)-1) shorter time to divorce than couples where the

husband was black. The hazard ratio of 0.856 means that, on average, after holding all other covariates constant, couples were the husband isn't black have a 14.4% lower hazard of divorce compared to couples where the husband was black. The third coefficient, -0.21535, means that on average, with all covariates held constant, couples that are not of the same races/ethnicities are expected to have a 19.374% (exp(-0.21535)-1) shorter time to divorce than couples of the same race/ethnicity. The hazard ratio of 0.806 means that, on average, after holding all other covariates constant, couples of different races/ethnicities have a 19.4% lower hazard of divorce compared to couples of the same race/ethnicity. The coefficient of -0.0433 means that on average, with all covariates held the same between two couples, a couple who has a one point older average age at the date of the wedding will be expected to have a 4.24% (exp(-0.0433)-1) shorter time to divorce than a couple with a one point younger average age. The age hazard ratio of 0.958 means that, on average, holding other variables constant, a couple being one point older in average age is associated with a 0.042% decrease in hazard of divorce compared to couples one point younger. Lastly, the coefficient of -0.00515 means that on average, with all covariates held constant, a couple who has a one dollar higher annual income will be expected to have a 0.514% (exp(-0.00515)-1) shorter time to divorce than a couple with a one dollar lower income. The income hazard ratio of 0.995 means that, on average, holding other variables constant, for every one dollar increase in income, hazard of divorce is predicted to decrease by 0.005%.

As for parametric modeling, I believe the best model for examining risk factors for divorce is **log[S(t)] = 1.1132 + 0.2961*(mixed) + 0.0747*(age) + 0.0043*(income),** which is outlined in Table 14**.** For the intercept of 1.1132, it means that on average, we predict the expected time to divorce is exp(1.1132) = 3.044 years, given all covariates are set to take the value of zero. The first coefficient, 0.2961, means that on average, with all covariates held

Table 14

Analysis of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept		1	1.1132	0.5618	0.0122 2.2143	3.93	0.0475
mixed	No	1	0.2961	0.0974	0.1052 0.4871	9.24	0.0024
mixed	Yes	0	0.0000	.	. .	.	.
age		1	0.0747	0.0205	0.0346 0.1148	13.30	0.0003
income		1	0.0043	0.0024	-0.0005 0.0090	3.08	0.0794
Scale		1	2.0301	0.0663	1.9043 2.1643		
Shape		1	-0.6686	0.1416	-0.9460 -0.3912		

constant, couples of the same race/ethnicity are expected to have a 34.46% (exp(0.2961)-1) longer time to divorce than couples of different races/ethnicities. The coefficient of 0.0747 means that on average, with all covariates held the same between two couples, a couple who has a one point older average age at the date of the wedding will be expected to have a 77.56% (exp(0.0747)-1) longer time to divorce than a couple with a one point younger average age. Lastly, the coefficient of 0.0043 means that on average, with all covariates held constant, a couple who has a one dollar higher annual income will be expected to have a 0.43% (exp(0.0043)-1) longer time to divorce than a couple with a one dollar lower income.

I originally considered a model with all variables included, but the P-values for husband's education and heblack, shown in Table 15, were insignificant. So, I tried two more models, one containing education and not heblack, and the other containing heblack but not education. These models also had insignificant P-values for these variables as well. So, I decided to try the model without them and it worked very well, as seen above.

Table 15

Analysis of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept		1	1.2407	0.5524	0.1581 2.3234	5.05	0.0247
heduc	12-15 years	1	-0.3026	0.0907	-0.4804 -0.1248	11.12	0.0009
heduc	16+ years	1	-0.0312	0.1428	-0.3112 0.2487	0.05	0.8268
heduc	< 12 years	0	0.0000	.	. .	.	.
heblack	No	1	0.0885	0.1002	-0.1078 0.2849	0.78	0.3767
heblack	Yes	0	0.0000	.	. .	.	.
mixed	No	1	0.2930	0.1012	0.0946 0.4915	8.38	0.0038
mixed	Yes	0	0.0000	.	. .	.	.
age		1	0.0720	0.0201	0.0327 0.1114	12.88	0.0003
income		1	0.0058	0.0027	0.0006 0.0110	4.76	0.0292
Scale		1	1.9727	0.0732	1.8343 2.1216		
Shape		1	-0.5302	0.1470	-0.8184 -0.2421		

The assumption of parametric models is whether the model fits one of the standard parametric model distributions (i.e. Gamma, Weibull, Exponential, or Log-normal). Tables 16 and 18 show log-likelihood and AIC values for the four parametric distributions for inclusion and exclusion of the education and heblack variables, respectively. Comparisons of these models, their log-likelihood ratios, and conclusions are in Tables 17 and 19, respectively as well. If the ratio was greater than the critical value of Chi-square, the first model would fit better. The second model would fit if the ratio was less than that value. They all say the Gamma distribution fits the parametric model I made the best, meeting the assumption.

Table 16

Model	Log-likelihood	AIC
Exponential	-3158.467	6324.934
Weibull	-3144.471	6298.943
Log-Normal	-3079.143	6168.286
Gamma	-3067.566	6147.132

Table 17

Comparison	$2(L_1-L_2)$	Critical Value	Conclusion
Gamma vs. Log-Normal	23.154	3.84	Gamma
Gamma vs. Weibull	153.81	3.84	Gamma
Weibull vs. Exponential	27.992	3.84	Weibull
Gamma vs. Exponential	181.802	5.99	Gamma

Table 18

Model	Log-likelihood	AIC
Exponential	-3133.822	6281.644
Weibull	-3124.193	6264.386
Log-Normal	-3067.590	6151.223
Gamma	-3060.650	6139.299

Table 19

Comparison	$2(L_1-L_2)$	Critical Value	Conclusion
Gamma vs. Log-Normal	13.88	3.84	Gamma
Gamma vs. Weibull	127.086	3.84	Gamma
Weibull vs. Exponential	19.258	3.84	Weibull
Gamma vs. Exponential	146.344	5.99	Gamma

Overall, I believe the final Cox proportional hazards model I found, **log[h(t)] = 0.30047*(heduc12to15) – 0.15557*(heblack) – 0.21535*(mixed) - 0.0433*(age) - 0.00515*(income),** was the best model compared to the final parametric model with the gamma

distribution. The assumption for the Cox model was definitely more clearly proven to be valid, with the P-values of the Martingale residuals all being less than 0.05, compared to the complex log-likelihood ratios used to choose the Gamma model for the assumption of the parametric model. Plus, the Gamma model is much more generic and is the go-to model when the parameters and ratios don't match up with the other model types. I do not think it is trustworthy compared to the Cox model. The Cox model, done with forward selection, actually removed the insignificant variables from the model instead of me having to guess and remove them manually like with the parametric model. I do also believe splitting up the levels of husband's education into separate variables helped this model greatly since it was possible to remove the insignificant value level of 16+ years. The variable heblack also acted significant in every Cox model I made whereas it was more insignificant with the parametric models. It should also be said that there is greater flexibility in the estimates of the Cox model compared to the parametric model.