Marriage Survival Analysis: Secrets Behind "Happily Ever Afters"

C240B / STAT C245B Project Final Report

Z. Tom Hu
Sijin Wu

Berkeley
UNIVERSITY OF CALIFORNIA

Content

1. Overview

2. Research Goal Motivations

3. TMLE and COX

4. Conclusion

1. Overview

China Family Panel Studies（CFPS） of 2010, 2012, 2014, 2016, 2018.

Each of them has about 30000 samples with 1000 covariates. Sample distribution is balanced in gender and marriage state, unbalanced in region.

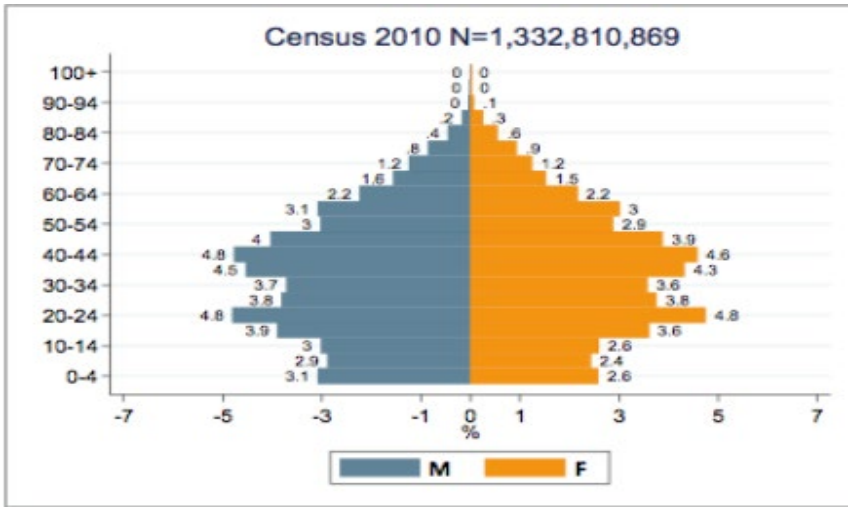
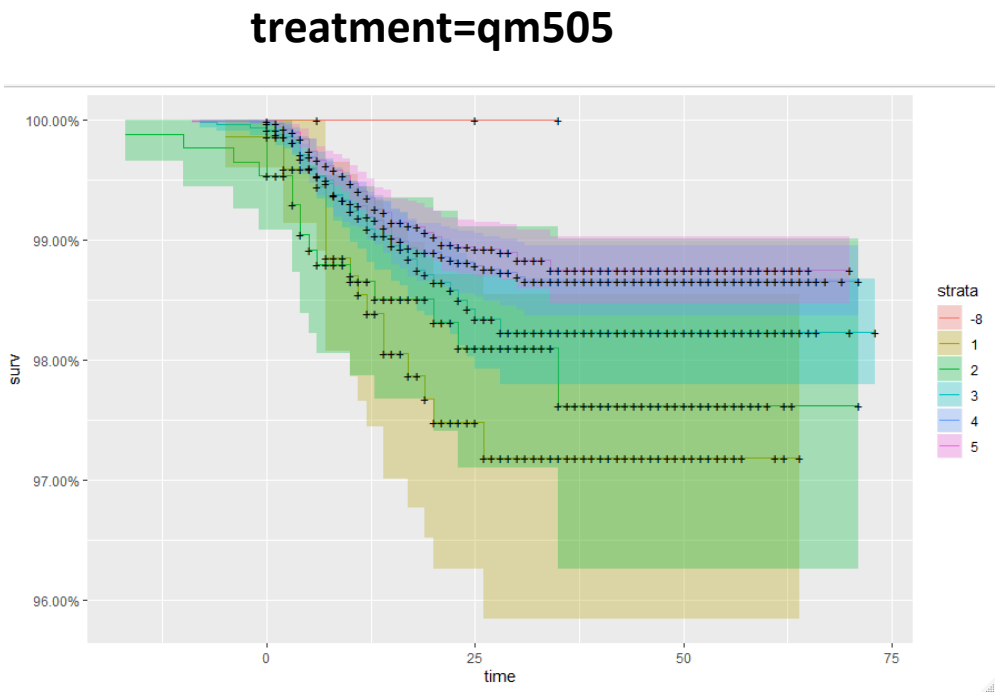


Table 49. Distribution of Marital Status, Ages 15 or above (%)

		CFPS 2010				
		T1 members	Individual respondents	2010 Census	CHFS 2011[a]	CGSS 2010[b]
Total	Unmarried	20.8	14.6	21.6	18.2	8.1
	Married	72.2	78.5	71.3	76.5	82.8
	Divorced	1.2	1.2	1.4	1.3	2.1
	Widowed	5.8	5.8	5.7	4.0	7.0
	N	30,642	22,197	105,542,243	24,693	10,154
Male	Unmarried	24.1	17.0	21.6	21.1	10.1
	Married	71.2	78.2	71.3	75.3	83.6
	Divorced	1.4	1.4	1.4	1.2	2.1
	Widowed	3.3	3.4	5.7	2.3	4.1
	N	15,454	10,732	52,943,450	12,352	4,932
Female	Unmarried	17.3	12.3	18.5	15.2	6.3
	Married	73.2	78.8	72.3	77.7	82.0

Berkeley
UNIVERSITY OF CALIFORNIA

2. Research Goal Motivations



Motivation 1： Recall from our last presentation, we have shown that using some machine learning methods it can be shown that covariate qm505 is highly correlated with wedding time. qm505 is measuring how intolerable the interviewee thinks about loneliness, in a scale of 1 to 5.

2. Research Goal Motivations

Motivation 2： At the beginning we set all covariates except wedding time as baseline covariates. However, only MOSS package supports censored data among current TMLE packages.

But we have a large dataset, and different covariates have different patterns of missingness. Thus we have to do a variable selection based on missingness.

We then used the `initial_sl_fit()` in (MOSS) to specify the data (as defined it above) and the SuperLearner library for initial estimation. The data size we used is of 25000 samples and 1400 features. It took about 4 hours of training and resulted in a vector whose size was larger than 7M, which meant the result is invalid.

We need to "slim down" the data again.

2. Research Goal Motivations

From Motivation 1, we realize that ”internal” covariates might be more important, such as:

☑ qm501	Importance: Having lots of money
☑ qm502	Importance: Not being disliked by others
☑ qm503	Importance: Having fun in life
☑ qm504	Importance: Intimate relationship with spouse
☑ qm505	Importance: Not lonely
☑ qm506	Importance: Feeling successful
☑ qm507	Importance: Being missed posthumously
☑ qm508	Importance: A happy and harmonious family
☑ qm509	Importance: Having children to carry on the family name
☑ qm510	Importance: Children being successful

All of them are in a scale of 1 to 5.

We also select income rank, education level, appearance, IQ, expression ability and social status as “external” covariates for comparison. These covariates are in a scale of 1 to 8.

We want to infer the effects of these covariates as treatment through simultaneous inference curve.

2. Research Goal Motivations

Moss requires binary treatments…

- Internal: We separate into two classes (<=3) and (>= 4)
- External: We separate into two classes (<=5, >= 6)

These cutoffs are selected by our background knowledge. For example, as for degree, 6 = bachelor, 7 = master and 8 = PhD.

We then did a conditional sampling based on censored and complete data with finally 2500 samples.

3. TMLE

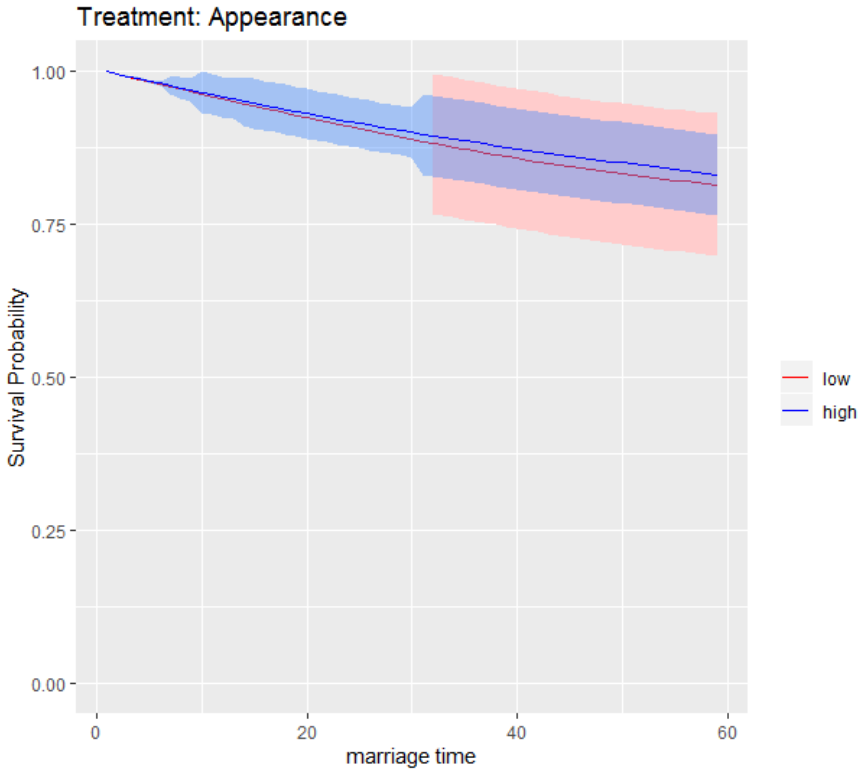
Since MOSS have some problems to deal with estimating the exact average treatment effect:

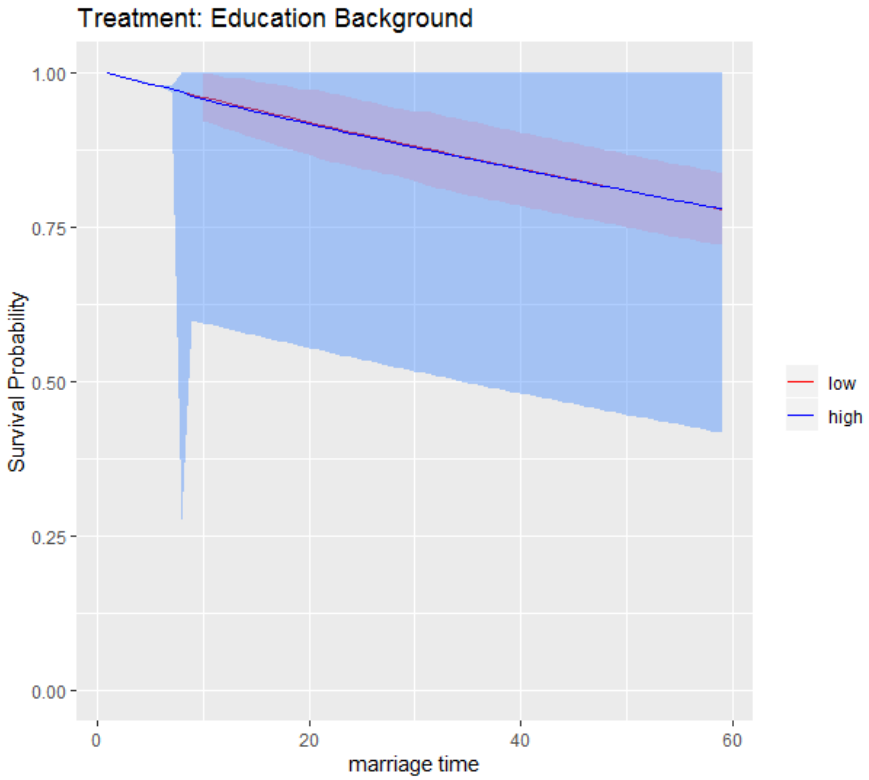
$$ATE = \mathbb{E}_0(Y(1) - Y(0)) = \mathbb{E}_0\left(\mathbb{E}_0[Y \mid A = 1, W] - \mathbb{E}_0[Y \mid A = 0, W]\right)$$

We plot the Simultaneous Confidence Survival Curve by TMLE with different treatment on the same plot and observe if there is a distinct difference.

3. TMLE: External Covariates

The outcomes of external covariates: We can say that they nearly have no obvious effect on marriage lifespan.

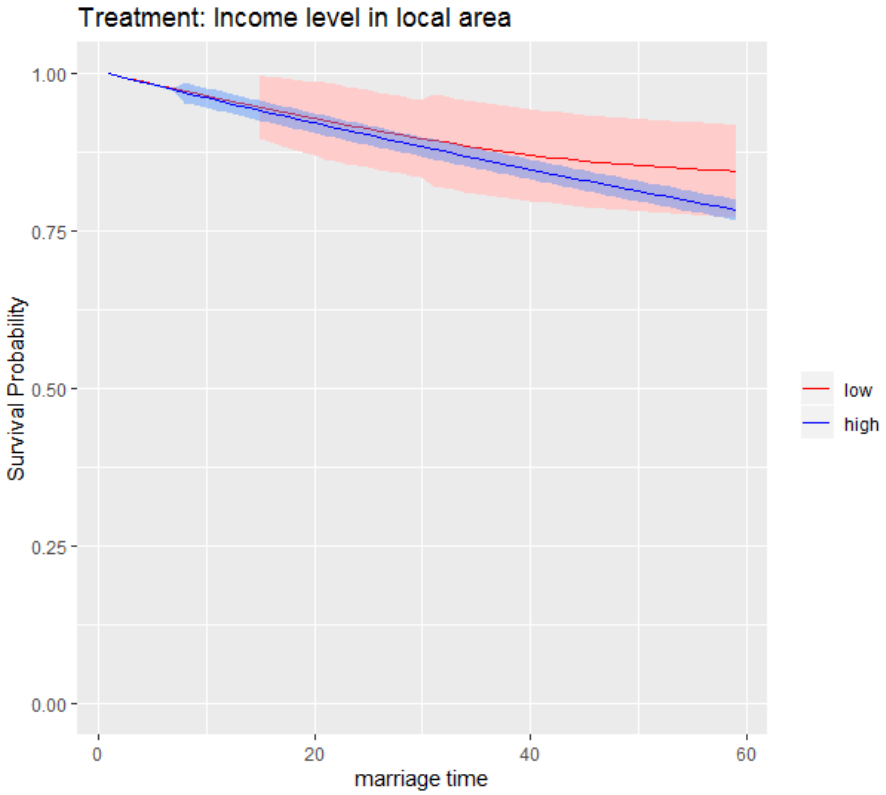


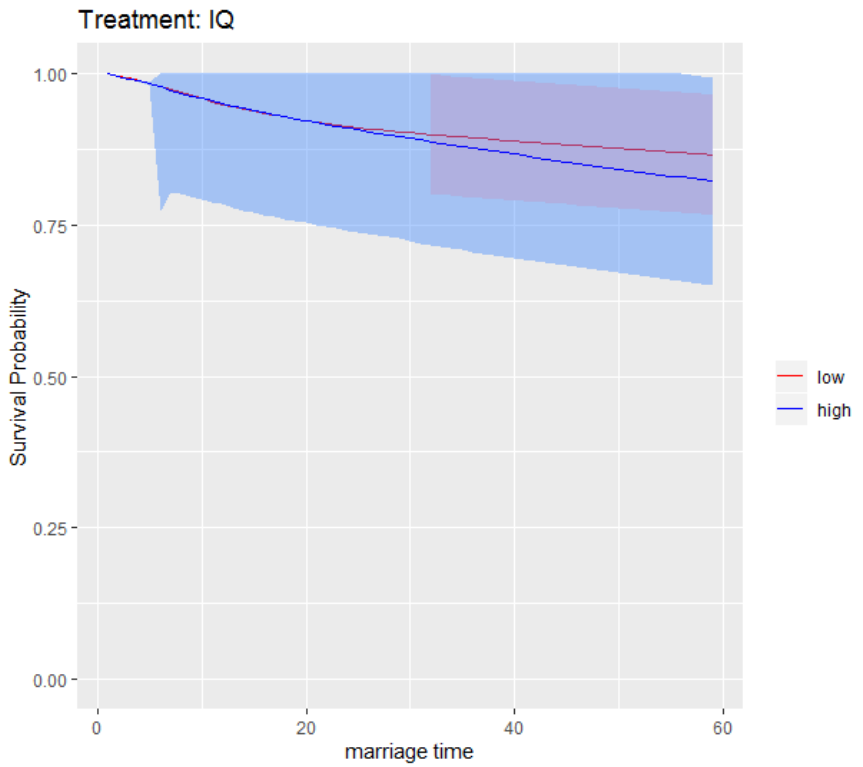


Berkeley
UNIVERSITY OF CALIFORNIA

3. TMLE: External Covariates

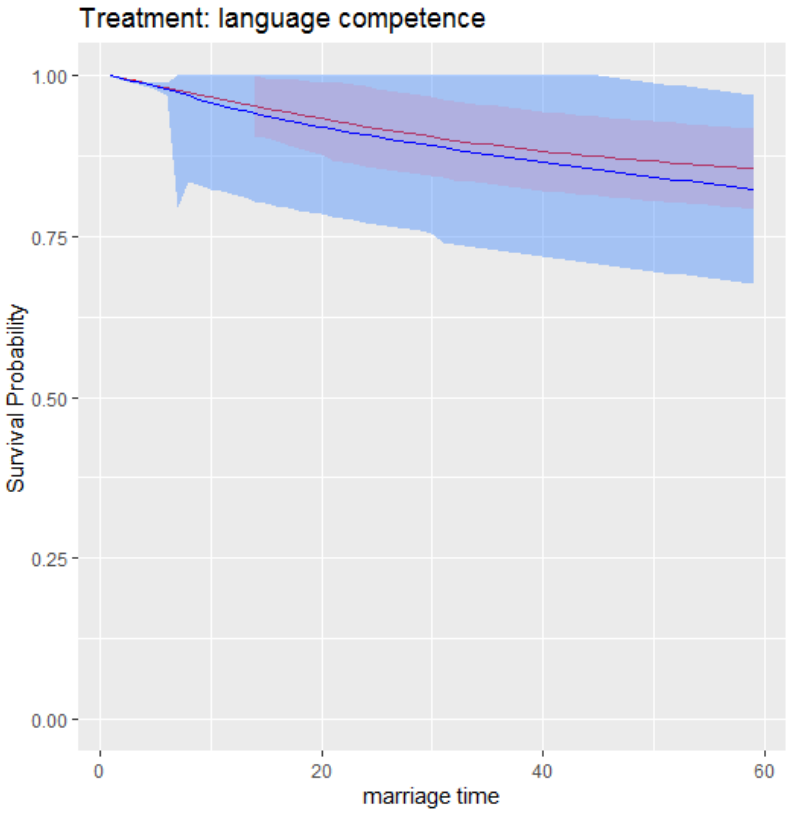
The outcomes of external covariates: We can say that they nearly have no obvious effect on marriage lifespan.

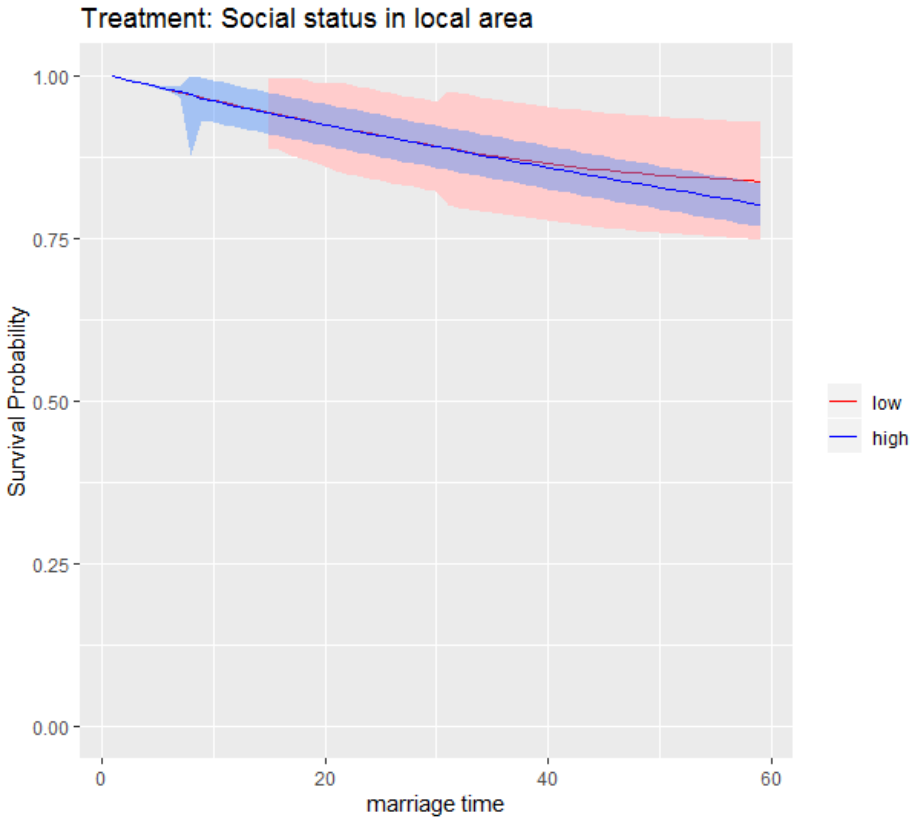




3. TMLE: External Covariates

The outcomes of external covariates: We can say that they nearly have no obvious effect on marriage lifespan.

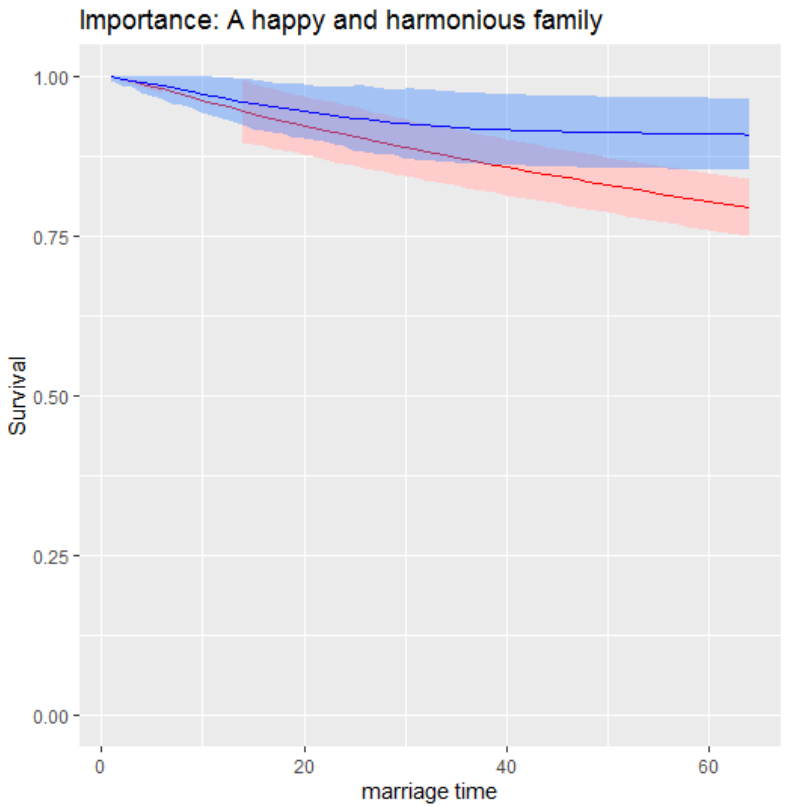


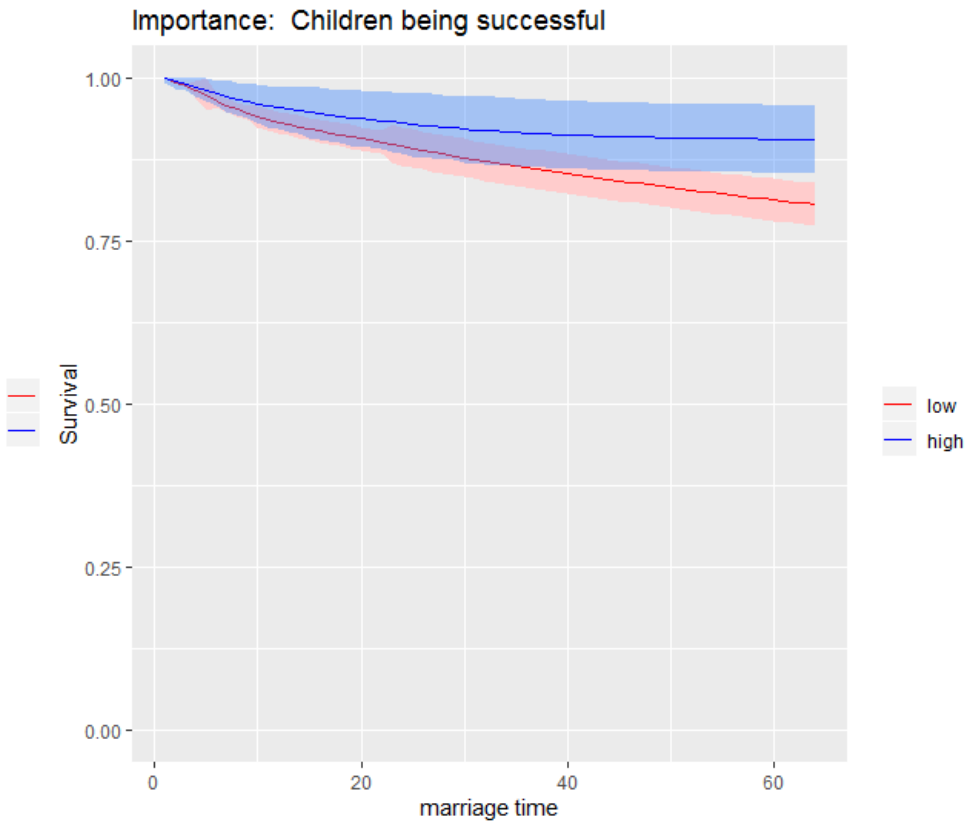


Berkeley
UNIVERSITY OF CALIFORNIA

3. TMLE: Internal Covariates

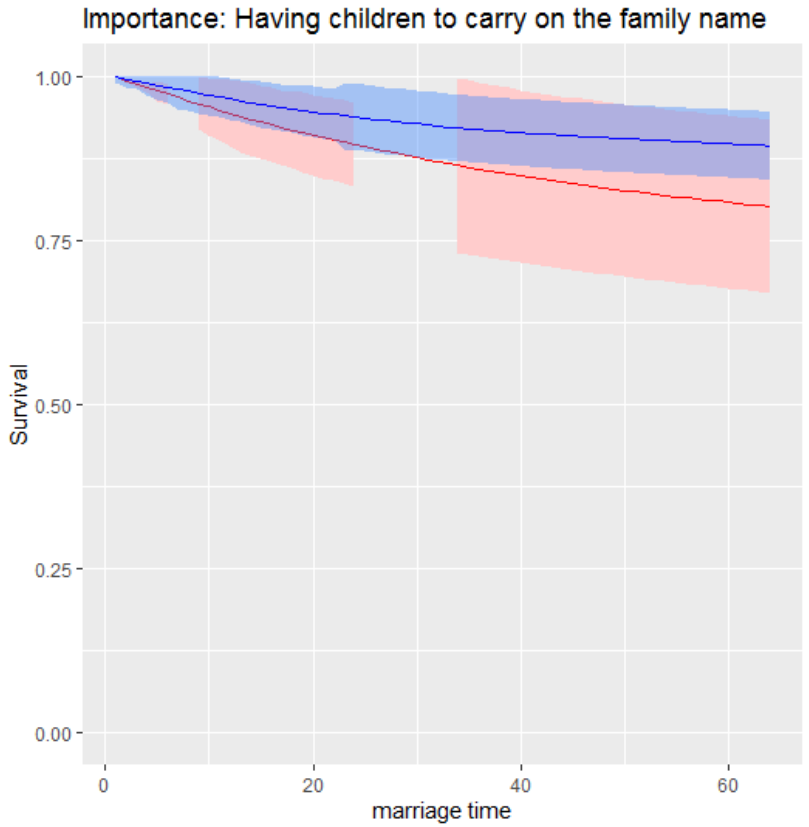
These internal covariates as treatments cause a distiction between different treatments.

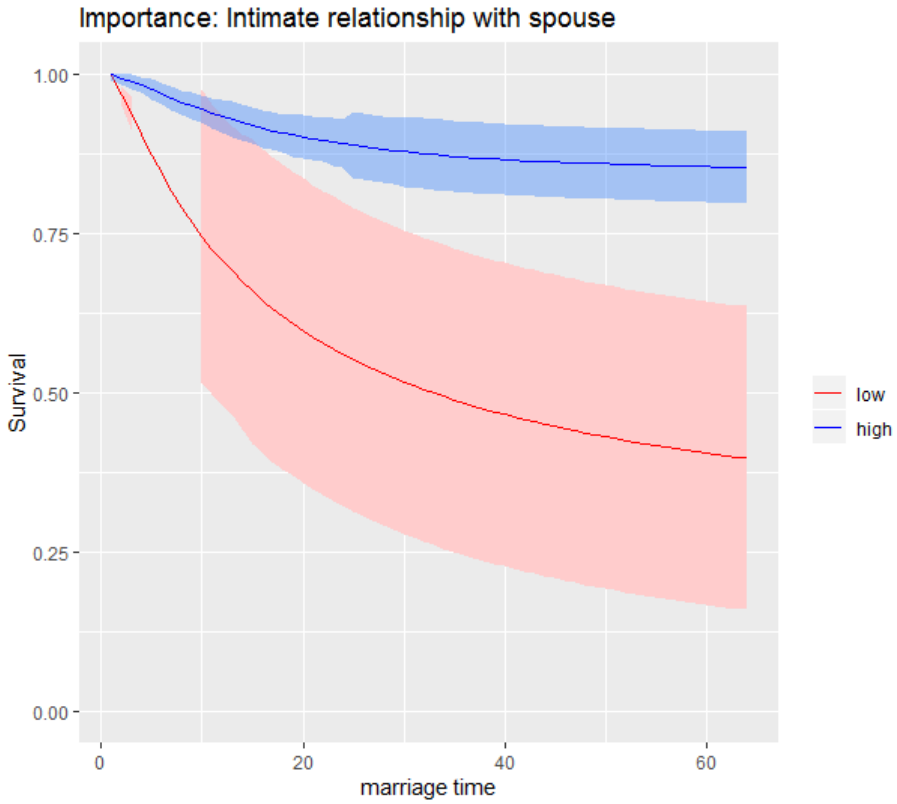




3. TMLE: Internal Covariates

Another set of internal covariates.





3. COX

STHDA
Statistical tools for high-throughput data analysis

In this project, we also tried the COX model to see the difference of the results through 2 different approaches.

Here, we'll disscuss three types of diagonostics for the Cox model:

- Testing the proportional hazards assumption.
- Examining influential observations (or outliers).
- Detecting nonlinearity in relationship between the log hazard and the covariates.

In order to check these model assumptions, Residuals method are used. The common residuals for the Cox model include:

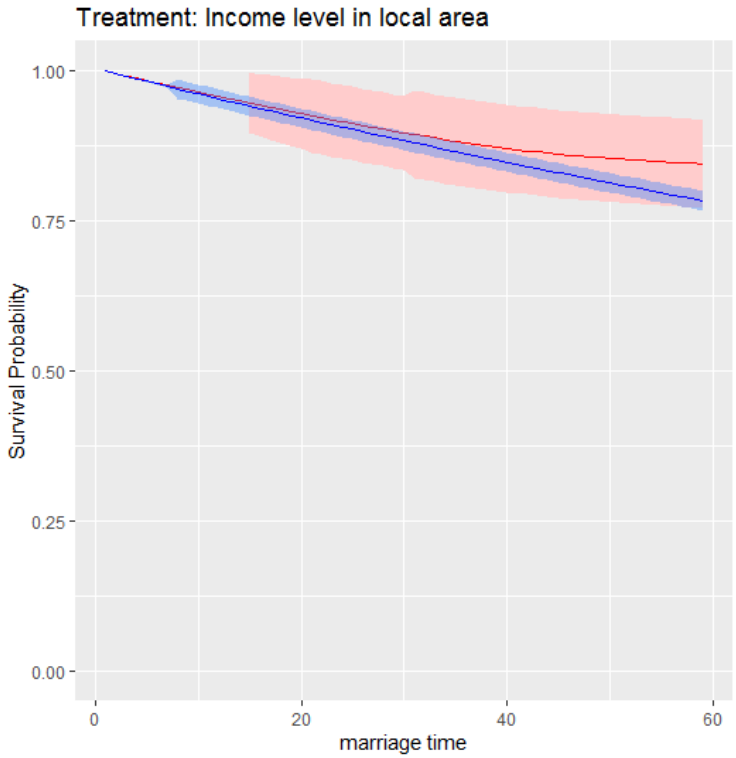
- Schoenfeld residuals to check the proportional hazards assumption
- Martingale residual to assess nonlinearity
- Deviance residual (symmetric transformation of the Martinguale residuals), to examine influential observations

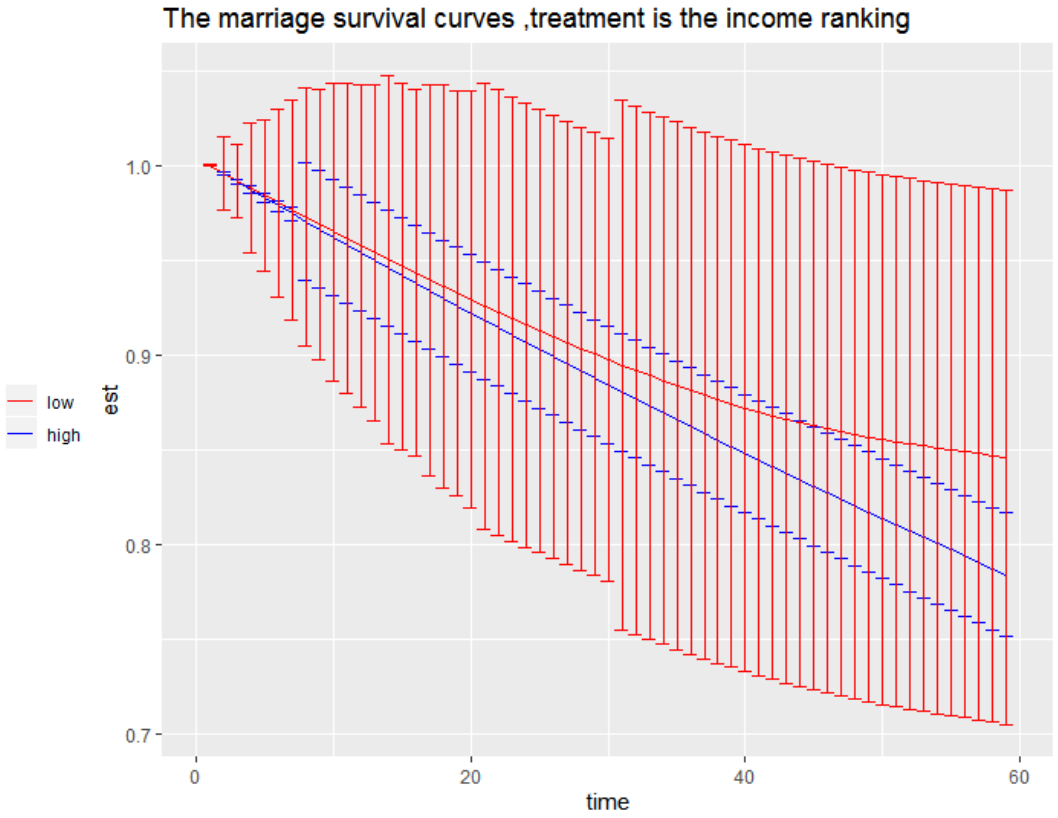
We will use the "survival" and "survminer" to do the cox regression and cox assumption tests for external and internal covariates seperately.

Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: External Covariates

Compared with TMLE's results: the local income rank

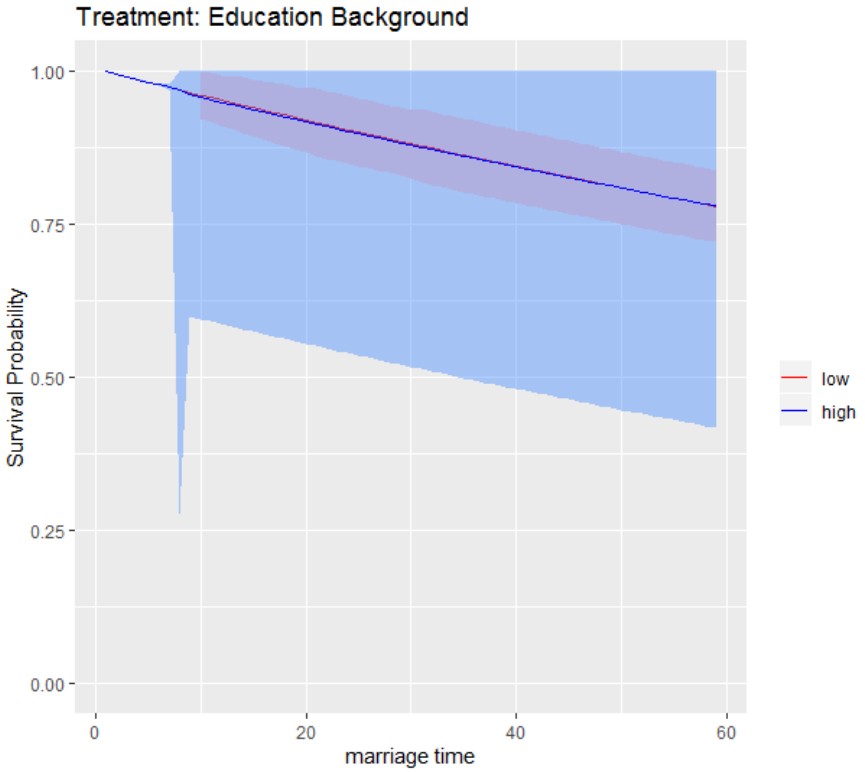




Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: External Covariates

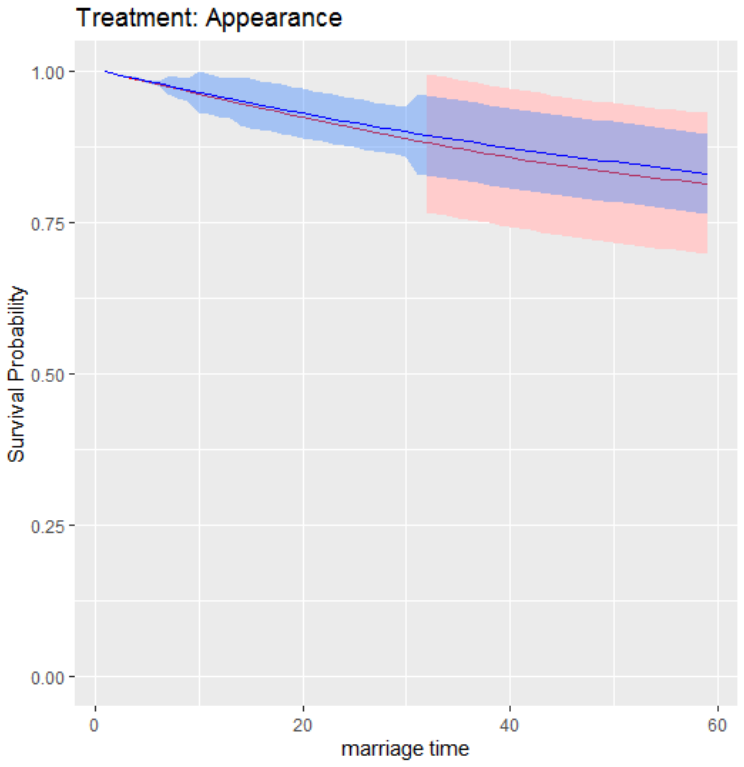
Compared with TMLE's results: education



Berkeley
UNIVERSITY OF CALIFORNIA

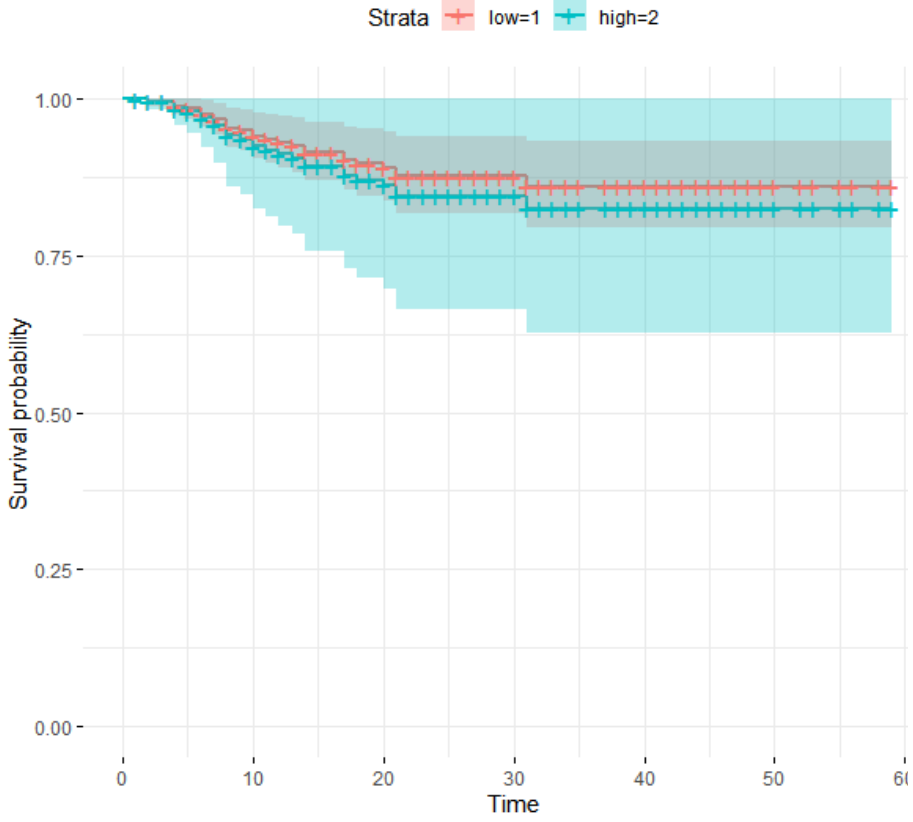
3. COX: External Covariates

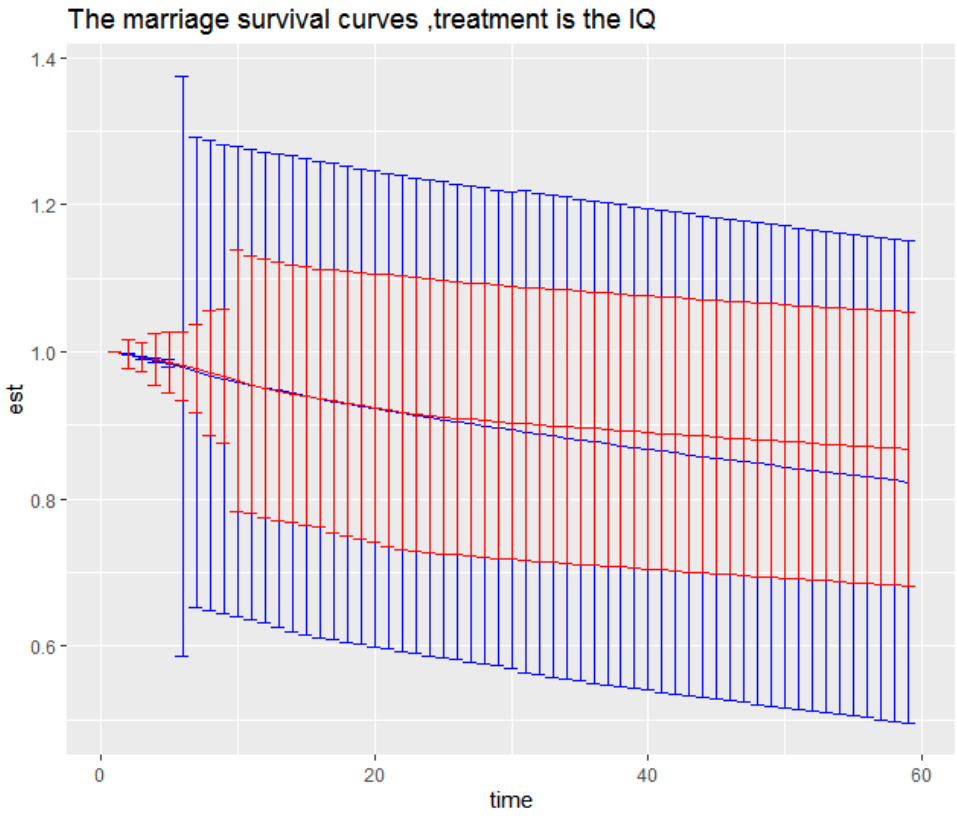
Compared with TMLE's results: appearence



3. COX: External Covariates

Compared with TMLE's results: IQ

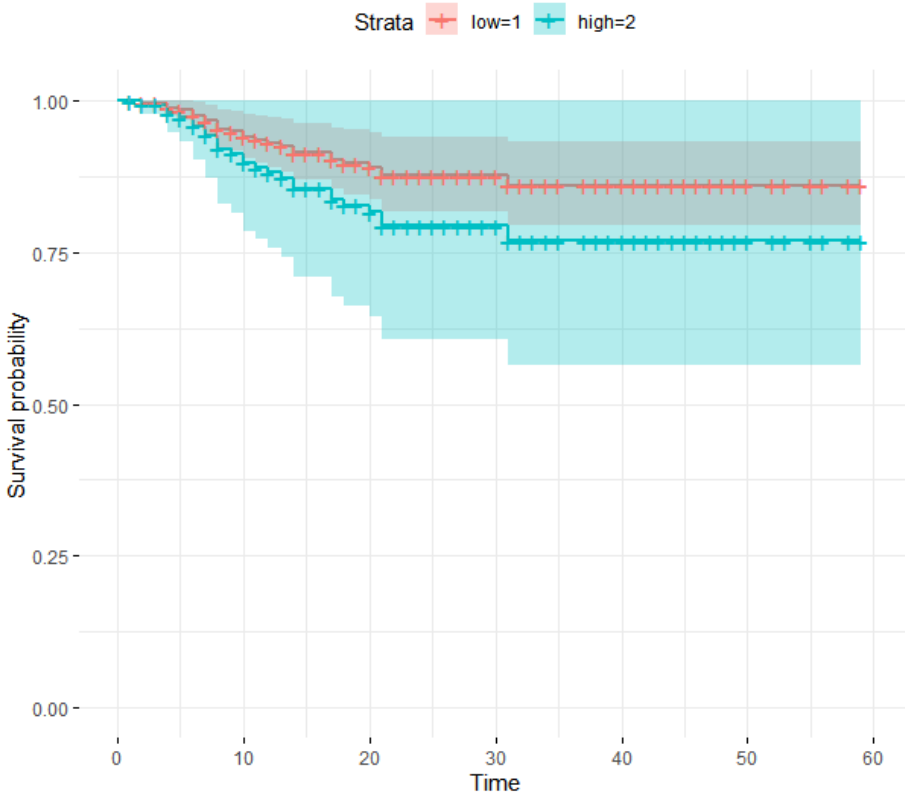


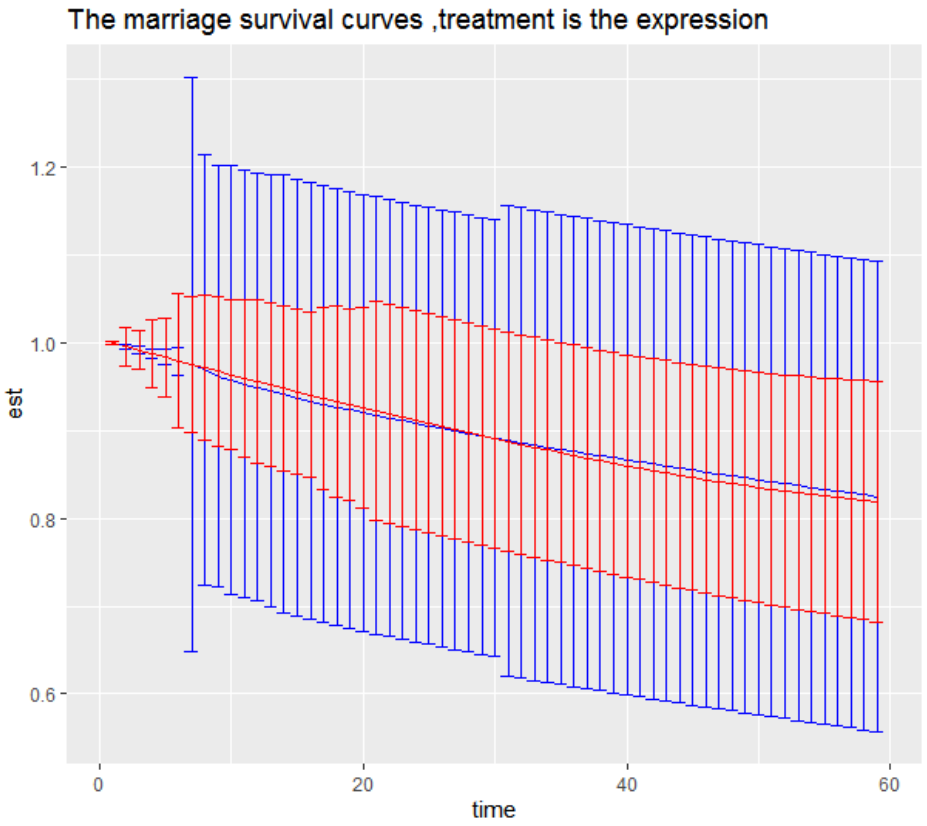


Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: External Covariates

Compared with TMLE's results : expression ability

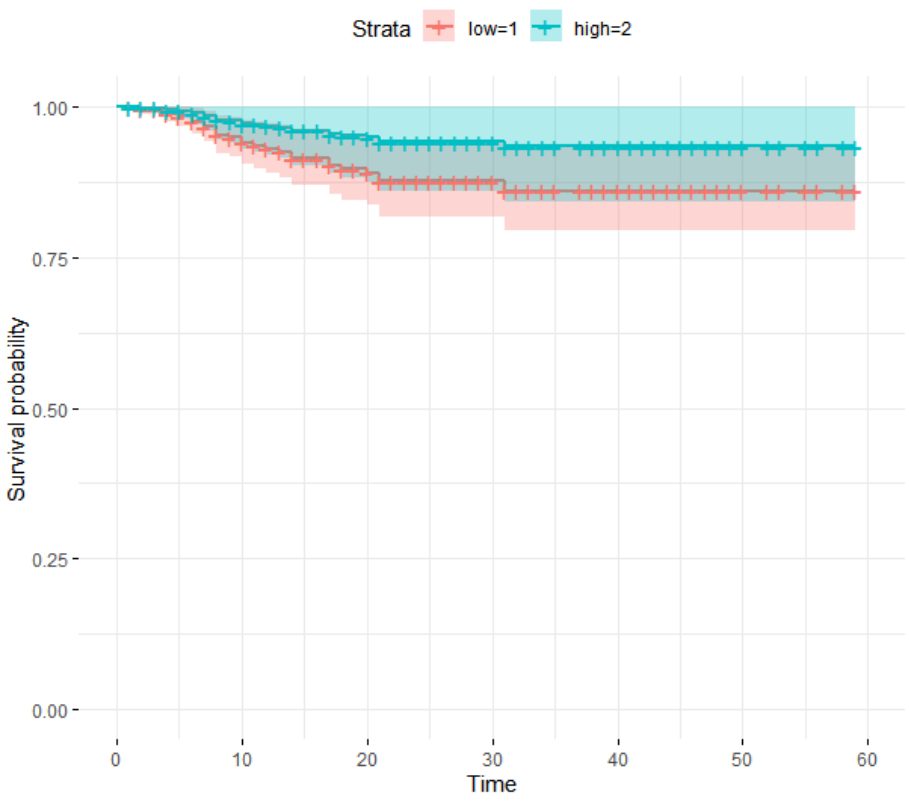




Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: External Covariates

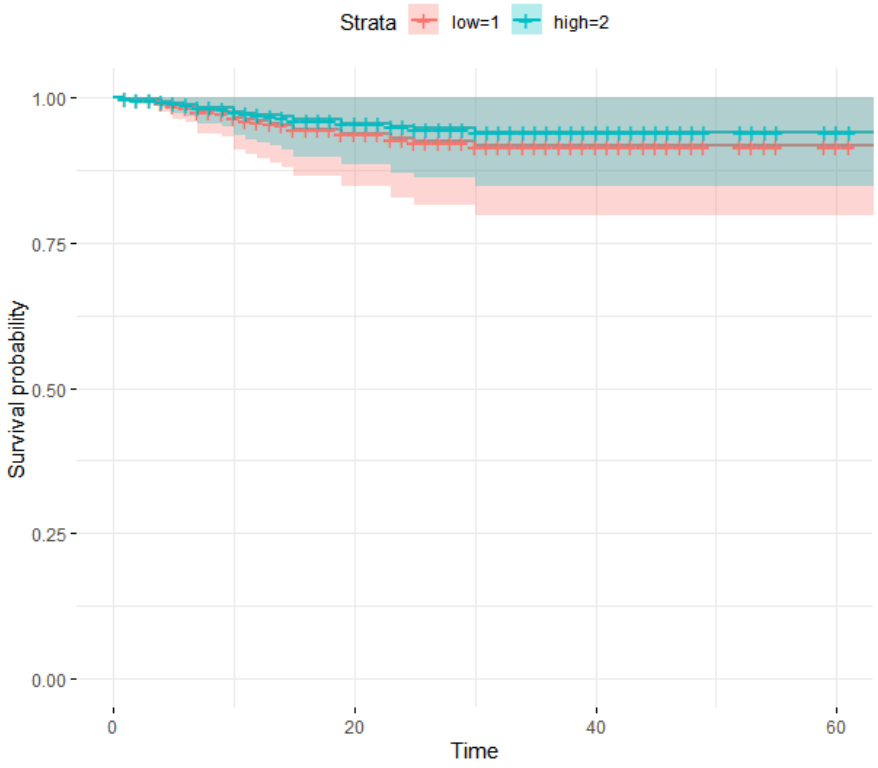
Compared with TMLE's results: local social status

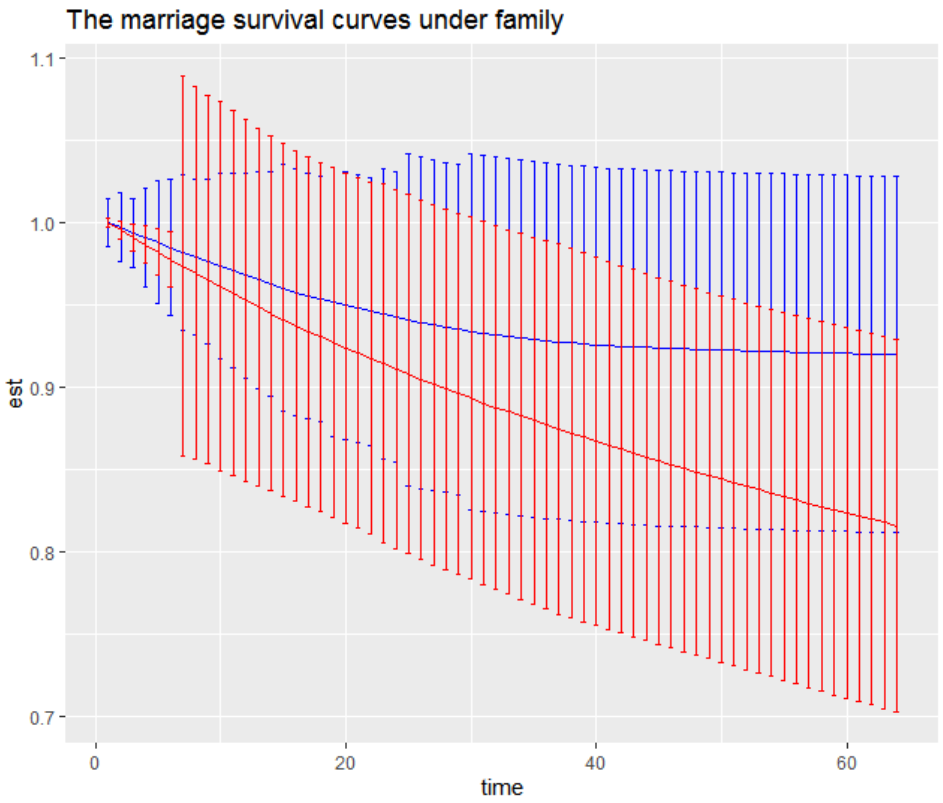


Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: Internal Covariates

Compared with TMLE's results: importance of family

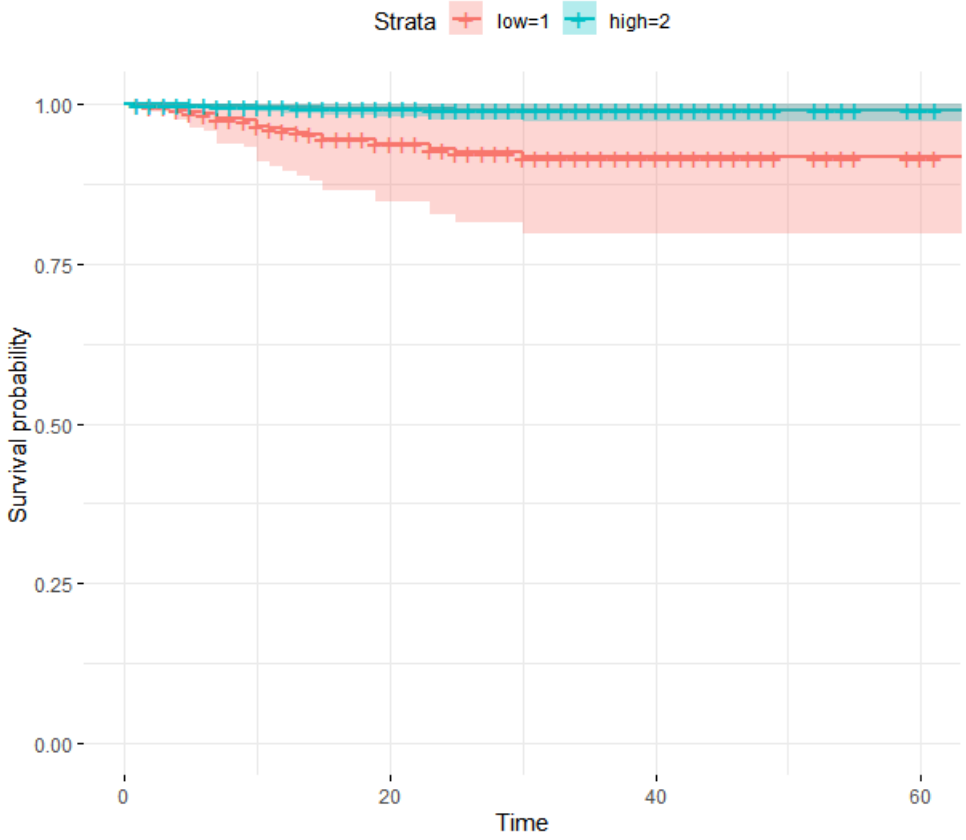


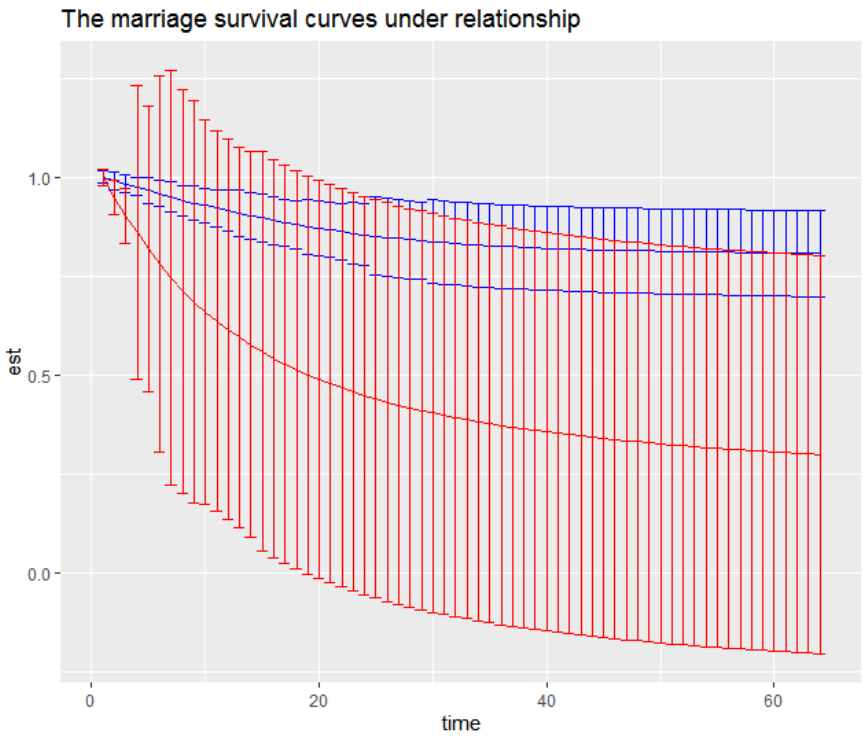


Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: Internal Covariates

Compared with TMLE's results: importance of couple's relation

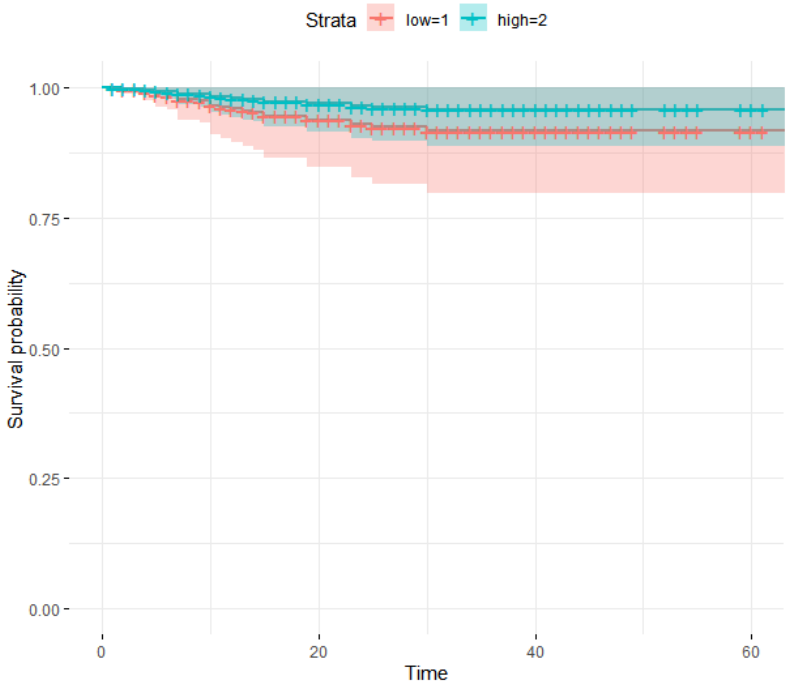


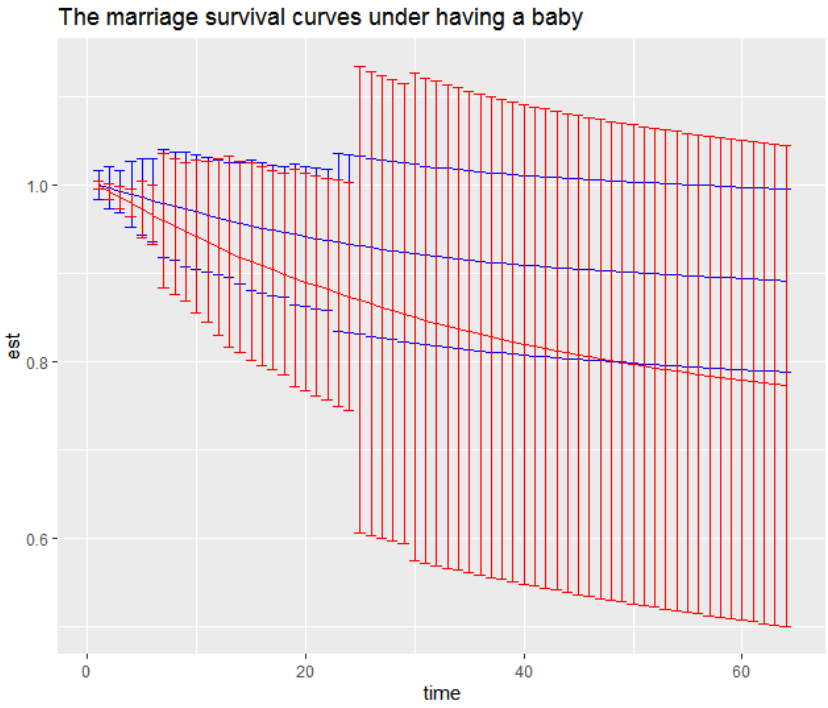


Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: Internal Covariates

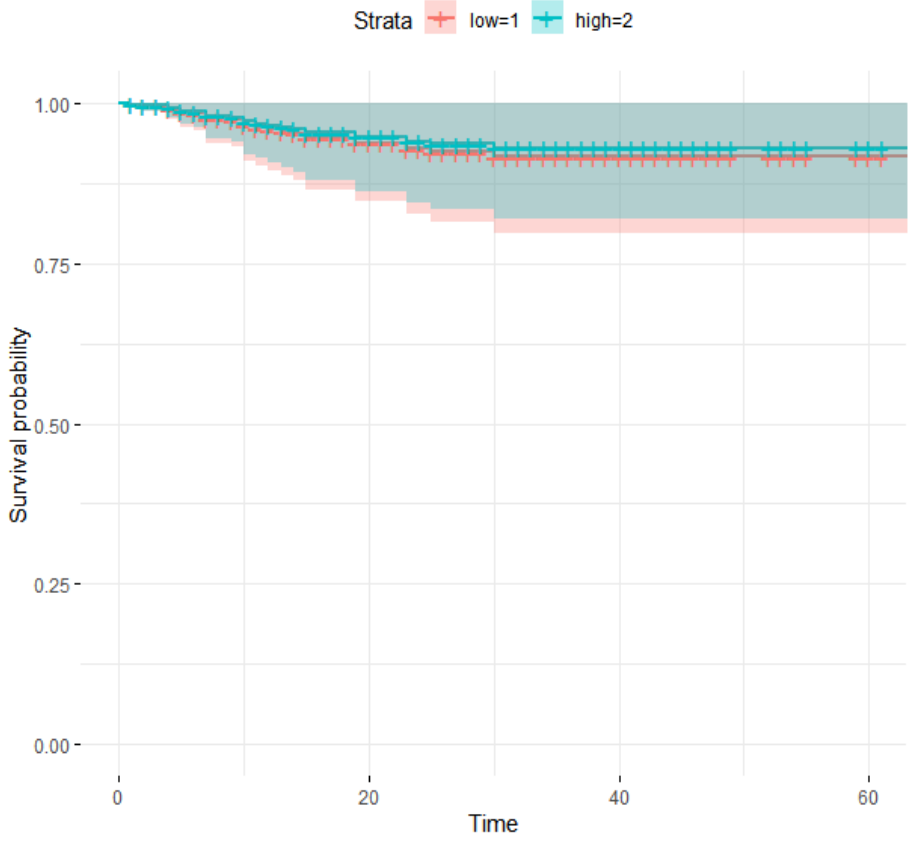
Compared with TMLE's results: importance of having the offspring

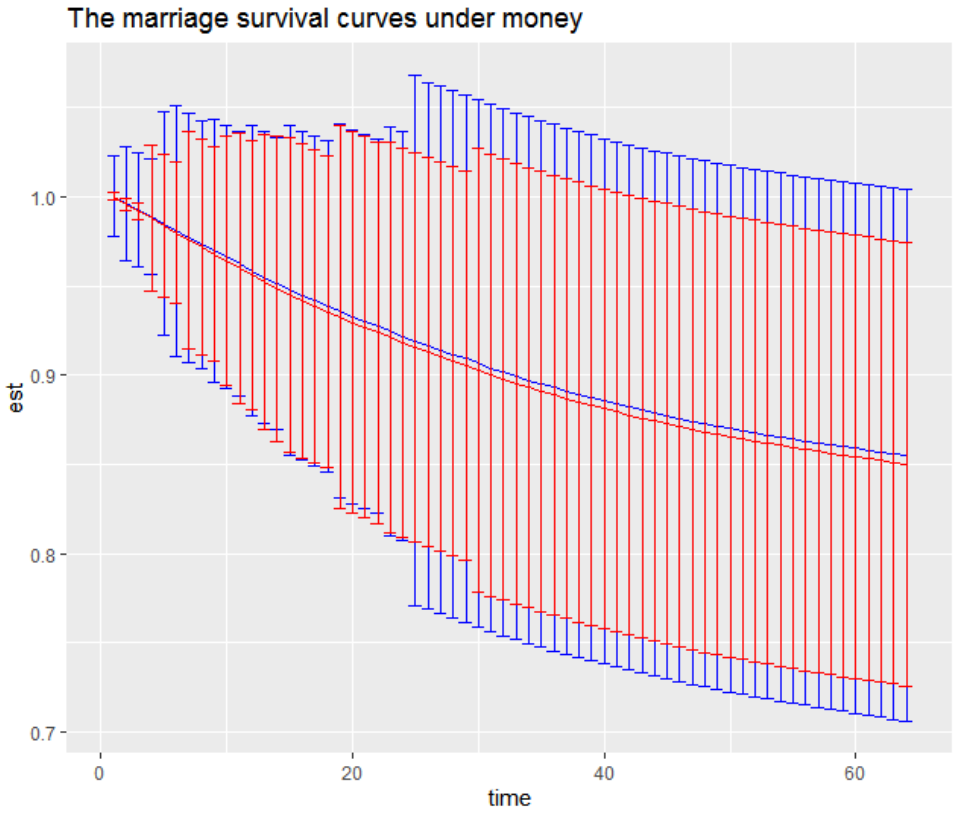




3. COX: Internal Covariates

Compared with TMLE's results: importance of money

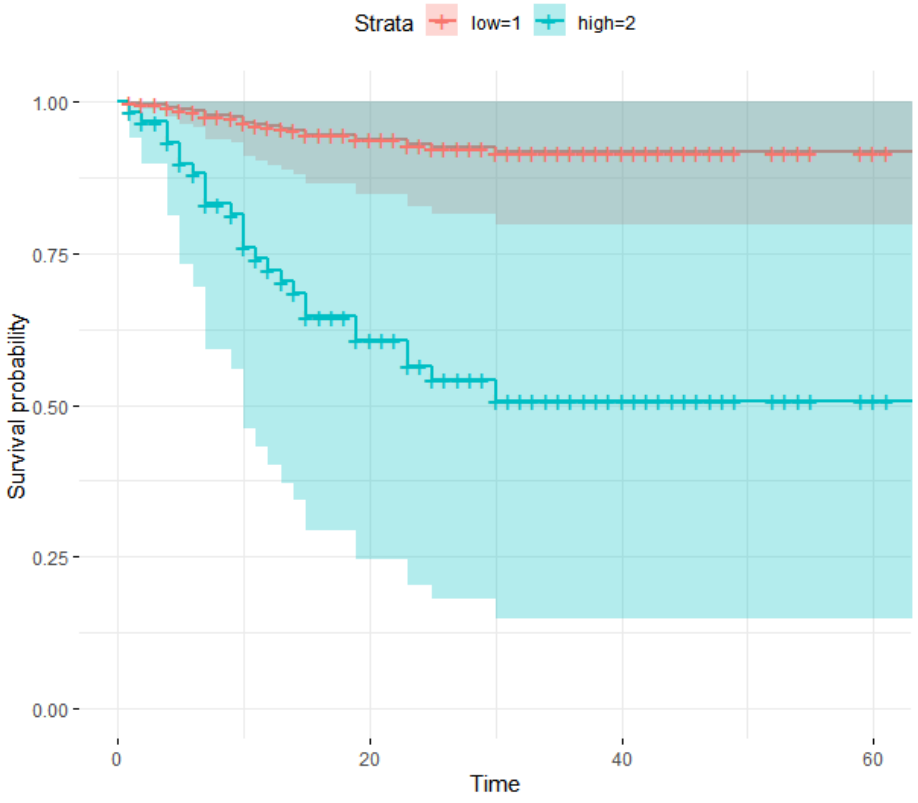




Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: Internal Covariates

Compared with TMLE's results: importance of humor



4. Conclusion

- Two approaches do have different survival curves, possibly due to different assumptions

- But no mater TMLE or COX, we can see that internal covariates have much bigger impact on your wedding length, especially those who think relations are important / families are important have a far longer survived wedding than those who do not.

- Due to time and knowledge constraint, we exclude some of our data and must conduct variable selection by hand. We would be interested to see TMLE applied to large scale data.

Thanks for your attention!
Questions?

3. COX: External Covariates

Model Output:

	coef	exp(coef)	se(coef)	z	Pr(>|z|)
qm401	0.07081	1.07337	1.22007	0.058	0.954
qc1	0.50615	1.65889	1.13753	0.445	0.656
qz204	-0.91297	0.40133	0.62725	-1.456	0.146
qz207	0.24194	1.27371	0.68591	0.353	0.724
qz212	0.55712	1.74564	0.62895	0.886	0.376
qm402	-0.79319	0.45240	0.79801	-0.994	0.320

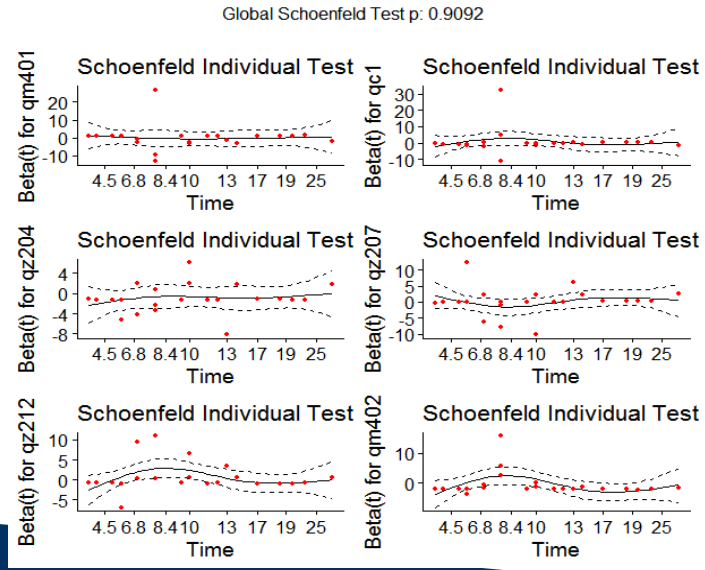
	exp(coef)	exp(-coef)	lower .95	upper .95
qm401	1.0734	0.9316	0.09823	11.729
qc1	1.6589	0.6028	0.17846	15.420
qz204	0.4013	2.4917	0.11738	1.372
qz207	1.2737	0.7851	0.33206	4.886
qz212	1.7456	0.5729	0.50885	5.988
qm402	0.4524	2.2104	0.09468	2.162

Testing proportional Hazards assumption:

	rho	chisq	p
qm401	-0.0292	0.0232	0.879
qc1	-0.0606	0.1358	0.712
qz204	0.1480	0.3968	0.529
qz207	0.0867	0.2532	0.615
qz212	-0.0722	0.1597	0.689
qm402	-0.1790	0.7297	0.393
GLOBAL	NA	2.1108	0.909

The proportional hazard assumption is supported by a non-significant relationship between residuals and time and refuted by a significant relationship. So the test is not statistically significant for each of the covariates, and the global test is also not statistically significant. Therefore, we can assume the proportional hazards.

From the graphical inspection, there is no pattern with time. The assumption of proportional hazards appears to be supported



Berkeley
UNIVERSITY OF CALIFORNIA

3. COX: Internal Covariates

Model Output

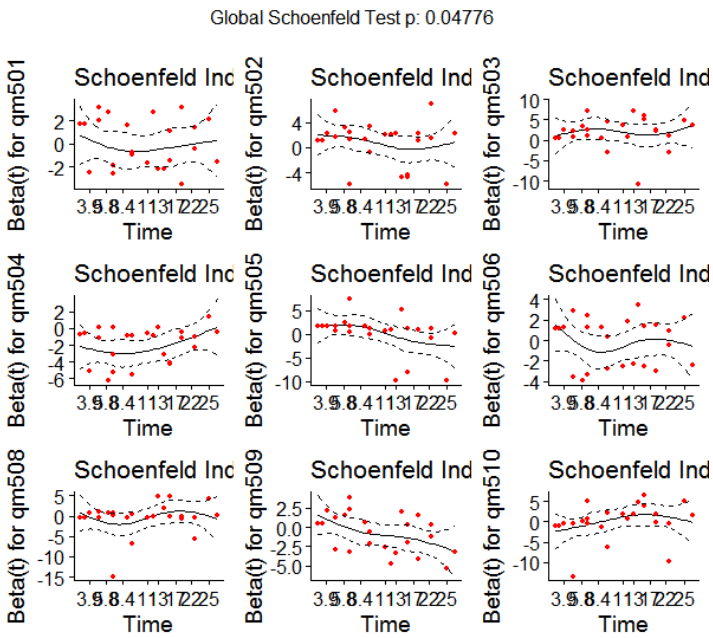
```
        coef exp(coef) se(coef)      z Pr(>|z|)
qm501 -0.1832    0.8326   0.4332 -0.423  0.67238
qm502  0.8938    2.4443   0.5487  1.629  0.10335
qm503  2.0573    7.8246   0.7464  2.756  0.00585 **
qm504 -2.1732    0.1138   0.4446 -4.888 1.02e-06 ***
qm505  0.2707    1.3109   0.6293  0.430  0.66710
qm506 -0.2008    0.8181   0.4303 -0.467  0.64072
qm508 -0.3519    0.7033   0.7491 -0.470  0.63849
qm509 -0.7082    0.4925   0.4437 -1.596  0.11047
qm510  0.1107    1.1171   0.7340  0.151  0.88011
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
      exp(coef) exp(-coef) lower .95 upper .95
qm501    0.8326    1.2010   0.35622    1.9461
qm502    2.4443    0.4091   0.83386    7.1651
qm503    7.8246    0.1278   1.81180   33.7923
qm504    0.1138    8.7863   0.04761    0.2721
qm505    1.3109    0.7629   0.38185    4.5001
qm506    0.8181    1.2224   0.35201    1.9012
qm508    0.7033    1.4218   0.16200    3.0535
qm509    0.4925    2.0304   0.20641    1.1752
qm510    1.1171    0.8952   0.26502    4.7085

Concordance= 0.788  (se = 0.043 )
Likelihood ratio test= 33.71  on 9 df,   p=1e-04
Wald test            = 36.22  on 9 df,   p=4e-05
Score (logrank) test = 37.49  on 9 df,   p=2e-05
```

Testing proportional Hazards assumption:

```
        rho   chisq      p
qm501 -0.0310 0.02291 0.87969
qm502 -0.2159 1.65367 0.19846
qm503  0.0127 0.00374 0.95122
qm504  0.3000 2.24812 0.13378
qm505 -0.4213 6.36991 0.01161
qm506 -0.0676 0.12877 0.71971
qm508  0.1412 0.50605 0.47685
qm509 -0.4719 6.84962 0.00887
qm510  0.2541 2.20108 0.13791
GLOBAL     NA 17.06153 0.04776
```



Berkeley
UNIVERSITY OF CALIFORNIA