Sociological Statistical Inference by Survival Analysis

C240B / STAT C245B Project Progress Report

Z. Tom Hu

Sijin Wu

Berkeley
UNIVERSITY OF CALIFORNIA

Index

1. Explanatory Data Analysis (EDA)

2. Feature Selection

3. Some results on K-M and COX estimators

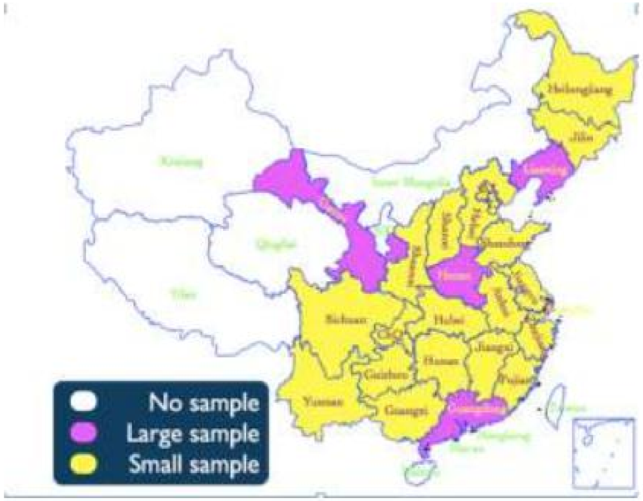
4. Future works

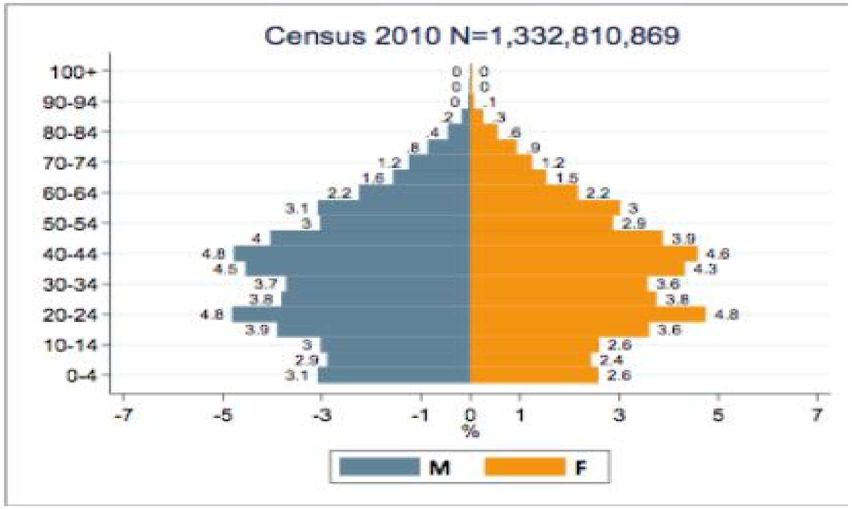
Berkeley
UNIVERSITY OF CALIFORNIA

1. EDA

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

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





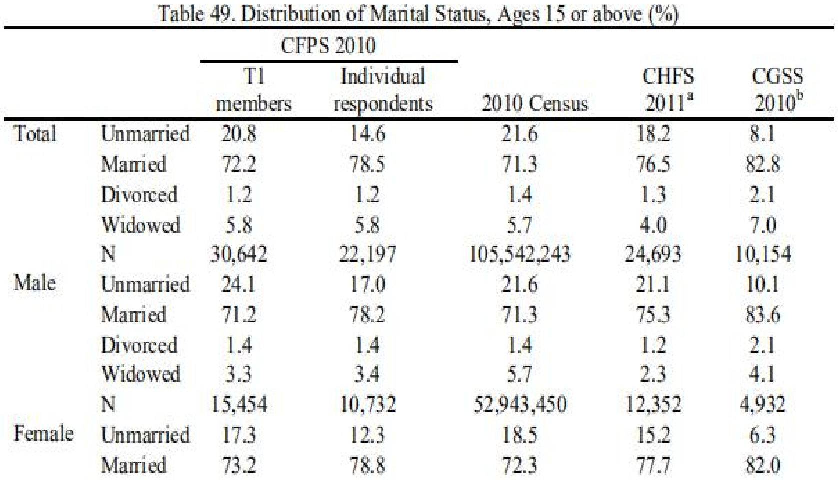


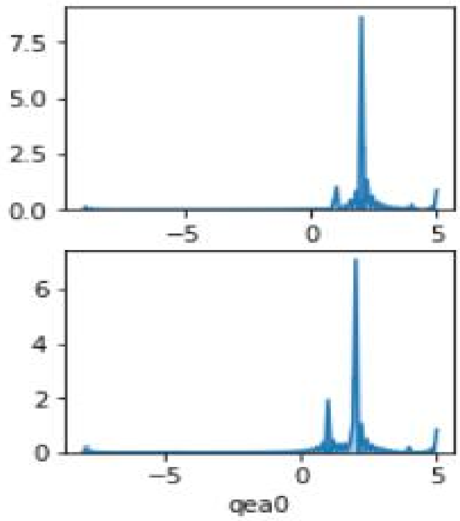
Table 49. Distribution of Marital Status, Ages 15 or above (%)

		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

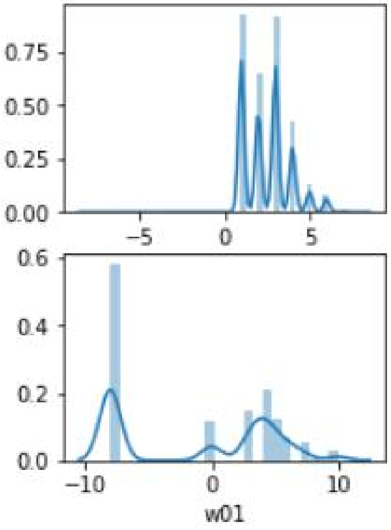
Berkeley
UNIVERSITY OF CALIFORNIA

1. EDA

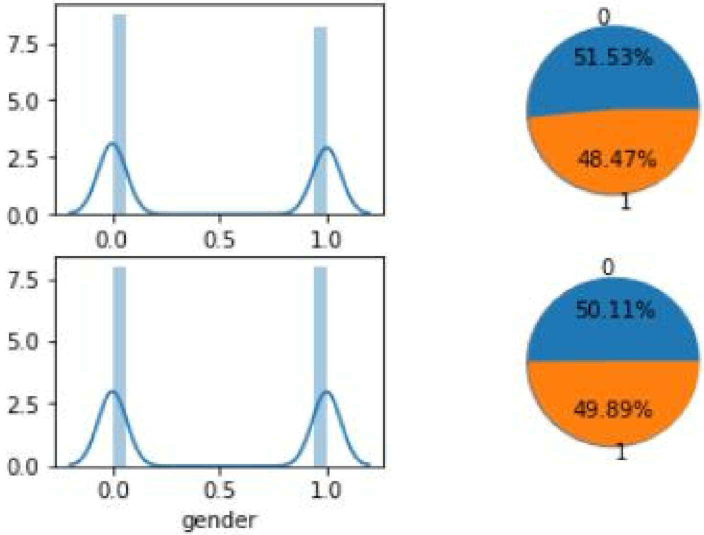
Comparing the 2010’s and 2018’s sample, we can see that the distribution of marriage status and gender are mostly stable, while the distribution of education has a clear shift.



1.marriage



2.education



3.gender

1. EDA

Note that according to the stucture of questionaires, we can also consider data sets as a non-longitudinal data --- we can assume that year 2010 and year 2018 can represent all marriage information.

Next, our idea is to find proper covariates that can be modeled.

2 Feature selection

Note that we have more than 1000 covariates. It will be problematic if we throw all of them into a causal model as it will highly likely result in interpretation error of the data. So we need to do some feature selections.

First, we can mine useful features from initial data sets by hands based on our intuition and experience. Then, we can apply some machine learning tools to further explore meaningful features.

2 Feature selection

1 Compute the marriage time by people who divorced. This data set becomes the training set.

2 There are some possible variables like age, birthplace, job, education status that may effect the marriage time. We select them manually and compute the correlation effecients.

3 Alternative variables :

w01: highest degree

egc1011: monthly after-tax income (yuan/month)

qg401: job income satisfaction qg12 total income

qn8011: income in the local state

qkz204: appearance

....

2 Feature selection

According to the correlation coefficient matrix, the relationship between marital longevity and these variables is not very strong.

Then, we focused on the intrinsic information of the sample, such as personality, appearance, intelligence, outlook, age difference and other potential factors.

Date of birth: eeb402y

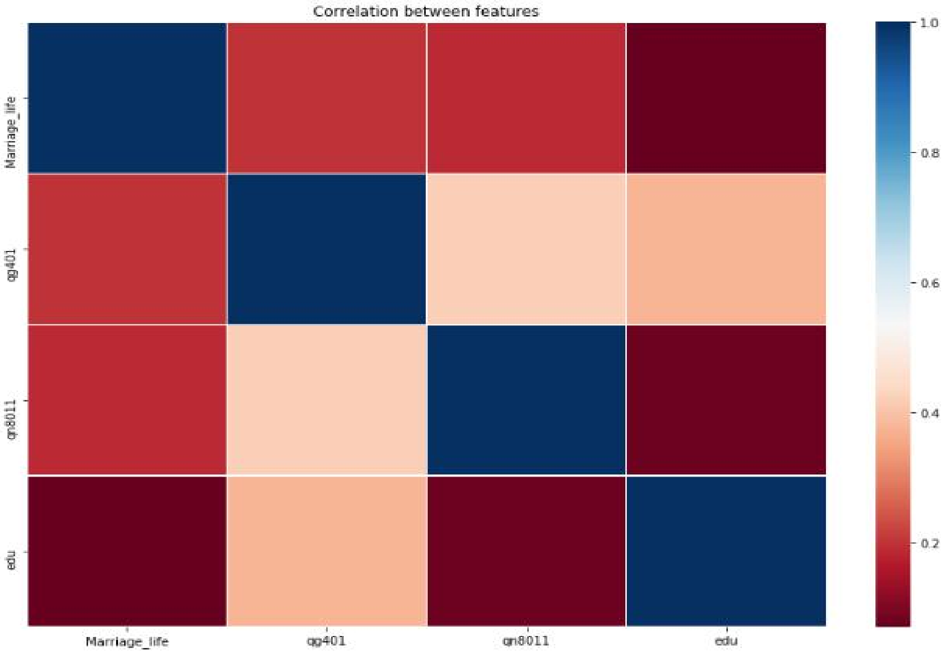
His/her highest education: eeb4021_a_1

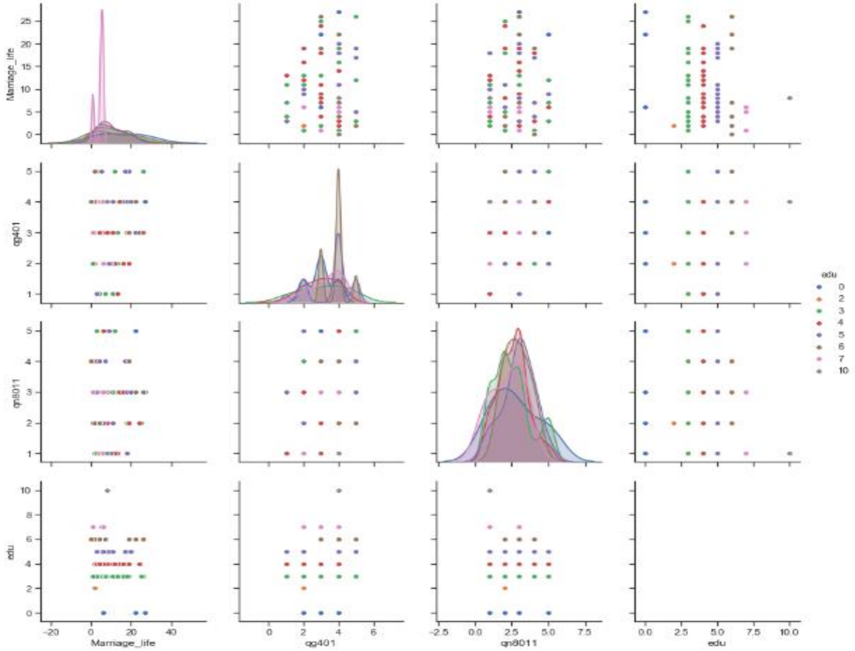
Premarital cohabitation: eeb404_a_1

Premarital cohabitation months: qeb405_1_a_1

Intelligence level: kz207

Personality: wm

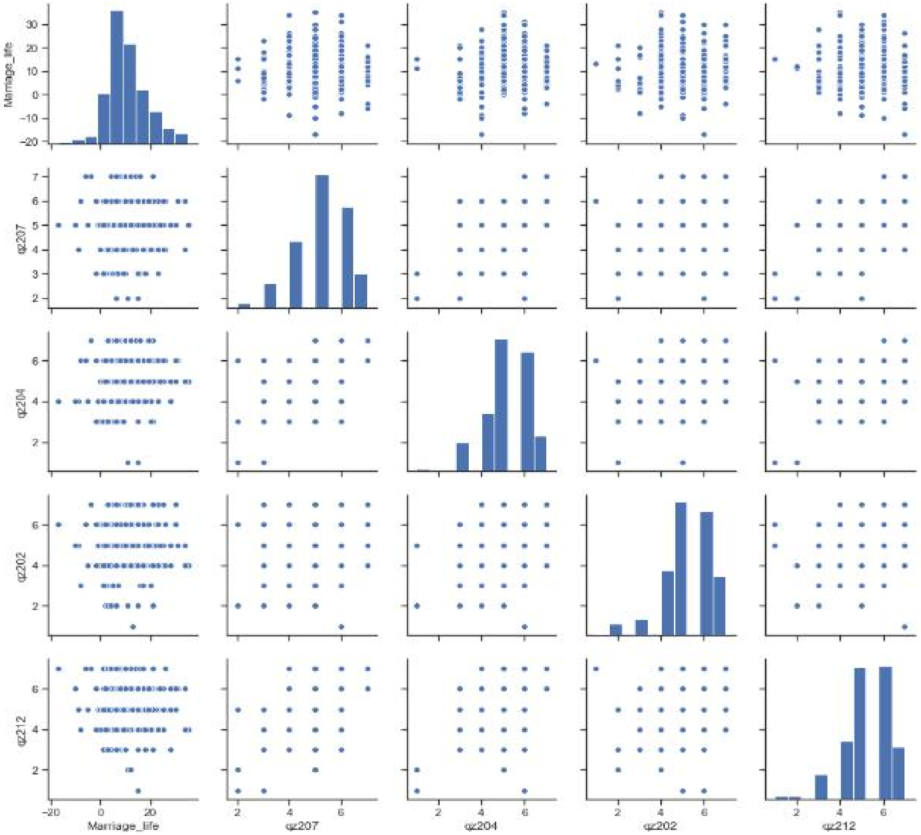


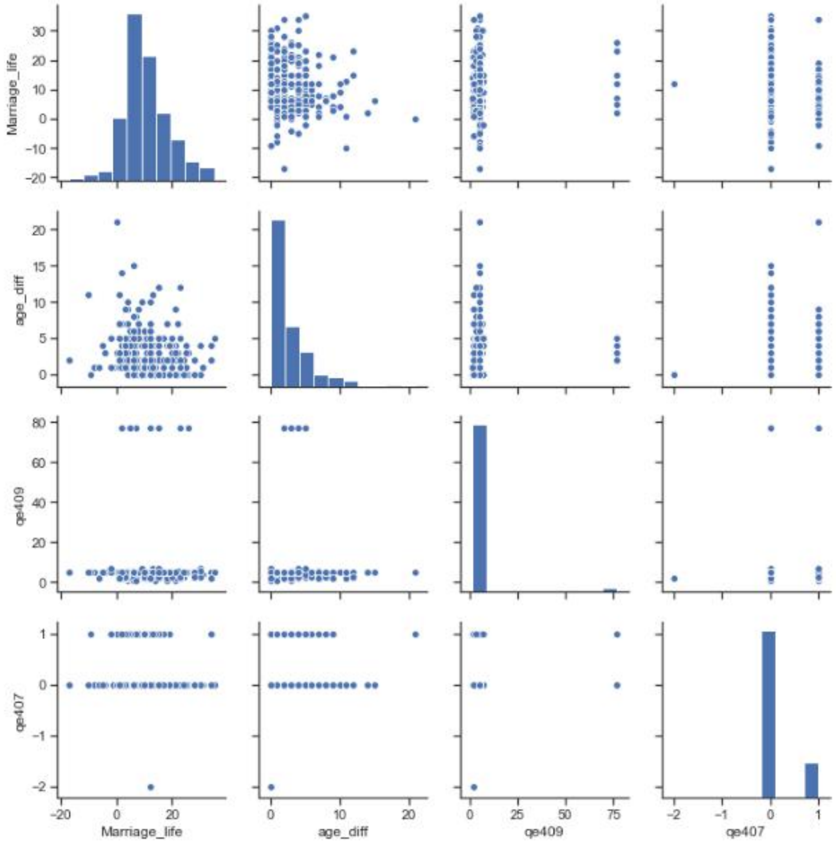


2 Feature selection

The results showed that there was no high correlation coefficient.

We also tried to get the “matching dgree” from the training data sets, like the character difference computed by the Psychological scale or attitudes toward different opnions. However, there are only 2 couples in the dataset.





2 Feature selection

Feature selector (python package) is written by William Koehrsen, a data scientist in feature Labs. Feature selector mainly selects the following types of features:

Set the marriage time as labels.

(1) features with high missing-values percentage

(2) features with high correlation

(3) features have no contribution to model prediction results (namely zero importance).

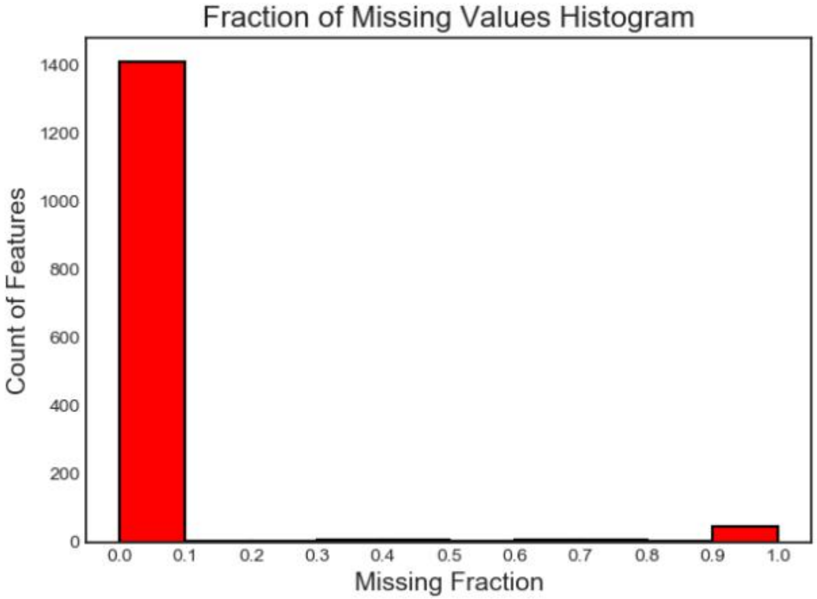
(4) features that make only a small contribution to the prediction results of the model (i.e., low importance)

(5) features with a single value feature (i.e., there is only one element in the set of values of the feature in the data set).

2 Feature selection

(1) Identify_missing

This method is used to select a feature where the missing value percentage is greater than the specified value (specified by missing_threshold). The method can be applied to feature selection of supervised learning and unsupervised learning.



2 Feature selection

(2) Identify_collinear

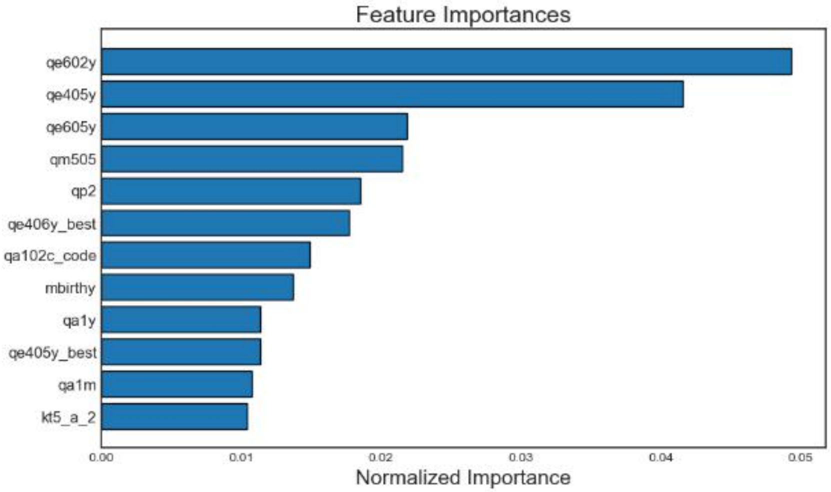
This method is used to select features whose correlation is greater than the specified value (through correlation_threshold).

2 Feature selection

(3) Identify_zero_importance

This method is used to select features that do not contribute to the prediction results of the model .

Feature selector trains a Gradient Boosting machine (GBM) with the data set, and then gets the importance score of each feature by GBM, normalizes the importance score of all features, and selects the feature whose importance score is equal to zero.



2 Feature selection

Meaningless variables: qe602 qe405y qe605y qe406y_best qe405y_best (because these variables are used to calculate the marriage life)

Potentially significant variables:

Qm505: the importance of avoiding loneliness

qp2: the current weight

qa102c_code: the area code of the birth location

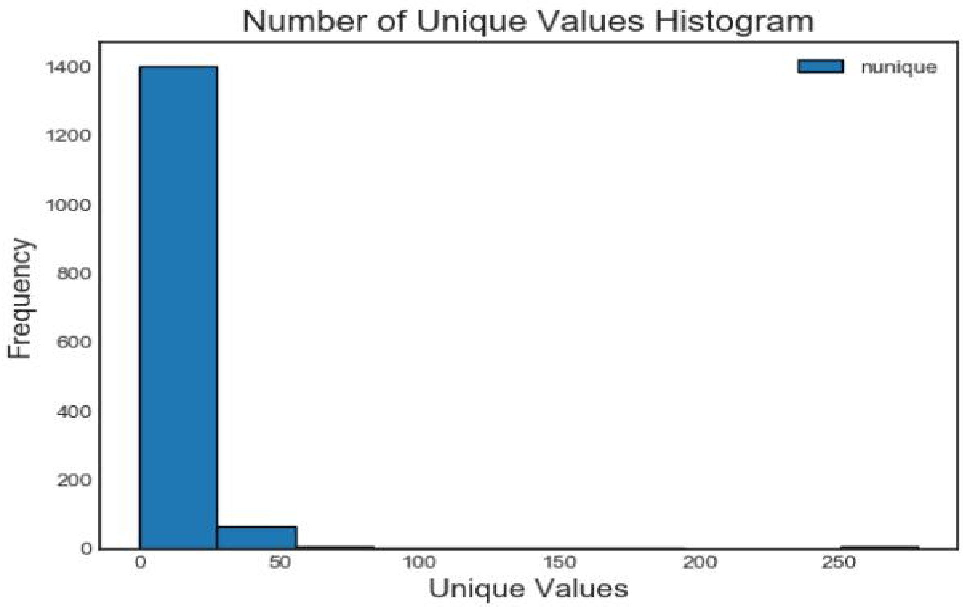
mbirth: the birth year of the mother

qa1y (m) : birth year (month)

2 Feature selection

(4) Identify_single_unique

This method is used to select a feature with only a single value. The variance of a feature with a single value is 0, which will have no effect on the training of the model (from the perspective of information entropy, the entropy of the feature is 0).



2 Feature selection

In conclusion, first we construct some new features using original data according to our experience like age difference, character score. Then, we used ML tools to reomove 1277 features. Combing them together, we decide variables for future modeling.

3. Some results on K-M and COX

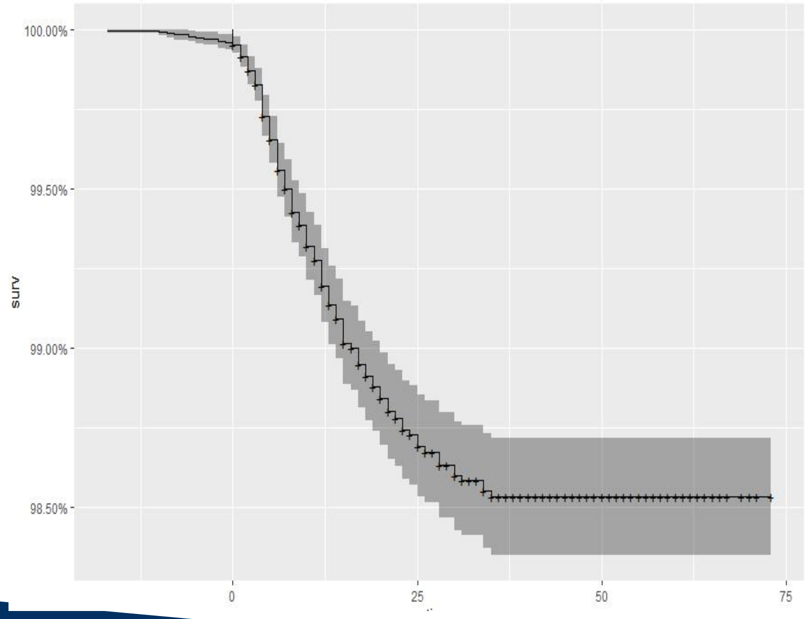
1. Transform the data sets into survival data, which means we introduce the censored data. Until 2010, people who were marriaged represent the right-censored data. And we use divorced data to get the complete time data.

3. Some results about K-M and COX

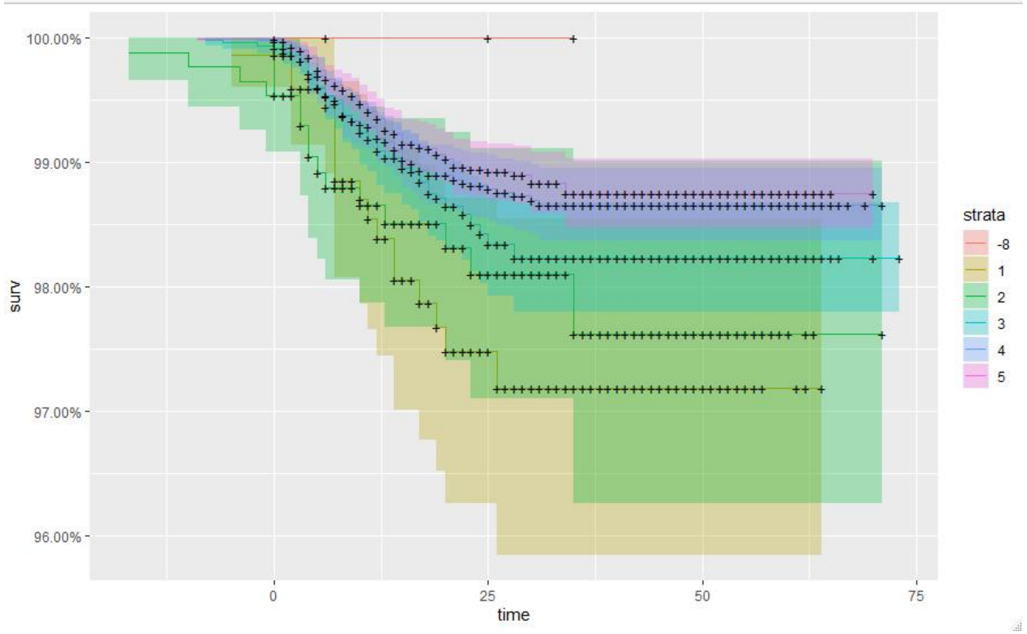
2. Try to use K-M estimator for taking treatment = qm505 which has the highest importance in ML features selection process and = age_difference artificially constructed to see the difference.

(R package didn't work in this situation.)

overall surviaval

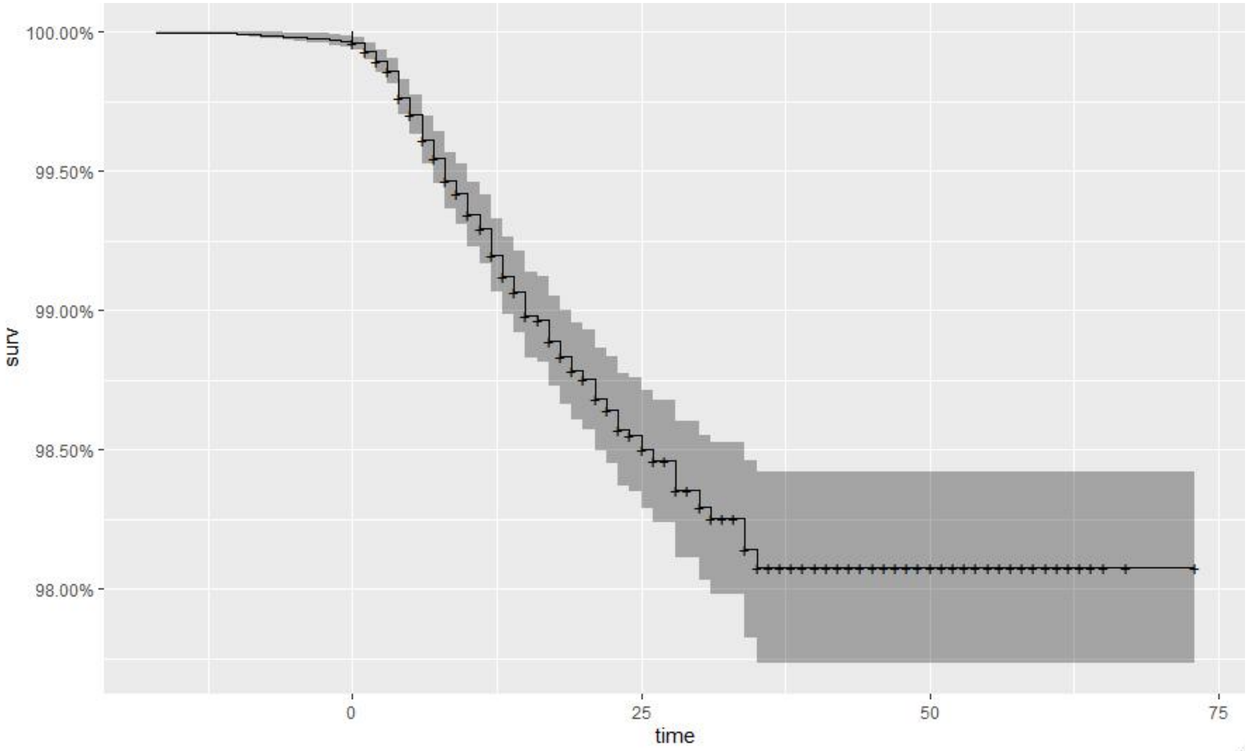


treatment=qm505



3. Some results about K-M and COX

3. Run COX for selected variables: qm505 + qp2 + qa1y +mbirthy. These variables are selected by their high feature importance.



4. Future work

1 Based on the feature selection, use selected variables to check the univariate model and Cox.

2 Change the marriage_time from continuous variable into classification variable, then use

$$\Psi\left(P_0\right) = E_{W,0}\left[E_0(Y|A=1,W) - E_0(Y|A=0,W)\right]$$

as targted parameter to apply super learner and TMLE. (Maybe we will do a comparision between targted learning and traditional Ml.

3 Study more about the model inferencce.