April 2021

Module 6 Assignment: Time Series and Survival Analysis

Survival Analysis on Echocardiogram heart attack data

This dataset consists of 132 instances of patients for 12 variables describing the patient's heart attack and condition. (https://archive.ics.uci.edu/ml/datasets/echocardiogram)

MAIN GOAL

In this work we want to predict **survival month** ("survival") based on the covariates in **Echocardiogram - UCI data**. This data includes **censored** data and we will perform a **Kaplan-Meier estimate**, **Cox proportional hazards model** and The Weibull AFT model with single and multiple variables.

Data Description

1. **survival** -- the number of months patient survived (has survived, if patient is still alive).
2. **still-alive** -- **a binary variable**. 0=dead at end of survival period, 1 means still alive
3. **age-at-heart-attack** -- age in years when heart attack occurred
4. **pericardial-effusion -- binary**. Pericardial effusion is fluid around the heart. 0=no fluid, 1=fluid
5. **fractional-shortening** -- a measure of contractility around the heart lower numbers are increasingly abnormal
6. **epss** -- E-point septal separation, another measure of contractility. Larger numbers are increasingly abnormal.
7. **lvdd** -- left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts.
8. **wall-motion-score** -- a measure of how the segments of the left ventricle are moving
9. **wall-motion-index** -- equals wall-motion-score divided by number of segments seen. Usually 12-13 segments are seen in an echocardiogram.
10. **mult** -- a derivate var which can be ignored
11. **name** -- the name of the patient, replaced with "name"
12. **group** -- meaningless, we can ignore it
13. **alive-at-1** -- Boolean-valued. Derived from the first two attributes. 0 means patient was either dead after 1 year or had been followed for less than 1 year. 1 means patient was alive at 1 year.

Data Shape and Missing Values

	survival	alive	age	pericardialeffusion	fractionalshortening	epss	lvdd	wallmotion-score	wallmotion-index	mult	name	group	aliveat1
0	11.0	0.0	71.0	0.0	0.260	9.000	4.600	14.0	1.00	1.000	name	1	0.0
1	19.0	0.0	72.0	0.0	0.380	6.000	4.100	14.0	1.70	0.588	name	1	0.0
2	16.0	0.0	55.0	0.0	0.260	4.000	3.420	14.0	1.00	1.000	name	1	0.0
3	57.0	0.0	60.0	0.0	0.253	12.062	4.603	16.0	1.45	0.788	name	1	0.0
4	19.0	1.0	57.0	0.0	0.160	22.000	5.750	18.0	2.25	0.571	name	1	0.0

```
print(df.isnull().sum())
print(df.shape)
```

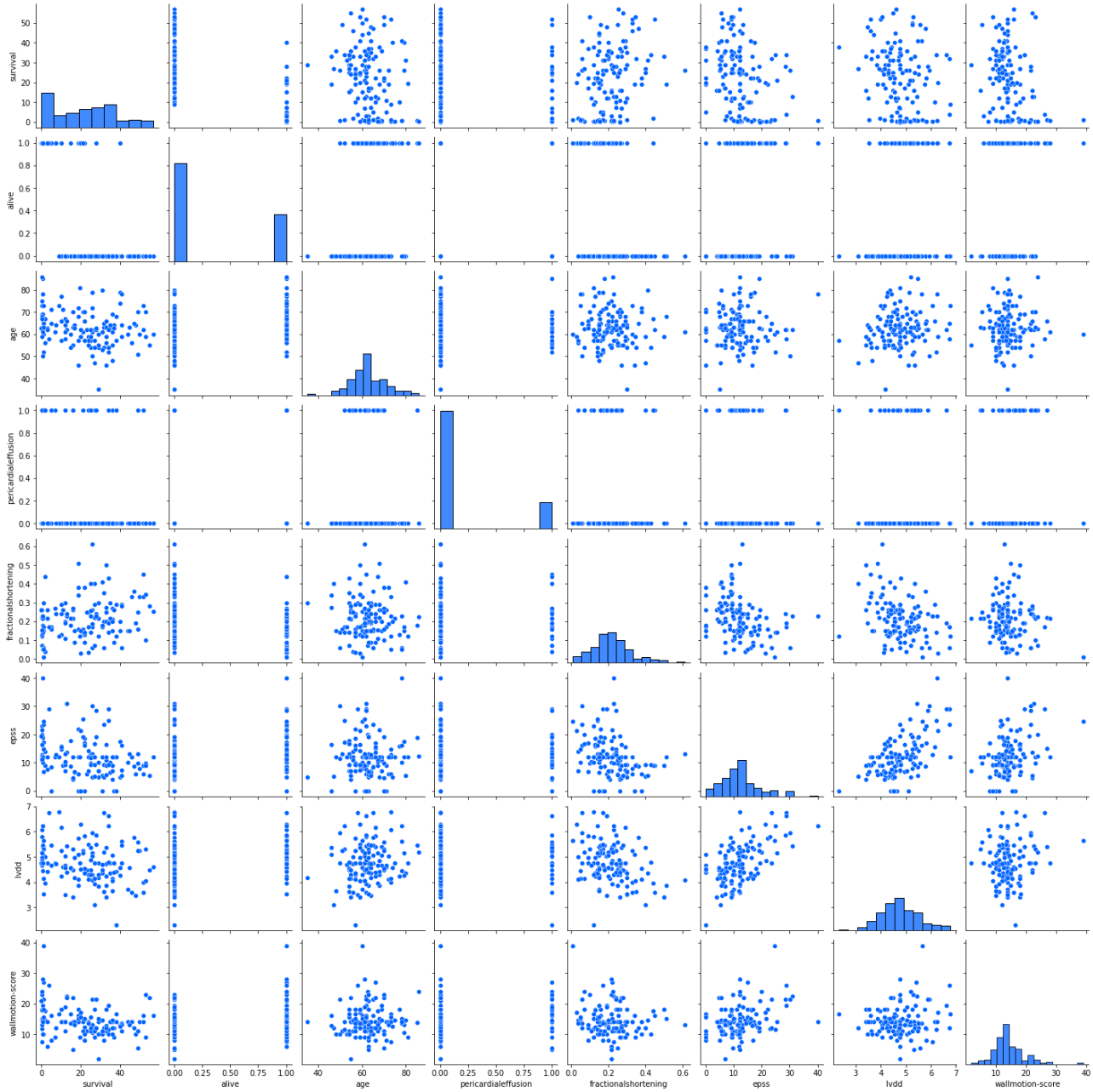
```
survival                   3
alive                      2
age                        7
pericardialeffusion        1
fractionalshortening       9
epss                      16
lvdd                      12
wallmotion-score           5
wallmotion-index           3
mult                       4
name                       2
group                     23
aliveat1                  58
dtype: int64
(133, 13)
```

We have only 133 observations and a few missing values across variables. We will replace them with mean value for each column and we will drop the not relevant features

```
df = pd.concat([df_keep, df_X], axis = 1)
df = df.dropna()
print(df.isnull().sum())
print(df.shape)
```

```
survival                  0
alive                     0
age                       0
pericardialeffusion       0
fractionalshortening      0
epss                      0
lvdd                      0
wallmotion-score          0
dtype: int64
(130, 8)
```

Scatter plots between survival and covariates



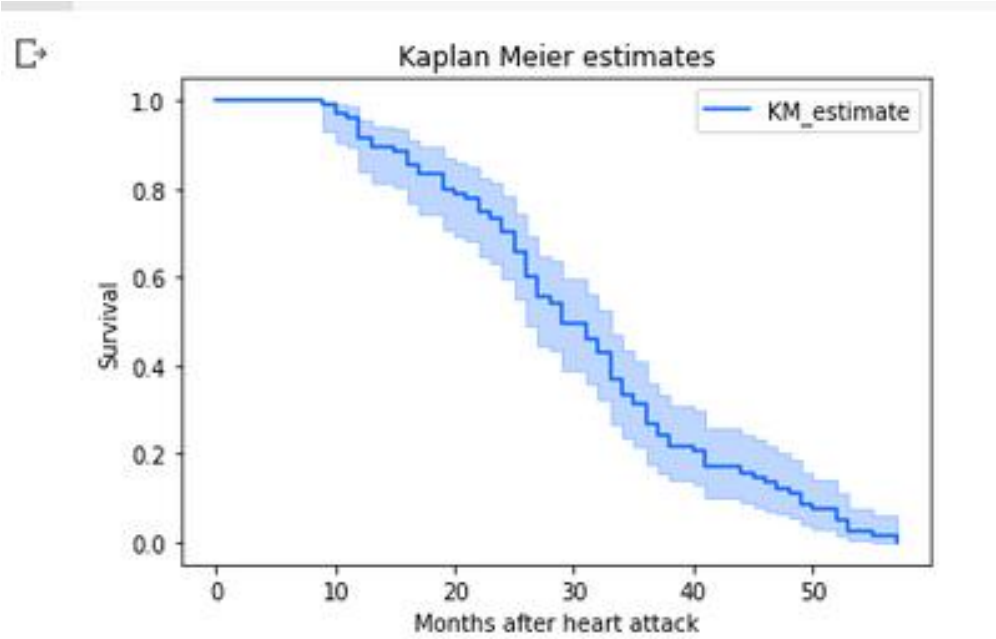
Check censored data

We have censored data, because for alive (=1) patients during data collection period, we do not know their survival months after the data collection. Hence, the following analysis needs to consider the censored data creating dead variable below: with 42 Censored data and 88 non censored data:

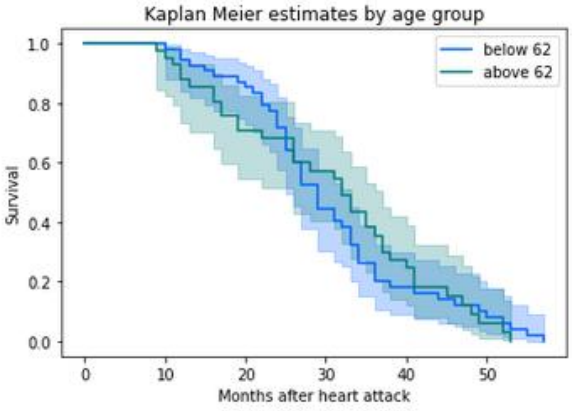
```
[9] df.loc[df.alive == 1, 'dead'] = 0
    df.loc[df.alive == 0, 'dead'] = 1
    df.groupby('dead').count()
```

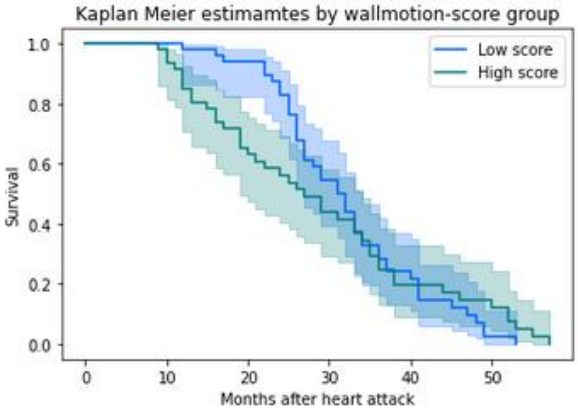
dead	survival	alive	age	pericardialeffusion	fractionalshortening	epss	lvdd	wallmotion-score
0.0	42	42	42	42	42	42	42	42
1.0	88	88	88	88	88	88	88	88

Kaplan Meier estimate



And let's see also how other variables influence the survival rate, I've found interesting Age and Wallmotion-score:



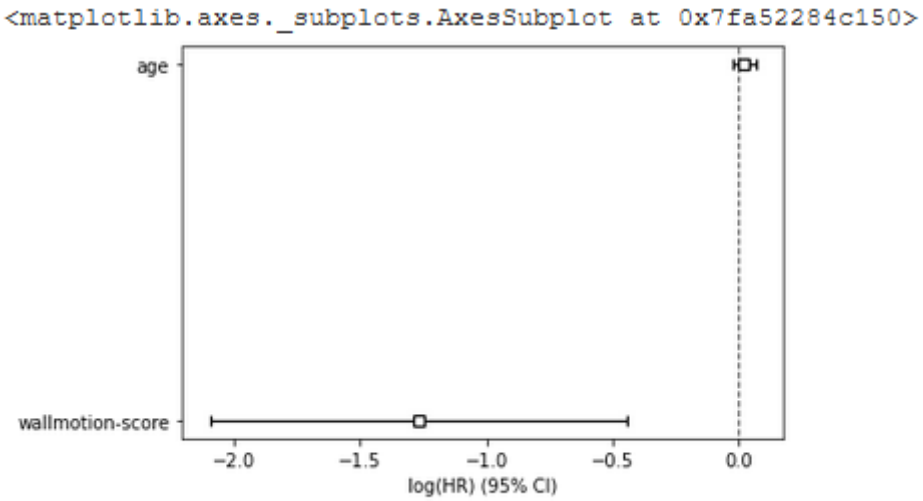


The **first** picture is self-explaining, from the **second** picture we can see there's **an interesting point of "inversion**" around the 28th month after Heart Attack where "above 62" is more likely to survive than "below 62" until the 40th month where they have more or less the same rate, the **third** picture shows a **significant difference** in survival time by wallmotion score (a measure of how the segments of the left ventricle are moving, lower is better) group for the first 2 years after heart attack.

Having looked at our data and related Kaplan-Meier curves, we can formalize the analysis by running survival regression. the Cox model estimates a baseline hazard rate, and assumes features impact this hazard rate proportionally and The Weibull AFT model

Cox proportional hazards model

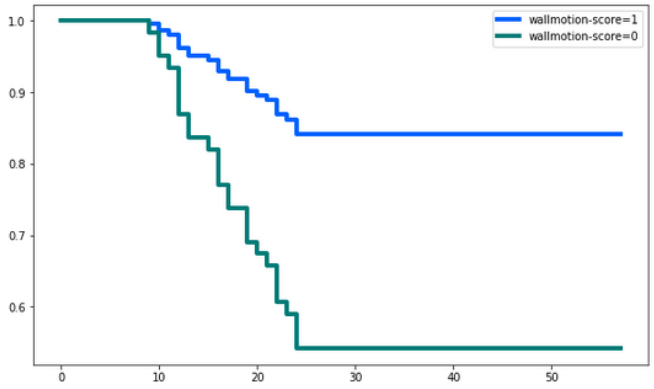
```
<lifelines.CoxPHFitter: fitted with 130 total observations, 102 right-censored observations>
             duration col = 'survival'
                event col = 'censored'
      baseline estimation = breslow
   number of observations = 130
number of events observed = 28
   partial log-likelihood = -117.36
         time fit was run = 2021-04-18 22:29:49 UTC

---
                    coef  exp(coef)   se(coef)   coef lower 95%   coef upper 95%  exp(coef) lower 95%  exp(coef) upper 95%
covariate
age                 0.02       1.02       0.02            -0.03             0.07                 0.97                 1.07
wallmotion-score   -1.27       0.28       0.42            -2.09            -0.45                 0.12                 0.64

                      z      p   -log2(p)
covariate
age                0.88   0.38       1.39
wallmotion-score -3.02 <0.005       8.63
---
Concordance = 0.70
Partial AIC = 238.71
log-likelihood ratio test = 10.68 on 2 df
-log2(p) of ll-ratio test = 7.71
```



And for the statistically significant variable :

```
cph.plot_partial_effects_on_outcome('wallmotion-score', [1,0], plot_base
```

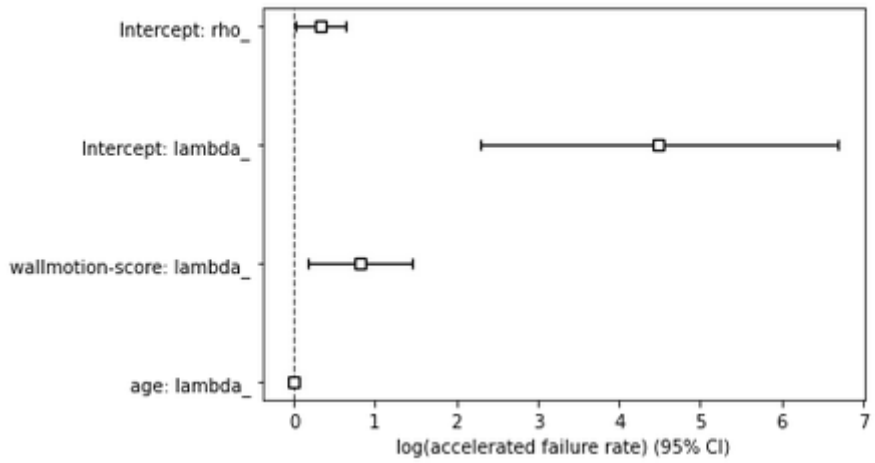
```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa5224d95d0>
```



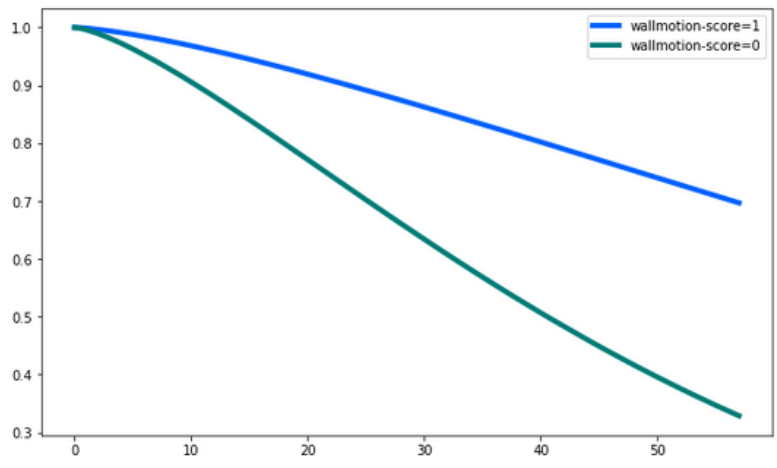
The Weibull AFT model

```
<lifelines.WeibullAFTFitter: fitted with 130 total observations, 102 right-censored observations>
             duration col = 'survival'
                event col = 'censored'
   number of observations = 130
number of events observed = 28
           log-likelihood = -151.86
         time fit was run = 2021-04-18 22:53:47 UTC

---
                            coef  exp(coef)   se(coef)   coef lower 95%   coef upper 95%  exp(coef) lower 95%  exp(coef) upper 95%
param   covariate
lambda_ age                -0.01       0.99       0.02            -0.04             0.03                 0.96                1.03
        wallmotion-score    0.81       2.24       0.33             0.17             1.45                 1.18                4.26
        Intercept           4.49      89.28       1.12             2.30             6.68                 9.96              799.87
rho_    Intercept           0.33       1.39       0.16             0.02             0.64                 1.02                1.90

                              z      p   -log2(p)
param   covariate
lambda_ age               -0.48   0.63       0.66
        wallmotion-score   2.47   0.01       6.22
        Intercept          4.02 <0.005      14.04
rho_    Intercept          2.08   0.04       4.75
---
Concordance = 0.70
AIC = 311.72
log-likelihood ratio test = 8.21 on 2 df
-log2(p) of ll-ratio test = 5.92
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa524fa9c90>
```




```
<matplotlib.axes._subplots.AxesSubplot at 0x7fa522478450>
```



Conclusions

We've tested 3 Models :

1)Kaplan Maier (a statistical overview)

2)Cox proportional hazards model

3)The Weibull AFT model

Both models are valid as we can see from their statistic summary. The statistics follow the Chi-square distribution with 2 degree of freedom. We can base our statistical inference on both models. Kaplan Maier is a **good starting point** as its overview is confirmed by both models.

From **both models** we can see that **Wallmotion-score group is a risk factor for survival time,** at the contrary **Age** (even if it shows interesting behaviors around the 28^{th} month) doesn't pass the p-value test in both the models.

Negative sign of wallmotion-score variable (-1.28 in CPH) shows that the patients with low wallmotion score **reduce the risk of death**. Hazard ratio of wallmotion-score is 0.28, which means it **reduces hazard risk** (since it's less than 1) by **72%**. Same conclusions from AFT model can be inferred.

To conclude we can say that for the first 2 years after heart attack, **people** with **Higher Wallmotion score** could have a Higher Risk of death, hence **we should focus clinical cares** on this group of patients.

Next Steps and Issues

The main issue here it seems the limited size of the data set, the problem addressed by past researchers was to predict from the other variables whether the patient will survive at least one year. The most difficult part of this problem is correctly predicting that the patient will NOT survive, but we focused on a Survival Analysis, and we gave an interesting insight on some variables effect on Survival Rate.

It could be interesting to explore the data set with multiple supervised and unsupervised learning techniques for predictive purposes.