Survival Analysis:

Data Analysis for the Mayo Clinic Primary Biliary Cirrhosis Data

Author: Mathieu Bernardi Date: 17/04/2019

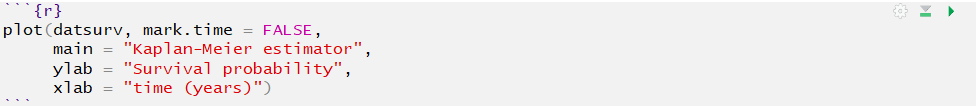
**Description**

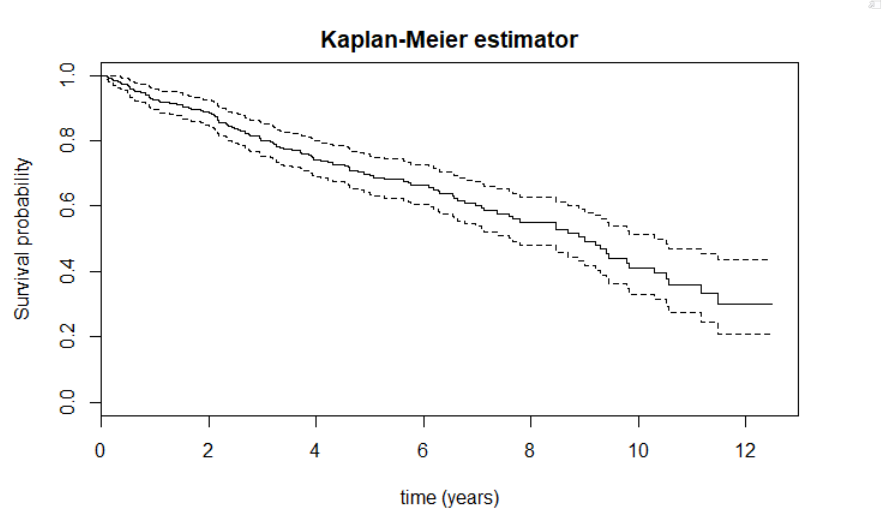
D This data is from the Mayo Clinic trial in primary biliary cirrhosis (PBC) of the liver conducted between 1974 and 1984. A total of 424 PBC patients, referred to Mayo Clinic during that ten-year interval, met eligibility criteria for the randomized placebo controlled trial of the drug D-penicillamine. The first 312 cases in the data set participated in the randomized trial and contain largely complete data. The additional 112 cases did not participate in the clinical trial, but consented to have basic measurements recorded and to be followed for survival. Six of those cases were lost to follow-up shortly after diagnosis, so the data here are on an additional 106 cases as well as the 312 randomized participants.

A nearly identical data set found in appendix D of Fleming and Harrington; this version has fewer missing values.

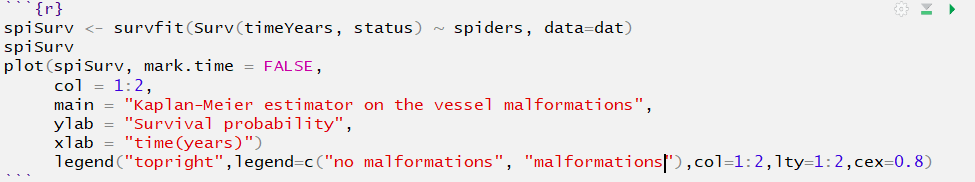
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Format**   |  |  | | --- | --- | | age: | in years | | albumin: | serum albumin (g/dl) | | alk.phos: | alkaline phosphotase (U/liter) | | ascites: | presence of ascites | | ast: | aspartate aminotransferase, once called SGOT (U/ml) | | bili: | serum bilirunbin (mg/dl) | | chol: | serum cholesterol (mg/dl) | | copper: | urine copper (ug/day) | | edema: | 0 no edema, 0.5 untreated or successfully treated | |  | 1 edema despite diuretic therapy | | hepato: | presence of hepatomegaly or enlarged liver | | id: | case number | | platelet: | platelet count | | protime: | standardised blood clotting time | | sex: | m/f | | spiders: | blood vessel malformations in the skin | | stage: | histologic stage of disease (needs biopsy) | | status: | status at endpoint, 0/1/2 for censored, transplant, dead | | time: | number of days between registration and the earlier of death, | |  | transplantion, or study analysis in July, 1986 | | trt: | 1/2/NA for D-penicillmain, placebo, not randomised | | trig: | triglycerides (mg/dl) | |  |
|  |  |
| **Data set PBC** |  |
| On the data set pbc we have 418 observations and 20 variables, with some missing values. |  |
| The data set contains the main variable time, which is the observed, potentially censored survival time and the variable status, which indicates if the observation has been censored.  **preprocessing on the data**  For my analysis I will remove from my data set all the rows with missing data "NA’s".  I exclude the 25 patients who have been transplanted.  I create a variable timeYears corresponding to the variable time in years.      After the preprocessing, we have 258 observations with 21 variables. |  |
|  |  |
| In the status variable, there are 258 patients, and the variable shows that 111 have died and 147 have been censored. The data set is right-censored.  **Estimation of the survival function using the Kaplan-Meier estimator** |  |

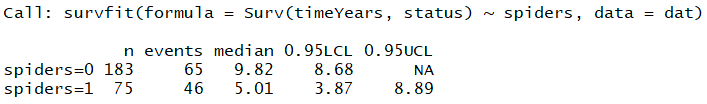
The survival median for the patients is of 8.99 years, with a confidence interval 7.66 and 10.31.



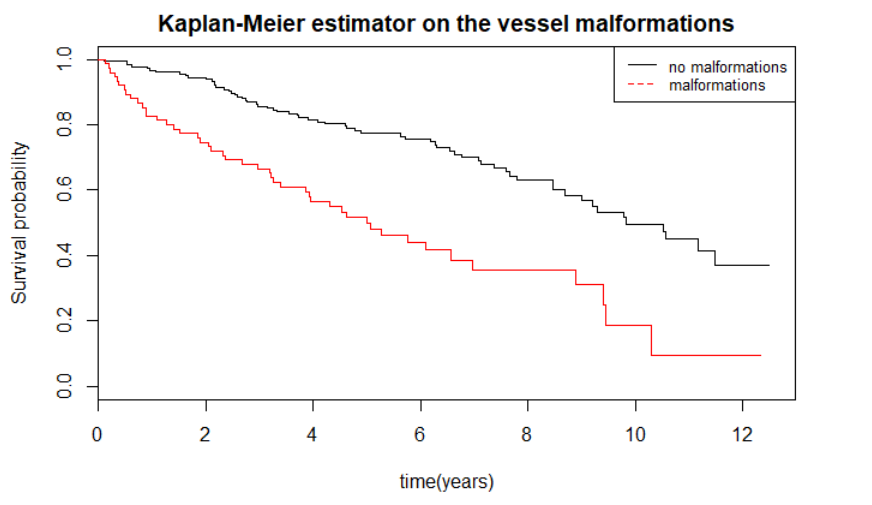


**Does the malformation of the blood vessels in the skin play a role in patient’s survival?**



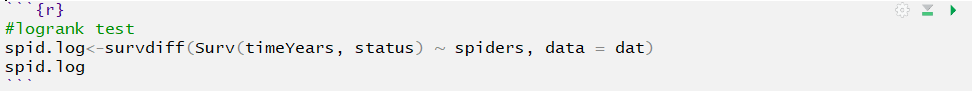


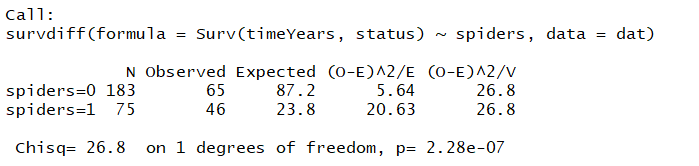
The survival median for the patients with no blood vessel malformations is of 9.82 years and the patients who have some blood vessel malformations is of 5.01 years.



I use the Kaplan-Meier estimator to estimate the survival curves, we can see that the two curves are almost parallel. We can already see that the absence of the blood vessels malformations in the skin improves the survival of the patients.

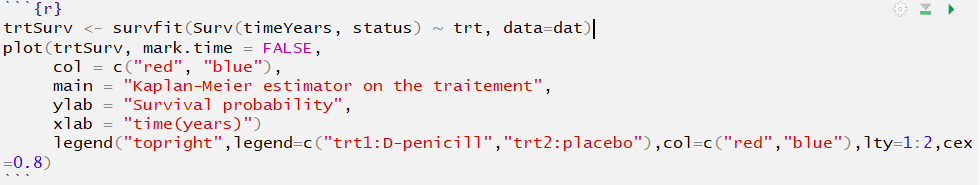
we will be able to use the logrank test to verify the hypothesis with the function survdiff.

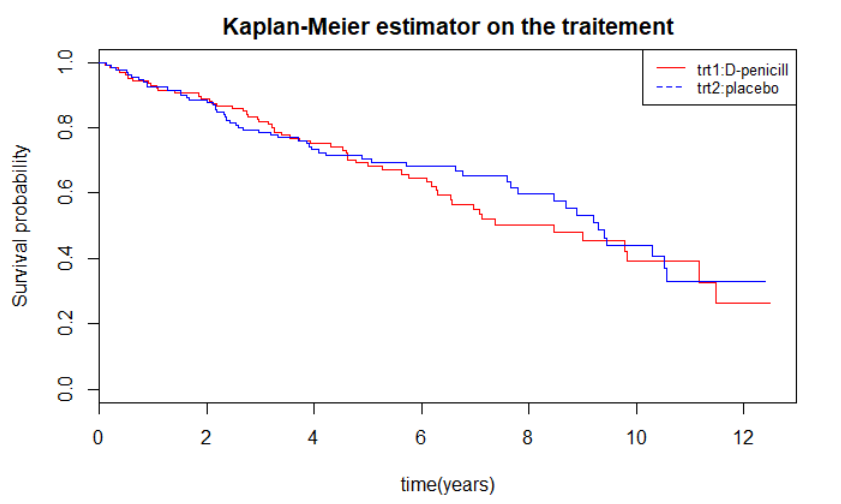




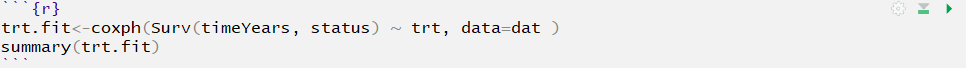
The p value is very small, so we accept the hypothesis that if there are some blood vessel malformations in the skin, there is a significant impact on the survival of patients.

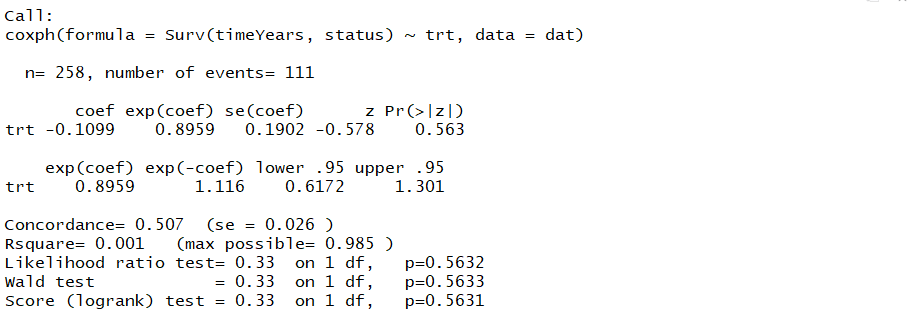
**Does treatment improve patient survival?**

****

****

I use the Kaplan-Meier estimator to estimate the survival curves of the patients, the two curves for the treatments are almost identical, I verify the hypothesis using the cox regression model.





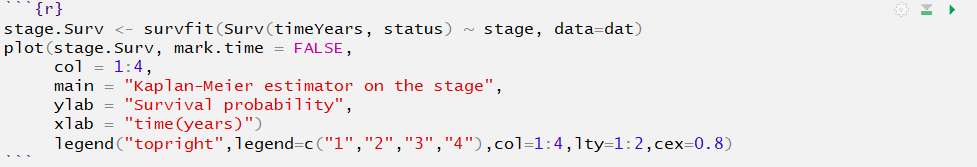
The P value is high p = 0.563, the beta is negative -0.11, the confidence interval is small 0.61 and 1.3, so we reject the hypothesis, we can say that the D-penicill has no effect on the patient’s survival.

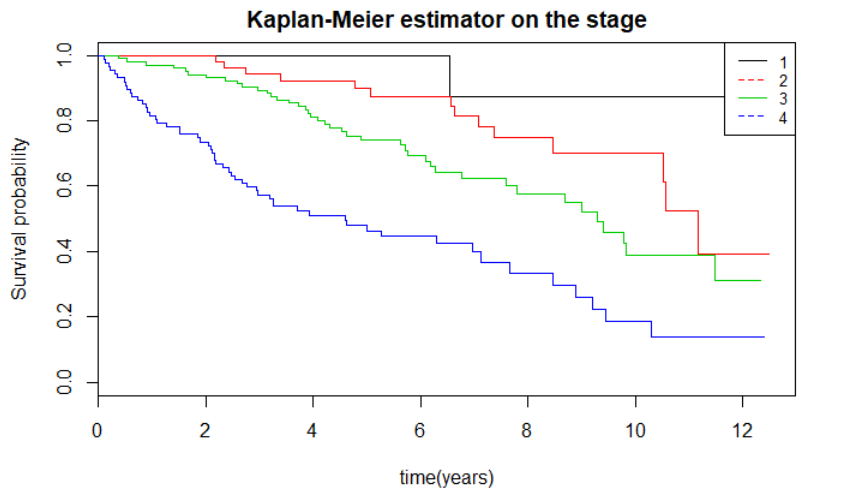
**Does the histologic stage disease play a role in patients’ survival?**

****

****

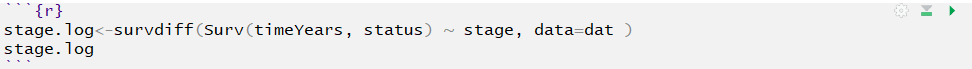
We can see that 12 patients are at stage 1, 56 patients are stage 2, 103 patients are stage 3 and 87 patients are stage 4.

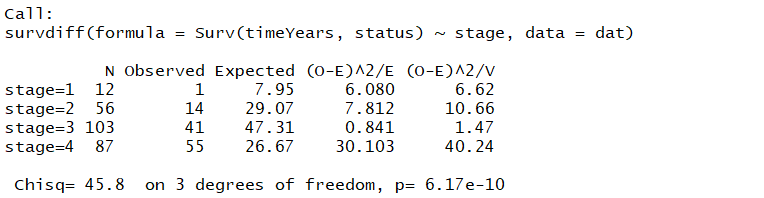
****



I use the Kaplan-Meier estimator to estimate the survival curves for the four histologic stage.

I use the logrank test to verify the hypothesis with the function survdiff.

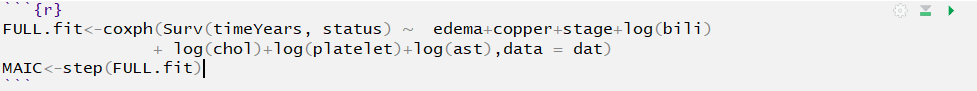




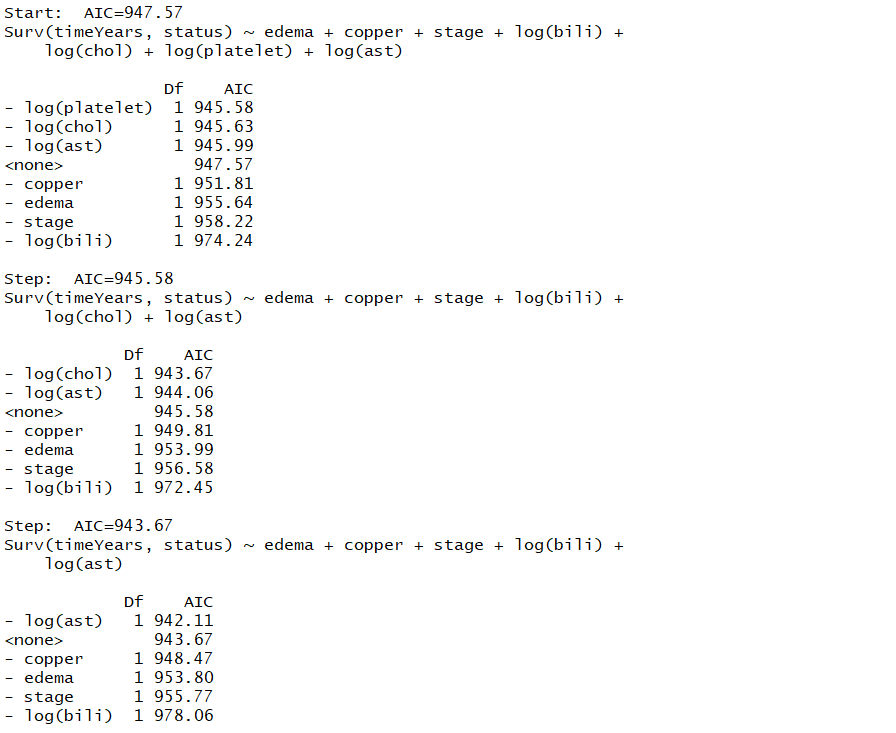
The P value is very low, so we accept the hypothesis, if the histological stage of the patient is higher, the patient has a lower chance of survival.

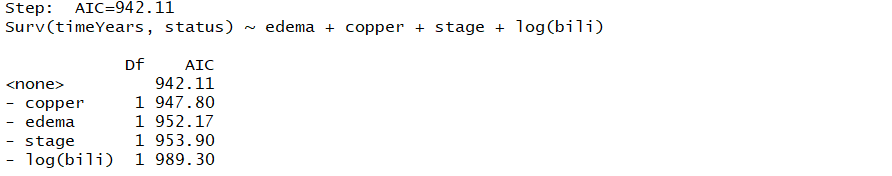
**Stepwise model selection based on AIC**

**What it is the best covariate for the cox regression model?**

****

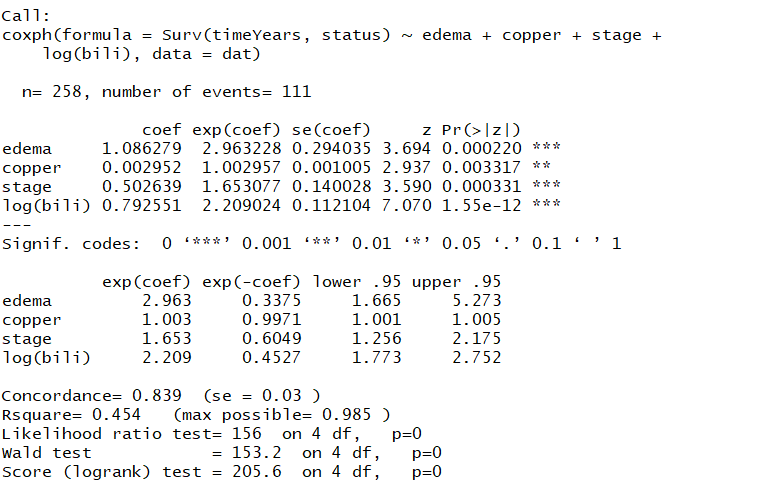
In this example I will use 7 covariates on the cox regression survival model. Thanks to the automatic selection model based on the AIC, I will be able to define the best model with the best covariates to study.

****

****

The best covariates to study on the 7 are edema, copper, stage and bili.



****

The best variable of our model is bili with a very low Pvalue, the z is high, the risk for the bili variable is 2 times higher for the survival of the patient. The variable stage and edema have a low p-value and a good z, for a high stage the risk is 1.6 times higher for the survival of the patient. For the variable edema the risk is 2.96 times higher for the survival of the patients, the confidence interval is large 1.66 and 5.27. The coper variable have a p-value inferior at 0.05 with a good z, the confidence interval is too short to have an impact on the survival of patient.