

Survival Analysis of TCGA Liver cancer expressed gene transcript

Sailendra Pradhananga (sailendra.pradhananga@scilifelab.se)

Project Report: Algorithms in Bioinformatics

February 15, 2018

1 Introduction

Survival analysis are statistical toolkits that are used to analyzing the outcome of variables based on time period of event occurrence. The event could be either be death, outbreak of disease or failure of automobile parts which are then followed specified time period in days, weeks or years. The major part of the these type of data involves two kinds of variables: first is the time to event and second whether the event have occurred or not. Based on different methods such as non parametric model such Kaplan Meier method or regression model such as cox proportional hazards two function, survival function and hazard function. The survival function gives the probability for each time-points of surviving up to that events while hazard function whether the event has occurred or not given the individual has survived or not. The main aim of these type of analysis is find the relationship between the survival time and variables under study such as drugs, gene expression, treatment either with different covariates such as age, gender, race.

In the current project, we are provided a liver cancer patients of different races dataset survival time and gene expression dataset of various grades. The main aim was to model the data with survival model in order find the potential prognostic marker for survival of patients using Cox Proportional Hazard models. Additionally, using the multiple hypothesis correction we predicted 4698 gene transcripts at 0.05 FDR.

2 Method

2.1 TCGA Dataset

The TCGA dataset for the analysis was download from the (<http://kaell.org/files/survivalLIHC.txt>) site. As already stated in the project guidelines, these are the cancer patient

expression data from the sequenced liver tumor. Additionally there are few other features implication each cancer patient ethnicity, sex and cancer types. These data were first read into python notebook as pandas data frame as done in jupyter notebook. Additional exploratory analysis was done in the notebook itself. The dataset consists of both categorical features and numerical data of gene expression dataset.

2.2 Preprocessing of Dataset

In the survival analysis, the different categorical features were provided of liver cancer patients. These categorical features includes Gender, Race, Stage , Status, Age and LivingDays of different cancer patients. These categorical features were then coded into different numerical codes . Originally, the dataset consists of gene expression profile os 19,571. Furthermore, only those gene expression data were taken which had at least an expression in one of the patients.

2.3 Cox proportional Hazard model

Cox proportional hazards regression, used to relate several risk factors or exposures, considered simultaneously, to survival time. In the current project we used the implementation of Cox regression hazard model implemented on lifelines packages. Although I had tried including other covariates such as age and temperature to be used in the model, at this phase we have used individual expression of each gene transcript iteratively in order to get hazard function and survival rates of each gene transcripts. Based on proportional hazard model we get an p-value estimates of gene transcript. However not all the gene transcripts fitted into the Cox survival model thus we excluded to get final list of pvalues for each transcripts.

2.4 Multiple testing correction

As discussed in earlier exercises we have implemented multiple testing correction for all the p.value estimates of gene expression for liver cancer to occur. However since we are testing multiple features, the pvalue < 0.05 might be too lenient estimates. Hence based on storey and tribasni papser [1] an q value estimates were made of all the pvalues calculated. We fitted the cubic spline in the π_0 estimates in the given λ values.

3 Result

3.1 TCGA Liver cancer characteristics

The data consists of 6 categorical features as shown in table 1 . Additionally since we are interested to know the dependent variable of event that in our case alive and death as a function of Living day we can observe the marked difference of alive and death events in patient cohort . Additionally, there is markedly difference in gender distribution in cancer cohort with more than half of the patient are male as shown in fig 2 and table 1.

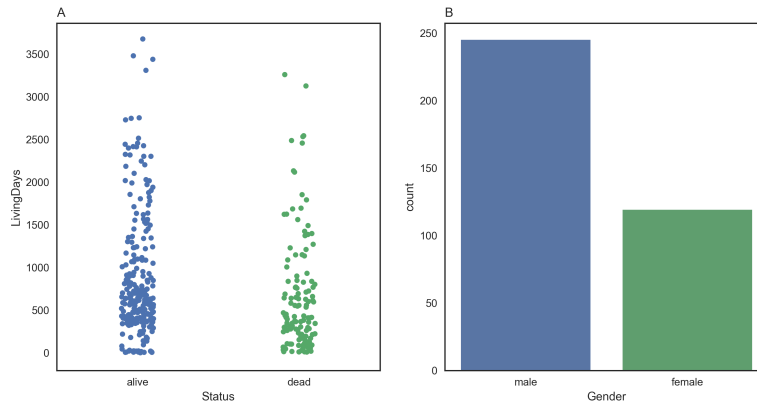


Figure 1: Pictures of animals

Summary statistics of TCGA liver cancer cohorts	
Gender	
male	245
female	119
Race	
white	182
asian	154
black or african american	17
not reported	10
american indian or alaska native	1
Liver cancer stage	
stage i	170
stage ii	83
stage iiia	63
not reported	24
stage iiic	9
stage iiib	8
stage iii	3
stage ivb	2
stage iv	1
stage iva	1
Status	
alive	234
dead	130

Table 1: Pictures of animals

3.2 Survival analysis of dataset

As discussed earlier , the cox proportional hazard survival analysis gives us estimates of survival probabilities and cumulative hazard of each transcript. As

shown in the figure 2 for one of the transcripts.. The model yields the p-values for significance of the regression model. There were 6032 features significant at $pvalue < 0.05$ which is quite alot. Thus in order to control the multiple testing of features , we need used the multiple hypothesis correction as indicated in Storey and tibsachi paper.

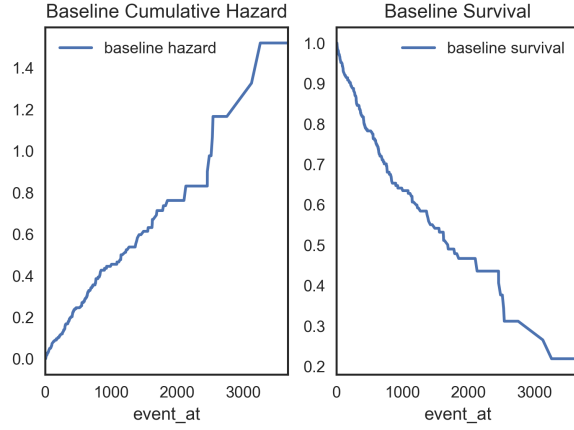


Figure 2: Pictures of animals

3.3 Estimating λ estimates on pvalues calculated

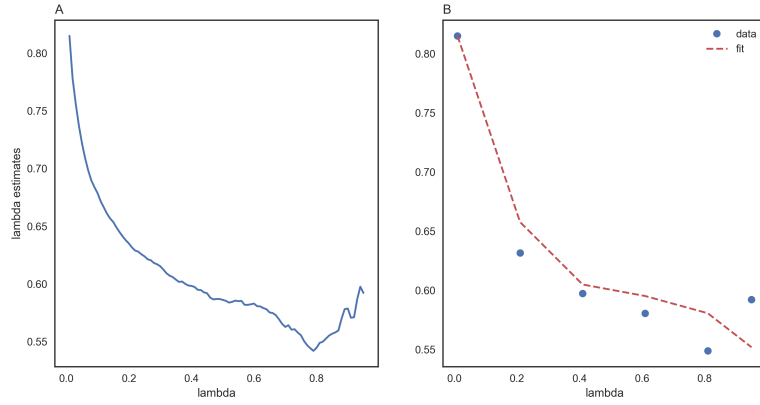


Figure 3: Pictures of animals

3.4 Estimation of expression levels based on FDR

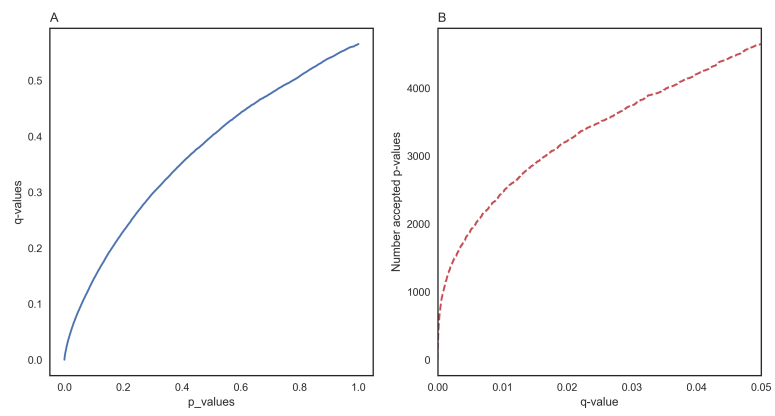


Figure 4: Pictures of animals