OPTIMIZING COX P.H. MODEL THROUGH FACTOR ANALYSIS OF MIXED DATA IN PBC SURVIVAL PREDICTION

A PROJECT REPORT

Submitted for Summer Internship at ISI, KOLKATA

MASTER OF SCIENCE
IN
STATISTICS

by

SUJASH KRISHNA BASAK

Under the guidance of

PROF. BISWABRATA PRADHAN

SQC & OR UNIT
INDIAN STATISTICAL INSTITUTE, KOLKATA
Barrackpore Trunk Road, Kolkata 700108, India

May-July 2024

Acknowledgment

I would like to express my thanks and acknowledgement to **Prof. Gopal Krishna Basak**, the Dean of Studies, Indian Statistical Institute, Kolkata; for giving me the opportunity to work on this project.

Specially, I would like to express my thanks and gratitude to my project supervisor **Prof. Biswabrata Pradhan (SQC & OR unit)**, for his immense support, guidance, and valuable advice to complete this dissertation paper, titled as – **"Optimizing Cox P.H. Model Through Factor Analysis of Mixed Data in PBC Survival Prediction"** This project has helped me to explore so many things related to the topic. It has been a great experience to work under the supervision of **Prof. Biswabrata Pradhan**, whose valuable lessons and suggestions have really enriched the content of my dissertation work.

I would also like to express my gratitude to **Prof. Mohua Banerjee**, Head of the Department of Mathematics and Statistics, Indian Institute of Technology, Kanpur, and the authority of the Indian Institute of Technology, Kanpur, for allowing me to work on this project.

Contents

1 Introduction 5

2 Some Basic Terminologies 7

2.1 Survival Function 7

2.2 Survival and Censoring Times 8

2.2.1 Right Censoring 8

2.3 Hazard Rate 9

2.4 The Kaplan–Meier Survival Curve 10

2.5 The Log-Rank Test 11

3 Regression Models With a Survival Response 12

3.1 Construction of the Likelihood and Proportional Hazard Model 12

3.2 Introduction to Cox's Proportional Hazard Model and Partial Likelihood to Estimate β 13

3.3 Testing of Hypothesis 15

3.3.1 Wald's Test 15

3.3.2 Likle-hood Ratio Test 15

3.3.3 Scores Test 15

3.4 Harrell's Concordance Index (or C-index) 16

4 Illustration of Kaplan–Meier Survival Curve and Fitting Cox Proportional Hazard Model in Data 17

4.1 Description of the data 17

4.2 Data manipulation 18

4.3 Kaplan–Meier Survival Curve 19

4.3.1 Survival Curves for Different Covariate 20

4.4 Interpretation of Log-Rank Test 20

4.5 Fitting of Cox Proportional Hazard Model 21

4.5.1 Defining the training set and test set 21

4.5.2 Summary of the Fitted Cox Model 21

4.5.3 Model Assumption Checking 22

4.5.4 Implementing K-fold cross-validation: 24

5 Variable (Predictor) Selection Methods in Cox Proportional Hazard Model 24

5.1 Subset Selection 25

5.2 Stepwise Variable Selection Procedure for Cox's Proportional Hazards Model: 26

5.3 Illustration of Stepwise Selection in Cox Proportional Hazard Model in "pbc" data 26

6 Shrinkage Method(LASSO) in Cox Proportional Hazard Model 28

6.1 Shrinkage Method 28

6.2 Basic Algorithm 29

6.3 Finding Optimal λ 30

6.4 Illustration of Shrinkage Method(LASSO) in Cox Model through "pbc" data 30

6.5 Interpretation of the summary: 32

7 Factor Analysis of Mixed Data (FAMD) 33

7.1 Theory 33

7.2 Geometrical Interpretation: 36

7.3 Illustration of FAMD on "pbc" dataset 38

7.4 Cox Regression with Principal Components from FAMD 41

8 Conclusions 43

9 References 44

Abstract

This article presents an application of the Kaplan-Meier estimator and a real data, the primary biliary cirrhosis collected in Mayo clinic, which conntains several time dependent covariates and the observations is measured repeatedly fits for the Cox proportional hazard models because the Cox PH model is the most popular method for survival data. We find principal components from the data using factor analysis for mixed data and then fit cox PH model on this. Also, by employing the model fit criterion to determine a suitable model for the real data and the criterion considered were the Akaike information criterion, called AIC for short. We find some factors, which have time to event, are agreed the assumptions of PH model. So the Cox PH model is appropriate to the data. Moreover, whether patients took D-penicillamine or not would not affect clealy the lifetime of patients with PBC.

Keywords: Cox proportional hazard model, Kaplan-Meier estimate, Factor analysis of mixed data, AIC, Time-dependent.

1. Introduction

The problem of analysing time to event data arises in several applied fields, such as medicine, biology, public health, epidemiology, engineering, economics, and demography. This kind of analysis, which arise in a unique kind of outcome variable: the time until an event occurs or sometimes referred to as failure is widely known as Survival Analysis.

Though the phrase "survival analysis" evokes a medical study, the applications of survival analysis extend far beyond medicine. For example, consider a company that wishes to model churn, the process by which customers cancel subscription to a service. The company might collect data on customers over some time period, in order to model each customer's time to cancellation as a function of demographics or other predictors. However, presumably not all customers will have cancelled their subscription by the end of this time period; for such customers, the time to cancellation is censored. In fact, survival analysis is relevant even in application areas that are unrelated to time. For instance, suppose we wish to model a person's weight as a function of some covariates, using a dataset with measurements for many people. Unfortunately, the scale used to weigh those people is unable to report weights above a certain number. Then, any weights that exceed that number are censored.

The 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. The main purpose of this study is to investigate the impact of D-penicillamine to lifetime of patients with Primary Biliary Cirrhosis (PBC). For individuals involved in medical research and those who are at risk due of PBC, it is of great interest to identify key factors that will lead to the progression of the disease, and those that lead to death. Though there is no clear cure to PBC, identifying these important factors can assist in determining how severe the disease is on a quantitative scale. This may provide assistance for research studies and understanding factors to prioritize in the development of PBC. The data set contains the covariates which are drug, patient's age at registration, patient's sex, presence of ascites, hepatomegaly, spiders and edema, serum bilirubin, albumin, alkaline phosphatase, serum glutamicoxaloacetic transaminase (SGOT), platelets per cubic, prothrombin time and histologic stage of disease, etc...

This article gives a summarized account of the techniques for analysis, such as the Kaplan-Meier estimator, AIC and Cox's PH model for time dependent covariate. Checking the PH assumption and applying the methodologies in to the PBC data, explanations of the suitable Cox's PH model by extracting principal components using Factorial analysis of mixed data obtained by the criteria AIC. Finally, we make some conclusions about the model.

2. Some Basic Terminologies

Let X be the time until some specified event. This event may be death, the appearance of a tumour, the development of some disease, recurrence of a disease, equipment breakdown, cessation of breast feeding, and so forth. Furthermore, the event may be a good event, such as remission after some treatment, conception, cessation of smoking, and so forth. More precisely, in this chapter, X is a non-negative random variable from a homogeneous population. Four functions characterize the distribution of X, namely, the **survival function**, which is the probability of an individual surviving to time x; the **hazard rate** (function), sometimes termed **risk function**, which is the chance an individual of age x experiences the event in the next instant in time; the probability density (or probability mass) function, which is the unconditional probability of the event's occurring at time x; and the **mean residual life (m.r.l.)** at time x, which is the mean time to the event of interest, given the event has not occurred at x.

2.1 Survival Function

The basic quantity employed to describe time-to-event phenomena is the survival function, the probability of an individual surviving beyond time x (experiencing the event after time x). It is defined as,

$$S(x) = \Pr(X > x)$$

In the context of equipment or manufactured item failures, $S(x)$ is referred to as the reliability function. If X is a continuous random variable, then, $S(x)$ is a continuous, strictly decreasing function.

When X is a continuous random variable, the survival function is the complement of the cumulative distribution function, that is,

$$S(x) = 1 - F(x),$$

where

$$F(x) = \Pr(X \leq x).$$

Also, the survival function is the integral of the probability density function, $f(x)$, that is,

$$S(x) = \Pr(X > x) = \int_{x}^{\infty} f(t)\, dt.$$

Thus,

$$f(x) = -\frac{dS(x)}{dx}.$$

Note that $f(x)\, dx$ may be thought of as the "approximate" probability that the event will occur at time x and that $f(x)$ is a nonnegative function with the area under $f(x)$ being equal to one.

2.2 Survival and Censoring Times

For each individual, we suppose that there is a true survival time, X, as well as a true censoring time, C. (The survival time is also known as the failure time or the event time.) The survival time represents the time at which the event of interest occurs: for instance, the time at which the patient dies, or the customer cancels his or her subscription. By contrast, the censoring time is the time at which censoring occurs: for example, the time at which the patient drops out of the study or the study ends. There are various categories of censoring, such as right censoring, left censoring, and interval censoring. Mainly, right censoring has been used in this project.

2.2.1 Right Censoring

We will consider right censoring where the event is observed only if it occurs prior to some prespecified time. These censoring times may vary from individual to individual. Because of time or cost considerations, the investigator will terminate the study or report the results before all subjects realize their events. In this instance, if there are no accidental losses or subject withdrawals, all censored observations have times equal to the length of the study period.

In right censoring, we observe either the survival time X or else the censoring time C. The X's are assumed to be independent and identically distributed with probability density function f(x) and survival function S(x). Specifically, we observe the random variable

$$T = \min(X, C).$$

In other words, if the event occurs before censoring (i.e. $X < C$) then we observe the true survival time X; however, if censoring occurs before the event ($X > C$) then we observe the censoring time. We also observe a status indicator,

$$\delta = \begin{cases} 1 & \text{if } X \leq C \\ 0 & \text{if } X > C \end{cases}.$$

Thus, $\delta = 1$ if we observe the true survival time, and $\delta = 0$ if we instead observe the censoring time. The data from this experiment can be conveniently represented by pairs of random variables (T, δ).

2.3 Hazard Rate

A basic quantity, fundamental in survival analysis, is the hazard function. This function is also known as the conditional failure rate in reliability, the force of mortality in demography, the intensity function in stochastic processes, the age-specific failure rate in epidemiology, the inverse of the Mill's ratio in economics, or simply as the hazard rate. The hazard rate is defined by,

$$h(x) = \lim_{\Delta x \to 0} \frac{P[x \leq X < x + \Delta x \mid X \geq x]}{\Delta x}$$

If X is a continuous random variable, then,

$$h(x) = \frac{f(x)}{S(x)} = -\frac{d \ln[S(x)]}{dx}$$

A related quantity is the cumulative hazard function $H(x)$, defined by,

$$H(x) = \int_0^x h(u)\, du = -\ln[S(x)]$$

Thus, for continuous lifetimes,

$$S(x) = \exp[-H(x)] = \exp\left[-\int_0^x h(u)\,du\right]$$

$h(x)\Delta x$ may be viewed as the "approximate" probability of an individual of age x experiencing the event in the next instant. This function is particularly useful in determining the appropriate failure distributions utilizing qualitative information about the mechanism of failure and for describing the way in which the chance of experiencing the event changes with time. There are many general shapes for the hazard rate. The only restriction on $h(x)$ is that it be nonnegative, i.e., $h(x) \geq 0$.

2.4 The Kaplan–Meier Survival Curve

The survival curve, or survival function, is defined as *survival curve survival function*

$$S(t) = \Pr(T > t).$$

This decreasing function quantifies the probability of surviving past time t. For example, suppose that a company is interested in modeling customer churn. Let T represent the time that a customer cancels a subscription to the company's service. Then $S(t)$ represents the probability that a customer cancels later than time t. The larger the value of $S(t)$, the less likely that the customer will cancel before time t.

We have seen that estimating $S(t)$ is complicated by the presence of censoring. We now present an approach to overcome these challenges. We

let $d_1 < d_2 < \cdots < d_K$ denote the K unique death times among the non-censored patients, and we let q_k denote the number of patients who died at time d_k. For $k = 1, \ldots, K$, we let r_k denote the number of patients alive and in the study just before d_k; these are the at-risk patients. The set of patients that are at risk at a given time are referred to as the *risk set. risk set* By the law of total probability,

$$\Pr(T > d_k) = \Pr(T > d_k | T > d_{k-1})\Pr(T > d_{k-1}) + \Pr(T > d_k | T \leq d_{k-1})\Pr(T \leq d_{k-1}).$$

The fact that $d_{k-1} < d_k$ implies that $\Pr(T > d_k | T \leq d_{k-1}) = 0$ (it is impossible for a patient to survive past time d_k if he or she did not survive until an earlier time d_{k-1}). Therefore,

$$S(d_k) = \Pr(T > d_k) = \Pr(T > d_k | T > d_{k-1}) \Pr(T > d_{k-1}).$$

Plugging in survival function again, we see that

$$S(d_k) = \Pr(T > d_k | T > d_{k-1}) S(d_{k-1}).$$

This implies that

$$S(d_k) = \Pr(T > d_k | T > d_{k-1}) \times \cdots \times \Pr(T > d_2 | T > d_1) \Pr(T > d_1).$$

We now must simply plug in estimates of each of the terms on the right-hand side of the previous equation. It is natural to use the estimator

$$\Pr(T > d_j | T > d_{j-1}) = \frac{r_j - q_j}{r_j},$$

which is the fraction of the risk set at time d_j who survived past time d_j. This leads to the Kaplan–Meier estimator of the survival curve: *Kaplan– Meier estimator*

$$S_W(d_k) = \prod_{j=1}^{k} \left(\frac{r_j - q_j}{r_j} \right).$$

For times t between d_k and d_{k+1}, we set $S_W(t) = S_W(d_k)$. Consequently, the Kaplan–Meier survival curve has a step-like shape.

2.5 The Log-Rank Test

We wish to test whether the survival functions of two or more samples could have significant difference or not. Here we could test whether the mean survival time among the two samples are equal or not. But the presence of censoring again creates a complication. To overcome this challenge, we will conduct test,which is called log rank test. At risk in combined sample at time t_i. Let O_i be the observed numbers and E_i be the expected numbers of failures in group i, $i = 1, \cdots, K$. Then the log rank test statistic is defined as

$$X^2 = \sum_{i=1}^{K} \frac{(O_i - E_i)^2}{E_i}$$

which has approximately Chi-square distribution with degree of freedom $K - 1$. A large value X^2 could lead to reject the testing hypothesis that there are discrepancies in survivor among the K families.

3. Regression Models With a Survival Response

3.1 Construction of the Likelihood and Proportional Hazard Model

Let Y denote the time to some event. Our data, based on a sample of size n, consists of the triple $(Y_j, \delta_j, X_j), j = 1, \ldots, n$ where Y_j is the time on study for the j-th patient, δ_j is the event indicator for the j-th patient ($\delta_j = 1$, if the event has occurred and $\delta_j = 0$, if the lifetime is right-censored) and $X_j = (X_{j1}, \ldots, X_{jp})^t$ is the vector of covariates or risk factors for the j-th individual at time y which may affect the survival distribution of Y. Here the X_{jk}'s, $k = 1, \ldots, p$, are the covariates for the j-th individual.

Now, the likelihood associated with the j-th observation is,

$$L_i = \begin{cases} f(y_i) & \text{if the } i\text{th observation is not censored} \\ S(y_i) & \text{if the } i\text{th observation is censored} \end{cases}$$

The intuition behind this is as follows: if $Y = y_j$ and the j-th observation is not censored, then the likelihood is the probability of dying in a tiny interval around time y_j. If the j-th observation is censored, then the likelihood is the probability of surviving at least until time y_j. Assuming that the n observations are independent, the likelihood for the data takes the form,

$$\mathcal{L} = \prod_{i=1}^{n} f(y_i)^{\delta_i} S(y_i)^{1-\delta_i} = \prod_{i=1}^{n} h(y_i)^{\delta_i} S(y_i) \quad \text{................(1)}$$

However, if we want to model the survival time as a function of the covariates, then it is convenient to work directly with the hazard function, instead of the probability density function. One possible approach is to assume a functional form for the hazard function $h(t|x_i)$, such as

$$h(t|x_i) = \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})$$

where the exponent function guarantees that the hazard function is non-negative. The exponential hazard function does not vary with time. Given $h(t|x_i)$, we could calculate $s(t|x_i)$. Plugging these equations into (1), we could then maximize the likelihood in order to estimate the parameter $\beta = (\beta_1, \dots, \beta_p)^t$.

The proportional hazards assumption states that,

$$h(t|x_i) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right) \quad \text{................(2)}$$

where $h_0(t) \geq 0$ is an unspecified function, known as the baseline hazard. It is the hazard function for an individual with features $x_{i1} = \ldots = x_{ip} = 0$. The name "proportional hazards" arises from the fact that the hazard function for an individual with feature vector x_i is some unknown function $h_0(t)$ times the factor $\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$. The quantity $\exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$ is called the relative risk for the feature vector $x_j = (x_{j1}, \dots, x_{jp})^t$, relative to that for the feature vector $x_j = (0, \dots, 0)^t$.

Basically, we make no assumptions about its functional form of baseline hazard. We allow the instantaneous probability of death at time t, given that one has survived at least until time t, to take any form. This means that the hazard function is very flexible and can model a wide range of relationships between the covariates and survival time. Only assumption is that a one-unit increase in x_{ij} corresponds to an increase in $h(t|x_i)$ by a factor of $\exp(\beta_j)$.

3.2 Introduction to Cox's Proportional Hazard Model and Partial Likelihood to Estimate β

Because the form of $h_0(t)$ in the proportional hazard assumption is unknown, we cannot simply plug $h(t|x_i)$ into the likelihood (1) and then estimate $\beta = (\beta_1, \dots, \beta_p)^t$ by maximum likelihood. The interesting thing about Cox's proportional hazards model (Cox 1972) lies in the fact that it is possible to estimate β without having to specify the form of $h_0(t)$.

For simplicity, assume that there are no ties among the failure, or death, times: i.e., each failure occurs at a distinct time. Assume that $\delta_i = 1$, i.e., the i-th observation is uncensored, and thus y_i is its failure time. Then the hazard function for the i-th observation at time y_i is $h(y_i|x_i) = h_0(y_i)\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)$, andthe total hazard at time y_i for the at-risk observations is

$$\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right).$$

Therefore, the probability that the i-th observation is the one to fail at time y_i (as opposed to one of the other observations in the risk set) is

$$\frac{h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i} h_0(y_i)\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} = \frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i}\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)}.$$

Notice that the unspecified baseline hazard function $h_0(y_i)$ cancels out of the numerator and denominator!

The partial likelihood is simply the product of these probabilities over all the uncensored observations,

$$L(\beta) = \prod_{i:\delta_i=1}\frac{\exp\left(\sum_{j=1}^{p} x_{ij}\beta_j\right)}{\sum_{i':y_{i'}\geq y_i}\exp\left(\sum_{j=1}^{p} x_{i'j}\beta_j\right)} \quad \text{.................(3)}$$

Critically, the partial likelihood is valid regardless of the true value of $h_0(t)$, making the model very flexible and robust. To estimate β, we simply maximize the partial likelihood (3) with respect to β. But no closed form solution is available, and so iterative algorithms are required.

(In general, the partial likelihood is used in settings where it is difficult to compute the full likelihood for all the parameters. Instead, we compute a likelihood for just the parameters of primary interest: in this case, $\beta_1, \ldots, \beta_p$. It can be shown that maximizing (3) provides good estimates for these parameters).

3.3 Testing of Hypothesis

There are three main tests for hypotheses about regression parameters β. Let $b = (b_1, ..., b_p)^t$ denote the (partial) maximum likelihood estimates of β and let $I(\beta)$ be the $p \times p$ information matrix evaluated at β.

3.3.1 Wald's Test

The first test is the usual test based on the asymptotic normality of the (partial) maximum likelihood estimates, referred to as Wald's test. It is based on the result that, for large samples, b has a p-variate normal distribution with mean β and variance-covariance estimated by $I^{-1}(b)$. A test of the global hypothesis of $H_0 : \beta = \beta_0$ is,

$$\chi^2_W = (b - \beta_0)^t I(b)(b - \beta_0)$$

which has a chi-squared distribution with p degrees of freedom if H_0 is true for large samples.

3.3.2 Likle-hood Ratio Test

The second test is the likelihood ratio test of the hypothesis of $H_0 : \beta = \beta_0$ and uses,

$$\chi^2_{LR} = 2\left[LL(b) - LL(\beta_0)\right]$$

which has a chi-squared distribution with p degrees of freedom under H_0 for large n. Here $LL(\beta)$ is the logarithm of the partial likelihood $PL(\beta)$ mentioned in (3).

3.3.3 Scores Test

The third test is the scores test. It is based on the efficient scores, $U(\beta) = (U_1(\beta), ..., U_p(\beta))^t$, where $U_k(\beta)$ is defined by $U_k(\beta) = \frac{\partial LL(\beta)}{\partial \beta_k}$. For large samples, $U(\beta)$ is asymptotically p-variate normal with mean 0 and covariance $I(\beta)$ when H_0 is true. Thus, a test of $H_0 : \beta = \beta_0$ is,

$$\chi^2_{SC} = (U(\beta_0))^t I^{-1}(\beta_0)(U(\beta_0))$$

which has a large sample chi-squared distribution with p degrees of freedom under H_0.

3.4 Harrell's Concordance Index (or C-index)

We use the area under the ROC curve — often referred to as the "AUC" — to quantify the performance of a two-class classifier. Define the score for the i-th observation to be the classifier's estimate of $\Pr(Y = 1|X = x_i)$. It turns out that if we consider all pairs consisting of one observation in Class 1 and one observation in Class 2, then the AUC is the fraction of pairs for which the score for the observation in Class 1 exceeds the score for the observation in Class 2.

This suggests a way to generalize the notion of AUC to survival analysis. We calculate an estimated risk score, $\hat{\eta}_i = \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}$, for $i = 1, \ldots, n$, using the Cox model coefficients. If $\hat{\eta}_{i'} > \hat{\eta}_i$, then the model predicts that the i'-th observation has a larger hazard than the i-th observation, and thus that the survival time t_i will be greater than $t_{i'}$. Thus, it is tempting to try to generalize AUC by computing the proportion of observations for which $t_i > t_{i'}$ and $\hat{\eta}_{i'} > \hat{\eta}_i$.

However, things are not quite so easy, because we do not observe $t_1, \ldots, t_n$; instead, we observe the (possibly-censored) times $y_1, \ldots, y_n$, as well as the censoring indicators $\delta_1, \ldots, \delta_n$.

Therefore, Harrell's concordance index (or C-index) computes the proportion of observation pairs for which $\hat{\eta}_{i'} > \hat{\eta}_i$ and $y_i > y_{i'}$:

$$C = \frac{\sum_{i,i':y_i>y_{i'}} I(\hat{\eta}_{i'} > \hat{\eta}_i)\delta_{i'}}{\sum_{i,i':y_i>y_{i'}} \delta_{i'}}$$

where the indicator variable $I(\hat{\eta}_{i'} > \hat{\eta}_i)$ equals one if $\hat{\eta}_{i'} > \hat{\eta}_i$, and equals zero otherwise. The numerator and denominator are multiplied by the status indicator $\delta_{i'}$, since if the i'-th observation is uncensored (i.e., if $\delta_{i'} = 1$), then $y_i > y_{i'}$ implies that $t_i > t_{i'}$. By contrast, if $\delta_{i'} = 0$, then $y_i > y_{i'}$ does not imply that $t_i > t_{i'}$.

4. Illustration of Kaplan–Meier Survival Curve and Fitting Cox Proportional Hazard Model in Data

4.1 Description of the data

The datasets are sourced from the Vanderbilt University Department of Biostatistics.[1] The PBC dataset is from a Mayo Clinical trial studying PBC 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.

The variable descriptions are listed below:

- **fu.days**: number of days between registration and the earlier of death, transplantation, or study analysis time in July, 1986
- **status**: 0 = alive, 1 = dead
- **drug**: 1 = D-penicillamine, 2 = placebo
- **age**: age in days
- **sex**: 0 = male, 1 = female
- **ascites**: presence of ascites - 0 = no, 1 = yes
- **hept**: presence of hepatomegaly - 0 = no, 1 = yes
- **spiders**: presence of spiders - 0 = no, 1 = yes

[1] https://hbiostat.org/data/

- **edema**: presence of edema - 0 = no edema and no diuretic therapy for edema; 0.5 = edema present without diuretics, or edema resolved by diuretics; 1 = edema despite diuretic therapy

- **bili**: serum bilirubin in mg/dl

- **chol**: serum cholesterol in mg/dl

- **albumin**: albumin in gm/dl

- **copper**: urine copper in μg/day

- **alk.phos**: alkaline phosphatase in U/liter

- **sgot**: SGOT in U/ml

- **trig**: triglycerides in mg/dl

- **platelet**: platelets per cubic ml / 1000

- **protime**: prothrombin time in seconds

- **stage**: histologic stage of disease

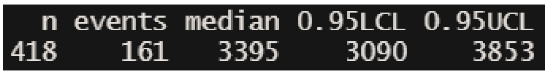
4.2 Data manipulation

To clean up the data and factors that will model the status of the patient (0 - alive, or 1 - deceased), some are not going to make practical sense. For brevity, the dataset will be filtered to remove the patients who are not randomized, which are inherently missing many measurements. This will limit the dataset to those treated with a placebo or the drug of interest.

For this initial study, there will be no imputed data points. For brevity, and to reduce model noise, very few NA's and their related rows will be removed. This will further contribute to a more complete dataset.

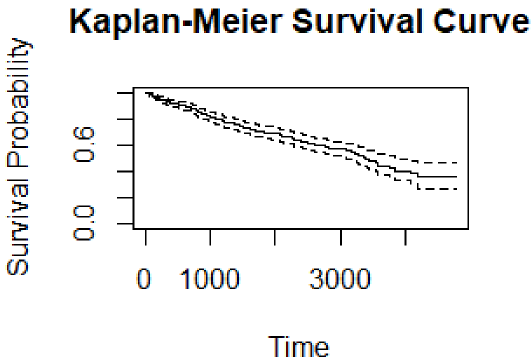
4.3 Kaplan–Meier Survival Curve

Kaplan–Meier Survival Curve has been fitted on the dataset using R software. The summary is shown below.



```
   n events median 0.95LCL 0.95UCL
418    161   3395    3090    3853
```

(a)



(b)

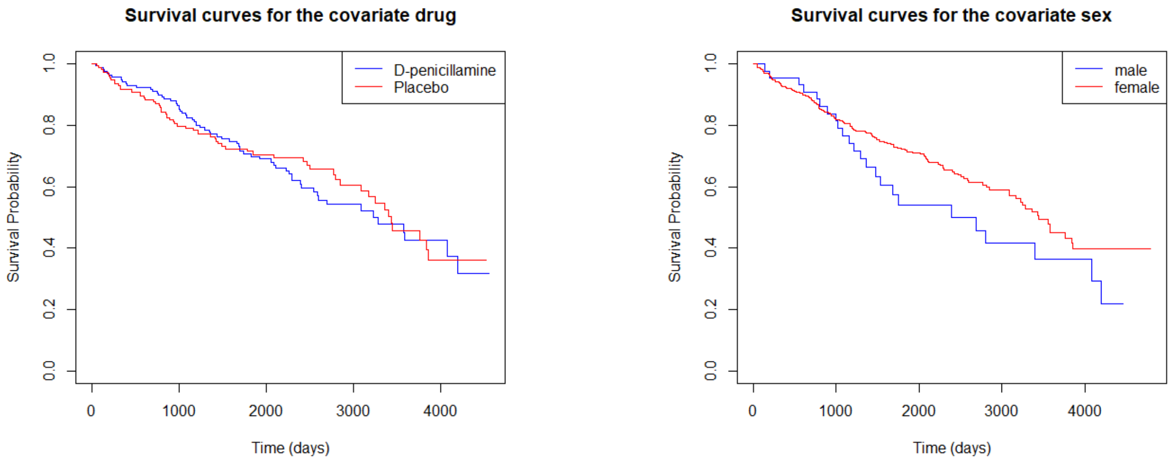
Figure 1: (a) R output of Kaplan-Meier estimate (b) Kaplan-Meier survival curve

Interpretation:

- **n**: The number of observations included in the analysis (418 patients).
- **events**: The number of events observed (161 events)
- **median**:The median survival time, which is the time by which 50% of the patients have experienced the event. In this case, the median survival time is 3395 days.
- **0.95LCL and 0.95UCL**: The lower and upper limits of the 95% confidence interval for the median survival time. These intervals give an estimate of the precision of the median survival time. Here, the 95% confidence interval ranges from 3090 to 3853 days.

4.3.1 Survival Curves for Different Covariate

Survival curves for different covariates has been plotted below to see if there is any difference between repective levels of that covariates.



(a)

(b)

Figure 2: Survival curves for different covariates

Interpretation:

Figure 1 plots the estimators of survival function for each level of drug and sex with the Kaplan-Meier estimator. In figure 2(b) we know that there are discrepancies for the patients between male and female and the female patients have more larger survival probability than male. Contrary to drug, the survival functions of the two drug levels look like similar to each other classification. There are a little cross over within each level of the two covaraites, but we need further evidence to conclude this.

4.4 Interpretation of Log-Rank Test

We have computed the values of log rank test for drug and sex by using the code **survdiff** in R program and their p-values obtained by log rank test are 0.5 (larger than 0.05) and 0.03 (less than 0.05), respectively. So the covariate drug might not have differences between two drug levels but there are significantly discrepancies for sex.

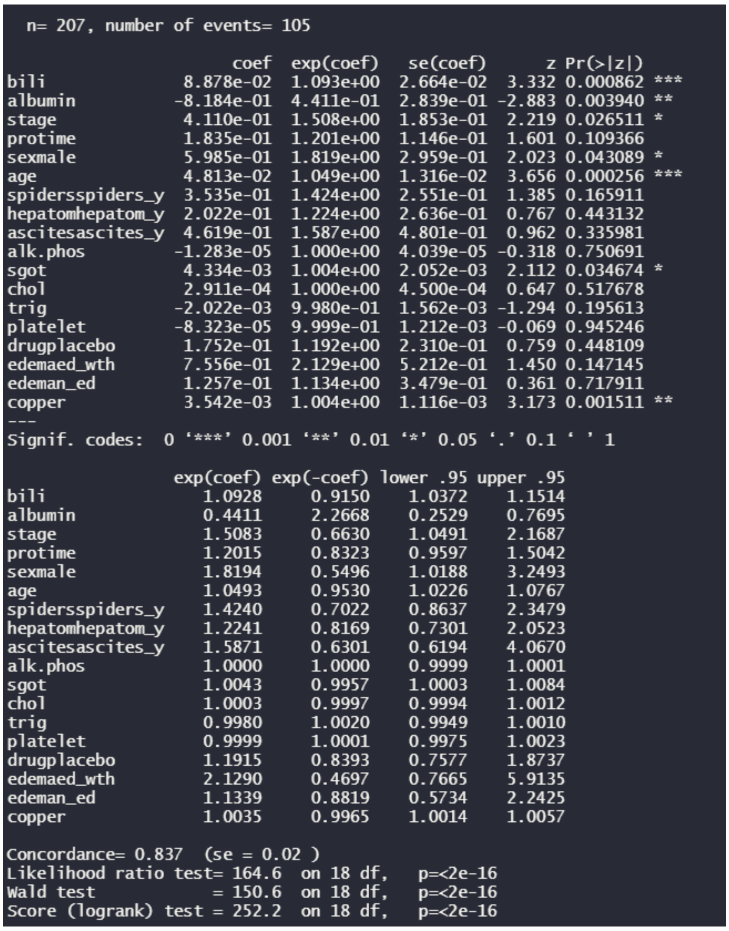
4.5 Fitting of Cox Proportional Hazard Model

4.5.1 Defining the training set and test set

The dataset has been divided into 80% and 20% as the members of training set and testing set respectively at random. Now we will fit Cox model on training dataset and check accuracy on test data.

4.5.2 Summary of the Fitted Cox Model

A Cox proportional hazard model has been fitted on the training data set using R software. The summary of the model is shown below.



```
  n= 207, number of events= 105

                       coef  exp(coef)   se(coef)      z Pr(>|z|)
bili              8.878e-02  1.093e+00  2.664e-02  3.332 0.000862 ***
albumin          -8.184e-01  4.411e-01  2.839e-01 -2.883 0.003940 **
stage             4.110e-01  1.508e+00  1.853e-01  2.219 0.026511 *
protime           1.835e-01  1.201e+00  1.146e-01  1.601 0.109366
sexmale           5.985e-01  1.819e+00  2.959e-01  2.023 0.043089 *
age               4.813e-02  1.049e+00  1.316e-02  3.656 0.000256 ***
spidersspiders_y  3.535e-01  1.424e+00  2.551e-01  1.385 0.165911
hepatomhepatom_y  2.022e-01  1.224e+00  2.636e-01  0.767 0.443132
ascitesascites_y  4.619e-01  1.587e+00  4.801e-01  0.962 0.335981
alk.phos         -1.283e-05  1.000e+00  4.039e-05 -0.318 0.750691
sgot              4.334e-03  1.004e+00  2.052e-03  2.112 0.034674 *
chol              2.911e-04  1.000e+00  4.500e-04  0.647 0.517678
trig             -2.022e-03  9.980e-01  1.562e-03 -1.294 0.195613
platelet         -8.323e-05  9.999e-01  1.212e-03 -0.069 0.945246
drugplacebo       1.752e-01  1.192e+00  2.310e-01  0.759 0.448109
edemaed_wth       7.556e-01  2.129e+00  5.212e-01  1.450 0.147145
edeman_ed         1.257e-01  1.134e+00  3.479e-01  0.361 0.717911
copper            3.542e-03  1.004e+00  1.116e-03  3.173 0.001511 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                 exp(coef) exp(-coef) lower .95 upper .95
bili                1.0928     0.9150    1.0372    1.1514
albumin             0.4411     2.2668    0.2529    0.7695
stage               1.5083     0.6630    1.0491    2.1687
protime             1.2015     0.8323    0.9597    1.5042
sexmale             1.8194     0.5496    1.0188    3.2493
age                 1.0493     0.9530    1.0226    1.0767
spidersspiders_y    1.4240     0.7022    0.8637    2.3479
hepatomhepatom_y    1.2241     0.8169    0.7301    2.0523
ascitesascites_y    1.5871     0.6301    0.6194    4.0670
alk.phos            1.0000     1.0000    0.9999    1.0001
sgot                1.0043     0.9957    1.0003    1.0084
chol                1.0003     0.9997    0.9994    1.0012
trig                0.9980     1.0020    0.9949    1.0010
platelet            0.9999     1.0001    0.9975    1.0023
drugplacebo         1.1915     0.8393    0.7577    1.8737
edemaed_wth         2.1290     0.4697    0.7665    5.9135
edeman_ed           1.1339     0.8819    0.5734    2.2425
copper              1.0035     0.9965    1.0014    1.0057

Concordance= 0.837  (se = 0.02 )
Likelihood ratio test= 164.6  on 18 df,   p=<2e-16
Wald test            = 150.6  on 18 df,   p=<2e-16
Score (logrank) test = 252.2  on 18 df,   p=<2e-16
```

Figure 3: Summary of the Fitted Cox Model

Interpretation of the summary:

- **Predictor Interpretations**: The exp(coef) indicates hazard ratio (HR). It indicates how the hazard changes with a one-unit increase in the predictor variable. For example exp(coef) of age is 1.0493. It means the hazard increases by about 4% for each additional year of age

- All the three tests Likelihood ratio test, Wald test and Score (log rank) test has p-values $<$ 2e-16, which means that in the light of the given data it seems that all the coefficients of the corresponding predictors are not zero simultaneously

- The model yielded Concordance = 0.837 (se = 0.02), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 83.7% accuracy. For test data Concordance = 0.9073482

- Also, the survival probabilities of the individuals of the test data set have been computed. The average standard error in the prediction is found to be 0.04976

4.5.3 Model Assumption Checking

In this project, the Schoenfeld residuals test was utilized to evaluate the proportional hazards assumption in the Cox proportional hazards model. The test's null hypothesis (H_0) states that the Schoenfeld residuals are independent of time, indicating that the proportional hazards assumption holds. The alternative hypothesis (H_A) posits that the residuals are dependent on time, suggesting a violation of the assumption. For each covariate X_j, the test calculates the correlation ρ_j between Schoenfeld residuals r_{ij} and event times t_i, resulting in a chi-squared test statistic (χ_j^2). The test statistic for each covariate is given by:

$$\chi_j^2 = \frac{\left(\sum_i (t_i - \bar{t}) r_{ij}\right)^2}{\sum_i (t_i - \bar{t})^2 \sum_i r_{ij}^2 / n}$$

where $\bar{t}$ is the mean of the event times, and n is the number of events. The global test aggregates these individual statistics into a single chi-squared statistic:

$$\chi^2_{\text{GLOBAL}} = \sum_j \chi_j^2$$

This global statistic follows a chi-squared distribution with degrees of freedom equal to the number of covariates p. In this analysis, the Schoenfeld residuals test was implemented using the `cox.zph` function in R, which provided a global p-value of 0.07099. Since this p-value exceeds the common significance threshold of 0.05, there is no significant evidence against the proportional hazards assumption. Thus, the assumption holds true for the Cox model applied to the Primary Biliary Cirrhosis (PBC) data, affirming the model's validity.

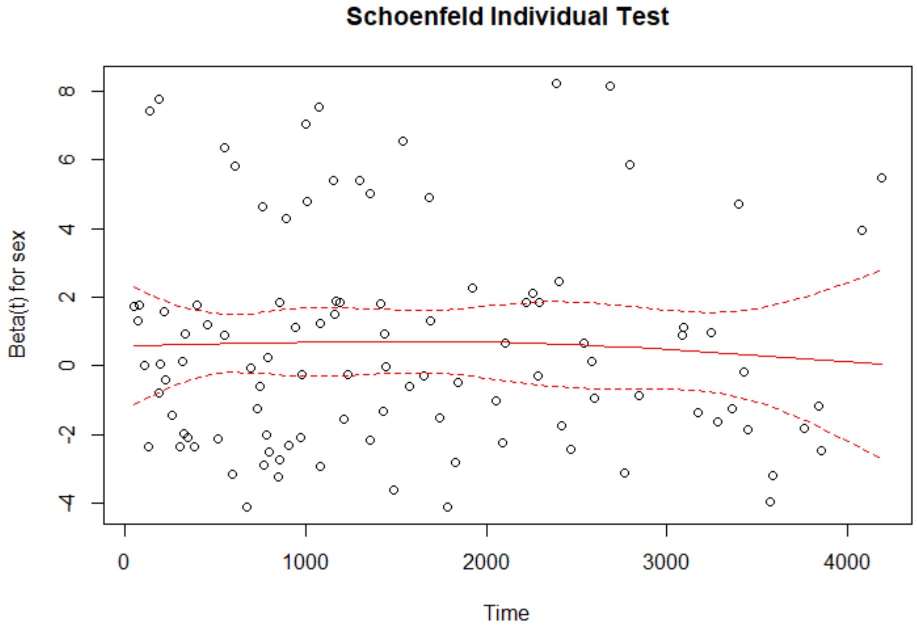


Figure 4: Schoenfeld residual plot for covariate sex

Systematic departures from a horizontal line are indicative of non-proportional hazards, since the proportional hazards assumes that the estimates β do not vary significantly over time. Each line represents the smoothed residuals and confidence intervals for covariate sex. From graphical inspection, there is no discernible pattern with time observed for the covariate sex. This suggests that the assumption of proportional hazards is supported for the covariate sex and also it is suggested by its pvalue $= 0.4359$ from the Schoenfeld residuals test.

4.5.4 Implementing K-fold cross-validation:

K-fold cross-validation is a technique used to assess the performance and generalizability of a machine learning model. In this method, the dataset is divided into K equally sized folds. The model is trained K times, each time using K-1 folds for training and the remaining fold for validation. This process ensures that every data point is used for both training and validation exactly once. This helps in obtaining a more reliable estimate of model performance by reducing variance and mitigating the risk of overfitting compared to a single train-test split.

In this data, we have applied K-fold cross-validation with K = 4 and we found mean concordance = 0.827 and mean standard error = 0.10349 for test data.

5. Variable (Predictor) Selection Methods in Cox Proportional Hazard Model

Provided that the true relationship between the response and the predictors is approximately linear, the least squares estimates will have low bias. If n > p that is, if n, the number of observations, is much larger than p, the number of variables—then the least squares estimates tend to also have low variance, and hence will perform well on test observations. However, if n is not much larger than p, then there can be a lot of variability in the least squares fit, resulting in overfitting and consequently poor predictions on future observations not used in model training. And if p > n, then there is no longer a unique least squares coefficient estimate: the variance is infinite so the method cannot be used at all. By constraining or shrinking the estimated coefficients, we can often substantially reduce the variance at the cost of a negligible increase in bias. This can lead to substantial improvements in the accuracy with which we can predict the response for observations not used in model training. It is often the case that some or many of the variables used in a multiple regression model are in fact not associated with the response. Including such irrelevant variables leads to unnecessary complexity in the resulting model. By removing these variables, that is, by setting the corresponding coefficient estimates to zero, we can obtain a model that is more easily interpreted. Here in this project, we have considered two approaches for automatically performing feature selection or variable selection that is, for excluding irrelevant variables from a multiple regression model.

5.1 Subset Selection

This approach involves identifying a subset of the p predictors that one can believe to be related to the response. Then model is fitted corresponding to the reduced set of variables.

Stepwise Selection: When p is large, i.e., larger the search space, the higher the chance of finding models that look good on the training data, even though they might not have any predictive power on future data. Thus, an enormous search space can lead to overfitting and high variance of the coefficient estimates. For both reasons, stepwise methods, which explore a far more restricted set of models.

- **Forward Stepwise Selection:** Forward stepwise selection begins with a model containing no predictors, and then adds predictors to the model, oneat-a-time, until all the predictors are in the model. At each step the variable that gives the greatest additional improvement to the fit is added to the model. Lastly, select a single best model from among all the models consisting different number of predictors using cross-validated prediction error, Cp (AIC), BIC or adjusted R2 .

- **Backward Stepwise Selection:** Unlike forward stepwise selection, it begins with the full least squares model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.

- **Hybrid Approach (Mixture of Both Forward and Backward Selection):** The best subset, forward stepwise, and backward stepwise selection approaches generally give similar but not identical models. As another alternative, hybrid versions of forward and backward stepwise selection are available, in which variables are added to the model sequentially, in analogy to forward selection. However, after adding each new variable, the method may also remove any variables that no longer provide an improvement in the model fit.

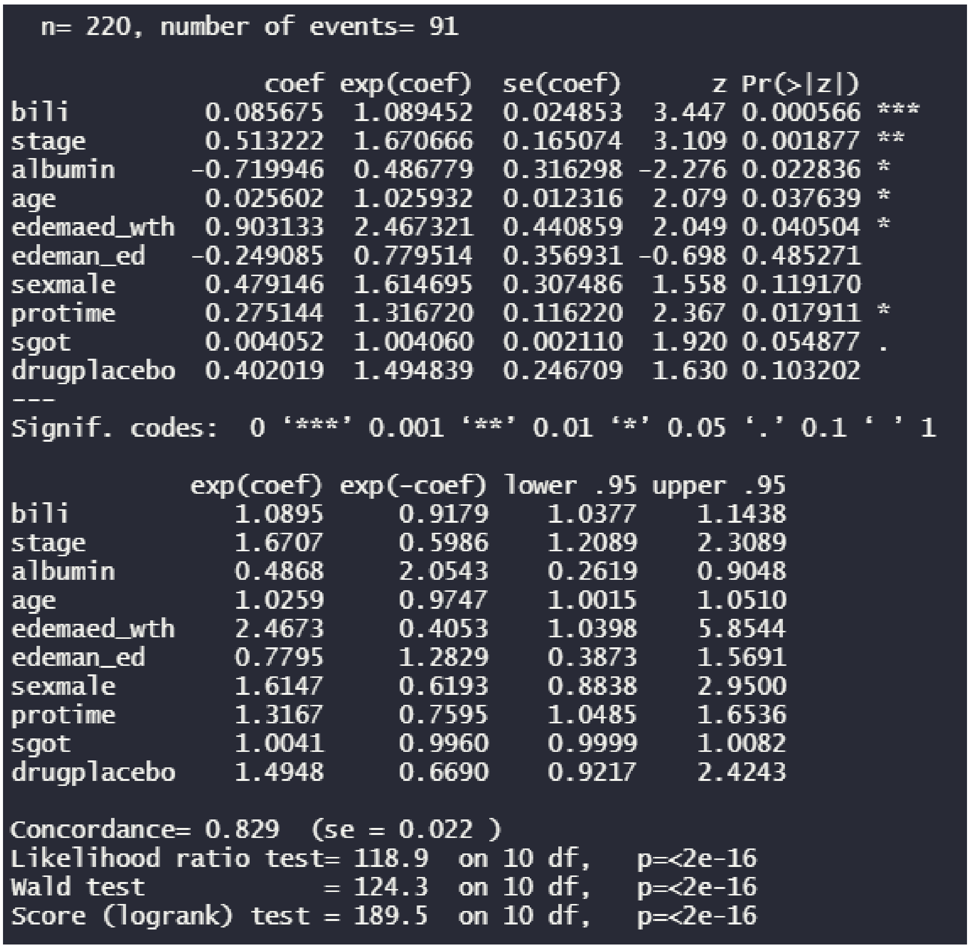
5.2 Stepwise Variable Selection Procedure for Cox's Proportional Hazards Model:

This stepwise variable selection procedure (with iterations between the 'forward' and 'backward' steps) can be applied to obtain the best candidate final Cox's proportional hazards model. The goal of regression analysis is to find one or a few parsimonious regression models that fit the observed data well for effect estimation and/or outcome prediction. To ensure a good quality of analysis, the model-fitting techniques for **(1) variable selection, (2) goodness-of-fit assessment, and (3) regression diagnostics** and remedies should be used in regression analysis. The stepwise variable selection procedure (with iterations between the 'forward' and 'backward' steps) is one of the best ways to obtaining the best candidate final regression model. All the bivariate significant and non-significant relevant covariates and some of their interaction terms (or moderators) are put on the variable list to be selected. The significance levels for entry (SLE) and for stay (SLS) may be set at 0.15 or larger for being conservative. Then, with the aid of substantive knowledge, the best candidate final regression model is identified manually by dropping the covariates with p value $>$ 0.05 one at a time until all regression coefficients are significantly different from 0 at the chosen alpha level of 0.05. Since the statistical testing at each step of the stepwise variable selection procedure is conditioning on the other covariates in the regression model, the multiple testing problem is not of concern. Any discrepancy between the results of bivariate analysis and regression analysis is likely due to the confounding effects of uncontrolled covariates in bivariate analysis or the masking effects of intermediate variables (or mediators) in regression analysis.

5.3 Illustration of Stepwise Selection in Cox Proportional Hazard Model in "pbc" data

The summary of the cox model fitted using the covariates selected by the stepwise selection method is given below,

Interpretation of the summary



```
  n= 220, number of events= 91

                 coef exp(coef)  se(coef)      z Pr(>|z|)    
bili         0.085675  1.089452  0.024853  3.447 0.000566 ***
stage        0.513222  1.670666  0.165074  3.109 0.001877 ** 
albumin     -0.719946  0.486779  0.316298 -2.276 0.022836 *  
age          0.025602  1.025932  0.012316  2.079 0.037639 *  
edemaed_wth  0.903133  2.467321  0.440859  2.049 0.040504 *  
edeman_ed   -0.249085  0.779514  0.356931 -0.698 0.485271    
sexmale      0.479146  1.614695  0.307486  1.558 0.119170    
protime      0.275144  1.316720  0.116220  2.367 0.017911 *  
sgot         0.004052  1.004060  0.002110  1.920 0.054877 .  
drugplacebo  0.402019  1.494839  0.246709  1.630 0.103202    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

            exp(coef) exp(-coef) lower .95 upper .95
bili           1.0895     0.9179    1.0377    1.1438
stage          1.6707     0.5986    1.2089    2.3089
albumin        0.4868     2.0543    0.2619    0.9048
age            1.0259     0.9747    1.0015    1.0510
edemaed_wth    2.4673     0.4053    1.0398    5.8544
edeman_ed      0.7795     1.2829    0.3873    1.5691
sexmale        1.6147     0.6193    0.8838    2.9500
protime        1.3167     0.7595    1.0485    1.6536
sgot           1.0041     0.9960    0.9999    1.0082
drugplacebo    1.4948     0.6690    0.9217    2.4243

Concordance= 0.829  (se = 0.022 )
Likelihood ratio test= 118.9  on 10 df,   p=<2e-16
Wald test            = 124.3  on 10 df,   p=<2e-16
Score (logrank) test = 189.5  on 10 df,   p=<2e-16
```

Figure 5: Summary of stepwise cox regession

- All the three tests Likelihood ratio test, Wald test and Score (log rank) test has p-values < 0.05, which means that in the light of the given data it seems that all the coefficients of the corresponding predictors are not zero simultaneously

- The model yielded Concordance = 0.829 (se = 0.022), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 82.9% accuracy

- The concordance on the test data is found to be 0.846. That means given two random individuals of the test set, the model can predict who has gre ater risk of dying with 84.6% accuracy

- Also, the survival probabilities of the individuals of the test data set have been computed. The average standard error in the prediction is found to be 0.069

- Proportional hazard assumption of the model has been checked using Schoenfeld residuals test which provides a global p-value of 0.07099. Since this p-value exceeds the common significance threshold of 0.05, there is no significant evidence against the proportional hazards assumption

- After applying cross-validation we found mean concordance = 0.834 and mean standard error = 0.0294 for test data. As the covariates have multicollinearity among themselves, some predictors may result complexity in the model. So, by using stepwise selection procedure, an improvement in the concordance value is observed on the test data, where the selected covariates are smaller in number but appropriate and effective, with respect to the full model

6. Shrinkage Method(LASSO) in Cox Proportional Hazard Model

6.1 Shrinkage Method

This approach involves fitting a model involving all p predictors. However, the estimated coefficients are shrunken towards zero relative to the least squares estimates. This shrinkage (also known as regularization) has the effect of reducing variance. Depending on what type of shrinkage is performed, some of the coefficients may be estimated to be exactly zero. Hence, shrinkage methods can also perform variable selection. Consider the framework mentioned in deriving the partial likelihood in equation (3). Moreover, let $t_1 < t_2 < \ldots < t_m$ be the increasing list of unique failure times, and $j(i)$ denote the index of the observation failing at time t_i. Then another form of the partial likelihood mentioned in (3) can be written as,

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(x_{j(i)}^t \beta)}{\sum_{j \in R_i} \exp(x_j^t \beta)}$$

where R_i is the set of indices, j, with $y_j \geq t_i$ (those at risk at time t_i). Inference made with the partial likelihood ignores all information between failure times. For ease of notation the above formula assumes that the y_i are unique. By maximizing the partial likelihood, one can estimate β. For classical problems, with many more observations than predictors, the Cox model performs well. However, problems with $p > n$, lead to degenerate behaviour; to maximize the partial likelihood, all β_i are sent to $\pm\infty$. To combat this problem, Tibshirani (1997) proposed the use of an L_1 (LASSO: Least Absolute Shrinkage and Selection Operator) penalty in the Cox model. This both provides a well-defined solution, and a solution with few nonzero β_i. Even in the $n > p$ case, if p is sufficiently close to n, this may better estimate β than the unpenalized Cox model. Gui and Li (2005) developed an algorithm to fit this model using Newton Raphson approximations and the lasso path solution to the penalized least squares problem.

6.2 Basic Algorithm

Let X denote the design matrix, β the coefficient vector, and $\eta = X\beta$. Let $\dot{l}(\beta), \ddot{l}(\beta), l'(\beta)$ and $l''(\beta)$ denote the gradient and Hessian of the log-partial likelihood with respect to β and η respectively. A two-term Taylor series expansion of the log-partial likelihood centred at $\tilde{\beta}$ has the form,

$$l(\beta) \approx l(\tilde{\beta}) + (\beta - \tilde{\beta})^t \dot{l}(\tilde{\beta}) + \frac{(\beta - \tilde{\beta})^t \ddot{l}(\tilde{\beta})(\beta - \tilde{\beta})}{2}$$

$$= l(\tilde{\beta}) + (X\beta - \tilde{\eta})^t l'(\tilde{\eta}) + \frac{(X\beta - \tilde{\eta})^t l''(\tilde{\eta})(X\beta - \tilde{\eta})}{2}$$

where, $\tilde{\eta} = X\tilde{\beta}$. By algebraic calculations, one can have,

$$l(\beta) \approx \frac{(z(\tilde{\eta}) - X\beta)^t l''(\tilde{\eta})(z(\tilde{\eta}) - X\beta)}{2} + c(\tilde{\eta}, \tilde{\beta})$$

where, $z(\tilde{\eta}) = \tilde{\eta} - l''(\tilde{\eta})^{-1} l'(\tilde{\eta})$ and $c(\tilde{\eta}, \tilde{\beta})$ does not depend on β.

One difficulty arises in the computation of $l''(\tilde{\eta})$. Because this is a full matrix it would require computation of $O(n^2)$ entries. In order to speed up the algorithm, we instead replace $l''(\tilde{\eta})$ by a diagonal matrix with the diagonal entries of $l''(\tilde{\eta})$. We denote the i-th diagonal entry of $l''(\tilde{\eta})$ by $w(\tilde{\eta})_i$. Thus, the algorithm is,

1. Initialize $\tilde{\beta}$, and set $\tilde{\eta} = X\tilde{\beta}$.

2. Find $\hat{\beta}$ minimizing
$$\frac{1}{n}\sum_{i=1}^{n} w(\tilde{\eta})_i (z(\tilde{\eta})_i - x_i^t \beta)^2 + \lambda P_\alpha(\beta)$$

3. Set $\tilde{\beta} = \hat{\beta}$ and $\tilde{\eta} = X\hat{\beta}$

4. Repeat steps 2-4 until convergence of $\hat{\beta}$

The minimization in step 2 is done by cyclical coordinate descent.

6.3 Finding Optimal λ

For choosing λ, the method of cross-validation has been used. For this, let the data be split into k parts. Then the goodness of fit estimate for a given part i and λ is,

$$\hat{CV}_i(\lambda) = l(\beta_{-i}(\lambda)) - l_{-i}(\beta_{-i}(\lambda))$$

where l_{-i} is the log-partial likelihood excluding part i of the data, and $\beta_{-i}(\lambda)$ is the optimal β for the non-left out data, found from maximizing $l_{-i} + \lambda\|\beta\|_1$. Our total goodness of fit estimate, $\hat{CV}(\lambda)$, is the sum of all $\hat{CV}_i(\lambda)$. Then, choose the value of λ which maximizes $\hat{CV}(\lambda)$.

6.4 Illustration of Shrinkage Method(LASSO) in Cox Model through "pbc" data

Defining the training set and test set:
80% of the 144 individuals have been selected at random as the members of the training set along with the corresponding covariates and the rest of the 20% of the observations have been selected as the members of the test set.

Then by using cross validation method mentioned above, the best λ for which the partial likelihood deviance (which is treated as the cross-validation error) is minimum, is found to be 0.0636 The plot of the cross-validated error rates is given below:

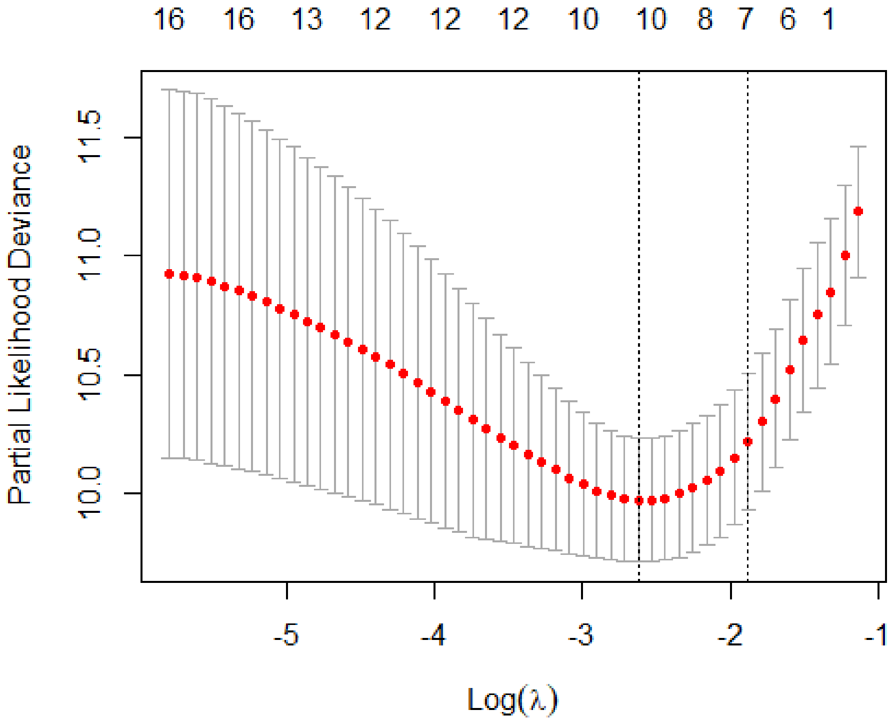


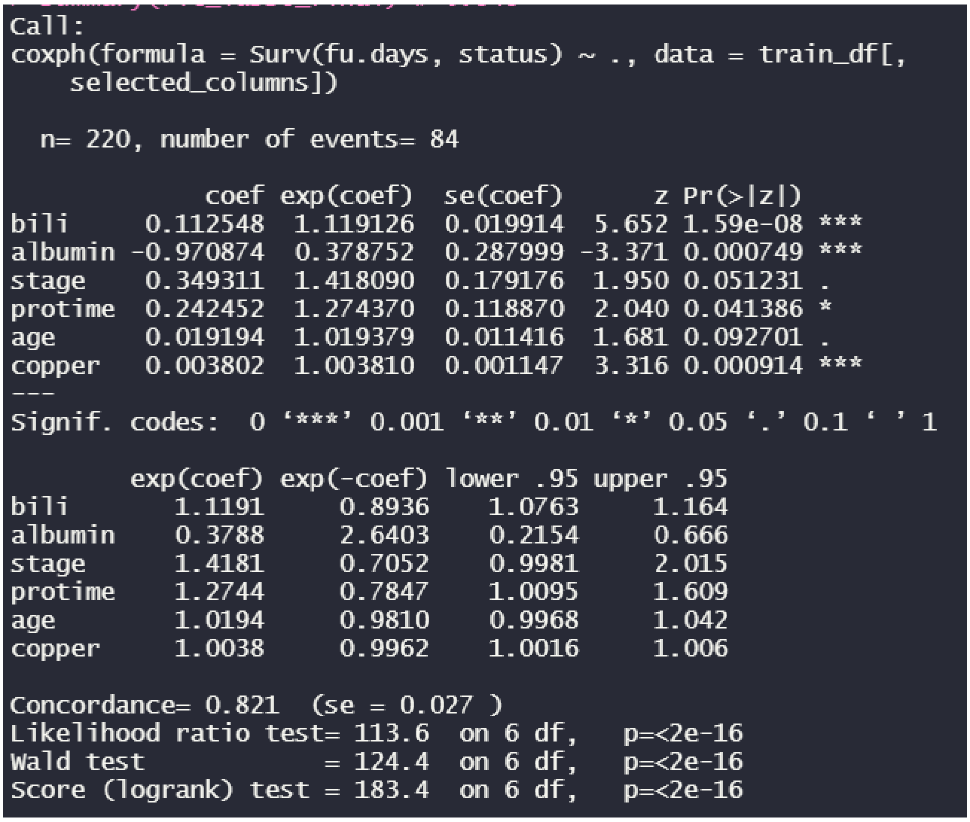
Figure 6: Plot of the cross-validated error rates

Each dot represents a value along our path, with error bars to give a confidence interval for the cross-validated error rate. The left vertical bar indicates the minimum error while the right shows the largest value of such that the error is within one standard deviation of the minimum. The top of the plot gives the size of each model. After this, a penalized cox proportional hazard model with Lasso penalty and $\lambda = 0.05552002$ has been fitted.

Only 10 covariates have been selected by the model which are shown below:

"bili", "albumin", "stage", "protime", "age", "hepatom", "ascites", "edemaed$_{w}th$”,”$copper$”

Rest of the covariates have been ignored by the model i.e., they got zero as their corresponding coefficients. So, considering only these mentioned covariates which have been selected by the Lasso regression, a new Cox proportional hazard model has been fitted, whose summary is shown below:



```
Call:
coxph(formula = Surv(fu.days, status) ~ ., data = train_df[, 
    selected_columns])

  n= 220, number of events= 84 

             coef exp(coef)  se(coef)      z Pr(>|z|)    
bili     0.112548  1.119126  0.019914  5.652 1.59e-08 ***
albumin -0.970874  0.378752  0.287999 -3.371 0.000749 ***
stage    0.349311  1.418090  0.179176  1.950 0.051231 .  
protime  0.242452  1.274370  0.118870  2.040 0.041386 *  
age      0.019194  1.019379  0.011416  1.681 0.092701 .  
copper   0.003802  1.003810  0.001147  3.316 0.000914 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
bili       1.1191     0.8936    1.0763     1.164
albumin    0.3788     2.6403    0.2154     0.666
stage      1.4181     0.7052    0.9981     2.015
protime    1.2744     0.7847    1.0095     1.609
age        1.0194     0.9810    0.9968     1.042
copper     1.0038     0.9962    1.0016     1.006

Concordance= 0.821  (se = 0.027 )
Likelihood ratio test= 113.6  on 6 df,   p=<2e-16
Wald test            = 124.4  on 6 df,   p=<2e-16
Score (logrank) test = 183.4  on 6 df,   p=<2e-16
```

Figure 7: Summary of lasso with cox regression

6.5 Interpretation of the summary:

- All the three tests Likelihood ratio test, Wald test and Score (log rank) test has p-values < 0.05, which means that in the light of the given data it seems that all the coefficients of the corresponding predictors are not zero simultaneously.

- The model yielded Concordance = 0.821 (se = 0.027), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 82.1% accuracy.

- The concordance on the test data after cross validation is found to be 0.869. That means given two random individuals of the test set, the model can predict who has gre ater risk of dying with 86.9% accuracy.

- Also, the survival probabilities of the individuals of the test data set have been computed. The average standard error in the prediction is found to be 0.06109.
- As the covariates have multicollinearity among themselves, some predictors may result complexity in the model. So, by using penalized Cox proportional hazard regression procedure, an improvement in the concordance value is observed on the test data, where the selected covariates are smaller in number but appropriate and effective, with respect to the full model. The average standard error in the prediction is also decreased.

7. Factor Analysis of Mixed Data (FAMD)

7.1 Theory

Factor Analysis of Mixed Data (FAMD) is a statistical technique designed to analyze datasets containing both continuous and categorical variables. FAMD extends the principles of traditional factor analysis and principal component analysis to accommodate the different scales of measurement inherent in mixed data types. By identifying underlying structures and reducing dimensionality, FAMD provides a way to uncover relationships and patterns among variables, making it useful in fields like social sciences and bioinformatics.

Some Notation:

Let we have I individuals. Each individual i is attributed a weight p_i such as

$$\sum_i p_i = 1.$$

To simplify matters, except when explicitly stated, we suppose that the individuals are of the same weight, thus $p_i = \frac{1}{I} \forall i$. These individuals are described by:

- K_1 quantitative variables $\{k = 1, K_1\}$; these variables are standardised (centred and reduced); this is not merely for convenience but is necessary due to the presence of

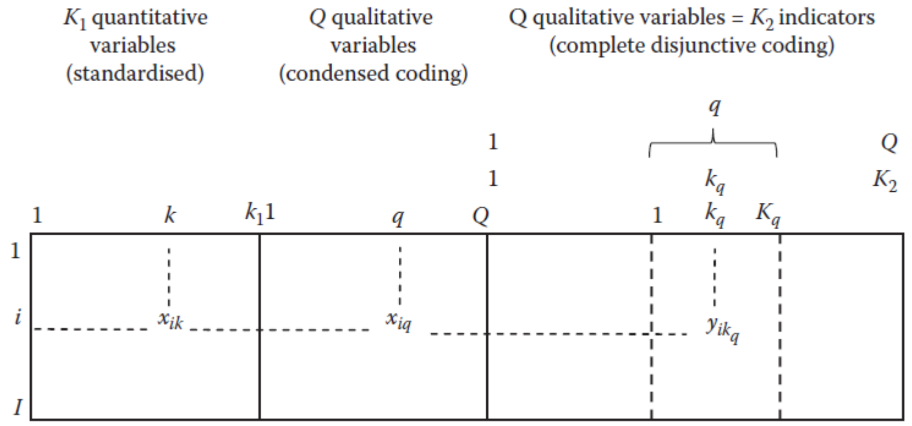

K_1 quantitative variables (standardised)
Q qualitative variables (condensed coding)
Q qualitative variables = K_2 indicators (complete disjunctive coding)
q
1
1
Q
k_q
K_2
1
k
$k_1$1
q
Q
1
k_q
K_q
1
i
I
x_{ik}
x_{iq}
y_{ik_q}

Figure 8: Data structure and principal notations: x_{ik} : Value of i for variable (centred-reduced) k; x_{iq} : Category of i for variable q; y_{ikq}: 1 if i possesses k_q of variable q and otherwise 0.

two types of variables.

- Q qualitative variables $\{q = 1, Q\}$; the qth variable presents K_q categories $\{k_q = 1, K_q\}$; the overall number of categories is $\sum_q K_q = K_2$; we denote p_{k_q} the proportion of individuals possessing category k_q.

Let $K = K_1 + K_2$ be the total number of quantitative variables and indicator variables.

These notations can be brought together in the table in Figure 8 in which the qualitative variables appear both in their condensed form and in their complete disjunctive form.

Representing Variables:

Let R_I be the space of functions on I. This space is endowed with the diagonal metric of the weights of the individuals, denoted D:

$$D(i,j) = \begin{cases} 0 & \text{if } j \neq i \\ p_i & \text{if } j = i \end{cases}$$

Generally, the individuals have the same weights: $D = \left(\frac{1}{I}\right) I_d$ (where I_d is the identity matrix of appropriate dimensions).

As in standardised PCA, the quantitative variables are represented by vectors with a length of 1.

As in MCA, variable q is represented by cloud N_q of its centred indicators K_q. This cloud generates subspace E_q of dimension $K_q - 1$; E_q is the set of centred functions constant on the classes of the partition defined by q.

For N_q to possess the same inertial properties as in an MCA, if we perform an unstandardised PCA on it, the indicator k_q must be divided by p_{kq} and Attributed a weight p_{kq} (Obtaining the exact inertia of the MCA requires the weight $\frac{p_{kq}}{J}$.

Specifically, by proceeding in this way, we obtain a fundamental property of MCA: the projected inertia of N_q on a centred variable y is equal to the squared correlation ratio $\eta^2(q, y)$ between q and y.

When looking for direction v of R_I which maximises the projected inertia of cloud N_K (made up of the quantitative variables and the indicators), we maximise the criterion:

$$\sum_{k \in K_1} \eta^2(k, v) + \sum_{q \in Q} \eta^2(q, v)$$

This is the starting point of the method put forward by Gilbert Saporta in 1990. Geometrically, as variables k are standardised, the projection coordinate of variable k on v is worth $\cos^2(\theta_{kv}) = \eta^2(k, v)$, where θ_{kv} is the angle between vectors k and v. Similarly, as v is centred, $\eta^2(q, y) = \cos^2(\theta_{qv})$ where θ_{qv} is the angle between v and its projection on E_q. The criterion is thus expressed:

$$\sum_{k \in K_1} \cos^2 \theta_{kv} + \sum_{q \in Q} \cos^2 \theta_{qv}.$$

This is the starting point of the method put forward by Brigitte Escofier in 1979.

The influence of a variable must be explained according to the dimension of the subspace it generates. Thus, in space R_I: - A quantitative variable is represented by a vector associated with an inertia of 1. - A qualitative variable with K_q categories is represented by K_q vectors generating a subspace E_q of dimension $K_q - 1$, all of which are associated with an inertia of $K_q - 1$.

As in MCA, the total inertia of a qualitative variable increases with the number of categories. However, when projected onto any dimension of E_q, this inertia is normalized to 1. Thus, when searching for directions of maximum inertia, these two types of variables are balanced, which is highlighted by one or another of the two expressions of the criterion below.

7.2 Geometrical Interpretation:

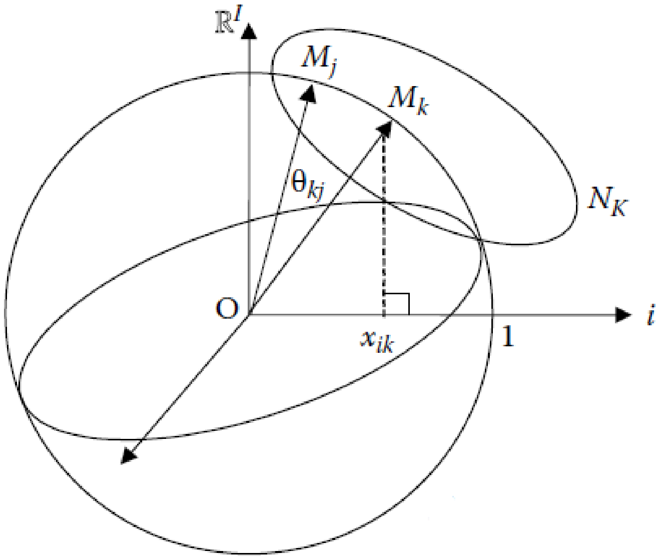


Figure 9: The cloud of variables: centred and reduced data. kj is the angle formed by the two vectors representing variables k and j ($\overrightarrow{OM_k}$ and $\overrightarrow{OM_j}$)

Cloud of Variables N_K:

To variable k, we attribute its values for all of the individuals studied $\{x_{ik}; i = 1, I\}$. This set corresponds to point M_k (and to vector v_k) in space R_I, in which each dimension corresponds to an individual. R_I is called the variables' space or, more generally, the space of functions on I (a function on I attributes a numerical value to each individual i). The set of points M_k constitutes the cloud of variables denoted N_K (Figure 9).

When the variables are centred, as is always the case in PCA, this space has two remarkable properties:

1. The cosine of the angle θ_{kj} formed by the two variables k and j is equal to their correlation coefficient. This geometrical interpretation of the correlation coefficient

justifies the use of this space to study the relationships between variables. It also explains that we represent variable k by the vector linking the origin to point M_k.

2. The distance between M_k and O is equal to the variance of variable k. Here we are interpreting variance as a squared length. A centred-reduced variable has a length of 1; cloud N_K is then situated on a hypersphere (with a radius of 1).

To obtain these two properties, it is necessary, when calculating a distance in R_I, to attribute to each dimension i the weight p_i of the corresponding individual. Thus, we obtain:

$$d^2(O, M_k) = \sum_i p_i (x_{ik} - \bar{x}_k)^2 = \mathrm{Var}[k].$$

The total inertia of cloud N_K in R_I, with respect to the origin O, is easy to calculate as the variables all have an equal weight of 1.

$$\mathrm{Inertia}(N_K/O) = \sum_k 1 \cdot d^2(O, M_k) = \sum_k \mathrm{Var}[k].$$

This total inertia is equal to those of cloud N_I in R_K: the number K of variables in the centred-reduced case.

Now, geometrically, as variables k are standardised, the projection coordinate of variable k on v is worth $\cos(\theta_{kv}) = r(k, v)$, where θ_{kv} is the angle between vectors k and v. Similarly, as v is centred, $\eta^2(q, y) = \cos^2(\theta_{qv})$ where θ_{qv} is the angle between v and its projection on E_q.

So, total variance captured by quantitative variable is

$$\sum_{k \in K_1} \cos^2 \theta_{kv}.$$

For qualitative variable with K_q categories is represented by K_q vectors generating a subspace E_q of dimension $K_q - 1$, all of which are associated with an inertia of $K_q - 1$. Similarly, as v is centred, $\eta^2(q, y) = \cos^2(\theta_{qv})$ where θ_{qv} is the angle between v and its projection on E_q.

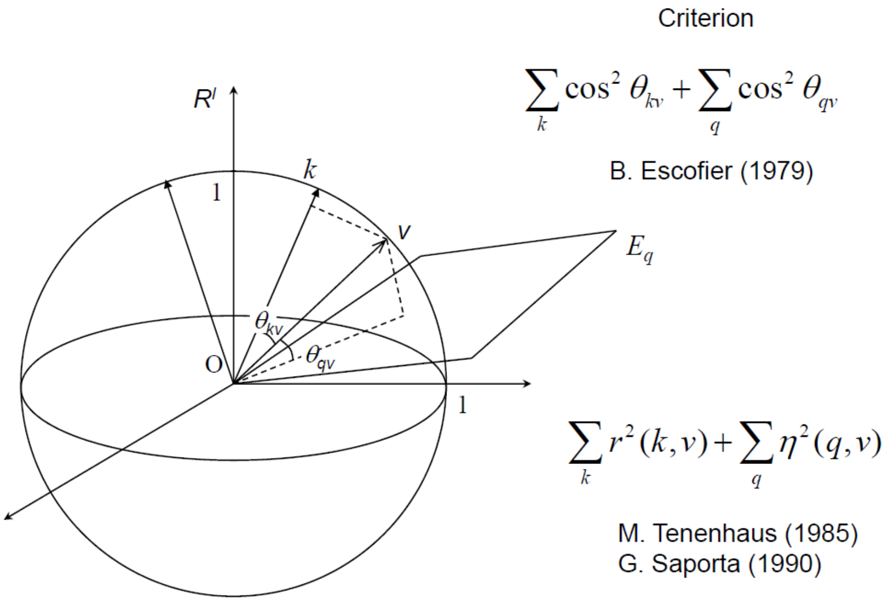


Figure 10: Geometric Interpretation of FAMD

So, total variance captured by quantitative variable is

$$\sum_{q \in Q} \cos^2 \theta_{qv}.$$

When looking for direction v of R_I which maximises the projected inertia of cloud N_K (made up of the quantitative variables and the indicators), we maximise the criterion:

$$\sum_{k \in K_1} \cos^2 \theta_{kv} + \sum_{q \in Q} \cos^2 \theta_{qv}.$$

7.3 Illustration of FAMD on "pbc" dataset

- We have extracted the first 5 principal components using "factomineR" package in Rstudio from this data and found first 5 pc's can explain 53.892% of the variance in the data. Here's a Scree plot indicating the % variance explained by each component.

- Top contributing variables to the first few PCs can provide insights into which variables underlie variations in the dataset, and may help with feature selection for

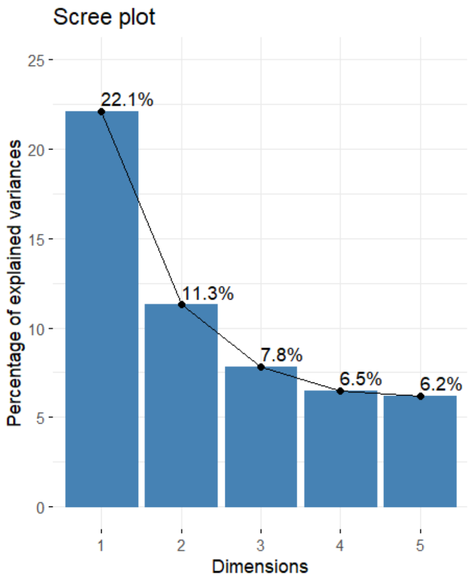


Figure 11: % of variance explained by PC's

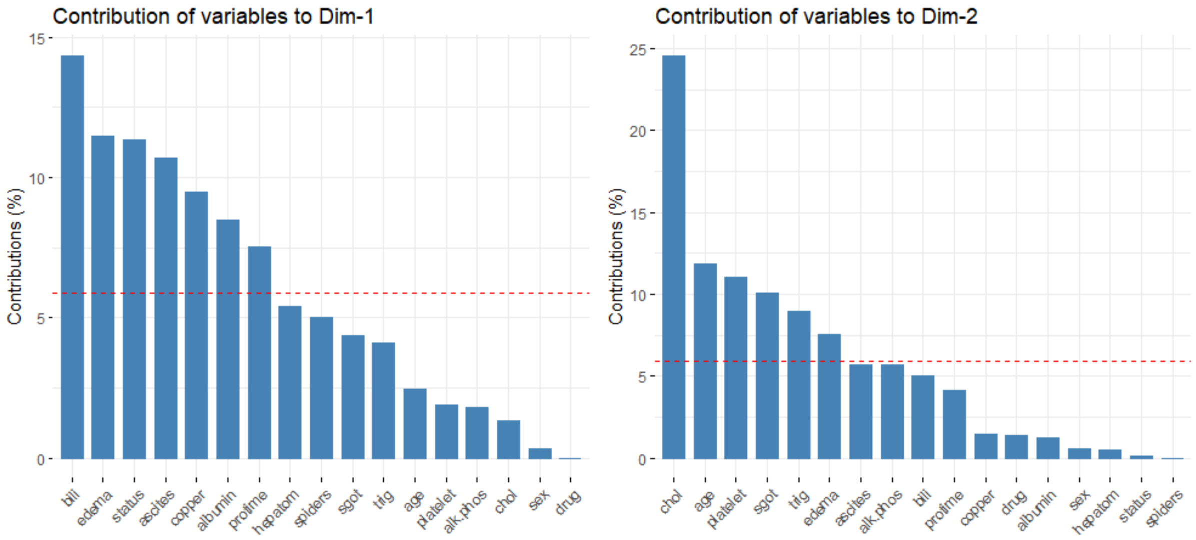


Figure 12: Variable Contributions Of Different PC's

downstream analyses. The red dashed line indicates the expected average contribution (100% contribution divided the total number of variables avaiable in the dataset). So variables meeting the cut-off would be considered as important in contributing to the PD.

From the variables that meet the cut-off, we can glean some insights into what are the most important variables in this dataset, such as bili, edema and chol. So, FAMD can also be a handy tool for variable selection.

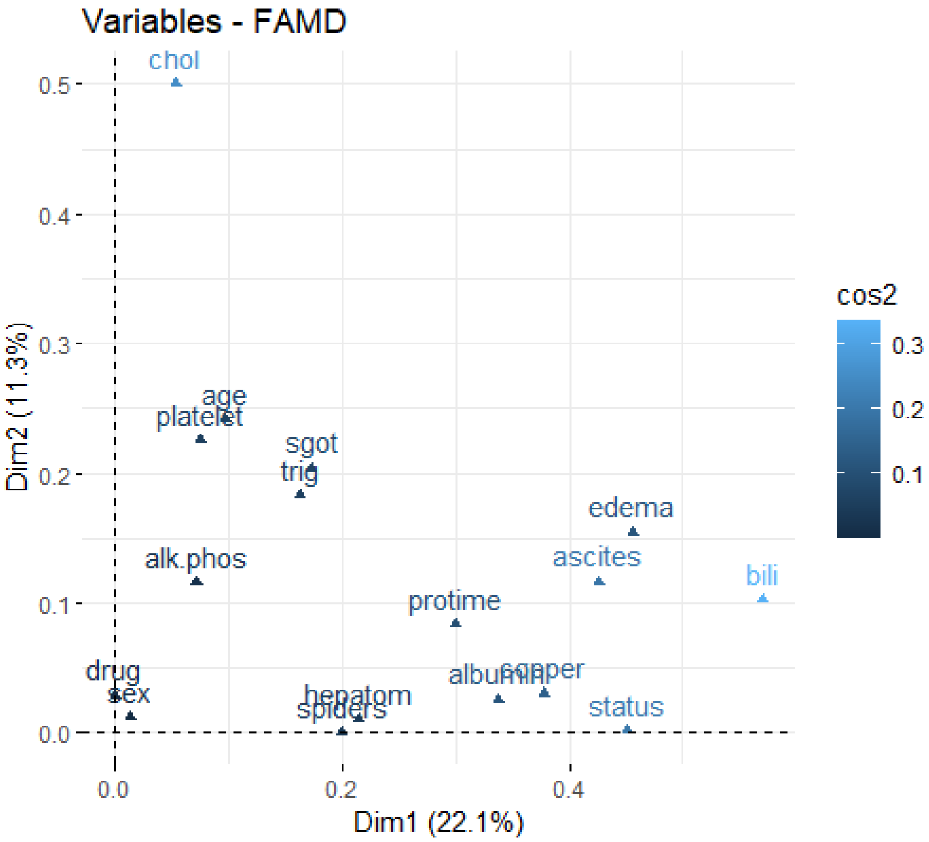


Figure 13: Squared Loading Plot

- Squared loading plots allow us to visualize qualitative and quantitative variables together in the new feature space. The coordinates here are to be interpreted as measuring the links between variables and principal components. This may be interpreted as the coordinates of each variable being the absolute value its squared loading.

  The factor loading of a variable describes the correlation, i.e. information shared, between it and a given PD. By squaring the factor loading for a variable, we also get its squared loading (which you may see also called squared cosine or cos2). This provides a measure of the proportion of variance in a variable that is captured by a particular PD. For each variable, the sum of its squared loading across all PDs equals to 1.

  From the above plot, we can see that variable ascites is more closely correlated with PD1 than with PD2, whereas trig or sgot is described by a more even combination of PD1 and PD2. Being furthest from the origin, the variables chol, bili have the highest squared loading values and so are more important in explaining the variance captured by PD1 and PD2 than variables clustered near the origin, such as sex, drug.

7.4 Cox Regression with Principal Components from FAMD

Now, after extracting the principal components we will implement cox proportional hazard model using these principal components. Here, we can see from the plot that, cox regression with more no of principal components give better concordance value and after a certain no of PC's the concordance value is constant. That is probably because of additional PC's can't capture more variance from the data. We also have to check proportional hazard assumption for that model.

Based on the concordance value and model assumption we have select first 14 principal components and applied cox regression mmodel with proper model assumptions.

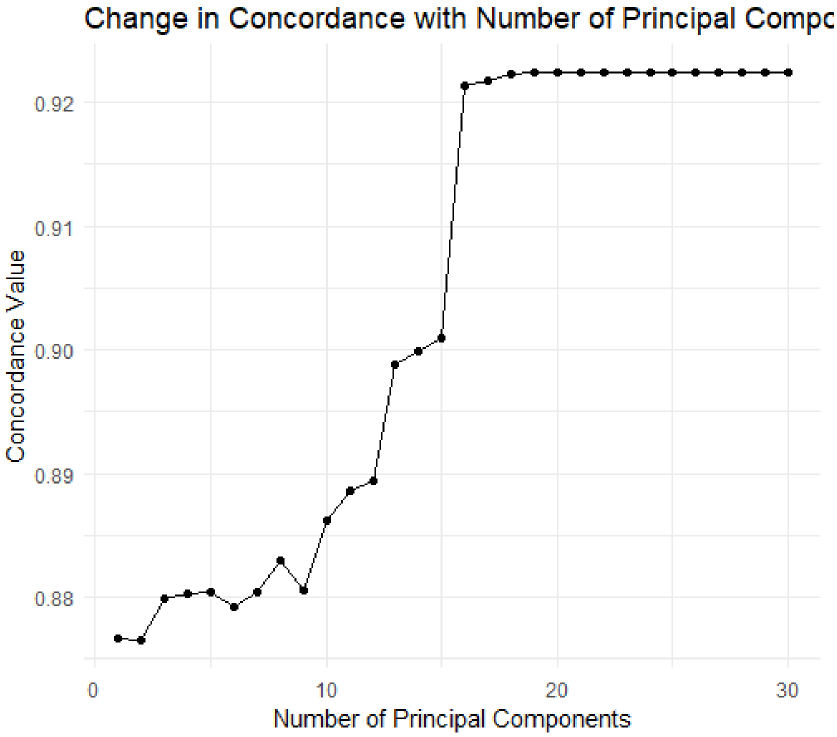
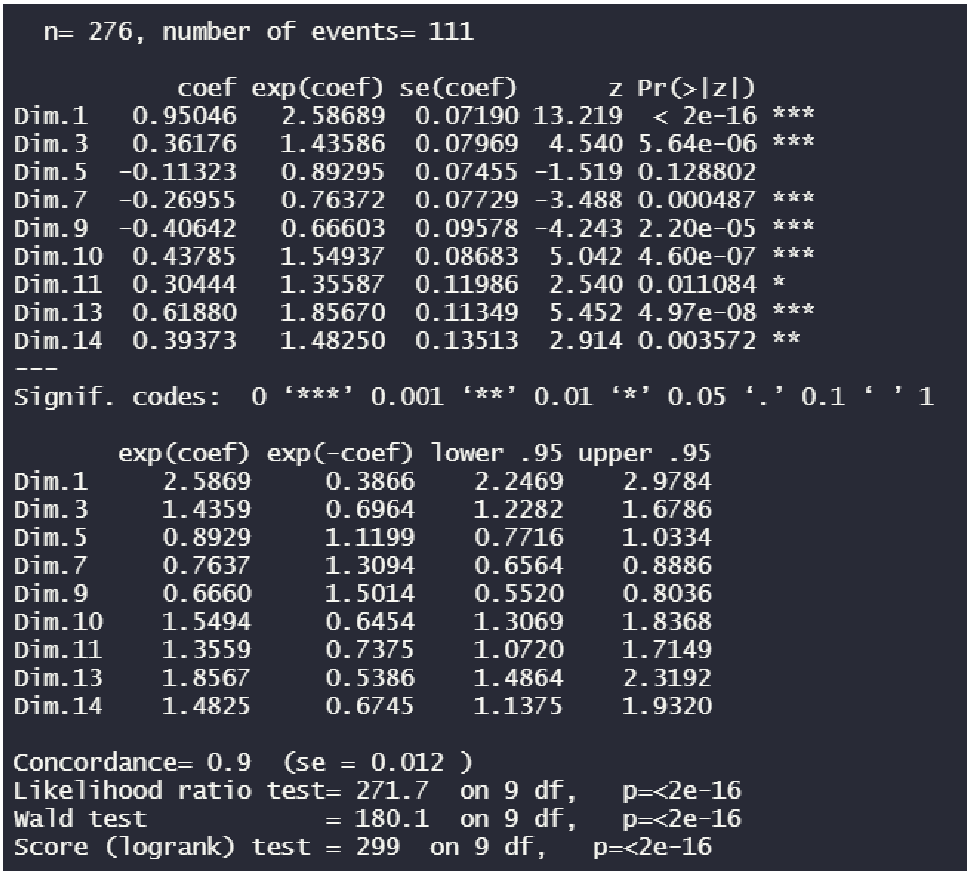


Figure 14: No of principal component vs concordance value in cox model

The summary of the fitted model is given below:

- All the three tests Likelihood ratio test, Wald test and Score (log rank) test has



```
  n= 276, number of events= 111

           coef exp(coef) se(coef)      z Pr(>|z|)    
Dim.1   0.95046   2.58689  0.07190 13.219  < 2e-16 ***
Dim.3   0.36176   1.43586  0.07969  4.540 5.64e-06 ***
Dim.5  -0.11323   0.89295  0.07455 -1.519 0.128802    
Dim.7  -0.26955   0.76372  0.07729 -3.488 0.000487 ***
Dim.9  -0.40642   0.66603  0.09578 -4.243 2.20e-05 ***
Dim.10  0.43785   1.54937  0.08683  5.042 4.60e-07 ***
Dim.11  0.30444   1.35587  0.11986  2.540 0.011084 *  
Dim.13  0.61880   1.85670  0.11349  5.452 4.97e-08 ***
Dim.14  0.39373   1.48250  0.13513  2.914 0.003572 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
Dim.1     2.5869     0.3866    2.2469    2.9784
Dim.3     1.4359     0.6964    1.2282    1.6786
Dim.5     0.8929     1.1199    0.7716    1.0334
Dim.7     0.7637     1.3094    0.6564    0.8886
Dim.9     0.6660     1.5014    0.5520    0.8036
Dim.10    1.5494     0.6454    1.3069    1.8368
Dim.11    1.3559     0.7375    1.0720    1.7149
Dim.13    1.8567     0.5386    1.4864    2.3192
Dim.14    1.4825     0.6745    1.1375    1.9320

Concordance= 0.9  (se = 0.012 )
Likelihood ratio test= 271.7  on 9 df,   p=<2e-16
Wald test            = 180.1  on 9 df,   p=<2e-16
Score (logrank) test = 299  on 9 df,   p=<2e-16
```

Figure 15: Summary of fitted cox model with PC's

p-values < 0.05, which means that in the light of the given data it seems that all the coefficients of the corresponding predictors are not zero simultaneously.

- The model yielded Concordance = 0.9 (se = 0.012), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 90% accuracy which is better than the previous models.

- The concordance on the test data after cross validation is found to be 0.8932. That means given two random individuals of the test set, the model can predict who has gre ater risk of dying with 89.32% accuracy means it working well on unknown data also.

- Also, the survival probabilities of the individuals of the test data set have been computed. The average standard error in the prediction is found to be 0.0574.

8. Conclusions

By the analysis of "pbc" data and from fitted Cox's PH model, we can conclude the following.

From the stepwise selection model, the p-values of covariates drug, sgot, sex, edema are > 0.05. Thus, the three covariates do not exist significant effect obviously for the patients of survival time. Here, we are interested in the impact of D-penicillamine (drug$=$ 1) to lifetime of patients with PBC data. Therefore we would still consider their effect for the patients' lifetime in model.

Next, the relative risk for Cox's PH model is considered and it is defined as

$$\frac{\hat{h}(t|Z_i = 1)}{\hat{h}(t|Z_i = 0)} = \frac{h_0(t)\exp\{\hat{\beta}_i\}}{h_0(t)} = \exp\{\hat{\beta}_i\}.$$

So the relative risk for the patients using placebo as compared to using D-penicillamine is exp0.402 $=$ 1.4948 that is, the patients using placebbo would have 1.498 times death risk as using D-penicillamine. Although the effect of drug is not significantly, patients given D-penicillamine would have longer lifetime than patients given placebo. Similarly, if bilirubin is equal to x $+$ 1 and x then its relative risk is equal to exp0.0856 $=$ 1.0894 which means that if the bilirubin increases one milliliter then the individual's death rate would increase 1.0894 times. And the normal bilirubin rage is between 0.2 and 1.2 milliliter in the medical science. Consequently, the bilirubin is larger than 20 milliliter then the individual would fall sick of serious illness, such as Liver Cirrhosis (L.C. for short). Thus, if the patient has high value of bilirubin then whether given Dpenicillamine or not would have a small inference even if non-impact to lifetime of patients. Also, the covariates drug(D-penicillamine), edema, albumin would make death rate decreasing, i.e., these covariates would cause longer survival time. And the rest would reduce patients' lifetime.

Using factor analysis, we can say that bilirubin, edema and cholesterol are the most important features to survival time of the patients. Among all the COX's PH models(stepwise selection, lasso), we can see that model using the principal components gives us the maximum concordance value of .8932 in test data. So, this model is giving higher accuracy in predicting survival probablity among the patients.

9. References

- **Tibshirani R (1996)**. "Regression Shrinkage and Selection via the Lasso." Journal of the Royal Statistical Society B, 58, 267–288.
- **Tibshirani R (1997)**. "The Lasso Method for Variable Selection in the Cox Model." Statistics in Medicine, 16, 385–395.
- **Gui J, Li H (2005)**. "Penalized Cox Regression Analysis in the High-Dimensional and Low Sample Size Settings, with Applications to Microarray Gene Expression Data." Bioinformatics, 25(13), 3001–2008.
- An Introduction to Statistical Learning with Applications in R, By Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani.
- Variable Selection For Cox's Proportional Hazard Model And Frailty Mod el; By JIANQING FAN1 and RUNZE LI ,Chinese University of Hong Ko ng and Pennsylvania State University
- Variable selection techniques for the Cox proportional hazards model: A comparative study, By Simon Petersson and Klas Sehlstedt, University of Gothenburg School of Bussiness, Economics and Law 2018-02-21.
- An Introduction to Statistical Learning with Applications in R, By Gareth James • Daniela Witten • Trevor Hastie • Robert Tibshirani.
- Survival Analysis Techniques for Censored and Truncated Data, By John P. Klein and Melvin L. Moeschberger.
- Multiple Factor Analysis by Example Using R
- Principal component analysis
- Using Time Dependent Covariates and Time Dependent Coefficients in the Cox Model
- GitHub Link for R (software) Scripts