APPLICATION OF LASSO IN COX PROPORTIONAL HAZARD MODEL IN PRESENCE OF MULTICOLLINEARITY

NAME: ROHIT DUTTA

ROLL: 221396

PROGRAMME: M.Sc. STATISTICS (1ST YEAR)

DEPARTMENT: MATHEMATICS & STATISTICS

INSTITUTE: INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

SUPERVISOR:

PROF. BISWABRATA PRADHAN (H.O.D),

SQC & OR UNIT,

INDIAN STATISTICAL INSTITUTE, KOLKATA

Acknowledgement

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 – "**Application LASSO in Cox Proportional Hazard Model in Presence of Multicollinearity**." 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 the **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.

CONTENT

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**, 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.

A common feature of these data sets is they contain either censored or truncated observations. Censored data arises when an individual's life length is known to occur only in a certain period. Possible censoring schemes are **right censoring,** where all that is known is that the individual is still alive at a given time, **left censoring** when all that is known is that the individual has experienced the event of interest prior to the start of the study, or **interval censoring**, where the only information is that the event occurs within some interval.

Survival analysis is a very well-studied topic within statistics, due to its critical importance in a variety of applications, both in and out of medicine.

Objective:

The objective of this project is to analyse the performance of Cox proportional hazard model by applying LASSO in the presence of multicollinearity.

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

nonnegative 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 (mrl)** at time x, which is the mean time to the event of interest, given the event has not occurred at x.

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.

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,

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

If X is a continuous random variable, then,

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

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., $h(x) \geq 0$.

Censoring:

Time-to-event data present themselves in different ways which create special problems in analysing such data. One peculiar feature, often present in time-to-event data, is known as censoring, which, broadly speaking, occurs when some lifetimes are known to have occurred only within certain intervals. The remainder of the lifetimes are known exactly. There are various categories of censoring, such as right censoring, left censoring, and interval censoring. Mainly, right censoring has been used in this project.

Right Censoring:

First, 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, it is convenient to use the following notation. For a specific individual under study, we assume that there is a lifetime X and a fixed censoring time, C (C for "right" censoring time). The X's are assumed to be independent and identically distributed with probability density function $f(x)$ and survival function S (x). The exact lifetime X of an individual will be known if, and only if, X is less than or equal to C. If X is greater than C, the individual is a survivor, and his or her event time is censored at Cr. The data from this experiment can be conveniently represented by pairs of random variables (T, δ), where δ indicates whether the lifetime X corresponds to an event (δ = 1) or is censored (δ = 0), and T is equal to X, if the lifetime is observed, and to Cr if it is censored, i.e., T = min(X, C).

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 (δ_j= 1, if the event has occurred and δ_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,...,p, are the covariates for the j-th individual.

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

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

The intuition behind 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,

$$\boldsymbol{L} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \mathrm{f}(y_i)^{\delta_i} \mathrm{S}(y_i)^{\mathbf{1}-\delta_i} = \prod_{\boldsymbol{i=1}}^{\boldsymbol{n}} \mathrm{h}(y_i)^{\delta_i} \mathrm{S}(y_i) \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

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 $h(t|x_i) = \exp(\beta_0 + \sum_{j=1}^{p} \beta_i x_{ij})$, 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 $\beta = (\beta_1, \ldots, \beta_p)^t$.

The proportional hazards assumption states that,

$$h(t|x_i) = h_0(\mathrm{t}) \exp\left(\sum_{j=1}^{p} \beta_i x_{ij}\right) \ldots\ldots\ldots\ldots\ldots\ldots (2),$$

where $h_0(\mathrm{t}) \geq 0$ 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 $h_0(\mathrm{t})$ times the factor $\exp\left(\sum_{j=0}^{p} \beta_j x_{ij}\right)$. The quantity $\exp\left(\sum_{j=1}^{p} \beta_i x_{ij}\right)$ is called the relative risk for the feature vector $x_j = (x_{j1}, \ldots, x_{jp})^t$, relative to that for the feature vector $x_j = (0, \ldots, 0)^t$.

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 $exp(\beta_j)$.

Introduction to Cox's Proportional Hazard Model and Partial Likelihood to Estimate $\boldsymbol{\beta}$:

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

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 $h(y_i|x_i) = h_0(y_i)\exp\left(\sum_{j=1}^{p}\beta_j x_{ij}\right)$, and the total hazard at time y_i for the at-risk observations is,

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

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,

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

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

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

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)

Testing of Hypothesis:

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

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

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

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

The third test is the scores test. It is based on the efficient scores, $U(\beta) = (U_1(\beta), \ldots, U_p(\beta))^t$, 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(β) when H_0 is true. Thus, a test of H_0: $\beta = \beta_0$ is,

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

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

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, $\hat{\eta}_i = \widehat{\beta_1} x_{i1} + \cdots + \widehat{\beta_p} x_{ip}$, 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,..., t_n; instead, we observe the (possibly-censored) times y_1,..., y_1, as well as the censoring indicators δ_1,..., δ_n.

Therefore, Harrell's concordance index (or C-index) computes the pro- Harrell's concordance index portion of observation pairs for which $\widehat{\eta_{i'}} > \widehat{\eta_i}$and $y_i > y_{i'}$:

$$C = \frac{\sum_{i,i':\, y_i > y_{i'}} I(\widehat{\eta_{i'}} > \widehat{\eta_i})\delta_{i'}}{\sum_{i,i':\, y_i > y_{i'}} \delta_{i'}}$$

where the indicator variable $I(\widehat{\eta_{i'}} > \widehat{\eta_i})$ equals one if $\widehat{\eta_{i'}} > \widehat{\eta_i}$, 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 $t_i > t_{i'}$.

Illustration of Fitting a Cox Proportional Hazard Model in a Data:

Description of the data: The data set named "**nki70**", taken from Netherlands Cancer Institute, consists of 144 individuals where every one of them is lymph node positive breast cancer patients. The data set contains information on the following variables,

- Time: Metastasis-free follow-up time (months)
- Event: Event indicator, 1 = metastasis or death; 0 = censored
- Diam: Diameter of the tumour (two levels)
- N: Number of affected lymph nodes (two levels)
- ER: Estrogen receptor status (two levels)
- Grade: Grade of the tumour (three ordered levels)
- Age: Patient age at diagnosis (years)

Also, the data contains gene expression measurements for 70 prognostic genes for each of the individuals.

Data manipulation for fitting Cox Proportional Hazard Model:

In the data frame, the covariates named "grade", "diam", "N" and "ER", have multiple levels. So, for fitting the Cox model, first we need to make them numeric.

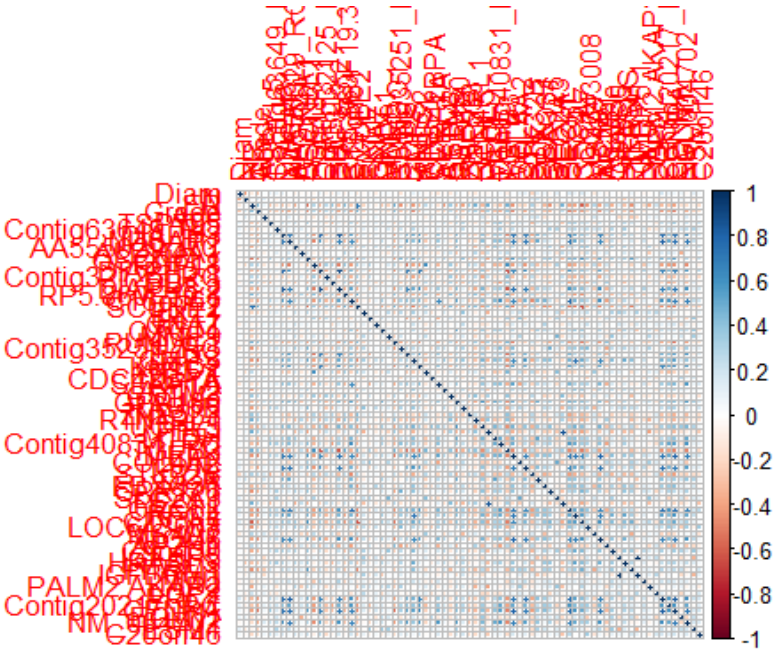
The three ordered levels of the covariate "grade", namely Poorly diff < Intermediate < Well diff, has been assigned the values "1", "2" and "3" respectively.

The two levels of the covariate "diam", namely "<=2cm" and ">2cm", has been assigned the values "1", and "2" respectively. The two levels of the covariate "N", namely "1-3" and ">=4", has been assigned the values "2", and "1" respectively.

The two levels of the covariate "ER", namely "Negative" and "Positive" has been assigned the values "1", and "2" respectively.

Detection of Multicollinearity by Graphical Display of Correlation Matrix:

A graphical display of the correlation matrix is given below, from which it is observed that there are number of predictors having high correlation among them.



The pairs of predictors who have correlation greater than 0.5 are shown below:

predictor1	predictor2	predictor1	predictor2	predictor1	predictor2
DIAPH3	NUSAP1	FGF18	TGFB3	C16orf61	ORC6L
DIAPH3	DIAPH3.1	RP5.860F19.3	TGFB3	KNTC2	ORC6L
NUSAP1	DIAPH3.1	DIAPH3	MELK	Contig40831_RC	ORC6L
DIAPH3	DIAPH3.2	NUSAP1	MELK	MELK	ORC6L
NUSAP1	DIAPH3.2	QSCN6L1	MELK	DTL	ORC6L
DIAPH3.1	DIAPH3.2	DIAPH3.1	MELK	DIAPH3	RFC4
DIAPH3	C16orf61	DIAPH3.2	MELK	GMPS	RFC4
NUSAP1	C16orf61	C16orf61	MELK	MELK	RFC4
DIAPH3.2	C16orf61	GMPS	MELK	DTL	RFC4
ER	SCUBE2	KNTC2	MELK	ORC6L	RFC4
NUSAP1	ECT2	SERF1A	MELK	QSCN6L1	CDCA7
DIAPH3.1	ECT2	DIAPH3	DTL	MELK	CDCA7
DIAPH3.2	ECT2	NUSAP1	DTL	RFC4	CDCA7
DIAPH3	GMPS	DIAPH3.1	DTL	DIAPH3	MCM6
DIAPH3	KNTC2	DIAPH3.2	DTL	NUSAP1	MCM6
NUSAP1	KNTC2	GMPS	DTL	DIAPH3.1	MCM6
DIAPH3.1	KNTC2	KNTC2	DTL	DIAPH3.2	MCM6
DIAPH3.2	KNTC2	MELK	DTL	C16orf61	MCM6
ECT2	KNTC2	NUSAP1	DCK	GMPS	MCM6
FGF18	WISP1	GPR180	DCK	KNTC2	MCM6
C16orf61	SERF1A	MTDH	DCK	MELK	MCM6
C16orf61	GPR180	PECI	PECI.1	DTL	MCM6
OXCT1	GPR180	DIAPH3	ORC6L	ORC6L	MCM6
GPR180	UCHL5	NUSAP1	ORC6L	RFC4	MCM6
GPR180	MTDH	DIAPH3.1	ORC6L	CDCA7	MCM6
C16orf61	Contig40831_RC	DIAPH3.2	ORC6L	RFC4	HRASLS

predictor1	predictor2	predictor1	predictor2
QSCN6L1	PITRM1	ECT2	CENPA
IGFBP5	IGFBP5.1	GMPS	CENPA
DIAPH3	PRC1	KNTC2	CENPA
NUSAP1	PRC1	Contig40831_RC	CENPA
DIAPH3.1	PRC1	MELK	CENPA
DIAPH3.2	PRC1	DTL	CENPA
C16orf61	PRC1	ORC6L	CENPA
KNTC2	PRC1	RFC4	CENPA
MELK	PRC1	CDCA7	CENPA
DTL	PRC1	MCM6	CENPA
ORC6L	PRC1	PRC1	CENPA
RFC4	PRC1	Contig20217_RC	CENPA
MCM6	PRC1	DIAPH3	NM_004702
NUSAP1	Contig20217_RC	NUSAP1	NM_004702
C16orf61	Contig20217_RC	DIAPH3.1	NM_004702
KNTC2	Contig20217_RC	DIAPH3.2	NM_004702
Contig40831_RC	Contig20217_RC	C16orf61	NM_004702
MELK	Contig20217_RC	ECT2	NM_004702
ORC6L	Contig20217_RC	KNTC2	NM_004702
PRC1	Contig20217_RC	SERF1A	NM_004702
DIAPH3	CENPA	GPR180	NM_004702
NUSAP1	CENPA	MTDH	NM_004702
QSCN6L1	CENPA	Contig40831_RC	NM_004702
DIAPH3.1	CENPA	MELK	NM_004702
DIAPH3.2	CENPA	DTL	NM_004702
C16orf61	CENPA	ORC6L	NM_004702

The pair of predictors who have correlation less than -0.5 are shown below:

predictor1	predictor2
QSCN6L1	ER
CDCA7	ER
GMPS	Grade
MELK	Grade
ORC6L	Grade
CENPA	Grade
TGFB3	DIAPH3
ER	QSCN6L1
SCUBE2	QSCN6L1
DIAPH3.2	FGF18
ECT2	FGF18
TGFB3	DIAPH3.1
FGF18	DIAPH3.2
TGFB3	DIAPH3.2
QSCN6L1	SCUBE2
CDCA7	SCUBE2
MELK	RUNDC1
CDCA7	RUNDC1
CENPA	RUNDC1
FGF18	ECT2
TGFB3	ECT2
Grade	GMPS
TGFB3	KNTC2
DIAPH3	TGFB3
DIAPH3.1	TGFB3
DIAPH3.2	TGFB3
ECT2	TGFB3
KNTC2	TGFB3
CDCA7	TGFB3
Grade	MELK
RUNDC1	MELK
Grade	ORC6L
ER	CDCA7
SCUBE2	CDCA7
RUNDC1	CDCA7
TGFB3	CDCA7
Grade	CENPA
RUNDC1	CENPA

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.

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.

```
Call:
coxph(formula = Surv(time, event) ~ ., data = train)

  n= 115, number of events= 39

                    coef  exp(coef)   se(coef)       z Pr(>|z|)
Diam           -4.569e+00  1.037e-02  4.316e-01 -10.586  < 2e-16 ***
N              -1.081e+00  3.394e-01  4.577e-01  -2.361 0.018219 *
ER              8.107e+00  3.318e+03  5.744e-01  14.114  < 2e-16 ***
Grade          -6.177e+00  2.076e-03  3.489e-01 -17.704  < 2e-16 ***
Age            -6.273e-01  5.340e-01  4.606e-02 -13.621  < 2e-16 ***
TSPYL5          1.210e+01  1.800e+05  6.676e-01  18.125  < 2e-16 ***
Contig63649_RC  5.084e+01  1.196e+22  1.057e+00  48.083  < 2e-16 ***
DIAPH3          2.040e+01  7.232e+08  1.345e+00  15.166  < 2e-16 ***
NUSAP1         -3.889e+00  2.047e-02  1.162e+00  -3.346 0.000821 ***
AA555029_RC     2.258e+01  6.377e+09  1.314e+00  17.182  < 2e-16 ***
ALDH4A1         4.798e+01  6.847e+20  1.614e+00  29.719  < 2e-16 ***
QSCN6L1        -2.566e+01  7.211e-12  1.093e+00 -23.469  < 2e-16 ***
FGF18          -8.345e+00  2.376e-04  1.005e+00  -8.301  < 2e-16 ***
DIAPH3.1       -7.685e+01  4.216e-34  1.409e+00 -54.547  < 2e-16 ***
Contig32125_RC  1.070e+01  4.445e+04  1.674e+00   6.393 1.62e-10 ***
BBC3           -6.676e+00  1.261e-03  1.506e+00  -4.434 9.24e-06 ***
DIAPH3.2        9.323e+01  3.084e+40  2.147e+00  43.428  < 2e-16 ***
RP5.860F19.3   -6.670e+00  1.268e-03  1.572e+00  -4.242 2.22e-05 ***
C16orf61       -3.351e+01  2.797e-15  1.661e+00 -20.174  < 2e-16 ***
SCUBE2         -8.047e+00  3.201e-04  4.249e-01 -18.940  < 2e-16 ***
EXT1           -5.309e+01  8.814e-24  1.565e+00 -33.925  < 2e-16 ***
FLT1            1.873e+01  1.365e+08  1.612e+00  11.621  < 2e-16 ***
GNAZ           -1.671e+00  1.880e-01  1.330e+00  -1.256 0.209065
OXCT1           2.282e+01  8.105e+09  1.448e+00  15.756  < 2e-16 ***
MMP9            1.368e+01  8.724e+05  1.022e+00  13.382  < 2e-16 ***
RUNDC1          5.320e+01  1.272e+23  1.314e+00  40.491  < 2e-16 ***
Contig35251_RC  2.096e+01  1.270e+09  1.089e+00  19.242  < 2e-16 ***
ECT2           -6.787e+00  1.128e-03  1.390e+00  -4.884 1.04e-06 ***
GMPS            3.197e+01  7.639e+13  1.414e+00  22.613  < 2e-16 ***
KNTC2          -1.313e+01  1.985e-06  1.668e+00  -7.872 3.48e-15 ***
WISP1           2.197e+01  3.495e+09  1.315e+00  16.705  < 2e-16 ***
CDC42BPA       -2.299e+01  1.039e-10  1.572e+00 -14.624  < 2e-16 ***
```

SERF1A	9.318e-01	2.539e+00	1.811e+00	0.514	0.606985	
AYTL2	-3.079e+01	4.252e-14	1.510e+00	-20.388	< 2e-16	***
GSTM3	-1.436e+00	2.380e-01	8.210e-01	-1.749	0.080357	.
GPR180	-5.112e+01	6.303e-23	1.294e+00	-39.518	< 2e-16	***
RAB6B	-2.756e+01	1.076e-12	7.937e-01	-34.722	< 2e-16	***
ZNF533	-1.883e+00	1.521e-01	5.382e-01	-3.499	0.000467	***
RTN4RL1	-4.037e+01	2.945e-18	1.461e+00	-27.637	< 2e-16	***
UCHL5	2.780e+00	1.612e+01	1.795e+00	1.549	0.121392	
PECI	-5.196e+01	2.729e-23	1.464e+00	-35.478	< 2e-16	***
MTDH	-6.057e-01	5.457e-01	1.306e+00	-0.464	0.642790	
Contig40831_RC	-1.163e+01	8.904e-06	1.362e+00	-8.541	< 2e-16	***
TGFB3	1.169e+01	1.198e+05	1.288e+00	9.079	< 2e-16	***
MELK	-7.413e+01	6.391e-33	1.413e+00	-52.451	< 2e-16	***
COL4A2	5.656e+01	3.679e+24	1.561e+00	36.233	< 2e-16	***
DTL	-5.007e+01	1.796e-22	1.534e+00	-32.633	< 2e-16	***
STK32B	-6.267e+01	6.034e-28	2.190e+00	-28.617	< 2e-16	***
DCK	4.216e+01	2.040e+18	1.561e+00	27.007	< 2e-16	***
FBXO31	-3.243e+01	8.277e-15	1.479e+00	-21.923	< 2e-16	***
GPR126	-1.074e+01	2.158e-05	6.839e-01	-15.709	< 2e-16	***
SLC2A3	-5.145e+01	4.511e-23	1.323e+00	-38.880	< 2e-16	***
PECI.1	-8.449e+00	2.141e-04	1.334e+00	-6.332	2.42e-10	***
ORC6L	-1.955e+01	3.243e-09	1.088e+00	-17.958	< 2e-16	***
RFC4	-1.409e+01	7.563e-07	1.711e+00	-8.236	< 2e-16	***
CDCA7	2.532e+01	9.955e+10	7.042e-01	35.961	< 2e-16	***
LOC643008	-2.045e+01	1.314e-09	7.134e-01	-28.668	< 2e-16	***
MS4A7	-2.438e+01	2.592e-11	8.946e-01	-27.247	< 2e-16	***
MCM6	1.165e+02	3.910e+50	1.739e+00	66.988	< 2e-16	***
AP2B1	-1.855e+01	8.748e-09	1.546e+00	-11.999	< 2e-16	***
C9orf30	-7.473e+00	5.682e-04	2.014e+00	-3.711	0.000206	***
IGFBP5	6.307e+00	5.486e+02	5.458e-01	11.555	< 2e-16	***
HRASLS	-3.723e+01	6.755e-17	1.218e+00	-30.579	< 2e-16	***
PITRM1	-2.667e+01	2.622e-12	1.725e+00	-15.460	< 2e-16	***
IGFBP5.1	1.882e+00	6.569e+00	6.040e-01	3.116	0.001832	**
NMU	-2.982e+00	5.072e-02	1.094e+00	-2.726	0.006409	**
PALM2.AKAP2	1.546e+01	5.195e+06	1.399e+00	11.053	< 2e-16	***
LGP2	4.495e+01	3.312e+19	1.512e+00	29.736	< 2e-16	***
PRC1	2.651e+01	3.265e+11	1.222e+00	21.687	< 2e-16	***
Contig20217_RC	2.449e+01	4.305e+10	1.141e+00	21.456	< 2e-16	***

```
Contig20217_RC  2.449e+01  4.305e+10  1.141e+00  21.456  < 2e-16 ***
CENPA           3.140e+00  2.311e+01  1.027e+00   3.057 0.002236 **
EGLN1          -1.041e+01  3.015e-05  1.537e+00  -6.771 1.28e-11 ***
NM_004702       4.490e+01  3.174e+19  1.115e+00  40.270  < 2e-16 ***
ESM1           -3.121e+01  2.793e-14  9.373e-01 -33.298  < 2e-16 ***
C20orf46        6.492e+00  6.596e+02  1.275e+00   5.090 3.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
Diam           1.037e-02  9.646e+01 4.449e-03 2.416e-02
N              3.394e-01  2.947e+00 1.384e-01 8.323e-01
ER             3.318e+03  3.014e-04 1.076e+03 1.023e+04
Grade          2.076e-03  4.816e+02 1.048e-03 4.114e-03
Age            5.340e-01  1.873e+00 4.879e-01 5.845e-01
TSPYL5         1.800e+05  5.557e-06 4.863e+04 6.659e+05
Contig63649_RC 1.196e+22  8.359e-23 1.506e+21 9.501e+22
DIAPH3         7.232e+08  1.383e-09 5.180e+07 1.010e+10
NUSAP1         2.047e-02  4.885e+01 2.098e-03 1.997e-01
AA555029_RC    6.377e+09  1.568e-10 4.855e+08 8.377e+10
ALDH4A1        6.847e+20  1.460e-21 2.893e+19 1.620e+22
QSCN6L1        7.211e-12  1.387e+11 8.463e-13 6.145e-11
FGF18          2.376e-04  4.209e+03 3.312e-05 1.704e-03
DIAPH3.1       4.216e-34  2.372e+33 2.665e-35 6.670e-33
Contig32125_RC 4.445e+04  2.249e-05 1.671e+03 1.182e+06
BBC3           1.261e-03  7.933e+02 6.591e-05 2.411e-02
DIAPH3.2       3.084e+40  3.242e-41 4.590e+38 2.072e+42
RP5.860F19.3   1.268e-03  7.884e+02 5.818e-05 2.765e-02
C16orf61       2.797e-15  3.575e+14 1.079e-16 7.255e-14
SCUBE2         3.201e-04  3.124e+03 1.392e-04 7.361e-04
EXT1           8.814e-24  1.135e+23 4.104e-25 1.893e-22
FLT1           1.365e+08  7.327e-09 5.794e+06 3.215e+09
GNAZ           1.880e-01  5.319e+00 1.386e-02 2.551e+00
OXCT1          8.105e+09  1.234e-10 4.744e+08 1.385e+11
MMP9           8.724e+05  1.146e-06 1.176e+05 6.469e+06
RUNDC1         1.272e+23  7.863e-24 9.684e+21 1.670e+24
Contig35251_RC 1.270e+09  7.871e-10 1.502e+08 1.075e+10
ECT2           1.128e-03  8.863e+02 7.406e-05 1.719e-02
```

GMPS	7.639e+13	1.309e-14	4.783e+12	1.220e+15
KNTC2	1.985e-06	5.039e+05	7.550e-08	5.217e-05
WISP1	3.495e+09	2.862e-10	2.653e+08	4.604e+10
CDC42BPA	1.039e-10	9.629e+09	4.768e-12	2.262e-09
SERF1A	2.539e+00	3.939e-01	7.290e-02	8.843e+01
AYTL2	4.252e-14	2.352e+13	2.204e-15	8.205e-13
GSTM3	2.380e-01	4.202e+00	4.761e-02	1.189e+00
GPR180	6.303e-23	1.586e+22	4.995e-24	7.955e-22
RAB6B	1.076e-12	9.298e+11	2.270e-13	5.096e-12
ZNF533	1.521e-01	6.575e+00	5.296e-02	4.367e-01
RTN4RL1	2.945e-18	3.395e+17	1.682e-19	5.157e-17
UCHL5	1.612e+01	6.202e-02	4.782e-01	5.437e+02
PECI	2.729e-23	3.664e+22	1.547e-24	4.815e-22
MTDH	5.457e-01	1.833e+00	4.220e-02	7.057e+00
Contig40831_RC	8.904e-06	1.123e+05	6.174e-07	1.284e-04
TGFB3	1.198e+05	8.346e-06	9.597e+03	1.496e+06
MELK	6.391e-33	1.565e+32	4.004e-34	1.020e-31
COL4A2	3.679e+24	2.718e-25	1.726e+23	7.846e+25
DTL	1.796e-22	5.569e+21	8.875e-24	3.633e-21
STK32B	6.034e-28	1.657e+27	8.248e-30	4.414e-26
DCK	2.040e+18	4.903e-19	9.567e+16	4.348e+19
FBXO31	8.277e-15	1.208e+14	4.559e-16	1.503e-13
GPR126	2.158e-05	4.633e+04	5.649e-06	8.246e-05
SLC2A3	4.511e-23	2.217e+22	3.372e-24	6.036e-22
PECI.1	2.141e-04	4.670e+03	1.566e-05	2.927e-03
ORC6L	3.243e-09	3.083e+08	3.841e-10	2.738e-08
RFC4	7.563e-07	1.322e+06	2.642e-08	2.165e-05
CDCA7	9.955e+10	1.005e-11	2.504e+10	3.958e+11
LOC643008	1.314e-09	7.613e+08	3.245e-10	5.317e-09
MS4A7	2.592e-11	3.858e+10	4.488e-12	1.497e-10
MCM6	3.910e+50	2.557e-51	1.294e+49	1.182e+52
AP2B1	8.748e-09	1.143e+08	4.224e-10	1.812e-07
C9orf30	5.682e-04	1.760e+03	1.098e-05	2.941e-02
IGFBP5	5.486e+02	1.823e-03	1.882e+02	1.599e+03
HRASLS	6.755e-17	1.480e+16	6.211e-18	7.346e-16
PITRM1	2.622e-12	3.814e+11	8.922e-14	7.707e-11
IGFBP5.1	6.569e+00	1.522e-01	2.011e+00	2.146e+01
NMU	5.072e-02	1.972e+01	5.946e-03	4.326e-01

```
PALM2.AKAP2     5.195e+06  1.925e-07 3.348e+05 8.061e+07
LGP2            3.312e+19  3.020e-20 1.712e+18 6.407e+20
PRC1            3.265e+11  3.063e-12 2.974e+10 3.585e+12
Contig20217_RC 4.305e+10  2.323e-11 4.598e+09 4.030e+11
CENPA           2.311e+01  4.328e-02 3.086e+00 1.730e+02
EGLN1           3.015e-05  3.317e+04 1.481e-06 6.137e-04
NM_004702       3.174e+19  3.151e-20 3.568e+18 2.823e+20
ESM1            2.793e-14  3.580e+13 4.449e-15 1.753e-13
C20orf46        6.596e+02  1.516e-03 5.415e+01 8.034e+03

Concordance= 0.999  (se = 0.001 )
Likelihood ratio test= 262.9  on 75 df,   p=<2e-16
Wald test            = 44318  on 75 df,   p=<2e-16
Score (logrank) test = 221.4  on 75 df,   p=2e-16
```

Warning massage shown by the model:

```
In coxph.fit(X, Y, istrat, offset, init, control, weights = weights,  :
  Ran out of iterations and did not converge
```

The warning massage means that, the coefficients of the model parameters are not fully optimized as the optimization ran out of iterations and did not converge.

Interpretation of the summary:

- Most of the covariates are significant at 5% level of significance.
- 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** = 0.999 (se = 0.001), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 99.9% accuracy.
- After that I have identified the covariates which was found to be significant at 5% level of significance and again by fitting the cox proportional hazard model on the train data, the **concordance** on the test data is found to be 0.755. That means given two random individuals of the test set, the model can predict who has greater risk of dying with 75.5% 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 0.01452291.

Estimating Model Concordance on the Test Set:

It is observed that depending on the data corresponding to the individuals selected at random in the training set, the model concordance on the test data set changes and the choice of the significant predictors at 5% level of significance also changes.

That is why, I have performed the whole process for 47 times, where at every iteration, the training and test data sets were randomly chosen and the cox proportional hazard model is fitted corresponding to those covariates which were found to be significant at 5% level of significance. For every iteration, the model concordance on the test data set is computed and ultimately the average concordance on the test data set is found to be 0.6078984. That means given two random individuals of the test set, ordinary Cox proportional hazard model can predict who has greater risk of dying with 60.7% accuracy.

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 $n \gg p$—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, I have considered two approaches for automatically performing feature selection or variable selection—that is, for excluding irrelevant variables from a multiple regression model.

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, one-at-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 R^2.
- **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.

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.

Illustration of Stepwise Selection in Cox Proportional Hazard Model through the data “nki70”:

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

```
Call:
coxph(formula = Surv(time, event) ~ KNTC2 + ESM1 + ER + Contig32125_RC +
    RUNDC1 + COL4A2 + PITRM1 + ORC6L + GPR126 + SLC2A3 + STK32B +
    MMP9 + GPR180 + RTN4RL1 + RAB6B + SERF1A + QSCN6L1 + Age +
    SCUBE2 + MELK + IGFBP5.1, data = train)

  n= 115, number of events= 38

                    coef  exp(coef)   se(coef)      z Pr(>|z|)
KNTC2          -4.358e+00  1.280e-02  1.521e+00 -2.866 0.004162 **
ESM1            2.427e+00  1.132e+01  9.570e-01  2.536 0.011225 *
ER             -1.601e+00  2.017e-01  8.867e-01 -1.805 0.071006 .
Contig32125_RC  6.252e+00  5.190e+02  1.333e+00  4.690 2.73e-06 ***
RUNDC1          3.482e+00  3.253e+01  1.191e+00  2.923 0.003468 **
COL4A2          6.954e+00  1.048e+03  1.986e+00  3.501 0.000464 ***
PITRM1         -1.412e+01  7.399e-07  2.761e+00 -5.112 3.18e-07 ***
ORC6L           9.651e+00  1.554e+04  1.831e+00  5.271 1.35e-07 ***
GPR126         -2.568e+00  7.670e-02  7.282e-01 -3.526 0.000422 ***
SLC2A3         -8.656e+00  1.741e-04  2.131e+00 -4.063 4.85e-05 ***
STK32B         -7.553e+00  5.243e-04  2.594e+00 -2.912 0.003591 **
MMP9            4.129e+00  6.210e+01  1.193e+00  3.460 0.000541 ***
GPR180         -4.295e+00  1.364e-02  1.588e+00 -2.705 0.006826 **
RTN4RL1        -5.187e+00  5.588e-03  1.800e+00 -2.882 0.003954 **
RAB6B          -2.867e+00  5.685e-02  9.127e-01 -3.142 0.001679 **
SERF1A         -8.206e+00  2.731e-04  2.781e+00 -2.951 0.003169 **
QSCN6L1         4.645e+00  1.040e+02  2.166e+00  2.144 0.032008 *
Age            -1.036e-01  9.016e-01  5.021e-02 -2.063 0.039119 *
SCUBE2         -1.570e+00  2.080e-01  6.635e-01 -2.367 0.017952 *
MELK            5.215e+00  1.839e+02  2.078e+00  2.509 0.012108 *

IGFBP5.1        3.546e+00  3.468e+01  8.196e-01  4.327 1.51e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

               exp(coef) exp(-coef) lower .95 upper .95
KNTC2          1.280e-02  7.812e+01 6.497e-04 2.523e-01
ESM1           1.132e+01  8.834e-02 1.735e+00 7.387e+01
ER             2.017e-01  4.958e+00 3.548e-02 1.147e+00
Contig32125_RC 5.190e+02  1.927e-03 3.806e+01 7.077e+03
RUNDC1         3.253e+01  3.075e-02 3.149e+00 3.359e+02
COL4A2         1.048e+03  9.546e-04 2.135e+01 5.141e+04
PITRM1         7.399e-07  1.351e+06 3.302e-09 1.658e-04
ORC6L          1.554e+04  6.433e-05 4.296e+02 5.625e+05
GPR126         7.670e-02  1.304e+01 1.840e-02 3.196e-01
SLC2A3         1.741e-04  5.742e+03 2.676e-06 1.134e-02
STK32B         5.243e-04  1.907e+03 3.248e-06 8.463e-02
MMP9           6.210e+01  1.610e-02 5.987e+00 6.441e+02
GPR180         1.364e-02  7.331e+01 6.075e-04 3.063e-01
RTN4RL1        5.588e-03  1.789e+02 1.641e-04 1.903e-01
RAB6B          5.685e-02  1.759e+01 9.503e-03 3.401e-01
SERF1A         2.731e-04  3.661e+03 1.173e-06 6.359e-02
QSCN6L1        1.040e+02  9.612e-03 1.491e+00 7.260e+03
Age            9.016e-01  1.109e+00 8.171e-01 9.948e-01
SCUBE2         2.080e-01  4.808e+00 5.666e-02 7.635e-01
MELK           1.839e+02  5.437e-03 3.130e+00 1.081e+04
IGFBP5.1       3.468e+01  2.884e-02 6.956e+00 1.729e+02
```

```
Concordance= 0.918  (se = 0.027 )
Likelihood ratio test= 125  on 21 df,   p=<2e-16
Wald test            = 53.1  on 21 df,   p=1e-04
Score (logrank) test = 112  on 21 df,   p=2e-14
```

Interpretation of the summary:

- All the covariates are significant at 5% level of significance except one, "ER".
- 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** = 0.918 (se = 0.027), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 91.8% accuracy.
- The **concordance** on the test data is found to be 0.799. That means given two random individuals of the test set, the model can predict who has gre ater risk of dying with 79.9% 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 0.1032933.
- 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.

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,

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

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 L1 **(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.

Algorithm:

Let us assume no ties in failure/censoring time. Here we wish to find β which maximizes,

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

subject to our constraint: $\alpha \sum|\beta_i| + (1-\alpha)\sum \beta_i^2 \leq c$. Maximizing the partial likelihood is equivalent to maximizing a scaled log partial likelihood,

$$\frac{2}{n}l(\beta) = \frac{2}{n}[\sum_{i=1}^{m} x_{j(i)}^{t}\beta - \log(\sum_{j \in R_i} \exp(x_j^t\beta))]$$

The scaling factor 2/n is used for convenience. Hence, if we consider the **Lagrangian** formulation, our problem becomes,

$$\hat{\beta} = argmax_{\beta}[\frac{2}{n}\left[\sum_{i=1}^{m} x_{j(i)}^{t}\beta - \log\left(\sum_{j \in R_i} \exp(x_j^t\beta)\right)\right] - \lambda P_{\alpha}(\beta)]$$

where,

$$\lambda P_{\alpha}(\beta) = \lambda(\alpha\sum_{i=1}^{p}|\beta_i| + (1-\alpha)\sum_{i=1}^{p}\beta_i^2)$$

is known as the elastic net penalty. It is a mixture of the $\alpha = 1$ (lasso) and $\alpha = 2$ (ridge regression) penalties. The lasso penalty (Tibshirani 1996) tends to choose only a few nonzero coefficients. While often desirable, this can cause problems. If two predictors are very correlated, the lasso will pick one and entirely ignore the other.

On the other hand, ridge regression scales all the coefficients towards 0, but sets none to exactly zero. This helps to regularize in problems with p > n, but does not give a sparse solution. However, ridge regression better handles correlated predictors. If two predictors are very correlated, ridge regression will tend to give them equal weight.

Basic Algorithm:

Let X denote the design matrix, β the coefficient vector, and η = Xβ. 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 $\hat{\beta}$ minimising
$$\frac{1}{n}\sum_{i=1}^{n} w(\tilde{\eta})_i(z(\tilde{\eta})_i - x_i^t\beta)^2 + \lambda P_\alpha(\beta)$$
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 3 is done by cyclical coordinate descent.

Finding optimal λ:

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

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

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, $\widehat{CV}(\lambda)$, is the sum of all $\widehat{CV}_i(\lambda)$. Then, choose the value of λ which maximizes $\widehat{CV}(\lambda)$.

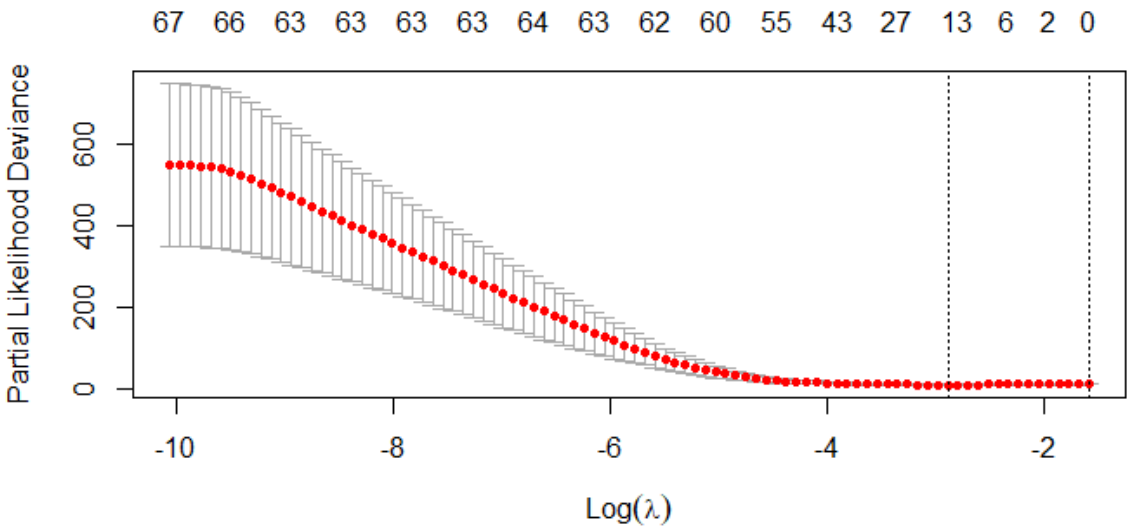
Illustration of Shrinkage Method (Lasso) in Cox Proportional Hazard Model through the data "nki70"

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 **0.05552002**.

The plot of the cross-validated error rates is given below:



Each dot represents a λ 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 λ 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 16 covariates have been selected by the model which are shown below:

```
"N"                "ALDH4A1"          "QSCN6L1"
"Contig32125_RC"   "MMP9"             "RUNDC1"
"KNTC2"            "GPR180"           "ZNF533"
"RTN4RL1"          "Contig40831_RC"   "STK32B"
"ORC6L"            "IGFBP5.1"         "LGP2"
"PRC1"
```

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(time, event) ~ ., data = train.glm[, c(1,
    2, index_lasso)])

  n= 115, number of events= 39

                    coef  exp(coef)   se(coef)      z Pr(>|z|)
N                0.032850   1.033396   0.405766  0.081 0.935475
ALDH4A1          1.311360   3.711218   1.278961  1.025 0.305207
QSCN6L1          0.207200   1.230229   1.046281  0.198 0.843018
Contig32125_RC   4.560346  95.616590   1.247674  3.655 0.000257 ***
MMP9             3.065011  21.434702   1.034688  2.962 0.003054 **
RUNDC1           2.842915  17.165737   1.006316  2.825 0.004727 **
KNTC2           -6.467183   0.001554   1.933989 -3.344 0.000826 ***
GPR180          -1.482387   0.227095   1.259147 -1.177 0.239078
ZNF533          -1.945114   0.142971   0.585039 -3.325 0.000885 ***
RTN4RL1         -3.234234   0.039390   1.427912 -2.265 0.023512 *
Contig40831_RC   2.192802   8.960283   1.047020  2.094 0.036231 *
STK32B          -1.217038   0.296106   1.540685 -0.790 0.429567
ORC6L            0.701990   2.017764   1.211130  0.580 0.562174
IGFBP5.1         1.388977   4.010745   0.560040  2.480 0.013133 *
LGP2             2.080253   8.006492   1.236677  1.682 0.092544 .
PRC1             6.889160 981.576458   1.732919  3.975 7.02e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
                exp(coef) exp(-coef) lower .95 upper .95
N               1.033e+00  9.677e-01 4.665e-01 2.289e+00
ALDH4A1         3.711e+00  2.695e-01 3.026e-01 4.552e+01
QSCN6L1         1.230e+00  8.129e-01 1.583e-01 9.563e+00
Contig32125_RC  9.562e+01  1.046e-02 8.289e+00 1.103e+03
MMP9            2.143e+01  4.665e-02 2.821e+00 1.629e+02
RUNDC1          1.717e+01  5.826e-02 2.388e+00 1.234e+02
KNTC2           1.554e-03  6.437e+02 3.509e-05 6.879e-02
GPR180          2.271e-01  4.403e+00 1.925e-02 2.679e+00
ZNF533          1.430e-01  6.994e+00 4.542e-02 4.500e-01
RTN4RL1         3.939e-02  2.539e+01 2.399e-03 6.469e-01
Contig40831_RC  8.960e+00  1.116e-01 1.151e+00 6.975e+01
STK32B          2.961e-01  3.377e+00 1.445e-02 6.066e+00
ORC6L           2.018e+00  4.956e-01 1.879e-01 2.167e+01
IGFBP5.1        4.011e+00  2.493e-01 1.338e+00 1.202e+01
LGP2            8.006e+00  1.249e-01 7.092e-01 9.039e+01
PRC1            9.816e+02  1.019e-03 3.287e+01 2.931e+04

Concordance= 0.889  (se = 0.023 )
Likelihood ratio test= 90.1  on 16 df,   p=2e-12
Wald test            = 60.29  on 16 df,   p=5e-07
Score (logrank) test = 89.39  on 16 df,   p=3e-12
```

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** = 0.889 (se = 0.023), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 88.9% accuracy.
- The **concordance** on the test data is found to be 0.745. That means given two random individuals of the test set, the model can predict who has gre ater risk of dying with 74.5% 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 0.08888743.
- 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.

Estimating Model Concordance on the Test Set:

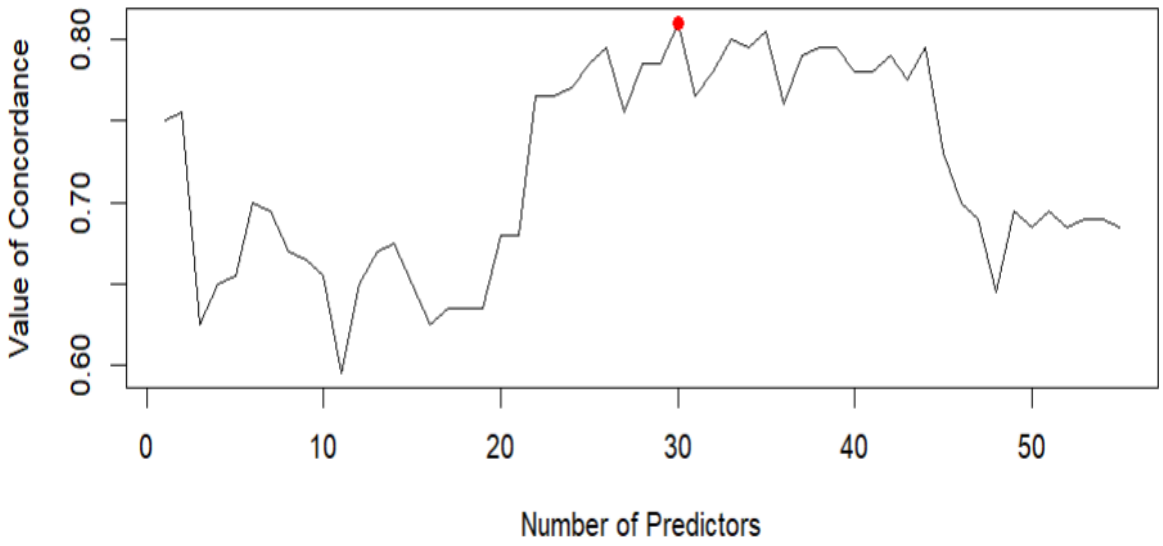
- It is observed that depending on the data corresponding to the individuals selected at random in the training set, the optimal value of λ in the fitted model also changes and the model concordance on the test data set changes along with the choice of the significant predictors.
- That is why, I have performed the whole process for 50 times, where at every iteration,
    1. The training and test data sets were randomly chosen.
    2. Then using cross-validation, optimal value of λ has been obtained.
    3. Considering the obtained λ, the non-zero coefficients have been noted by maximizing lasso penalized Cox partial likelihood.
    4. Using the covariates corresponding to these non-zero coefficients, a new Cox proportional hazard model has been fitted.
    5. Corresponding to the fitted Cox model, the concordance value on the test set have been computed and stored.
- In 22 out of 50 iterations, none of the predictors was selected by Lasso penalized Cox proportional hazard model.
- Considering only those models for which at least one predictor was select ed by Lasso penalized Cox proportional hazard model, the concordance v alues has been obtained on the corresponding test set of those iterations. F inally, the mean concordance value for those 28 models has been compute d, which is found to be 0.5809248. That means on an average, given two r andom individuals of the test set, the model can predict who has greater ri sk of dying with 58.09% accuracy.

In each of those 28 models, different covariates were selected. Some of the cova riates were selected in multiple models. These covariates have been arranged in decreasing order with respect to their corresponding count i.e., the number of m odels that include those covariates which is shown below.

predictors	predictor_count
MS4A7	5
TSPYL5	4
ALDH4A1	3
MMP9	3
RUNDC1	3
GMPS	3
GSTM3	3
MTDH	3
DCK	3
SLC2A3	3
LGP2	3
EGLN1	3
Contig63649_RC	2
FGF18	2
Contig32125_RC	2
EXT1	2
OXCT1	2
KNTC2	2
AYTL2	2
GPR180	2
RAB6B	2
RTN4RL1	2
PECI	2
MELK	2
COL4A2	2
DTL	2
PECI.1	2
ORC6L	2
AP2B1	2
IGFBP5.1	2
PALM2.AKAP2	2
C20orf46	2
N	1
Grade	1
Age	1
QSCN6L1	1
DIAPH3.1	1
FLT1	1
GNAZ	1
Contig35251_RC	1
ECT2	1
WISP1	1
SERF1A	1
ZNF533	1
Contig40831_RC	1
STK32B	1
FBXO31	1
CDCA7	1
LOC643008	1
IGFBP5	1
HRASLS	1
NMU	1
PRC1	1
Contig20217_RC	1
ESM1	1

After obtaining this table, first, a Cox proportional hazard model has been fitted with the predictor ("MS4A7") which occurred maximum number of times. Next , along with "MS4A7", the next most occurred predictor ("TSPYL5") has been c onsidered to fit a new Cox proportional hazard model. In this way, one-at-a-time , the next most occurred predictor has been selected to fit another Cox model.

Every time a Cox proportional hazard model has been fitted using different num ber of predictors, the concordance value has been computed on the test set. It is f ound that the model gives highest value of concordance (0.81) when first 30 pre dictors from the above table have been selected. A graph of concordance value c orresponding to the number of predictors used is shown below.



So, if we consider the above mentioned first 30 predictors to fit a Cox proportio nal hazard model, then given two random individuals of the test set, the model c an predict who has greater risk of dying with 81% accuracy.

Illustration of Shrinkage Method (LASSO) in Cox Proportional Hazard Model through the data "Breast Cancer Gene Expression Data (Meabric_RNA_Mutation)"

Source of the Data:

The Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) database is a Canada-UK Project which contains targeted sequencing data of 1,980 primary breast cancer samples. Clinical and genomic data was downloaded from cBioPortal.
The dataset was collected by Professor Carlos Caldas from Cambridge Research Institute and Professor Sam Aparicio from the British Columbia Cancer Centre in Canada and published on Nature Communications (Pereira et al., 2016).

This dataset has been downloaded from "kaggle".

Link: https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric.

Description of the Data:

The data includes 31 clinical attributes, m-RNA levels z-score for 331 genes, and mutation in 175 genes for 1904 breast cancer patients.

Data Cleaning:

After removing the rows which had at least one missing value corresponding to any covariate, a complete data of 854 patients has been obtained.

Three columns namely "patient_id", "cancer_type", "death_from_cancer" have been dropped off, as they are of no use to model the hazard rate.

Also, all the covariates which had multiple levels, had been transformed into numeric variable by assigning different numbers to the corresponding levels.

At last, we get a complete data of 854 patients with 690 covariates.

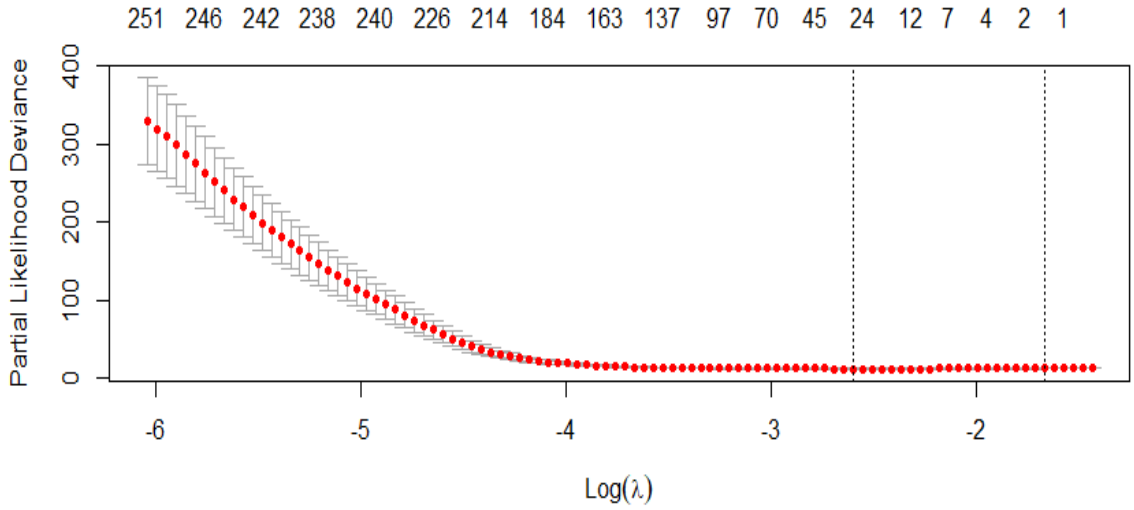
It is also observed that out of 688 predictors, there are 328 pairs who have correlation greater than 0.5 and 102 pairs who have correlation less than -0.5, which confirms the presence of multicollinearity.

Defining the training set and test set:

40% of the 854 individuals have been selected at random as the members of the training set along with the corresponding covariates and the rest of the observations have been selected as the members of the test set to make the training set to be a high dimensional data.

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 **0.07425321**.

The plot of the cross-validated error rates is given below:



Each dot represents a λ 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 λ 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 $\lambda =$ **0.07425321**has been fitted.

Only 28 covariates have been selected out of 688, by the model which are shown below:

```
"age_at_diagnosis"                "type_of_breast_surgery"
"primary_tumor_laterality"        "nottingham_prognostic_index"
"tumor_size"                      "tumor_stage"
"bard1"                           "pms2"
"stat5a"                          "notch3"
"bcl2"                            "gsk3b"
"mmp15"                           "smad6"
"agmo"                            "asxl2"
"ctcf"                            "muc16"
"ncoa3"                           "rpgr"
"sik1"                            "smarcd1"
"ttyh1"                           "hsd17b2"
"shbg"                            "st7"
"ros1_mut"                        "prkce_mut"
```

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(overall_survival_months, overall_survival) ~
    ., data = train.glm[, c(1, 2, index_lasso)])

  n= 341, number of events= 149

                              coef exp(coef)  se(coef)      z Pr(>|z|)
age_at_diagnosis          0.045699  1.046759  0.007861  5.814 6.11e-09 ***
type_of_breast_surgery    0.508628  1.663008  0.213895  2.378  0.01741 *
primary_tumor_laterality -0.488172  0.613747  0.181973 -2.683  0.00730 **
nottingham_prognostic_index 0.341093  1.406484  0.131129  2.601  0.00929 **
tumor_size                0.015272  1.015389  0.005491  2.781  0.00542 **
tumor_stage               0.277960  1.320434  0.199589  1.393  0.16372
bard1                     0.160993  1.174677  0.115796  1.390  0.16443
pms2                     -0.064877  0.937183  0.125701 -0.516  0.60577
stat5a                   -0.227249  0.796722  0.107791 -2.108  0.03501 *
notch3                    0.148296  1.159857  0.101745  1.458  0.14497
bcl2                      0.029679  1.030124  0.129602  0.229  0.81887
gsk3b                     0.106614  1.112505  0.139024  0.767  0.44315
mmp15                    -0.036980  0.963696  0.124320 -0.297  0.76612
smad6                     0.087341  1.091269  0.102594  0.851  0.39459
agmo                      0.330849  1.392149  0.104644  3.162  0.00157 **
asxl2                    -0.105039  0.900290  0.102466 -1.025  0.30531
ctcf                      0.283336  1.327551  0.110362  2.567  0.01025 *
muc16                     0.217891  1.243452  0.092413  2.358  0.01838 *
ncoa3                     0.161273  1.175005  0.112920  1.428  0.15324
rpgr                     -0.158708  0.853245  0.113373 -1.400  0.16155
sik1                     -0.439156  0.644580  0.106229 -4.134 3.56e-05 ***
smarcd1                   0.101695  1.107046  0.102211  0.995  0.31976
```

```
ttyh1                          0.279132  1.321982  0.124955  2.234  0.02549 *
hsd17b2                        0.238189  1.268949  0.088586  2.689  0.00717 **
shbg                          -0.224211  0.799147  0.092763 -2.417  0.01565 *
st7                           -0.234036  0.791333  0.118947 -1.968  0.04912 *
ros1_mut                       0.079988  1.083274  0.045724  1.749  0.08023 .
prkce_mut                      0.742252  2.100661  0.292267  2.540  0.01110 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

                             exp(coef) exp(-coef) lower .95 upper .95
age_at_diagnosis                1.0468     0.9553    1.0308    1.0630
type_of_breast_surgery          1.6630     0.6013    1.0935    2.5291
primary_tumor_laterality        0.6137     1.6293    0.4296    0.8768
nottingham_prognostic_index     1.4065     0.7110    1.0877    1.8187
tumor_size                      1.0154     0.9848    1.0045    1.0264
tumor_stage                     1.3204     0.7573    0.8929    1.9526
bard1                           1.1747     0.8513    0.9362    1.4740
pms2                            0.9372     1.0670    0.7325    1.1990
stat5a                          0.7967     1.2551    0.6450    0.9841
notch3                          1.1599     0.8622    0.9502    1.4158
bcl2                            1.0301     0.9708    0.7990    1.3280
gsk3b                           1.1125     0.8989    0.8472    1.4610
mmp15                           0.9637     1.0377    0.7553    1.2296
smad6                           1.0913     0.9164    0.8925    1.3343
agmo                            1.3921     0.7183    1.1340    1.7091
asxl2                           0.9003     1.1108    0.7365    1.1005
ctcf                            1.3276     0.7533    1.0693    1.6481
muc16                           1.2435     0.8042    1.0374    1.4904
ncoa3                           1.1750     0.8511    0.9417    1.4661
rpgr                            0.8532     1.1720    0.6832    1.0656
sik1                            0.6446     1.5514    0.5234    0.7938
smarcd1                         1.1070     0.9033    0.9061    1.3526
ttyh1                           1.3220     0.7564    1.0348    1.6888
hsd17b2                         1.2689     0.7881    1.0667    1.5096
shbg                            0.7991     1.2513    0.6663    0.9585
st7                             0.7913     1.2637    0.6268    0.9991
ros1_mut                        1.0833     0.9231    0.9904    1.1848
prkce_mut                       2.1007     0.4760    1.1846    3.7251

Concordance= 0.813  (se = 0.018 )
Likelihood ratio test= 211.4  on 28 df,   p=<2e-16
Wald test            = 197.2  on 28 df,   p=<2e-16
Score (logrank) test = 229.3  on 28 df,   p=<2e-16
```

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** = 0.813 (se = 0.018), which means given two random individuals of the training set, the model can predict who has greater risk of dying with 81.3% accuracy.

- The **concordance** on the test data is found to be 0.838. That means given two random individuals of the test set, the model can predict who has gre ater risk of dying with 83.8% 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 0.08591582.
- As the covariates have multicollinearity among themselves, some predicto rs may result complexity in the model. So, using penalized Cox proportio nal hazard regression procedure is found to be feasible method, where the selected covariates are smaller in number but appropriate and effective, w ith respect to the full model. Here, ordinary cox proportional hazard mode l cannot be fitted and stepwise selection is also not possible here due to th e high dimension of the data.

Estimating Model Concordance on the Test Set:

- It is observed that depending on the data corresponding to the individuals selected at random in the training set, the optimal value of λ in the fitted model also changes and the model concordance on the test data set changes along with the choice of the significant predictors.
- That is why, I have performed the whole process for 50 times, where at every iteration,
    6. The training and test data sets were randomly chosen.
    7. Then using cross-validation, optimal value of λ has been obtained.
    8. Considering the obtained λ, the non-zero coefficients have been noted by maximizing lasso penalized Cox partial likelihood.
    9. Using the covariates corresponding to these non-zero coefficients, a new Cox proportional hazard model has been fitted.
    10. Corresponding to the fitted Cox model, the concordance value on the test set have been computed and stored.
- In 12 out of 50 iterations, none of the predictors was selected by Lasso penalized Cox proportional hazard model.
- Considering only those models for which at least one predictor was select ed by Lasso penalized Cox proportional hazard model, the concordance v alues has been obtained on the corresponding test set of those iterations. F inally, the mean concordance value for those 38 models has been compute d, which is found to be 0.5421146. That means on an average, given two r andom individuals of the test set, the model can predict who has greater ri sk of dying with 54.21% accuracy.

In each of those 38 models, different covariates were selected. Some of the cova riates were selected in multiple models. These covariates have been arranged in decreasing order with respect to their corresponding count i.e., the number of m odels that include those covariates which is shown below.

predictor_index	predictors	predictor_count
19	primary_tumor_laterality	4
153	bmp10	3
263	peg3	3
287	smad6	3
363	ctcf	3
380	hdac9	3
425	rpgr	3
490	hsd3b1	3
501	rdh5	3
607	setdb1_mut	3
4	type_of_breast_surgery	2
34	atm	2
92	ctbp2	2
105	hes1	2
120	notch3	2
122	numb	2
168	casp7	2
181	egfr	2
187	erbb4	2
209	izumo1r	2
239	mmp15	2
269	rab25	2
291	terc	2
304	wwox	2
311	kmt2d	2
321	abcc1	2
324	bmf	2
339	afdn	2
340	aff2	2
341	agmo	2
346	alk	2
351	asxl2	2
398	muc16	2
400	myo3a	2
403	nf2	2
409	nt5e	2
420	prps2	2
426	ryr2	2
444	syne1	2
447	tbl1xr1	2
453	ush2a	2
470	cyp17a1	2
484	hsd17b2	2
509	srd5a2	2
516	ugt2b17	2
521	ahnak2_mut	2
531	ush2a_mut	2
532	ryr2_mut	2
546	ncor2_mut	2
549	pten_mut	2
554	stab2_mut	2
558	map2k4_mut	2
559	ros1_mut	2
562	erbb2_mut	2
565	ep300_mut	2
569	setd1a_mut	2
579	pik3r1_mut	2

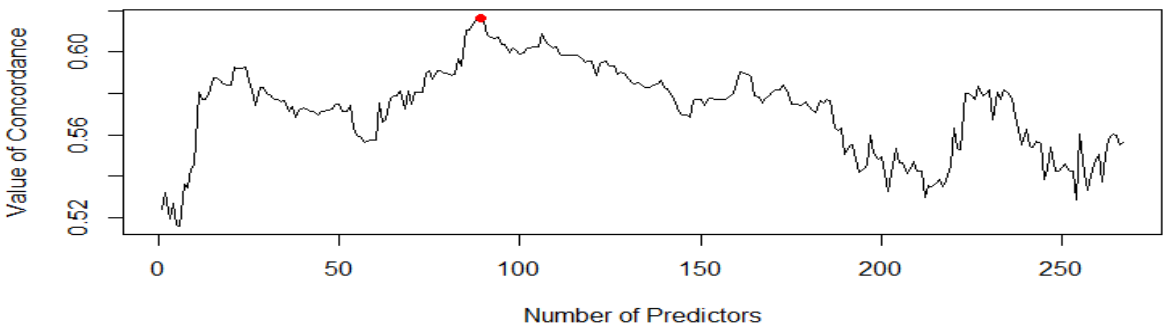
580	myo3a_mut	2
649	prkce_mut	2
657	acvrl1_mut	2
3	age_at_diagnosis	1
5	cancer_type_detailed	1
6	cellularity	1
7	chemotherapy	1
9	cohort	1
13	her2_status_measured_by_snp6	1
14	her2_status	1
18	integrative_cluster	1
20	lymph_nodes_examined_positive	1
21	mutation_count	1
22	nottingham_prognostic_index	1
23	oncotree_code	1
25	radio_therapy	1
26	X3.gene_classifier_subtype	1
27	tumor_size	1
28	tumor_stage	1
29	brca1	1
32	pten	1
33	tp53	1
35	cdh1	1
37	nbn	1
40	bard1	1
44	pms2	1
46	rad51c	1
53	ccnb1	1
55	ccne1	1
56	cdk2	1
67	e2f1	1
75	src	1
76	jak1	1
80	stat3	1
81	stat5a	1
85	adam10	1
93	cul1	1
96	dll4	1
103	hdac1	1
106	hes5	1
109	jag1	1
114	maml2	1
119	notch2	1
124	psen1	1
126	psenen	1
127	rbpj	1
128	rbpjl	1
129	rfng	1
132	hes2	1
133	hes4	1
134	hes7	1
141	acvr2b	1
144	akt1s1	1
146	apaf1	1
148	atr	1
151	bcl2	1
156	bmp3	1
157	bmp4	1
158	bmp5	1
159	bmp6	1
164	braf	1
166	casp3	1
172	csf1	1
178	diras3	1
180	dph1	1
182	eif4e	1
184	eif5a2	1
193	folr3	1
196	gdf11	1
197	gdf2	1
198	gsk3b	1
199	hif1a	1
201	hras	1
208	itgb3	1
215	map2k3	1

219	map3k3	1
221	map3k5	1
224	mapk14	1
231	mdc1	1
235	mmp11	1
243	mmp2	1
246	mmp24	1
248	mmp26	1
249	mmp27	1
250	mmp28	1
258	pdgfa	1
259	pdgfb	1
260	pdgfra	1
266	pik3r2	1
268	ptk2	1
270	rad51	1
272	rassf1	1
274	rictor	1
275	rps6	1
277	rps6ka2	1
278	rps6kb1	1
279	rps6kb2	1
280	rptor	1
286	smad5	1
288	smad7	1
293	tgfb1	1
294	tgfb2	1
297	tgfbr2	1
298	tgfbr3	1
301	vegfa	1
306	arid1a	1
307	arid1b	1
308	cbfb	1
318	tbx3	1
320	abcb11	1
326	cyp3a4	1
328	fn1	1
330	map4	1
334	tubb1	1
345	akap9	1
347	apc	1
353	bcas3	1
355	cacna2d3	1
356	ccnd3	1
357	chd1	1
358	clk3	1
365	ctnna3	1
368	dnah5	1
376	gh1	1
382	hist1h2bc	1
383	kdm3a	1
393	magea8	1
394	map3k10	1
396	men1	1
399	myo1a	1
401	ncoa3	1
406	nr3c1	1
408	nrg3	1
412	pbrm1	1
416	prkce	1
419	prkg1	1
422	ptpn22	1
424	rasgef1b	1
433	shank2	1
435	sik1	1
437	smarcb1	1
438	smarcc1	1
440	smarcd1	1
443	stmn2	1
451	ttyh1	1
452	ubr5	1
457	ackr3	1
459	akr1c2	1
460	akr1c3	1
467	cyb5a	1
469	cyp11b2	1
474	cyp3a5	1
475	cyp3a7	1
476	ddc	1
477	hes6	1
478	hsd17b1	1
481	hsd17b12	1
482	hsd17b13	1
483	hsd17b14	1
485	hsd17b3	1
489	hsd17b8	1
491	hsd3b2	1
492	hsd3b7	1
494	met	1
497	pik3r3	1
504	shbg	1
505	slc29a1	1
510	srd5a3	1
511	st7	1
512	star	1
513	tnk2	1
514	tulp4	1
517	ugt2b7	1
519	tp53_mut	1

527	dnah11_mut	1
530	kmt2d_mut	1
533	dnah5_mut	1
534	herc2_mut	1
543	lama2_mut	1
547	col12a1_mut	1
550	akt1_mut	1
551	atr_mut	1
552	thada_mut	1
555	myh9_mut	1
556	runx1_mut	1
557	nf1_mut	1
560	lamb3_mut	1
561	arid1b_mut	1
568	setd2_mut	1
573	rb1_mut	1
583	ctcf_mut	1
592	usp28_mut	1
594	brca2_mut	1
603	l1cam_mut	1
609	arid5b_mut	1
610	egfr_mut	1
614	npnt_mut	1
615	nek1_mut	1
617	zfp36l1_mut	1
618	smad4_mut	1
626	cdkn1b_mut	1
628	men1_mut	1
631	ptpn22_mut	1
636	ttyh1_mut	1
638	or6a2_mut	1
639	map3k13_mut	1
646	prkcz_mut	1
653	nf2_mut	1
659	cdkn2a_mut	1
669	nr3c1_mut	1

After obtaining this table, first, a Cox proportional hazard model has been fitted with the predictor ("primary_tumor_laterality") which occurred maximum number of times. Next, along with "primary_tumor_laterality", the next most occurred predictor ("bmp10") has been considered to fit a new Cox proportional hazard model. In this way, one-at-a-time, the next most occurred predictor has been selected to fit another Cox model.

Every time a Cox proportional hazard model has been fitted using different number of predictors, the concordance value has been computed on the test set. It is found that the model gives highest value of concordance (0.616) when first 89 predictors from the above table have been selected. A graph of concordance value corresponding to the number of predictors used is shown below.



So, if we consider the above mentioned first 89 predictors to fit a Cox proportional hazard model, then given two random individuals of the test set, the model can predict who has greater risk of dying with 61% accuracy.

References

- **Gui J, Li H (2005)**. "Penalized Cox Regression Analysis in the High-Dim ensional and Low Sample Size Settings, with Applications to Microarray Gene Expression Data." Bioinformatics, 25(13), 3001–2008.
- **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 Co x Model." Statistics in Medicine, 16, 385–395.
- **Cox DR (1972)**. "Regression Models and Life Tables." Journal of the Roy al Statistical Society B, 34, 187–220.
- 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
- Regularization Paths for Cox's Proportional Hazards Model via Coordinat e Descent, By Noah Simon, Jerome Friedman, Trevor Hastie, Rob Tibshir ani.
- Regularized Cox Regression, By Noah Simon, Jerome Friedman, Trevor H astie, Rob Tibshirani, Kenneth Tay, Balasubramanian Narasimhan.
- Variable selection techniques for the Cox proportional hazards model: A c omparative study, By Simon Petersson and Klas Sehlstedt, University of G othenburg 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.

GitHub Link for R (software) Scripts: https://github.com/rohitdutta22/Cox.git