Echocardiogram Dataset

**Survival Analysis Problem**

# Objective –

The problem addressed by past researchers was to predict from the other variables whether the patient will survive at least one year.

# Data Summary –

Number of instances: 132

|  |  |
| --- | --- |
| **Variable Name** | **Description** |
| Survival | The number of months patient survived (has survived, if patient is still alive) |
| Still-alive | A binary variable. 0=dead at end of survival period, 1 means still alive |
| Age | Age in years when heart attack occurred |
| Pericardial-effusion | Binary. Pericardial effusion is fluid around the heart. 0=no fluid, 1=fluid |
| Fractional-shortening | A measure of contracility around the heart lower numbers are increasingly abnormal |
| Epss | E-point septal separation, another measure of contractility. Larger numbers are increasingly abnormal. |
| LVDD | Left ventricular end-diastolic dimension. This is a measure of the size of the heart at end-diastole. Large hearts tend to be sick hearts. |
| Wall-motion-score | A measure of how the segments of the left ventricle are moving |
| Wall-motion-index | Equals wall-motion-score divided by number of segments seen. Usually 12-13 segments are seen in an echocardiogram. To be used INSTEAD of the wall motion score. |

# Theory (Cox PH Model) –

The purpose of the model is to evaluate simultaneously the effect of several factors on survival. In other words, it allows us to examine how specified factors influence the rate of a particular event happening (e.g., infection, death) at a particular point in time. This rate is commonly referred as the hazard rate. Predictor variables (or factors) are usually termed *covariates* in the survival-analysis literature.

The Cox model is expressed by the *hazard function* denoted by h(t). Briefly, the hazard function can be interpreted as the risk of dying at time t. It can be estimated as follow:

h(t)=h0(t)×exp(b1x1+b2x2+...+bpxp)

where,

* *t* represents the survival time
* h(t)h(t) is the hazard function determined by a set of p covariates (x1,x2,...,xpx1,x2,...,xp)
* the coefficients (b1,b2,...,bpb1,b2,...,bp) measure the impact (i.e., the effect size) of covariates.
* the term h0 is called the baseline hazard. It corresponds to the value of the hazard if all the xi are equal to zero (the quantity exp(0) equals 1). The ‘t’ in h(t) reminds us that the hazard may vary over time.

The Cox model can be written as a multiple linear regression of the logarithm of the hazard on the variables xi, with the baseline hazard being an ‘intercept’ term that varies with time.

The quantities exp(bi) are called hazard ratios (HR). A value of bi greater than zero, or equivalently a hazard ratio greater than one, indicates that as the value of the ith covariate increases, the event hazard increases and thus the length of survival decreases.

Put another way, a hazard ratio above 1 indicates a covariate that is positively associated with the event probability, and thus negatively associated with the length of survival.

In summary,

* HR = 1: No effect
* HR < 1: Reduction in the hazard
* HR > 1: Increase in Hazard

## Steps:

1. Treat missing values, replaced NAs with mean value
2. Fit univariate models to assess significance of the covariates
3. Fit a cox PH Model on Train Data to obtain the significant result
4. Predict probabilities of Survival for Test data

## Code & Output:

**Univariate Analysis:**

> covariates <- c("age", "pericardialeffusion", "fractionalshortening", "epss", "lvdd","wallmotion.index")

> univ\_formulas <- sapply(covariates,

+ function(x) as.formula(paste('Surv(survival, alive)~', x)))

>

> univ\_models <- lapply( univ\_formulas, function(x){coxph(x, data)})

> # Extract data

> univ\_results <- lapply(univ\_models,

+ function(x){

+ x <- summary(x)

+ p.value<-signif(x$wald["pvalue"], digits=2)

+ wald.test<-signif(x$wald["test"], digits=2)

+ beta<-signif(x$coef[1], digits=2);#coeficient beta

+ HR <-signif(x$coef[2], digits=2);#exp(beta)

+ HR.confint.lower <- signif(x$conf.int[,"lower .95"], 2)

+ HR.confint.upper <- signif(x$conf.int[,"upper .95"],2)

+ HR <- paste0(HR, " (",

+ HR.confint.lower, "-", HR.confint.upper, ")")

+ res<-c(beta, HR, wald.test, p.value)

+ names(res)<-c("beta", "HR (95% CI for HR)", "wald.test",

+ "p.value")

+ return(res)

+ #return(exp(cbind(coef(x),confint(x))))

+ })

> res <- t(as.data.frame(univ\_results, check.names = FALSE))

> as.data.frame(res)

beta HR (95% CI for HR) wald.test p.value

age 0.074 1.1 (1-1.1) 14 0.00019

pericardialeffusion 0.72 2.1 (1-4.2) 4.1 0.043

fractionalshortening -5.6 0.0039 (0.00011-0.14) 9.2 0.0024

epss 0.069 1.1 (1-1.1) 13 0.00036

lvdd 0.55 1.7 (1.2-2.5) 8.6 0.0034

wallmotion.index 1.4 4.2 (2.3-7.6) 21 3.6e-06

**Cox PH Model (Multivariate Analysis)**

> library(rms)

> dd<- datadist(train.data)

> options(datadist="dd")

> cox\_model <- cph(Surv(survival,alive) ~ age+ pericardialeffusion+fractionalshortening

+ +epss + lvdd + wallmotion.index, data = train.data)

> summary(cox\_model)

Effects Response : Surv(survival, alive)

Factor Low High Diff. Effect S.E. Lower 0.95 Upper 0.95

age 57.000 68.0000 11.00000 0.76277 0.23725 0.29777 1.22780

Hazard Ratio 57.000 68.0000 11.00000 2.14420 NA 1.34690 3.41360

pericardialeffusion 0.000 1.0000 1.00000 1.07170 0.40091 0.28589 1.85740

Hazard Ratio 0.000 1.0000 1.00000 2.92020 NA 1.33090 6.40730

fractionalshortening 0.150 0.2705 0.12050 -0.30292 0.26760 -0.82741 0.22158

Hazard Ratio 0.150 0.2705 0.12050 0.73866 NA 0.43718 1.24800

epss 8.650 15.1500 6.50000 0.18757 0.22440 -0.25224 0.62738

Hazard Ratio 8.650 15.1500 6.50000 1.20630 NA 0.77706 1.87270

lvdd 4.285 5.2700 0.98500 0.11214 0.28754 -0.45143 0.67571

Hazard Ratio 4.285 5.2700 0.98500 1.11870 NA 0.63672 1.96540

wallmotion.index 1.000 1.5762 0.57625 0.76257 0.20979 0.35139 1.17380

Hazard Ratio 1.000 1.5762 0.57625 2.14380 NA 1.42100 3.23410

> fast\_bw

Deleted Chi-Sq d.f. P Residual d.f. P AIC

lvdd 0.15 1 0.6965 0.15 1 0.6965 -1.85

epss 1.20 1 0.2730 1.35 2 0.5082 -2.65

fractionalshortening 3.70 1 0.0543 5.06 3 0.1677 -0.94

Approximate Estimates after Deleting Factors

Coef S.E. Wald Z P

age 0.06893 0.0208 3.314 9.212e-04

pericardialeffusion 0.86168 0.3867 2.228 2.586e-02

wallmotion.index 1.62608 0.3373 4.822 1.425e-06

**Factors in Final Model**

**[1] age pericardialeffusion wallmotion.index**

Fitting Model on only Significant Variables:

> cox\_model\_selected<-cph(Surv(survival,alive)~age + pericardialeffusion + wallmotion.index, data = train.data,

+ x = T, y = T, surv = TRUE)

> print(cox\_model\_selected)

Cox Proportional Hazards Model

cph(formula = Surv(survival, alive) ~ age + pericardialeffusion +

wallmotion.index, data = train.data, x = T, y = T, surv = TRUE)

Model Tests Discrimination

Indexes

Obs 100 LR chi2 33.66 R2 0.302

Events 34 d.f. 3 Dxy 0.570

Center 6.9202 Pr(> chi2) 0.0000 g 1.215

Score chi2 36.98 gr 3.369

Pr(> chi2) 0.0000

**Coef** S.E.Wald **Z Pr(>|Z|)**

**age 0.0712** 0.02173.28 **0.0010**

**pericardialeffusion 0.8650** 0.38222.26 **0.0236**

**wallmotion.index 1.6498** 0.34114.84 **<0.0001**

Predictions for Test Data:

> survfit\_testx<-survfit(cox\_model\_selected, test.data, conf.int = 0.9)

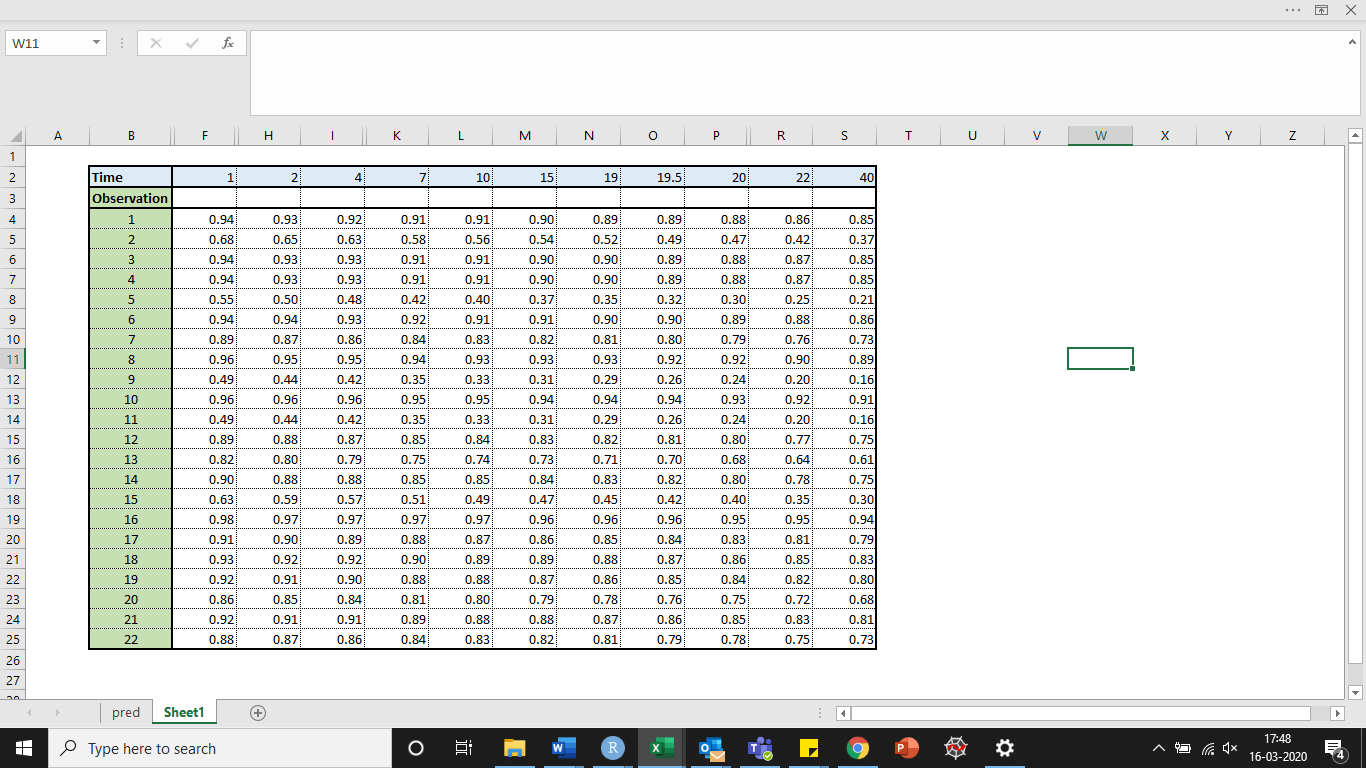
> View(survfit\_testx)

> summary(survfit\_testx)

Call: survfit(formula = cox\_model\_selected, newdata = test.data, conf.int = 0.9)

### FINAL OUTPUT:

The below table gives the probability of survival at different values of time (say, t=1,2,4,7…) for 22 observations passed as train data



## Result:

The problem statement is a Survival Analysis problem, with censored & uncensored data both.

3 Significant variables were obtained from CoxPH ie. **Age**, **Pericardial Effusion** and **Wall-motion Index**