Cox Proportional Hazard (CPH) Model

Md. Moyazzem Hossain

Professor,

Department of Statistics and Data Science

Jahangirnagar University

History

- David Cox, a British statistician, published a paper in 1972 entitled “Regression Models and Life-Tables (with Discussion),” *Journal of the Royal Statistical Society, Series B*, 34:187-220
- It is one of the most frequently cited journal articles in statistics and medicine
- Introduced “maximum partial likelihood”

Cox PH regression equation

$$h_i(t) = h_0(t)\exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots\ldots + \beta_n x_{in})$$

$h_i(t)$ is the hazard function for individual *i*

$h_0(t)$ is the baseline hazard function and can take any form
It is estimated from the data (non parametric)

$x_{i1}, x_{i2}, \ldots, x_{in}$ are the covariates

$\beta_1, \beta_2, \ldots, \beta_n$ are the regression coefficients estimated from the data

Effect of covariates is constant over time (parameterised)
This is the proportional hazards assumption

Therefore, Cox regression referred to as a semi-parametric model

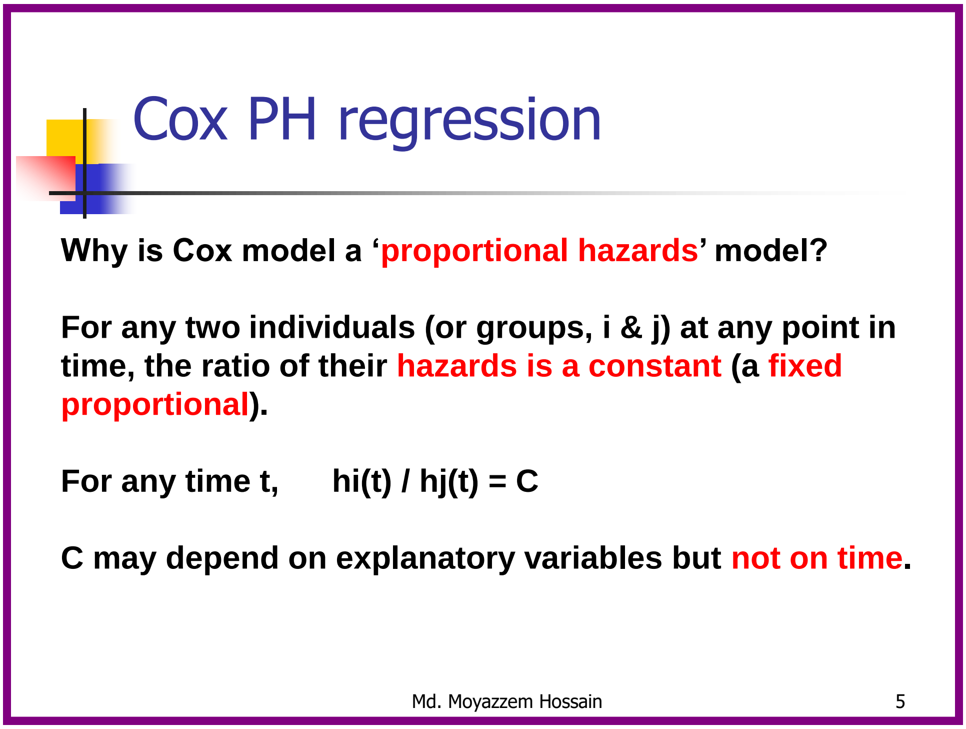
Cox PH regression

Why is Cox model a semiparametric model?

$$\boldsymbol{h(t|x_i) = h_0(t)\ \exp\ (\beta_i x_i)}$$

h_0(t): nonparametric baseline hazard function, this function does not have to be specified, the hazard may change as a function of time.

exp ($\beta_i x_i$): parametric form for the effects of the covariates, the hazard function changes as a exponential function of covariates



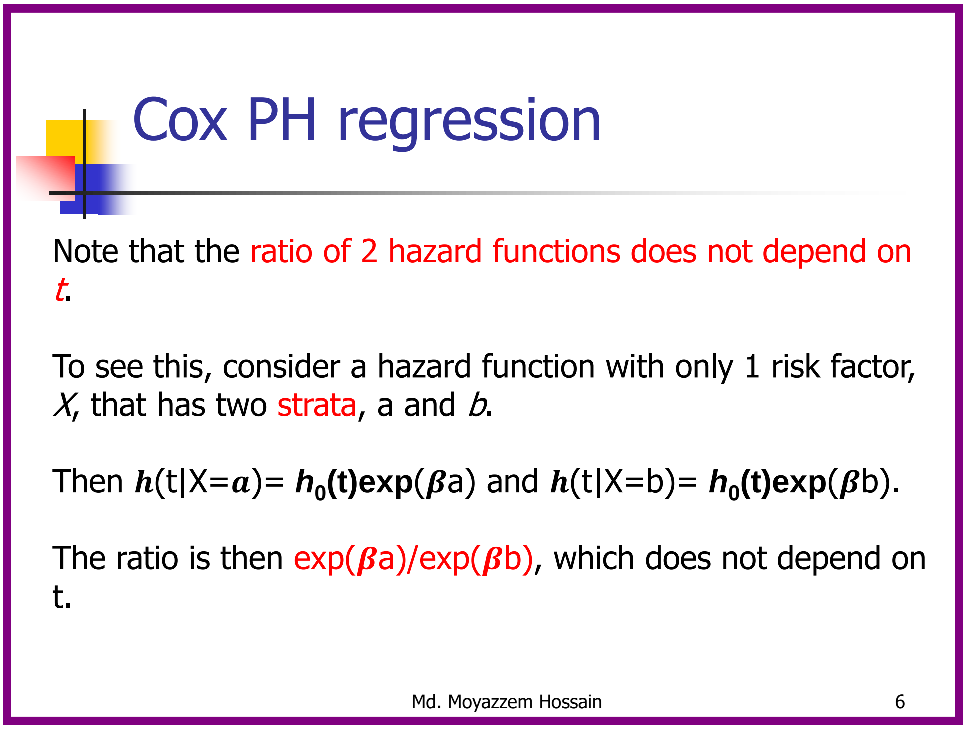
Cox PH regression

Why is Cox model a 'proportional hazards**' model?**

For any two individuals (or groups, i & j) at any point in time, the ratio of their hazards is a constant (a fixed proportional).

For any time t,   hi(t) / hj(t) = C

C may depend on explanatory variables but not on time.



Cox PH regression

Note that the ratio of 2 hazard functions does not depend on *t*.

To see this, consider a hazard function with only 1 risk factor, *X*, that has two strata, a and *b*.

Then $h(t|X=a)= h_0(t)\exp(\beta a)$ and $h(t|X=b)= h_0(t)\exp(\beta b)$.

The ratio is then $\exp(\beta a)/\exp(\beta b)$, which does not depend on t.

Cox PH regression

Why partial likelihood (PL)?

It is easy to write down a model:

$$h(t|x_i) = h_0(t) \exp(\beta_i x_i)$$

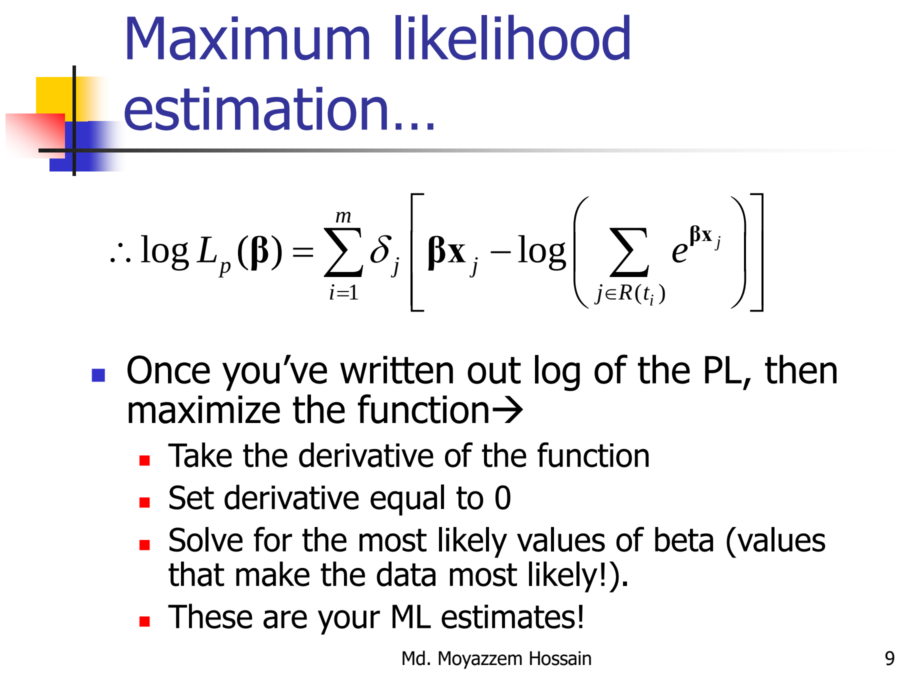
Cox's most important contribution was to propose a method called partial likelihood because it does not include the baseline hazard function $h_0(t)$.

The PL

$$\therefore L_p(\boldsymbol{\beta}) = \prod_{i=1}^{m} \left(\frac{e^{\boldsymbol{\beta}\mathbf{x}_j}}{\sum_{j \in R(t_i)} e^{\boldsymbol{\beta}\mathbf{x}_j}} \right)^{\delta_j}$$

Where, δ_j is the censoring variable (1=if event, 0 if censored)

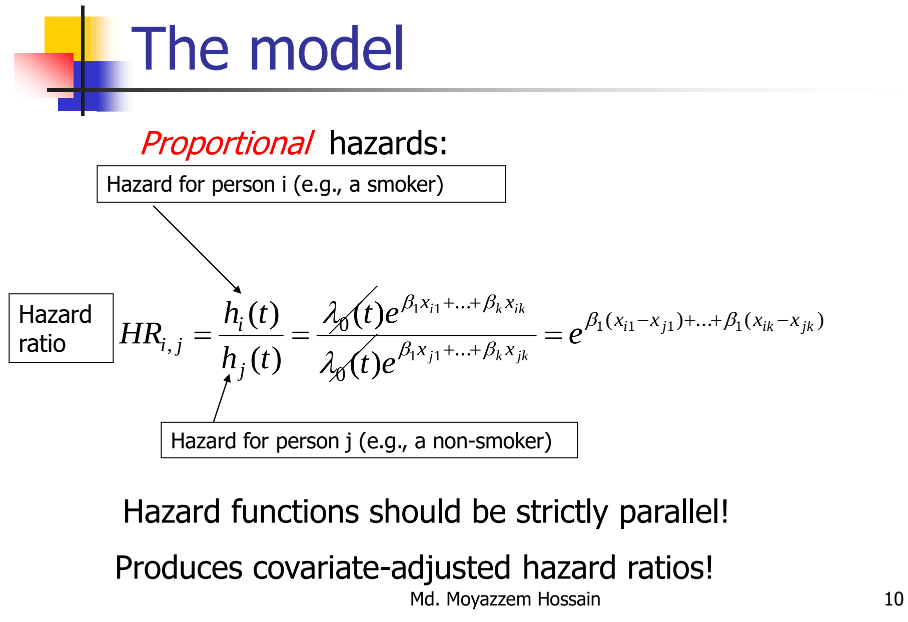
$$\therefore \log L_p(\boldsymbol{\beta}) = \sum_{i=1}^{m} \delta_j \left[\boldsymbol{\beta}\mathbf{x}_j - \log\left(\sum_{j \in R(t_i)} e^{\boldsymbol{\beta}\mathbf{x}_j} \right) \right]$$



Maximum likelihood estimation…

$$\therefore \log L_p(\boldsymbol{\beta}) = \sum_{i=1}^{m} \delta_j \left[\boldsymbol{\beta}\mathbf{x}_j - \log\left(\sum_{j\in R(t_i)} e^{\boldsymbol{\beta}\mathbf{x}_j} \right) \right]$$

- Once you’ve written out log of the PL, then maximize the function→
  - Take the derivative of the function
  - Set derivative equal to 0
  - Solve for the most likely values of beta (values that make the data most likely!).
  - These are your ML estimates!



The model

Proportional hazards:

Hazard for person i (e.g., a smoker)

Hazard ratio

$$HR_{i,j} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{\lambda_0}(t)e^{\beta_1 x_{i1}+\ldots+\beta_k x_{ik}}}{\cancel{\lambda_0}(t)e^{\beta_1 x_{j1}+\ldots+\beta_k x_{jk}}} = e^{\beta_1(x_{i1}-x_{j1})+\ldots+\beta_1(x_{ik}-x_{jk})}$$

Hazard for person j (e.g., a non-smoker)

Hazard functions should be strictly parallel!

Produces covariate-adjusted hazard ratios!

The model: binary predictor

$$HR_{lung\ cancer/smoking} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{\lambda_0}(t)e^{\beta_{smoking}(1)+\beta_{age}\cancel{(60)}}}{\cancel{\lambda_0(t)}e^{\beta_{smoking}(0)+\beta_{age}\cancel{(60)}}} = e^{\beta_{smoking}(1-0)}$$

$$HR_{lung\ cancer/smoking} = e^{\beta_{smoking}}$$

This is the hazard ratio for smoking adjusted for age.

The model: continuous predictor

$$HR_{lung\ cancer/10-years\ \text{increase in age}} = \frac{h_i(t)}{h_j(t)} = \frac{\cancel{\lambda_0(t)}e^{\cancel{\beta_{smoking}(0)}+\beta_{age}(70)}}{\cancel{\lambda_0(t)}e^{\cancel{\beta_{smoking}(0)}+\beta_{age}(60)}} = e^{\beta_{age}(70-60)}$$

$$HR_{lung\ cancer/10-years\ \text{increase in age}} = e^{\beta_{age}(10)}$$

This is the hazard ratio for a 10-year increase in age, adjusted for smoking.

Exponentiating a continuous predictor gives you the hazard ratio for a 1-unit increase in the predictor.

Assumptions of CPH Model

Assumption 1: Independent observations.

• This assumption means that there is no relationship between the subjects in your data set and that information about one subject's survival does not in any way inform the estimated survival of any other subject.

• That is, they are not related to each genetically or in other types of 'clusters', such as health care systems, neighborhoods, places of work, etc.

• This is a key assumption in most statistical models.

Assumptions of CPH Model

Assumption: The survival curves for two different strata of a risk factor must have hazard functions that are proportional over time.

- This assumption is satisfied when the change in hazard from one category to the next does not depend on time.
- That is, a person in one stratum has the same instant relative risk compared to a person in a different stratum, irrespective of how much time has passed.
- This is why the model is called the *proportional* hazards model.

Typical Output of a Cox Regression Model

Estimates						
Variable	Beta	Standard Error	P value	Hazard Ratio exp(Beta)	95% confidence interval	
Diabetes (yes/no)	0.537	0.068	<.0001	**1.711**	**1.498**	**1.954**

- Diabetes was coded as a binary variable (1=yes/ 0=no)
- In this case the beta is positive (0.54) which means that the log of the incidence rate for death in diabetics is higher than in non diabetic patients
- Thus, the hazard ratio (HR) for diabetics compared to non diabetics is $e^{0.54} = 1.71$
- The 95% confidence interval for the hazard ratio is [1.498-1.954]

Conclusion: the mortality of patients with diabetes is higher than in patients without diabetics

Note: should we account for known ***confounders*** (e.g. age) before drawing any conclusions

Typical Output of a Cox Regression Model

Estimates						
Variable	Beta	Standard Error	P value	Hazard Ratio exp(Beta)	95% confidence interval	
Age (continuous)	0.047	0.002	<.0001	1.048	1.043	1.053
Diabetes (yes/no)	0.662	0.068	<.0001	**1.939**	**1.696**	**2.216**

- Cox regression is the tool to account for confounding effects when performing survival analysis
- The output of this multiple Cox regression model shows an effect of both age and diabetes
- The hazard ratio for diabetes increased from 1.711 (see previous slide) to 1.939. This change shows that after accounting for the confounding effect of age, the impact of diabetes on survival is even stronger
- Our previous conclusion: "the mortality of patients with diabetes is higher than in patients without diabetics" is still valid. Moreover, age is indeed an important confounder in the diabetes-mortality relationship

Typical Output of a Cox Regression Model

Estimates						
Variable	Beta	Standard Error	P value	Hazard Ratio exp(Beta)	95% confidence interval	
Age (continuous)	0.047	0.002	<.0001	1.048	1.043	1.053
Diabetes (yes/no)	0.662	0.068	<.0001	**1.939**	**1.696**	**2.216**

- The hazard ratio (also known as risk ratio) is the ratio of the hazard functions that correspond to a change of one unit of the given variable and conditional on fixed values of all other variables.
- An increase in one unit of Age increases the hazard of dying by 4.8% (1.048-1).

Note: (if beta is negative, i.e., **-0.11899 and the corresponding hazard ratio is 0.888.** Then increase in one unit of Age decreases the hazard of dying by 11.2% (1-0.888).

Example

- We consider the “lung” dataset in the `survival` package of R that contains information about survival times and censoring status for patients with advanced lung cancer. In this dataset, the following variables are available:
- inst: Institution code.
- time: This variable represents the survival time or the time until death (measured in days).
- status: This variable indicates the censoring status. A value of 1 represents an observed event (death), and a value of 0 represents censoring (individuals who were still alive at the end of the study).
- sex: The gender of the patient, coded as 1 for male and 2 for female.
- age: The age of the patient at the time of diagnosis.
- ph.ecog: The performance status of the patient, measured on the ECOG scale (Eastern Cooperative Oncology Group). It is a categorical variable representing the overall health and activity level of the patient. Common values include 0 (fully active), 1 (restricted activity but ambulatory), 2 (ambulatory but unable to work), and so on.
- ph.karno: The Karnofsky performance score, another measure of the patient's ability to perform normal daily activities.
- pat.karno: The Karnofsky performance score for the patient's spouse or partner.
- meal.cal: The number of calories consumed during a meal.
- wt.loss: Weight loss in the last six months.

Example

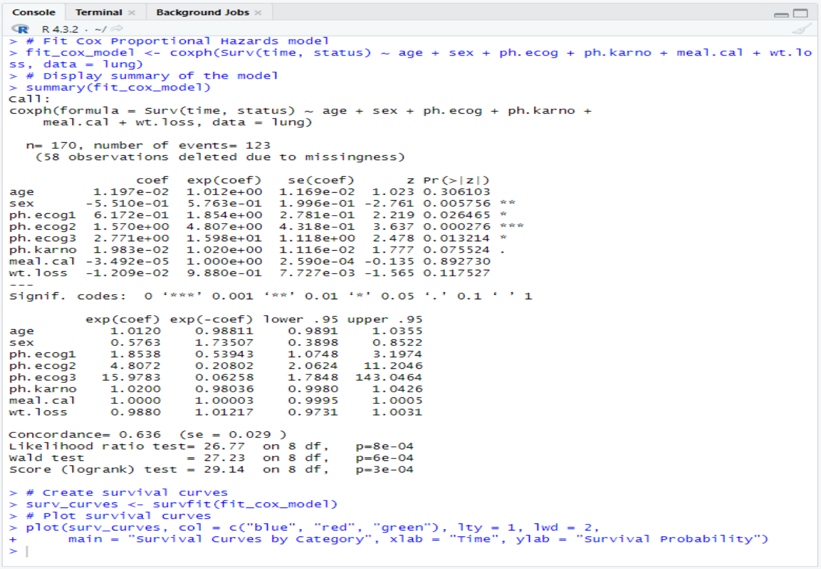
- Suppose we want to include the variables age, sex, the performance status of the patient (ph.ecog), the Karnofsky performance score of the patient, number of calories consumed during a meal, and Weight loss in the last six months to fit the Cox Proportional Hazards model.

- We use `coxph` function to fit the Cox Proportional Hazards model and the commands used in R, are given below:

Example

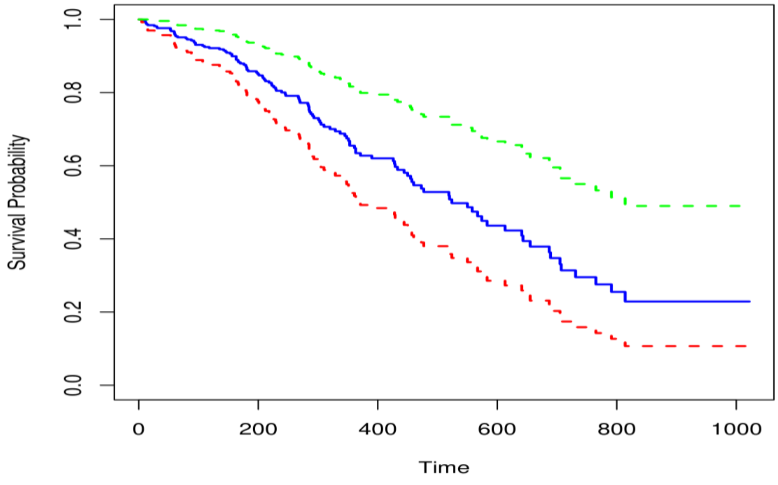
```
# Fit Cox Proportional Hazards model
fit_cox_model <- coxph(Surv(time, status) ~ age + sex +
ph.ecog + ph.karno + meal.cal + wt.loss, data = lung)
# Display summary of the model
summary(fit_cox_model)
# Create survival curves
surv_curves <- survfit(fit_cox_model)
# Plot survival curves
plot(surv_curves, col = c("blue", "red", "green"), lty
= 1, lwd = 2, main = "Survival Curves by Category", xlab
= "Time", ylab = "Survival Probability")
```

Example



```
> # Fit Cox Proportional Hazards model
> fit_cox_model <- coxph(Surv(time, status) ~ age + sex + ph.ecog + ph.karno + meal.cal + wt.lo
ss, data = lung)
> # Display summary of the model
> summary(fit_cox_model)
Call:
coxph(formula = Surv(time, status) ~ age + sex + ph.ecog + ph.karno + 
    meal.cal + wt.loss, data = lung)

  n= 170, number of events= 123 
   (58 observations deleted due to missingness)

                coef  exp(coef)   se(coef)      z Pr(>|z|)    
age        1.197e-02  1.012e+00  1.169e-02  1.023 0.306103    
sex       -5.510e-01  5.763e-01  1.996e-01 -2.761 0.005756 ** 
ph.ecog1   6.172e-01  1.854e+00  2.781e-01  2.219 0.026465 *  
ph.ecog2   1.570e+00  4.807e+00  4.318e-01  3.637 0.000276 ***
ph.ecog3   2.771e+00  1.598e+01  1.118e+00  2.478 0.013214 *  
ph.karno   1.983e-02  1.020e+00  1.116e-02  1.777 0.075524 .  
meal.cal  -3.492e-05  1.000e+00  2.590e-04 -0.135 0.892730    
wt.loss   -1.209e-02  9.880e-01  7.727e-03 -1.565 0.117527    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

         exp(coef) exp(-coef) lower .95 upper .95
age         1.0120    0.98811    0.9891    1.0355
sex         0.5763    1.73507    0.3898    0.8522
ph.ecog1    1.8538    0.53943    1.0748    3.1974
ph.ecog2    4.8072    0.20802    2.0624   11.2046
ph.ecog3   15.9783    0.06258    1.7848  143.0464
ph.karno    1.0200    0.98036    0.9980    1.0426
meal.cal    1.0000    1.00003    0.9995    1.0005
wt.loss     0.9880    1.01217    0.9731    1.0031

Concordance= 0.636  (se = 0.029 )
Likelihood ratio test= 26.77  on 8 df,   p=8e-04
Wald test            = 27.23  on 8 df,   p=6e-04
Score (logrank) test = 29.14  on 8 df,   p=3e-04

> # Create survival curves
> surv_curves <- survfit(fit_cox_model)
> # Plot survival curves
> plot(surv_curves, col = c("blue", "red", "green"), lty = 1, lwd = 2,
+      main = "Survival Curves by Category", xlab = "Time", ylab = "Survival Probability")
> |
```

Example



Example

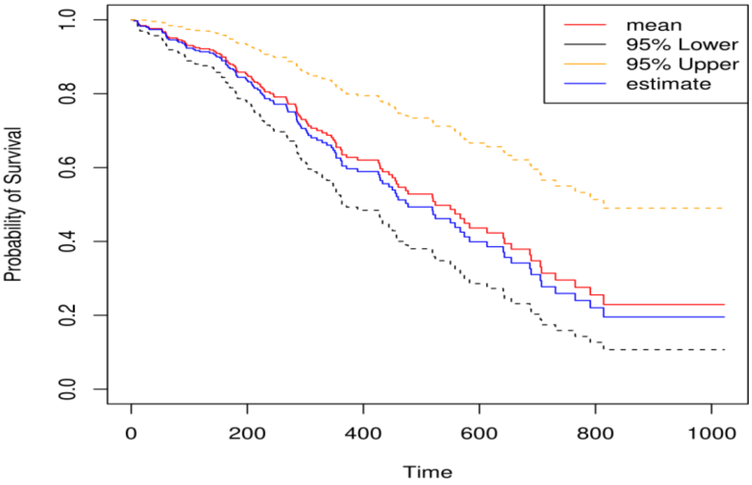
Suppose we want to estimate the survival curve for the specific covariates, i.e., sex= 1, age= 55, ph.ecog = "1", ph.karno = 50, meal.cal = 1150, and wt.loss = 11.

The following R-code produced the survival curve:

Example

```
#create a survival curve given your specified covariates
newdata <- data.frame(age = 57,
                      sex= 1,
                      ph.ecog  = "1",
                      ph.karno = 50,
                      meal.cal = 1175,
                      wt.loss  = 11)

estimate <- survfit(fit_cox_model, newdata = newdata)

plot(survfit(fit_cox_model), ylab = "Probability of Survival",
     xlab = "Time", col = c("red", "black", "orange"))
lines(estimate$time, estimate$surv, col = "blue", type = 's')
legend("topright", legend = c("mean", "95% Lower", "95% Upper",
"estimate"), col = c("red", "black", "orange", "blue"), lty=1)
```

Example



Thank you all.