

# RShiny app: Interactive Survival Analysis of the survival::lung dataset

Shoumi Sarkar

PHC6068 Final Project (Fall 2021)

## Accessing the app

The app is hosted on the shinyapps.io server at <https://shoumi.shinyapps.io/lungsurvanalysis/> (<https://shoumi.shinyapps.io/lungsurvanalysis/>).

## Introduction

The objective of this Rshiny app is to provide an interactive demonstration of survival analysis tools using the `lung` dataset in the `survival` package in R. Survival analysis is a branch of statistics for analyzing the expected duration of time until one event occurs (such as death or relapse of cancer, termed as a “failure”). It attempts to answer questions, such as: what is the proportion of a population which will survive past a certain time? Of those that survive, at what rate will they die or fail? Can multiple causes of death or failure be taken into account? How do particular circumstances or characteristics increase or decrease the probability of survival?

Ideally, to carry out a survival analysis, we would need complete information on times until failure (or “lifetimes”). But rarely do we obtain complete data. In studies such as those following up patients until the failure event, there may be issues like patients dropping out or failing to turn up for their scheduled check-ups. This case of missing data is known as right-censoring. We do not know the actual failure times for such cases but do know the fact that a failure had simply not occurred up till that time.

Mathematically, the data we will work with will be of the following structure:  $(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)$ , denoting the failure times  $t_i$  and censoring status  $\delta_i$  ( $\delta_i = 0$  if censored,  $\delta_i = 1$  if not censored) on  $n$  individuals, along with additional explanatory variables. If  $F(t)$  denotes the CDF of the failure time at time-point  $t$ , then  $S(t) = 1 - F(t)$  is the survival function at time  $t$ , synonymous with the probability of surviving past time  $t$ .  $H(t) = -\ln S(t)$  is known as the cumulative hazard function at time  $t$ . Its derivative  $h(t) = H'(t)$  denotes the hazard function at time  $t$ , and can be interpreted as a speedometer of risk - when  $h(t)$  is high, the subject is likely to fail if they haven't already.  $H(t)$  is thus a cumulative sum of this risk.

We will start with demonstrating **some basic measures of survival** (point prevalence, period prevalence and incidence rate ratios) which help us have an idea of how rampant or prevalent the disease/condition is in the population and how its incidence rate compares among two groups (such as a comparison between females to males). The mathematics behind it are described in the next section. In our app, the user can select any time-point or time interval and get reactive outputs and interpretations for these measures.

Next, we will use **parametric distributions to fit Accelerated Failure Time (AFT) models** to model the survival data, with the user's choice of covariates. Depending on the user inputs, a model is built and **significant predictors affecting survival** (at 0.05 level) are reported, along with a **Cox-Snell residual plot to judge the goodness of fit** for the built model as a diagnostic measure.

However, the problem with parametric models is that incorrect specification of the true model can lead to erroneous fitting. So we also fit the **nonparametric Nelson-Aalen estimator** of cumulative hazard of failure and **compare them** with the cumulative hazards obtained from the parametric models.

Another alternative is **building a (semi-parametric) Cox proportional hazards model** that assumes that the ratio of the hazards for any two individuals is constant over time. We can obtain significant predictors under this assumption and judge the fit with Cox-Snell residuals.

Lastly, we can have a **log-rank test** comparing survival rates between two groups (sex groups: male and female). As this comparison may differ along different intervals of time, we let the user choose the time interval to carry out this test at. We report the test results and plot the survival curves by sex.

## Methods

In this section we describe the theory behind our suggested measures and R functions used to implement the RShiny app. It will be useful to denote the concept of the **risk set** in survival analysis. The risk set  $R(t)$  denotes all subjects who fail at time  $t$ , are censored at time  $t$ , or who survive past time  $t$ . Similarly, **a subject is said to be at risk at time  $t$**  if he fails at time  $t$ , is censored at time  $t$ , or survives past time  $t$ . Another term is person-time: it is an estimate of the actual time-at-risk – in years, months, or days – that all participants contributed to a study, that is, the sum of all their failure times.

## Prevalence Measures

The **point prevalence** of disease at time  $t$  is the proportion of the population that has disease/condition at time  $t$ . The **period prevalence** of a disease in time interval  $(t_a, t_b]$  is the proportion of of the population having the disease/condition in  $(t_a, t_b]$ .

The **incidence rate ratio** of a disease/condition between two groups is the ratio of incidence of the disease (incidence being defined as number of deaths divided by total person-time) in the two groups. Mathematically, it is  $IRR = \frac{m_2/T_2}{m_1/T_1}$  with  $m_i$  being the number of deaths/events in the group and  $T_i$  being the total person-time (sum of failure times) for that group.

In our app, we implement these calculations reactively so that the user can select the timepoints and get relevant interpretations.

## Parametric Fitting: AFT Models

In an accelerated failure time (AFT) regression model, the covariates act multiplicatively on the survival time:

$$T_i = \exp(\beta_1 X_{i1} + \dots + \beta_k X_{ik}) * \tau_i$$

where  $\tau_i$  is a random sample from the survival time distribution for a person with  $X_{ij} = 0, i = 1, 2, \dots, n$ . Possible distributions include the exponential, Weibull and log-logistic distributions.

Let  $\lambda$  be the rate parameter of the exponential distribution. In an **exponential AFT regression model**, the scale and rate parameters are respectively given by

$$\sigma(X, \beta) = \exp(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)$$

and

$$\lambda(X, \beta) = 1/\sigma(X, \beta) = \exp(-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k))$$

The cumulative hazard function is  $H(t, X, \beta) = t\lambda(X, \beta)$ .

Let  $\lambda$  and  $\gamma$  be the rate and shape parameters of a Weibull distribution. For a **Weibull AFT regression model**, the scale and rates are given by the same expressions as above. The Weibull cumulative hazard function is

$$H(t, X, \beta, \gamma) = t^\gamma \exp(-\gamma(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)).$$

Let  $\lambda$  and  $\gamma$  be the rate and shape parameters of a log-logistic distribution. For a **Log-logistic AFT regression model**, the scale and rates are given by the same expressions above. The cumulative hazard function is

$$H(t, X, \beta, \gamma) = \ln(1 + t^\gamma \exp(-\gamma(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k))).$$

In R, AFT models can be fitted using the `survreg` function in the `survival` package.

## Non-parametric Fitting: Nelson-Aalen Estimator

Let  $n_j$  be the number of people at risk in  $(t_{j-1}, t_j]$  and let  $d_j \geq 0$  be the number of failures at  $t_j$ . The **Nelson-Aalen (NA) estimator** of the cumulative hazard function  $H(t)$  is  $\tilde{H}(t) = \sum_{j:t_j \leq t} \frac{d_j}{n_j}$ . The `survfit()` function allows us to compute the NA estimator.

## Semiparametric Fitting: Cox Proportional Hazards Model

The Cox Proportional Hazards model assumes that the hazard function is of the form

$$h(t, X, \beta) = \exp(\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k) h_0(t)$$

where  $h_0(t)$  is the unspecified baseline hazard function. The parametric component of the model is the first term, where each covariate  $X_j$  has a multiplicative effect on the hazard function (hence the name proportional hazards). The nonparametric component of the model is the second term, the unspecified baseline hazard function. The Cox model is fit using the method of partial likelihood, which involves maximizing the partial likelihood

$$L(\beta) = \prod_{i=1}^m \frac{\exp(\beta X_i)}{\sum_{j \in R(t_i)} \exp(\beta X_j)}$$

In R, it can be implemented with the function `coxph` in the `survival` library.

In case there are tied survival times, the ties can be corrected for by using either of the following three methods: exact (which works with the exact partial likelihood), Efron's approximation or Breslow's approximation. The algebra behind it is a bit complicated, but it can be easily implemented in R by specifying `ties="exact"`, `ties="efron"` or `ties="breslow"` in `coxph()`.

## Cox-Snell Residuals

If  $X$  is a continuous random variable with CDF  $F(x)$ , the random variable  $F(X)$  has a uniform(0,1) distribution. If  $T$  has a failure time with survival function  $S(t)$ , then  $S(T)$  has a uniform(0,1) distribution. The negative logarithm of a  $U(0, 1)$  random variable has an exponential distribution, so the cumulative hazard  $H(T) = -\ln S(T)$  has an exponential(1) distribution.

For both AFT and Cox models, the Cox-Snell residuals  $\hat{H}_i$  is the estimated cumulative hazard for person  $i$  at his or her follow-up time. Under right-censoring, the data  $(\hat{H}_1, \delta_1), \dots, (\hat{H}_n, \delta_n)$  acts as a right-censored sample from an exponential(1) distribution. To test the goodness-of-fit of a model, we can calculate the Nelson-Aalen estimate for this data and compare it to the graph of  $H(t) = t$  from the exponential(1) distribution.

# Log-Rank Test

The log-rank test has the null hypothesis  $H_0$  that two groups (in the `lung` example, 1=males and 2=females) have the same failure time distribution. Let  $t_1 < t_2 < \dots < t_m$  be the distinct analysis times where failures occur. Let  $d_i$  denote the number of failures at time  $t_i$ , and let  $n_i$  denote the number of people in the risk set  $R_i$  (i.e. the set of people at risk of failure at time  $t_i$ ). Let  $X$  be a binary covariate (1=male, 2=female). Let  $n_{2i}$  denote the number of people in  $R_i$  with  $X = 2$ , and let  $d_{2i}$  denote the number of these that fail at time  $t_i$ . Denoting  $p_{2i} = n_{2i}/n_i$ ,  $e_{2i} = d_i p_{2i}$ ,  $U = \sum_{i=1}^m (d_{2i} - e_{2i})$ , the log-rank test statistic is given by

$$S_{log-rank} = \frac{U^2}{V} \sim \chi^2_1$$

where  $V = \sum_{i=1}^m p_{2i}(1 - p_{2i}) = \sum_{i=1}^m \frac{n_{2i}(n_i - n_{2i})}{n_i^2}$  in the case of no ties and  $V = \sum_{i=1}^m p_{2i}(1 - p_{2i}) \frac{d_i(n_i - d_i)}{n_i - 1}$  when there are ties (in our `lung` dataset, there are ties and hence the latter  $V$  will be used).

## Results

Here is a snapshot of the RShiny dashboard:

Interactive Survival Analysis of the survival::lung dataset

We present an interactive survival analysis tool to analyze the NCCTG "lung" data in the survival package: It contains survival data on patients with advanced lung cancer from the North Central Cancer Treatment Group.

The variables in this dataset include:  
**inst**: institution code,  
**time**: survival time in days,  
**status**: censoring status (1=censored, 2=dead),  
**age**: age in years,  
**sex**: male=1, female=2,  
**ph.ecog**: ECOG performance score as rated by the physician. (0=asymptomatic, 1= symptomatic but completely ambulatory, 2= in bed <50% of the day, 3= in bed > 50% of the day but not bedbound, 4 = bedbound),  
**ph.karno**: Karnofsky performance score (bad=0-good=100) rated by physician,  
**pat.karno**: Karnofsky performance score as rated by patient,  
**meal.cal**: Calories consumed at meals,  
**wt.loss**: Weight loss in last six months (pounds).

Performance scores rate how well the patient can perform usual daily activities. We present a preview of the data, provide interactive tools to measure prevalence, build parametric (we cover exponential, Weibull and log-logistic distributions), non-parametric and semi-parametric models of our choice, assess their goodness of fit, and finally compare groups (males and females) across user-specified time intervals to see if the survivals differ by sex.

PREVALENCE MEASURESPARAMETRIC FITTINGNON-PARAMETRIC FITTINGSEMIPARAMETRIC (COX PH) FITTINGLOG-RANK TEST

A preview of the NCCTG Lung Cancer dataset

Show10entries

Search:

	inst	time	status	age	sex	ph.ecog	ph.karno	pat.karno	meal.cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90		15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90		0
6	12	1022	1	74	1	1	50	80	513	0
7	7	310	2	68	2	2	70	60	384	10
8	11	361	2	71	2	2	60	80	538	1
9	1	218	2	53	1	1	70	80	825	16
10	7	166	2	61	1	2	70	70	271	34

Showing 1 to 10 of 228 entries

Previous12345...23Next

Point Measures of Prevalence:

Play around with the slider bars to get interpretations specific to your choice!

Choose a timepoint:

5

107

209

311

413

515

617

719

821

923

1,022

520

1,022

Point Prevalence

The point prevalence of death among the lung cancer patients at time 520 is 0.004386. It is the number of people dying at time 520 (which is 1), divided by the total size of the population, 228.

Our Rshiny dashboard has a sidebar with information on the NCCTG `lung` dataset, and the following tabs.

PREVALENCE MEASURESPARAMETRIC FITTINGNON-PARAMETRIC FITTINGSEMIPARAMETRIC (COX PH) FITTINGLOG-RANK TEST

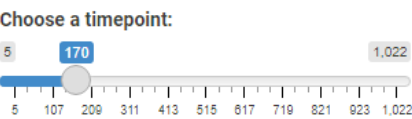
We address the buttons in each tab one by one.

## Tab 1: Prevalence Measures

In this tab, we first present a display of the lung cancer data. Then, we have interactive slider bars to choose a time-point as per the user's choice to calculate point prevalence. The calculation and interpretation is reactive and refreshes for each new choice entered by the user.

# Point Measures of Prevalence:

Play around with the slider bars to get interpretations specific to your choice!



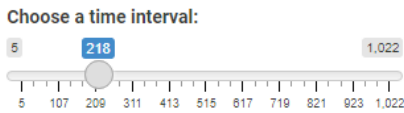
## Point Prevalence

The point prevalence of death among the lung cancer patients at time 170 is 0.004386. It is the number of people dying at time 170 (which is 1), divided by the total size of the population, 228.

A sample choice of time-point and its corresponding output message.

Similarly, we have a slider bar to select time periods for calculations like period prevalence and incidence rate ratio that require a range of time. This is reactive as well, and displays error messages if the start and endpoints of the time intervals are not the same, or if the interval does not contain subjects of two groups (i.e. of two sexes) to be able to calculate incidence rate ratio of females to males.

# Period Measures of Prevalence:



## Period Prevalence

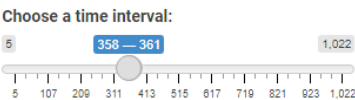
Try again with distinct time points!

## Incidence Rate Ratio

Try again with distinct time points!

Error message when time interval is of length zero.

# Period Measures of Prevalence:



## Period Prevalence

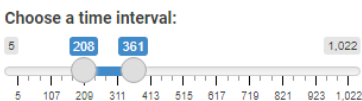
The period prevalence of death among the lung cancer patients in the time interval (358, 361] is 0.004386. It is the number of people dying in the time interval (358,361], (which is 1), divided by the total size of the population, 228.

## Incidence Rate Ratio

For the chosen time period (358, 361], the incidence rate of death among female lung cancer patients is 1 and there are no male observations. As we do not have observations on both genders for this time interval, it is not meaningful to compute the incidence rate ratio. Try setting a wider time interval!

Error message for Incidence Rate Ratio when time interval is too short to contain subjects of both sexes.

# Period Measures of Prevalence:



## Period Prevalence

The period prevalence of death among the lung cancer patients in the time interval (208, 361] is 0.179825. It is the number of people dying in the time interval (208,361], (which is 1), divided by the total size of the population, 228.

## Incidence Rate Ratio

For the chosen time period (208, 361], the incidence rate of death among female lung cancer patients is 0.521739 and that among male patients is 0.69697. The incidence rate ratio comparing females to males is 0.748582, with confidence interval (0.372491, 1.504399).

Output with no warnings.

# Tab 2: Parametric Fitting

In this tab we can choose multiple predictors from the selector menu to build an AFT regression model of our choice:

PREVALENCE MEASURES

PARAMETRIC FITTING

NON-PARAMETRIC FITTING

SEMIPARAMETRIC (COX PH) FITTING

LOG-RANK TEST

Here we fit Accelerated Failure Time (AFT) regression models to the lung data.

Select covariates to keep in the regression - by default, all regressors are selected (the covariates "time" and "status" are always selected as it defines the censored survival times):

Choose variable(s):

age status inst sex pat.karno

time						
ph.ecog						
ph.karno						
meal.cal						
wt.loss						
4	210.00	2.00	57.00	5.00	1.00	60.00
5	883.00	2.00	60.00	1.00	1.00	90.00
6	1022.00	1.00	74.00	12.00	1.00	80.00
7	310.00	2.00	68.00	7.00	2.00	60.00
8	361.00	2.00	71.00	11.00	2.00	80.00

Further, we can specify the parametric distribution for the AFT model: exponential, Weibull or log-logistic from the drop down menu.

Select parametric distribution for the AFT model:

Distribution:

Weibull

Exponential

Weibull

Log-logistic

```
survreg(formula = Surv(time, status) ~ ., data = subsetted, dist = "weib")
              Value Std. Error      z      p
(Intercept)  5.19478    0.59048  8.80 < 2e-16
age          -0.00832    0.00687 -1.21 0.22627
inst          0.00738    0.00779  0.95 0.34361
sex2          0.37097    0.12532  2.96 0.00308
pat.karno     0.01436    0.00416  3.45 0.00056
Log(scale)   -0.30998    0.06207 -4.99 5.9e-07

Scale= 0.733

Weibull distribution
Loglik(model)= -1115.2  Loglik(intercept only)= -1127.6
      Chisq= 24.79 on 4 degrees of freedom, p= 5.6e-05
Number of Newton-Raphson Iterations: 5
n=224 (4 observations deleted due to missingness)
```

We get a list of significant predictors (0.05 level) for the model that we build.

### Significant predictors in the above model

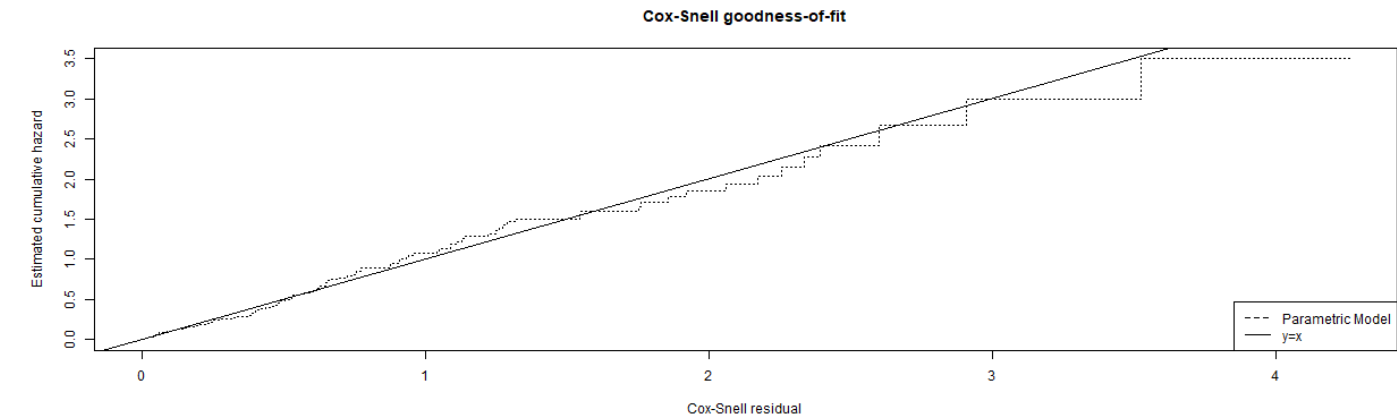
(Play around with the predictor and distribution choices to see what predictors turn out to be significant!)

The Weibull AFT regression model with covariates: (time, status, age, status, inst, sex, pat.karno) yields the following predictors significant at the 0.05 level: sex2, pat.karno.

For the model built, the Cox-Snell residual plot is generated:

### Diagnostics: Cox-Snell Goodness of Fit plots

This refreshes for every new model that we build! The closer the plot is to the line  $y=x$ , the better the fit.



## Tab 3: Non-parametric Fitting

In this tab, a plot of the Nelson-Aalen estimator of cumulative hazard is calculated and the user is given a choice to overlay one or more of the parametric estimates (exponential, weibull, log-logistic) on it for comparison. The legend and plot updates for each new entry entered by the user.

PREVALENCE MEASURESPARAMETRIC FITTINGNON-PARAMETRIC FITTINGSEMIPARAMETRIC (COX PH) FITTINGLOG-RANK TEST

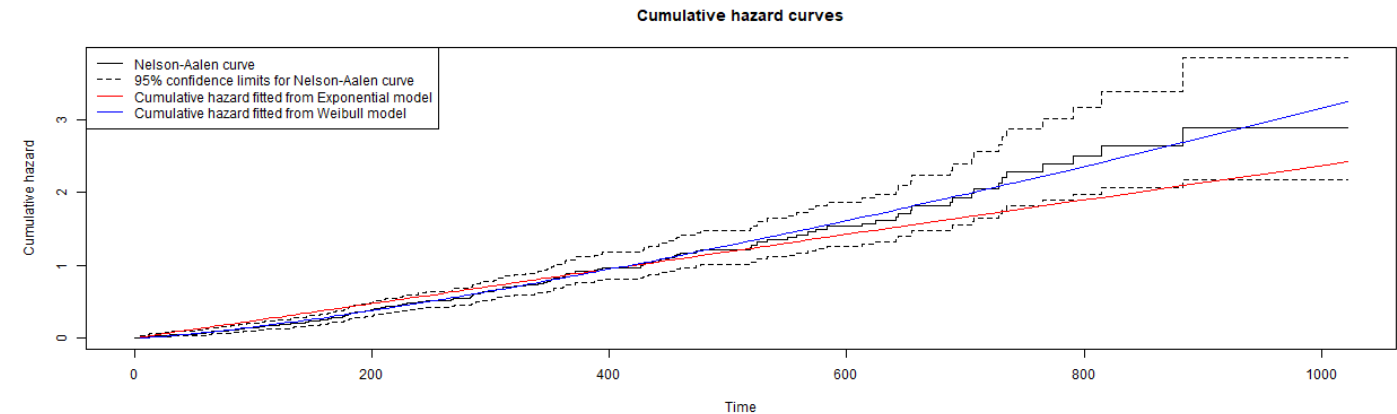
Here we find the Nelson-Aalen estimates of the cumulative hazard function and compare it with its estimated (parametric) curves.

The Nelson-Aalen curve is plotted below. Select distributions from the menu below to compare it with AFT regression models.

Select parametric distribution(s) to compare with:

ExponentialWeibull

Log-logistic



# Tab 4: Semiparametric (Cox PH) Fitting

In this panel, we can once again select regressors to build a Cox Proportional Hazards model of choice.

PREVALENCE MEASURES

PARAMETRIC FITTING

NON-PARAMETRIC FITTING

SEMIPARAMETRIC (COX PH) FITTING

LOG-RANK TEST

We now implement a semiparametric approach: the Cox Proportional Hazards model.

Select covariates below for the Cox PH regression. By default, all covariates are selected (the covariates "time" and "status" are always selected as they define the censored survival times).

Choose variable(s):

inst time status age meal.cal wt.loss

sex

ph.ecog

ph.karno

pat.karno

meal.cal

wt.loss

3	1010.00	1.00	3.00	56.00	NA	15.00
4	210.00	2.00	5.00	57.00	1150.00	11.00
5	883.00	2.00	1.00	60.00	NA	0.00
6	1022.00	1.00	12.00	74.00	513.00	0.00
7	310.00	2.00	7.00	68.00	384.00	10.00
8	361.00	2.00	11.00	71.00	538.00	1.00

We can specify which method to employ to handle ties: exact, Efron's approximation or Breslow's approximation using the radio buttons.

Of the 228 observed survival times, there are only 186 unique values, meaning that there are ties in the data. We can handle ties by the Exact method, Efron's method or Breslow's method. Choose one below!

Select method to handle ties:

☐ Exact

☐ Efron

☒ Breslow

```
Call:
coxph(formula = Surv(time, status) ~ ., data = subsetted2, ties = "breslow")

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

      coef exp(coef) se(coef)      z Pr(>|z|)
inst -1.404e-02  9.861e-01  1.220e-02 -1.151  0.2498
age   2.204e-02  1.022e+00  1.106e-02  1.994  0.0462 *
meal.cal -3.805e-05  1.000e+00  2.460e-04 -0.155  0.8771
wt.loss  5.705e-05  1.000e+00  6.848e-03  0.008  0.9934
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

      exp(coef) exp(-coef) lower .95 upper .95
inst    0.9861    1.0141    0.9628    1.010
age     1.0223    0.9782    1.0004    1.045
meal.cal 1.0000    1.0000    0.9995    1.000
wt.loss  1.0001    0.9999    0.9867    1.014

Concordance= 0.565 (se = 0.031 )
Likelihood ratio test= 5.61 on 4 df,  p=0.2
Wald test            = 5.37 on 4 df,  p=0.3
Score (logrank) test = 5.4 on 4 df,  p=0.2
```

We get the list of significant variables (0.05 level) for the user's choice of model.

Significant predictors in the above Cox PH model

(Play around with the predictor and tie-handling method choices to see what predictors turn out to be significant!)

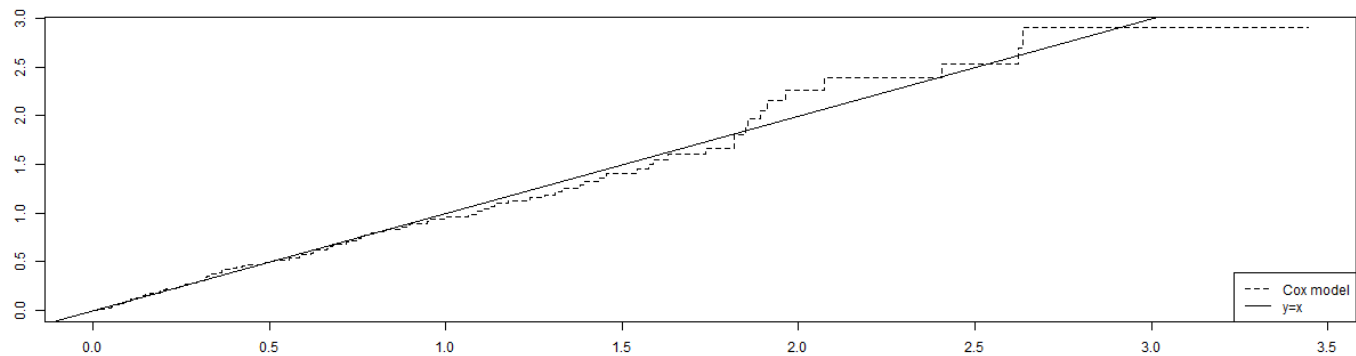
The Cox Proportional Hazards model with covariates: (time, status, inst, time, status, age, meal.cal, wt.loss) yields the following predictors significant at the 0.05 level: age.



We get a Cox-Snell's residual plot specific to the model we just built.

Diagnostics: Cox-Snell Goodness of Fit plots

This refreshes for every new model that we build! The closer the plot is to the line  $y=x$ , the better the fit.



Tab 5: Log-Rank Test

We let the user specify a time interval to consider for carrying out a log-rank test. If the endpoints of the time interval coincide, the following message is printed.

PREVALENCE MEASURES

PARAMETRIC FITTING

NON-PARAMETRIC FITTING

SEMIPARAMETRIC (COX PH) FITTING

LOG-RANK TEST

We compare survival between females (sex=2) and males (sex=1) using the log-rank test. Specify a time interval to consider for this comparison:

Choose a time interval:

54931,022

51072093114135156177198219231,022

Try again with two distinct timepoints!

In case the selected time-interval is too short to include subjects of either group under comparison, the following message is displayed.

We compare survival between females (sex=2) and males (sex=1) using the log-rank test. Specify a time interval to consider for this comparison:

Choose a time interval:

5123 — 1451,022

51072093114135156177198219231,022

There is only one gender group in this interval. Try again with a wider interval!

Without the above two issues, we get a proper output, like this:

We compare survival between females (sex=2) and males (sex=1) using the log-rank test.  
Specify a time interval to consider for this comparison:

Choose a time interval:

5

123

383

1,022

5

107

209

311

413

515

617

719

821

923

1,022

Call:  
survdifff(formula = Surv(time, status) ~ sex, data = subsetted,  
rho = 0)

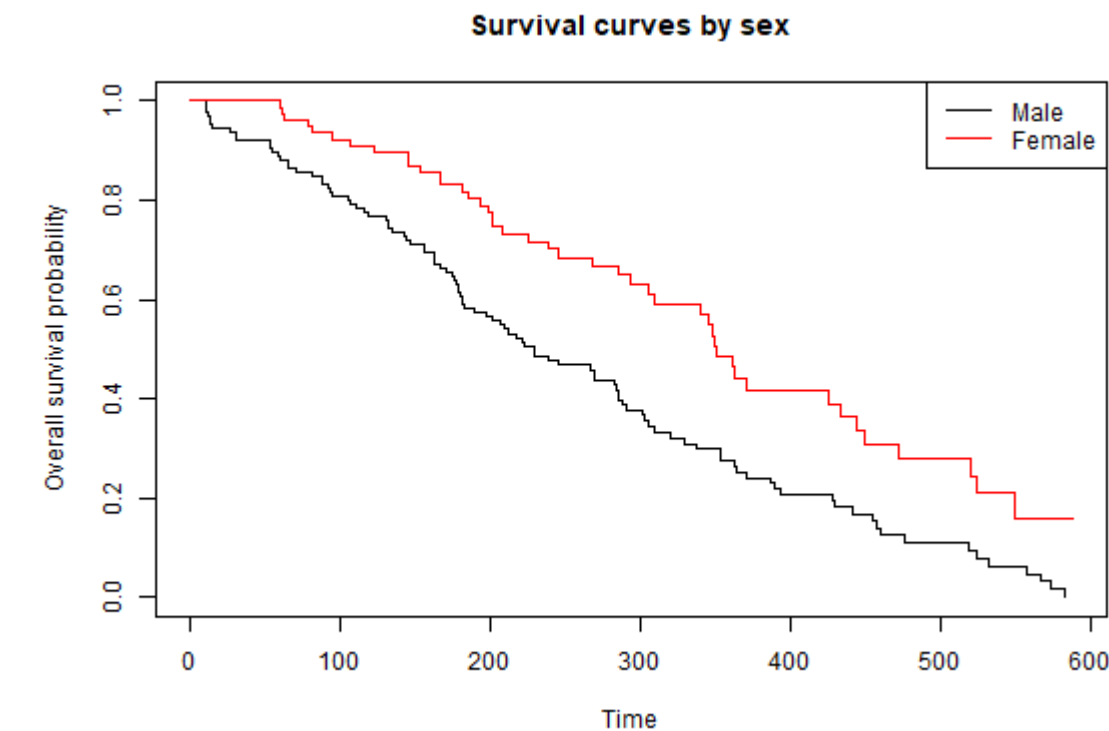
	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
sex=1	75	57	42.2	5.21	10.9
sex=2	52	28	42.8	5.13	10.9

Chisq= 10.9 on 1 degrees of freedom, p= 0.001

With proper input of time interval (i.e. unique time-points, and wide enough to have subjects of both groups), we get a reactive message interpreting the p-value of the test:

The p-value for the log-rank test is 0.000266. As it is < 0.05, we reject the null hypothesis that the survival is same across male and female groups. A plot of the survival curves by sex is as below.

If proper time intervals are entered by the user, survival curves for each group (sex) are also plotted depending on the choice of the time interval for the log-rank test.



# Conclusion

This dashboard allows an interactive analysis of the survival data in the `lung` dataset. The user can play around with choices of time-points, time intervals, regressors, tie handling methods and also explore parametric, non-parametric, semi-parametric model building choices. For every choice of model, the corresponding list of significant regressors, model summary, p-value interpretation, plots and diagnostic residual plots are generated. This helps develop intuition about the popular methods in survival analysis.

There is always scope for improvement and future work. Further methods in survival analysis could be implemented. Besides that, as an advanced improvement to this dashboard, one could add a feature of elegantly allowing the app to accept an uploaded dataset from the user's end. A challenge with this would be variable identification (as one needs to identify the variables for `time` and `status` to be able to work with the survival data, but the naming of these covariates vary from dataset to dataset - some popular datasets call the time variable `times` or `t`, and the variable name for censoring status is also sometimes called `delta`, `stat` etc instead). It becomes difficult to work with the dataset without being able to identify the time and status variables each time. A workaround could be asking the user to explicitly enter variable names or choose columns corresponding to these variables, but it does not seem very elegant. This needs further thought and can be implemented in a future improvement.