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: $(t_1, \delta_1), (t_2, \delta_2), \ldots, (t_n, \delta_n)$, denoting the failure times t_i and censoring status δ_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 probablity 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 accelearated failure time (AFT) regression model, the covariates act multiplicatively on the survival time:

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

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

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

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

and

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

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

Let λ and γ 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

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

Let λ and γ 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

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

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

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

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 continous random variable with CDF $F(x)$, the random variable $F(X)$ has a uniform(0,1) distribution. If T ha 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 $(\hat{H}_1, \delta_1), \ldots, (\hat{H}_n, \delta_n)$ 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

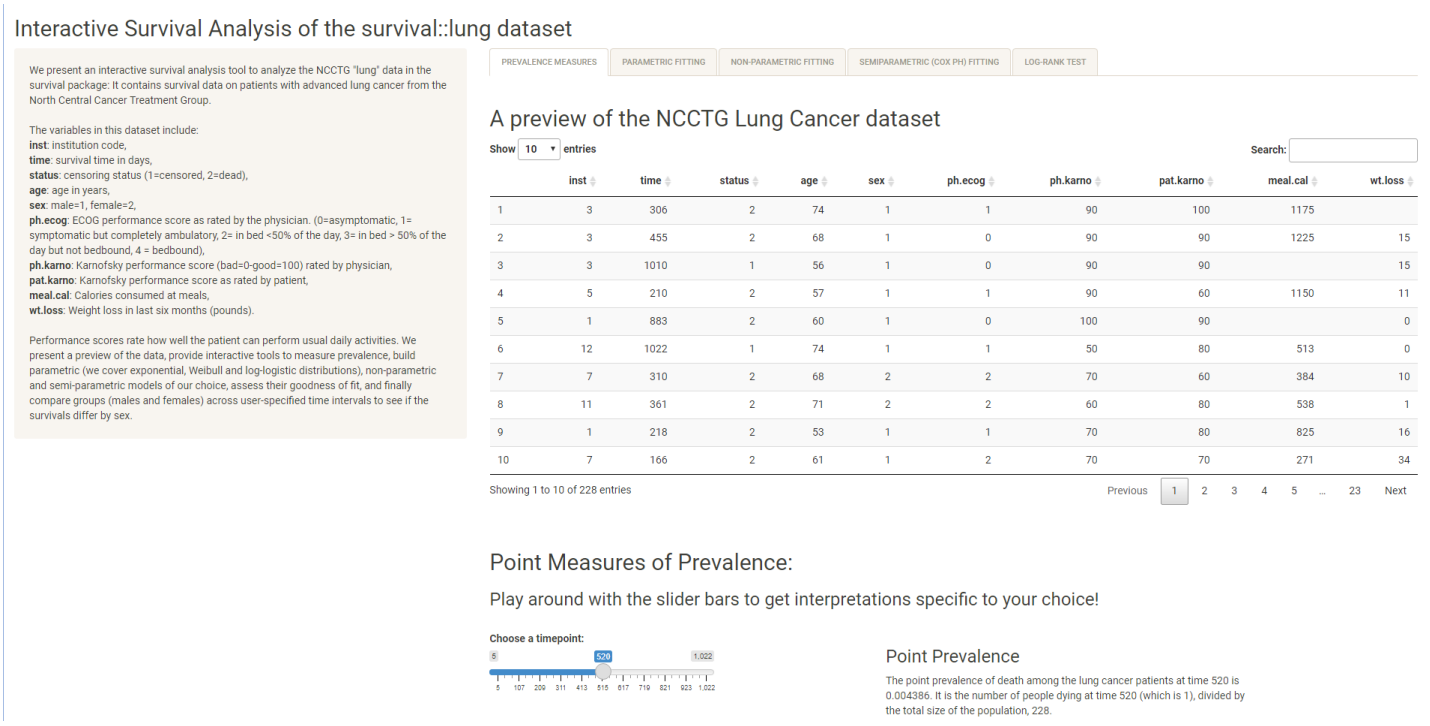
The log-rank test has the null hypothesis H_0 that two groups (in the `lung` example, 1=males and 2=females) have the same failure time distribution. Let $t_1 < t_2 < \ldots < t_m$ be the distinct analysis times where failures occur. Let d_i denote the number of failures at time t_i, and let n_i denote 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

$$S_{log-rank} = \frac{U^2}{V} \sim \chi_1^2$$

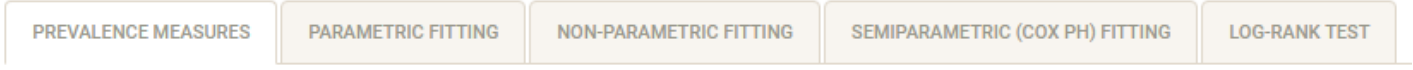
where $V = \sum_{i=1}^{m} p_{2i}(1 - p_{2i}) = \sum_{i=1}^{m} \frac{n_{2i}(n_i - n_{2i})}{n_i^2}$ in the case of no ties and $V = \sum_{i=1}^{m} p_{2i}(1 - p_{2i}) \frac{d_i(n_i - d_i)}{n_i - 1}$ when there are ties (in our `lung` dataset, there are ties and hence the latter V will be used).

Results

Here is a snapshot of the RShiny dashboard:



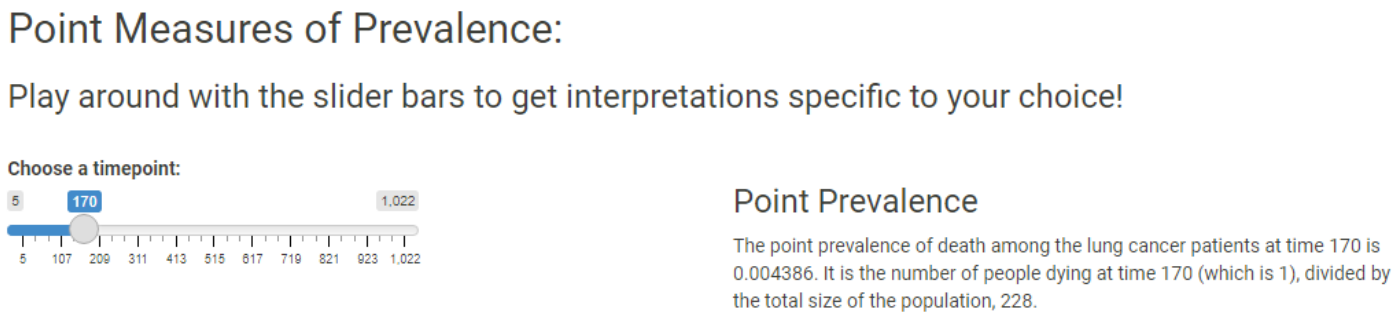
Our Rshiny dashboard has a sidebar with information on the NCCTG `lung` dataset, and the following tabs.



We address the buttons in each tab one by one.

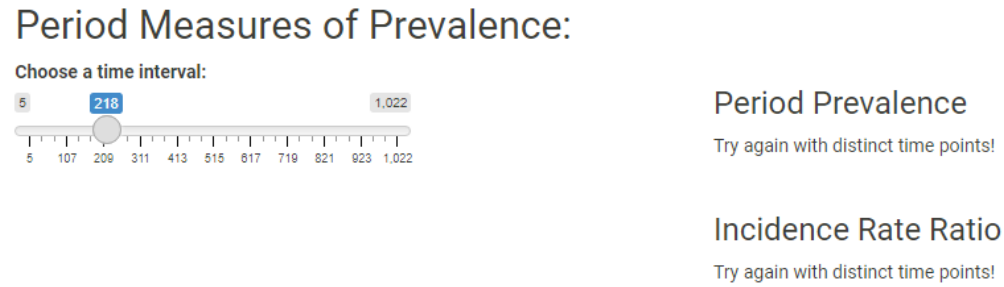
Tab 1: Prevalence Measures

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.

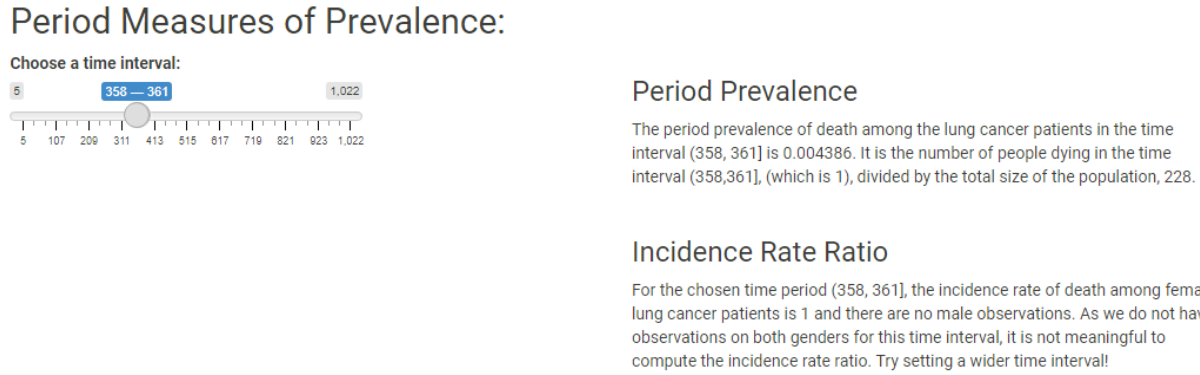


A sample choice of time-point and its corresponding output message.

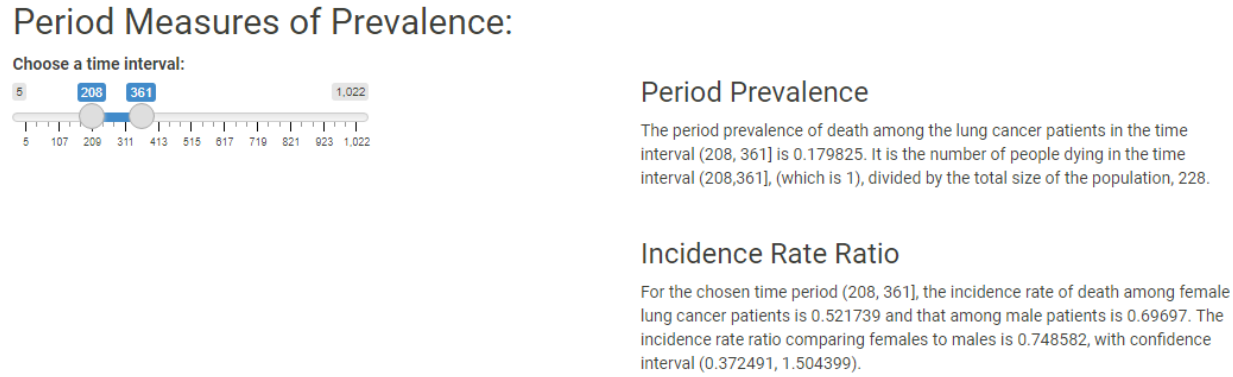
Similarly, we have a slider bar to select time periods for calculations like period prevalence and incidence rate ratio that require a range of time. This is reactive as well, and displays error messages if the start and endpoints of the time intervals are not the same, or if the interval does not contain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



Error message when time interval is of length zero.



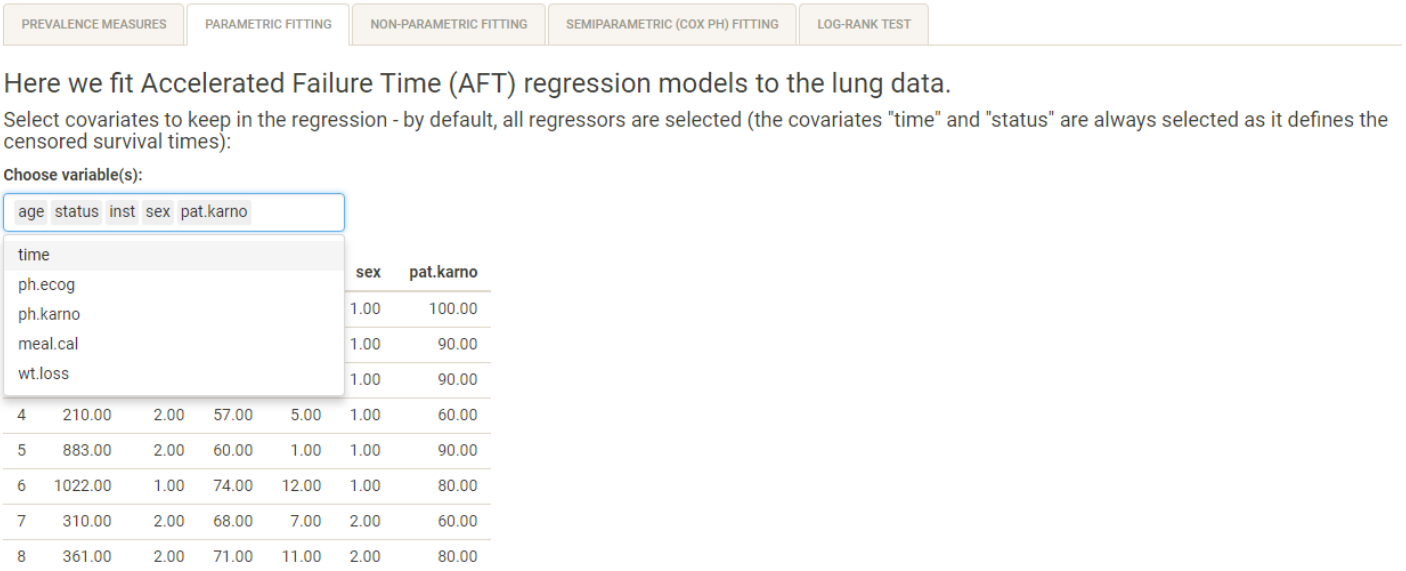
Error message for Incidence Rate Ratio when time interval is too short to contain subjects of both sexes.



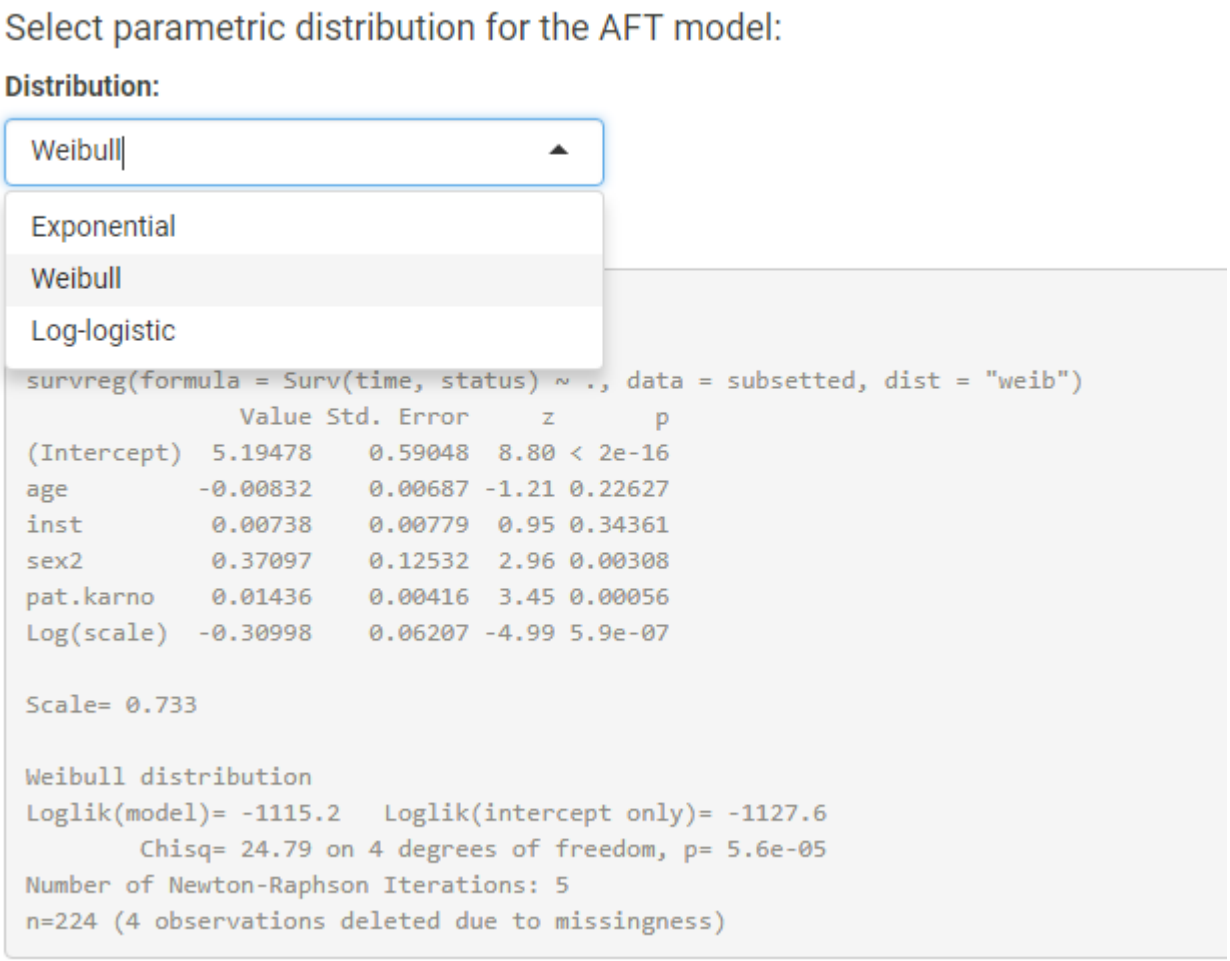
Output with no warnings.

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:



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



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

Significant predictors in the above model

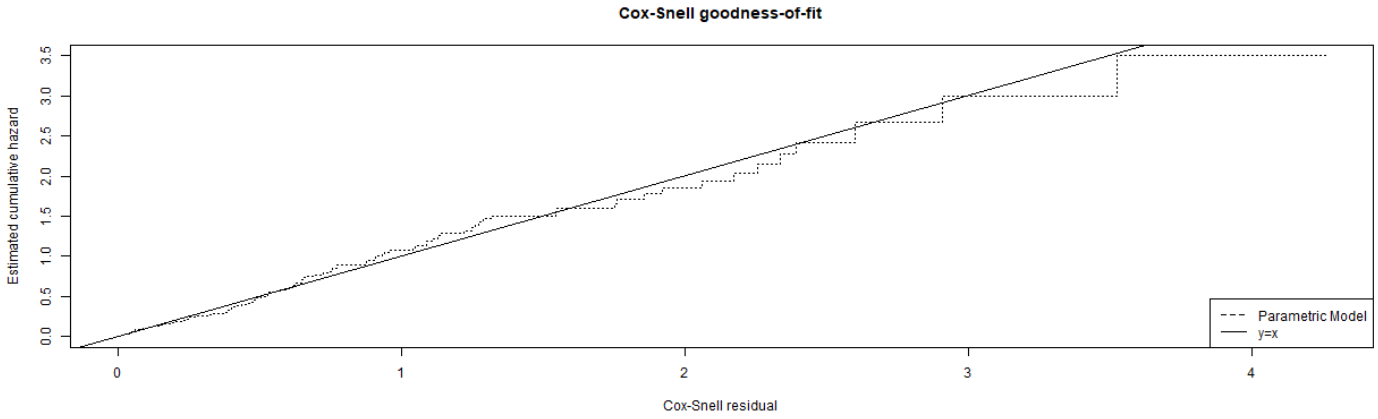
(Play around with the predictor and distribution choices to see what predictors turn out to be significant!)

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 MEASURES | PARAMETRIC FITTING | NON-PARAMETRIC FITTING | SEMIPARAMETRIC (COX PH) FITTING | LOG-RANK TEST

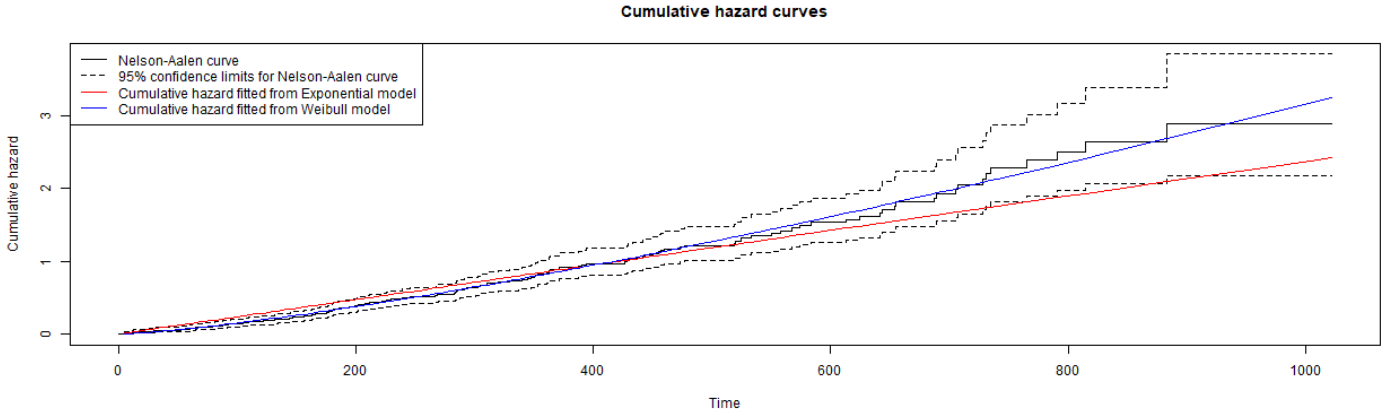
Here we find the Nelson-Aalen estimates of the cumulative hazard function and compare it with its estimated (parametric) curves.

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:

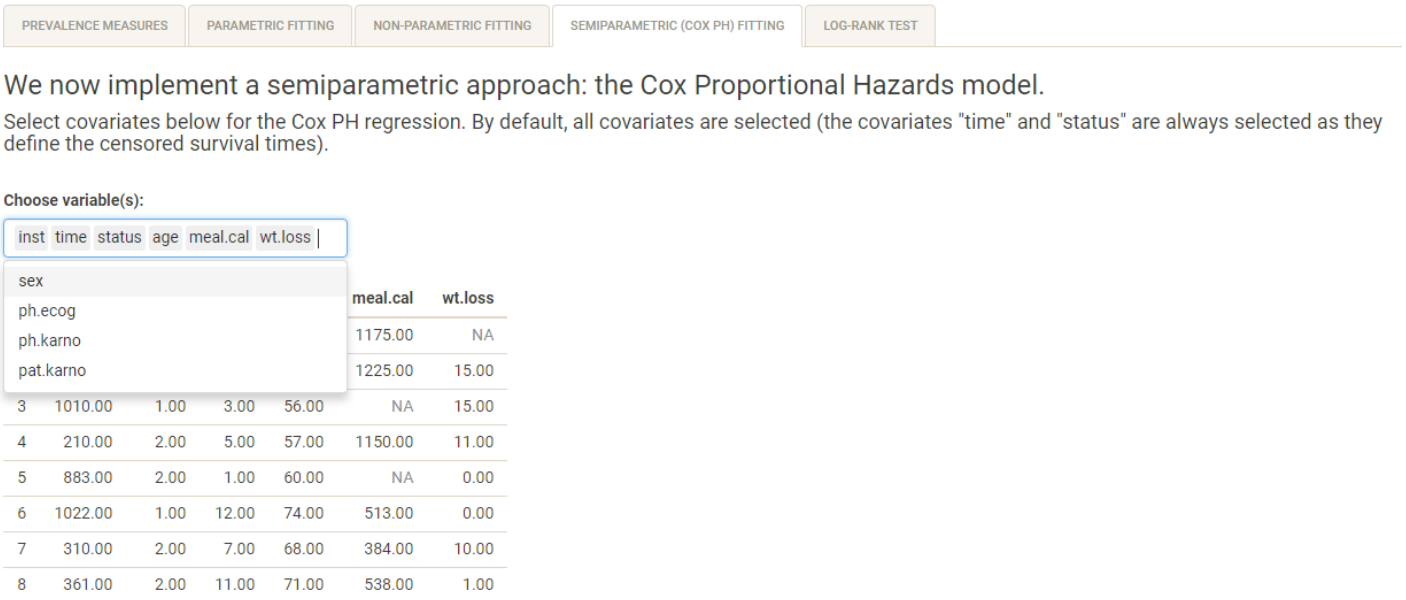
Exponential Weibull

Log-logistic

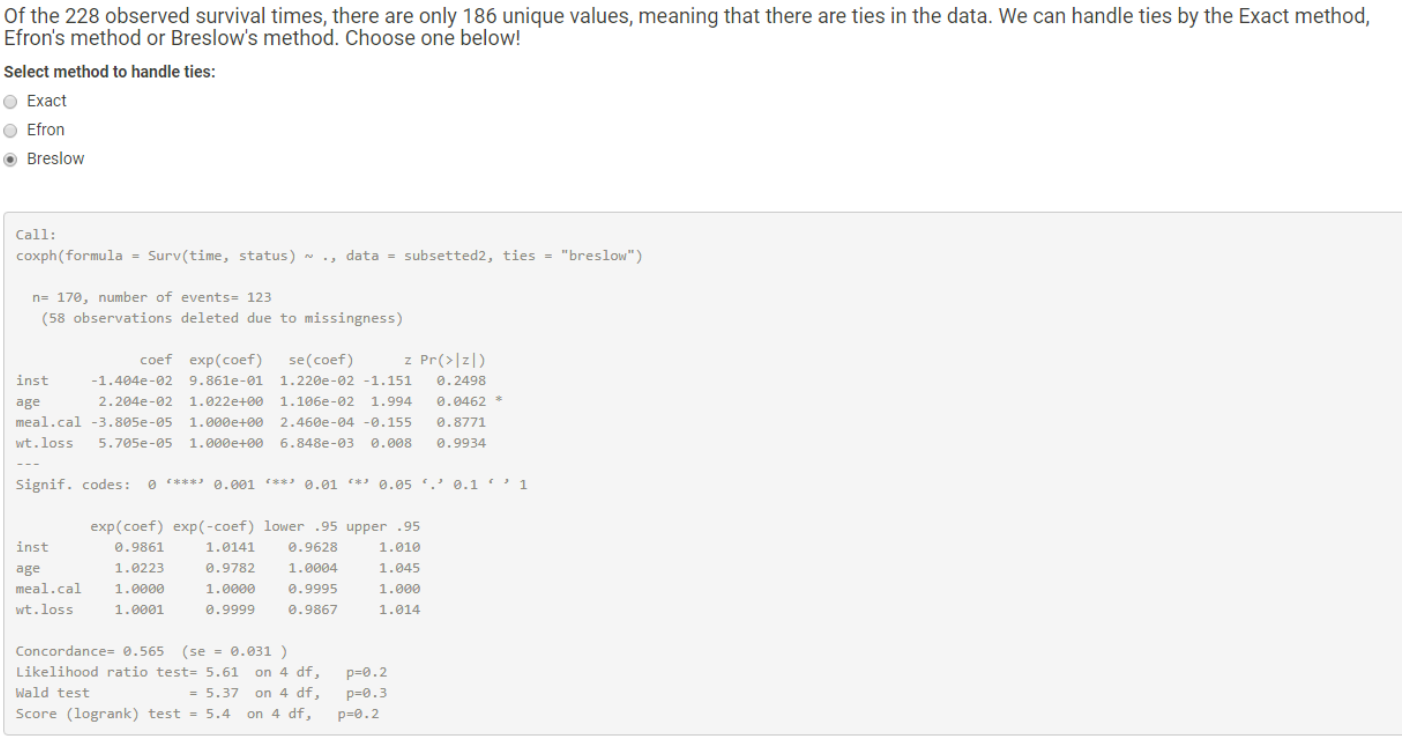


Tab 4: Semiparametric (Cox PH) Fitting

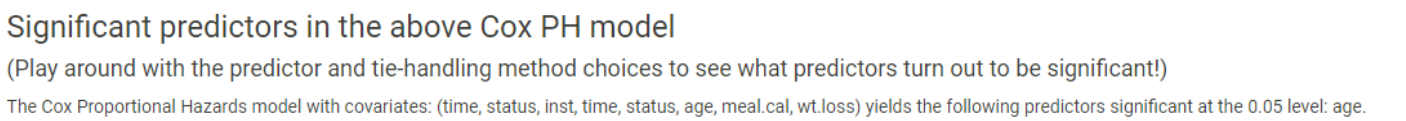
In this panel, we can once again select regressors to build a Cox Proportional Hazards model of choice.



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



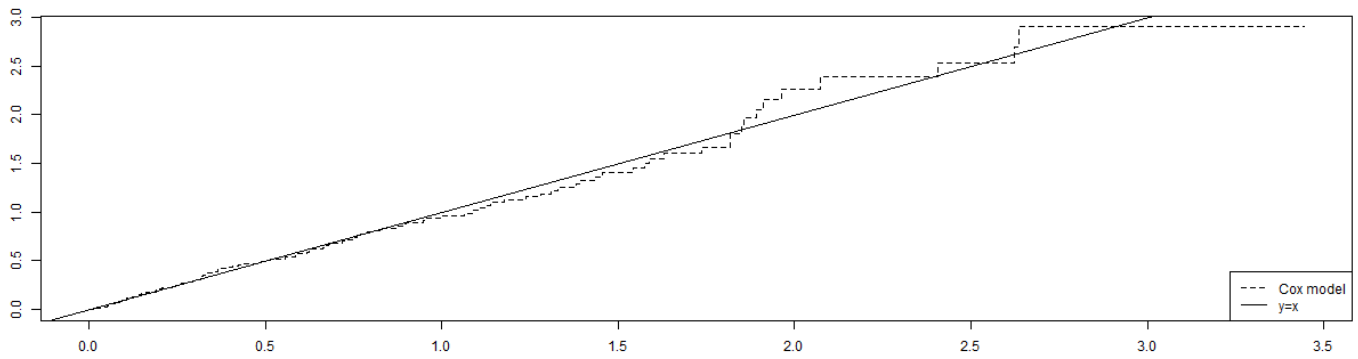
We get the list of significant variables (0.05 level) for the user’s choice of model.



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:

5 493 1,022

5 107 209 311 413 515 617 719 821 923 1,022

```
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:

5 123 — 145 1,022

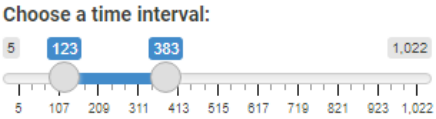
5 107 209 311 413 515 617 719 821 923 1,022

```
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:

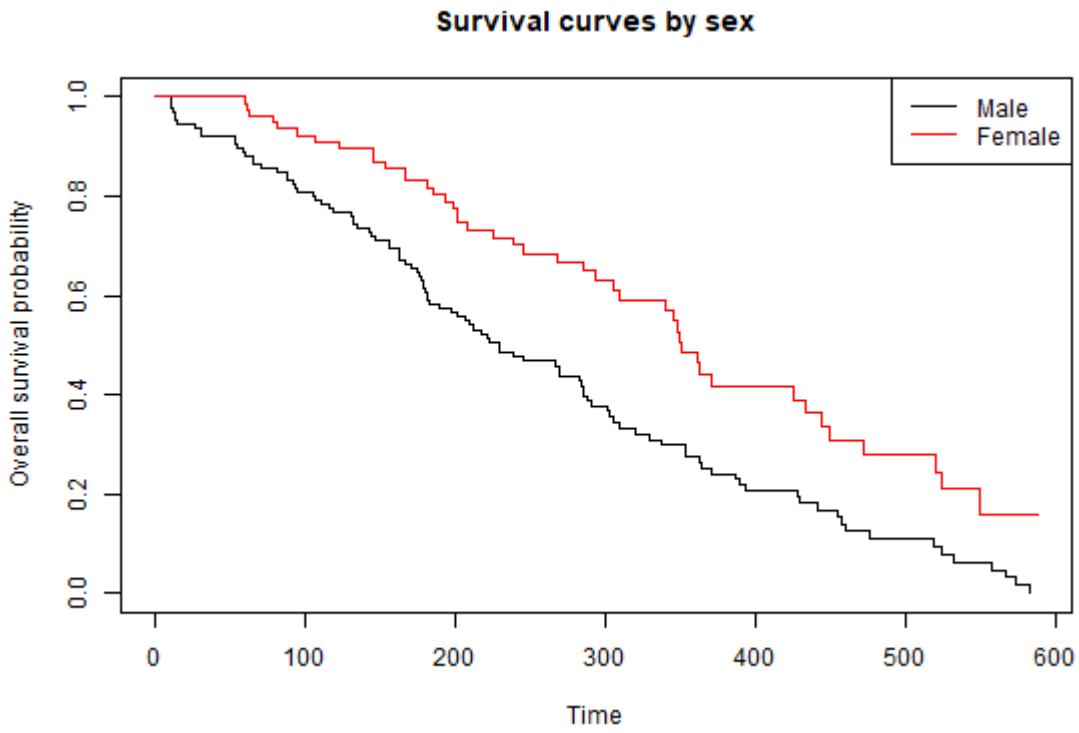



```
Call:
survdiff(formula = Surv(time, status) ~ sex, data = subsetted,
    rho = 0)

       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 need 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.