

Introduction

Lung cancer is one of the most common types of cancer for both men and women. The exploration of survival of patients with lung cancer is crucial for controlling the disease development, obtaining right treatment methods, understanding what influences the disease progression. To accomplish these purposes, accurate survival analysis methods are needed.

Survival analysis is the combination of different statistical methods for analyzing time to event data. Exploring survival models can be challenging, since different models from non-parametric to parametric can be used, various distributions, like exponential, Weibull, log-normal, are applicable for each concrete case. The most common model is Cox hazard model, however, it is too simple and proposes constant effect of predictor variables on survival duration throughout time.

This study examines the way Bayesian approach proceeds to fit Weibull model for lifetime data of patients with advanced lung cancer analysis. Weibull approach is more flexible and hazard rate is not constant through time. For simulation Bayesian inference with MCMC is used, providing us with satisfying approximation of uncertainty and ability to use priors as domain knowledge. The model is implemented and tested with the help of R and Stan package.

The code with model implementation is provided in the following GitHub repository: <https://github.com/yannmclatchie/bda-project>.

Data description

General description

The data, used in the study, shows survival time of patients with advanced lung cancer from the North Central Cancer Treatment Group. It is provided in the R package: <https://stat.ethz.ch/R-manual/R-devel/library/survival/html/lung.html>.

Dataset contains 9 features and 228 observations, which are assumed to be independent and identically distributed. The target variable is the survival time in days. The covariates are presented by both categorical and numerical values.

Special attention has to be paid for “censoring status” feature. It indicates if the patient had an event (=1) or not (=0). If patient is censored, true survival time for him is not known. Right censoring approach is used, meaning incompleteness of survival time at the right side of the follow-up period. We can get rid of it.

There are three variables, needed to be explained: `ph.ecog` - ECOG performance score (0-4). 0-good condition, 4-the worst condition, much time in bad. `ph.karno` - Karnofsky performance score (bad=0-good=100). Provided by physician. `pat.karno` - Karnofsky performance score. Provided by patient.

Exploratory data analysis

It is better to provide some descriptive statistics to familiarize with data.

```
library(dplyr)
library(ggplot2)
data("cancer", package = "survival")
```

There were 108 observations with missed or censored values, which have been removed from dataset. Now it consists of 120 rows.

```
data = cancer %>% dplyr::filter(status == 2) %>% na.omit()
```

Institutions are considered as variables, useful for hierarchical model.

```
table(data$inst)
```

```
##
##  1  2  3  4  5  6  7 10 11 12 13 15 16 21 22 26 32
## 21  3  9  4  6 10  5  4  7 12  7  4  6  8 10  2  2
```

Moving to categorical variables (fig.1), the number of men prevails over women. The majority of patients are ambulatory with symptoms.

```
par(mfrow=c(1,2))
barplot(table(data$sex), main="Sex statistics", names.arg=c("male", "female"),
        col=c("steelblue", "cornflowerblue"))
barplot(table(data$ph.ecog), main="ECOG score statistics",
        legend = c("asymptom.", "ambulatory", "in bed <50% of t", "in bed >50% of t"),
        args.legend = list(x = "topright", inset = c(-0.15, 0)),
        col=c("steelblue", "cornflowerblue", "blue", "darkblue"))
```

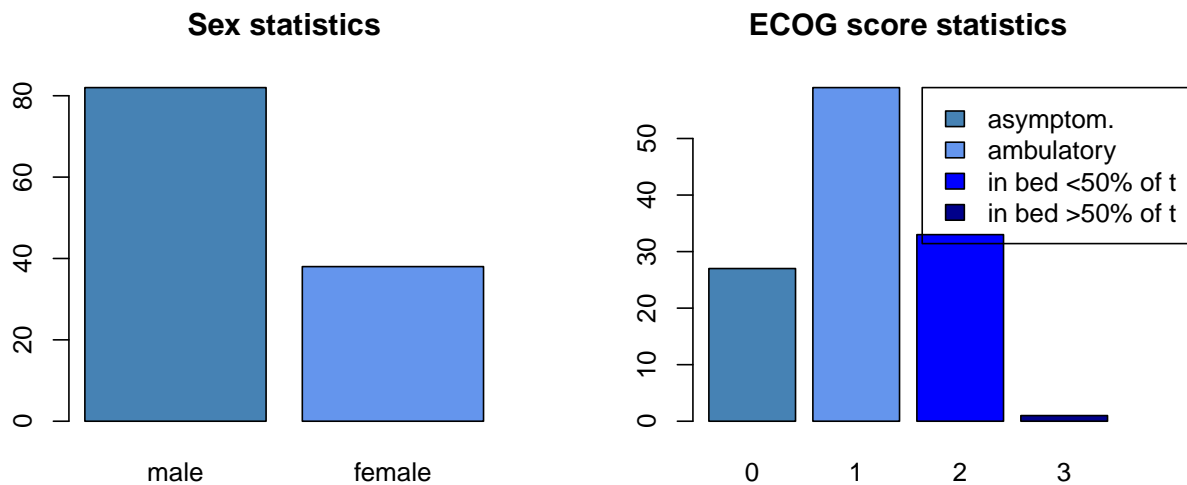


Figure 1: Categorical variables

The distribution for continuous variables is shown in fig. 2. The features do not follow normal distribution.

```
par(mfrow=c(3,2))
hist(data$age, freq=FALSE, col="cornflowerblue", main="Histogram of age", xlab="")
hist(data$ph.karno, freq=FALSE, col="cornflowerblue", main="Histogram of ph.karno", xlab="")
hist(data$pat.karno, freq=FALSE, col="cornflowerblue", main="Histogram of pat.karno", xlab="")
hist(data$meal.cal, freq=FALSE, col="cornflowerblue", main="Histogram of meal.cal", xlab="")
hist(data$wt.loss, freq=FALSE, col="cornflowerblue", main="Histogram of wt.loss", xlab="")
```

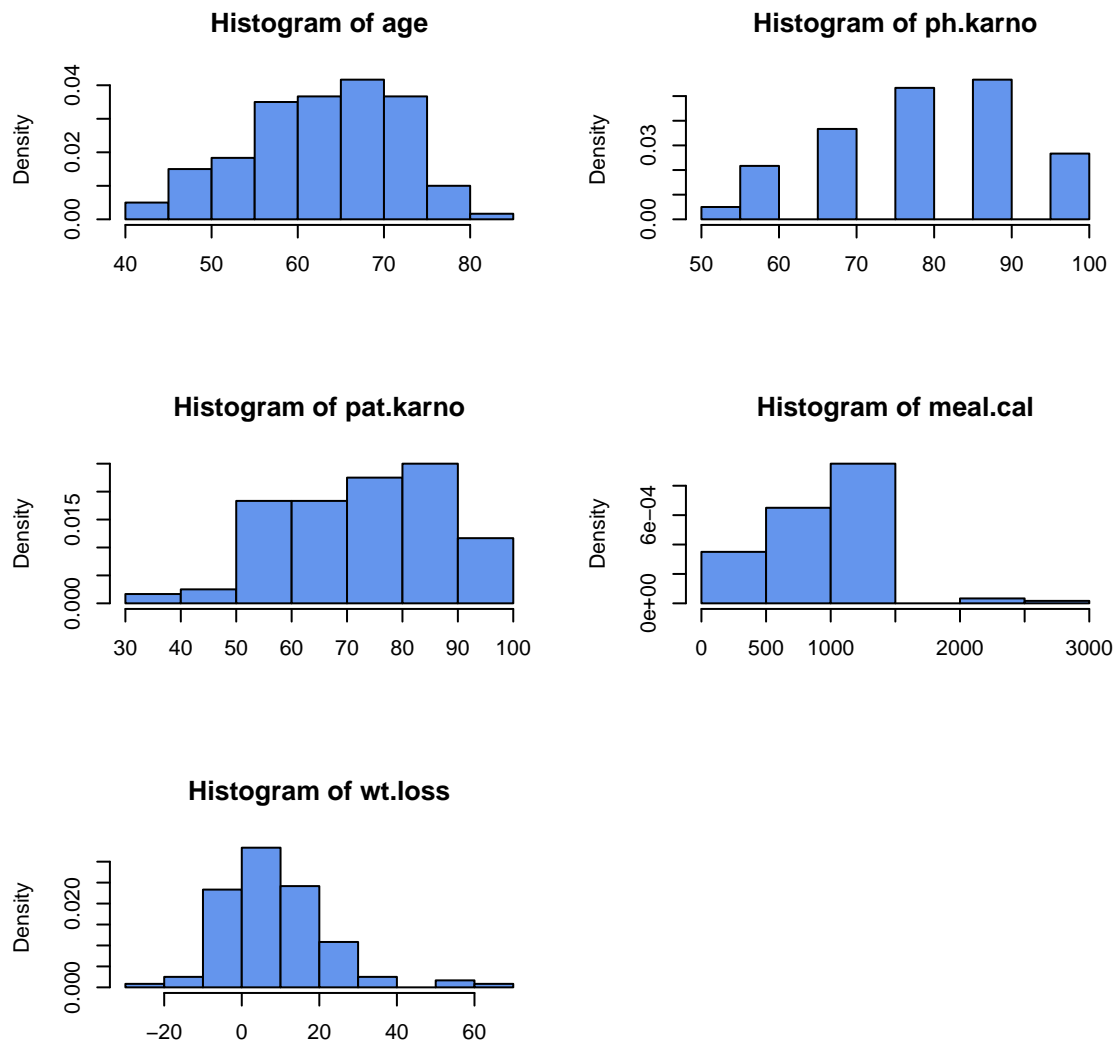


Figure 2: Continuous variables

It is also useful to identify if there is linear correlation between variables. The correlation is not that high, the only one is between ph.karno and pat.karno (0.525), but it is reasonable, as, eventually, patient and doctor measure the same quantity.

```
cor(data[c(4,7,8,9,10)], method=c("pearson"))
```

```
##           age    ph.karno    pat.karno    meal.cal    wt.loss
## age      1.0000000 -0.25559603 -0.15400053 -0.23775148 -0.04536374
## ph.karno -0.25559603  1.00000000  0.52548155  0.08428390 -0.06216308
## pat.karno -0.15400053  0.52548155  1.00000000  0.16526427 -0.08670497
## meal.cal  -0.23775148  0.08428390  0.16526427  1.00000000 -0.07973735
## wt.loss   -0.04536374 -0.06216308 -0.08670497 -0.07973735  1.00000000
```

Relative studies

The original data was presented in the work [1]. In [2] the comparison of semi-parametric and non-parametric models for survival analysis was presented, but different dataset was used, also there was no deep description of Weibull model implementation, i.e. more attention was payed to Cox regression. Study [3] provided the description of exponential models, applied to lung cancer data analysis with Stan code. The Weibull distribution was just mentioned their as a possibility, but no formulas and conclusions were derived for Weibull model.

References

1. Loprinzi CL. Laurie JA. Wieand HS. Krook JE. Novotny PJ. Kugler JW. Bartel J. Law M. Bateman M. Klatt NE. et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*. 12(3):601-7, 1994.
2. Kumar M, Sonker PK, Saroj A, Jain A, Bhattacharjee A, Saroj RK. Parametric survival analysis using R: Illustration with lung cancer data. *Cancer Rep (Hoboken)*. 2020;3(4):e1210. doi:10.1002/cnr2.1210
3. Mohammed H AbuJarad and Athar Ali Khan. 2018, Exponential Model: A Bayesian Study With Stan. *Int J Recent Sci Res*. 9(8), pp. 28495-28506. DOI: <http://dx.doi.org/10.24327/ijrsr.2018.0908.2470>