Lung Cancer Survival Prediction with Bayesian Generalised Linear Models

Yann McLatchie, Arina Odnoblyudova

Contents

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 development. 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 McLatchie and Odnoblyudova [2021].

Data description

General description

The data used in the study shows survival of patients with advanced lung cancer from the North Central Cancer Treatment Group. It is provided in the `survival` R package, Therneau [2021].

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)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

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

There were 67 observations with missed values or censored, 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$sex), main="Sex statistics", names.arg=c("male", "female"),col=c("steelblue","cornflower
barplot(table(data$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"))
```

The distribution for continuous variables is shown in fig. 2.

```
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="")
```

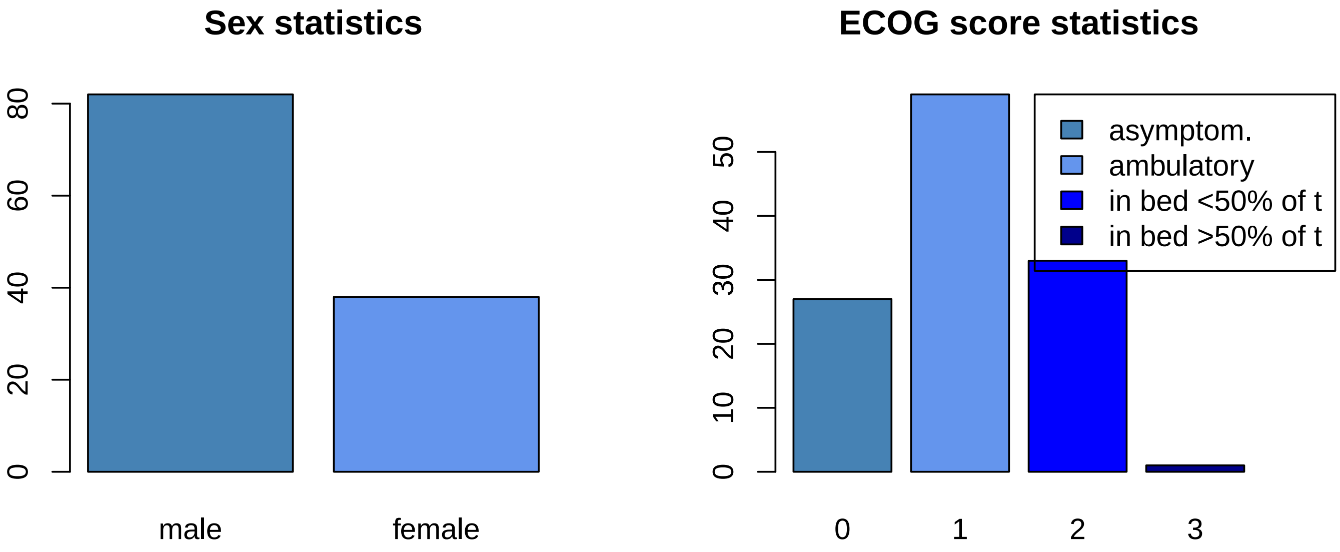


Figure 1: Sex variable

```
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="")
```

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, 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        1.00000000 -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
```

Weibull Generalised Linear Model

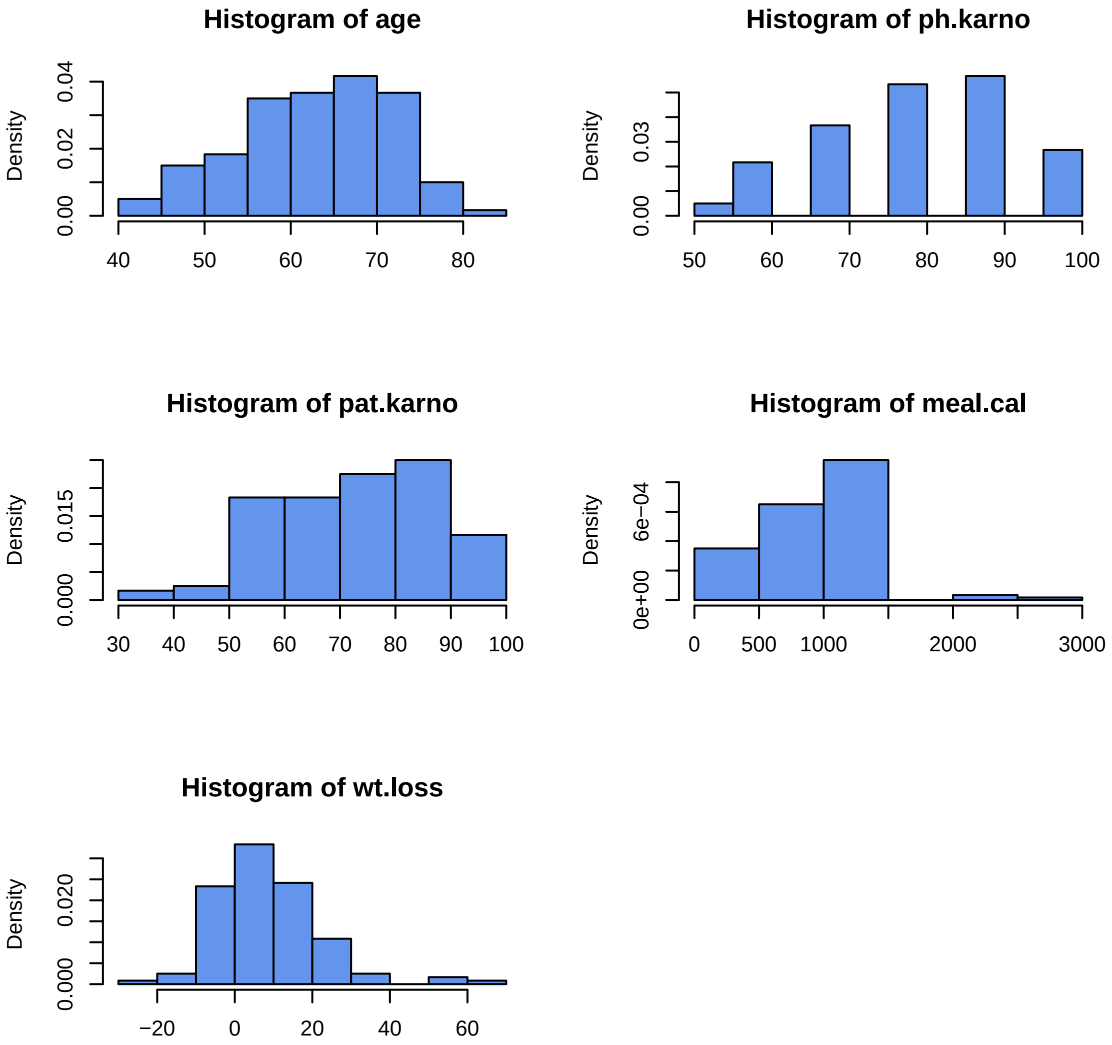
Let $y \sim \text{Weibull}(\alpha, \sigma)$, so that

$$\text{Weibull}(y|\alpha,\sigma) = \frac{\alpha}{\sigma}\left(\frac{y}{\sigma}\right)^{\alpha-1}\exp\left(-\left(\frac{y}{\sigma}\right)^{\alpha}\right),$$

for $y \in [0,\infty), \alpha \in \mathbb{R}^+$, and $\sigma \in \mathbb{R}^+$.

Motivating the distribution

The Weibull distribution is often used as a more flexible and complex alternative to the semi-parametric proportional hazard Cox model for modelling time to failure events, since the hazard rate is not taken to be constant with time.



Histogram of age

Histogram of ph.karno

Histogram of pat.karno

Histogram of meal.cal

Histogram of wt.loss

Figure 2: Continuous variables

The Weibull distribution as a member of the exponential family

Now take α fixed and finite, then it can be shown that this distribution belongs to the exponential family since we can write it's probability density function

$$\text{Weibull}(y|\sigma) = \alpha y^{\alpha-1}\exp(-y^{\alpha}\sigma^{-\alpha} - \alpha\log\sigma),$$

with

$$\begin{aligned} b(y) &= \alpha y^{\alpha-1} \\ \eta &= \sigma^{-\alpha} \\ T(y) &= -y^{\alpha} \\ a(\eta) &= \alpha\log\sigma. \end{aligned}$$

Defining the link function

Looking at our sufficient statistic $\eta = \sigma^{-\alpha}$, it can be shown that

$$\sigma = \exp\left(\frac{\log\eta}{-\alpha}\right)$$

where we construct $\eta = \exp(\boldsymbol{X}\beta)$ so that η is strictly positive. Thus we choose a log link function for our GLM such that

$$\sigma = \exp(-\frac{\boldsymbol{X}\beta}{\alpha}).$$

References

Yann McLatchie and Arina Odnoblyudova. Bda project. https://github.com/yannmclatchie/bda-project, 2021.

Terry M Therneau. *A Package for Survival Analysis in R*, 2021. URL https://CRAN.R-project.org/package=survival. R package version 3.2-13.