Survival Analysis for Clinicians

Dimitris Rizopoulos
Erasmus University Medical Center

Survival - Practical 1: Standard Survival Analysis

The purpose of this practical is to illustrate how standard statistical analysis of survival data can be performed in R.

You are strongly encouraged to do this practical interactively by opening the file:

Survival_Practical1.Rmd

(found in CANVAS under Files → Weeks 2 & 3 → Computer_Practicals) in Rstudio and clicking on the button `Run Document`, or online by pointing your web browser to the following url:

https://emcbiostatistics.shinyapps.io/Survival_Practical1

The following questions are based on the AIDS dataset. This dataset is available as object `aids` from package **JM**. If you decide to directly work in Rstudio instead of the online tutorial, before continuing you will need to load packages **survival** and **JM** using the commands `library("survival")` and `library("JM")`, respectively[1].

From this dataset we will use the following variables:

* `Time`: the observed time-to-death in months.
* `death`: the event indicator; '1' denotes death and '0' censored observation.
* `drug`: the treatment indicator with values 'ddC' and 'ddI'.
* `gender`: the sex indicator with values 'male' and 'female'.

Perform the following analysis:

Q1 Calculate and plot the Kaplan-Meier estimator of the survival function based on all the data. Which is the median survival time and its 95% confidence interval? (hint: Section 2.1, Survival Analysis in R Companion)

[1] If you have not already installed package **JM** in your machine, you will need to do so using the command `install.packages("JM")`.

Q2 Calculate and plot the Breslow estimators of the survival functions for ddC and ddI, separately. Calculate also the estimates of the 50%, 60% and 70% percentiles of the survival distribution with their 95% confidence intervals. (hint: Section 2.1, Survival Analysis in R Companion)

Q3 Calculate the 8- and 10-month survival probability with its corresponding 95% confidence interval. You will need to use the `summary()` function for `survfit` objects. (hint: Section 2.1, Survival Analysis in R Companion)

Q4 Compare with the log-rank Peto & Peto modified Gehan-Wilcoxon tests if the survival curves for the two treatment groups differ statistically significantly. Before doing the analysis, which of the two tests you expect to yield the smaller p-value and why? (hint: Section 2.2, Survival Analysis in R Companion)

Q5 Do the same for gender, i.e., calculate the Kaplan-Meier (or Breslow) estimators of the survival functions for males and females, and compare the results from the log-rank Peto & Peto modified Gehan-Wilcoxon tests. Which test you should trust more in this case and why? (hint: Section 2.2, Survival Analysis in R Companion)

Survival - Practical 2: AFT Models for Time-to-Event Data

The purpose of this practical is to illustrate how Accelerated Failure Time model can be fitted in R.

You are strongly encouraged to do this practical interactively by opening the file:

Survival_Practical2.Rmd

(found in CANVAS under Files → Weeks 2 & 3 → Computer_Practicals) in Rstudio and clicking on the button `Run Document`, or online by pointing your web browser to the following url:

https://emcbiostatistics.shinyapps.io/Survival_Practical2

The following questions are based on the Lung data set. This data set is available as object `lung` from package **survival**. If you decide to directly work in Rstudio instead of the online tutorial, before continuing you will need to load package **survival** using the command `library("survival")`.

From this data set we will use the following variables:

* `time`: the observed time-to-death in days.
* `status`: the event indicator; '1' denotes censored and '2' denotes death.
* `age`: age in years.
* `ph.karno`: Karnofsky performance score rated by the physician.
* `sex`: the sex indicator with values 'male' and 'female'.

Perform the following analysis:

Q1 Our initial hypothesis is that the time-to-death is affected by `sex`, `age` and `ph.karno`. In addition, we also believe that the effects of `age` and `ph.karno` are not the same for males and females. Transform this initial hypothesis into a suitable AFT model. For the error terms assume the extreme value distribution, which as we have seen corresponds to the Weibull distribution for the time-to-death. (hint: Section 3.1, Survival Analysis in R Companion)

Q2 We would like to test whether some aspects of our initial hypothesis are supported by the data. In particular, we are interested in testing: (a) whether `sex` has at all an effect in the time-to-death, and (b) whether the effects of `age` and `ph.karno` are equal for the males and females. Based on the results of these two hypotheses, simplify the model appropriately. (hint: Section 3.3, Survival Analysis in R Companion)

Q3 For the final model obtained in Q2 create an effects plot depicting how the average failure time changes with increasing values of `ph.karno`, for males and females at median age of their respective groups, i.e., for the median age for males and the median age for females. (hint: Section 3.2, Survival Analysis in R Companion)

Q4 Check whether the assumption of the extreme value distribution for the error terms is violated using the AFT residuals. What is your conclusion? (hint: Section 3.4, Survival Analysis in R Companion)

Survival - Practical 3: Cox PH Models for Time-to-Event Data

The purpose of this practical is to illustrate how the Cox proportional hazards model can be fitted in R.

You are strongly encouraged to do this practical interactively by opening the file:

Survival_Practical3.Rmd

(found in CANVAS under Files → Weeks 2 & 3 → Computer_Practicals) in Rstudio and clicking on the button `Run Document`, or online by pointing your web browser to the following url:

https://emcbiostatistics.shinyapps.io/Survival_Practical3

The following questions are based on the AIDS data set. This data set is available as object `aids` from package **JM**. If you decide to directly work in Rstudio instead of the online tutorial, before continuing you will need to load packages **survival** and **JM** using the commands `library("survival")` and `library("JM")`, respectively[2].

From this data set we will use the following variables:

* `Time`: the observed time-to-death in months.
* `death`: the event indicator; '1' denotes death and '0' censored observation.
* `CD4`: baseline CD4 cell count measurement.
* `drug`: the treatment indicator with values 'ddC' and 'ddI'.
* `AZT`: indicator denoting whether the patient was enrolled because of AZT 'intolerance' or AZT 'failure'.

Perform the following analysis:

Q1 Fit a Cox model that relaxes the linearity assumption for the effect of `CD4` using natural cubic splines with 3 degrees of freedom. In addition, include the main effects of `drug` and `AZT`, and the interaction effects of `CD4` with both `drug` and `AZT`. (hint: Section 4.1, Survival Analysis in R Companion)

Q2 Use a likelihood ratio test to test whether the model can be reduced by dropping all interaction terms. Depending on the result choose the model that you will use for the remaining questions unless otherwise stated. (hint: Section 4.3, Survival Analysis in R Companion)

[2] If you have not already installed package **JM** in your machine, you will need to do so using the command `install.packages("JM")`.

Q3 Use the `summary()` method to obtain a detailed summary of the fitted model. What is the interpretation of the estimated coefficient for `drug`? In addition, in the output you have values for `exp(coef)` and `exp(-coef)`. What do these values represent? (hint: Section 4.1, Survival Analysis in R Companion)

Q4 Using the model of Q1, create an effects plot depicting how the average log hazard ratio changes with increasing values of `CD4`, for 'ddI' and 'ddC' patients who had enrolled because of either AZT 'intolerance' or AZT 'failure'. What do you observe? (hint: Section 4.2, Survival Analysis in R Companion)

Q5 Using the Kaplan-Meier estimator to compare whether the proportional hazards assumption is justified for `AZT`. (hint: Section 4.4, Survival Analysis in R Companion)

Survival - Practical 4: Extensions of the Cox Model

The purpose of this practical is to illustrate how to a representative Cox PH regression analysis including the extensions seen in the last sections of Chapter 4 and in Chapter 5.

You are strongly encouraged to do this practical interactively by opening the file:

Survival_Practical4.Rmd

(found in CANVAS under Files → Weeks 2 & 3 → Computer_Practicals) in Rstudio and clicking on the button `Run Document`, or online by pointing your web browser to the following url:

https://emcbiostatistics.shinyapps.io/Survival_Practical4

The following questions are based on the Lung data set. This data set is available as object `lung` from package **survival**. If you decide to directly work in Rstudio instead of the online tutorial, before continuing you will need to load package **survival** using the command `library("survival")`. In addition, you will also need package **splines** that can be similarly loaded with the command `library("splines")`.

From this data set we will use the following variables:

* `time`: the observed time-to-death in days.
* `status`: the event indicator; '1' denotes censored and '2' denotes death.
* `age`: age in years.
* `ph.karno`: Karnofsky performance score rated by the physician.
* `sex`: the sex indicator with values 'male' and 'female'.
* `ph.ecog`: ECOG performance score (0=good 5=dead).

Perform the following analysis:

Q1 Our initial hypothesis is that the time-to-death is affected by `sex`, `age` and `ph.karno`. Also, the physicians believe that the effect of `ph.karno` and `age` may be nonlinear in the log-hazard scale. Moreover, the (possibly nonlinear – model using natural cubic splines with 3 degrees of freedom) effects of `age` and `ph.karno` on the log-hazard scale are not the same for males and females. Transform this initial hypothesis into a suitable Cox PH model. (hint: Section 4.1, Survival Analysis in R Companion)

The aim here is to do a realistic analysis of a survival dataset with a Cox PH model. This involves the following steps:

a. We first translate our initial hypothesis into a full model that contains all terms of interest. This includes all covariates we are interested in and also possibly nonlinear and interaction terms.

b. We then first test the important assumption behind the model. In the case of the Cox model that is the proportional hazards assumption. (In the case of an AFT model that is the distribution of the error terms). We need to do that first and rectify any problems with these assumptions **before** proceeding to simplify the model using hypothesis testing.

c. Then we continue by performing an omnibus test for all interaction terms in the model and see if we can drop them. Typically, using a p-value threshold higher than 5%, e.g., we can use 15%. This is to ensure that we do not miss any potentially interesting interactions. If the test suggests that some interactions may seem to improve the fit of the model, then we can proceed to see which interaction terms specifically achieve that. We test then each interaction separately, and at the end we can correct the p-values for multiple testing. Hence, at the final stage of this step we will know which interaction terms we will keep in the model.

d. We do the same for the nonlinear terms. Namely, first, we start by the omnibus test, and if the p-value is smaller than 15%, we are going to see which nonlinear terms we need. Hence, at the final stage of this step, we will know our final model. Note, that unless the aim is to do prediction, it is not advisable to remove non statistically significant covariates from the final model.

e. Finally, we interpret the results using the table of coefficients and effect plots if necessary.

Q2 We are interested in estimating survival probabilities for males and females with the median age and with the average Karnofsky score. (hint: Section 5.1, Survival Analysis in R Companion)

- Which are the median survival times and their 95% confidence limits for males and females with median age and average Karnofsky score?
- Plot the corresponding survival curves.
- What are the corresponding survival probabilities for 200, 400, 600 and 800 days?

Q3 For the rest of the questions we consider the additive Cox PH model with `sex`, `age` and `ph.karno` fitted in the original `lung` database (i.e., not the two databases before and after 170 days). It is believed that the baseline hazard of death has a completely different shape for patients with ECOG score greater than 0 compared to patients with ECOG equal to 0, i.e., the hazard functions of the two groups is not analogous. First, from the `ph.ecog` variable that takes values from 0 to 3, and construct the variable `ph.ecog2` that is 0 if `ph.ecog` was 0, and 1 otherwise. Then, fit an appropriate Cox model that takes the feature described above into account, and then interpret the results. (hint: Section 5.2, Survival Analysis in R Companion)

Q4 The team of physicians of the North Central Cancer Treatment Group (who are responsible for the Lung study) believe that the effects of `sex`, `age` and `ph.karno` in the risk of

death are different for the two ECOG groups. Extend the model of Q3 accordingly and test whether this hypothesis is supported by the data for each of the two predictors. (hint: Section 5.2, Survival Analysis in R Companion)