Biostatistics Methods II – Software Practicals
Likelihood and Survival Analysis

Dimitris Rizopoulos
Erasmus University Medical Center

A Very Basic Introduction to R

The aim of this short section is to provide a very fast introduction in the R statistical programming language that will be used during the practical sessions of this course.

* Data Structures
  - we have two basic types of variables, namely variables of type `numeric`, which are usually continuous (e.g., weight, age, etc.), and variables of type `factor`, which are categorical (e.g., sex, treatment, etc.).
  - data sets in R are usually stored in `data.frames`. Typically, each row of the data frame denotes a subject and the columns denote the different variables recorded on that patient. These variables can be continuous (i.e., `numeric`) or categorical (i.e., `factor`).

* Manipulating Data
  - in the following we will work with the data frame `aids.id`. To access this data frame you need to load package **JM** using the command `library(JM)`.
  - type `aids.id` to see how this data set looks like. For instance, you can see that variable `Time` is a continuous variable and is of type `numeric` whereas variable `gender` is a `factor`.
  - to extract a variable from the data frame we use the symbol `$`. For instance, the following code extracts the variable `sex` from the data frame `aids.id`:

    ```
    R> aids.id$sex
    ```

    With a similar code you can also transform a variable. For instance, the following code defines a new variable in the data frame which equals the natural logarithm of the `Time` variable:

    ```
    R> aids.id$logTime <- log(aids.id$Time)
    ```

To convert a `numeric` variable into a `factor` we use the `factor()` function. For example, the following code, transforms variable `death` into a `factor`:

```
R> aids.id$death <- factor(aids.id$death, levels = 0:1,
        labels = c("alive", "dead"))
```

* Using `formula` to Define Regression Models

  – the way to define in R the relationship between a response variable and a set of predictors is via a `formula`. In the following examples, we will denote the response variable by `y`, a continuous predictor by `x` and a categorical predictor, i.e., a factor by `f`

  – A model that postulates that the average `y` is related to the main effect of `x` and the main effect of `f`:

  ```
  R> y ~ x + f
  ```

  – A model that postulates that the average `y` is related to the main effects of `x` and `f`, and the interaction effect between `x` and `f`:

  ```
  R> y ~ x + f + x:f
  R> # equivalently the above can be shortened to
  R> y ~ x * f
  ```

  – A model that postulates that the average `y` is related to the linear and quadratic effects of `x`:

  ```
  R> y ~ x + I(x^2)
  ```

Likelihood - Practical 1: Linear Regression Models in R

The purpose of this practical is to illustrate how standard linear regression models can be fitted in R.

The following questions are based on the PBC dataset. This dataset is available as object `pbc2.id`. If you decide to directly work in Rstudio instead of the online tutorial, before continuing you will need to load package **JM** using the command `library("JM")`[1].

From this dataset we will use the following variables:

* `serBilir`: baseline serum bilirubin in mg/dl.
* `drug`: the treatment indicator with values 'placebo' and 'D-penicil'.
* `age`: baseline age in years.
* `sex`: the sex indicator with values 'male' and 'female'.
* `serChol`: baseline serum cholesterol in mg/dl.
* `prothrombin`: baseline prothrombin time in sec.

Perform the following analysis:

Q1 We want to see how the log-transformed serum bilirubin time is related with the rest of the variables. Start by fitting an additive linear regression model, using function `lm()`, and interpret the results you obtain from the `summary()` method.

Q2 Check the residuals of this model using the 'plot()' method; before calling `plot()`, set `par(mfrow = c(2, 2))` in order to obtain the first basic plots in one figure.

Q3 We believe that the association between `serBilir` and each one of `age`, `serChol` and `prothrombin` could be different between males and females. Extend the previous model to accommodate this. Use again the `summary()` method to get a detailed output and interpret the results.

Q4 We would like to statistically test whether the extra interaction terms really improve the fit of the model. Using an F-test (the analogues likelihood ratio test for linear regression models), compare the interaction model with the additive model. In R this is done using the `anova()` function.

Q5 Given that the F-test suggests that some of the interaction terms are significant, we could proceed to look whether we need all of them or some of them. To find these, we could directly look at the individual p-values from the `summary()` output of the last model. However, one issue is that these p-values are not corrected for multiple testing. Using the `p.adjust()` function obtain the adjusted p-values.

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

Q6 In a second analysis, the researchers are interested in studying the relationship between the natural logarithm of serum bilirubin and serum cholesterol corrected for age and sex. It is believed that the relationship may be nonlinear. Use a 3rd degree polynomial of serum cholesterol to explore this.

Q7 Investigate whether the relationship is truly nonlinear but first fitting the model that assumes linearity (null hypothesis), and following comparing this model with the previous model (alternative hypothesis) using an F-test and the `anova()` function.

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.

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[2].

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 2.2, R Code Appendix)

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 2.2, R Code Appendix)

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.

Q4 Compare with the log-rank Peto & Peto modified Gehan-Wilcoxon tests if the survival curves for the two treatment groups differ statistically significantly. <u>Before</u> doing the analysis, which of the two tests you expect to yield the smaller p-value and why? (hint: Section 2.3, R Code Appendix)

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?

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

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.

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, R Code Appendix)

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, R Code Appendix)

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, R Code Appendix)

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, R Code Appendix)

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`.

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[3].

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. In addition, include the main effects of `drug` and `AZT`, and the interaction effects of `CD4` with both `drug` and `AZT`. (hint: Sections 4.1 & 4.8, R Code Appendix)

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, R Code Appendix)

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, R Code Appendix)

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, R Code Appendix)

Q5 Using the Kaplan-Meier estimator to compare whether the proportional hazards assumption is justified for `AZT`. (hint: Section 4.5, R Code Appendix)

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

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.

The following questions are based on the Lung data set. This data set is available as object `lung` from package **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, the physician believe that the effect of `ph.karno` and `age` may be nonlinear in the log-hazard scale. Moreover, the (possibly nonlinear) 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, R Code Appendix; to allow for nonlinear effects using function `ns()` from package **splines** – you will need first to load the package using the command `library("splines")`)

The aim here is to do a realistic analysis of a survival dataset with a Cox PH model. Hence,

a. Start by checking if the interaction effects can be dropped using a likelihood ratio test.

b. Similarly check if the nonlinear effects can be dropped using a likelihood ratio test.

c. For the final model check the PH assumption for the terms in the model.

Q2 We are interested in estimating survival probabilities for males and females at their median respective age and with their average respective Karnofsky score. (hint: Section 5.1, R Code Appendix / `R` commander: (a) to compute numerical summaries: Statistics → Summaries → Numerical summaries. . . ; (b) to estimate the survival function: Models → Graphs → Cox-model survival function. . .)

– which are the median survival times and their 95% confidence limits for males and females with median age and average Karnofsky score? (with commands)

– plot the corresponding survival curves. (with the `R` commander)

  – what are the corresponding survival probabilities for 200, 400, 600 and 800 days? (with commands)

Q3 It is believed that the baseline hazard for death has a completely different shape for males and females, i.e., the hazard function of males is not analogous to the hazard function females[4]. Fit an appropriate Cox model that takes this feature into account. (hint: Section 5.2, R Code Appendix / R commander: Statistics → Fit models → Cox regression model...)

Q4 The team of physicians of the North Central Cancer Treatment Group (who are responsible for the Lung study) believe that the effects of `age` and `ph.karno` in the risk for death are different for males and females. 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, R Code Appendix / R commander: Statistics → Fit models → Cox regression model...)

Q5 Fit two separate Cox models for males and females adjusting in each one for `age` and `ph.karno`, and compare the results with Q4. What do you observe? (hint: Section 5.2, R Code Appendix / R commander: Statistics → Fit models → Cox regression model...)

[4] or to put it also otherwise, the hazard function of males is not equal to the hazard function of females times a constant.