TMA4275 Life time analysis
Obligatory project 3, Spring 2024

Out: Thursday April 4
In: Thursday April 18 at (latest) 21.00

Important information: Parts of this project are to be done using R. An introduction to R can be found in the course web page (see Statistical software). The project report should consist of one (and only one) pdf-file, and should be uploaded via Blackboard. The project report should include derivation of formulas that you are using in your implementations. The project report should also include the R code you have used to solve the project and the plots you have generated. Associated to the various plots there should be captions explaining the contents of the plots, and in addition all the plots should be explained and discussed in the main text of the report.

The project report should be formulated as a scientific report. In particular, it should be possible to understand what you have done without reading the questions in this problem text. Moreover, the text in the project report should consists of full sentences and proper punctuation should by used throughout, also in equations! All results you present should be discussed. What can you (and the world) learn from your results? The project text should be written so that it is easy to follow by your fellow students in TMA4275 Lifetime analysis.

The report can be written in English or Norwegian. You should do the project alone! In your solution, specify your (full) name, NOT student or candidate numbers.

Your solution should be handed in in Blackboard. After having logged in to Blackboard click on "course information" to find where to hand in your solution.

In this problem we will study data used in Kristov et al (2002). The article can be downloaded from

`https://www.sciencedirect.com/science/article/pii/S0951832002001692`

(if you are at the NTNU network) or from

`https://www.math.ntnu.no/emner/TMA4275/2024v/Papers/CoxReliabilityPaper.pdf`

(does not require you to be at the NTNU network). The dataset discussed in the article can be downloaded from

`https://www.math.ntnu.no/emner/TMA4275/2024v/Datasets/tire.txt`.

Before starting to work with the problems specified below you should read the article so that you get familiar with the background for the data set, including the set of possible covariates.

a) Use the R function *coxph* to redo the two Cox regression analyses reported in Tables 2 and 3 in Kritsov et al (2002). For each of the two analyses, discuss the importance of the various covariates. *Note: You may not get exactly the same results as in Kritsov et al. (2002). The difference might be from different treatment of ties, or other sources like rounding errors.*

b) For each of the two models considered in **a**) and using the Breslow estimator, write your own

R code and estimate the integrated hazard rate for a component with all covariate values equal to unity, $A_0(t, x_0)$. Make a plot of the two estimated integrated hazard rates (in the same plot so that they are easy to compare).
Find also the corresponding estimated survival functions and visualise them in a plot.
Discuss what you see in the two plots.

c) Redo the two analysis in **a**) using a Weibull regression model. For this you can use the *survreg* function in R. Note that the interpretation of the estimated parameter values computed by *survreg* is different from the interpretation of the estimated parameter values in *coxph*. [*Section 10.3.7 in Moore (2016) describes in an example how one from the* survreg *output can compute parameter values corresponding to the ones estimated in* coxph.] For each of the two Weibull regression analysis results, compute estimated regression coefficients that have the same interpretation as for the Cox regression models and compare the estimated values for the Cox and Weibull regression models.

d) For each of the two Weibull regression models estimated in **c**), form and plot the resulting estimated survival functions in the time interval $[0, 1.3]$ for a component with all covariate values equal to unity. Plot both curves in the same plot to make it easier to compare, and include in this plot also the corresponding estimated quantities based on the Cox model (which you found in **b**)).
For each of the two Weibull regression models make also plots of the martingale residuals.
Based on your analysis results, what would you say about the model fit for the two fitted Weibull models? Would you expect also an exponential regression model to give a good fit?

e) The data set has seven potential covariates, which gives $2^7 = 128$ possible regression models. Fit a Weibull regression model for each of these models and, using a reasonable criterion of your choice, find which of the 128 models that gives the best fit to the data set. Specify and briefly discuss your choice of criterion.

Problem 2:

In this problem you will perform a simulation study related to your analysis in Problem 1. Based on the dataset studied in Problem 1 you will simulate multiple new data sets and fit a model to each of these. The number of cases and the values of the covariates will in the simulated data sets be identical to what they are in the real dataset, it is just the (possibly censored) survival times and the censoring indicators that will be given new values.

a) Make an R function that takes as input the dataset used in Problem 1 and the best model found in Problem 1 **e**). The output of the function should be a new (simulated) dataset. For each case in the dataset, new survival and censoring times should be simulated. The survival times should be simulated according to the fitted model from Problem 1 **e**). Note that the survival times are then all simulated from a Weibull distribution, but with different parameter values because the different cases have different covariate values. To simulate from the Weibull distribution you can use the built-in R function for this or you can make your own R function. If you use the built-in R function please check carefully the parameterisation used by this R function, since this parameterisation does not need to be the same as the parameterisation used

by *survreg*. The censoring times you should simulate from an exponential distribution. Choose a parameter value for the exponential distribution so that the mean fraction of censored survival times in the simulated data sets are approximately equal to the number of censored survival times in the original data set.

b) If you have not already done it in Problem 1 **e**), make an R function that fits all possible regression models to a specific data set and finds the best model according to your criterion chosen in Problem 1 **e**). Note that you should try all regression models based on all seven covariates in the original data set, not only the covariates you used to generate your data set.

Generate a simulated data set by your R function from Problem 2 **a**) and use this simulated data set in your R function for finding the best model. For three of the cases in the data set, make plots with the estimated intensity processes (from the simulated data set) and the corresponding "true" intensity processes (from the model used to generate the simulated data set). Make also a plot of the martingale residuals.

Make two more simulated data sets, use your R function to find the best model for each of these, and make the same plots as you did for the first simulated data set. Compare your results here with your results in Problem 1, and comment. Do the simulation results here give you more or less confidence to the estimated results in Problem 1?

c) Repeat the process of simulating data sets and finding the best model many times. You must decide exactly how many times, it should be large but computationally doable within a reasonable computation time. From the results, find for each of the seven covariates the proportion of the fitted models that includes the covariate. Are these numbers reasonable considering what covariates that are in the model used to simulate the data sets?

From the results, make also a plot with estimated survival functions for an individual with all covariate values equal to unity, i.e. one curve for each of the fitted models. In the same plot, include also one (thicker) curve for the survival function of the model used to generate the simulated data sets. What can you learn from this plot? *Note: If you are familiar with bootstrapping, note that what you have done in this problem can be seen as a parametric bootstrapping procedure for estimating the uncertainty of the fitted model in Problem 1* **e**).

References

Kristov, V.V., Tananko, D.E. and Davis, T.P. (2002). Regression approach to tire reliability analysis, Reliability Engineering and System Safety, **78**, 267–273.

Moore, D.F. (2016). Applied Survival Analysis Using R, Springer.