TMA4275 Life time analysis
Obligatory project 2, Spring 2021

Out: Monday March 29
In: Wednesday April 28 at (latest) 18.00

Important information: The 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 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 content of the plots, and in addition all the plots and other results should be explained and discussed in the main text of the project text. You should also note that you should have enough text in your report so that it is possible for a reader to follow what you have done without having the problem text available.

Problem 1:

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/TMA4275/2021v/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/TMA4275/2021v/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 and make a plot of the estimated values for the relative risk function, i.e. $r(\widehat{\beta}, x_i)$, for each of the 34 cases in the data set. *(Note that for this dataset the estimated relative risk function $r(\widehat{\beta}, x_i)$ is just a number and not a function since the covariates do not vary with time.)* Try to design a plot for the values of the estimated relative risk function values so that it is easy to compare the results of the two estimated models. *Note: You will 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) Write your own R code and estimate the integrated baseline hazard rate $A_0(t)$, using the Breslow estimator, for each of the two models considered in **a**). Make a plot of the two estimated integrated hazard rates (in the same plot so that they are easy to compare). *Note: you will need*

the estimated integrated baseline hazard rates also in **d***) below.*

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*. Read for example Section 10.3.7 in Moore (2016) for a description of the difference. For each of the two Weibull regression analysis, compute estimated regression coefficients that have the same interpretation as for the Cox regression models and compare the estimated values for Cox and Weibull regression models.

d) For each of the two Weibull regression models estimated in **c**), make a plot of the estimated integrated hazard rates, and include in the same plot also the estimated integrated hazard rates for the Cox regression models 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) Starting with the model with all potential covariates included, perform a stepwise elimination procedure. So you must choose a reasonable criterion for choosing what (if any) covariate to remove from the model at each step, and you must choose a reasonable criterion for deciding when to stop the elimination process. Specify what criteria you have chosen and report you results in each step of the procedure. *Note: As you should report the fitted models during the elimination process you here need to run the elimination process "by hand". In Problem 2***b***) you need to run a similar elimination process multiple times for different data values, but with the same set of possible covariates as used here. In Problem 2 you therefore need to make an R function that runs the elimination process for you and returns the final estimated model. You should code this function yourself, not use an R function that already exists. You may include calls to* survreg *in your function, but the code that decides what covariate to eliminate and when to stop you should code yourself. You may of course use your "hand run" solution here to verify that your function is working properly.*

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 fitted model found by the backward elimination process 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 the parameter value of the exponential distribution so that the mean number 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) Make an R function for running the backward elimination process for a given dataset, as discussed in Problem 1**e**). Use the same criteria for choosing what covariate to eliminate and when to stop the elimination process as you used in Problem 1**e**). Note that you should start the elimination process with a model with all seven covariates included, i.e. also the covariates that are not included in the fitted model used to generate your simulated data.

Generate a simulated data set by your R function from Problem 2**a**) and use this simulated data set in your R function for the backward elimination process. 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, fit models by backward elimination 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 fitting a model to the simulated data set may times. You must decide exactly how many times, it should be 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 all the estimated baseline hazard rates, i.e. one curve for each of the fitted models. In the same plot, include also one (thicker) curve for the baseline hazard rate used to generate the simulated data sets. What can you learn from this plot?*Note: For those who are taking TMA4300, note that what we have done in this problem is 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.