Obligatory excercise 2, TMA4275 Lifetime analysis, Spring 2017

March 21, 2017

You may use either Minitab or R to solve the exercise (instructions for doing the analysis in R is given below). For MINITAB, use the help system in MINITAB and the slides on the course web page. Write your **candidate** (or student number if you have not yet been assigned a candidate number), not names, on the report. We recommend using latex + knitr (menu option 'File/New file/R Sweave' in Rstudio) for writing the report (see for example, `https://support.rstudio.com/hc/en-us/articles/200552056-Using-Sweave-and-knitr`). You may work on your own or write the report jointly with another student.

First do

```
data <- read.table("https://www.math.ntnu.no/emner/TMA4275/2017v/data.dat")
```

to load the data into R and then do `attach(data)`. This dataset contains right censored observations `y` with right censoring indicator `delta` and two potential covariates `x1` (continuous) and `x2` (binary).

1. First fit a Cox proportional hazards model without this any covariates included and plot martingale residuals from this model against each covariate and possible transformations using methods in Moore Ch. 7 including the function `smoothSEcurve` included in the Appendix.

2. Next fit the Cox proportional hazard model using only the binary variable `x2` as covariate in the model. Assess the proportional hazard assumption by plotting the Schoenfeld residuals against the observed failure times. See `residuals.coxph` and the `type="schoenfeld"` argument. The ordered failure times can be computed by `sort(y[delta==1])`. See Moore 2017, cp. 7. Again, estimate the mean using `smoothSEcurve`.

   In addition, make a log-minus-log plot of the Kaplan-Meier estimates of the survival function $R(t)$ against time aswell as the log of time for each value of the binary variable `x2` to assess the proportional hazard assumption. Lastly, discuss if the curves in the second plot appear consistent with Weibull distributions? (See slides 12 and Moore ch. 7.2.1.)

3. Then fit a Cox proportional hazards model with your choice of transformed variables both included as covariates in the model. Examine how the p-values based on the likelihood ratio test of the hypothesis $H_0 : \beta_k = 0$ vs. $H_0 : \beta_k \neq 0$ compares to the Wald-test.

4. Reassess the proportional hazards assumptions for the extended model by plotting Schoenfeld residuals against the observed failure times. Also carry out the formal hypothesis test of the proportional hazard assumption available via the `cox.zph` function. (See Moore 2017, cp. 7 for details).

5. Estimate and plot the baseline survival function $R_0(t)$ of the fitted Cox model using the `survfit.coxph` function with default options. Also compute and plot corresponding estimates of the survival functions for subjects within each group defined by `x2` and for values of `x1` equal to 0.2, 1 and 5. See the `newdata` argument of `survfit.coxph`.

6. What form would the baseline survival function take if the lifetimes conditional on the covariate values follow a Weibull distribution? Find a transformation of the $\hat{R}_0(t)$ and t which would make the relationship linear under the Weibull model and plot this to examine if the estimated baseline hazard model agrees with the Weibull model.

7. Estimate and compute the associated standard errors of β_1 and β_2 in the above Cox proportional hazard model from parameter estimates of a Weibull survival regression (`survreg`). Note that the covariance matrix that can be obtained using `vcov( )` uses the parameterization $\beta_0, \beta_1, \beta_2, \ln \sigma$ where σ is the scale parameter in the log-location-scale representation of the Weibull distribution. How do the standard errors compare to those from the semi-parametric Cox model? How does this agree with your intuition?

8. If the distribution of the lifetimes are Weibull given both covariates, why are they not Weibull conditional on only `x2` as seen in point 2? Write out a formula in integral form for pdf of the resulting mixture distribution conditional on x_2 assuming that `x1` is standard lognormal.

9. Suppose you need to analyse some data where you know that the data generating process generates a Weibull distribution but an important covariate is missing. Based on your findings in this exercise, do you think a Cox proportional hazard model, a parametric Weibull survival regression model, or perhaps some other model would be appropriate for analysis of the data?