Exercise 1:

Show that, under the Cox model, the survival function is

$$S(t, \boldsymbol{x}_i) = [S_0(t)]^{\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})} \quad ,$$

where $S_0(t)$ denotes the baseline survival function.

Exercise 2:

The data frame `tongue`, which is given in the **R** package `KMsurv`, contains death times (in weeks) of patients with cancer of the tongue. The variable `type` gives information whether the tumour had an aneuploid (abnormal) or diploid (normal) DNA profile. Make use of `help(tongue)` to become acquainted with the data set.

(a) Fit a Cox proportional hazards regression model to this data set to estimate the impact of the variable tumor type (`type`) on the survival time. Use the function `coxph()` from the **R** package `survival`.

(b) Compare your results obtained in (a) with the results you obtain when fitting a parametric Weibull-AFT model to these data.

(c) Use different methods for tie handling (see the option `ties` in the function `coxph()`), to test whether there is a significant effect of the tumor type.

Exercise 3:

Given data $D_n = \{(t_i, \delta_i, \boldsymbol{x}_i), i = 1 \ldots, n\}$ and a non-informative (random) censoring scheme, the likelihood for the Cox model is given by

$$L(\boldsymbol{\beta}, h_0) = \prod_{i=1}^{n} h(t_i, \boldsymbol{x}_i)^{\delta_i} S(t_i, \boldsymbol{x}_i)$$

with $h(t, \boldsymbol{x}_i) = h_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta})$ and $S(t, \boldsymbol{x}_i) = \exp(-H_0(t) \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}))$, where $H_0(t) = \int_0^t h_0(u)\, \mathrm{d}u$ denotes the cumulative baseline hazard.

(a) Show that the likelihood can be written as

$$L(\boldsymbol{\beta}, h_0) = \prod_{i=1}^{n} \exp(\eta_i)^{\delta_i} \exp(-\exp(\eta_i)) \left(\frac{h_0(t_i)}{H_0(t_i)} \right)^{\delta_i} \quad ,$$

where $\eta_i = \ln(H_0(t_i)) + \boldsymbol{x}_i^\top \boldsymbol{\beta}$.

(b) Assume that the censoring indicator δ_i is Poisson distributed with parameter $\mu_i = \exp(\eta_i)$, that is, $\delta_i \sim \mathcal{P}(\mu_i)$. Specify the likelihood for this log-linear Poisson model.

(c) Making use of the results obtained in (a) and (b) show that the likelihood of the Cox model is proportional to the likelihood of the log-linear Poisson model with offset $\log(H_0(t_i))$.

(d) Consider the special case of a Cox model with constant baseline hazard rate $h_0(t) = h$. Show that, for purposes of conducting statistical inference for this model, one can use the log-linear Poisson model with offset $\log(t_i)$.

(e) Recall the data file `melanoma.dat` that has been analysed previously. Fit a Cox model with constant baseline hazard rate to these data, using the function `glm()`. Compare your results to the results obtained when using the `survreg()` function.

Exercise 4:

The data frame `kidtran`, which is given in the **R** package `KMsurv`, contains time to death of 863 kidney transplant patients. All patients had their transplant performed at The Ohio State University Transplant Center during the period 1982-1992. The maximum follow-up time for this study was 9.47 years. Patients were censored if they moved from Columbus (lost-to follow-up) or if they were alive on June 30, 1992. In the sample, there were 432 white males, 92 black males, 280 white females, and 59 black females. Patient ages at transplant ranged from 9.5 months to 74.5 years with a mean age of 42.8 years. Make use of `help(kidtran)` to become acquainted with the data set.

(a) Estimate a Cox model with the covariates `gender`, `race`, the interaction `gender * race` and `age`. Compute the hazard ratio of (i) a 40 year old black male, (ii) a 40 year old white male, and (iii) a 40 year old black female, compared to a 40 year old white female.

(b) Start with a Cox model containing `age` as the only covariate. Use the Akaike Information Criterion (AIC), which is implemented in the function `AIC()`, to check whether `gender`, `race` and the interaction `gender * race` should also be included in the model.

(c) Use the function `basehaz()` in the **R** package `survival()` to plot the cumulative baseline hazard for the best model (according to the AIC) and the null model (containing no covariates). Compare the results with the results obtained when using the option `centered=FALSE`.

Exercise 5:

(a) Recall the inverse transform sampling method for generating random numbers from a distribution with invertible cumulative hazard function $H(t)$ (see exercise 5 of study sheet 1). Generalize the method developed in exercise 5 (a) to build up a scheme for generating survival times from a Cox model with given and invertible cumulative baseline hazard rate $H_0(t)$.

(b) Simulate $r = 100$ samples with sizes $n = 50$ from a Cox model with hazard rate

$$h(t; x) = t \exp(\beta x)$$

and $\beta = 0.5$. In order to do this, simulate the covariate x from a uniform distribution on the interval [0,1] and use these values in the $r = 100$ samples. Simulate the censoring times for the $r = 100$ samples from a uniform distribution on the interval [0,5].

(c) The bias of the estimator $\hat{\beta}$ is defined as $\text{Bias}(\hat{\beta})=\text{E}(\hat{\beta})-\beta$. For the analysis of simulation studies one considers frequently the empirical bias

$$\widehat{\text{Bias}}\left(\hat{\beta}\right) = \frac{1}{r}\sum_{k=1}^{r}\hat{\beta}^{(k)} - \beta \ ,$$

where $\hat{\beta}^{(k)}$ denotes the estimator arising from the kth simulation run. Compute the empirical bias for β and draw a histogram (together with a kernel density estimate) of the 100 estimated $\hat{\beta}^{(k)}$ $(k = 1 \ldots, r)$.

(d) Repeat the steps (b) and (c) using sample sizes $n = 100$, $n = 250$ and $n = 500$.