Exercise 1:

Consider the accelerated failure time (AFT) model

$$\ln(T) = Y = \boldsymbol{x}^\top \boldsymbol{\beta} + \sigma\epsilon,$$

where $\boldsymbol{x} = (1, x_1, \ldots, x_p)^\top$, $\boldsymbol{\beta} = (\beta_0, \beta_1 \ldots, \beta_p)^\top$ and the error ϵ being distributed according to the standard Gumbel (minimum) distribution with density $f_\epsilon(\epsilon) = \exp(\epsilon - \exp(\epsilon))$ for $\epsilon \in \mathbb{R}$.

(a) Derive the cumulative distribution function (cdf) $F_\epsilon(\epsilon)$ and the survivor function $S_\epsilon(\epsilon)$.

(b) Show that for $\sigma \neq 1$, T has a Weibull distribution.

(c) Show that the Weibull distribution has the property of proportional hazards, that is, for arbitrary vectors of explanatory variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ the quotient $h(t; \boldsymbol{x}_1)/h(t; \boldsymbol{x}_2)$ is independent of t.

(d) For a parametric regression model the baseline hazard rate is defined as $h_0(t) = h(t; 0)$, that is, $h_0(t)$ is the hazard with all explanatory variables being set to zero. For the Weibull transformation model specify the baseline hazard rate and the baseline survival function $S_0(t) = S(t; 0)$.

(e) For an arbitrary vector of predictors $\boldsymbol{x}$, determine the relationship between the survivor function $S(t; \boldsymbol{x})$ (hazard function $h(t; \boldsymbol{x})$) and the baseline survivor function $S_0(t)$ (baseline hazard function $h_0(t)$).

(f) Derive the density $f_Y(y)$, survival function $S_Y(y)$ and hazard rate $h_Y(y)$ for the transformed data $Y = \ln(T)$. In general, describe the relationship between the survival function and the hazard function of the variables Y and ϵ (of the variables T and Y).

(g) In **R**, one can fit parametric regression models using the function `survreg()` in the `survival` package. Fit a Weibull regression model (using an intercept term only) to the `melanoma.dat` data that have been analyzed previously. Give estimates for the parameters α and λ of the Weibull distribution. Plot the estimates for the baseline survivor function together with the Kaplan-Meier estimate of the survivor function.

(h) Fit a Weibull regression model again using also the explanatory variables `sex`, `thick`, `ulcer` and `age` and give estimates for the parameters of the Weibull distribution. Give an interpretation of the influence of the corresponding explanatory variable on the hazard rate. Which of the variables has a significant impact on the survival time ($\alpha = 0.05$)?

(i) Plot the estimated hazard rates for men and women together with the baseline hazard rate. For that purpose, set all covariables equal to the corresponding group mean value. Then, plot the survival curves for men and women together with the baseline survival curve and the Kaplan-Meier estimates.

(j) Fit the Weibull AFT model again using `age` as the only predictor (apart from the intercept). Use a B-spline basis for a polynomial spline of degree 2 (implemented in the **R** function `bs()` in the `splines` package) to model impact of the covariate `age`.

Exercise 2:

(a) A random variable ϵ has a logistic distribution with parameters $\alpha, \lambda > 0$ if it has the cdf

$$F_\epsilon(\epsilon) = \frac{\lambda \exp(\alpha\epsilon)}{1 + \lambda \exp(\alpha\epsilon)} \ , \quad \epsilon \in \mathbb{R} \ .$$

[Note that the software **R** uses a different parameterization for the logistic distribution.] The case $\alpha = \lambda = 1$ is known as the standard logistic distribution. For the logistic distribution calculate the survivor function $S_\epsilon(\epsilon)$, the density $f_\epsilon(\epsilon)$ and the hazard function $h_\epsilon(\epsilon)$. Use your results to derive the corresponding functions for the standard logistic distribution as well.

(b) The transformed random variable $X = \exp(\epsilon)$ has a log-logistic distribution with parameters α and λ. For the transformed random variable X, compute the cdf $F_X(x)$, the survivor function $S_X(x)$, the density $f_X(x)$, the hazard rate $h_X(x)$ and the cumulative hazard rate $H_X(x)$.

(c) Consider the regression model $\ln(T) = Y = \boldsymbol{x}^\top\boldsymbol{\beta} + \sigma\epsilon$ with standard logistic distributed error ϵ. Derive the density $f_Y(y)$, survivor function $S_Y(y)$ and hazard rate $h_Y(y)$ for the transformed data $Y = \ln(T)$.

(d) Establish a relationship between the parameters α, λ of the logistic distributed random variable ϵ in (a) and the parameters $\boldsymbol{\beta}, \sigma$ of the log survival time $\ln(T)$ in (c).

(e) Suppose that right censored data $D_{(n)} = \{(t_i, \delta_i, \boldsymbol{x}_i), i = 1 \ldots, n\}$ is available. For the transformation $Y = \ln(T)$, compute the log-likelihood $l(\boldsymbol{\beta}, \sigma)$. Show that the score vector for the parameters $\boldsymbol{\beta}$ and σ is

$$s(\boldsymbol{\beta}, \sigma) = \left(\frac{\partial l(\boldsymbol{\beta}, \sigma)}{\partial \boldsymbol{\beta}}, \frac{\partial l(\boldsymbol{\beta}, \sigma)}{\partial \sigma} \right)^\top$$

with

$$\begin{aligned} \frac{\partial l(\boldsymbol{\beta}, \sigma)}{\partial \boldsymbol{\beta}} &= -\frac{1}{\sigma} \sum_{i=1}^{n} [\delta_i S_\epsilon(\epsilon_i) - F_\epsilon(\epsilon_i)] \boldsymbol{x}_i \\ \frac{\partial l(\boldsymbol{\beta}, \sigma)}{\partial \sigma} &= -\frac{1}{\sigma} \sum_{i=1}^{n} \delta_i (1 + \epsilon_i S_\epsilon(\epsilon_i)) - \epsilon_i F_\epsilon(\epsilon_i) \ . \end{aligned}$$

Hint: Use the notation $\epsilon_i = (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})/\sigma$, $i = 1, \ldots, n$.

(f) Show that the log-logistic model has the property of proportional odds. For that purpose, compute and interpret the quotient

$$\frac{S_T(t; \boldsymbol{x}_1)/(1 - S_T(t; \boldsymbol{x}_1))}{S_T(t; \boldsymbol{x}_2)/(1 - S_T(t; \boldsymbol{x}_2))} \ .$$