TMA 4275 Lifetime Analysis
June 2004
Solution

Problem 1

a) Observation of the outcome is censored, if the time of the outcome is not known exactly and only the last time when it was observed being intact, is presented.

The typical causes leading to a dataset containing censored observations are: the object under consideration may withdraw from the study, be lost to follow-up, or economic or practical reasons may require that the study ends before the outcome has occurred.

The intuitive content of the assumption of independent censoring can be formulated as follows: items cannot be censored (withdrawn) because they have a higher or lower risk than the average.

In the given situation there is an opportunity to get dependent censored data, for example, in this case: the owner of the car decides that the brake pad is already worn and replaces (or just stops the observation).

b) The method of computing the value of the KM-plot for Year = 0 at the time 45.1 is as follows:

1. Order all the observed survival times (only uncensored) for the observations with Year = 0:
$$Y_{(0)} = 0, \quad Y_{(1)} = 22.6, \quad Y_{(2)} = 22.7, \ldots, \; Y_{(17)} = 86.2$$

2. Define n_i as the number of subjects (both censored and uncensored), that are still intact and participating in the study, until just before the time $Y_{(i)}$ $i = 1, 2, \ldots, 17$:
$$n_1 = 20, \quad n_2 = 19, \ldots$$

3. Define d_i as the number of subjects (only the uncensored ones) that fail at the time Y_i, $i = 1, 2, \ldots, 17$:
$$d_1 = 1, \quad d_2 = 1, \ldots$$

4. For every $i = 1, 2, \ldots, 17$ compute the value of the KM-plot $\hat{R}(Y_{(i)})$:
$$\hat{R}(Y_{(i)}) = \hat{S}_i = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j}$$

Following this scheme, we have got the table:

i	$Y_{(i)}$	n_i	d_i
1	22.6	20	1
2	22.7	19	1
3	34.4	18	1
4	36.7	17	1
5	38.4	16	1
6	38.8	15	1
7	40.0	14	1
	41.0∗		
	42.2∗		
8	45.1	11	1

Due to $45.1 = Y_{(8)}$, the KM-plot at time 45.1 is computed by:

$$\hat{R}(45.1) = \hat{S}_8 = \frac{19}{20} \cdot \frac{18}{19} \cdot \frac{17}{18} \cdot \frac{16}{17} \cdot \frac{15}{16} \cdot \frac{14}{15} \cdot \frac{13}{14} \cdot \frac{10}{11} = \frac{13}{22} = 0.5909$$

The expression for the estimate of the variance of $\hat{R}(45.1)$ is:

$$\widehat{Var}[\hat{R}(45.1)] = [\hat{R}(45.1)]^2 \cdot \sum_{t_i \leq 45.1} \frac{d_i}{n_i(n_i - d_i)} = \left(\frac{13}{22}\right)^2 \cdot \left[\frac{1}{20 \cdot 19} + \cdots + \frac{1}{11 \cdot 10}\right] = 0.0126,$$

and therefore, the standard deviation estimate is:

$$\widehat{SD}[\hat{R}(45.1)] = \sqrt{\widehat{Var}[\hat{R}(45.1)]} = 0.1122.$$

The median lifetime can be roughly estimated using the KM-plot: the median corresponds to the abscissa of the point on the plot, whose ordinate corresponds to 50%.
From Figure 2 one can find the median:
for Region=0: ≈ 48,
for Region=1: ≈ 51.

c) The distribution of survival times can also be presented using the *hazard function*, which is defined as:

$$z(t) = \lim_{\Delta t \to 0} \frac{\Pr[(t \leq T < t + \Delta t) | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)}.$$

In terms of the hazard function, the lifetime distribution for the two year models Year=0 and Year=1 (ignoring other factors) can be given by $z_0(t)$ for Year=0 and $z_1(t)$ for Year=1. Therefore, the hypotheses are formulated as follows:

$$\begin{array}{ll} H_0: & z_0(t) = z_1(t) \quad \text{for all } t \\ H_1: & \text{otherwise} \end{array}$$

The Cox model supposes the special type of hazard function, namely:

$$z(t;x) = z_0(t) \cdot e^{\beta x},$$

where $z_0(t)$ is some function and x is the covariate Year. In terms of the Cox model, the hypotheses can be reformulated like this:

$$\begin{aligned} H_0: &\quad \beta = 0, \\ H_1: &\quad \beta \neq 0. \end{aligned}$$

The test was performed; p-value for $H_0 : \beta = 0$ is $P = 0.206$. P is high and H_0 should be accepted. (For example, $P = 0.206 > 0.05$, therefore H_0 cannot be rejected at 5% level). The plots also seem to be similar.

A similar analysis for the factor *Region* gave the p-value $P = 0.269$. H_0 also should be accepted.

Problem 2

a) One should prefer a model where the factor *Driving* is represented by two covariates x_3 and x_4, because the values of *Driving* are categorical and do not measure anything, and splitting also simplifies the model.

$$\ln T_{\boldsymbol{x}} \sim N(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,\ \sigma^2),$$

and so,

$$T_{\boldsymbol{x}} \sim lognormal(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4,\ \sigma^2).$$

The median and expectation for $T_{\boldsymbol{x}}$ are, respectively,

$$\text{median}(T_{\boldsymbol{x}}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)$$

and

$$ET_{\boldsymbol{x}} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \frac{\sigma^2}{2}).$$

b) Suppose that the data are written as $(y_i, d_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$, where y_i are the observed values of the response variable, d_i gives the censoring status and $\boldsymbol{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})$ is the i-th covariate vector. The likelihood function is

$$\prod_{i:\, d_i=1} \phi\left(\frac{\ln y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \beta_4 x_4}{\sigma}\right) \frac{1}{\sigma y_i} \times$$

$$\times \prod_{i:\, d_i=0} \left(1 - \Phi\left(\frac{\ln y_i - \beta_0 - \beta_1 x_1 - \beta_2 x_2 - \beta_3 x_3 - \beta_4 x_4}{\sigma}\right)\right)$$

where $\phi(x)$ and $\Phi(x)$ are the density and cumulative distribution function, respectively, of the standard normal distribution.

The reason for the Φ is that censored observations are represented by their reliability function rather than the density.

c) In the given model, a brake pad for a car with the parameters $Year = 0$, $Region = 1$, *mixed driving* corresponds to the covariate vector

$$\boldsymbol{x}^* = (x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 1).$$

To find the point estimate of the *median lifetime* of such a car, one needs to substitute the estimated regression coefficients given by the MINITAB output into the expression for the median of $T_{\boldsymbol{x}}$:

$$\mu(T_{\boldsymbol{x}^*}) = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4) \approx e^{3.45156+0.173254+0.496124} \approx$$

$$\approx 61.61701 \approx 61.62.$$

The *expected lifetime* of that car is,

$$ET_{\boldsymbol{x}^*} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \frac{\sigma^2}{2}) \approx$$

$$\approx \mu(T_{\boldsymbol{x}^*}) e^{0.319886^2/2} \approx 64.85.$$

To find the 0.10-quantile of the car with the covariates vector $\boldsymbol{x}^*$, one needs to find the value $t_{0.10}$, which satisfies the equation:

$$P(T_{\boldsymbol{x}^*} \leq t_{0.10}) = 0.10.$$

But

$$P(T_{\boldsymbol{x}^*} \leq t_{0.10}) = P(\ln T_{\boldsymbol{x}^*} \leq \ln t_{0.10}) =$$

$$= \Phi\left(\frac{\ln t_{0.10} - \beta_0 - \beta_2 - \beta_4}{\sigma}\right) = 0.10;$$

using the table of quantiles of the standard normal distribution,

$$\frac{\ln t_{0.10} - \beta_0 - \beta_2 - \beta_4}{\sigma} \approx -1.282,$$

hence,

$$t_{0.10} \approx e^{-1.282\sigma + \beta_0 + \beta_2 + \beta_4}.$$

And, substituting the results of MINITAB,

$$t_{0.10} \approx \exp(-1.282 \cdot 0.319886 + 3.45156 + 0.173254 + 0.496124) \approx 40.89$$

d) For this model the standardized residuals are defined as:

$$\hat{r}_i = \frac{\ln y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta}_3 x_{i3} - \hat{\beta}_4 x_{i4}}{\hat{\sigma}}, \quad i = 1, 2, \ldots, n.$$

If the model is correct, the distribution of $\hat{r}_i$ is approximately $N(0, 1)$.

$$\hat{r}_1 = \frac{\ln 22.6 - 3.45156}{0.319886} \approx -1.04.$$

According to the *Figure 3*, the model fits the data well: a significant part (60 %) of the residuals are in range of $[-1, 1]$ and 40% - in range of $[-0.5, 0.5]$, that shows a good accordance with the model, the median of the residual values is approximately 0.

e) The p-values of x_1 and x_2 in the table of MINITAB output are bigger than, say, 0.05; therefore these covariates do not have significant effect on the model.

In the given model, the situation when a covariate does not have significant effect implies that its coefficient is close to zero. Therefore, to decide if to remove the covariates x_1 and x_2 one should test the null hypothesis:

$$H_0 : \quad \beta_1 = \beta_2 = 0.$$

To decide whether there are reasons to include only x_3 and x_4 to the regression model, the likelihood- ratio statistic W should be investigated:

$$W = 2(\ln L_f - \ln L_c),$$

where $\ln L_c$ is the log-likelihood of the model under H_0, and $\ln L_f$ is the log-likelihood of the full model. The likelihood ratio statistic has a Chi-Square distribution with $K_2 - K_1$ degrees of freedom, where K_2 and K_1 denote the number of parameters in the full model and the model under H_0, respectively. If $W < \chi^2_{0.05,2}$, where $\chi^2_{0.05,2}$ is the 0.05-quantile of the $\chi^2_{K_2-K_1}$ distribution, H_0 should be rejected. But

$$W = 2(-132.772 - (-135.335)) = 5.126 < 5.991 = \chi^2_{0.05,2},$$

therefore, H_0 is not rejected for $\alpha = 5\%$.

Problem 3

a) A cumulative distribution function of a lifetime T must be a distribution function of a continuous and positive random variable. That can be reformulated as the following requirements:

1. $F(0) = 0$;

2. $F(t)$ is an increasing function on $(0;\ \infty)$;

3. $\lim\limits_{t\to+\infty} F(t) = 1$;

4. $F(t)$ has a derivative on $(0;\ \infty)$.

Check them:

1. $F(0) = \lim\limits_{t\to 0+} F(t) = \lim\limits_{t\to 0+} e^{-\left(\frac{\theta}{t}\right)^{\alpha}} = 0$;

2. $F(t)\ \nearrow$, because as $t \to +\infty$: $\left(\frac{\theta}{t}\right) \searrow \implies \left(\frac{\theta}{t}\right)^{\alpha} \searrow \implies -\left(\frac{\theta}{t}\right)^{\alpha} \nearrow \implies e^{-\left(\frac{\theta}{t}\right)^{\alpha}} \nearrow$;

3. $\lim\limits_{t\to+\infty} F(t) = \lim\limits_{t\to+\infty} e^{-\left(\frac{\theta}{t}\right)^{\alpha}} = 1$;

4. The derivative of $F(t)$ can be found as follows:

$$f(t) = F'(t) = e^{-\left(\frac{\theta}{t}\right)^{\alpha}}(-1)\alpha\left(\frac{\theta}{t}\right)^{\alpha-1}\theta\left(-\frac{1}{t^2}\right) = \alpha\,\theta^{\alpha}\left(\frac{1}{t}\right)^{\alpha+1} e^{-\left(\frac{\theta}{t}\right)^{\alpha}}.$$

The expression for the hazard rate is by definition:

$$z(t) = \frac{f(t)}{1-F(t)}.$$

b) $-\ln F(t) = \left(\frac{\theta}{t}\right)^{\alpha}$, therefore

$$\ln(-\ln F(t)) = \alpha(\ln\theta - \ln t).$$

Assume given a right censored dataset. Let $\hat{R}(t_i)$ be the Kaplan-Meier estimator in each failure time t_i. Then plot the points

$$(\ln t_i, \ln(-\ln(1-\hat{R}(t_i))))$$

This is then ideally a straight line with slope $-\alpha$ and intercept $\alpha\ln\theta$.
An even better plot is obtained by replacing $\hat{R}(t_i)$ by

$$\frac{\hat{R}(t_i)+\hat{R}(t_{i-1})}{2}$$

c) If $Y = \ln T$, then

$$F_Y(y) = P(Y \le y) = P(\ln T \le y) = P(T \le e^y) = e^{-\left(\frac{\theta}{e^y}\right)^{\alpha}} =$$

$$= e^{-\theta^{\alpha}e^{-\alpha y}} = e^{-e^{-\alpha(y-\ln\theta)}} = \Phi_0\left(\frac{y-\mu}{\sigma}\right),$$

where

$$\Phi_0(x) = e^{-e^{-x}}, \quad \mu = \ln\theta \quad \text{and} \quad \sigma = \frac{1}{\alpha}.$$