Norges teknisk–
naturvitenskapelige universitet
Institutt for matematiske fag

Contact during exam:
Bo Lindqvist 975 89 418

EXAM IN TMA4275 LIFETIME ANALYSIS

Thursday 3 June 2004

Time: 09:00–14:00

Aids:
All printed and handwritten sources, and the calculator HP30S.

Grading: 18 June 2004

The table on the last page in the problem set shows observed lifetimes for brake pads for a random sample of 40 cars of the same make. The variable Y (with unit 1000 km) gives the observed lifetimes, while D gives the censoring status ($D = 1$ means non-censored, $D = 0$ means right censored). Three factors are, furthermore, given for each car:

Year Possible values are 0 and 1, which correspond to two different year models

Region Possible values are 0 and 1, which correspond to two different geographical regions

Driving Possible values are 0, 1 and 2 with the following meaning: 0 = mainly city driving, 1 = mainly highway driving, 2 = mixed driving

The 40 observations, each represented by a line in the table, are assumed to be stochastically independent. In addition we assume independent censoring.

The data are used in Problems 1 and 2.

Problem 1

a) What does it mean that an observation is censored?

What are typical causes leading to a dataset containing censored observations?

Explain briefly what is the intuitive content of the assumption of independent censoring for the brake pad data of this problem set. Can one in this situation also suspect that there is dependent censoring?

b) In order to conclude whether the lifetime of brake pads depend on year model and region, one first performs separate analyses for each of these factors.

Figure 1 shows a Kaplan-Meier plot (KM-plot) when the data are divided into two groups according to the value of the factor *Year*. One here ignores the values of the factors *Region* and *Driving*.

Similarly, Figure 2 shows the KM-plot when the data are divided in two groups according to the value of the factor *Region*.

Show how one computes the value of the KM-plott for Year=0 at the time 45.1. (Use the data at the end of the problem set).

Also write down an expression for estimated standard deviation for the value you found at time 45.1. (Complete computation is not necessary).

How will you from Figure 2 roughly estimate the median lifetime for brake pads for, respectively, Region=0 and Region=1?

c) One wants to test the null hypothesis that the lifetime distribution for the brake pads is the same for the two year models Year=0 and Year=1, when the other factors are ignored.

Formulate a null hypothesis and alternative hypothesis for this. Then write down a simple Cox-model which leads to a test for this situation. (You are not asked to perform the testing, nor writing down the partial likelihood.)

This test was performed, and the p-value was 0.206. Use this information together with Figure 1 to conclude whether the lifetime of the brake pads reasonably can be assumed to be independent of the car's year model (*Year*).

A similar analysis for the factor *Region* gave the p-value 0.269. Which conclusion do you draw from this information, possibly also using Figure 2?

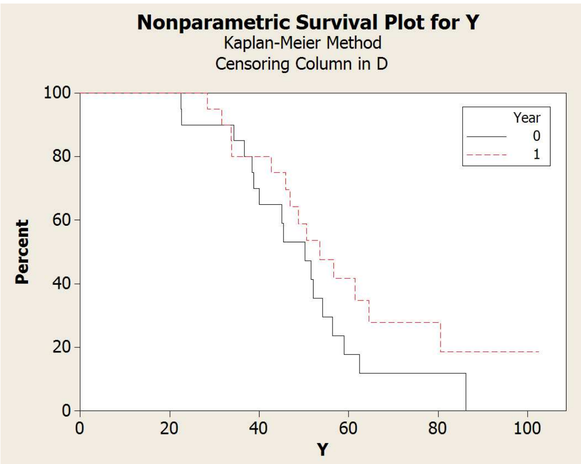

Figure 1

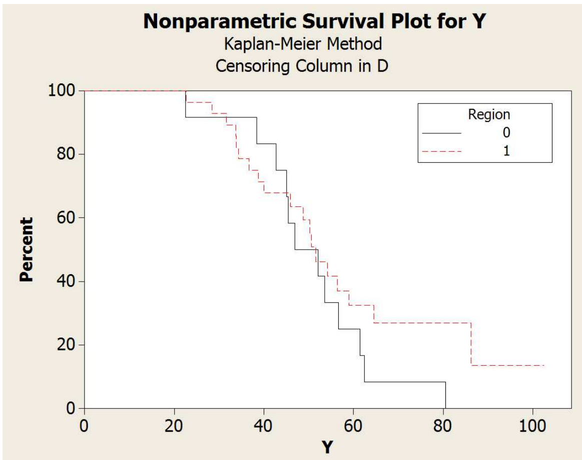

Figure 2

Problem 2

In order to find out how the lifetime of the brake pads depend on the three factors *Year*, *Region* and *Driving*, one performs a regression analysis.

Let the covariates x_1 and x_2 give the value of the factors *Year* and *Region*, respectively. Further define x_3 and x_4 as covariates with values 0 and 1 such that

$x_3 = 1$ if the factor *Driving* has the value 1, and $x_3 = 0$ otherwise

$x_4 = 1$ if the factor *Driving* has the value 2, and $x_4 = 0$ otherwise

Let $T_{\boldsymbol{x}}$ be the lifetime of a brake pad with covariate vector $\boldsymbol{x} = (x_1, x_2, x_3, x_4)$. The following model is suggested:

$$\ln T_{\boldsymbol{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \sigma Z \tag{1}$$

where Z is standard normally distributed, and $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ and σ are unknown parameters.

a) Explain briefly why one should prefer a model where the factor *Driving* is represented by two covariates x_3 and x_4, instead of a model where one lets x_3 be the value of the factor *Driving* (0, 1 or 2) and omits x_4.

What is the distribution of the lifetime $T_{\boldsymbol{x}}$?

What are, respectively, median and expectation for $T_{\boldsymbol{x}}$ expressed by the parameters of the model?

b) Write down, and briefly give the reasons for, an expression for the likelihood function for the model in (1) when the data are written as $(y_i, d_i, \boldsymbol{x}_i)$, $i = 1, 2, \ldots, n$. (Her $\boldsymbol{x}_i = (x_{1i}, x_{2i}, x_{3i}, x_{4i})$ is the covariate vector for car number i.) You may express the likelihood function by the density ϕ and the cumulative distribution function Φ of the standard normal distribution.

c) MINITAB gives the following output when the model in (1) is used and the data are as given at the end of the problem set.

Regression Table

Predictor	Coef	Standard Error	Z	P	95,0% Normal CI Lower	Upper
Intercept	3,45156	0,138739	24,88	0,000	3,17964	3,72349
x1	0,192203	0,107675	1,79	0,074	-0,0188358	0,403243
x2	0,173254	0,117168	1,48	0,139	-0,0563908	0,402898
x3	0,389199	0,139015	2,80	0,005	0,116736	0,661663
x4	0,496124	0,127148	3,90	0,000	0,246919	0,745329
Scale	0,319886	0,0413712			0,248261	0,412175

Log-Likelihood = -132,772

Use these results to find point estimates of the *median lifetime* and *expected lifetime* for a brake pad for a car with *Year*=0 and *Region*=1, when the car is used in *mixed driving.*

Also give a point estimate of the 0.10-quantile of the corresponding lifetime. (Recall that the p-quantile t_p for a lifetime T is defined by $P(T \leq t_p) = p$.)

(*Hint*: *Scale* in the MINITAB output is the estimate for σ.)

d) How are the standardized residuals defined for the model and data in this problem?

Which distribution is, approximately, expected for the standardized residuals when the model is correct?

Compute the standardized residual for the observation (*Row*) number 1 in the dataset.

A probability plot of the standardized residuals is given in Figure 3. Give a comment regarding the model fit for the given data.

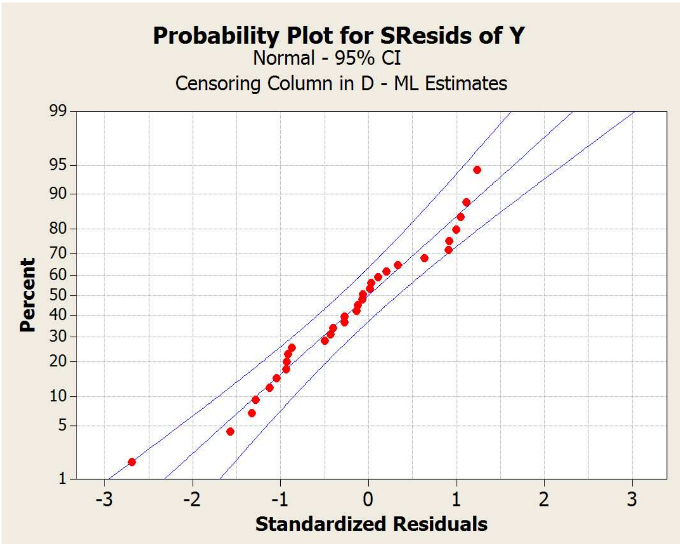

Figure 3

e) The MINITAB output of question 2(c) indicates that the covariates x_1 and x_2 do not have significant effect in the model. How can you read this from the output?

One therefore considers to remove the two covariates from the model. Explain why this problem can be viewed as a hypothesis testing situation where one tests the null hypothesis

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

A session in MINITAB using the model where only x_3 and x_4 are included as covariates, gave a log likelihood of -135.335. Does this give reasons to reject the null hypothesis if the level is set to 5%, and you use a likelihood test?

Problem 3

Let the lifetime T have the cumulative distribution function

$$F(t) = P(T \leq t) = e^{-\left(\frac{\theta}{t}\right)^{\alpha}} \quad \text{for } t > 0, \tag{2}$$

where $\alpha > 0$, $\theta > 0$ are parameters.

This distribution is known as the Frechet distribution, and is also called the inverse Weibull distribution because of the form of the cumulative distribution function.

a) Which requirements does a cumulative distribution function of a lifetime T have to satisfy? Check that these requirements are satisfied for $F(t)$ in (2).

Find the density $f(t)$ of T and write down an expression for the hazard rate $z(t)$ for T.

b) Compute $\ln(-\ln F(t))$.

Explain how this can be used to construct a probability plot for checking whether a right censored dataset can be assumed to come from the distribution (2).

Also explain how the parameters α and θ can be estimated graphically from this plot.

c) Show that (2) defines a log-location-scale family, i.e. that $Y = \ln T$ has a cumulative distribution function of the form

$$F_Y(y) = P(Y \leq y) = \Phi_0\left(\frac{y - \mu}{\sigma}\right)$$

for a cumulative distribution function Φ_0 and constants μ (location) og σ (scale).

Find the function Φ_0, and find the constants μ and σ expressed in terms of α and θ.

Row	Y	D	Year	Region	Driving
1	22,6	1	0	0	0
2	22,7	1	0	1	0
3	28,4	1	1	1	1
4	31,7	1	1	1	0
5	33,8	1	1	1	0
6	33,9	1	1	1	0
7	34,4	1	0	1	0
8	36,7	1	0	1	1
9	38,4	1	0	0	2
10	38,8	1	0	1	0
11	40,0	1	0	1	0
12	41,0	0	0	1	2
13	42,2	0	0	1	1
14	42,7	1	1	0	1
15	42,8	0	1	1	2
16	45,1	1	0	0	2
17	45,5	1	0	0	1
18	45,9	1	1	1	0
19	46,9	1	1	0	0
20	48,8	1	1	1	2
21	50,2	1	0	1	0
22	50,6	1	1	1	0
23	50,7	0	1	1	2
24	51,6	1	0	1	0
25	52,1	1	0	0	2
26	53,6	1	1	0	2
27	54,2	1	0	1	2
28	56,4	1	0	1	2
29	56,7	1	1	0	0
30	59,0	1	0	1	2
31	59,8	0	1	1	0
32	61,5	1	1	0	2
33	62,4	1	0	0	1
34	64,5	1	1	1	1
35	73,1	0	1	1	1
36	80,6	1	1	0	1
37	81,3	0	0	1	1
38	81,7	0	1	1	2
39	86,2	1	0	1	2
40	102,5	0	1	1	2