Reliability Data and Survival Analysis
Homework 7

Chenghua Liu
liuch18@mails.tsinghua.edu.cn
Department of Computer Science
Tsinghua University

1

Suppose that the hazard function $\lambda(y)$ for a variable Y satisfies

$$\log \lambda(y) = \alpha + \beta y.$$

Show that $T = e^Y$ follows a Weibull distribution. State the parameters of that Weibull distribution (λ and γ) in terms of α and β.
solution:

$$\begin{aligned} S_Y(y) &= \exp\left\{-\int_{-\infty}^{y} \lambda(t)dt\right\} \\ &= \exp\left\{-\int_{-\infty}^{y} e^{\alpha+\beta t}dt\right\} \\ &= \exp\left\{-\frac{e^{\alpha+\beta y}}{\beta}\right\} \end{aligned}$$

So, we have

$$S_T(t) = P(T > t) = P\left(e^Y > t\right) = P\left(Y > \log t\right) = S_Y(\log t) = \exp\left\{-\frac{e^{\alpha}t^{\beta}}{\beta}\right\}$$

Therefore, T follows Weibull distribution($S(t) = e^{-(\lambda t)^p}$) with parameters

$$p = \beta, \quad \lambda = \left(\frac{e^{\alpha}}{\beta}\right)^{\frac{1}{\beta}}$$

2

Analyze the data GVHD using exponential regression.

(a) Obviously, the failure times do not even remotely resemble an exponential distribution, since they all occur within the first 60 days and the survival function is completely flat past that point. To make the distribution more exponential-like, consider all times past 60 days to be censored (to be clear: do not throw any observations out, and do not touch any subjects with times on study less than 60 days). Make a linear diagnostic plot and comment on whether the modified data look reasonably exponential in distribution.

(b) Fit an exponential regression model with three covariates: Group, Age (as a continuous numeric quantity), and the interaction between Group and Age. Carry out a Wald test of the interaction term. Report your result and comment on what it means.

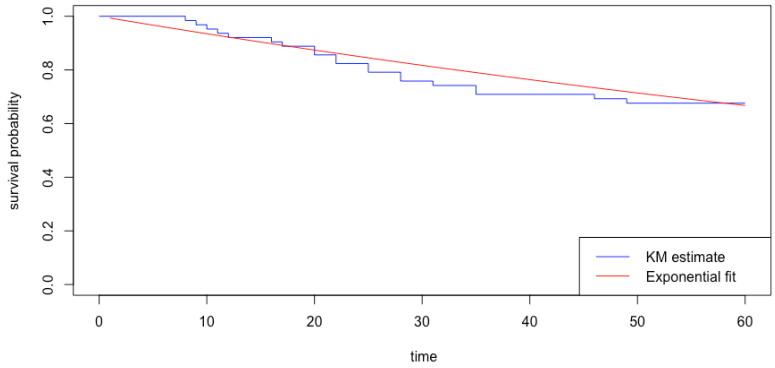
(c) For the model in (b), estimate the hazard ratio for the Group term and provide a 95% confidence interval. Comment on the meaning of these results.

(d) Re-fit the model, only this time, subtract 21 (the median age) from Age. Repeat part (c) for this new model. Comment on the meaning of the hazard ratios you obtain, and how and why they differ from (c). What is the λ parameter (i.e., the "baseline" hazard) of the exponential regression model?

solution:
(a)

```
library(survival)
gvhd <- read.table('/Users/liuchenghua/Downloads/gvhd.txt', header = TRUE)
gvhd_a <- gvhd
gvhd_a$Time[gvhd_a$Time>60] <- 60
fit.KM <- survfit(Surv(Time, Status) ~ 1, gvhd_a)

t_int <- 1:max(gvhd_a$Time)

fit.exp <- survreg(Surv(Time, Status) ~ 1, gvhd_a, dist = "exponential")

prob_exp <- 1 - pexp(t_int, exp(-fit.exp$coefficients))

plot(fit.KM, conf.int = FALSE, xlab = "time", ylab = "survival probability", col="blue")
lines(prob_exp ~ t_int, col="red")
legend("bottomright", c("KM estimate","Exponential fit"), lty = 1,col = c("blue", "red"))
```

We make a linear diagnostic plot as follow.And the modified data look reasonably exponential in distribution.



(b)

```
fit.exp <- survreg(Surv(Time,Status) ~ Group*Age, gvhd, dist = 'exponential')
summary(fit.exp)
#####################################
Call:
survreg(formula = Surv(Time, Status) ~ Group * Age, data = gvhd,
    dist = "exponential")
                    Value Std. Error     z        p
(Intercept)        9.8081     0.8824 11.12 < 2e-16
GroupMTX+CSP      -2.8297     1.3607 -2.08  0.03756
Age               -0.1791     0.0328 -5.46  4.7e-08
GroupMTX+CSP:Age   0.2171     0.0587  3.70  0.00022

Scale fixed at 1

Exponential distribution
Loglik(model)= -136.6   Loglik(intercept only)= -157.8
        Chisq= 42.44 on 3 degrees of freedom, p= 3.2e-09
Number of Newton-Raphson Iterations: 6
n= 64
```

Wald test for interaction term produces have $p = 0.00022$, which means that we supposed that treatment effect differs for different age.

(c)

The hazard ratio for the Group term is

$$\frac{\lambda(t \mid \text{ Group } = \text{MTX} + \text{CSP}, \text{ Age } = 0)}{\lambda(t \mid \text{ Group } = \text{MTX}, \text{ Age } = 0)}$$

```
> exp(-coef(fit.exp)[2])
GroupMTX+CSP
    16.94061
> exp(-confint(fit.exp)[2, ])
     2.5 %     97.5 %
```

```
6  243.862257    1.176829
```

So the estimate of hazard ratio is 16.94061 and the 95% confidence interval is $[1.177, 243.862]$.We found that MTX+CSP is worse than MTX when Age equals to 0.The result is usefuless.

(d)

```
1  gvhd$Age_handle <- gvhd$Age - 21
2  fit.exp.Age_handle <- survreg(Surv(Time,Status) ~ Group*Age_handle, gvhd, dist = 'exponential')
3  summary(fit.exp.Age_handle)
4  exp(-coef(fit.exp.Age_handle)[2])
5  exp(-confint(fit.exp.Age_handle)[2, ])
```

The estimate of hazard ratio is 0.1772577 and the 95% confidence interval is $[0.06, 0.52]$.We found that MTX+CSP is better than MTX when Age equals to 23.The result is reasonable.

3

Now fit a Weibull AFT model to the data GVHD, again including Age, Group, and the Group by Age interaction.

(a) Is there significant evidence that the Weibull model provides a superior fit to the data than an exponential model?

(b) Provide an estimate along with a 95% confidence interval for the Weibull shape parameter γ. Provide some explanation for how you arrived at this interval.

solution:

(a)

```
1  > fit.weibull <- survreg(Surv(Time,Status) ~ Group*Age, gvhd, dist = 'weibull')
2  > summary(fit.weibull)
3
4  Call:
5  survreg(formula = Surv(Time, Status) ~ Group * Age, data = gvhd,
6      dist = "weibull")
7                       Value Std. Error     z        p
8  (Intercept)        11.4801     1.7713  6.48 9.1e-11
9  GroupMTX+CSP       -3.8045     2.6026 -1.46 0.14380
10 Age                -0.2332     0.0634 -3.68 0.00023
11 GroupMTX+CSP:Age    0.3077     0.1172  2.63 0.00864
12 Log(scale)          0.6719     0.1956  3.44 0.00059
13
14 Scale= 1.96
15
16 Weibull distribution
17 Loglik(model)= -128.3   Loglik(intercept only)= -139.9
18         Chisq= 23.22 on 3 degrees of freedom, p= 3.6e-05
19 Number of Newton-Raphson Iterations: 7
20 n= 64
```

We found that p = 0.00059 which means log(scale) is significant.Therefore, the Weibull model provides a superior fit to the data than an exponential model.

(b)

The estimate of shape parameter γ is $\frac{1}{Scale} = \frac{1}{1.96} = 0.51$ and the 95% confidence interval is $[0.35, 0.75]$

```
> exp(-log(fit.weibull$scale)+ c(-1, 1)*qnorm(0.975)*sqrt(fit.weibull$var[5, 5]))
 [1] 0.3481125 0.7492879
```

4

Consider the Veterans' administration lung cancer data (the data can be found in R: library(survival) and then data(veteran). There are six covariates collected for each patient: treatment, cell type (four categories which need three dummy variables), Karnofsky score, months from diagnosis, age and indicator of prior therapy. Fit an exponential AFT model to the data with these six covariates in the model. Answer the following questions:

(a) Compare the mean survival times for two treatments.

(b) Since the exponential AFT model is also a proportional hazards model, find the estimates and 95% CIs for the hazard ratios comparing two treatments.

solution:

(a)

```
library(survival)
data(veteran)
fit.exp <- survreg(Surv(time, status) ~ ., veteran, dist = "exponential")
summary(fit.exp)
##########################
Call:
survreg(formula = Surv(time, status) ~ ., data = veteran, dist = "exponential")
                    Value Std. Error     z       p
(Intercept)      3.408176   0.734373  4.64 3.5e-06
trt             -0.219565   0.198634 -1.11  0.2690
celltypesmallcell -0.820245   0.262111 -3.13  0.0018
celltypeadeno    -1.113121   0.275825 -4.04 5.4e-05
celltypelarge    -0.377220   0.272626 -1.38  0.1665
karno             0.030624   0.005108  6.00 2.0e-09
diagtime         -0.000297   0.008970 -0.03  0.9736
age               0.006108   0.009161  0.67  0.5049
prior            -0.004948   0.022687 -0.22  0.8273

Scale fixed at 1

Exponential distribution
```

```
Loglik(model)= -716.2   Loglik(intercept only)= -751.2
        Chisq= 70.12 on 8 degrees of freedom, p= 4.6e-12
Number of Newton-Raphson Iterations: 5
n= 137
```

The mean of trt 2 / mean of trt 1 is 0.8028677.

```
> exp(coef(fit.exp)[2])
      trt
0.8028677
```

(b)

```
> exp(-coef(fit.exp)[2])
    trt2
1.245535
> exp(-confint(fit.exp)[2, ])
    2.5 %    97.5 %
1.8383723 0.8438758
```

The estimate of hazard ratio is 1.245535 and the 95% confidence interval is $[0.84, 1.84]$.

5

For data in section 4, fit a Weibull AFT model, and answer the following questions:

(a) Compare the mean survival times for two treatments.

(b) Since the Weibull AFT model is also a proportional hazards model, find the estimates and 95% CIs for the hazard ratios comparing two treatments.

solution:

(a)

```
> fit.wei <- survreg(Surv(time, status) ~ ., veteran, dist = "weibull")
> summary(fit.wei)
Call:
survreg(formula = Surv(time, status) ~ ., data = veteran, dist = "weibull")
                     Value Std. Error     z       p
(Intercept)       3.262014   0.662531  4.92 8.5e-07
trt2             -0.228523   0.186844 -1.22  0.2213
celltypesmallcell -0.826185   0.246312 -3.35  0.0008
celltypeadeno     -1.132725   0.257598 -4.40 1.1e-05
celltypelarge     -0.397681   0.254749 -1.56  0.1185
karno              0.030068   0.004828  6.23 4.7e-10
diagtime          -0.000469   0.008361 -0.06  0.9553
age                0.006099   0.008553  0.71  0.4758
prior             -0.004390   0.021228 -0.21  0.8362
Log(scale)        -0.074599   0.066311 -1.12  0.2606

```

```
Scale= 0.928

Weibull distribution
Loglik(model)= -715.6   Loglik(intercept only)= -748.1
        Chisq= 65.08 on 8 degrees of freedom, p= 4.7e-11
Number of Newton-Raphson Iterations: 6
n= 137
```

The mean of trt 2 / mean of trt 1 is 0.7957083.

```
> exp(coef(fit.wei)[2])
     trt2
0.7957083
```

(b)

```
> exp(-coef(fit.wei)[2]/fit.wei$scale)
    trt2
1.279184
> f <- log(-coef(fit.wei)[2]/fit.wei$scale)
> df <- c(1/coef(fit.wei)[2], -1)
> varf = as.vector(t(df) %*% fit.wei$var[c(2, 10), c(2, 10)] %*% df)
> exp(exp(f + c(-1, 1)*qnorm(0.975)*sqrt(varf)))
[1] 1.050269 3.442110
```

The estimate of hazard ratio is 1.279184 and the 95% confidence interval(Delta Method) is $[1.05, 3.44]$.

6

For data in section 4, fit a log-logistic model, and answer the following questions:

(a) Compare the mean survival times for two treatments.

(b) Find the estimates and 95% CIs of odds-ratios comparing two treatments.

(c) Extend to individual frailty model to account for the possible heterogeneity. Please write down the model explicitly. Derive the corresponding unconditional (or marginal) hazard function and survival function. Based on your analysis, is it better to include frailty factor or not?

solution:

(a)

```
> fit.log <- survreg(Surv(time, status) ~ ., veteran, dist = "loglogistic")
> summary(fit.log)
Call:
survreg(formula = Surv(time, status) ~ ., data = veteran, dist = "loglogistic")
```

```
                      Value Std. Error     z        p
(Intercept)         2.02426    0.67836  2.98   0.0028
trt2               -0.08846    0.17933 -0.49   0.6218
celltypesmallcell  -0.70798    0.24925 -2.84   0.0045
celltypeadeno      -0.74255    0.27213 -2.73   0.0064
celltypelarge       0.01663    0.26956  0.06   0.9508
karno               0.03606    0.00448  8.04 8.8e-16
diagtime            0.00211    0.01027  0.21   0.8375
age                 0.00855    0.00892  0.96   0.3381
prior              -0.01020    0.02112 -0.48   0.6291
Log(scale)         -0.54645    0.07419 -7.37 1.8e-13

Scale= 0.579

Log logistic distribution
Loglik(model)= -711.9   Loglik(intercept only)= -750.3
        Chisq= 76.65 on 8 degrees of freedom, p= 2.3e-13
Number of Newton-Raphson Iterations: 4
n= 137
```

The mean of trt 2 / mean of trt 1 is 0.9153378.

```
> exp(coef(fit.log)[2])
     trt2
0.9153378
```

(b)

```
> exp(-coef(fit.log)[2]/fit.log$scale)
    trt2
1.165074
> f <- log(-coef(fit.log)[2]/fit.log$scale)
> df <- c(1/coef(fit.log)[2], -1)
> varf = as.vector(t(df) %*% fit.log$var[c(2, 10), c(2, 10)] %*% df)
> exp(exp(f + c(-1, 1)*qnorm(0.975)*sqrt(varf)))
[1]     1.002879 3363.817996
```

The estimate of hazard ratio is 1.165074 and the 95% confidence interval(Delta Method) is $[1.002879, 3363.817996]$.

(c) Hazard and survival conditioned on individual frailty:

$$h(t \mid \alpha) = \alpha h(t)$$
$$S(t \mid \alpha) = S(t)^{\alpha}$$

Suppose α is gamma distribution with variance θ. Then we have

$$\mathrm{S_U(t)} = \int_0^{\infty} S(t \mid \alpha) g\{\alpha\} d\alpha = [1 - \theta \ln S(t)]^{-1/\theta}$$
$$h_{\mathrm{U}}(t) = \frac{-d\left[S_{\mathrm{U}}(t)\right]/dt}{S_{\mathrm{U}}(t)} = \frac{h(t)}{1 - \theta \ln[S(t)]}$$

I don't konw how to analysis that is it better to include frailty factor or not.