HW#2 Survival Analysis I

NAME: JIA-HAN SHIH

Problem 1. (Exercise 2, p.30)

Weibull regression: Let $\log T_i = \alpha_0 + \boldsymbol{\alpha}'\mathbf{x}_i + \sigma\varepsilon_i$, where $\Pr(\varepsilon_i > x) = \exp(-e^x)$ for $-\infty < x < \infty$.

(1) Derive the survival function $S(t \mid \mathbf{x}_i)$ and the hazard function $h(t \mid \mathbf{x}_i)$.

Solution (1).

By straightforward calculations, we have

$$\begin{aligned} S(t \mid \mathbf{x}_i) &= \Pr(T_i > t \mid \mathbf{x}_i) = \Pr(\log T_i > \log t \mid \mathbf{x}_i) = \Pr(\alpha_0 + \boldsymbol{\alpha}'\mathbf{x}_i + \sigma\varepsilon_i > \log t \mid \mathbf{x}_i) \\ &= \Pr\left(\varepsilon_i > \frac{\log t - \alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_i}{\sigma} \,\middle|\, \mathbf{x}_i\right) = \exp\left(-e^{\frac{\log t - \alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_i}{\sigma}}\right) \\ &= \exp\left\{-t^{1/\sigma}\exp\left(\frac{-\alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_i}{\sigma}\right)\right\} \end{aligned}$$

and

$$h(t \mid \mathbf{x}_i) = -\frac{\partial}{\partial t}\log S(t \mid \mathbf{x}_i) = \frac{t^{1/\sigma - 1}}{\sigma}\exp\left(\frac{-\alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_i}{\sigma}\right).$$

We obtained the desired results.

(2) Show that the model can be expressed as $h(t \mid \mathbf{x}_i) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i)$.

Solution (2).

The hazard function obtained in (1) can be expressed as

$$h(t \mid \mathbf{x}_i) = \frac{t^{1/\sigma - 1}}{\sigma}\exp\left(\frac{-\alpha_0 - \boldsymbol{\alpha}'\mathbf{x}_i}{\sigma}\right) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_i),$$

where $\boldsymbol{\beta}' = -\boldsymbol{\alpha}'/\sigma$ and $h_0(t) = (1/\sigma)t^{1/\sigma - 1}\exp(-\alpha_0/\sigma)$ is the baseline hazard function.

(3) Show $\Pr(T_i > t+w \mid T_i > t, \mathbf{x}_i) < \Pr(T_i > w \mid \mathbf{x}_i)$ for $0<\sigma<1$ and $w>0$. What does this inequality imply?

Solution (3).

By straightforward calculations,

$$
\begin{aligned}
&\Pr(T_i > t+w \mid T_i > t, \mathbf{x}_i) < \Pr(T_i > w \mid \mathbf{x}_i) \\
\Leftrightarrow\ &\frac{\Pr(T_i > t+w \mid \mathbf{x}_i)}{\Pr(T_i > t \mid \mathbf{x}_i)} < \Pr(T_i > w \mid \mathbf{x}_i) \\
\Leftrightarrow\ &\Pr(T_i > t+w \mid \mathbf{x}_i) < \Pr(T_i > t \mid \mathbf{x}_i)\Pr(T_i > w \mid \mathbf{x}_i) \\
\Leftrightarrow\ &\exp\left\{-(t+w)^{1/\sigma}\exp\left(\frac{-\alpha_0-\boldsymbol{\alpha}'\mathbf{x}_i}{\sigma}\right)\right\} < \exp\left\{-(t^{1/\sigma}+w^{1/\sigma})\exp\left(\frac{-\alpha_0-\boldsymbol{\alpha}'\mathbf{x}_i}{\sigma}\right)\right\} \\
\Leftrightarrow\ &(t+w)^{1/\sigma} > t^{1/\sigma}+w^{1/\sigma}.
\end{aligned}
$$

Thus, it suffices to show the last inequality holds for all $0<\sigma<1$ and $w>0$. Consider the case of $w<t$, we have

$$\left(1+\frac{w}{t}\right)^{1/\sigma} > 1+\frac{w}{t} > 1+\left(\frac{w}{t}\right)^{1/\sigma}.$$

The above inequality holds since $w/t<1$ and $1<1/\sigma<\infty$. Then we obtain the desired result by multiplying $t^{1/\sigma}$ to both sides of the inequality. The case of $t<w$ can be proven in the a similar fashion.

This inequality implies that for any object with life time distribution following T_i. Its survival probability at time w is larger than its survival probability at time $t+w$ given it survived at time t. This can be interpreted as the aging property. For surviving the same amount of time, the early survival probability is larger than the late survival probability.

Problem 2.

Obtain 63 training samples from *Lung* in *compound.Cox* package.

(1) Is there any tie in the data? Explain it.

Solution (1).

Attached below is 63 training samples from *Lung* in *compound.Cox* package

0.231023	2.508251	2.937294	4.719472	4.846686	5.280528	7.557756	7.887789	9.438944	9.636964
10.82508	10.89109	11.25413	12.21122	13.86139	13.86139	14.12541	14.40818	14.68647	16.20462
16.46865	16.46865	16.89769	17.9538	18.21782	18.44884	18.64686	18.67987	19.83498	20.03300
20.06601	20.16502	20.33003	20.56106	20.62706	20.82508	21.51815	22.21122	22.73927	26.27063
26.33663	26.53465	26.83168	26.93069	26.99670	27.02970	27.12871	27.78878	27.79426	29.07591
29.63696	29.66997	30.9901	31.35314	32.14521	32.64026	35.93802	36.27063	36.69967	40.85809
41.32013	45.28053	49.27393							

For the tie at time ***13.86139***, both of these two events are censored. For the tie at time ***16.46865***, one event is death and another is censored. There are no two or more events of death at the same time. Thus, the no tie assumption is hold under this data set.

(2) Detail computation of the log-rank test for comparing two groups ($ERBB3 \geq 3$ versus $ERBB3 < 3$).

Solution (2).

We define $ERBB3 \geq 3$ and $ERBB3 < 3$ as $x = 1$ and $x = 0$, respectively. Let t_i be the event time and δ_i be the censoring indicator. We also define n_{i1} and n_{i0} as the number of $x = 1$ and $x = 0$ at-risk at time t_i, respectively. Now, one can derive the log-rank test statistic and its variance as

$$S = \sum_{i=1}^{n} \delta_i \left(x_i - \frac{n_{i1}}{n_{i1} + n_{i0}} \right) \quad \text{and} \quad \text{var}(S) = \sum_{i=1}^{n} \delta_i \frac{n_{i1} n_{i0}}{(n_{i1} + n_{i0})^2}.$$

The log-rank test for no effect on gene *ERBB*3 is based on $z^2 = S^2 / \text{var}(S)$. The *p*-value is computed as $\Pr(\chi^2_{\text{df}=1} > z^2)$, where $\chi^2_{\text{df}=1}$ is the chi-squared distribution with 1 degree of freedom. Detail computation of the log-rank test is summarized in Table 1. The log-rank test statistic is ***3.89044*** with *p*-value ***0.0486***. Thus, we reject the null hypothesis that there is no effect on gene *ERBB*3 at level $\alpha = 0.05$.

Table 1. Detail computation of the log-rank test statistic (display only $\delta_i = 1$).

t_i	x_i	n_{i1}	n_{i0}	$\frac{n_{i1}}{n_{i1}+n_{i0}}$	$x_i - \frac{n_{i1}}{n_{i1}+n_{i0}}$	$\frac{n_{i1}n_{i0}}{(n_{i1}+n_{i0})^2}$
0.2310231	1	43	20	0.68254	0.31746	0.21668
2.5082508	0	42	20	0.67742	-0.6774	0.21852
2.9372937	1	42	19	0.68852	0.31148	0.21446
4.7194719	1	41	19	0.68333	0.31667	0.21639
5.2805281	1	39	19	0.67241	0.32759	0.22027
7.8877888	1	37	19	0.66071	0.33929	0.22417
9.4389439	1	36	19	0.65455	0.34545	0.22612
10.891089	1	33	19	0.63462	0.36538	0.23188
11.254125	1	32	19	0.62745	0.37255	0.23376
14.125413	1	28	19	0.59574	0.40426	0.24083
14.686469	1	27	18	0.60000	0.40000	0.24000
16.468647	0	25	18	0.58140	-0.58140	0.24337
17.953795	0	23	17	0.57500	-0.57500	0.24438
18.646865	1	21	16	0.56757	0.43243	0.24543
20.033003	1	18	16	0.52941	0.47059	0.24913
20.066007	1	17	16	0.51515	0.48485	0.24977
20.330033	1	16	15	0.51613	0.48387	0.24974
26.930693	1	10	10	0.50000	0.50000	0.25000
26.996700	0	9	10	0.47368	-0.47370	0.24931
31.353135	1	3	7	0.30000	0.70000	0.21000
			Sum	11.7356	$S = 4.26436$	$\text{var}(S) = 4.67421$
			$z^2 = S^2 / \text{var}(S) = 3.89044$		(*p*-value = 0.0486)	

(3) Detail computation of the two Kaplan-Meier survival curves.

Solution (3).

Using the same notations as in (2), one can compute the Kaplan-Meier survival curves for two groups as

$$\hat{S}_j(t) = \prod_{t_i \le t, \delta_i = 1} \left(1 - \frac{1}{n_{ij}}\right), \qquad j = 1, 2.$$

Detail computation of two Kaplan-Meier survival curves are given in Tables 2 – 3. Figure 1 plots two curves and it shows that the grouping based on gene expression $ERBB3 \ge 3$ or $ERBB3 < 3$ can separate the two survival curves very well.

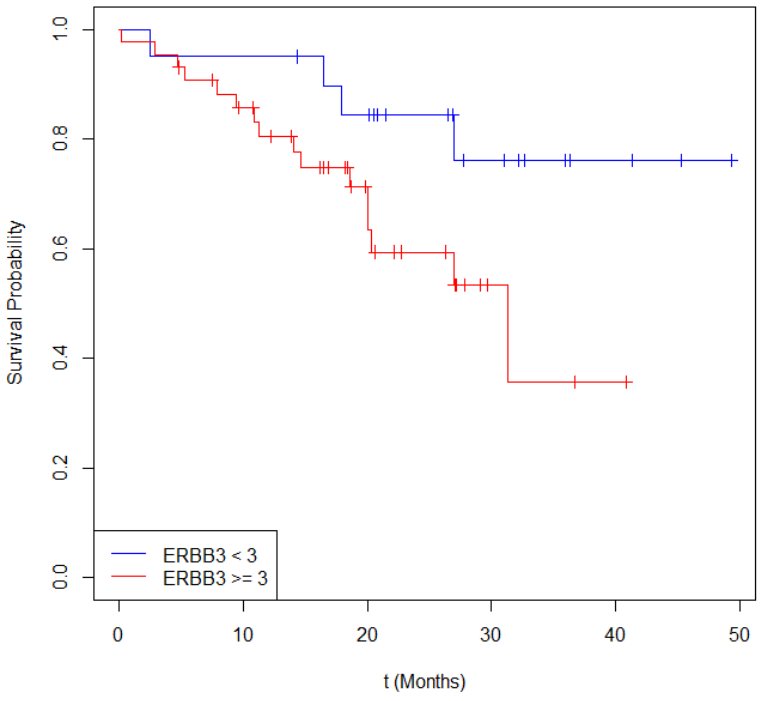

Figure 1. Two Kaplan-Meier survival curves based on gene expression $ERBB3 \ge 3$ or $ERBB3 < 3$.

Table 2. Detail computation of the Kaplan-Meier survival curve on the group $ERBB3 < 3$ (display only $\delta_i = 1$).

t_i	n_{i0}	$1-\frac{1}{n_{i0}}$	$\prod_{t_i \le t, \delta_i = 1}\left(1-\frac{1}{n_{i0}}\right)$
2.508251	20	0.950000	0.950000
16.468647	18	0.944444	0.897222
17.953795	17	0.941176	0.844444
26.996700	10	0.900000	0.760000

Table 3. Detail computation of the Kaplan-Meier survival curve on the group $ERBB3 \ge 3$ (display only $\delta_i = 1$).

t_i	n_{i1}	$1-\frac{1}{n_{i1}}$	$\prod_{t_i \le t, \delta_i = 1}\left(1-\frac{1}{n_{i1}}\right)$
0.231023	43	0.976744	0.976744
2.937294	42	0.976190	0.953488
4.719472	41	0.975610	0.930233
5.280528	39	0.974359	0.906380
7.887789	37	0.972973	0.881884
9.438944	36	0.972222	0.857387
10.89109	33	0.969697	0.831405
11.25413	32	0.968750	0.805424
14.12541	28	0.964286	0.776659
14.68647	27	0.962963	0.747894
18.64686	21	0.952381	0.712280
20.03300	18	0.944444	0.672709
20.06601	17	0.941176	0.633138
20.33003	16	0.937500	0.593566
26.93069	10	0.900000	0.534210
31.35314	3	0.666667	0.356140

(4) Check (2) and (3) by using the *survival* package.

Solution (4).

We first check the result in (2). The result produced by the *survival* package is attached below.

```
> survdiff(Surv(t.vec,d.vec) ~ ERBB3 < 3,data = Lung,subset = train)
Call:
survdiff(formula = Surv(t.vec, d.vec) ~ ERBB3 < 3, data = Lung,
    subset = train)

                N Observed Expected (O-E)^2/E (O-E)^2/V
ERBB3 < 3=FALSE 43       16    11.74      1.55      3.89
ERBB3 < 3=TRUE  20        4     8.26      2.20      3.89

 Chisq= 3.9  on 1 degrees of freedom, p= 0.0486
```

This result agrees with the result in (2).

Next, we check the result in (3). Similarly, the result produced by the *survival* package is attached below.

```
> res = survfit(Surv(t.vec,d.vec) ~ ERBB3 < 3,data = Lung,subset = train)
> summary(res)
Call: survfit(formula = Surv(t.vec, d.vec) ~ ERBB3 < 3, data = Lung,
    subset = train)

                ERBB3 < 3=FALSE
  time n.risk n.event survival std.err lower 95% CI upper 95% CI
 0.231     43       1    0.977  0.0230        0.933        1.000
 2.937     42       1    0.953  0.0321        0.893        1.000
 4.719     41       1    0.930  0.0388        0.857        1.000
 5.281     39       1    0.906  0.0446        0.823        0.998
 7.888     37       1    0.882  0.0497        0.790        0.985
 9.439     36       1    0.857  0.0540        0.758        0.970
10.891     33       1    0.831  0.0583        0.725        0.954
11.254     32       1    0.805  0.0620        0.693        0.937
14.125     28       1    0.777  0.0661        0.657        0.918
14.686     27       1    0.748  0.0696        0.623        0.898
18.647     21       1    0.712  0.0749        0.580        0.875
20.033     18       1    0.673  0.0805        0.532        0.850
20.066     17       1    0.633  0.0849        0.487        0.824
20.330     16       1    0.594  0.0884        0.443        0.795
26.931     10       1    0.534  0.0974        0.374        0.764
31.353      3       1    0.356  0.1592        0.148        0.856

                ERBB3 < 3=TRUE
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
 2.51     20       1    0.950  0.0487        0.859            1
16.47     18       1    0.897  0.0689        0.772            1
17.95     17       1    0.844  0.0826        0.697            1
27.00     10       1    0.760  0.1093        0.573            1
```

This result again agrees with the result in (3).

Figure 2 plots the Kaplan-Meier survival curves by using *survival* package. It again agrees with Figure 1.

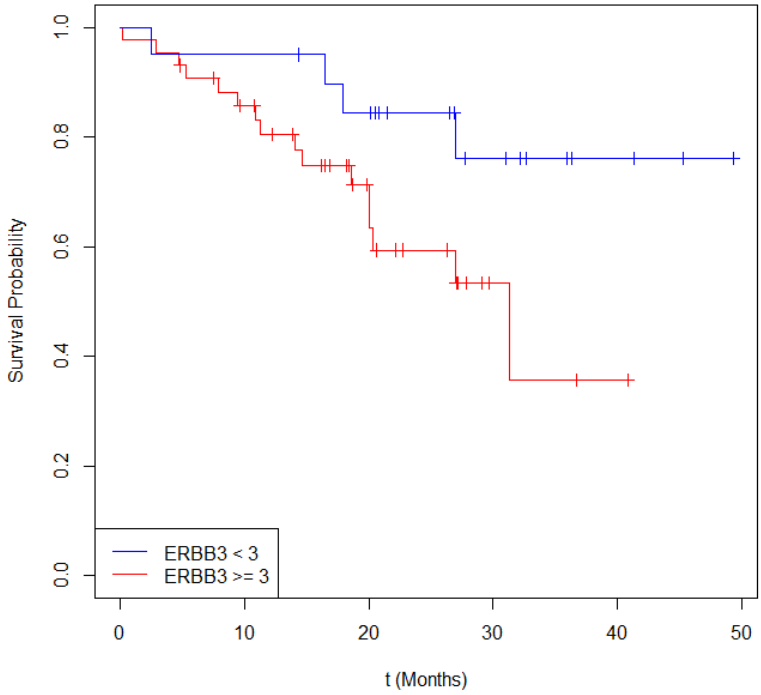

Figure 2. Two Kaplan-Meier survival curves based on gene expression $ERBB3 \geq 3$ or $ERBB3 < 3$ (produced by *survival* package).

(5) In *Lung*, which gene most significantly separates the two groups? Show the results of the log-rank test and Kaplan-Meier curves for the most significant gene.

Solution (5).

We first simply try the top 16 genes (both univariate selection and their proposed method) selected in Emura and Chen (2016). After investigation, we separate the two groups based on the gene expression $HCK \geq 3$ or $HCK < 3$. The result of the log-rank test is given below.

```
> survdiff(Surv(t.vec,d.vec) ~ HCK < 3,data = Lung,subset = train)
Call:
survdiff(formula = Surv(t.vec, d.vec) ~ HCK < 3, data = Lung,
    subset = train)

                N Observed Expected (O-E)^2/E (O-E)^2/V
HCK < 3=FALSE 32        7    11.66      1.86      4.84
HCK < 3=TRUE  31       13     8.34      2.60      4.84

 Chisq= 4.8  on 1 degrees of freedom, p= 0.0279
```

This choice produces a lower *p*-value = 0.0279 which means that it separates more significantly than gene *ERBB3* in the aspect of log-rank test.

However, in the aspect of Kaplan-Meier survival curves (Figure 3), the advantage seems not very clear.

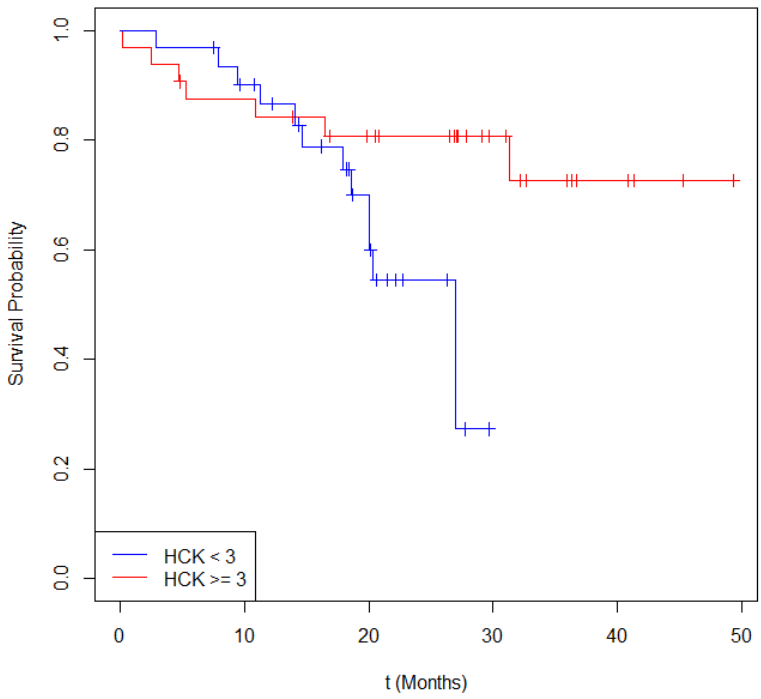

Figure 3. Two Kaplan-Meier survival curves based on gene expression $HCK \geq 3$ or $HCK < 3$ (produced by *survival* package).

There are total 97 genes in addition with different grouping method (e.g., $ERBB3 \geq 3$ versus $ERBB3 < 3$ or $ERBB3 \geq 2$ versus $ERBB3 < 2$). The number of samples in each group is also a very important issue to obtain reasonable results. Thus, to find the "most" significant gene, it seems that we require a more systematic method.

Appendix R codes for Problem 2

```
library(compound.Cox)
library(survival)
data(Lung)

# (1)
sort(Lung$t.vec[Lung$train == "TRUE"])
# 13.8613861
# 16.4686469

# (3)
t.vec2 = Lung$t.vec[Lung$train == "TRUE" & Lung$ERBB3 < 3]
d.vec2 = Lung$d.vec[Lung$train == "TRUE" & Lung$ERBB3 < 3]
data2 = data.frame(cbind(t.vec2,d.vec2))
order.data2 = data2[order(data2$t.vec2),]
n2 = c(20:1)
n2.d = n2[which(order.data2$d.vec2 == 1)]
S2.KM = cumprod(1-1/n2.d)
plot(c(0,order.data2$t.vec2[order.data2$d.vec2 == 1],max(t.vec2)),
        c(1,S2.KM,S2.KM[length(S2.KM)]),type = "s",ylim = c(0,1),
        xlim = c(0,max(t.vec2)),ylab = "Survival Probability",
        xlab = "t (Months)",col = "blue")
f2 = function(x) {

   for (i in length(data2$t.vec2[data2$d.vec2 == 1]):1) {

      if (x > order.data2$t.vec2[order.data2$d.vec2 == 1][i]) {return(S2.KM[i])}

   }

}
points(data2$t.vec2[data2$d.vec2 == 0],
            sapply(data2$t.vec2[data2$d.vec2 == 0],f2),
            pch = 3,col = "blue")

t.vec1 = Lung$t.vec[Lung$train == "TRUE" & Lung$ERBB3 >= 3]
```

```
d.vec1 = Lung$d.vec[Lung$train == "TRUE" & Lung$ERBB3 >= 3]
data1 = data.frame(cbind(t.vec1,d.vec1))
order.data1 = data1[order(data1$t.vec1),]
n1 = c(43:1)
n1.d = n1[which(order.data1$d.vec1 == 1)]
S1.KM = cumprod(1-1/n1.d)
lines(c(0,order.data1$t.vec1[order.data1$d.vec1 == 1],
        max(t.vec1)),c(1,S1.KM,S1.KM[length(S1.KM)]),type = "s",col = "red")
f1 = function(x) {

  for (i in length(data1$t.vec1[data1$d.vec1 == 1]):1) {

    if (x > order.data1$t.vec1[order.data1$d.vec1 == 1][i]) {return(S1.KM[i])}

  }

}
points(data1$t.vec1[data1$d.vec1 == 0],
         sapply(data1$t.vec1[data1$d.vec1 == 0],f1), pch = 3,col = "red")
legend("bottomleft",c("ERBB3 < 3","ERBB3 >= 3"),col = c("blue","red"),lty = 1)

# (4)
survdiff(Surv(t.vec,d.vec) ~ ERBB3 < 3,data = Lung,subset = train)
res = survfit(Surv(t.vec,d.vec) ~ ERBB3 < 3,data = Lung,subset = train)
summary(res)
plot(res,mark.time = TRUE, xaxs = "r",ylab = "Survival Probability",xlab = "t (Months)",col
= c("red","blue"))
legend("bottomleft",c("ERBB3 < 3","ERBB3 >= 3"),col = c("blue","red"),lty = 1)

# (5)

survdiff(Surv(t.vec,d.vec) ~ HCK < 3,data = Lung,subset = train)
res = survfit(Surv(t.vec,d.vec) ~ HCK < 3,data = Lung,subset = train)
summary(res)
plot(res,mark.time = TRUE, xaxs = "r",ylab = "Survival Probability",xlab = "t (Months)",col
= c("red","blue"))
legend("bottomleft",c("HCK < 3","HCK >= 3"),col = c("blue","red"),lty = 1)
```