https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_surv.html

HW1 1.6 Exercises -A. Applications

Identify the data types of the following cases:

1.1 Suppose that six rats have been exposed to carcinogens by injecting tumor cells into their foot-pads. The times to develop a tumor of a given size are observed. The investigator decides to terminate the experiment after 30 weeks. Rats A, B, and D develop tumors after 10, 15, and 25 weeks, respectively. Rats C and E do not develop by the end of the study. Rat F died accidentally without any tumors after 19 weeks of observation.(Source: Lee, E.T. (1992, page 2). Statistical Methods for Survival Data Analysis, 2nd ed., New York: John Wiley & Sons.)

- **Type I censoring**. The experiment terminated at a prespecified time. The endpoint is a fixed value and the number of observed failure times is a random variable which assumes a value in the set $\{0, 1, 2, .., n\}$

1.2 In Exercise 1.1, the investigator may decide to terminate the study after four of the six rats have developed tumors. Rats A, B ,and D develop tumors after 10, 15, and 25 weeks, respectively. Rat F died accidentally without any tumors after 19 weeks of observation. Rat E develops tumor after 35 weeks but Rats C does not develop by that time. How would the data set in Exercise 1.1 change? (Source: pages 2-3).

- **Type II censoring**. The observations terminate after the r^{th} failure occurs. The number of failure time T_r is a fixed value whereas the endpoint is a random observation. Hence we could wait possibly a very long time to observe the r failures.

1.3 Suppose that six patients with acute leukemia enter a clinical study during a total study period of one year. Suppose also that all six respond to treatment and achieve remission. Patients A, C, and E achieve remission at the beginning of the second, fourth, and ninth months and relapse after four, six, and three months, respectively. Patient B achieves remission at the beginning of the third month but is lost to follow-up four months later. Patients D and F achieve remission at the beginning of the fifth and tenth month, respectively, and are still in remission at the end of the study. Find out the remission times of the six patients. (Source: pages 3-4).

- **Right censoring**. The censoring occurred for dropping out ($t_A = 4, t_C = 6, t_E = 3$), losing to follow-up (remission times $t_B = 4+$ months), termination of study ($t_D = 8+, t_F = 3+$).

1.4 Survival/sacrifice experiments are designed to determine whether a suspected agent accelerates the time until tumor onset in experimental animals. For such studies, each animal is assigned to a prespecified dose of a suspected carcinogen, and examined at sacrifice or death, for the presence or absence of a tumor. Since a lung tumor is occult, the time until tumor onset is not directly observable. Instead, we observe only a time of sacrifice or death. (Source: Hoel, D.G. and Walburg, H.E., Jr. (1972). Statistical analysis of survival experiments. J. Natl. Cancer Inst., 49, 361-372.)

- **Case I Interval censored data: current status data**. The only available observed time is U, the censoring time. The endpoint of interest is T.

1.5 An annual survey on 196 girls recorded whether or not, at the time of the survey, sexual maturity had developed. Development was complete in some girls before the first survey, some girls were lost before the last survey and before development was complete, and some girls had not completed development at the last survey. (Source: Peto, R. (1973). Empirical survival curves for interval censored data. Appl. Statist., 22, 86-91.)

- **Interval censoring**. The time-to-event T is known only to occur within an interval. Such censoring occurs when the longitudinal study has periodic follow-up.

1.6 Woolson (1981) has reported survival data on 26 psychiatric inpatients admitted to the University of Iowa hospitals during the years 1935 – 1948. This sample is part of a larger study of psychiatric inpatients discussed by Tsuang and Woolson (1977). Data for each patient consists of age at first admission to the hospital, sex, number of years of follow-up (years from admission to death or censoring), and patient status at the followup time. The main goal is to compare the survival experience of these 26 patients to the standard mortality of residents of Iowa to determine if psychiatric patients tend to have shorter lifetimes. (Source: Klein, J.P. and Moeschberger, M.L. (1997, page 15). Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer.)

- **Left-truncated & right censored data**. The data include: T exact death time is observed (uncensored), T^- the main event of interest, death time, is left-truncated as the death patient are excluded from the study, T^+ death time is right-censored since the patient did not die during the study period.

1.7 The US Centers for Disease Control maintains a database of reported AIDS cases. We consider the 1,927 cases who were infected by contaminated blood transfusions and developed AIDS by November 1989. For our data, the earliest reported infection date was January 1975. For our analysis, we give a code of 0 for young children (ages 0 – 12) and 1 for older children and adults (ages 13 and up). We wish to test whether the induction periods for the two groups have the same latency distribution. (Source: Finkelstein, D.M., Moore, D.F., and Schoenfeld, D.A. (1993). A proportional hazards model for truncated AIDS data. Biometrics, 49, 731-740.)

- **Right-truncated**. Only individuals who have developed AIDS prior to the end of the study are included in the study. Infected individuals who have yet to develop AIDS are excluded from the sample; unknown to the investigator.

1.8 Leiderman et al. wanted to establish norms for infant development for a community in Kenya in order to make comparisons with known standards in the United States and the United Kingdom. The sample consisted of 65 children born between July 1 and December 31, 1969. Starting in January 1970, each child was tested monthly to see if he had learned to accomplish certain standard tasks. Here the variable of interest T would represent the time from birth to first learn to perform a particular task. Late entries occurred when it was found that, at the very first test, some children could already perform the task, whereas losses occurred when some infants were still unsuccessful by the end of the study. (Source: Leiderman, P.H., Babu, D., Kagia, J., Kraemer, H.C., and Leiderman, G.F. (1973). African infant precocity and some social influences during the first year. Nature, 242, 247-249.)

- **Doubly-censored data**. The recored values include: T^- age is left-censored as the child already knew the task when s/he was initially tested in the study, T exact age is observed u(ncensored), and T^+ age is right-censored since the child did not learn the task during the study period.

HW2

- 1.9 Show expression (1.8)

$$E(T) = \int_{t=0}^{\infty} (\int_{x=0}^{t} 1dx) f(t)dt = \int_{x=0}^{\infty} \left\{ \int_{t=x}^{\infty} f(t)dt \right\} dx = \int_{x=0}^{\infty} \{F(\infty) - F(x)\} dx$$
$$= \int_{x=0}^{\infty} \{1 - F(x)\} dx = \int_{0}^{\infty} S(x)dx = \int_{0}^{\infty} S(t)dt$$

- 1.10 Cerify expression (1.9) Mean residual life

For $\int_0^\infty tf(t)dt = E[T] = \int_0^\infty S(t)dt$

$$\begin{aligned}
mrl(u) &= E[T-u|T>u] = E[T|T>u] - E[u|T>u] = E[f(t|T>u)] - u \\
&= \int_u^\infty t\frac{f(t)}{S(u)}dt - u = \frac{1}{S(u)}\int_u^\infty tf(t)dt - u \\
&= \frac{1}{S(u)}\left\{ \int_0^\infty tf(t)dt - \int_0^u tf(t)dt \right\} - u \\
&= \frac{1}{S(u)}\left\{ \int_0^\infty S(t)dt - \left[tF(t)|_0^u - \int_0^u F(t)dt \right] \right\} - u \\
&= \frac{1}{S(u)}\left\{ \int_0^\infty S(t)dt - \left[uF(u) - \int_0^u (1-S(t))dt \right] \right\} - u \\
&= \frac{1}{S(u)}\left\{ \int_0^\infty S(t)dt - \left[u(1-S(u)) - \int_0^u dt + \int_0^u S(t)dt \right] \right\} - u \\
&= \frac{1}{S(u)}\left\{ \int_0^\infty S(t)dt + uS(u) - \int_0^u S(t)dt \right\} - u \\
&= \frac{1}{S(u)}\left\{ \int_0^\infty S(t)dt - \int_0^u S(t)dt \right\} = \frac{\int_u^\infty S(t)dt}{S(u)}
\end{aligned}$$ ■

Proof: $f(t|T>u) = \frac{f(t)}{S(u)}$

$$f(t|T>u) = \lim_{\Delta t \to 0^+} \frac{P(t<T<t+\Delta t|T>u)}{\Delta t} = \lim_{\Delta t \to 0^+} \frac{P(t<T<t+\Delta t \cap T>u)}{\Delta t \cdot P(T>u)}$$

$$= \frac{1}{S(u)} \lim_{\Delta t \to 0^+} \frac{P(t<T<t+\Delta t)}{\Delta t} = \frac{1}{S(u)} \lim_{\Delta t \to 0^+} \frac{F(t+\Delta t) - F(t)}{\Delta t} = \begin{cases} \frac{f(t)}{S(u)} & t>u \\ 0 & o.w. \end{cases}$$ □

- 1.11 Derive expression (1.11)

$$f(t_{(1)},\cdots,t_{(n)}) = \begin{cases} n!f(t_{(1)})\cdots f(t_{(n)}) & 0 < t_{(1)} < \cdots < t_{(n)} \\ 0 & o.w. \end{cases}$$

$\frac{[1-F(t_{(n-1)})]^1}{1!} = \int_{n-1}^{\infty} \left\{ f(t_{(n)}) \right\} dt_{(n)}$

$\frac{[1-F(t_{(n-2)})]^2}{2!} = \int_{n-2}^{\infty} \left\{ f(t_{(n-1)}) \frac{[1-F(t_{(n-1)})]^1}{1!} \right\} dt_{(n-1)}$

$\frac{[1-F(t_{(n-3)})]^3}{3!} = \int_{n-3}^{\infty} \left\{ f(t_{(n-2)}) \frac{[1-F(t_{(n-2)})]^2}{2!} \right\} dt_{(n-2)}$

...

$\frac{[1-F(t_{(r+1)})]^{n-r-1}}{(n-r-1)!} = \int_{r+1}^{\infty} \left\{ f(t_{(r+2)}) \frac{[1-F(t_{(r+2)})]^{n-r-2}}{(n-r-2)!} \right\} dt_{(r+2)}$

$\frac{[1-F(t_{(r)})]^{n-r}}{(n-r)!} = \int_{r}^{\infty} \left\{ f(t_{(r+1)}) \frac{[1-F(t_{(r+1)})]^{n-r-1}}{(n-r-1)!} \right\} dt_{(r+1)}$

Integrate out the $n-r$ order statistics

$$\int_r^{\infty} \int_{r+1}^{\infty} \cdots \int_{n-2}^{\infty} \int_{n-1}^{\infty} \left\{ f(t_{(n)}) f(t_{(n-1)}) \cdots f(t_{(r+2)}) f(t_{(r+1)}) \right\} dt_{(n)} dt_{(n-1)} \cdots dt_{(r+2)} dt_{(r+1)} = \frac{[1-F(t_{(r)})]^{n-r}}{(n-r)!}$$

Hence, the Type II censoring likelihood function is:

$$L = n!f(t_{(1)})\cdots f(t_{(r)}) \frac{[1-F(t_{(r)})]^{n-r}}{(n-r)!} = \frac{n!}{(n-r)!} f(t_{(1)})\cdots f(t_{(r)}) [S(t_{(r)})]^{n-r}$$

- Proof:$S(y) = \frac{mrl(0)}{mrl(y)} e^{-\int_0^y \frac{1}{mrl(t)} dt}$

$$-\frac{1}{mrl(t)} = \frac{d}{dt} \log \int_t^{\infty} S(u)du$$

$$\implies -\int_0^t \frac{1}{mrl(u)} du + c = \log \int_t^{\infty} S(u)du$$

$$\implies c = \log \int_0^{\infty} S(u)du = \log\left(mrl(0)\right)$$

$$\implies \log \frac{\int_t^{\infty} S(u)du}{mrl(0)} = -\int_0^t \frac{1}{mrl(u)} du$$

$$\implies \frac{\int_t^{\infty} S(u)du}{mrl(0)} = e^{-\int_0^t \frac{1}{mrl(u)} du}$$

$$\implies \int_t^{\infty} S(u)du = mrl(t)S(t) = mrl(0) \cdot e^{-\int_0^t \frac{1}{mrl(u)} du}$$

$$\implies S(t) = \frac{mrl(0)}{mrl(t)} \cdot e^{-\int_0^t \frac{1}{mrl(u)} du}$$

- Extra Exercise 1

```
SC_Estimator <- function(time,status){
df <- data.frame(time,status)
df.distinct<- df %>% mutate(status=1-status)%>% group_by_all %>% count %>% as.data.frame()
names(df.distinct) <- c("y_i","status","event")
df.distinct<- df.distinct %>% mutate(status=1-status)
df.distinct<- df.distinct %>% mutate(d_i= ifelse(status==1,event,0))%>%
  mutate(c_i= ifelse(status==0,event,0))%>%select(-2,-3)%>%
  add_row(y_i=0,d_i=0,c_i=0, .before = 1)
y_i <- df.distinct[,1]
d_i <- df.distinct[,2]
c_i <- df.distinct[,3]
n <- length(status)
k <- nrow(df.distinct)
n_i <- N_i <- n
for (i in 2:k) {
n_i[i] <- n_i[i-1]-d_i[i-1]-c_i[i-1]
N_i[i] <- N_i[i-1]-d_i[i]-c_i[i]
}
p_i <- (n_i-d_i)/n_i  # Probability of surviving through I_i | alive at beginning I_i
s_i <- 1              # K-M estimator of the survivor function
for (i in 1:k) {s_i[i]<- prod(p_i[1:i])}

sc_i <-  1            # Initial self-consistency estimator
cum<-0     #  (1-d_i[1])/sc_i
sc_i[1] <-(N_i[1])/(n-sum(cum))
for (i in 2:k) {    # The rest of the self-consistency estimator
  cum[i] <- (c_i[i])/(sc_i[i-1])
  sc_i[i]<- (N_i[i])/(n-sum(cum)) }
if (d_i[k]==0){
cum[k] <- (1-c_i[k])/(sc_i[k-1]) # force the largest observed time to be uncensored
sc_i[k]<- (n_i[k])/(n-sum(cum)) }# and calculate separately

return(cbind(y_i,d_i,c_i,n_i,N_i,p_i,s_i,sc_i))
}
```

```
aml1 <- read_excel("~/qushen26/stat2019_website/static/stat578/aml1.xls")
pander(SC_Estimator(aml1$weeks,aml1$status))
```

y_i	d_i	c_i	n_i	N_i	p_i	s_i	sc_i
0	0	0	11	11	1	1	1
9	1	0	11	10	0.9091	0.9091	0.9091
13	1	0	10	9	0.9	0.8182	0.8182
13	0	1	9	8	1	0.8182	0.8182
18	1	0	8	7	0.875	0.7159	0.7159
23	1	0	7	6	0.8571	0.6136	0.6136
28	0	1	6	5	1	0.6136	0.6136
31	1	0	5	4	0.8	0.4909	0.4909
34	1	0	4	3	0.75	0.3682	0.3682
45	0	1	3	2	1	0.3682	0.3682
48	1	0	2	1	0.5	0.1841	0.1841
161	0	1	1	0	1	0.1841	0.1841

```
time <- c(1, 1 , 2, 4, 4, 4, 6, 9)
status <- c(1, 0, 1, 1, 1, 0, 1, 1)
pander(SC_Estimator(time,status))
```

y_i	d_i	c_i	n_i	N_i	p_i	s_i	sc_i
0	0	0	8	8	1	1	1
1	1	0	8	7	0.875	0.875	0.875
1	0	1	7	6	1	0.875	0.875
2	1	0	6	5	0.8333	0.7292	0.7292
4	2	0	5	3	0.6	0.4375	0.4375
4	0	1	3	2	1	0.4375	0.4375
6	1	0	2	1	0.5	0.2188	0.2188
9	1	0	1	0	0	0	0

```
diabetes <- read_excel("~/qushen26/stat2019_website/static/stat578/diabetes.xls")
time <- diabetes[diabetes$diab==1,]$lzeit
status <- diabetes[diabetes$diab==1,]$tod
pander(SC_Estimator(time,status))
```

y_i	d_i	c_i	n_i	N_i	p_i	s_i	sc_i
0	0	0	40	40	1	1	1
20	1	0	40	39	0.975	0.975	0.975
35	1	0	39	38	0.9744	0.95	0.95
37	1	0	38	37	0.9737	0.925	0.925
61	1	0	37	36	0.973	0.9	0.9
82	1	0	36	35	0.9722	0.875	0.875
93	1	0	35	34	0.9714	0.85	0.85
130	1	0	34	33	0.9706	0.825	0.825
180	1	0	33	32	0.9697	0.8	0.8
232	0	1	32	31	1	0.8	0.8
242	0	1	31	30	1	0.8	0.8
244	0	1	30	29	1	0.8	0.8
264	1	0	29	28	0.9655	0.7724	0.7724
321	1	0	28	27	0.9643	0.7448	0.7448
435	1	0	27	26	0.963	0.7172	0.7172
436	1	0	26	25	0.9615	0.6897	0.6897
487	1	0	25	24	0.96	0.6621	0.6621
538	1	0	24	23	0.9583	0.6345	0.6345
547	1	0	23	22	0.9565	0.6069	0.6069
595	1	0	22	21	0.9545	0.5793	0.5793
630	1	0	21	20	0.9524	0.5517	0.5517
636	1	0	20	19	0.95	0.5241	0.5241
796	1	0	19	18	0.9474	0.4966	0.4966
819	1	0	18	17	0.9444	0.469	0.469
921	0	1	17	16	1	0.469	0.469
924	1	0	16	15	0.9375	0.4397	0.4397
950	0	1	15	14	1	0.4397	0.4397
1006	1	0	14	13	0.9286	0.4083	0.4083
1038	1	0	13	12	0.9231	0.3768	0.3768
1039	1	0	12	11	0.9167	0.3454	0.3454

y_i	d_i	c_i	n_i	N_i	p_i	s_i	sc_i
1052	1	0	11	10	0.9091	0.314	0.314
1179	1	0	10	9	0.9	0.2826	0.2826
1379	0	1	9	8	1	0.2826	0.2826
1504	1	0	8	7	0.875	0.2473	0.2473
1663	0	1	7	6	1	0.2473	0.2473
1786	0	1	6	5	1	0.2473	0.2473
1966	0	1	5	4	1	0.2473	0.2473
2071	1	0	4	3	0.75	0.1855	0.1855
2124	1	0	3	2	0.6667	0.1237	0.1237
2208	1	0	2	1	0.5	0.06183	0.06183
2850	1	0	1	0	0	0	0

The result shows that the self-consistency estimator coincides with the K-M estimator.

- Extra Exercise 2

For current status data, the self-consistency estimator of the survivor function is

$$\widehat{SC(t)} = \frac{1}{n}\left[\sum_{i=1}^{n} 1\cdot\mathbb{1}(u_{(i)} > t, \delta_{(i)} = 0) + \sum_{i=1}^{n}\frac{\widehat{SC(t)} - \widehat{SC(u_{(i)})}}{1-\widehat{SC(u_{(i)})}}\mathbb{1}(u_{(i)} > t, \delta_{(i)} = 1)\right.$$
$$\left. + \sum_{i=1}^{n}\frac{\widehat{SC(t)}}{\widehat{SC(u_{(i)})}}\mathbb{1}(u_{(i)} \le t, \delta_{(i)} = 0) + \sum_{i=1}^{n} 0\cdot\mathbb{1}(u_{(i)} \le t, \delta_{(i)} = 1)\right]$$

where $\frac{\widehat{SC(t)}}{\widehat{SC(u_{(i)})}}$ estimates the conditional probability of surviving beyond t given alive at u_i, $\frac{\widehat{SC(t)}-\widehat{SC(u_{(i)})}}{1-\widehat{SC(u_{(i)})}}$ estimates the conditional probability of surviving inside t given dead at $u_{(i)}$.

I don't get a converge code yet. Quote Klein J.P. and Moeschberger M.L. (2003, p.143)

In some applications the data may be interval-censored. Here the only information we have for each individual is that their event time falls in an interval $(L_i, R_i], i, 1, ..., n$, but the exact time is unknown. An estimate of the survival function can be found by a modification of above iterative procedure as proposed by Turnbull (1976).

Let $0 = \tau_0 < \tau_1 < .. < \tau_m$ be a grid of time points which includes all the points L_i, R_i for $i = 1, ..., n$. For the i^{th} observation, define a weight α_{ij} to be 1 if the interval $(\tau_{j-1}, \tau_j]$ is contained in the interval $(L_i, R_i]$, and 0 otherwise. Note that $\alpha_{ij}$is an indicator of whether the event which occurs in the interval $(L_i, R_i]$ could have occurred atτ_j. An initial guess at $S(\tau_j)$ is made. The algorithm is as follows:

Step 1: Compute the probability of an event's occurring at time τ_j, $p_j = S(\tau_{j-1}) - S(\tau_j), j = 1, ..., m$

Step 2: Estimate the number of events which occurred at τ_i by $d_i = \sum_{i=1}^{n}\frac{\alpha_{ij}p_j}{\sum_k \alpha_{ik}p_k}$

Note the denominator is the total probability assigned to possible event times in the interval $(L_i, R_i]$.

Step 3: Compute the estimated number at risk at time τ_i by $Y_i = \sum_{k=j}^{m} d_k$.

Step 4: Compute the updated Product-Limit estimator using the pseudo data found in steps 2 and 3. If the updated estimate of S is close to the old version of S for all τ_i's, stop the iterative process, otherwise repeat steps 1–3 using the updated estimate of S.

- Extra Exercise 3

Proof $Var(\hat{H}(t)) = \sum_{y_{(i)} \le t} \frac{d_i}{(n_i - d_i)n_i}$; $Var(\hat{S}(t)) \approx S^2(t) \left[\sum_{i=1}^{k} \frac{d_i}{(n_i - d_i)n_i}\right]$

Let $f(x) = \log(\hat{p}_i)$, $f(\mu) = \log(p_i)$, $f'(\mu) = \frac{1}{p_i}$

$Var(\log(\hat{p}_i)) = \frac{1}{p_i^2} Var(\hat{p}_i)$

$$Var(\hat{H}(t)) = Var(\log(\hat{S}(t))) = Var(\log(\prod_{i=1}^{k} \hat{p}_i)) = \sum_{i=1}^{k} Var(\log \hat{p}_i) = \sum_{i=1}^{k} \frac{1}{p_i^2} Var(\hat{p}_i) = \sum_{i=1}^{k} \frac{1}{p_i^2} \frac{n_i p_i (1 - p_i)}{n_i^2}$$

$$= \sum_{i=1}^{k} \frac{q_i}{p_i n_i} = \sum_{i=1}^{k} \frac{\frac{d_i}{n_i}}{(1 - \frac{d_i}{n_i}) n_i} = \sum_{i=1}^{k} \frac{d_i}{(n_i - d_i) n_i}$$

Let $f(x) = \log(\hat{S}(t)) = f(\mu) + f'(\mu)(x - \mu) = \log(S(t)) + \frac{1}{S(t)}(\hat{S}(t) - S(t))$

$Var(\log(\hat{S}(t))) = \frac{1}{S^2(t)} Var(\hat{S}(t))$

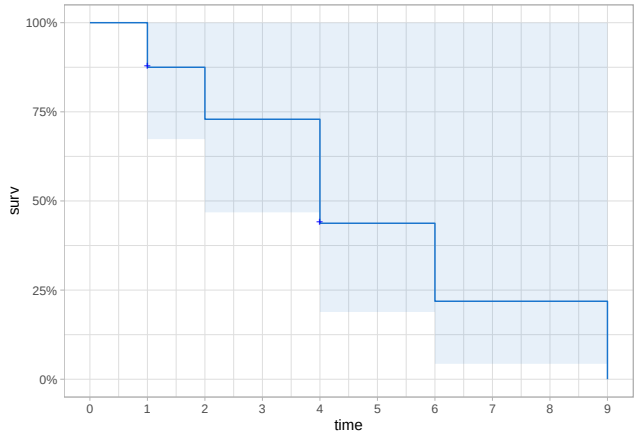
$Var(\hat{S}(t)) = S^2(t) Var(\log(\hat{S}(t))) = S^2(t) \sum_{i=1}^{k} \frac{d_i}{(n_i - d_i) n_i}$

HW3 2.1, 2.3, 2.4, 2.5, and 2.7.

2.1 Use only hand-held calculator. No need for computer.

(a) Calculate the following table and sketch the Kaplan-Meier (K-M) estimate of survival for the data set y: 1, 1 +, 2, 4, 4, 4 +, 6, 9. ("+" denotes censored observation.) The s.e.($\hat{S}(t)$) is computed using Greenwood's formula (2.3) for the estimated (asymptotic) variance of the K-M curve at time t.

y_i	d	n	p	s	se
0	0	8	1	1	0
1	1	8	0.875	0.875	0.1169
1	0	7	1	0.875	0.1169
2	1	6	0.8333	0.7292	0.165
4	2	5	0.6	0.4375	0.1879
4	0	3	1	0.4375	0.1879
6	1	2	0.5	0.2188	0.181
9	1	1	0	0	NA



(b) Calculate a 95% confidence interval for $S(t)$ at $t = 3$. Use the **default interval** given in Remark 4, expression (2.16). Is is necessary to use the C.I. in expression (2.18)? If yes, use it to report a 95% C.I.

$\widehat{Var}(\hat{H}(3)) = \sum_{y_{(i)} \leq 3} \frac{d_i}{n_i(n_i - s_i)} = 0.0512$

$\exp(\log(\hat{S}(3)) \pm 1.96 se(\hat{H}(3))) = \exp(\log(0.7292) \pm 1.96 * \sqrt{0.0512} = (0.468, 1.1361)$

The interval exceed (0,1). Use the C.I. in expression (2.18) by *log-log* transformation

$W = \log(-\log(\hat{S}(3)))$ = -1.1525

$\widehat{Var}(W) = \frac{1}{(\log(\hat{S}(3))^2} \widehat{Var}(\hat{H}(3)) = 0.5131$

$\hat{S}(3)^{\exp(\pm 1.96 se(W))} = (0.2764, 0.9254)$

(c) Compute the estimated hazard (2.5) $\tilde{h}(t_i)$ at$t_i = 2$. Then compute a 95% C.I. for $H(t)$ at $t = 3$ using the Nelson-Aalen estimate (2.9).

```
tilde.h.t3 <- d[4]/n[4]  # Nelson-Aalen
tilde.H.t3 <- sum(d[1:4]/n[1:4])
Var.tilde.H.t3 <- sum(d[1:4]/n[1:4]^2)
CI.tilde.H.t3 <- c(exp(log(s[4])-qnorm(0.975)*sqrt(Var.tilde.H.t3)),exp(log(s[4])+qnorm(0.975)*sqrt(Var
Var.tilde.W <- Var.tilde.H.t3/(log(s[4])^2)
CI.tilde.H.t3.log <- c(s[4]^(exp(qnorm(0.975)*sqrt(Var.tilde.W))),s[4]^(exp(-qnorm(0.975)*sqrt(Var.tild
```

At $t_3 = 2$, $\tilde{h}(t_3) = d_3/n_3 = 0.1667$

$\tilde{H}(3) = \sum_{y_{(i)} \le 3} \frac{d_i}{n_i} = 0.2917$

$\widehat{Var}[\tilde{H}(3)] = \sum_{y_{(i)} \le 3} \frac{d_i}{n_i^2} = 0.0434$

$\exp(\log(\hat{S}(3)) \pm 1.96 se(\tilde{H}(3))) = (0.4847, 1.0969)$

The interval exceed (0,1). Using *log-log* transformation,

$\widehat{Var}(W) = \frac{1}{(\log(\hat{S}(3))^2} \widehat{Var}(\tilde{H}(3)) = 0.4351$

$\hat{S}(3)^{\exp(\pm 1.96 se(W))} = (0.3164, 0.9169)$

(d) Provide a point and 95% C.I. estimate of the median survival time. See page 34.

$\widehat{Var}[\hat{t}_p] = \frac{\widehat{Var}[\hat{S}(\hat{t}_p)]}{(\hat{f}(\hat{t}_p))^2}$, where $\widehat{Var}[\hat{S}(\hat{t}_p)]$ is Greenwood's formula for the estimate of the variance of the K-M estimator, and $\hat{f}(\hat{t}_p)$ is the estimated probability density at $\hat{t}_p$.

$\hat{f}(\hat{t}_p) = \frac{\hat{S}(\hat{u}_p) - \hat{S}(\hat{l}_p)}{\hat{l}_p - \hat{u}_p}$, where $\hat{u}_p = \max\{t_i | \hat{S}(t_i) \ge 1 - p + \epsilon\}$, $\hat{l}_p = \min\{t_i | \hat{S}(t_i) \le 1 - p - \epsilon\}$, for $i = 1, .., r \le n$ with r being the number of distinct death times, and take $\epsilon = 0.05$.

The median $\hat{t}_{0.5} = 4$.

$\hat{u}_{0.5} = \max\{t_i | \hat{S}(t_i) \ge 0.55\} = 2$

$\hat{l}_{0.5} = \min\{t_i | \hat{S}(t_i) \le 0.45\} = 4$

$se(4) = \frac{se(\hat{S}(4))}{\frac{\hat{S}(2) - \hat{S}(4)}{4-2}} = 1.2887$

$4 \pm 1.96 se(4) = (1.4742, 6.5258)$

2.3 Use S or R for this exercise.

In this study the survival times (in days) of 66 patients after a particular operation were observed. The data frame diabetes contains for each patient the following variables: Variable Key sex gender (m=0, f=1) diab diabetic (1=yes, 0=no) alter age in years altgr age group in years = 0 if age 64, or 1 if age > 64 lzeit survival times in days (number of days to death) after operation tod 0 = censored, 1 = uncensored (dead)

(a) Following the S code on page 35 of the text, obtain a summary of the K-M survival curve for the diabetic group only. survfit is the main function.

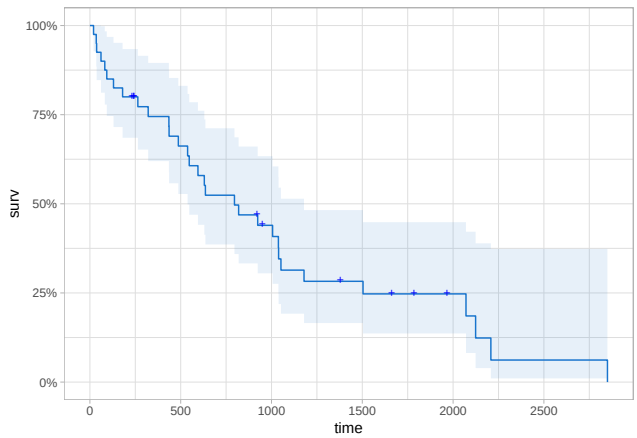
```
## Call: survfit(formula = Surv(lzeit, tod) ~ 1, data = diabetes1, type = "kaplan-meier")
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##    20     40       1   0.9750  0.0247       0.9278        1.000
##    35     39       1   0.9500  0.0345       0.8848        1.000
```

```
##    37     38       1   0.9250  0.0416       0.8469        1.000
##    61     37       1   0.9000  0.0474       0.8117        0.998
##    82     36       1   0.8750  0.0523       0.7783        0.984
##    93     35       1   0.8500  0.0565       0.7462        0.968
##   130     34       1   0.8250  0.0601       0.7153        0.952
##   180     33       1   0.8000  0.0632       0.6852        0.934
##   264     29       1   0.7724  0.0668       0.6520        0.915
##   321     28       1   0.7448  0.0699       0.6197        0.895
##   435     27       1   0.7172  0.0725       0.5883        0.874
##   436     26       1   0.6897  0.0748       0.5576        0.853
##   487     25       1   0.6621  0.0767       0.5275        0.831
##   538     24       1   0.6345  0.0783       0.4981        0.808
##   547     23       1   0.6069  0.0796       0.4693        0.785
##   595     22       1   0.5793  0.0807       0.4410        0.761
##   630     21       1   0.5517  0.0814       0.4132        0.737
##   636     20       1   0.5241  0.0819       0.3859        0.712
##   796     19       1   0.4966  0.0821       0.3591        0.687
##   819     18       1   0.4690  0.0820       0.3329        0.661
##   924     16       1   0.4397  0.0820       0.3051        0.634
##  1006     14       1   0.4083  0.0819       0.2755        0.605
##  1038     13       1   0.3768  0.0814       0.2468        0.575
##  1039     12       1   0.3454  0.0805       0.2188        0.545
##  1052     11       1   0.3140  0.0790       0.1918        0.514
##  1179     10       1   0.2826  0.0771       0.1656        0.482
##  1504      8       1   0.2473  0.0751       0.1363        0.449
##  2071      4       1   0.1855  0.0777       0.0816        0.422
##  2124      3       1   0.1237  0.0723       0.0393        0.389
##  2208      2       1   0.0618  0.0567       0.0102        0.374
##  2850      1       1   0.0000     NaN           NA           NA
```

(b) Report the mean and median survival times.

```
## Call: survfit(formula = Surv(lzeit, tod) ~ 1, data = diabetes1, type = "kaplan-meier")
##
##          n     events      *rmean *se(rmean)     median    0.95LCL    0.95UCL
##         40         31       1015        143        796        538       1179
##     * restricted mean with upper limit =  2850
```

(c) Plot the K-M curve for this group.



(d) Use the function hazard.km (page 38) to give a summary of the various estimates of hazard and cumulative hazard.

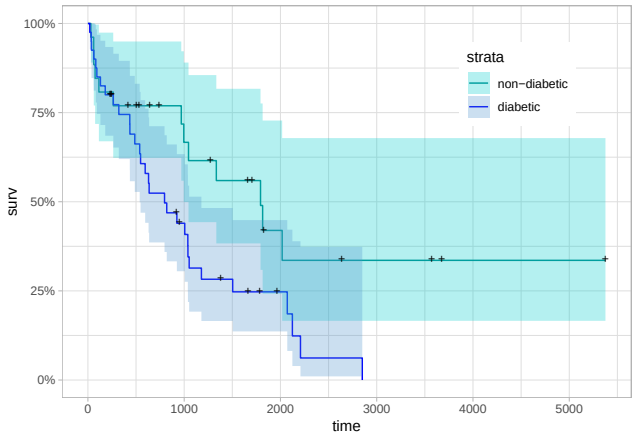
time	ni	di	hihat	hitilde	Hhat	se.Hhat	Htilde	se.Htilde
20	40	1	0.0017	0.025	0.0253	0.0253	0.025	0.025
35	39	1	0.0128	0.0256	0.0513	0.0363	0.0506	0.0358
37	38	1	0.0011	0.0263	0.078	0.045	0.077	0.0444
61	37	1	0.0013	0.027	0.1054	0.0527	0.104	0.052
82	36	1	0.0025	0.0278	0.1335	0.0598	0.1318	0.059
93	35	1	0.0008	0.0286	0.1625	0.0664	0.1603	0.0655
130	34	1	0.0006	0.0294	0.1924	0.0728	0.1897	0.0718
180	33	1	0.0004	0.0303	0.2231	0.0791	0.22	0.078
264	29	1	0.0006	0.0345	0.2582	0.0865	0.2545	0.0852
321	28	1	0.0003	0.0357	0.2946	0.0938	0.2902	0.0924
435	27	1	0.037	0.037	0.3323	0.1011	0.3273	0.0996
436	26	1	0.0008	0.0385	0.3716	0.1085	0.3657	0.1067
487	25	1	0.0008	0.04	0.4124	0.1159	0.4057	0.114
538	24	1	0.0046	0.0417	0.4549	0.1235	0.4474	0.1214
547	23	1	0.0009	0.0435	0.4994	0.1312	0.4909	0.1289
595	22	1	0.0013	0.0455	0.5459	0.1392	0.5363	0.1367
630	21	1	0.0079	0.0476	0.5947	0.1475	0.584	0.1447
636	20	1	0.0003	0.05	0.646	0.1562	0.634	0.1531
796	19	1	0.0023	0.0526	0.7001	0.1653	0.6866	0.1619
819	18	1	0.0005	0.0556	0.7572	0.1749	0.7421	0.1712
924	16	1	0.0008	0.0625	0.8218	0.1864	0.8046	0.1823
1006	14	1	0.0022	0.0714	0.8959	0.2006	0.8761	0.1957
1038	13	1	0.0769	0.0769	0.9759	0.216	0.953	0.2103
1039	12	1	0.0064	0.0833	1.063	0.2329	1.036	0.2262
1052	11	1	0.0007	0.0909	1.158	0.2517	1.127	0.2438
1179	10	1	0.0003	0.1	1.264	0.2728	1.227	0.2635
1504	8	1	0.0002	0.125	1.397	0.3038	1.352	0.2917
2071	4	1	0.0047	0.25	1.685	0.4191	1.602	0.3841
2124	3	1	0.004	0.3333	2.09	0.5851	1.936	0.5086
2208	2	1	0.0008	0.5	2.783	0.9178	2.436	0.7132
2850	1	1	NA	1	Inf	NA	3.436	1.228

(e) Use the function quantile.km (page 38) to provide point and 95% confidence interval estimates for the .25th and .80th quantiles.

	qp	se.S.qp	f.qp	se.qp	LCL	UCL
25th	321	0.0699	0.0004	158	11.3	630.7
80th	2071	0.0777	0.0002	462	1165	2977

2.4 We continue with the diabetes data.

(a) Plot the K-M curves for the data of the diabetic group and the nondiabetic. Comment briefly! Be sure to give a legend so we know which line corresponds to which group. See page 39.



The non-diabetic group has longer survival time than diabetic group.

(b) Is there a statistically significant difference in survival between the two groups - diabetic and nondiabetic? What weaknesses (shortcomings) does this global analysis have? See page 46 for example.

```
## Call: survfit(formula = Surv(lzeit, tod) ~ diab, data = diabetes, type = "kaplan-meier")
##
##         n events *rmean *se(rmean) median 0.95LCL 0.95UCL
## diab=0 26     13   2055        363   1793     996      NA
## diab=1 40     31   1015        143    796     538    1179
##     * restricted mean with upper limit =  4113
```

```
## Call:
## survdiff(formula = Surv(lzeit, tod) ~ diab, data = diabetes)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## diab=0 26       13     20.6      2.83      5.48
## diab=1 40       31     23.4      2.50      5.48
##
##  Chisq= 5.5  on 1 degrees of freedom, p= 0.02
```

There is significant difference in survival between the two groups (the p-value is 0.02/2=0.01). The mean and median of survival time in non-diabetic group is longer than diabetic group.

(c) Stratifying on known prognostic factor(s) can improve an analysis.

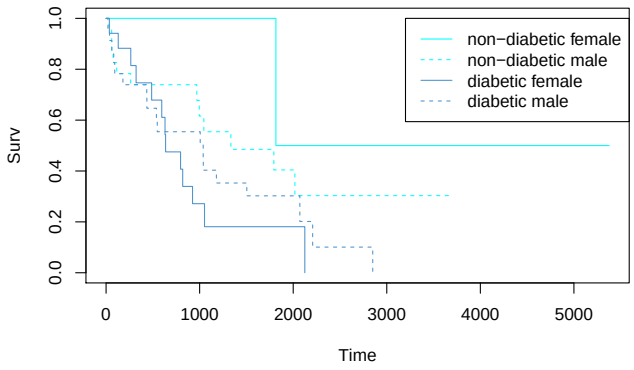
i. Stratify by sex. Judge now the difference in survival between the diabetic and nondiabetic groups. Tips:

```
##             diabetes.diab
## diabetes.sex  0  1
##            0 23 23
##            1  3 17
```

```
## Call:
## survdiff(formula = Surv(lzeit, tod) ~ diab + strata(sex), data = diabetes)
##
```

```
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## diab=0 26       13       20      2.45      4.85
## diab=1 40       31       24      2.05      4.85
##
##  Chisq= 4.9  on 1 degrees of freedom, p= 0.03
```

```
## Call: survfit(formula = Surv(lzeit, tod) ~ diab + strata(sex), data = diabetes)
##
##                             n events *rmean *se(rmean) median 0.95LCL 0.95UCL
## diab=0, strata(sex)=sex=0 23     12   1642        292   1334     996      NA
## diab=0, strata(sex)=sex=1  3      1   2539        512   1815    1815      NA
## diab=1, strata(sex)=sex=0 23     18   1120        206   1038     436      NA
## diab=1, strata(sex)=sex=1 17     13    861        173    636     487      NA
##     * restricted mean with upper limit =  3262
```



There is significant difference in survival between the two groups (the p-value is 0.03/2=0.015). The survival time in non-diabetic group is longer than diabetic group. However, taking into account the variation due to sex, the difference between two groups is small in male, while it is large in female.

It should be noted that the data are unbalanced. There are only three non-diabetic female. That may explain why diabetic female has less survival time than diabetic male.

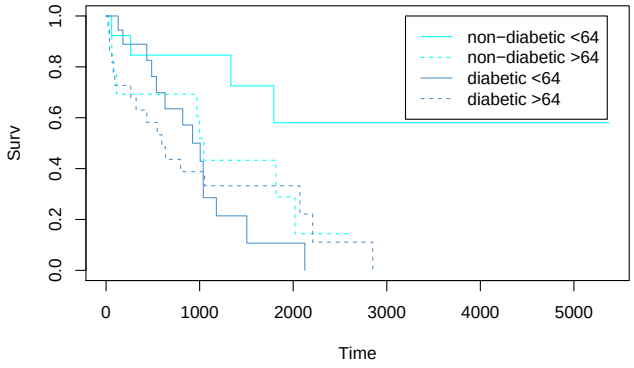
ii. Stratify by altgr (age group). Judge now the difference in survival between the diabetic and nondiabetic groups.

```
##                diabetes.diab
## diabetes.altgr  0  1
##              0 13 18
##              1 13 22
```

```
## Call:
## survdiff(formula = Surv(lzeit, tod) ~ diab + strata(altgr), data = diabetes)
##
##         N Observed Expected (O-E)^2/E (O-E)^2/V
## diab=0 26       13     19.6      2.22      4.32
## diab=1 40       31     24.4      1.78      4.32
##
##  Chisq= 4.3  on 1 degrees of freedom, p= 0.04
```

```
## Call: survfit(formula = Surv(lzeit, tod) ~ diab + strata(altgr), data = diabetes)
```

```
##
##                                n events *rmean *se(rmean) median 0.95LCL
## diab=0, strata(altgr)=altgr=0 13      4   2038        288     NA    1334
## diab=0, strata(altgr)=altgr=1 13      9   1231        275   1045     114
## diab=1, strata(altgr)=altgr=0 18     14    961        142    965     630
## diab=1, strata(altgr)=altgr=1 22     17   1026        218    595     321
##                               0.95UCL
## diab=0, strata(altgr)=altgr=0      NA
## diab=0, strata(altgr)=altgr=1      NA
## diab=1, strata(altgr)=altgr=0      NA
## diab=1, strata(altgr)=altgr=1      NA
##     * restricted mean with upper limit =  2743
```

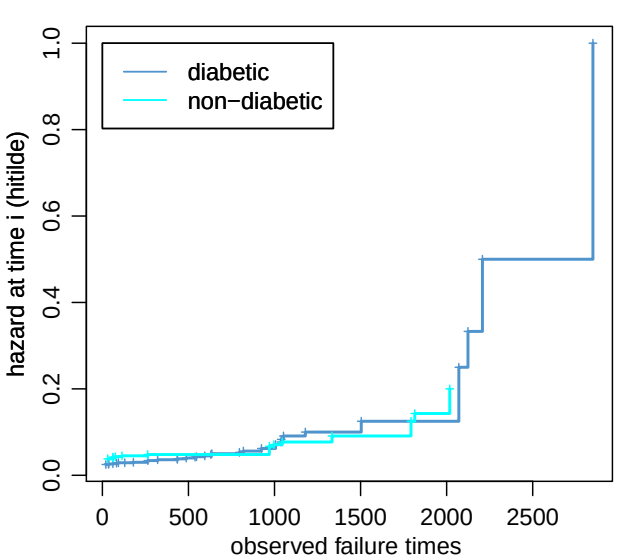


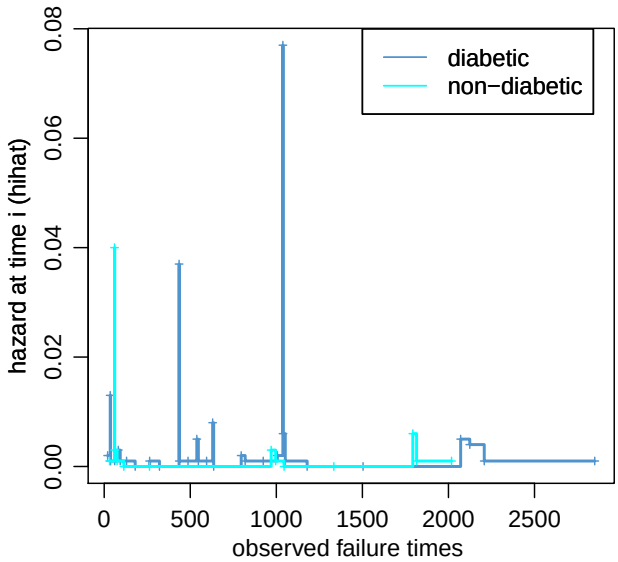
There is significant difference in survival between the two groups (the p-value is 0.04/2=0.02). Based on the figure, taking into account the variation due to age, the younger (≤ 64) non-diabetic group is much longer than other groups but survival time in the younger (≤ 64) diabetic group get down faster than other groups.

(d) Refer to the Hazard ratio as a measure of effect discussion starting on page 45. Does it appear that the hazard ratio between the two groups, diabetic and nondiabetic, is constant over time? That is, are the two empirical hazard functions proportional?

time	hitilde	hihat
31	0.038	0.001
60	0.04	0.04
61	0.042	0.003
75	0.043	0.001
114	0.045	0
263	0.048	0
970	0.067	0.003
996	0.071	0.001
1045	0.077	0
1334	0.091	0
1793	0.125	0.006
1815	0.143	0.001
2018	0.2	0.001

time	hitilde	hihat
20	0.025	0.002
35	0.026	0.013
37	0.026	0.001
61	0.027	0.001
82	0.028	0.003
93	0.029	0.001
130	0.029	0.001
180	0.03	0
264	0.034	0.001
321	0.036	0
435	0.037	0.037
436	0.038	0.001
487	0.04	0.001
538	0.042	0.005
547	0.043	0.001
595	0.045	0.001
630	0.048	0.008
636	0.05	0
796	0.053	0.002
819	0.056	0.001
924	0.062	0.001
1006	0.071	0.002
1038	0.077	0.077
1039	0.083	0.006
1052	0.091	0.001
1179	0.1	0
1504	0.125	0
2071	0.25	0.005
2124	0.333	0.004
2208	0.5	0.001
2850	1	0.001



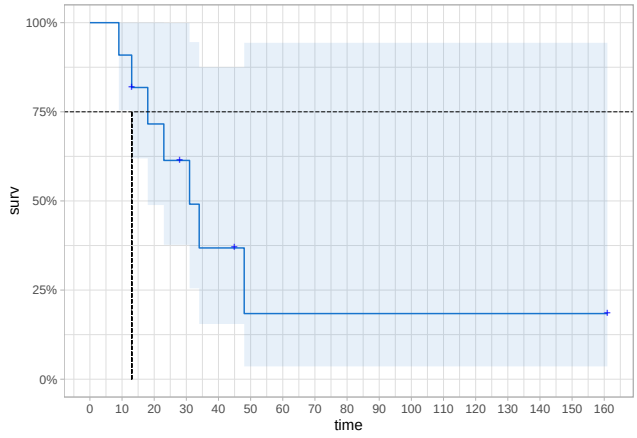


Both plots cross over time and don't display roughly parallel curves over time, which implies one group's risk is not always lower than another with respect to time.

Both plots indicate the hazard ratio between the two groups, diabetic and non-diabetic, is not constant with respect to time, which says the two empirical hazard functions of the two groups are not proportional.

As the April-30th lecture introduced, the kernel estimator of $h(t)$ can examine the hazard ratio better with larger data sets. It raises a question, how straight a hazard ratio curve is can be considered as constant? Since most of realistic data cannot give a perfect straight line.

2.5 Answer the WHY! on page 39.



Use survfit graphical method to find a confidence interval for the 25^{th} quantile, draw a horizontal line at 0.75 on the graph of the survival curve, and use intersections of this line with the curve and its upper and lower confidence bands. The lower confidence limit is the smallest time (t=13) at which the lower confidence limit for $S(t) \leq 0.75$.

However, the upper confidence limit for S(t) never less than 0.75, then the corresponding confidence limit for the 25^{th} quantile is unknown and it is represented as an NA.

2.7 On the data given in Exercise 2.1, compute by hand the truncated mean survival time (2.15) and its estimated variance (2.19). Check your answer using the appropriate S function.

y_i	d	n	p	s	se
0	0	8	1	1	0
1	1	8	0.875	0.875	0.1169
1	0	7	1	0.875	0.1169
2	1	6	0.8333	0.7292	0.165
4	2	5	0.6	0.4375	0.1879
4	0	3	1	0.4375	0.1879
6	1	2	0.5	0.2188	0.181
9	1	1	0	0	NA

```
## Call: survfit(formula = Surv(y, status) ~ 1, type = "kaplan-meier")
##
##          n     events      *rmean *se(rmean)     median    0.95LCL    0.95UCL
##       8.00       6.00       4.86       1.05       4.00       2.00         NA
##     * restricted mean with upper limit =  9
```

$$\widehat{mean} = \sum_{i=1}^{n'} \left(y_{(i)} - y_{(i-1)}\right) \hat{S}(y_{(i-1)})$$

$$= (1-0)\times 1 + \underbrace{(2-1)\times 0.8750}_{e_1} + \underbrace{(4-2)\times 0.7292}_{e_2} + \underbrace{(6-4)\times 0.4375}_{e_3} + \underbrace{(9-6)\times 0.2188}_{e_4} = 4.865$$

$$\widehat{Var}(\widehat{mean}) = \sum_{i=1}^{n'} \left(\int_{y_{(i)}}^{y_{(n)}} \hat{S}(u)du\right)^2 \frac{d_i}{n_i(n_i-d_i)}$$

$$= (\sum_{i=1}^{4} e_i)^2 \frac{1}{8(8-1)} + (e_2+e_3+e_4)^2 \frac{1}{6(6-1)} + (e_3+e_4)^2 \frac{2}{5(5-2)} + (e_4)^2 \frac{1}{2(2-1)} = 1.093$$

HW4 3.1 and 3.2.

3.1 Let T denote survival time of an experimental unit with survivor function $S(t) = \exp(-t/\theta)$ for $t > 0$ and $\theta > 0$. In this experiment n experimental units were observed and their lengths of life (in hours) were recorded. Let $t_1, ..., t_k$ denote the completely observed (uncensored) lifetimes, and let $c_{k+1}, c_{k+2}, ..., c_n$ denote the $n-k$ censored times. That is, this data set contains randomly right-censored data points.

(a) Derive the maximum likelihood estimate (MLE) $\hat{\theta}_{ML}$ for θ. Describe this in words. Refer to expression (3.8) and pages 70- 71.

$S(t) = \exp(-\frac{t}{\theta})$, $f(t) = \frac{1}{\theta}\exp(-\frac{t}{\theta})$

$$\log L(\theta) = \log \prod_{i=1}^{n} f^{\delta_i}(t_i|\theta) S_f^{1-\delta_i}(c_i|\theta) = \log\left(\prod_{i=1}^{k} f(t_i|\theta) \prod_{i=k+1}^{n} S_f(c_i|\theta)\right) = \sum_k \log f(t_i|\theta) + \sum_{n-k} \log S_f(c_i|\theta)$$

$$= \sum_k \log(\frac{1}{\theta}\exp(-\frac{t_i}{\theta})) + \sum_{n-k} \log\exp(-\frac{c_i}{\theta}) = -k\log\theta - \frac{\sum_k t_i}{\theta} - \frac{\sum_{n-k} c_i}{\theta}$$

$$\frac{\partial \log L(\theta)}{\partial\theta} = -\frac{k}{\theta} + \frac{\sum_k t_i + \sum_{n-k} c_i}{\theta^2} \overset{set}{=} 0 \implies \hat{\theta}_{ML} = \frac{\sum_k t_i + \sum_{n-k} c_i}{k}$$

The maximum likelihood estimator (MLE) $\hat{\theta}_{ML}$ is the total sum of survival time with censored and uncensored obervations divided by the number of uncensored observations.

(b) Referring to the observed information matrix $i(\theta)$ (3.11), we derive the following expression for the (estimated) asymptotic variance of $\hat{\theta}_{ML}$: $\widehat{var}_a(\hat{\theta}_{ML}) = \frac{(\hat{\theta}_{ML})^2}{k}$, where k is the number of uncensored data points n_u.

i. Calculate for $t_1, ..., t_5 = 1, 4, 7, 9, 12$ and $c_6, c_7, ..., c_{10} = 3, 3, 4, 6, 6$ the value of the estimate $\hat{\theta}_{ML}$ and $\widehat{var}_a(\hat{\theta}_{ML})$.

$\hat{\theta}_{ML} = \frac{1+4+7+9+12+3+3+4+6+6}{5} = 11$

$\widehat{var}_a(\hat{\theta}_{ML}) = \frac{(\hat{\theta}_{ML})^2}{k} = \frac{11^2}{5} = 24.2$

ii. Provide an asymptotic 95% confidence interval for θ, the true mean lifetime of the experimental units. Refer to expression (3.18) and Table 3.2.

$\log(\hat{\theta}) \overset{a}{\sim} N(\log(\theta), \frac{1}{n_u}) = N(\log(11), \frac{1}{5}) = N(2.398, \frac{1}{5})$

$\log(11) \pm 1.96 \times \frac{1}{\sqrt{5}} = (1.521, 3.274)$

The asymptotic 95% confidence interval for θ is $(e^{1.521}, e^{3.274}) = (4.578, 26.43)$.

(c) Refer to expression (3.14). Give the expression for Neyman-Pearson/ Wilks Likelihood Ratio Test (LRT) statistic $r^*(t)$ to test the hypothesis $H_0 : \theta = \theta_0$. Then calculate its value on the data in part (b) with $\theta_0 = 10$. Use the asymptotic distribution of $r^*(t)$ to test the hypothesis $H_0 : \theta = 10$ against $H_A : \theta \neq 10$. Also see page 75.

$\log(L(\theta_0)) = -k \log \theta_0 - \frac{\sum_k t_i}{\theta_0} - \frac{\sum_{n-k} c_i}{\theta_0} = -5 \log 10 - \frac{55}{10} = -17.01$

$\log(L(\hat{\theta})) = -k \log \hat{\theta} - \frac{\sum_k t_i}{\hat{\theta}} - \frac{\sum_{n-k} c_i}{\hat{\theta}} = -5 \log 11 - \frac{55}{11} = -16.99$

Under H_0, the Likelihood Ratio Test (LRT) $r^*(t) \overset{a}{\sim} \chi^2_d$ where the degrees of freedom equal to the difference in dimensionality of Θ and Θ_0.

$r^*(t) = -2 \log(L(\theta_0)) + 2 \log(L(\hat{\theta})) = -2(-17.01) + 2(-16.99) = 0.0469$

$T_1^* \sim \chi^2_1$. (Casella and Berger, 2002, Theorem 10.3.1) The p-value $= P(r^*(t) \geq 0.0469) = 0.8285$.

Therefore, we fail to reject $H_0 : \theta = 10$ at 95% confidence level.

(d) Suppose now that all n lifetimes are completely observed; that is, no censored times. Then $t1, ..., t5 = 1, 4, 7, 9, 12$ and $t6, ..., t10 = 3, 3, 4, 6, 6$. Compute $\hat{\theta}_{ML}$ and $\widehat{var}_a(\hat{\theta}_{ML})$. See page 68 and expression (3.11).

$\log L(\theta) = \log \prod_{i=1}^n f(t_i|\theta) = \sum_{i=1}^n \log(\frac{1}{\theta} \exp(-\frac{t_i}{\theta})) = -n \log \theta - \frac{\sum_{i=1}^n t_i}{\theta}$

$\frac{\partial \log L(\theta)}{\partial \theta} = -\frac{n}{\theta} + \frac{\sum_{i=1}^n t_i}{\theta^2} \overset{set}{=} 0 \implies \hat{\theta}_{ML} = \frac{\sum_{i=1}^n t_i}{n} = \frac{55}{10} = 5.5$

$i(\theta) = -\frac{\partial^2 \log L(\theta)}{\partial \theta^2} = -\frac{n}{\theta^2} + \frac{2 \sum_{i=1}^n t_i}{\theta^3}$

$\widehat{var}_a(\hat{\theta}_{ML}) = (i(\hat{\theta}))^{-1} = \frac{5.5^3}{2 \times 55 - 10 \times 5.5} = 3.025$

(e) For the complete data case in part (d), calculate the LRT statistic to test $H_0 : \theta = \theta_0$. Denote the LRT statistic by T_1^*. Use the asymptotic distribution of T_1^* to test $H_0 : \theta = 10$ against $H_A : \theta \neq 10$ with data from part (d). See page 71. Note that $T_1^* = r^*(\underline{t})$.

$\log(L(\theta_0)) = -10 \log \theta_0 - \frac{\sum_{i=1}^n t_i}{\theta_0} = -10 \log 10 - \frac{55}{10} = -17.01$

$\log(L(\hat{\theta})) = -n \log \hat{\theta} - \frac{\sum_{i=1}^n t_i}{\hat{\theta}} = -10 \log 5.5 - \frac{55}{5.5} = -27.05$

Under H_0, $r^*(t) = -2 \log(L(\theta_0)) + 2 \log(L(\hat{\theta})) = -2(-28.53) + 2(-27.05) = 2.957$

$T_1^* \sim \chi^2_1$. The p-value $= P(r^*(t) \geq 2.957) = 0.08551$.

Therefore, we fail to reject $H_0 : \theta = 10$ at 95% confidence level.

(f) In the complete data case there exists a test statistic T_2^*, equivalent to the LRT statistic T_1^*, to test $H_0 : \theta = \theta_0$: $T_2^* = \frac{2 \sum_{i=1} n t_i}{\theta_0}$. The exact distribution of T_2^* under H_0 is $\chi^2_{(2n)}$.

i. Conduct an analogous test to part (e) and compare results.

$T_2^* = \frac{2\sum_{i=1} nt_i}{\theta_0} = \frac{2\times 55}{10} = 11$

Using two-sided test, $P(T_2^* \geq 11) = 0.9462$, $P(T_2^* < 11) = 0.05378$. We use exact confidence interval.

$\chi^2_{(0.025,20)} = 9.591$, $\chi^2_{(0.975,20)} = 34.17$.

$9.591 < T_2^* < 34.17$, then we fail to reject $H_0 : \theta = 10$ at 95% confidence level.

Hence, T_2^* is equivalent to the LRT statistic T_1^*.

ii. Construct an exact 95% confidence interval for the true mean lifetime θ. Hint: Refer to page 70.

$T_2^* = \frac{2\sum_{i=1} nt_i}{\theta} = \frac{2n\hat{\theta}}{\theta} \in [9.591, 34.17]$

The exact 95% confidence interval for the true mean lifetime θ is

$\theta \in [\frac{2n\hat{\theta}}{34.17}, \frac{2n\hat{\theta}}{9.591}] = [3.219, 11.47]$.

3.2 The diabetes data in Exercise 2.3. We consider only the diabetic group. Refer to Section 3.4.2.

(a) Fit the data to the Weibull model. Then:

```
##
## Call:
## survreg(formula = Surv(lzeit, tod) ~ 1, data = diabetes1, dist = "weib")
##               Value Std. Error     z      p
## (Intercept) 6.9847     0.1824 38.29 <2e-16
## Log(scale)  0.0123     0.1482  0.08   0.93
##
## Scale= 1.01
##
## Weibull distribution
## Loglik(model)= -247.6   Loglik(intercept only)= -247.6
## Number of Newton-Raphson Iterations: 6
## n= 40
```

i. Obtain point and 95% C.I. estimates for the three quartiles.

	estimator	CI-L	CI-U
1st Quartile	305.9	179.7	520.7
2nd Quartile	745.2	508.8	1091
3rd Quartile	1503	1046	2161

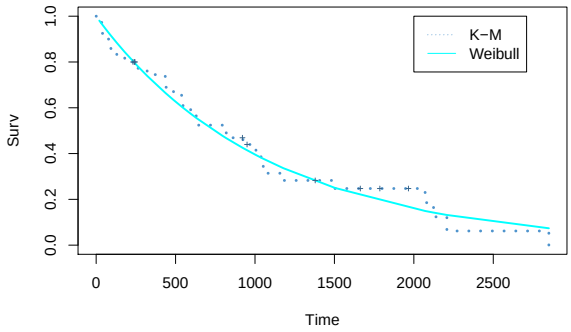
ii. Compute point estimates for S(t) at the uncensored survival times lzeit.

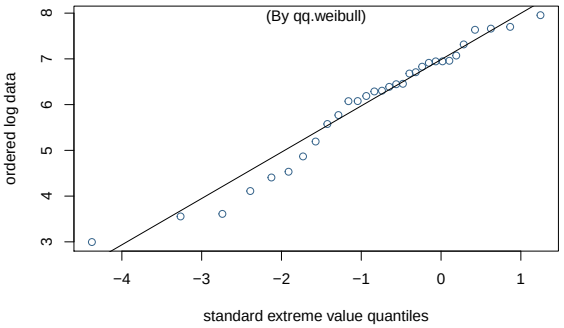
lzeit.u	20	35	37	61	82	93	130	180	264	321	435	436	487	538	547	595
weib.Shat	0.98	0.97	0.96	0.94	0.92	0.92	0.88	0.84	0.78	0.74	0.67	0.66	0.63	0.61	0.6	0.57

lzeit.u	630	636	796	819	924	1006	1038	1039	1052	1179	1504	2071	2124	2208	2850
weib.Shat	0.56	0.55	0.48	0.47	0.42	0.39	0.38	0.38	0.38	0.34	0.25	0.15	0.14	0.13	0.07

iii. Plot the Kaplan-Meier curve and the estimated Weibull survivor function ($\hat{S}_W(t)$) Shat on the same plot.

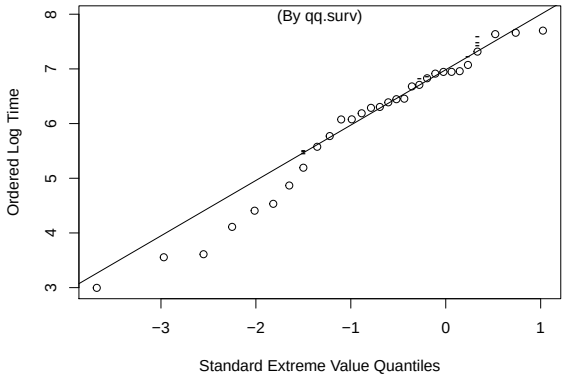
iv. Produce a Q-Q plot.






```
##    logtime     sevq
## 1    2.996 -3.67625
## 2    3.555 -2.97020
## 3    3.611 -2.55154
## 4    4.111 -2.25037
## 5    4.407 -2.01342
## 6    4.533 -1.81696
## 7    4.868 -1.64832
## 8    5.193 -1.49994
## 9    5.447 -1.49994
## 10   5.489 -1.49994
## 11   5.497 -1.49994
## 12   5.576 -1.35389
## 13   5.771 -1.22213
## 14   6.075 -1.10159
## 15   6.078 -0.99004
## 16   6.188 -0.88580
## 17   6.288 -0.78758
## 18   6.304 -0.69435
## 19   6.389 -0.60529
## 20   6.446 -0.51969
## 21   6.455 -0.43696
## 22   6.680 -0.35658
## 23   6.708 -0.27809
## 24   6.825 -0.27809
## 25   6.829 -0.19630
## 26   6.856 -0.19630
## 27   6.914 -0.10996
## 28   6.945 -0.02438
## 29   6.946  0.06103
## 30   6.958  0.14690
## 31   7.072  0.23396
## 32   7.229  0.23396
## 33   7.316  0.33442
## 34   7.416  0.33442
## 35   7.488  0.33442
## 36   7.584  0.33442
## 37   7.636  0.52165
```

```
## 38   7.661  0.73730
## 39   7.700  1.02368
## 40   7.955      Inf
```



(b) Fit the data to the log-normal model. Repeat all of part (a).

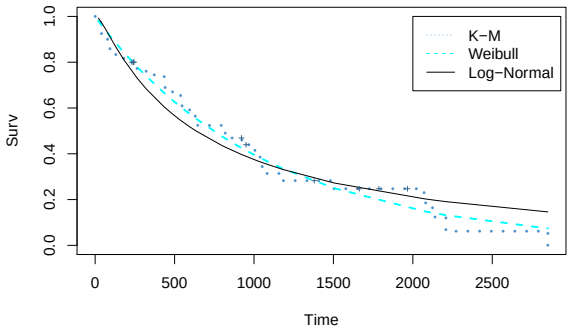
```
##
## Call:
## survreg(formula = Surv(lzeit, tod) ~ 1, data = diabetes1, dist = "lognormal")
##             Value Std. Error     z      p
## (Intercept) 6.453      0.238 27.17 <2e-16
## Log(scale)  0.355      0.129  2.74 0.0061
##
## Scale= 1.43
##
## Log Normal distribution
## Loglik(model)= -249.9   Loglik(intercept only)= -249.9
## Number of Newton-Raphson Iterations: 5
## n= 40
```

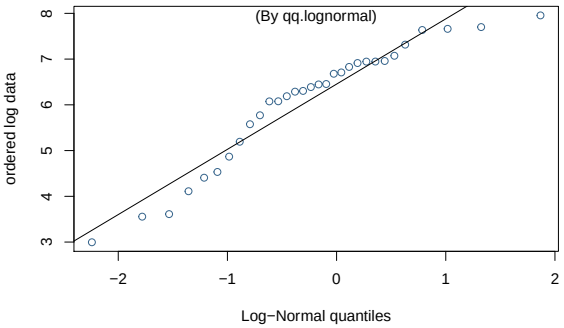
	estimator	CI-L	CI-U
1st Quartile	242.5	148.1	397.3
2nd Quartile	634.7	398.4	1011
3rd Quartile	1661	952.6	2895

lzeit.u	20	35	37	61	82	93	130	180	264	321	435	436	487	538	547	595
lognorm.Shat	0.99	0.98	0.98	0.95	0.92	0.91	0.87	0.81	0.73	0.68	0.6	0.6	0.57	0.55	0.54	0.52

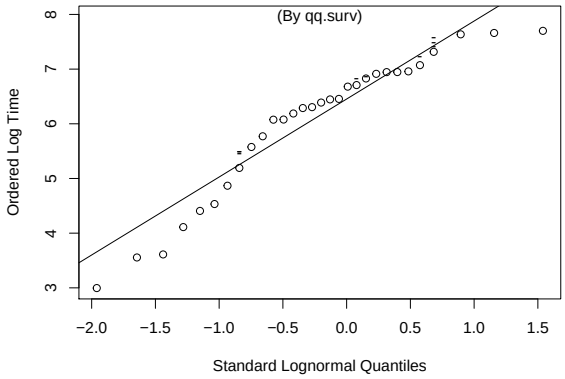
lzeit.u	630	636	796	819	924	1006	1038	1039	1052	1179	1504	2071	2124	2208	2850
lognorm.Shat	0.5	0.5	0.44	0.43	0.4	0.37	0.37	0.36	0.36	0.33	0.27	0.2	0.2	0.19	0.15

Plot Shat against lzeit.u (on the time axis). You must create your own qq.lognormal function. This is easy. Just read qq.loglogistic. Make minor changes.






```
##    logtime      sevq
## 1    2.996 -1.959964
## 2    3.555 -1.644854
## 3    3.611 -1.439531
## 4    4.111 -1.281552
## 5    4.407 -1.150349
## 6    4.533 -1.036433
## 7    4.868 -0.934589
## 8    5.193 -0.841621
## 9    5.447 -0.841621
## 10   5.489 -0.841621
## 11   5.497 -0.841621
## 12   5.576 -0.746820
## 13   5.771 -0.658301
## 14   6.075 -0.574666
## 15   6.078 -0.494873
## 16   6.188 -0.418116
## 17   6.288 -0.343750
## 18   6.304 -0.271239
## 19   6.389 -0.200129
## 20   6.446 -0.130019
## 21   6.455 -0.060542
## 22   6.680  0.008644
## 23   6.708  0.077871
## 24   6.825  0.077871
## 25   6.829  0.151844
## 26   6.856  0.151844
## 27   6.914  0.232046
## 28   6.945  0.313772
## 29   6.946  0.397652
## 30   6.958  0.484433
## 31   7.072  0.575030
## 32   7.229  0.575030
## 33   7.316  0.682992
## 34   7.416  0.682992
## 35   7.488  0.682992
## 36   7.584  0.682992
## 37   7.636  0.894678
## 38   7.661  1.156918
## 39   7.700  1.539620
## 40   7.955       Inf
```



(c) Compare your plots and comment as to how these models fit the data.

The plots of survival function show that Weibull distribution can fit the K-M curve better. The Log-Normal distribution underestimates the probability at the first half of time and overestimates it at the second half of time.

The QQ plots also show that Weibull model is closer to the straight line than Log-Normal model. Therefore, Weibull distribution is better than Log-Normal distribution for model adequacy in this case.

HW5 3.4, 3.5, and 3.8.

3.4 Derive the (estimated) asymptotic variance of the MLEθ_{ML} $\widehat{var}_a(\hat{\theta}_{ML}) = \frac{(\hat{\theta}_{ML})^2}{k}$, which was stated back in Exercise 3.1(b).

$\hat{\theta}_{ML} = \frac{\sum_k t_i + \sum_{n-k} c_i}{k} \overset{a}{\sim} MVN(\theta, I^{-1}(\theta))$, $\sum_k t_i + \sum_{n-k} c_i = k\hat{\theta}_{ML}$

$$\log L(\theta) = \log \prod_{i=1}^{n} f^{\delta_i}(t_i|\theta) S_f^{1-\delta_i}(c_i|\theta) = \sum_k \log(\frac{1}{\theta}\exp(-\frac{t_i}{\theta})) + \sum_{n-k} \log\exp(-\frac{c_i}{\theta}) = -k\log\theta - \frac{\sum_k t_i}{\theta} - \frac{\sum_{n-k} c_i}{\theta}$$

$$i(\theta) = -\frac{\partial}{\partial\theta}\frac{\partial \log L(\theta)}{\partial\theta} = -\frac{\partial}{\partial\theta}\left[-\frac{k}{\theta} + \frac{\sum_k t_i}{\theta} + \frac{\sum_{n-k} c_i}{\theta}\right] = -\frac{k}{\theta^2} + \frac{2(\sum_k t_i + \sum_{n-k} c_i)}{\theta^3}$$

$$\widehat{Var}_a(\hat{\theta}_{ML}) = i(\hat{\theta}_{ML})^{-1} = \left[-\frac{k}{\hat{\theta}_{ML}^2} + \frac{2k\hat{\theta}_{ML}}{\hat{\theta}_{ML}^3}\right]^{-1} = \frac{\hat{\theta}_{ML}^2}{k}$$

3.5 Show that the LRT based on the statistic T_1^* in Exercise 3.1(e) is equivalent to the test based on the test statistic T_2^* presented in Exercise 3.1(f). Hint: Show T_1^*, which is $r^*(t)$, is some convex function of T_2^*.

Let $\sum_k t_i + \sum_{n-k} c_i = k\hat{\theta}$; $T_2^* = 2\frac{\sum_k t_i + \sum_{n-k} c_i}{\theta_0} = \frac{2k\hat{\theta}}{\theta_0}$

$$T_1^* = r^*(t) = -2\log(L(\theta_0)) + 2\log(L(\hat{\theta})) = 2\left(k\log\theta_0 + \frac{\sum_k t_i + \sum_{n-k} c_i}{\theta_0} - k\log\hat{\theta} - \frac{\sum_k t_i + \sum_{n-k} c_i}{\hat{\theta}}\right)$$

$$= 2k\log(\frac{\theta_0}{\hat{\theta}}) + 2\frac{\sum_k t_i + \sum_{n-k} c_i}{\theta_0} - 2k = 2k\log(\frac{2k}{T_2^*}) + T_2^* - 2k = 2k\log(2k) - 2k\log(T_2^*) + T_2^* - 2k$$

$\frac{\partial^2 T_1^*}{\partial T_2^{*2}} = \frac{2k}{T_2^{*2}} > 0$, T_1^* is a convex function of T_2^*, the unique solution is

$$\frac{\partial T_1^*}{\partial T_2^*} = -\frac{2k}{T_2^*} + 1 \overset{set}{=} 0 \implies T_2^* = 2k, T_1^* = 0$$

$T_1^* = -2\log(L(\theta_0)) + 2\log(L(\hat{\theta})) = 0 \implies \hat{\theta} = \theta_0$

$T_2^* = \frac{2k\hat{\theta}}{\theta_0} = 2k \implies \hat{\theta} = \theta_0$

Hence, T_1^* is equivalent to T_2^*.

3.8 Show expression (3.15). Hint: Refer to the Example 6 in Hogg and Craig (1995, page 136).

$$T_i \sim Expo(\lambda) \implies \sum T_i \sim Gamma(n, \frac{1}{\lambda})$$
$$2\lambda \sum T_i \sim Gamma(n, 2) = \chi^2_{(2n)}$$

For $\sum_{i=1}^n T_i = \frac{n}{\hat{\lambda}}$

$$2\lambda \sum_{i=1}^{n} T_i = 2n\frac{\lambda}{\hat{\lambda}} \sim \chi^2_{(2n)}$$

- Note: By moment generating functions. Let $X_i \sim Expo(\lambda)$

$$M_{(2\lambda\sum_{i=1}^n x_i)}(t) \underset{\perp}{=} \prod_{i=1}^{n} (\underbrace{\frac{\lambda}{\lambda - 2\lambda t}}_{Expo(\lambda)}) = \underbrace{(1-2t)^{-n}}_{\Gamma(n,2)} = \underbrace{(1-2t)^{-\frac{2n}{2}}}_{\chi^2(2n)}$$

HW6 4.1 and 4.2.

4.1 We work with the diabetes data set again. Refer to Exercise 2.3. Consider the Weibull regression model

$$Y = log(\text{lzeit}) = \beta_0^* + \beta_1^* x^{(1)} + \beta_2^* x^{(2)} + \beta_3^* x^{(3)} + \sigma Z$$

where $Z \sim$ standard extreme value and

$$x^{(1)} = \begin{cases} 0 & \text{man} \\ 1 & \text{woman} \end{cases} \quad x^{(2)} = \begin{cases} 0 & \text{nondiabetic} \\ 1 & \text{diabetic} \end{cases} \quad x^{(3)} = \text{age in years}$$

(a) Estimate σ and the coefficients β_j^*. Which covariates are significant?

```
##
## Call:
## survreg(formula = Surv(lzeit, tod) ~ sex + diab + alter, data = diabetes,
##     dist = "weibull")
##               Value Std. Error     z      p
## (Intercept) 10.2306     1.0441  9.80 <2e-16
## sex         -0.0912     0.3772 -0.24  0.809
## diab        -0.6518     0.3979 -1.64  0.101
## alter       -0.0381     0.0156 -2.45  0.014
## Log(scale)   0.1226     0.1256  0.98  0.329
##
## Scale= 1.13
##
## Weibull distribution
## Loglik(model)= -359.3   Loglik(intercept only)= -366.1
##  Chisq= 13.57 on 3 degrees of freedom, p= 0.0036
## Number of Newton-Raphson Iterations: 6
## n= 66
```

The result shows that $\sigma = 1.13$, $\beta_0^* = 10.2306$, $\beta_1^* = -0.0912$, $\beta_2^* = -0.6518$, $\beta_3^* = -0.0381$. Age is significant at 95% significance level. ($P_{alter} = 0.014$)

(b) We now add two additional covariates which models in the possible dependence of diabetic or not with age. That is, we replace $x^{(3)}$ with the following interaction variables:

$$x^{(4)} = \begin{cases} \text{age}, & \text{if diabetic} \\ 0 & \text{otherwise} \end{cases} \quad x^{(5)} = \begin{cases} \text{age}, & \text{if nondiabetic} \\ 0 & \text{otherwise} \end{cases}$$

Describe the results of the analysis with the four covariates now. Which covariates are significant?

```
diabetes$x4 <- diabetes$x5 <- diabetes$alter
diabetes$x4[diabetes$diab==0] <- 0
diabetes$x5[diabetes$diab==1] <- 0
fit.b <- survreg(Surv(lzeit,tod) ~ sex+diab+x4+x5, data=diabetes)
summary(fit.b)
```

```
##
## Call:
## survreg(formula = Surv(lzeit, tod) ~ sex + diab + x4 + x5, data = diabetes)
```

```
##               Value Std. Error     z             p
## (Intercept) 12.1437     1.9115  6.35 0.00000000021
## sex         -0.1646     0.3760 -0.44         0.662
## diab        -3.8118     2.2859 -1.67         0.095
## x4          -0.0194     0.0195 -0.99         0.320
## x5          -0.0674     0.0280 -2.41         0.016
## Log(scale)   0.1104     0.1264  0.87         0.382
##
## Scale= 1.12
##
## Weibull distribution
## Loglik(model)= -358.2   Loglik(intercept only)= -366.1
##  Chisq= 15.78 on 4 degrees of freedom, p= 0.0033
## Number of Newton-Raphson Iterations: 6
## n= 66
```

The result shows that $\sigma = 1.12$. The covariates of age and diabetes is significant at 95% significance level. ($P_{x5} = 0.016$)

(c) Simplify the model fit in part (b) as much as possible. Draw conclusions (as much as possible as you are not diabetes specialists, etc.). For the remaining parts, use the fitted additive model fit.a (part (a)) with just sex, diab, and alter in the model.

```
## Start:  AIC=297.7
## Surv(lzeit, tod) ~ sex + diab + x4 + x5
##
##        Df AIC
## - sex   1 296
## - x4    1 296
## <none>    298
## - diab  1 299
## - x5    1 302
##
## Step:  AIC=296
## Surv(lzeit, tod) ~ diab + x4 + x5
##
##        Df AIC
## - x4    1 295
## <none>    296
## - diab  1 297
## - x5    1 300
##
## Step:  AIC=294.6
## Surv(lzeit, tod) ~ diab + x5
##
##        Df AIC
## <none>    295
## - x5    1 299
## - diab  1 302
```

```
## Call:
## coxph(formula = Surv(lzeit, tod) ~ diab + x5, data = diabetes)
##
```

```
##       coef exp(coef) se(coef) z    p
## diab  4.21     67.20     1.68 2 0.01
## x5    0.05      1.06     0.02 2 0.03
##
## Likelihood ratio test=12  on 2 df, p=0.003
## n= 66, number of events= 44
```

```
## Start:  AIC=728.3
## Surv(lzeit, tod) ~ sex + diab + x4 + x5
##
##        Df AIC
## - sex   1 727
## - x4    1 727
## <none>    728
## - diab  1 729
## - x5    1 734
##
## Step:  AIC=726.5
## Surv(lzeit, tod) ~ diab + x4 + x5
##
##        Df AIC
## - x4    1 726
## <none>    727
## - diab  1 728
## - x5    1 732
##
## Step:  AIC=725.5
## Surv(lzeit, tod) ~ diab + x5
##
##        Df AIC
## <none>    726
## - x5    1 731
## - diab  1 735
```

```
## Call:
## survreg(formula = Surv(lzeit, tod) ~ diab + x5, data = diabetes)
##
## Coefficients:
## (Intercept)        diab          x5
##    12.06265    -5.08582    -0.06651
##
## Scale= 1.12
##
## Loglik(model)= -358.8   Loglik(intercept only)= -366.1
##  Chisq= 14.57 on 2 degrees of freedom, p= 0.0007
## n= 66
```

The simplified model includes only diabetes and interaction between diabetes and age.

The survival time after the operation is significantly affect by whether you are diabetic or not, and older age with diabetes accelarate the change in survival time.

(d) Report the estimated hazard function for those who are men and nondiabetic.

Tip: See the summary on page 105.

```
hat.sigma <- fit.a$scale
(hat.alpha <- 1/hat.sigma)
```

```
## [1] 0.8846
```

```
(coef <- fit.a$coefficients)
```

```
## (Intercept)         sex        diab       alter
##    10.23056    -0.09116    -0.65183    -0.03811
```

Estimated hazard function $\hat{h}(t|age) = 0.8846t^{0.8846-1}(e^{-10.23+0.038age})^{0.8846}$

(e) Report the estimated hazard ratio comparing diabetic men to nondiabetic men all of whom have the same age. Interpret this ratio.

```
(beta3 <- -hat.alpha*coef[3])
```

```
##   diab
## 0.5766
```

```
exp((1-0)*beta3)
```

```
## diab
## 1.78
```

The estimated coefficient for diabetic status is 0.5766. The diabetic men has higher estimated hazard than non-diabetic men with the same age (1.78 times).

(f) A 50-year-old nondiabetic man is operated on today. What is the estimated probability that he is still alive in ten years?

```
lambda.tilde.h <- exp(-coef[1]-coef[2]*0-coef[3]*0-coef[4]*50)
1-pweibull(365*10,hat.alpha,1/lambda.tilde.h)
```

```
## [1] 0.4077
```

The estimated probability of surviving in ten years after operation is 40.77% for a 50-year-old non-diabetic man.

(g) With the help of the predict function, calculate the survival duration (in days or years) after the operation within which half (50%) of the 50-year-old diabetic men have died and then, similarly, for nondiabetic men. Report both point and interval estimates.

Tips: Use the preferred C.I. approach (Table 3.2 on Chapter 3); that is, type= "uquantile". Use the help routine in S or R to look up predict. Be sure to study the example given there. See the S code to compute the medhat for Model 2 on page 85.

```
hat.t.diab<- predict(fit.a,data.frame(sex=0,diab=c(1,0),alter=50),se.fit = T,type="uquantile",p=0.5)
CI <- matrix(exp(c(hat.t.diab[[1]],hat.t.diab[[1]]-qnorm(0.975)*hat.t.diab[[2]],
      hat.t.diab[[1]]+qnorm(0.975)*hat.t.diab[[2]])),2,3,
      dimnames=list(c("diabetic","non-diabetic"),c("estimator","CI-L","CI-U")))
pander(CI)
```

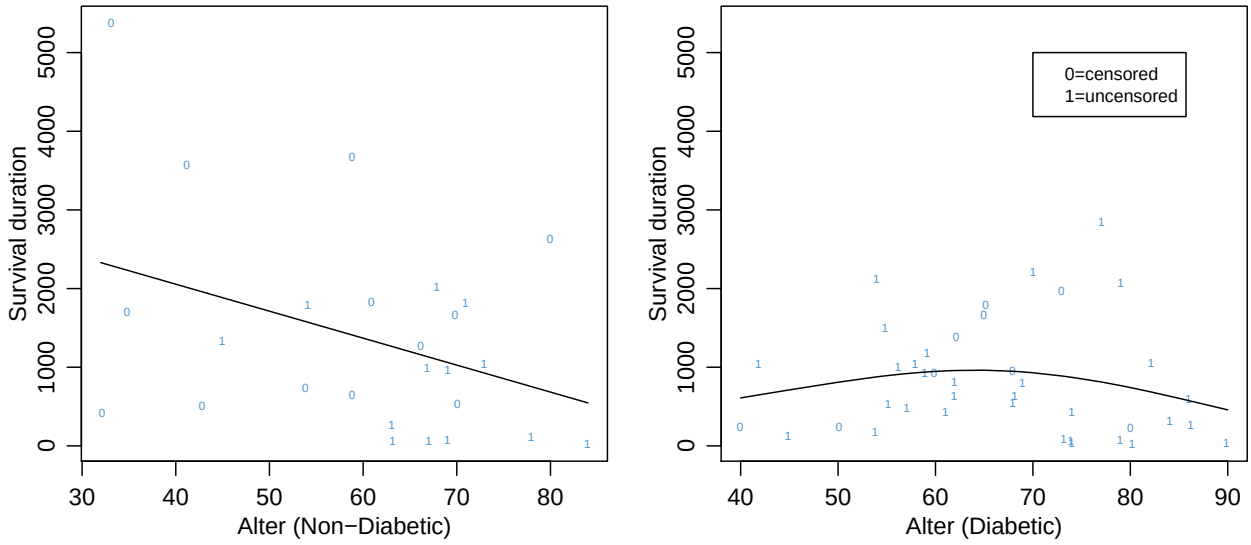
	estimator	CI-L	CI-U
diabetic	1421	663.8	3040
non-diabetic	2726	1299	5720

The predict survival duration is 1421 days after the operation within half of the 50-year-old non-diabetic men have died. The 95% confidence interval is $(1299, 5720)$ days.

It is 2726 days for non-diabetic men. The 95% confidence interval is $(663.8, 3040)$ days.

4.2 In order to better understand the age dependence of survival "lzeit", plot now the survival times against "$x^{(5)}$" and then against "$x^{(4)}$". Comment. Is there something here that helps explain what it is you are observing?

Tips: To investigate the age structure inherent in the raw data set, split the data set in to two sets: one with data corresponding to diabetics, the other with nondiabetics.



The plot show that, in the non-diabetic group, the younger patients can survive longer post operation. However, in the diabetic group, the longest survival duration happened for the 60-65 age old patients. It can be explained that the young diabetic patient is more vunerable than the 60-65 age old patients.

HW7 5.1.

5.1 We work with the diabetes data again. Refer to Exercise 2.3. Instead of a stratified analysis, we will now fit a Cox proportional hazards model: $h(t; x^{(1)}, x^{(2)}, x^{(3)}) = h_0(t) \cdot e^{\beta_1 x^{(1)} + \beta_2 x^{(2)} + \beta_3 x^{(3)}}$

where $x^{(1)} = \begin{cases} 0 & \text{man} \\ 1 & \text{woman} \end{cases}$ $x^{(2)} = \begin{cases} 0 & \text{nondiabetic} \\ 1 & \text{diabetic} \end{cases}$ $x^{(3)} = \text{age in years}$

(a) Describe the results of the analysis with all three covariates x(1), x(2), and x(3). Which covariates are significant?

```
## Call:
## coxph(formula = Surv(lzeit, tod) ~ sex + diab + alter, data = diabetes)
##
##   n= 66, number of events= 44
##
##         coef exp(coef) se(coef)    z Pr(>|z|)
## sex   0.1402    1.1505   0.3400 0.41    0.680
## diab  0.5302    1.6993   0.3507 1.51    0.131
## alter 0.0293    1.0298   0.0133 2.20    0.028 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##       exp(coef) exp(-coef) lower .95 upper .95
## sex        1.15      0.869     0.591      2.24
## diab       1.70      0.588     0.855      3.38
## alter      1.03      0.971     1.003      1.06
##
## Concordance= 0.641  (se = 0.049 )
## Likelihood ratio test= 10.8  on 3 df,   p=0.01
## Wald test            = 9.81  on 3 df,   p=0.02
## Score (logrank) test = 10.2  on 3 df,   p=0.02
```

The coefficient of 'sex', 'diab', and 'alter' is 0.1402, 0.5302, and 0.0293 respectively. Age (alter) is the significant covariate at 95% cignificance level.

(b) Stay with estimated full model in part (a) regardless of whether or not the coefficients of the covariates are statistically significantly different from zero.

i. Use the hazard ratio to estimate the gender effect when the other covariates are held constant. Put woman in the numerator. Interpret!

```
exp((1-0)*diab.cox$coefficients[1])
```

```
##   sex
## 1.151
```

When the other covariates are held constant, the hazard for women is 1.151 times of men.

ii. Use the hazard ratio to estimate the effect of $x^{(2)}$ and age together for the same gender. Take $x^{(2)} = 1$ and $x^{(3)} = age + 1$ in the numerator and $x^{(2)} = 0$ and $x^{(3)} = age$ in the denominator. Interpret!

```
exp(sum(diab.cox$coefficients[2:3]))
```

```
## [1] 1.75
```

$$HR = \frac{h(t|x_2)}{h(t|x_1)} = \frac{\exp(\beta_0 + \beta_1 * sex + \beta_2 * 1 = +\beta_3 * (age + 1))}{\exp(\beta_0 + \beta_1 * sex + \beta_2 * 0 + \beta_3 * age)} = \exp(\beta_2 + \beta_3)$$

The hazard for diabetic patients is 1.75 times of the non-diabetic and 1-year-younger patients with same gender.

(c) Simplify (reduce) the model in part (a) as far as possible. Comment!

```
diab.cox.AIC$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Surv(lzeit, tod) ~ sex + diab + alter
##
## Final Model:
## Surv(lzeit, tod) ~ diab + alter
##
##
##    Step Df Deviance Resid. Df Resid. Dev   AIC
## 1                          41      291.8 297.8
## 2 - sex  1   0.1677        42      292.0 296.0
```

```
diab.cox2.AIC$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Surv(lzeit, tod) ~ diab + alter
##
## Final Model:
## Surv(lzeit, tod) ~ diab + alter
##
##
##   Step Df Deviance Resid. Df Resid. Dev AIC
## 1                         42        292 296
```

```
diab.cox3 <- coxph(Surv(lzeit,tod) ~ alter, diabetes)
1-pchisq(c(2*diab.cox$loglik[2]-2*diab.cox2$loglik[2],2*diab.cox2$loglik[2]-2*diab.cox3$loglik[2]),c(1,
```

```
## [1] 0.68217 0.09201
```

There is not evidence against the models including covariates of 'diab' and 'alter' without any interactions (p-value=0.68217).

Further reducing model with only "alter" is not recommmended (p-value=0.09201).

(d) We now add two additional covariates, which model in the possible dependence of diabetic or not with age. That is, we replace x(3) with the following interaction variables:

$$x^{(4)} = \begin{cases} \text{age}, & \text{if diabetic} \\ 0 & \text{otherwise} \end{cases} \quad x^{(5)} = \begin{cases} \text{age}, & \text{if nondiabetic} \\ 0 & \text{otherwise} \end{cases}$$

Now do the analysis as in part (a).

```
diabetes$x4 <- diabetes$x5 <- diabetes$alter
diabetes$x4[diabetes$diab==0] <- 0
diabetes$x5[diabetes$diab==1] <- 0
diab.cox4 <- coxph(Surv(lzeit,tod) ~ sex+diab+x4+x5, data=diabetes)
summary(diab.cox4)
```

```
## Call:
## coxph(formula = Surv(lzeit, tod) ~ sex + diab + x4 + x5, data = diabetes)
##
##   n= 66, number of events= 44
##
##         coef exp(coef) se(coef)    z Pr(>|z|)
## sex   0.1968    1.2175   0.3428 0.57    0.566
## diab  3.3109   27.4089   2.0400 1.62    0.105
## x4    0.0137    1.0138   0.0171 0.80    0.424
## x5    0.0558    1.0573   0.0244 2.28    0.023 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##      exp(coef) exp(-coef) lower .95 upper .95
## sex       1.22     0.8214     0.622      2.38
## diab     27.41     0.0365     0.503   1494.05
## x4        1.01     0.9864     0.980      1.05
## x5        1.06     0.9458     1.008      1.11
##
## Concordance= 0.64  (se = 0.05 )
## Likelihood ratio test= 12.9  on 4 df,   p=0.01
## Wald test            = 8.74  on 4 df,   p=0.07
## Score (logrank) test = 10.4  on 4 df,   p=0.03
```

(e) Simplify the model in part (d) as much as possible. Draw conclusions (as much as possible as you are not diabetes specialists, etc.) and compare these results with those from your results in the stratified analysis you performed in Exercise 2.4(c).

```
diabetes$x6 <- diabetes$x7 <- diabetes$alter
diabetes$x6[diabetes$sex==0] <- 0
diabetes$x7[diabetes$sex==1] <- 0
diab.cox5 <- coxph(Surv(lzeit,tod) ~ sex*diab+x4+x5+x6+x7, data=diabetes)
```

```
diab.cox5.AIC$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
```

```
## Initial Model:
## Surv(lzeit, tod) ~ sex * diab + x4 + x5 + x6 + x7
##
## Final Model:
## Surv(lzeit, tod) ~ diab + x5
##
##
##         Step Df Deviance Resid. Df Resid. Dev   AIC
## 1                                38      288.5 300.5
## 2       - x7  0 0.000000        38      288.5 300.5
## 3       - x4  1 0.001974        39      288.5 298.5
## 4 - sex:diab  1 0.496497        40      289.0 297.0
## 5      - sex  1 0.943902        41      290.0 296.0
## 6       - x6  1 0.662365        42      290.7 294.7
```

```
diab.cox6 <- coxph(Surv(lzeit,tod) ~ diab+x5, diabetes)
1-pchisq(2*diab.cox5$loglik[2]-2*diab.cox6$loglik[2],5)
```

```
## [1] 0.8345
```

The simplified model includes 'diab' and the interation term between 'diab' and 'age'. The main effect of diabetic affects the survival duration after operation significantly. The non-diabetic older patients accelerate the change in survival time. The result is same with 2.4(c). There is not significant evident that gender affects the survival duration.

(f) Use stepAIC to select the best subset of variables from the four variables in part (d). Consider only main effects models. Do "backward" elimination starting with all four variables in the model. Refer to Remark 2 on page 129.

```
diab.cox4 <- coxph(Surv(lzeit,tod) ~ sex+diab+x4+x5, data=diabetes)
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Surv(lzeit, tod) ~ sex + diab + x4 + x5
##
## Final Model:
## Surv(lzeit, tod) ~ diab + x5
##
##
##    Step Df Deviance Resid. Df Resid. Dev   AIC
## 1                          40      289.7 297.7
## 2 - sex  1   0.3234        41      290.0 296.0
## 3  - x4  1   0.6432        42      290.7 294.7
```

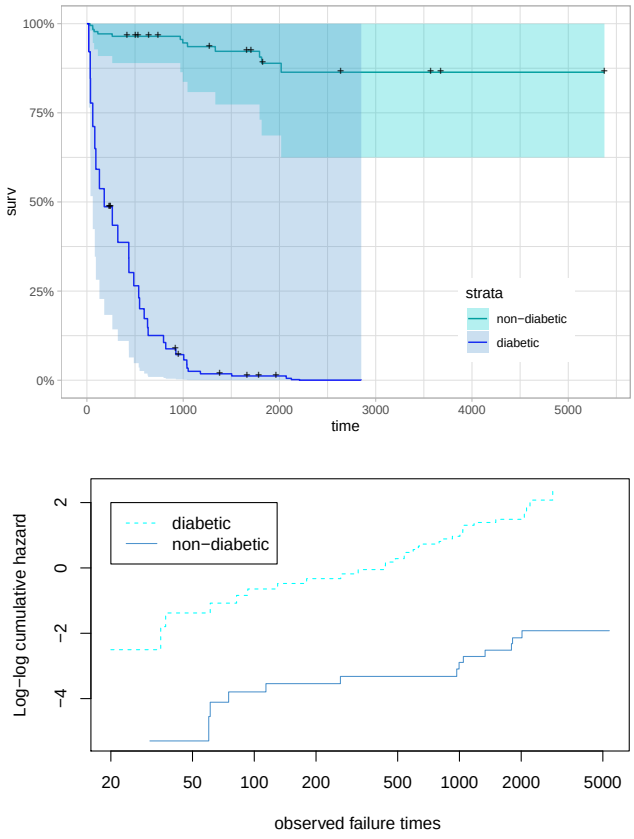
The backward AIC gives the same simplified model including 'diab', interaction between 'diab' and 'age'.

(g) Fit your selected model from stepAIC to a Cox regression model stratified on the diab variable. See page 133.

```
(diab.cox7 <- coxph(Surv(lzeit,tod) ~ strata(diab)+x5, data=diabetes) )
```

```
## Call:
## coxph(formula = Surv(lzeit, tod) ~ strata(diab) + x5, data = diabetes)
##
##    coef exp(coef) se(coef) z    p
## x5 0.05      1.05     0.02 2 0.05
##
## Likelihood ratio test=5  on 1 df, p=0.03
## n= 66, number of events= 44
```

(h) For the stratified fit in part (g), produce a plot of the survival curves and produce a plot of the log-log cumulative hazards. See pages 134 and 135.



Grube-Cavers, A., & Patterson, Z. (2015). Urban rapid rail transit and gentrification in Canadian urban centres: A survival analysis approach. Urban Studies, 52(1), 178–194. https://doi.org/10.1177/0042098014524287