Exercise 07: Survival Analysis
2018 Spring

1 Concepts of Survival Analysis

A survival analysis is a method for analyzing time to events, where the events can be "death" or "failure", etc.
The task of survival analysis is to

(1) Estimate and interpret survivor and/or hazard functions;

(2) Compare survivor and/or hazard functions

(3) Assess the relationship between explanatory variables and survival time

1.1 Concepts

Censoring The survival time is not exactly known due to

- A subject does not experience the event until the study ends.
- A subject is lost-to-follow-up during the study period.
- A subject withdraws from the study due to some other reason.

Right-censored Unknown but $T > t$

Left-censored Unknown but $T < t$

Interval-censored Unknown but $t_1 < T < t_2$

Survival time T: the outcome variable (time to event)

Risk set The set of subjects with $T \geq t$

1.1.1 Survivor function

- $S(t) = P(T > t)$
- Probability that the survival time T exceeds a specified time t
- We often use the empirical survivor function $\hat{S}(t)$.

1.1.2 Hazard function: Instantaneous hazard

- **Force of mortality**
- $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t)}{\Delta t}$
- $h(t) \geq 0$ and has no upper bounds
- Hazard function is also called **failure rate**.
- *Rate of events occurring per time unit*, e.g., 50 events per month = 600 events per year.

1.1.3 Relationship between $S(t)$ and $h(t)$

- $S(t) = \exp\left[-\int_0^t h(u)du\right]$
- $h(t) = -\frac{dS(t)/dt}{S(t)}$
- If $S(t) = e^{-\lambda t}$, $h(t) = \lambda$

1.1.4 Basic Descriptive Analysis

- Mean survival time $\overline{T}$ (平均生存时间，ignoring the censorship)
- Median survival time (中位生存时间，$t|\hat{S}(t) = 0.5$)
- Average hazard rate (平均风险率) $\overline{h} = \#\text{failures}/\sum_{i=1}^{n} t_i$

Example 1 (Descriptive Analysis of Survival Time)

individual	*t(weeks)*	δ *(failed=1;censored=0)*
1	*3.5*	*0*
2	*3.5*	*1*
3	*5*	*1*
4	*6*	*0*
5	*8*	*0*
6	*12*	*0*

- *The mean survival time is* $\overline{T} = \frac{3.5+5}{2} = 4.25$ *weeks.*
- *The average hazard rate is* $\overline{h} = 2/(3.5 + 3.5 + 5 + 6 + 8 + 12) = 0.0526$ *failures per week.*

2 Survival Analysis - Inference

The survival analysis can be classified into three main categories:

- Parametric methods: The survial times follow some parametric distribution
  - Lognormal distribution
  - Weibull distribution
  - Exponential distribution
  - Gamma distribution
- Nonparametric methods
  - Survival rate through Kaplan-Meier or life tables
  - Comparing n groups of survival rates through logrank test (n=2) or Breslow test (n=3+)
- Semi-parametric methods
  - Cox-proportional hazards model

2.1 Life tables

2.1.1 Assumptions

- There are no changes in survivorship over calendar time
- The experience of individuals who are lost to follow-up is the same as the experience of those who are followed.
- Withdrawal occurs uniformly within the interval.
- Event occurs uniformly within the interval.

Ordered failure times ($t_{(i)}$)	failures (m_i)	censored (q_i)	Risk set $R(t_{(i)})$
$t_{(0)} = 0$	m_0	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{(n)}$	m_n	q_n	$R(t_{(n)})$

2.2 Kaplan-Meier Estimation

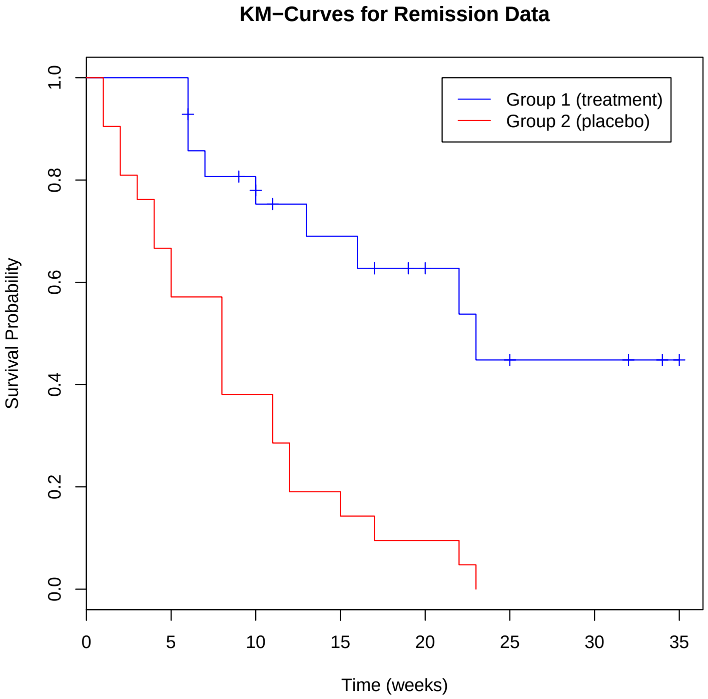
The Kaplan-Meier survival table is in the following form:

Table 1: Kaplan-Meier Table Example

t_j	n_j	d_j	$P(t_j)$	$S(t_j)$	SE
5	28	1	27/28=0.964	0.96	0.04
29	22	1	21/22=0.955	0.92	0.05
37	20	1	19/20=0.950	0.87	0.07
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

- The estimator $\hat{q}_i = d_j/n_j$ is the estimate of $h(t_j)$;
- The survival probability $P(t_j) = 1 - \hat{q}_j$
- The survival rate $S(t_j) = \prod_{t_i \leq t_j} P(t_i)$
- The standard error is $\text{SE}(S(t_j)) = S(t_j)\left[\sum_{i=1}^{j} \frac{d_j}{n_j(n_j - d_j)}\right]$

```
time1 <- c(6,6,6,7,10,13,16,22,23,6,9,10,11,17,19,20,25,32,32,34,35)
status1 <- c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0)
time2 <- c(1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23)
status2 <- c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1)
time <- c(time1, time2)
status <- c(status1, status2)
group <- factor(c(rep(0,21), rep(1,21)))
fit <- survfit(Surv(time,status) ~ group)
plot(fit, conf.int="none", col=c("blue", "red"),
        mark.time = TRUE,
        xlab="Time (weeks)", ylab="Survival Probability")
legend(21,1,c('Group 1 (treatment)', 'Group 2 (placebo)'),
        col = c('blue','red'), lty = 1)
title(main='KM-Curves for Remission Data')
```



2.3 Hypothesis testing for the survival data

There are three statistical test methods for comparing 2+ survival curves:

- **Log-rank test** using equal weights on different observed time points.
- **Breslow test** using the risk set size $|R_i|$ as the weight for each time point.
- **Tarone-Ware test** using the square root of the risk set size $\sqrt{|R_i|}$ as the weight for each time point.

2.3.1 Logrank test for comparing two survival curves/functions

- Sort the K unique times: $t_1 < t_2 < \cdots < t_K$
- n_{ij}: number of persons in group i at risk at t_j
- $n_j = \sum_i n_{ij}$ the total number of subjects at risk at t_j
- o_{ij}: number of failures in group i at t_j
- $o_j = \sum_j o_{ij}$: total number of failures at t_j

Under the null hypothesis $H_0 : S_1(t) = S_2(t), 0 < t < \infty$, o_{1j} has the hypergeometric distribution conditional on the margins $\{n_{1j}, n_{2j}, o_j, n_j - o_j\}$:

$$\mathbb{P}(o_{1j}) = \binom{o_j}{o_{1j}}\binom{n_j - o_j}{n_{1j} - o_{1j}} \Big/ \binom{n_j}{n_{1j}}$$

Then we can get the conditional expectation and variance:

$$\begin{aligned} e_{1j} &= E(o_{1j}\big|\text{marginals}) \\ &= \left(\frac{n_{1j}}{n_j}\right) o_j \\ V_j &= Var(o_{1j}\big|\text{marginals}) \\ &= \frac{n_j - n_{1j}}{n_j - 1} \times n_{1j} \left(\frac{o_j}{n_j}\right)\left(1 - \frac{o_j}{n_j}\right) \\ &= \frac{n_{1j} n_{2j} o_j (n_j - o_j)}{n_j^2 (n_j - 1)} \end{aligned}$$

Since

$$z = \frac{\sum_{j=1}^{K}(o_{1j} - e_{1j})}{\sqrt{\sum_{j=1}^{K} V_j}} \sim N(0,1) \text{ under } H_0$$

we can obtain the log-rank test statistic:

$$X^2 = \frac{\left(\sum_{j=1}^{K}(o_{1j} - e_{1j})\right)^2}{\sum_{j=1}^{K} V_j} \sim \chi^2(df = 1)$$

with χ^2 we can get the p-value to decide whether to reject the null hypothesis.
This can be executed in R:

```
fit <- survdiff(Surv(time,status) ~ group, data, rho=0)
summary(fit)
```

3 Exercises

1. (10 points) True or False

   (1) **F** The survival function $S(t)$ ranges between 0 and ∞.

   (2) **T** A hazard rate of one per day is equivalent to seven per week.

   (3) **T** If you know the form of hazard function, then you can determine the corresponding survivor curve, and vice versa.

   (4) **F** If the survival curve for group 1 lies completely above the curve for group2, the median survival time for group 2 is longer than that for group 1.

   (5) **F** The risk set at six weeks is the set of individuals whose survival time are less than or equal to six weeks.

   (6) **F** If the risk set at 6th week ocnsists of 22 persons, and 4 persons failed and 3 are censored by the 7th week, then the risk set at 7th week consists of 18 persons.

   (7) **T** If a hazard ratio comparing group 1 relative to group 2 equals 10, then the potential for failure is 10 times higher in group 1 than in group 2.

   (8) **T** Survivor function is a proportion metric, while hazard function is a rate metric.

   (9) **F** Compared to standard log-rank test, Peto-Prentice test place more emphasis on the late-occurred failures.

   (10) **F** Compared to life table, the Kaplan-Meier table is more commonly used in actuary.

2. (5 points) The **mean residual life time (mrl)** can be defined as

$$\text{mrl}(t_0) = E[T - t_0 | T \geq t_0],$$

i.e. the *average remaining survival time given the population has survived beyond* t_0. Prove that

$$\text{mrl}(t_0) = \frac{\int_{t_0}^{\infty} S(t)dt}{S(t_0)}.$$

Solution:

$$\begin{array}{rcl} \text{mrl}(t_0) & = & E[T - t_0 | T \geq t_0] \\ & = & \frac{E[(T-t_0)I(X>t)]}{P(X>t)} \\ & = & \frac{1}{1-F(t_0)} \int_{t_0}^{\infty} (t - t_0) dF(t) \end{array}$$

and since

$$\begin{array}{rcl} \int_{t_0}^{\infty} (t - t_0) dF(t) & = & \int_{t_0}^{\infty} \left(\int_{t_0}^{t} du \right) dF(t) \\ & = & \int_{t_0}^{\infty} \left(\int_{u}^{\infty} dF(t) \right) du \text{—(Tonellis' Theorem)} \\ & = & \int_{t_0}^{\infty} P(X > u) du \\ & = & \int_{t_0}^{\infty} (1 - F(u)) du \end{array}$$

Therefore,

$$\begin{array}{rcl} \text{mrl}(t_0) & = & \frac{1}{1-F(t_0)} \int_{t_0}^{\infty} (1 - F(t)) dt \\ & = & \frac{1}{S(t_0)} \int_{t_0}^{\infty} S(t) dt \end{array}$$

3. (10 points) The time (in days) to developing a tumor for rats exposed to a carcinogen follows a Weibull distribution with shape parameter $\lambda_0 = 0.5$ and scale parameter $\lambda_1 = 2$.

(1) (2 points) Compute the probability that a random rat will be tumor-free at the 30-th day.

Solution:

$$S(30) = Pr(T > 30) = e^{-\lambda_0 t^{\lambda_1}} = e^{-0.5 \times 30^2} = 3.69e - 196$$

(2) (2 points) What is the average time to tumor development?

Solution: The mean survival time

$$\int_0^{\infty} S(t)dt = \int_0^{\infty} e^{-\lambda_0 t^{\lambda_1}} dt = \int_0^{\infty} e^{-0.5t^2} dt$$

(3) (3 points) Find the hazard rate of time to tumor develpment at the 30-th day.

Solution:

$$h(t) = \lambda_1 \lambda_0 t^{\lambda_1 - 1} = t = 30/day$$

(4) (3 points) Find the median time to tumor development.

Solution:

$$t_{0.5} = \left(\frac{\log 2}{\lambda}\right)^{1/\alpha} = \sqrt{2\log 2} = 1.18\text{days}$$

4. (5 points) Suppose we have a small data set with different kinds of censoring: $2+, 3, 4, 5-, 6, 7+, [5, 7]$, Suppose the distribution of the underlying survival time is an exponential distribution with a constant hazard λ. Write down the likelihood function of λ for this given data set.

5. (20 points) A survival analysis was conducted to compare the survival times (in years) for two groups each with 25 participants. `CHR` is used to indicate whether the group has history of chronic disease ($CHR = 1/0$).

Group 1 (CHR=0)	12.3+,5.4,8.2,12.2,11.7,10.0,5.7,9.8,2.6, 11.0,9.2,12.1+,6.6,2.2,1.8,10.2,10.7,11.1, 5.3,3.5,9.2,2.5,8.7,3.8,3.0
Group 2 (CHR=1)	5.8,2.9,8.4,8.3,9.1,4.2,4.1,1.8,3.1,11.4, 2.4,1.4,5.9,1.6,2.8,4.9,3.5,6.5,9.9,3.6, 5.2,8.8,7.8,4.7,3.9

(1) (10 points) Make a life table and Kaplan-Meier table for each group, respectively.

(2) (5 points) Compute the average survival times ($\overline{T}$) and average hazard rates $\overline{h}$ for two groups. Which group has a better prognosis? Explain briefly.

Solution:

	$\overline{T}$	$\overline{h}$
Group 1	7.5	0.1165
Group 2	5.3	0.1894

Group 1 has better prognosis since $\overline{T}_1$ is larger and there are censored observations in group 1 which are not considered in calculating $\overline{T}$, but they are the largest 3 observations.

(3) (5 points) How would a comparison of survivor curves provide additional information to what is provided in the table?

Solution: A comparison of the two survivor curve allows to compare the survival curve within different time section and give a insight on how difference in a history of chronic disease will influence the overall survival time.

6. (20 points) Conduct the log-rank test procedure to compare these two survival data. You need to write down the details.

Group 1	6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+
Group 2	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Solution: Fill the Kaplan-Meier survival table,

t_j	o_{1j}	n_{1j}	o_{2j}	n_{2j}	e_{1j}	$e_{1j} - o_{1j}$
1	0	21	2	21		
2	0	21	2	19		
3	0	21	1	17		
4	0	21	2	16		
5	0	21	2	14		
6	3	21	0	12		
7	1	17	0	12		
8	0	16	4	12		
9	0	16	0	8		
10	1	15	0	8		
11	0	13	2	8		
12	0	12	2	6		
13	1	12	0	4		
15	0	11	1	4		
16	1	11	0	3		
17	0	10	1	3		
19	0	9	0	2		
20	0	8	0	2		
22	1	7	1	2		
23	1	6	1	1		
25	0	5	0	0		
32	0	4	0	0		
34	0	2	0	0		
35	0	1	0	0		

Since

$$\begin{aligned} e_{1j} &= \tfrac{n_{1j}(o_{1j}+o_{2j})}{n_{1j}+n_{2j}} \\ O_1 - E_1 &= \textstyle\sum_j (o_{1j} - e_{1j}) \\ v_j &= \tfrac{n_{1j}n_{2j}(o_{1j}+o_{2j})(n_{1j}+n_{2j}-o_{1j}-o_{2j})}{(n_{1j}+n_{2j})^2(n_{1j}+n_{2j}-1)} \end{aligned}$$

Compute the statistic:

$$X^2 = (O_1 - E_1)^2 / \sum_j v_j \sim \chi^2_{df=1}$$

7. (20 points) The dataset `veterans.dat` considers the survival times (days) for 137 patients from the Veterans Administration Lung Cancer Trial.

```
data <- read.table("http://cbb.sjtu.edu.cn/course/bi476/data/
    veterans.dat", header=F)
```

   (1) (5 points) Obtain the Kaplan-Meier plots for the two cell type (1=large, 0=other). Comment on how the two curves compare with each other. Moreover, draw a conclusion based on log-rank test.

   (2) (5 points) Obtain Kaplan-Meier plots for the four cell types (large, adeno, small, and squamous). Note that you will need to recode the data to define a single variable which numerically distinguishes the four categories.

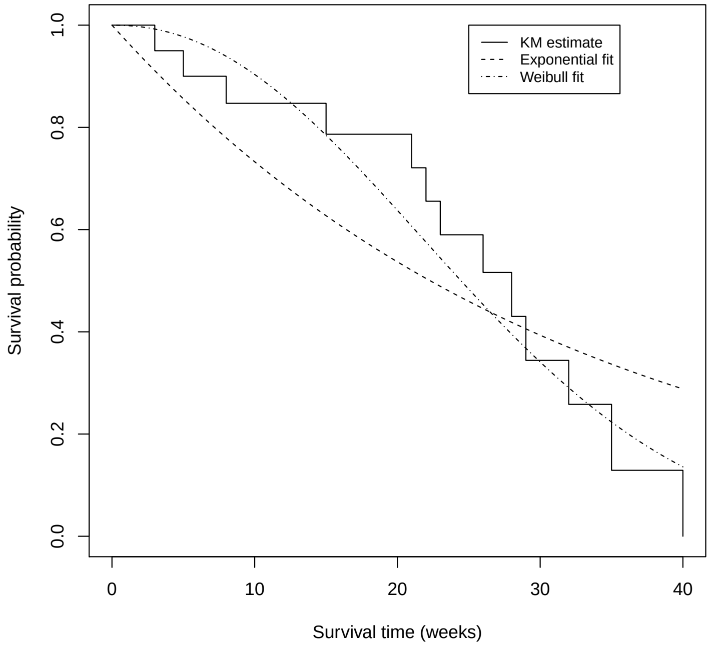
   (3) (10 points) Compare the curves and use log-rank test and weighted log-rank test to draw the final conclusions.

8. (20 points) The dataset `tempsurv.dat` contains a series of survival times. We can use nonparametric Kaplan-Meier method and also the parametric models (e.g., exponential model, Weibull model, etc.) in R.

```
library(survival)
example <- read.table("data/tempsurv.dat", header=TRUE)
# fit a Kaplan-Meier model
fit1 <- survfit(Surv(survtime, status)~1, data=example, conf.type="plain")
# plot the Kaplan-Meier curve
plot(0,0, type="n", xlim=c(0,40), ylim=c(0,1),
        xlab="Survival time (weeks)", ylab="Survival probability")
lines(fit1, conf.int="none", lty=1)

x <- seq(0, 40, by=0.5)
# Fit an exponential model
fit2 <- survreg(Surv(survtime, status)~1, data=example, dist="exponential")
lambda <- exp(-fit2$coef)
sx <- exp(-lambda * x)
lines(x, sx, lty=2)

# Fit a Weibull model
fit3 <- survreg(Surv(survtime, status)~1, data=example, dist="weibull")
lambda <- exp(-fit3$coef/fit3$scale)
alpha <- 1/fit3$scale
sx <- exp(-lambda * x^alpha)
lines(x, sx, lty=4)
legend(25,1, c("KM estimate", "Exponential fit", "Weibull fit"),
        lty=c(1,2,4), cex=0.8)
```



(1) (5 points) From the figure above, which model fits the data better? Exponential or Weibull? You can explain from both the theoretical and the observational perspective.

(2) (5 points) Here are the outputs for the two model fitting, which model is better? Why? Hint: use log likelihood-ratio test to check.

```
## Call:
## survreg(formula = Surv(survtime, status) ~ 1, data = example,
##     dist = "exponential")
##
## Coefficients:
## (Intercept)
##    3.470532
##
## Scale fixed at 1
##
## Loglik(model)= -58.1   Loglik(intercept only)= -58.1
## n= 20
```

```
## Call:
## survreg(formula = Surv(survtime, status) ~ 1, data = example,
##     dist = "weibull")
##
## Coefficients:
## (Intercept)
##     3.36717
##
## Scale= 0.4652515
##
## Loglik(model)= -54.1   Loglik(intercept only)= -54.1
## n= 20
```

(3) (5 points) You can also conduct the **Wald test** to check whether the data are from an exponential distribution.

(4) (5 points) Use **score test** to test whether or not the survival times are from an exponential distribution.

Solution: Suppose that our survival data are from a Weibull distribution with shape parameter λ and scale parameter α without censoring, the survival function $s(t) = e^{-\lambda t^{\alpha}}$. We need to construct a hypothesis

$$H_0 : \alpha = 1$$

i.e., the data are actually from an exponential distribution.

The likelihood function of (λ, α) is

$$\begin{aligned} L(\alpha, \lambda; \mathbf{t}) &= \textstyle\prod_{i=1}^{n} [\alpha \lambda t_i^{\alpha-1} e^{-\lambda t_i^{\alpha}}] \\ &= \lambda^n \alpha^n \exp\left[-\lambda \textstyle\sum_{i=1}^{n} t_i^{\alpha} + (\alpha - 1) \sum_{i=1}^{n} \log(t_i)\right] \end{aligned}$$

Therefore, the log-likelihood function of (α, λ) becomes

$$l(\alpha, \lambda; \mathbf{t}) = n \log(\alpha) + n \log(\lambda) - \lambda \sum_{i=1}^{n} t_i^{\alpha} + (\alpha - 1) \sum_{i=1}^{n} \log(t_i)$$

Then the scores are:

$$\begin{aligned} U_1(\alpha, \lambda) &= \tfrac{\partial l(\alpha, \lambda; \mathbf{t})}{\partial \alpha} \\ &= \tfrac{n}{\alpha} - \lambda \textstyle\sum_{i=1}^{n} t_i^{\alpha} \log(t_i) + \sum_{i=1}^{n} \log(t_i), \\ U_2(\alpha, \lambda) &= \tfrac{\partial l(\alpha, \lambda; \mathbf{t})}{\partial \lambda} \\ &= \tfrac{n}{\lambda} - \textstyle\sum_{i=1}^{n} t_i^{\alpha}, \end{aligned}$$

and the information matrix becomes

$$\begin{aligned} \tfrac{\partial^2 l(\alpha, \lambda; \mathbf{t})}{\partial \alpha^2} &= -\tfrac{n}{\alpha^2} - \lambda \textstyle\sum_{i=1}^{n} t_i^{\alpha} (\log(t_i))^2 \\ \tfrac{\partial^2 l(\alpha, \lambda; \mathbf{t})}{\partial \alpha \partial \lambda} &= -\textstyle\sum_{i=1}^{n} t_i^{\alpha} \log(t_i) \\ \tfrac{\partial^2 l(\alpha, \lambda; \mathbf{t})}{\partial \lambda^2} &= -\tfrac{n}{\lambda^2} \end{aligned}$$

For the data, we can calculate the above quantities under $H_0 : \alpha = 1$ and construct the score test and Wald test.

- **Wald test**:

  We can obtain the MLE $\hat{\alpha}$ and $\hat{\lambda}$

  $$\begin{aligned} \hat{\alpha} &= 1/\hat{\sigma} = 1/0.465 = 2.15 \\ \hat{\lambda} &= e^{-\hat{\beta}_0/\hat{\sigma}} = e^{-3.37/0.465} = 0.000719 \end{aligned}$$

  where $\hat{\beta}_0$ is the intercept of the fitted model, while $\hat{\sigma}$ is the `scale` in the output.

  Then we can compute the estimated information matrix $I_n(\hat{\alpha}, \hat{\lambda})$ and I_n^{α}, then we can compute the statistic:

  $$X^2 = (\hat{\alpha} - 1)^2 I_n^{\alpha} \sim \chi_1^2$$

  Reject the null hypothesis if $X^2 > \chi^2_{0.95,1}$.

- **Score test**:

  Under $H_0 : \alpha = 1$, the restricted MLE $\tilde{\lambda}$ can be obtained by:

  $$\tilde{\lambda} = e^{-\hat{\beta}_0} = e^{-3.47} = 0.0311$$

  With values of α and $\tilde{\lambda}$, we can get $\tilde{U}_1, \tilde{U}_2$ and $\tilde{I}_n$, therefore, we can get

  $$X^2 = \tilde{U}_1^2 / \tilde{I}_n^{11} \sim \chi_1^2$$

  Reject the null hypothesis if $X^2 > \chi^2_{0.95,1}$.

- **Likelihood ratio test**:

  Under $H_0 : \alpha = 1$, we can get

  $$X^2 = -2[l(\alpha = 1, \tilde{\lambda}; \mathbf{t}) - l(\hat{\alpha}, \hat{\lambda}; \mathbf{t})] \sim \chi_1^2$$

  Reject the null hypothesis if $X^2 > \chi^2_{0.95,1}$.