Exponential and Piecewise Exponential Distributions

Mai Zhou

Summary

For the purpose of learning *Survival Analysis*, we need to review some properties and some **extensions** of the exponential distribution. These extensions include the extreme value distribution, Weibull distribution and piece-wise exponential distribution.

Since lifetimes are almost always non-negative, the normal model/distribution may not be appropriate. An easy-to-use distribution is the exponential distribution. Exponential distribution to the survival analysis is sort of like normal distribution to the ANOVA.

Piecewise exponential distribution is the most flexible among the three, since we may have many pieces and thus many parameters. It is also used to bridge/connect the parametric and nonparametric method: when the number of pieces/parameters grows to infinite, the parametric model becomes a non-parametric model.

1 Exponential distribution, Extreme Value and Weibull Distribution

1. **Definition** If a random variable X satisfy

$$P(X > t) = P(X \geq t) = e^{-\lambda t}; \quad \text{for all} \quad t > 0,$$

then we say X is exponential, denote this by $X \sim \exp(\lambda)$.

In the above $\lambda > 0$ is a parameter (called hazard parameter). When $\lambda = 1$ we call X the standard exponential random variable. [forget about the mean parameter]

Exercise: How can you get a sample of iid (standard) exponential random variables from a sample of iid uniform(0,1) random variables?

Exercise: If X is $\exp(\lambda)$ then $c \times X, (c > 0)$ is [?].

3. **Definition** If X is $\exp(1)$ then $\log(X)$ is called a (standard) extreme value random variable/distribution.

$E \log(X) =$??.

Definition: Define a general extreme value random variable as: $\mu + \sigma \log(X)$. Here μ and σ are two parameters.

Notice μ is not the mean since mean of $\log X$ is not zero. Similarly σ is not the standard deviation. They are called location and scale parameters.

5. **Definition** If X is $\exp(1)$ then $X^{1/\beta}/\lambda$ is called a Weibull(λ, β) random variable.

Notice this definition is the same as Miller, Kalbfleish & Prentice, Cox & Oakes, ABGK but different from software R, Allison, and Klein & Moeschberger and some other text books. The difference lies in

the transformation

$$\frac{X^{1/\beta}}{\lambda} \quad \text{vs} \quad \left(\frac{X}{\lambda}\right)^{1/\beta} .$$

The two definitions are equivalent but when you use software or quote other books, be sure to check the definition.

In the above definition, $\beta > 0$ is called the shape parameter, and $\lambda > 0$ is called the scale parameter.

Exercise: Therefore, if X is exp(1) then $(X)^p$ is Weibull(?,?) for $(p > 0)$.

Therefore, if X is exp(1) then $\log[(cX)^p]$ is ..?...

If Y is Weibull(λ, β) then $\log Y$ is ..?....

2 Hazard and Cumulative Hazard function

To define the piecewise exponential distribution, we need to first define the hazard function.

Definition: The hazard function $h(t)$ (or sometimes denote by $\lambda(t)$), of a general r.v. Y (well, it needs to have density function $f(t)$), is

$$h(t) = \frac{f(t)}{1 - F(t-)} .$$

The minus in the denominator F is superfluous here but will be useful later when we talk about discrete r.v.s.

For X that is exp(λ), the hazard function $h(t)$ is a constant: $h(t) \equiv \lambda$, (this is memoryless, or forever young).

Definition: Cumulative hazard function $H(t)$ or $\Lambda(t)$ of a general r.v. Y is:

$$H(t) = \int_{-\infty}^{t} h(s)ds = \int_{-\infty}^{t} \frac{dF(s)}{1 - F(s-)} .$$

Notice the last expression is also valid even for discrete CDF $F(t)$. So, we can still have cumulative hazard function when the CDF of random variable Y is discrete.

The intuitive meaning of the hazard $h(t)$. We notice

$$h(t) = \lim_{\epsilon \downarrow 0} P(t \le Y < t + \epsilon | Y \ge t)/\epsilon$$

So $h(t)dt$ is a conditional probability.

2.1 The relation between density $f(t)$, cdf $F(t)$ and hazard $h(t)$, cumulative hazard $H(t)$

For continuous cdf $F(t)$, we have $H(t) = -\log[1 - F(t)]$. Thus $1 - F(t) = e^{-H(t)}$.

For (purely) discrete cdf $F(t)$ we can also **define** $H(t)$:

$$H(t) = \sum_{s \le t} \frac{\Delta F(s)}{1 - F(s-)} \ ;$$

where $\Delta F(t) = F(t+) - F(t-)$. Then we can show the following relation holds

$$1 - F(t) = \prod_{s \le t} (1 - \Delta H(s)) \ .$$

Exercise: If X_1 and X_2 are two independent exponential r.v.s (with parameter λ_1 and λ_2) then $\min(X_1, X_2)$ is also an exp r.v. (with parameter $\lambda_1 + \lambda_2$).

Generalizations. (If Y_1 and Y_2 are independent general r.v.s with hazard functions $h_1(t)$ and $h_2(t)$, then $\min(Y_1, Y_2)$ has hazard function .....)

(notice the two different ways of transform, both produce a new r.v.: a. transform on the r.v. b. transform on the hazard, or cumulative hazard.)

7. The expectation of order statistics from an independent exponential sample. (hint: Find the expectation of the minimum first).

2.2 Likelihood Function

Given an i.i.d. exp(λ) sample the MLE of the hazard/mean are .....

Do you know the distribution of the MLEs?

9. The Fisher information for λ in the sample is .....

10. In general, for any distribution, we have, based on an iid sample $x_1, \cdots x_n$ with density f, hazard h, cumulative hazard H,

$$\log lik(x_1, \cdots, x_n) = \sum_{i=1}^{n} [\log h(x_i) - H(x_i)] \quad = \sum_{i=1}^{n} \log f(x_i)$$

Generalize this to right censored case.

If (Y_i, δ_i) $i = 1, \cdots, n$ is a sample of right censored observations from a general distribution, ($Y_i = \min(X_i, C_i)$, with X_i iid from density f, hazard h, cumulative hazard H) the log likelihood is

$$\log lik(Y_1, \delta_1 \cdots Y_n \delta_n) = \sum_{i=1}^{n} [\delta_i \log h(Y_i) - H(Y_i)] \ .$$

In the above, if X_i is exponential, then h and H simplify to ...

3 Piecewise Exponential random variable

Definition: If a random variable Y's hazard function, $h_Y(t)$, is a piecewise constant function, then Y is called a piecewise exponential random variable. We suppose the boundary or the cut points of the pieces are given.

Exercise: how to estimate the parameters from a sample of piecewise exponential observations?

As an example a three piece (piecewise) exponential r.v. with cut points $0 < T_1 < T_2 < T_3 = \infty$ has hazard function

$$h(t) = \lambda_1 I[0 \le t < T_1] + \lambda_2 I[T_1 \le t < T_2] + \lambda_3 I[T_2 \le t] \ .$$

Its cumulative hazard is $H(t) =$??

The log likelihood function based on an iid sample of X_i from a piecewise exponential distribution can be written down. This is much easier using hazard function.

Theorem Based on a sample of n iid observations from a piecewise exponential distribution, the MLE of hazard of the i^{th} piece (for the interval $[T_{i-1}, T_i)$) is

$$\hat{\lambda}_i = \frac{\#\{x_j \in [T_{i-1}, T_i)\}}{\sum_{j=1}^n [\min(T_i, x_j) - T_{i-1}]^+} \quad \text{where} \quad [t]^+ = \max(0, t) \ .$$

The denominator in the above is identical to $\sum_j (x_j - T_{i-1}) I_{[T_{i-1} \le x_j \le T_i]} + \sum_j (T_i - T_{i-1}) I_{[x_j \ge T_i]}$.

9.5. Re-work the above theorem with right censored observations.

For right censored samples, the only modification is in the numerator: $\#\{\text{uncensored } x_j \in [T_{i-1}, T_i)\}$.

Define $R(t) = \sum I_{x_i \ge t}$ then the denominator above is approximately equal to $R(T_{i-1})(T_i - T_{i-1})$.

You see the Nelson-Aalen estimator here.

Theorem The (approximate) variance of the MLE in the above theorem $\hat{\lambda}_i$ is (by using Fisher information for MLE theory)

$$\frac{\hat{\lambda}_i^2}{\sum_j \delta_j I_{[T_{i-1} \le x_j < T_i]}} \ .$$

and $\hat{\lambda}_i$ is asymptotically independent of $\hat{\lambda}_j$ for $i \ne j$.

This theorem can be proved by standard calculation.

Notice the estimator $\hat{\lambda}_i$ for different i (or different pieces) can be considered independent, at least asymptotically. It is easy to verify that the information matrix for $\hat{\lambda}_1, \cdots, \hat{\lambda}_k$ is diagonal, therefore the

(approximate) variance-covariance matrix is also diagonal.

Although this is obtained under the assumption of piecewise exponential population, it is in fact true for a random sample from any distribution, asymptotically. [since a piecewise exponential can approximate any distribution, sort of] And it is in fact un-correlated for finite samples.

From here you get the Variance estimator for the Nelson-Aalen estimator.

Because the Nelson-Aalen estimator, $\hat{H}(t) = \hat{\Lambda}(t)$, is (approx.) equal to $\sum_{T \le t} \hat{\lambda}_i (T_i - T_{i-1})$. The variance of a sum is the sum of the variances (uncorrelated terms) and (please verify)

$$Var(\hat{\lambda}_i (T_i - T_{i-1})) \approx \frac{\#\{x_j \in [T_{i-1}, T_i)\}}{R(T_{i-1})^2}$$

Exercise: Given any continuous r.v. X with cdf $F(t)$, what transformation convert it into exp(1)?

Exercise: Any continuous r.v. can be thought of as obtained by a transformation of an exponential r.v.

10c. Any positive (continuous?) r.v. can be viewed as an exp() r.v. but with time-changing hazard rate. (instead of constant hazard) Crazy Clock!

11. The best rank test for testing the equality of 2 samples of exp data ..... Savage test.

4 Weibull Regression Models

12. Exponential regression model, MLE method.

Model Assumption (postulates) X_i are independent and

$$X_i \sim \exp(\lambda_i) \qquad \text{where} \qquad \log(\lambda_i) = \alpha + \beta z_i \quad i = 1, 2, \cdots n$$

We observe a sample of (X_i, z_i) and need to estimate α, β. WHere X_i are the survival times, z_i the covariate(s).

Let $Z_i \sim \exp(1)$, then $X_i \sim \frac{Z_i}{\lambda_i}$.

Equivalent/alternative modeling with log transformation and extreme value distribution.

$$\log X_i = -\log \lambda_i + \log Z_i = -(\alpha + \beta z_i) + \log Z_i$$

where $\log Z_i = \epsilon_i$ is standard extreme value r.v.

13. Generalization: Weibull regression i.e. $X_i \sim Weibull(b, \lambda_i)$. This can be achieved by adding a scale parameter in the above extreme value regression.

$$\log X_i = -\log \lambda_i + \sigma \log Z_i = -(\alpha + \beta z_i) + \sigma \log Z_i$$

We note $\sigma = 1/b$.

14. The estimation of α, β and σ can be obtained by the MLE methods, using numerical maximization. This can all be done in SAS or R. (code examples?)

15. Generalization (project): replace Z_i above by a piecewise exponential random variable. (problem: there is no 'standard' piecewise exponential; and we might as well take/replace (the exp of) $-\alpha + \sigma \log Z_i$ as a piecewise exponential) (This is a project topic) How MLE/LR test of β works out here?

Please note the two different way of viewing the Weibull regression model: (1) after taking the log, it is extreme value location regression. (2) without take transformation it is the scale parameter in the weibull distribution modeled as regression.

5 Exponential random variables and Poisson process

16. The relation between exp(λ) distributions and a Poisson process. (should be covered in sta624, the stochastic process)

17. Notations of the Poisson process: $N_\lambda(t)$.

Let

$$N_\lambda(t) - \lambda \times t = N_\lambda(t) - \int_0^t \lambda ds = M(t) \ .$$

The expectation of $M(t)$ is zero for any t, in fact, $M(t)$ is a so called *martingale*.

Some random Problems

0) Plot the hazard function $h(t)$ for

a. $N(\mu, \sigma)$ distributions with several different μ's and σ's.

b. log-normal distributions with several location/scale parameters.

c. Gamma distributions.

1). Use a discrete distribution (with 5-point mass) to verify the discrete formula connecting the CDF (F) and cumulative hazard function (H).

2). Work out the "?"s and "$\cdots$"s in the handout (on first page).

3). In computing the sum of the Savage scores, we can do a few change:

(a). change the scores to centered version:

$$a_{ni} = 1 - (\frac{1}{N} + \cdots + \frac{1}{N-i+1}).$$

(b). Originally we need to sum those scores correspond to sample 1. We could sum this differently as follows: when the score a_{ni} correspond to a sample 1 obs. we sum $1 - R_1 \times (\frac{1}{N-i+1})$; when the score a_{nj} correspond to a sample 2 obs. we sum $-R_1 \times (\frac{1}{N-j+1})$. Convince yourself that the resulting sum will be the same.

3). Write the sum as $\int (dN_1(t) - R_1(t) \cdot (\frac{d[N_1(t)+N_2(t)]}{R_1(t)+R_2(t)})$

Why proportional (cumulative) hazard is a point-wise property and shift model in probability is an interval property.

To compare $H_1(t_0)$ to $H_2(t_0)$, we can immediately find, for example, $H_1(t_0) = \eta * H_2(t_0)$. here η is a parameter. This only involve the H1 and H2 value at one point: t0

To model F1(t0) shift to F2(t0), if we use the model $F_1(t_0) = F_2(t_0 - \eta)$, then this not only involve the F2(t0), it involves all F2 values from $F_2(t_0)$ to $F_2(t_0 - \eta)$

6 Calculations related to Greenwood formula

For the right censored data: $(X_1, \delta_1), (X_2, \delta_2), \cdots, (X_n, \delta_n)$ define

$$N(t) = \sum_{i=1}^{n} \delta_i I_{[x_i \leq t]}$$

This counts the number of observed failures up to time t, whereas $dN(t) = \Delta N(t) =$ number of observed failures at time t.

Define also

$$R(t) = \sum_{i=1}^{n} I_{[x_i \geq t]}$$

This counts the number of subjects alive at time t. In medical term, those 'at risk' at time t.

Nelson-Aalen estimator is

$$\hat{\Lambda}(t) = \sum_{s \leq t} \frac{dN(s)}{R(s)}$$

We also have (from the computation of information with piecewise exponential random variables)

$$Var(\hat{\Lambda}(t)) \approx \sum_{s \leq t} \frac{dN(s)}{R(s)^2}.$$

The Kaplan-Meier estimator is

$$1 - \hat{F}(t) = \prod_{s \leq t} \left(1 - \frac{dN(s)}{R(s)}\right).$$

The variance of Kaplan-Meier estimator (Greenwood formula) can be derived as follows.

$$Var(\hat{F}(t)) = Var(1 - \hat{F}(t)) \approx Var(e^{-\hat{\Lambda}(t)})$$

since for continuous CDF we have $1 - F(t) = e^{-\Lambda(t)}$. Using the delta method, we have

$$Var(e^{-\hat{\Lambda}(t)}) \approx [e^{-\hat{\Lambda}(t)}]^2 Var(\hat{\Lambda}(t))$$

Finally, we get the Greenwood formula by replace $e^{-\hat{\Lambda}(t)}$ back with $1 - \hat{F}(t)$, and variance estimator of Nelson-Aalen

$$Var(\hat{F}(t)) \approx [1 - \hat{F}(t)]^2 \sum_{s \leq t} \frac{dN(s)}{R(s)[R(s) - dN(s)]}$$

In the denominator we use $R(R - dN)$ instead of R^2. This can be explained with binomial variance, with no censoring.

7 Some notes related to Cox model

Given k independent exponential random variables X_i with hazards $\lambda_1, \cdots, \lambda_k$.

We have

$$P(the\ failed\ one\ is\ X_i | one\ failed\ out\ of\ k\ X's) = \frac{\lambda_i}{\sum_{j=1}^k \lambda_j} \tag{1}$$

i.e.

$$P(X_i = t | \min_{1 \le j \le k} X_j = t) = \frac{\lambda_i}{\sum_{j=1}^k \lambda_j}$$

Now imaging the random variables X_i carry with them a value z_i, the covariate.

Then the conditional expectation of z value for the failed one, given there is one fail among the k is

$$\sum_{i=1}^k \frac{z_i \lambda_i}{\sum_{j=1}^k \lambda_j}$$

Or

$$\frac{\sum_{i=1}^k z_i \lambda_i}{\sum_{j=1}^k \lambda_j}.$$

If we identify λ_i with $\exp(\beta z_i)$, this is the term in the Partial likelihood.

$$\frac{\sum_{i=1}^k z_i \exp(\beta z_i)}{\sum_{j=1}^k \exp(\beta z_j)}.$$

So, at least the score function of the Cox partial likelihood has mean zero. (each term as mean zero, at true β.) Because it is in a form of 'observed' - 'expected'.

Also, using this we see the Cox partial likelihood is the product of many conditional probabilities, as in (1). So it is not a 'real' likelihood.

8 Notes on the implementation of piecewise exponential regression:

The example in Allison's book, he cut the time interval for each obs. T_i into pieces defined by intervals $a0 < a1 < a2 < ...$

For a genuine piecewise exponential regression, we should cut the time interval for the error variable, Z_i or $\log Z_i$, not the responses T_i.

AFT model: $T_i = exp(-\beta x_i)Z_i$

where Z_i is a standard exp, or exp(1) r.v. After taking log, we have

$$\log(T_i) = -\beta x_i + \log Z_i$$

To model $\log Z_i$ or Z_i as iid piecewise exponential we should divide the time interval for the variable Z_i to $0 = a0 < a1 < a2....$

These cut intervals in terms of the variable T_i is different:

So, for T_i we need to cut (denote $\phi_i = exp(-\beta x_i)$)

$$\phi_i a0 < \phi_i a1 < \phi_i a2 < \phi_i a3 < ...$$

If $a0 = 0$ then the first is always zero.

This (β dependent cut) could be done iteratively. [first estimate β, then refine the cut, ie. it lead to new cut, then update the β estimate with this new cut,....]

This way every subject has its own set of intervals.

This should lead to the AFT model with general error distribution. at least the estimator of beta should be comparable.

The hazard of AFT model: the hazard for T_i is seen given as

$$h_i(t) = h_0(t\phi_i)\phi_i$$

Question:

(1) understand what SAS is doing in the example.

(2) How do you implement the piecewise exp regression in R?

(3) use a cut that is different for each subject and is dependent on βx_i.

Can R function survreg() take (start, stop, event) type input? No.

9 Summary of likelihood based inference results

First we take the example of the Weibull distribution introduced earlier. How should one estimate the two parameters of Weibull distribution given a sample of iid observations? The answer is by the MLE. But unfortunately, the MLE of the shape parameter is not explicit, and can only be obtained by numerical methods.

The variance/covariance of the MLE is also hard to compute explicitly. But again the MLE theory told us the approximate method.

Suppose $X_1, X_2, \cdots, X_n$ are independent and $X_i \sim f(x_i|\theta)$ where $\theta \in \Theta$ is the parameter, when $\theta \in \Theta$ it defines a family of densitie functions $f(\cdot|\theta)$. The log likelihood function is

$$l_n(\theta) = \log Lik(\theta) = \sum_{i=1}^{n} \log f(X_i|\theta)$$

The θ value that maximize the $l_n(\theta)$ is the MLE, denoted by $\hat{\theta}_{MLE}$.

The MLE has the so called 'invariance property'.

Under mild regularity conditions (something like the two/three derivatives of the density exists and continuous etc.), we have, for large n

$$(\hat{\theta}_{MLE} - \theta_0) \approx N(0, \frac{1}{I(\theta)})$$

where $I(\theta)$ is the so called (expected) Fisher information, and θ_0 is the true value of the parameter. Based on this we can construct an approximate 95% confidence interval

$$\hat{\theta}_{MLE} \pm 1.96 \frac{1}{\sqrt{\hat{I}(\hat{\theta}_{MLE})}}$$

where $\hat{I}(\hat{\theta})$ is the so called observed Fisher information. For θ that are vectors, the similar thing also holds. There the Fisher information will be a matrix, and reciprocal will be inverse of the matrix, etc.

The above confidence intervals are called Wald confidence intervals. The Wald confidence intervals *DO NOT* have 'invariance property'. [due to what?]

There is also another way of constructing the confidence intervals via the equivalency of the confidence interval and testing hypothesis.

The likelihood ratio test for $H_0 : \theta = \theta_0$ vs. $H_A; \theta \neq \theta_0$ can be based on the

$$-2[l_n(\theta_0) - l_n(\hat{\theta}_{MLE})] = W$$

The rejection rule is to reject H_0 if the W is too large. (use chi square table to find the critical value).

This also leads to confidence intervals/regions if we invert the test. These confidence intervals are called Wilks confidence intervals. The Wilks confidence intervals *DO* have 'invariance property'.

An interesting relation: for large n,

$$-2[l_n(\theta_0) - l_n(\hat{\theta}_{MLE})] \approx \frac{(\hat{\theta}_{MLE} - \theta_0)^2}{\hat{I}(\hat{\theta}_{MLE})}$$

Notice the left hand side do not explicitly involve $I(\theta)$. *This can be a real advantage*, especially when the Fisher information is hard to calculate, or invert.

Expected information:

$$I_F = E\left(\frac{\partial}{\partial\theta}\log Lik(\theta, x)\right)^2$$

It is non random but is a function of θ.

Observed information:

$$\hat{I} = -\frac{\partial^2}{(\partial\theta)^2}\log Lik(\theta, x)|_{\theta=\hat{\theta}_{MLE}}$$

It is random but do not involve θ (at least not explicitly).

When we dealing with multi-dim parameters, we have similar results.

More complicated for nuisance parameter, parameter of interest etc.