Time to Event Analysis

Alina Levine

1 The Problem

You conduct a 5 year 500 person clinical trial to compare the effectiveness of two drugs in prolonging life. By the time your study ends, 20 people who completed the trial are still alive and 10 people withdrew from the study before it even ended (you assume that the reason people left the trial had nothing to do with the path of their disease). Therefore, you don't know how long some people lived, and you don't know how long some people will live. However, you do know something about these people. You know that their survival time must at least be as large as the time until they left the study [whether that be early departure or trial completion]. You need a way to analyze your data that takes this information into account. You hope to answer the following questions:

What's the chance that a person treated with either drug is still alive after five years? How long can someone treated with these drugs expect to survive, and maybe most importantly, is one of these drugs superior in its ability to provide patients with more time? These are the questions survival analysis answers.

2 The Models

The following models are used to describe survival

2.1 Parametric

Survival time is assumed to be distributed according to a known distribution. The parameters are the only missing components. Parametric estimators tend to be more efficient than non-parametric or semi-parametric estimators, especially with smaller sample sizes. However, parametric models can lead to inaccurate estimates because of the assumptions may not be true.

2.1.1 Exponential

The exponential distribution has the memoryless property, so by using the exponential distribution, you are assuming that a person's probability of dying in the next x time units, is independent of how long that person has already survived. This is also called the ageless property.

The model is T=$e^{\beta_0+\beta_1*x}\theta$, and $\theta \sim exp(1)$. Thus, T is exp($1/e^{\beta_0+\beta_1 x}$) (ie: the pdf of T). Let S(t) be the probability a person lives at least t time units.

$$S(t) = P(T \geq t) = e^{e^{-(\beta_0+\beta_1 x)}t} \tag{1.1}$$

The MLE method can be used to estimate β_0 and β_1 from the sample. The math works out better if you find the MLE of the natural log of time, which we will call y. If you take the natural log of T, you get $ln(T) = Y = \beta_0 + \beta_1 x + \mathcal{E}$, where $\epsilon = ln(\theta)$. ϵ has an extreme minimum value distribution with parameter one and mean 0 and variance 1 denoted $\epsilon \sim G(0, 1)$. Thus $Y \sim G(\beta_0 + \beta_1 x, 1)$ which has an expectation of $\beta_0 + \beta_1 x$.

To find the mle we must first find the likelihood function for Y, which is ln(T):

$$l(\beta) = \prod_{i=1}^{n} f(y_i, \beta, x_i)^{c_i} S(y_i, \beta, x_i)^{1-c_i} \tag{1.2}$$

Taking the natural log you get:

$$L(\beta) = \sum_{i=1}^{n} c_i ln(f(y_i, \beta, x_i)) + (1 - c_i) ln(S(y_i, \beta, x_i)) \tag{1.3}$$

Now we have to plug in the PDF of Y and survival function of Y. $Y \sim G(\beta_0 + \beta_1 x, 1)$. The minimum extreme value distribution has the following pdf.

$$f(x) = \frac{1}{\sigma^2} e^{\frac{x-\mu}{\sigma^2}} (-e^{e^{\frac{x-\mu}{\sigma^2}}}) \tag{1.4}$$

Therefore,

$$f(y, \beta, x) = e^{y-\beta_0+\beta_1 x} exp(-e^{y-\beta_0+\beta_1 x}) = e^{(y-\beta_0+\beta_1 x)-e^{y-\beta_0+\beta_1 x}} \tag{1.5}$$

The minimum extreme value distribution has the following surival function, which is just 1 minus the cdf.

$$S(x) = exp(-e^{\frac{x-\mu}{\sigma^2}}) \tag{1.6}$$

Plugging in the parameters for Y, $S(y)$ is the following

$$S(y) = exp(-e^{x-\beta_0+\beta_1 x}) \tag{1.8}$$

After plugging in, the log-likelihood of β becomes:

$$L(\beta) = \sum_{i=1}^{n} c_i ln(e^{(y_i-(\beta_0+\beta_1 x))-exp(y_i-(\beta_0+\beta_1 x))}) + (1 - c_i) ln(exp(-e^{y_i-(\beta_0+\beta_1 x)}))$$

Simplified, the log-likelood is

$$L(\beta) = \sum_{i=1}^{n} c_i (y_i - (\beta_0 + \beta_1 x_i)) - e^{y_i-(\beta_0+\beta_1 x_i)} \tag{1.9}$$

To get the MLE of β_1 and β_0, we must then take the derivatives of the log likelihood with respect to those variables. β_0 and β_1. β_0 and β_1 can then be used to predict t, using $\hat{t} = e^{\hat{\beta_0}+\hat{\beta_1} x}$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} -c_i + e^{y_i-(\beta_0+\beta_1 X_i)} \tag{1.10}$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} -c_i x_i - x_i e^{y_i-(\beta_0+\beta_1 x_i)} \tag{1.11}$$

This system of equations can be solved numerically.

The asymptotic normality of the MLE can be used to construct confidence intervals around $\hat{\beta}$ and conduct hypothesis tests for β. Asymptotic normality of the MLE, and in our case, $\hat{\beta}$ given the fact that $\hat{\beta}$ is the mle implies the following:

$$\hat{\beta} \approx N(\beta, \frac{1}{I(\beta)})$$

To use this information to find a confidence interval set up the following inequality (in this example a 95 percent confidence interval. It uses the estimator for the variance of an mle, $\frac{1}{I(\hat{\beta})}$

$$-1.96 \leq \frac{\hat{\beta} - \beta}{\frac{1}{\sqrt{I(\hat{\beta})}}} \geq 1.96$$

.

After multiplying each term by $\sqrt{I(\hat{\beta})}$, subtracting $\hat{\beta}$, and multiplying by -1, the confidence interval for β becomes:

$$\hat{\beta} \pm \hat{\beta} \sqrt{I(\hat{\beta})^{-1}} 1.96 \tag{1.12}$$

To conduct a hypothesis test, we can use the Wald Test for each parameter. If we are testing β_1, our null hypothesis is that the variable does not have an effect on survival time, meaning $\beta_1 = 0$. The test statistic is just the z-score of the $\hat{\beta}$ given β is 0. $z = \frac{\hat{\beta_1}}{\frac{1}{\sqrt{I(\hat{\beta})}}}$

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

2.1.2 Weibull

The exponential method can be altered by adding shape parameters. Earlier we said that $T = e^{\beta_0+\beta_1 x}\theta$, where $\theta \sim exp(1)$. Now we are changing this to $T = e^{\beta_0+\beta_1 x}x\theta^{\sigma}$ T is no longer exponentially distributed, but rather has a Weibull distribution with parameters $e^{\beta_0+\beta_1 x}$ and σ

In the exponential case, in order to find the MLE of of t, we took the natural log of t, which we called y, and got that $y = \beta_0 + \beta_1 x + \epsilon$. Now $y = \beta_0 + \beta_1 x + \sigma\epsilon$ and instead of the second term having an extreme minimum value distribution with mean=0 and variance=1, G(0,1), the variance now is σ, G(0, σ). Therefore, $Y \sim G(\beta_0 + \beta_1 x, \sigma)$ To get the pdf of Y, we can just plug the parameters into the pdf for the extreme minimum value distribution:

$$f(y, \beta, x, \sigma) = \frac{1}{\sigma} e^{\frac{y-\beta_0+\beta_1 x}{\sigma} - exp(\frac{y-\beta_0+\beta_1 x}{\sigma})} \tag{1.13}$$

and the survival function can be shown to be:

$$S(y, \beta, x, \sigma) = exp(-e^{\frac{y-(\beta_0+\beta_1 x)}{\sigma}}) \tag{1.14}$$

Plugging these into the log-likelihood, for which the general equation is:

$$L(\beta_0, \beta_1, \sigma) = \sum_{i=1}^{n} c_i ln(f(y_i, \beta, x_i)) + (1 - c_i) ln(S(y_i, \beta, x_i))$$

we get:

$$L(\beta_0, \beta_1, \sigma) = \sum_{i=1}^{n} c_i ln(\frac{1}{\sigma} e^{\frac{y_i-(\beta_0+\beta_1 x)}{\sigma} - exp(\frac{y_i-\beta_0+\beta_1 x}{\sigma})}) + (1 - c_i) ln(exp(-e^{\frac{y_i-(\beta_0+\beta_1 x)}{\sigma}}))$$

Simplified, the likelihood is

$$L(\beta_0, \beta_1, \sigma) = \sum_{i=1}^{n} c_i(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}) - e^{\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}} - c_i ln(\sigma) \tag{1.15}$$

. Taking the derivatives with respect to β_0 and β_1 you get the following system:

$$\frac{\partial L(\beta_0, \beta_1, \sigma)}{\partial \beta_0} = \sum_{i=1}^{n} -\frac{c_i}{\sigma} + \frac{1}{\sigma} e^{\frac{y_i-(\beta_0+\beta_1 X_i)}{\sigma}} \tag{1.16}$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} -\frac{c_i x_i}{\sigma} - \frac{1}{\sigma} x_i e^{\frac{y_i-(\beta_0+\beta_1 x_i)}{\sigma}} \tag{1.17}$$

$$\frac{\partial L(\beta)}{\partial \beta_0} = \sum_{i=1}^{n} \frac{1}{2} \frac{c_i(y_i - (\beta_0 + \beta_1 X_i))}{\sigma^2} - \frac{c_i}{\sigma} \tag{1.18}$$

These equations can be estimated numerically

2.1.3 Choosing the right model

In the Weibull modell, there is an extra parameter as compared to the exponential model. A likelihood ratio test can be used to determine whether the Weibull exponential model should be used to describe the data.

See this writeup for more information.

2.2 Non-Parametric

Non-parametric models do not run the risk of making misinformed assumptions. However, because there are not set parameters and there is not a set distribution, the calculations can be tedious if there is a large sample. Additionally, non-parametric methods are not suited for figuring out how many different variables affect survival and are not useful for figuring out how quantitative variables affect survival.

See this writeup for more information.

2.2.1 Kaplan Meir Estimations

The Kaplan Meier (KM) estimator is an estimator for the survival function. The KM estimator is $\hat{S}(t)$, is a point estimate for probability of living past time t. The KM estimate is known for every t from t=0 until the last time of death. It is a step function that changes value at each unique time of death. The estimates can be calculated through the multiplication of conditional probabilities.

The KMr estimate at t_i is the probability of a participant from the study living to t_1, which is a death time, multiplied by the probability of a participant living to the next death time given that that person survived past t_0 and so on until the conditional probability of living past t_i is reached. Censoring plays a role by lowering the number of people at risk of dying each time someone is censored, but the effect is not seen until the next time of death. When the next person dies after an interval in which censoring occurred, the drop in survival probability ends up being more drastic than if no censoring in that interval had occurred, as a larger proportion of the risk set has died when the risk set is smaller.

The estimator is usually written in the following way

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i} \tag{2.1}$$

where n_i is the number of people at risk of dying in the ith time interval and d_i is the number of people who died at time t_i. t_1, t_2, t_3... are the ranked times of death from smallest to largest.

Notice that if there were no censoring the Kaplan Meier estimator would be an empirical function estimator. Due to censoring, the probability of living past a certain time is not just the number of people still living divided by n, as the risk set decreases each time someone is censored.

If at least one person survives the total length of the study, the final multiplier will not be zero, since the number of people who died will never be the same as the number of people at risk in any interval. Therefore, the KM estimate will not reach zero. In order for the km estimator to reach zero, every participant must die.

Kaplan Meier estimators are asymptotically normal. We can use this fact to construct confidence intervals around $\hat{S}(t)$. The log-log survivorship function is the function most often used to find the confidence intervals for $S(t)$, when usingThe log-log survivorship function is the function most often used to find the confidence intervals for $S(t)$, when using the Kaplan Meier estimators, as finding the confidence intervals directly can sometimes lead to negative bounds or bounds greater than one, impossible for probabilities. This transformation is allowed because of the invariance principle of the mle. Since $\hat{S}(t)$ is an mle and is asymptotically normal, log(-$log(\hat{S}(t))$ is an mle and is asymptotically normal Therefore, an estimator for the variance of log(-$log(\hat{S}(t))$ needed. To get this estimator, you can use the delta method twice, using the fact that the $Var(ln(X)) \approx \frac{1}{\mu_X}\sigma_X^2$.T the Kaplan Meier estimators, as finding the confidence intervals directly can sometimes lead to negative bounds or bounds greater than one, impossible for probabilities. This transformation is allowed because of the invariance principle of the mle. Since $\hat{S}(t)$ is an mle and is

asymptotically normal, log(-$log(\hat{S}(t))$ is an mle and is asymptotically normal Therefore, an estimator for the variance of log(-$log(\hat{S}(t))$ needed. To get this estimator, you can use the delta method twice, using the fact that the $Var(ln(X)) \approx \frac{1}{\mu_X}\sigma_X^2$.T

$$\hat{Var}(ln(-ln(\hat{S}(t))) = \frac{1}{(ln(\hat{S}(t)))^2}\sum_{t_i<t}\frac{d_i}{n_i(n_i-d_i)}$$

Therefore, the confidence interval for $ln(-ln(S(t))$ is the following:

$$ln(-ln(\hat{S}(t)))-z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{(ln(\hat{S}(t)))^2}\sum_{t_i<t}\frac{d_i}{n_i(n_i-d_i)}} \leq S(t) \leq ln(-ln(\hat{S}(t)))+z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{(ln(\hat{S}(t)))^2}\sum_{t_i<t}\frac{d_i}{n_i(n_i-d_i)}} \tag{2.2}$$

To get the confidence interval for S(t), just take the exponential of the negative exponential of the interval.

$$exp(-exp(ln(-ln(\hat{S}(t))) - z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{(ln(\hat{S}(t)))^2}\sum_{t_i<t}\frac{d_i}{n_i(n_i-d_i)}})) \leq$$
$$S(t) \leq exp(-exp(ln(-ln(\hat{S}(t))) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{1}{(ln(\hat{S}(t)))^2}\sum_{t_i<t}\frac{d_i}{n_i(n_i-d_i)}})) \tag{1}$$

Kaplan Meier estimators are used to construct Kaplan Meier curves, which allow people to visualize estimated survival probabilities. The x axis represents time and the y axis represents the Kaplain Meier estimate of survival. Estimates for quantities such as the median survival time can be pulled just by looking at the curve. The median would be the smallest time at which the Kaplan Meier estimate is .5.

Kaplan Meier analysis is used to create descriptive statistics and visualizations of survival estimates. If you want to compare survival curves non-parametrically, you would use something along the lines of a log-rank test, which will be discussed later.

2.2.2 Deficiencies in Kaplan Meier Method

Even within the realm of non-parametric modeling, the Kaplan Meier Estimator has some deficiencies. One of these deficiencies is in its inability to use all the survival time information available. As noted before, between times of death, the Kaplan Meier estimate stays the same. It is the number of patients who are censored in between two times of death that determines how small the Kaplan Meier estimate will be at the next time of death. How far along in the interval someone is censored is completely ignored.Let's say someone died at t=2 and the next death occurred at t=8. Shouldn't survival prospects be slightly higher if the person who was censored dropped out at t=7 rather than at t=3? Valuable information is thrown away. Of course, this information would not be thrown away if censoring was just considered a death. However, lowering the probability of survival after anybody left the study would represent survival as being much lower than it already is. With km estimators, a person's getting censored by dropping out of a study early cannot be due to disease progression, so therefore, a person who is censored may die much later than the time of censoring. Automatically, saying that person died when he or she left the study is wrong. Another issue is that as time goes on, the sample gets smaller, so estimates later on in time have greater uncertainty. For these reasons, it is a good idea to include confidence bands on Kaplan Meier plots.

2.2.3 Non-Parametric Survival Comparison

ted Without making any assumptions about the survival time distribution, how can you say which treatment option will increase survival? A significance test is needed tha

Often times one will look at two Kaplan Meier curves, maybe the curves for two different drugs, and will see that one lies entirely above the other. However, this could just be due to chance, so significance tests must be conduct can take differences into account at many different points in time. The rank sum tests all

do this, although they differ in which differences carry more significance than others. The test statistic Z is asymptotically distributed $N(0,1)$

http://web.stanford.edu/ lutian/coursepdf/unitweek3.pdf

$$z = \frac{\sum_{i=1}^{m} w_i(d_{1i} - \hat{e}_{1i})}{\sqrt{\sum_{i=1}^{m} w_i^2 \hat{v}_{1i}}} \tag{2.3}$$

Each i in this statistic represents a time of death.Each term is taking the difference between the number of people who died in group one d_{i1}, which can be either of the groups, in the time period between t_{i-1} and t_i inclusive of t_i and the expected number of people would have died in group 1 if both groups had the same risk of death. This expected value, $\hat{e}_{1i}$, is $\frac{d_i}{n_i} n_{1i}$, where d_i is the total number of people who died in this time period and n_{1i} is the number of people at risk of dying in group 1 at the beginning of this time period.
$\hat{e}_{1i} = \frac{d_i}{n_i} n_1 i$, where n_i is the total number of people at risk at t_i in both groups and d_i is the total number of people who died at t_i.
The variance estimator of $d_1 i$, $\hat{v}_{1i}$, is $\frac{n_{1i} n_{0i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}$, where n_i is the total number of people at risk of dying at the beginning of the time period.

It is the weight coefficient, w_i that decides which survival differences will be taken most into account. For example, in the **Log Rank** test, the coefficient is 1, but in the **Wilcoxon** test, the w_i is the total number of people at risk of dying at t_i. Therefore, this test is going to have more power, when the differences are concentrated in the earlier deaths. The Wilcoxon test also requires that the hazard ratio between the two groups be constant. We look at hazard ratio later

2.3 Semi-Parametric

Semi-Parametric models are helpful because unlike a parametric model, you do not need to assume that survival time follows a specific statistical distribution. Instead, there is a non-parametric part, which is the baseline hazard and a parametric part which is e raised to a linear combination of each co-variable. In many cases with semi-parametric models, the non-parametric part does not need to be computed in order to compare survival or hazard, making these models useful.

2.3.1 Hazard

Before going into detail about semi-parametric models, the hazard function must be described. Hazard describes your chances of Adying in the next instant given that you survived up until that instant. It is not actually a probability because your chance of dying at a certain instant is actually zero and hazard, unlike probability, can be greater than one.

It is actually the instantaneous rate of deaths at a certain point in time. Hazard can be written as the following limit:

$$h(t) := \lim_{dt \to 0} \frac{P(t < T < t + dt | T > t)}{dt} = \frac{f(t)}{S(t)} \tag{3.1}$$

That is, it can also be represented as a ratio of the lifetime density function and survival function.

Using this equation we can get S(t) as a function of only hazard, which will be helpful later on. Integrating both sides of the equation from 0 to t, you get $\int_0^t h(t)dt = -ln(S(t))$, and then exponentiating both sides you get

$$S(t) = e^{-H(t)} \tag{3.2}$$

where H(t)=$\int_0^t h(t)dt$ and is called the "cumulative hazard".

Hazard can tell you whether aging occurs or not. Aging occurs when your chance of survival decreases as time goes on. A hazard function that has no time variable, signifies that the length of time a person has

survived does affect affect their chances of dying. The case where survival time is exponentially distributed would fall into the category of constant hazards

$$h(t) = e^{-(\beta_0+\beta_1 x)} \tag{3.3}$$

(note that $h(t)$ does not vary at all with t). On the other hand, when time has a Weibull distribution as in Section x,

$$h(t) = \frac{\frac{1}{\sigma} t^{\frac{1}{\sigma}-1}}{(e^{\beta_0+\beta_1 x})^{\frac{1}{\sigma}}} \tag{3.4}$$

and hazard does depend on time, so aging is occurring.

In a semi-parametric model, $h(t) = h_0(t)r(x,\beta)$, where $h_0(t)$ is the baseline hazard that is the same for any person regardless of x, the co-variates. Earlier, we found survival in terms of $H(t)$, or cumulative hazard, but we can go further and find survival in terms of $r(x,\beta)$ and $S_0(t)$). From the definition of $H(t)$, we know that

$$H(t) = \int_0^t h_0(u)r(x,\beta)du \tag{3.4}$$

$r(x,\beta)$ can be taken out of the integral since it is independent to time, and we are left with

$$H(t) = r(x,\beta)H_0(t) \tag{3.5}$$

We can plug in $r(x,\beta)H_0(t)$ for H(t) in the equation, $S(t) = e^{-H(t)}$, so

$$S(t) = (e^{-H_0(t)})^{r(x,\beta)} \tag{3.6}$$

.

We define $S_0(t) = e^{-H_0(t)}$ and we arrive at

$$S(t) = S_0(t)^{r(x,\beta)}. \tag{3.7}$$

2.3.2 Hazard Ratio

When dealing with two groups of patients, it is the relative hazard of one group to the other that determines how one group's survival fares with the survival of the other. The hazard ratio is

$$\frac{h_0(t)r(x_1,\beta)}{h_0(t)r(x_0,\beta)} \tag{3.8}$$

, The only difference between the two groups is in the covariates, and therefore, the baseline hazard cancels out. An actual hazard ratio less than one means that group 1 has better survival prospects than group 0 because its hazard is lower than group 0's.

2.3.3 Proportional Cox Model

In the proportional Cox model, we make a parametric assumption that

$$r(x,\beta) = e^{x\beta} \tag{3.9}$$

where β has the dimension of the number of co-variates.
If there is 1 co-variate, the hazard ratio with the Cox model for two people is $e^{\beta(x_1-x_2)}$.

The model assumes that the hazard ratio between two people stays constant over time, regardless of the values of their co-variables. Note that it is possible to do Cox regression where the co-variates are time dependent, although this is called an extended Cox model, even though the hazard ratio changes with time. An example would be having an indicator variables that is 1 if a patient has received a transplant and 0, if a patient was not. If the event occurs at death, this variable could certainly change.

Like usual, we must find an estimator for β. β determines how one person's hazard compares to another person's hazard. For example, if you give treatment A the value of 0 and treatment B the value of 1, the hazard ratio between a person getting treatment A and a person getting treatment B will be e^{β}.

First we can find the MLE of β. We will use the likelihood equation, equation 1.2, that we used for the parametric model,

$$l(\beta) = \prod_{i=1}^{n} f(t_i, \beta, x_i)^{c_i} S(t_i, \beta, x_i)^{1-c_i}$$

. In this case we will find the mle for when there is one variable that is being used to model survival. Using the fact that $f(t, x, \beta) = h(t, x, \beta) \times S(t, x, \beta)$ to plug into the likelihood function, we will get

$$l(\beta) = \prod_{i=1}^{n} h(t_i, \beta, x_i)^{c_i} S(t_i, \beta, x_i) \tag{3.10}$$

We can then plug in $S(t, x, \beta) = S_0(t)^{e^{x\beta}}$ and $h(t, x, \beta) = h_0(t)e^{x\beta}$ and we'll get

$$l(\beta) = \prod_{i=1}^{n} (h_0(t_i)e^{x\beta})^{c_i} S_0(t)^{e^{x\beta}} \tag{3.11}$$

However, we are missing the baseline survival and baseline hazard. We can get around that by using a partial likelihood that was created by Cox which does not contain these missing functions:

$$L_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_i\beta}}{\sum_{j \in R(t_i)} e^{x_j\beta}} \tag{3.12}$$

where $R(t_i)$ is the group of people with survival or censor times greater than or equal to t_i. The product is taken just over the survival times rather than survival times and censor times. This partial likelihood assumes that there are no ties. See this write up for a proof of the asymptotic normality of the mle of β when the partial likelihood is used.

Now the log of the partial likelihood must be taken.

$$l_p(\beta) = \sum_{i=1}^{m} [x_i\beta - ln(\sum_{j \in R(t_i)} e^{x_j\beta})] \tag{3.13}$$

Finally, set the derivative of the log-likelihood equal to zero and solve for $\hat{\beta}$

$$\sum_{i=1}^{m} [x_i - \frac{\sum_{j \in R(t_i)} x_j e^{x_j\beta}}{\sum_{j \in R(t_i)} e^{x_j\beta}}] = 0 \tag{3.14}$$

An estimator for the variance is the inverse of the observed information function evaluated at $\hat{\beta}$ which is the inverse of the negative of the second derivative of the likelihood evaluated at the mle.

Once we have the estimator, how can we make confidence intervals for β? Since the mle has asymptotic normality, if the sample size is large, the interval is $\hat{\beta} \pm z_{1-\alpha/2} s_{\hat{\beta}}$.

We can use $\hat{\beta}$ to make confidence intervals for the hazard ratio as well.In the case where you are comparing two drugs, in which case the hazard ratio is e^{β}, you can just exponentiate your confidence interval for β.

To test for the significance of $\hat{\beta}$, you can use either the Wald test, ratio-likelihood test, or score test.

1. The Wald Test, as used earlier just consists of using the asymptotic normality of the mle and calculating the z-score of $\hat{\beta}$ to find the p-value. The z-score is

$$z = \frac{\hat{\beta}}{s_{\hat{\beta}}} \tag{3.15}$$

2. The likelihood ratio test statistic is denoted, G, and it is the difference between the likelihoods when the model contains covariate, x_i, for which $\hat{\beta}_i$ is the coefficient and when the model does not contain this co-variate.

$$G = 2(l(\hat{\beta}, \ldots) - l(0, \ldots)) \tag{3.16}$$

where l in this case is the partial log likelihood. G has a chi-squared distribution with parameter, 1.

3. Score Test, unlike the likelihood ratio test does not need the the mle of β to be calculated. The test statistic has a standard normal distribution.

$$z = \frac{\frac{\partial L}{\partial \beta}}{\sqrt{I(\beta)}}|_{\beta=0} \tag{3.17}$$

, where $I(\beta)$ is the observed information and is the negative of the second derivative of the log-likelihood (in this case partial log-likelihood).

3 Example

Here we look at a clinical trial data set and will do all the tests from above.

I used parametric, non-parametric, and semi-parametric models to look at the relationship between treatment group and time to remission for a set of given data. With a significance level of .1, each test suggested that treatment group is significant in predicting time to remission.

Before doing any of the tests, I created a survival object and isolated the covariates on which I was regressing.

```
survival <- Surv(sampledata$TIME, sampledata$censor_indicator)
regression_variables <- sampledata[, 2 : 23]
```

3.1 Parametric

3.1.1 Exponential

```
exp_mod = survreg(survival ~ ., regression_variables, dist = "exponential")
summary(exp_mod)
```

The coefficient for the treatment variable is -.201 with a 2 sided p-value of .0889. Therefore, with a significance level of .1, we can reject the null hypothesis that there is no difference between the two treatments, but we cannot reject the null hypothesis with a significance level of .05.

This treatment variable, CONDAC, had a value of 1 or 2 depending on the treatment group.In the exponential case, the beta for the treatment variable is the approximate percentage change in expected time to remission when treatment 2 is used as compared to treatment 1. Therefore, treatment 2 is estimated to decrease time remission by 20.1 percent.

The coefficient can also be interpreted through looking at hazard. Since the hazard for the exponential distribution is $e^{-}(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + ...)$, the estimated hazard ratio between someone getting treatment 1 and someone getting treatment 2 is $e^{.201-.201*2} = e^{-.201} = .818$, which is less than one. Therefore, the remission rate for treatment 2 is estimated to be higher than the remission rate for treatment 1, making treatment 2 the preferable treatment.

3.1.2 Weibull

```
wei_mod = survreg(survival ~ ., regression_variables, dist = "weibull")
summary(wei_mod)
```

The 2-sided p-value for the Weibull distribution gives a p-value of .005 for the treatment group variable and estimates the treatment coefficient to be -.21436 and the scale factor to be .888 with a p-value of .00982. Unlike the exponential, we can reject the null hypothesis that the treatments are the same at the .05 significance level. The expected time for someone on treatment 2 to to achieve remission is about 21 percent less than the expected time for someone with the same co-variates on treatment 1.

The hazard for the Weibull distribution is

$$e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots)} t\sigma t^{\sigma-1}$$

If β_1 is the treatment variable, the hazard ratio of treatment 1 to treatment 2, with all other covariates the same is $e^{2*\beta_1-2*\beta_2}$Therefore the hazard ratio at any time between someone getting treatment 1 and someone getting treatment 2 is $e^{.21436-2*.21436} = e^{-.21436}$ which is less than 1, as is the case in the exponential case, and therefore, treatment 2 is estimated to have a higher remission rate.

Note that the weibull model provides a better fit of the data as measured by the likelihood ratio test and AIC:

```
anova(exp_mod, wei_mod)
AIC(exp_mod)
AIC(wei_mod)
```

which result in ...

The AIC for the exponential is 3841.961 and the AIC for the Weibull is 3837.66. The lower AIC value of the Weibull means the weibull is better.

3.2 Semi-Parametric

We run a cox regression below:

```
summary(coxph(survival ~ ., regression_variables))
```

The semi-parametric model gives an estimated coefficient of .2687 and a 2-sided p-value of .03. The hazard ratio between treatment 1 and treatment 2 is $e^{\beta_{tx}*1-\beta_{tx}*2}$ so the estimated hazard ratio is $e^{-.2687}$, which is similar to the estimated hazard ratios for the Weibull and exponential model.

3.3 Non-Parametric

We run the K-M test. See curve in Figure 1.

```
survdiff(survival ~ CONDAC, data = sampledata) #logrank test
survdiff(survival ~ CONDAC, data = sampledata, rho = 1) #wilcoxon rank test
#get K-M estimates
km_fit = survfit(survival ~ CONDAC, data = sampledata)
#now plot the K-M / survival curve
plot(km_fit,
  col = c("blue", "red"), ylab = "Probability of Not Achieving Remission", xlab = "Days")
```

I ran the logrank and Wilcoxon test. The two sided log rank p-value was .00972 and the two sided Wilcoxon test was .00824 so both these tests detected a difference in survival between the two treatment groups.

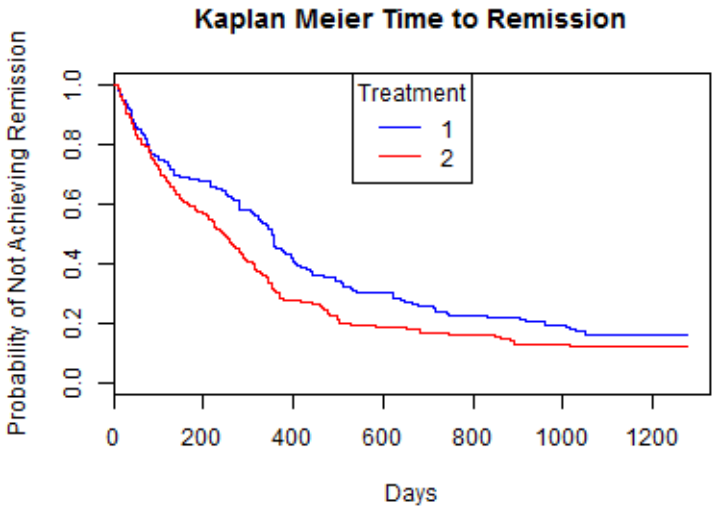

Figure 1: Survival plot for K-M curve.