Survival Analysis

Mingwei Fei

December 13, 2022

1 Sample Size

The $ln(HR)$ follows a normal distribution, we use this to calculate the sample size.

$$ln(\hat{\Delta}) \sim N\left(ln(\Delta), \frac{1}{d_1} + \frac{1}{d_2}\right)$$
$$\left(\frac{1}{d_1} + \frac{1}{d_2}\right)^{-1} = \left[\frac{(z_{\alpha/2} + z_\beta)^2}{(ln\Delta_0)^2}\right]$$

where d_i is the number of observed events.

If hazard ratio set at 2.1, then

$$\left(\frac{1}{d_1} + \frac{1}{d_2}\right)^{-1} = \left[\frac{(1.96 + 0.842)^2}{(ln2.1)^2}\right] = 14.26$$
$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{14.26} = 0.07, \qquad d_1 = d_2 = 28.5$$

The one-sided significance level 0.25, power is 0.8. Note that $Z_{\alpha/2}$ is the z score for the probability $1 - \alpha/2$, and z_β is the z score for the probability $1 - \beta$. Assume the overall event and censored rate is 20%, then the sample size is $57/0.2 = 285$. The total number in the paper is 276.

1.1 Non-inferiority margin Hazard ratio $\Delta_0 = 2.1$

The assumption is that control group (C) event rate 10% and treatment group (T) event rate 20% at 6 months. Assume survival function is an exponential distribution:

$$S_t(t) = exp(-\lambda_1 t), \qquad t = 0.5, S_t = 0.8, -\lambda_1 = ln(0.8)/0.5$$
$$S_c(t) = exp(-\lambda_2 t), \qquad t = 0.5, S_c = 0.9, -\lambda_2 = ln(0.9)/0.5$$
$$\Delta_0 = \frac{\lambda_1}{\lambda_2} = \frac{ln(0.8)}{ln(0.9)} = 2.117$$

1.2 Hazard ratio actual = 0.55

The control group survival 76.8% and treatment group survival 86.2% at 6 months. Assume survival function is an exponential distribution:

$$\begin{aligned} S_t(t) &= exp(-\lambda_1 t), \qquad t = 0.5, S_t = 0.862, -\lambda_1 = ln(0.862)/0.5 \\ S_c(t) &= exp(-\lambda_2 t), \qquad t = 0.5, S_c = 0.768, -\lambda_2 = ln(0.768)/0.5 \\ HR &= \frac{\lambda_1}{\lambda_2} \\ &= \frac{ln(0.862)}{ln(0.768)} = 0.56 \end{aligned}$$

2 Sample Size Formula

The test hypothesis is

$$\begin{aligned} H_0 &: \lambda_1 = \lambda_2 \\ H_1 &: \lambda_1 \neq \lambda_2 \end{aligned}$$

Or equivalently, in terms of hazard ratio, $\Delta = \lambda_1/\lambda_2$

$$\begin{aligned} H_0 &: \Delta = 1 \\ H_1 &: \Delta \neq 1 \end{aligned}$$

A much simpler and quite accurate approximation for a reasonably large number of events is based on the approximate normality of th natural logarithm of the estimated hazard ratio in each treatment group:

$$ln(\hat{\lambda}_i) \sim N(ln\lambda_i, \frac{1}{d_i})$$

where d_i is the number of observed events. Thus, the $ln\Delta = ln\lambda_1 - ln\lambda_2$ also follows a normal distribution with variance $\frac{1}{d_1} + \frac{1}{d_2}$.

$$\begin{aligned} ln(\hat{\Delta}) &\sim N\left(ln(\Delta), \frac{1}{d_1} + \frac{1}{d_2}\right) \\ \left(\frac{1}{d_1} + \frac{1}{d_2}\right)^{-1} &= \left[\frac{(z_{\alpha/2} + z_\beta)^2}{(ln\Delta_0)^2}\right] \end{aligned}$$

The calculation of sample size follows

$$\begin{aligned} Z &= \frac{ln(\hat{\Delta})}{\sigma}, \qquad \sigma = \sqrt{\frac{1}{d_1} + \frac{1}{d_2}}, \qquad \delta = ln(\Delta_0) \\ P(Z &\geq Z_{1-\alpha/2} | H_0) \leq \alpha/2 \\ P(Z &\leq Z_\beta | H_1 = \delta) \geq \beta \end{aligned}$$

So we set Z satisfy the below equation

$$\frac{ln(\hat{\Delta})}{\sigma} = Z_{1-\alpha/2}, \qquad \text{H}_0$$
$$\frac{ln(\hat{\Delta}) - \delta}{\sigma} = Z_{\beta}, \qquad \text{H}_1$$

So we have

$$ln(\hat{\Delta}) = Z_{1-\alpha/2}\sigma, \qquad ln(\hat{\Delta}) = Z_{\beta}\sigma + \delta, \qquad Z_{1-\alpha/2}\sigma = Z_{\beta}\sigma + \delta$$
$$\sigma = \frac{\delta}{Z_{1-\alpha/2} - Z_{\beta}}, \qquad \frac{1}{d_1} + \frac{1}{d_2} = \frac{\delta^2}{(Z_{1-\alpha/2} + Z_{1-\beta})^2}$$

3 Hazard Rate Asymptotic Distribution

3.1 Likelihood Function

If T_i and C_i are independent, which means non-informative censoring. We look at the cumulative conditional probability at time T:

$$p(T \leq s + \epsilon | T \geq s) \approx p(T < s + \epsilon | T \geq s, C \geq s)$$

Note that the above probability is not the hazard rate, it is the cumulative hazard rate. The hazard rate is as below

$$h(t) = \frac{p(t)}{S(t)} = p(s \leq T \leq s + \epsilon | T \geq s)$$

The key of success is to construct likelihood function. We use conditional probability in the situation when there are hidden variables that we can't or don't need to estimate. When there are censoring time, we don't know exactly what those censoring times are.

So in the presence of censoring, we only observe $(T_i, \delta_i), i = 1, ..n$. Let us suppose that T_i is the survival time, which may not be observed and we observe instead $U_i = min(T_i, C_i)$, where C_i is the potential censoring time.

$$\delta_i = \begin{cases} 1 & T_i \leq C_i, \qquad \text{Uncensored} \\ 0 & T_i > C_i, \qquad \text{Censored} \end{cases}$$

3.1.1 Likelihood under Censoring

The likelihood under censoring can be constructed using both the density and distribution functions or the hazard and cumulative hazard functions. Both are equivalent. The

loglikelihood will be a mixture of probabilities and densities, depending on whether the observation was censored or not.

Let us suppose that T_i has distribution $f(x, \theta_0)$, where f is known but θ_0 is unknown. The likelihood construction must be with respect to the bivariate, random variable (U_i, δ_i).

We observe (U_i, δ_i) where $U_i = min(T_i, C_i)$ and δ_i is the indicator variable. In this section we treat C_i as if they were deterministic, we consider the case that they are random later.

We first observe that if $\delta_i = 1$, then the log-likelihood of the individual observation U_i is $logf(U_i, \theta)$, since

$$\begin{aligned} P(U_i = x|\delta_i = 1) = P(T_i = x|T_i \leq c_i) &= \frac{f(x;\theta)}{1 - S(x,\theta)}dx \\ &= \frac{h(x)S(x,\theta)}{1 - S(x,\theta)}dx \end{aligned}$$

where $S(x, \theta)$ is the survival function $1 - F(T_i \leq x)$.

On the other hand, if $\delta_i = 0$, the log likelihood of the individual observation $U_i = c_i|\delta_i = 0$ is simply one, since if $\delta_i = 0$, then $U_i = c_i$ (it is given). Of course it is clear that $p(\delta_i = 1) = 1 - S(c_i, \theta)$ and $P(\delta_i = 0) = S(c_i; \theta)$. Thus altogether the joint density of U_i, δ_i is

$$\begin{aligned} p(U_i, \delta_i) &= \left(\frac{f(x;\theta)}{1 - S(c_i,\theta)}(1 - S(c_i, \theta))\right)^{\delta_i} (1 \times S(c_i, \theta))^{1-\delta_i} \\ &= f(x, \theta)^{\delta_i}[S(c_i, \theta)]^{1-\delta_i} \\ &= h(u_i)^{\delta_i} S(u_i) \\ p(\theta) &= \prod_{i=1}^{n} h(u_i)^{\delta_i} S(u_i) \end{aligned}$$

Therefore by using

$$\begin{aligned} f(U_i, \theta) &= h(U_i, \theta)S(U_i, \theta) \\ H(U_i, \theta) &= -logS(U_i, \theta) \end{aligned}$$

The joint log-likelihood of $(U_i, \delta_i)_{i=1}^n$ is

$$
\begin{aligned}
ln(\theta) &= \sum_{i=1}^{n} (\delta_i log f(\theta) + (1-\delta_i) log(1-F(\theta))) \\
&= \sum_{i=1}^{n} \delta_i \left[log h(T_i,\theta) - H(T_i,\theta) \right] - \sum_{i=1}^{n} (1-\delta_i) H(c_i,\theta) \\
&= \sum_{i=1}^{n} \delta_i log h(U_i,\theta) - \sum_{i=1}^{n} (1-\delta_i) H(U_i,\theta)
\end{aligned}
$$

We can get the MLE of θ by score function, Fisher Information to get the variance.

3.2 Exponential Distribution

Suppose that $T_1, T_2, ...T_n$ are i.i.d $Exp(\lambda)$ and subject to noninformative right censoring. The exponential distribution

$$f(x,\lambda) = \lambda exp(-\lambda x)$$

The survival function

$$S(x,\lambda) = 1 - F(x,\lambda) = 1 - \int_0^x \lambda exp(-\lambda x) dx = exp(-\lambda x)$$

The likelihood function

$$p(\lambda) = \prod_{i=1}^{n} \lambda^{\delta_i} exp(-\lambda u_i) = \lambda^r exp(-\lambda W)$$

where $r = \sum_{i=1}^{n} \delta_i$ are the number of failures; $W = \sum_{i=1}^{n} u_i$ is total followup time.

The Score function and observed information

$$
\begin{aligned}
\frac{\partial ln(\lambda)}{\partial \lambda} &= \frac{r}{\lambda} - W \\
-\frac{\partial^2 ln(\lambda)}{\partial \lambda \, \partial \lambda} &= \frac{r}{\lambda^2}
\end{aligned}
$$

$\hat{\lambda}$ approximately follows $N(\lambda, \lambda^2/r)$ for large n.

By delta method,

$$log(\hat{\lambda}) \sim N(log(\lambda), r^{-1})$$

The variance of log hazard ratio r^{-1} is free of the unknown parameter λ. Similarly, we see that the log of odds ratio is used more common than odds ratio.

4 Cox Proportional Hazard Assumption

There are a variety of techniques, both graphical and test-based, for assessing the validity of the proportional hazards assumption. One technique is to simply plot Kaplan–Meier survival curves if you are comparing two groups with no covariates. If the curves cross, the proportional hazards assumption may be violated. An important caveat to this approach must be kept in mind for small studies. There may be a large amount of error associated with the estimation of survival curves for studies with a small sample size, therefore the curves may cross even when the proportional hazards assumption is met.

The complementary log-log plot is a more robust test that plots the logarithm of the negative logarithm of the estimated survivor function against the logarithm of survival time. If the hazards are proportional across groups, this plot will yield parallel curves.

The second popular method is Schoenfeld residual plot. The Schoenfeld residuals have since become an indispensable tool in the field of Survival Analysis and they have found in a place in all major statistical analysis software such as STATA, SAS, SPSS, Statsmodels, Lifelines and many others. We will discuss that later in another post.

The third common method for testing the proportional hazards assumption is to include a time interaction term to determine if the HR changes over time, since time is often the culprit for non-proportionality of the hazards. Evidence that the group $\times$ time interaction term is not zero is evidence against proportional hazards.

The most popular graphical techniques for evaluating the PH assumption involves comparing estimated –ln(–ln) survival curves over different (combinations of) categories of variables being investigated.

A log–log survival curve is simply a transformation of an estimated survival curve that results from taking the natural log of an estimated survival probability twice.

4.1 Adjusting Survival Curves

What is the adjusted survival curve? What is the difference between the adjusted survival curve and Kaplan-Meier curve?

Remember that we don't use any model to get the Kaplan-Meier estimator, also we know that Kaplan-Meier curve couldn't adjust for covariates. If we use the Cox's model to fit the survival data, survival curves can be obtained adjusted for the explanatory variables used as predictors. What would the adjusted survival curves look like, are they close to or different from the Kaplan-Meier curve?

From a survival analysis point of view, we want to obtain both unadjusted and adjusted survival curve and estimates for comparison. And adjusted survival curve could be considered as Kaplan-Meier curve with adjusted covariates.

We will derive the adjusted survival curve using the hazard function definition. Hazard function describes the 'intensity of death' at the time t given that the individual has already survived past time t. There is another quantity that is also common in survival analysis, the cumulative hazard function, which is

$$h(t) = h_0(t)exp(x^T\beta), \qquad H(t) \qquad = \int_0^t h(s)ds$$

You can interpret $H(t)$ as the cumulative amount of hazard up to time t. The cumulative hazard function and survival function linked as follows:

$$h(t) = \frac{f(t)}{S(t)} = \frac{\frac{\partial S(t)}{\partial t}}{S(t)} = -\partial lnS(t), \qquad H(t) = -lnS(t)$$
$$S(t|x) = exp(-H(t|x)) = exp(-H_0(t)exp(x^T\beta)) = S_0(t)^{exp(x^T\beta)}$$

This means that if we know the Hazard function, we can solve this differential equation for survival function. And the adjusted survival function could be obtained when we fit the Cox's model. The estimates of $\hat{S}_0(t)$ and β are provided by the computer program that fits the Cox model. The $\mathbf{X}$'s, however, must first be specified by the investigator before the computer program can compute the estimated survival curve.

4.1.1 Exponential distribution adjusted survival curve

What is the survival function and hazard function of an exponential R.V.?

Let $T_1 \sim Exp(\lambda)$. Then

$$p(t) = \lambda exp(-\lambda t), \qquad F(t) = 1 - exp(-\lambda t), \quad t > 0$$

Thus,

$$S(t) = 1 - F(t) = exp(-\lambda t), \qquad h(t) = \frac{p(t)}{S(t)} = \lambda,$$
$$H(t) = \lambda t$$

Namely, in an exponential distribution, the hazard function is a constant and the cumulative hazard is just a linear function of time.

4.2 Cox's PH assumption

The Cox (proportional hazard) model is one of the most popular model combining the covariates and the survival function. It starts with modeling the hazard function $h(t|X = x)$:

$$h(t) = h_0(t)exp(x^T\beta),$$

where β is the vector of coefficients of each covariate. The function $h_0(t)$ is called the baseline hazard function. Namely, the Cox model assumes that the covariates have a

linear multiplication effect on the hazard function and the effect stays the same across time.

From above, we got the adjusted survival function

$$S(t|x) = S_0(t)^{exp(x^T\beta)}$$

where $S_0(t)$ is called the baseline survival function.

Suppose we have X_1, X_2 covariates in each group

$$\begin{aligned} HR &= exp\{(X_1 - X_2)^T\beta\} \\ -ln(-lnS(t)) &= -ln(-lnS_0(t)) - exp(X^T\beta) \\ -ln(-lnS_0(X_1))HR &= -ln(-lnS_0(X_2)) + exp[(X_1 - X_2)^T\beta] \end{aligned}$$

So under the proportional hazard ratio assumption, the log-log of $S(t)$ is hazard function should be parallel.

This expression indicates that if we use a Cox model (well-used) and plot the estimated log-log survival curves for individuals on the same graph, the two plots would be approximately parallel. The distance between the two curves is the linear expression involving the differences in predictor values, which does not involve time.

4.2.1 Schoenfeld Residual

The Assumptions of the Cox Proportional Hazards Model The Cox model makes three assumptions:

Common baseline hazard rate $\lambda(t)$: At any time t, all individuals are assumed to experience the same baseline hazard $\lambda(t)$. For example, if a study consists of males and females belonging to different races and age groups, then at any time t during the study, white males who entered the study when they were in the 18–34 years age range are assumed to experience the same baseline hazard $\lambda(t)$ as Asian females who entered the study when they 40–64 year old. This is clearly a strong assumption and one must validate it before accepting results of the Cox model for your data set.

Time invariant co-variates X: The effect of the regression variables X (a.k.a. co-variates) on the instantaneous hazard experienced by an individual is assumed to remain constant over time. For example, if a volunteer enters a study of cancer risk to smokers, and if the participant's genetic makeup is a co-variate, then the effect of their genes on the hazard of contracting cancer is assumed to remain constant throughout the study. Notation-wise, $X(t_i) = X$ for all t_i. This is again a rather strong assumption that should be validated. For example, gene expression can vary with time in response to a number of factors. Time invariant regression coefficients β: The regression coefficients β do not vary with time. In other words, $\beta(t_i) = \beta$ for all t_i. One can see that assumptions (2) and (3) are closely related. For example, the effect of gene expression increasing with time can be expressed as the coefficient of the gene regression variable increasing with time.

Schoenfeld Residuals are used to validate the above assumptions made by the Cox model.