ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

Research on the Test Method of Proportional hazard Hypothesis in Cox Model

Yu Gu*

Department of Statistics and Data Science, Jinan University, Guangzhou, 510632, China

Abstract

With the growth of world economy, the development of health care, the change of disease spectrum and the improvement of life expectancy, the follow-up studies on clinical trials and epidemiology of tumor, chronic diseases and senile diseases are becoming more and more important. The data of these clinical trials and follow-up studies can be sorted into survival data. Survival analysis is a subject that makes statistical inference on one or more non negative random variables, studies survival phenomenon and response time data and their statistical laws. At present, Cox proportional hazard regression model is still the most commonly used method for multivariate analysis of survival data. Due to the wide range of application of Cox model, analysts often ignore its application condition proportional hazard assumption, which directly affects the stability of the model. This paper systematically discusses the common methods of Cox model to test the proportional hazard hypothesis from two aspects: graphical method and hypothesis testing method. This paper studies six graphical methods, namely direct observation method, Cox & K-M comparison method, log-log image method, various graphical methods based on cumulative hazard function, Schoenfeld residual diagram and score residual diagram. In addition, five hypothesis testing methods are studied, which are linear correlation test, time covariate test, weighted residual score test, omnibus test and a new cubic spline function method. This paper explores the principles of these methods and compares their advantages and disadvantages.

Keywords

Cox model; Proportional hazard hypothesis; Schoenfeld residuals; Hypothesis test.

1. Introduction

1.1. Background

Survival analysis refers to the method of analyzing and inferring the survival time of organisms or people and studying the relationship between survival time and outcome and many influencing factors and the extent of it according to the data obtained from the experiment or investigation, also known as survival analysis or survival analysis. This "survival time" can be broadly understood as the duration of a certain state in nature, human society, technological process, and market behavior. Survival analysis is widely used in medicine, biology, insurance, engineering science, psychology, economics, market research and analysis and other scientific fields. In the past twenty or thirty years, in the statistical research of survival data, many methods and models have both important theoretical significance and wide application value. Cox model is the most commonly used semi-parametric model for multivariate analysis of survival data. Cox regression model, also known as "Cox Proportional Hazards Regression Model", is a semi-parametric regression model proposed by the British statistician D.R.Cox (1972)[1]. With survival outcome and survival time as dependent variables, the model can analyze the impact of many factors on survival time at the same time, it can also analyze data with truncated survival time and does not require to estimate the survival distribution type of

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

the data. Because of the above excellent properties, the model has been widely used in medical follow-up studies since its inception, and is the most widely used multifactor analysis method in survival analysis.

Since the Cox (1972) model was proposed, it has made breakthroughs in the analysis of deleted survival data. This model can be used to compare and test the treatment effect based on the occurrence of certain events while adjusting the influence of accompanying variables. A key assumption of the model is that the function of the two covariate values is independent of time, which is called the Proportional Hazards Assumption(PH assumption). Although Cox regression model makes multivariate analysis in survival data possible, since it depends on this strict assumption, if the data cannot be met, it will greatly affect the interpretation of the results, and even lead to wrong conclusions. Therefore, before using Cox model as an analysis tool, it is necessary to test the validity of model assumptions. If this assumption is violated, the simple Cox model is invalid and requires more complex analysis. At present, Cox regression model is abused to some extent. Most researchers ignore the test of PH assumption when using this method, which affects the authenticity and reliability of the results. This paper hopes to summarize the existing testing methods of PH assumption, so as to prompt readers to reasonably use Cox regression model, select appropriate methods to test the applicability of data, and establish a stable and effective model.

1.2. Literature Review

Since the introduction of Cox regression model, people have proposed several methods to test the proportional hazard hypothesis. Among them, some authors put forward the graphical method to check the proportional hazard hypothesis. For example, the Cox&K-M comparison method proposed by Cox himself (Cox 1972) in "Cox Regression Model and Life Table" is to compare survival curves based on Cox model and non-parametric methods such as Kaplan-Meier estimation. This method was extended by Harrel and Lee et al. (1986) and can be used for the analysis of counting data, grade data and measurement data. The measurement data shall be discretized first, and then the Cox and K-M curve results of each group shall be compared. Sehoenfeld (1982) defined a partial residual for the proportional hazard model, called Sehoenfeld residual. The Sehoenfeld residual plot can be used to test the PH hypothesis. Later, Grambsch and Therneau (1994) adjusted the scale of Schoenfeld residuals and proposed the scaled Schoenfeld residuals. The above two scholars (Grambsch and Therneau 1990) also mentioned the score residual in another document, "Survival model based on martingale residual", which can also be used in more complex situations as an extension of Schoenfeld residual. In addition, Hess (1995) summarized several commonly used graphical methods of cumulative hazard function, which can be used to test PH hypothesis. In addition to the methods mentioned above, there are also Aalen graphical method proposed by Aalen (1980), UCP (Updated Covariant Percentage) graphical method proposed by Anderson (1982), and Arjas graphical method proposed by Arjas (1988). However, due to the relatively complex calculation process and no corresponding software module can be directly drawn, it will not be described in detail in this paper.

Similarly, some statisticians put forward a variety of non-graphical methods based on hypothesis test and different test statistics. When Cox (1972) proposed the Cox model, he proposed a method to check the PH hypothesis by introducing a constructed time-dependent covariate into the model, that is, adding a time-dependent interaction term to the model, and then testing the significance of the interaction term. The common linear correlation test method originated from the concept of partial residual put forward by Schoenfeld (1982), and was improved by Harrel and Lee (1986), and gradually developed into a simple PH hypothesis test method. The linear correlation test method is also applicable to other forms of residual, such as martingale residual. The linear correlation test method is simple in principle, uses the

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

traditional hypothesis test method, provides statistics and p values, is simple and effective, and the results are objective and easy to judge. Some researchers, such as Kay (1984), Harrel and Lee (1986), believe that to test the PH hypothesis of the Cox regression model, we can also divide the survival time into several disjoint intervals in advance, and fit the Cox model in each interval to compare whether the regression coefficients in different intervals are consistent. According to this idea, Moreau et al. (1985) proposed the Omnibus method, which was extended by O 'Quigley and Pessione (1989) and became a special case of the time-dependent covariate method. Later, Grambsch and Therneau (1994) further proposed the weighted residual score method, which is similar to the synthesis of the time-dependent covariate method and the linear correlation test method. Other feasible test methods include the generalized moment test proposed by Horowitz et al. (1992), the linear rank test proposed by Chappell (1993), and the score test proposed by Lin DY et al. (1993). The simulation study by Nicholas (1997) shows that compared with other methods, the time-dependent covariate method, linear correlation test method and weighted residual score method have higher accuracy in testing the PH hypothesis, so this paper will focus on these methods.

1.3. Basic Concept

The basic form of Cox proportional hazard regression model is as follows:

$$h(t,X) = h_0(t)\exp\left(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right) \quad (1)$$

Where $\beta_1, \beta_2, \dots, \beta_p$ is the partial regression coefficient of the independent variable, which is a parameter to be estimated from the sample data. $h_0(t)$ is the baseline hazard rate of $h(t,X)$ when the vector X is 0, which is the quantity to be estimated from the sample data. This formula is called Cox regression model for short, and can also be written in the following form:

$$ln[h(t,X)/h_0(t)] = lnRR = \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \quad (2)$$

Cox regression model is a semi-parametric model because it contains $h_0(t)$ while the parameter β can still be estimated according to formula (1).

Hazard ratio is the ratio of the expected hazards of two individuals:

$$HR = \frac{h(t,X^*)}{h(t,X)} \quad (3)$$

$X^* = \left(X_1^*, X_{2,}^*, \dots, X_p^*\right)$ and $X = \left(X_1, X_2, \dots, X_p\right)$ represent two different individuals.

For Cox model, hazard ratio of different individuals X^* and X :

$$HR = \frac{h(t,X^*)}{h(t,X)} = \frac{h_0(t)\exp(\beta X^*)}{h_0(t)\exp(\beta X)} = exp[\beta(X^* - X)] \quad (4)$$

It is independent of the benchmark hazard function and does not change over time.

Proportional Hazards Assumption (PH Assumption): Assume that the risk ratio is fixed, that is, the influence of covariates on survival probability does not change over time, which means $h(t,X)/h_0(t)$ does not change over time. Because the hazard ratio does not change with time, it is proportional with time. Therefore, Cox model is also called proportional hazard rate model (PH model). This assumption is the prerequisite for establishing the Cox regression model. The Cox model is valid only when this assumption is satisfied.

ISSN: 2637-6067 | DOI: 10.6918/IJOSSER.202304_6(4).0004

1.4. Content

In this study, the inspection methods related to PH assumption are roughly divided into graphical method and normal hypothesis testing method. The graphical method is to determine whether the data meets or approximately meets the model assumptions by observing whether the distribution or trend of the scattered points in the scatter diagram meets the shape under the basic assumptions of the established model; The normal hypothesis test method is to determine whether the data meets or approximately meets the model assumption by constructing a test statistic that follows a known distribution under the basic assumption of the given model and using the p-value of the normal hypothesis test.

Six graphical methods are studied in this paper, including direct observation method, Cox&K-M comparison method, log-log image method, multiple graphic methods based on cumulative hazard function, Schoenfeld residual diagram, and Score residual diagram. This paper also studies five hypothesis testing methods, namely: linear correlation test, time-dependent covariate method, weighted residuals score method, Omnibus test method and a new cubic spline function method.

According to the idea of split survival time, this paper introduces a method to test the PH assumption of Cox regression model using cubic spline function. This method uses the spline function of time to avoid the problem of determining the specific function form of time on the whole time t and then modeling the Cox model in the time-dependent covariate method.

2. Graphical Method

2.1. Direct Observation Method

Observe whether the Kaplan-Meier survival curve of each qualitative independent variable at each level has a crossover. If there is a crossover, it means that the qualitative independent variable does not meet the proportional hazard assumption. However, when there is no obvious crossover, it does not mean that the independent variable must meet the PH assumption, and further verification is needed.

2.2. Cox&K-M Comparison Method

This method was first proposed by Cox himself in "Cox regression model and life table" in 1972, that is to compare the morphological differences of survival curves based on Cox model and other non-parametric methods such as Kaplan-meier estimation. If the trend of the two curves is basically the same and there is no intersection, it indicates that the PH assumption is met. Similarly, since $H(t) = -logS(t)$, the cumulative hazard function curve of the Cox model can be compared with the kaplan-meier cumulative hazard function curve. If there is a cross, it indicates that the PH assumption is violated. This method is extended by Harrel and Lee, and can be used for the analysis of continuous variables or hierarchical variables. For continuous variables, it is necessary to discretize the variables and then compare the Cox and K-M curve knots of each group[2].

This method skillfully utilizes the correlation between the risk function$h(t)$, cumulative hazard function $H(t) = \int_0^t h(u)du$ and survival function $S(t) = \exp(-H(t))$, and compares the observed "non-model-based" survival curve with the predicted "model-based" survival curve, which is simple, intuitive and easy to interpret. However, this method is difficult to determine whether the difference between Cox model and K-M curve is due to sampling error or real trend. Hess suggested that adding confidence interval to the survival curve might help, but the calculation and drawing of confidence interval would increase the difficulty of analysis and affect the subjectivity of judgment[10].

2.3. Log-log Plot Method

For each level group of specific qualitative independent variables, draw the relationship plot between $\log(-logS(t))$ and survival time or logarithm of survival time. If the line segments are obviously not parallel, it means that the qualitative independent variables do not conform to the proportional hazards assumption.

For 0-1 variables (i.e. survival data of two samples), the following expressions are valid when the PH assumption is satisfied:

$$h(t,X) = h_0(t)\exp(\beta X) \tag{5}$$

$$H(t,X) = \int_0^t h(u,X)du = \int_0^t h_0(u)\exp(\beta X)du = \exp(\beta X)\int_0^t h_0(u)du = H_0(t)\exp(\beta X) \tag{6}$$

$$logH(t) = logH_0(t) + \beta X \tag{7}$$

$$\log(-logS(t)) = \log(-logS_0(t)) + \beta X \tag{8}$$

Therefore, the two curves should be roughly parallel or equidistant from each other.

For multivariable Cox model, the data can be divided into several layers according to a certain variable in the model, and the Cox model can be fitted for each layer. If the value of each layer is close to the original model, and the graph of each layer to t is roughly parallel, the hazard rate is proportional, and it is appropriate to introduce this variable into the model.

2.4. Graphical Method Based on Cumulative Hazard Function

It can be observed that:

(1) the trend plot of the cumulative hazard function against time t, if the proportion is constant;

(2) The trend plot of the cumulative hazard function for the benchmark cumulative hazard function, if the slope is constant;

(3) The trend plot of the log of the cumulative hazard function against time t, if parallel;

(4) The trend plot of the log of the cumulative hazard function difference against time t is constant if it is constant.

The above four can all be considered that the data meets PH assumption[3].

According to formulas (1) - (4), when the PH assumption is true, the ratio between different groups in the cumulative hazard function curve and the ratio between different groups in the log of the cumulative hazard function curve should be constant, that is, displayed as proportional or parallel to each other.

2.5. Plotting Schoenfeld residual

For continuous independent variables, the Schoenfeld residual plot and Score residual plot can be used to judge, or the continuous independent variables can be qualitative, and then the graphical method mentioned before can be adopted.

Based on the correlation between the above methods and time variables, Schoenfeld defined a partial residual independent of t to test the PH assumption of Cox regression model.

Make R_i is the hazard set at time t_i (i.e. when the event of the ith individual occurs), then the partial residual of the individual can be expressed as $r_i = X_i - E(X_i|R_i)$, where

$$r_i = (r_{i1}, r_{i2}, \dots, r_{ik})' \tag{9}$$

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

$$E(X_i|R_i) = \frac{\sum_{k\in R_i} X_k \exp(X_k\beta)}{\sum_{k\in R_i} \exp(X_k\beta)} \tag{10}$$

$E(X_i|R_i)$ is the conditional expectation under the given risk set R_i. If the Schoenfeld residual value is positive, it means that the actual value of X is higher than the expected value at the corresponding time of death. $\hat{\beta}$ can be obtained from the maximum partial likelihood estimation under the PH assumption. Because $\hat{\beta}$ is the solution of $\sum\left(X_i - E(X_i|R_i)\right) = 0$, it can be proved that when PH assumption is satisfied, $E(\hat{r}_i) = \frac{1}{n}\sum\left(X_i - E(X_i|R_i)\right) = 0$, and $\hat{r}_i$ is approximately uncorrelated[4]. If we draw a generalized linear regression plot of Schoenfeld residual and survival time, i.e. the plot of $\hat{r}_i$ versus t_i, the plot should fluctuate around 0. If the curve presents a non-zero slope, it means that the variable does not meet the PH assumption.

Grambsch and Therneau (1994) adjusted the scale of Schoenfeld residuals, and proposed the scaled Schoenfeld residuals(also called standardized Schoenfeld residuals or weighted Schoenfeld residuals), which are expressed as follows:

$$r_i^* = r_i \cdot d \cdot S_\beta \tag{11}$$

Where, d is the number of all non-censored observations in the data. The purpose of scaling is to make r_i^* and $\hat{\beta}$ have the same scale. Under the PH assumption, the plot of $\hat{r}_i$ versus t_i should centre around 0, otherwise the PH assumption is violated[5].

However, the scatter trend in the Schoenfeld residual plot is difficult to evaluate, especially for binary variables. Lowess (locally weighted scatterplot smoothing) smoothing method and the estimation of its confidence interval can help estimate the trend of the scatter, play a greater role in the readability of the residual plot, and separate the actual PH test results from the random trend.

2.6. Plotting Score residual

Score residual is a decomposition of the log of the partial likelihood of Cox model to the first order partial derivative of β_k. It is a martingale transformation residual proposed by Therneau and Gramsch (1990). The score residual of the k-th variable of the i-th observation can be expressed as:

$$L_{ik} = \int_0^\infty (X_{ik} - \bar{X}_k)\, dM_i \tag{12}$$

$$\widehat{M}_i = \delta - \widehat{H}_0(t_i)\exp(\beta X) \tag{13}$$

Where $\widehat{M}_i$ is martingale residual, δ is the censoring indicator variable. A plot of L_{ik} versus t_i can evaluate whether the covariate effect monotonically increases or decreases with time[6].

The advantage of Score residual plot is that continuous variables do not need to be discretized, and only need to fit one Cox model.

3. Hypothesis Test

3.1. Linear Correlation Test

This method originated from the concept of partial residual put forward by Schoenfeld, and was improved by Harrell and Lee, and gradually developed into a simple method to test PH assumption[2]. The principle is that if the data meets the PH assumption, the Schoenfeld

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

residual does not depend on the survival time, and there is no correlation between the Schoenfeld residual and the rank of the survival time. So just check the correlation coefficient ρ between them and ensure $\rho = 0$, it can be proved that the data meets the PH assumption. Based on this, test statistics can be constructed:

$$z = \rho\sqrt{(d-2)/(1-\rho^2)} \sim N(0,1) \tag{14}$$

Where ρ is the correlation between Schoenfeld residuals and the rank of survival time, and d is the total number of uncensored observations. If the hazard ratio of the covariate value increases with time, the test statistic tends to be positive; If the hazard ratio decreases over time, the test statistics tend to be negative.

Specifically, the calculation steps are generally as follows:

(1) Calculate Schoenfeld residual;

(2) Sort the uncensoredsurvival time and create a new variable to record the event occurrence rank 1, 2, 3... If the same survival time (knot) occurs, record it with the average rank;

(3) Carry out hypothesis test to test the correlation between Schoenfeld residual and survival time rank in the first two steps. The original assumption is that the correlation between the two is zero($H_0: \rho = 0$). If the null hypothesis is rejected, it indicates that the data does not meet the PH assumption.

This method does not need to layer time and covariates, but can also use other types of residuals, such as martingale residuals and smooth Schoenfeld residuals. The test statistics are obtained by Fisher's z-transformation of Pearson correlation between residual and the rank order of failure time. When there is a knot failure time, the weighted correlation coefficient is calculated by using the number of knot failure times as the weight.

3.2. Time-dependent Covariate Method

When Cox (1972) proposed the Cox model, he proposed the method of checking PH assumption by introducing a constructed time-dependent covariate, that is, adding a time-dependent interaction term, such as $X \cdot g(t)$, to the model, and then testing its significance[1]. Where $g(t)$ is a time function. In this way, in order to check the PH assumption, an extended Cox model with time-dependent covariates determined by a certain time function is fitted:

$$z = \rho\sqrt{(d-2)/(1-\rho^2)} \sim N(0,1) \tag{15}$$

And there is:

$$lnHR = (X_1 - X_2)\big(\beta + \gamma \cdot g(t)\big) \tag{16}$$

$g(t)$ is a non-zero time function corresponding to X, such as $g(t) = t, log(t), rank(t)$, etc. The hazard ratio is constant at all times only when $\gamma = 0$. If $\gamma > 0$, the hazard ratio increases linearly with time; if $\gamma < 0$, the hazard ratio decreases linearly with time. The test of PH assumption can be transformed into the test of $\gamma = 0$.

To test $H_0: \gamma = 0$, i.e. PH assumption, the likelihood ratio test statistic $-2ln\big[L(\hat{\beta},0)/L(\hat{\beta},\gamma)\big] \sim \chi^2(1)$ can be calculated , at the same time score statistics and Wald statistics can also be obtained (using the inverse of the negative second order partial derivative matrix to obtain $V(\hat{\gamma})$). This test statistic does not need to partition time or covariates. In addition, as mentioned in the previous graphical method, the trend plot of lnHR against time t, called LHRF (Log Hazard

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

Ratio Function) plot, can also be drawn here to observe its levelness and stability, and also to determine whether the data meets the PH assumption.

In addition, this method is also very important for the selection of time function $g(t)$. In previous literature, monotonic functions such as linear functions, exponential functions and logarithmic functions are more common because they are easy to calculate and the interaction terms that monotonically change with time are easier to interpret.

The research on non-monotonic time function is relatively rare. Stablein and others have considered using quadratic function to model. Hess proposed that piecewise linear function can also be used to fit the time interaction term. In the early 1990s, nonparametric models were gradually proposed and used. Some studies used spline functions to estimate time-dependent covariates without determining the specific form of the time function, which can avoid the result deviation caused by the wrong selection of the model and improve the test efficiency. However, how to select the appropriate knot location and number will become a new problem. Some specific methods will be mentioned later.

3.3. Weighted Residuals Score Test

Grambsch and Therneau (1994) developed a score test based on weighted Schoenfeld residuals. The idea is that most common choices facing proportional hazard are represented by time-varying coefficient models. This method is similar to the synthesis of linear correlation test and time-dependent covariate method. Let $(t) = \beta + \gamma \cdot g(t)$, where g (t) is a measurable process. Grambsch and Therneau showed that $E[r_i(\beta)] \approx S_{\beta(t_i)} \cdot g(t_i) \cdot \gamma$.

The score test for $H_0: \beta(t) = \beta$ is equivalent to the generalized least squares test of Schoenfeld residuals. Let $r_i^* = r_i^*(\beta) = S_{\beta(t_i)} \cdot r_i(\beta)$ denotes the scale Schoenfeld residual, suppose β is unknown, $\hat{\beta}$ is the maximum partial likelihood estimation under H_0, let $\hat{r}_i = r_i(\hat{\beta})$, from the generalized least squares estimation we can obtain:

$$\hat{\gamma} = D^{-1} \sum g(t_i)\, \hat{r}_i \tag{17}$$

Where

$$D = \sum g(t_i)\, S_{\hat{\beta}(t_i)} \cdot g(t_i)' - \left(\sum g(t_i)\, S_{\hat{\beta}(t_i)}\right)\left(\sum S_{\hat{\beta}(t_i)}\right)^{-1}\left(\sum g(t_i)\, S_{\hat{\beta}(t_i)}\right)' \tag{18}$$

From this, we can construct an asymptotic χ^2 statistics with degree of freedom p:

$$\left(\sum g(t_i)\, \hat{r}_i\right) D^{-1} \left(\sum g(t_i)\, \hat{r}_i\right) \sim \chi^2(p) \tag{19}$$

If we can conclude that $\beta(t)$ is independent of time, it indicates that the data meets the PH assumption[5].

$\hat{r}$ is the one-step Newton-Raphson algorithm estimate of r, and the test statistic S is a Rao score test of $H_0: (\beta, \gamma) = (\beta, 0)$ based on partial likelihood, and also a test for the non-zero slope of the generalized linear regression of the selected time function.

One can also use $r_i^* = \hat{\beta} + d \cdot r_i \cdot S_{\hat{\beta}(t_i)}$ to calculate the scale residual, where d is the total number of deaths. It has been shown that $E(r_i^*) = \hat{\beta}(t)$, so r_i^* gives a direct estimate of $\hat{\beta}(t)$ for the smoothing graph of time. This test does not require the stratification of time or covariates, but only the Schoenfeld residual, along with the regression coefficient and covariance matrix of a standard time independent Cox model fit.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

3.4. Omnibus Test

Some researchers believe that the PH assumption of Cox regression model can also be tested by dividing the survival time, that is, dividing the time into several disjoint intervals in order in advance, and fitting the Cox model in each interval to compare whether the regression coefficients in different intervals are consistent.

Suppose time is divided into q intervals (τ_0, τ_1), ..., (τ_{q-1}, τ_q), where $\tau_0 = 0, \tau_q = \infty$, let

$$I_j = \begin{cases} 1, & \tau_{j-1} \le \tau_j \\ 0, & \text{elsewhere} \end{cases} \tag{20}$$

Then the piecewise Cox regression model can be expressed as:

$$h(t, X) = h_0(t)\exp\left(X\left(\beta + \sum \zeta_j I_j\right)\right) \tag{21}$$

That is, in the jth time interval, the value of the regression coefficient is always $\beta + \zeta_j$. The PH assumption of the model can be verified by testing $\zeta = 0$ (which means test the statistic $\sum \gamma_{(j)}{}' C_j^{-1} \gamma_{(j)} \sim \chi^2\left(p(q-1)\right)$, where $\gamma_{(j)} = \sum_{t_i \in (\tau_{j-1}, \tau_j)} \gamma_i$, $C_{(j)} = \sum_{t_i \in (\tau_{j-1}, \tau_j)} S_{\gamma i}$), or by plotting a piecewise linear function plot of regression coefficient against t.

This method was originally proposed by Moreau et al.[7], then extended by O 'Quigley and Pessione[8], and became a special case of the time-dependent covariate method. This method requires that the number of events in each interval should be balanced and comparable, and should not be too small. How to determine the location and number of interval partition points has been widely discussed. Some authors believe that it is appropriate to select the time-to-event quantiles, but this consideration is affected by the censoring mechanism, it is difficult to give a convincing conclusion if the survival data are not randomly censored; At the same time, the number of intervals also depends on the size of sample size and the distribution of survival time. Some authors suggest that post-analysis can be used to determine the partition point, but such an approach would inevitably lead to a decline in statistical efficiency and affect the credibility of the results.

3.5. Cubic Spine Function Method

The purpose of the spline function is to replace the unique function f defined in the entire time t with several low-order polynomials (splines) defined in the sub-interval of the time t, and the points dividing the sub-intervals are called knots. The splines are continuous piecewise polynomials of order m and continuously derivable on order m-1[9]. In this way, the spline function is a smooth joint piecewise polynomial. Generally, if there are k knots at time $t_i (i = 1, \dots, k)$, then time spline function can be written as

$$S(t) = \sum_{j=0}^{m} \beta_j t^j + \sum_{i=1}^{k} \theta_i (t - t_i)_+^m \tag{22}$$

Where if $u > 0$, $u_+ = u$. Otherwise, it is 0. So there are k+m+1 regression coefficients for this m-order spline function(β and θ).

Splines composed of piecewise cubic polynomials can provide greater flexibility for fitting data, which is more common because of its smooth number of continuous first and second derivatives and fewer parameters than higher-order splines.

The expression of constrained cubic spline function with k knots is :

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

$$S(t) = \beta_0 + \beta_1 t + \sum_{i=1}^{k-2} \theta_i \left[(t - t_i)_+^3 - \frac{(t-t_{k-1})_+^3 (t_k - t_i)}{(t_k - t_{k-1})} + \frac{(t-t_k)_+^3 (t_{k-1} - t_i)}{(t_k - t_{k-1})} \right] \quad (23)$$

In addition to t, k-2 new variables are introduced,

$$S_i = (t - t_i)_+^3 - \frac{(t-t_{k-1})_+^3 (t_k - t_i)}{(t_k - t_{k-1})} + \frac{(t-t_k)_+^3 (t_{k-1} - t_i)}{(t_k - t_{k-1})} \qquad i = 1, \ldots, k-2 \quad (24)$$

The function family that can be fitted by three-knot cubic spline function includes constant, linear function, monotone function and non-monotone function. The coefficient θ_i of new spline variable can be used to test the nonlinearity of f (t). Test θ= 0 is equivalent to test the linearity of $S(t) = \beta_0 + \beta_1 t$ or f (t).

The cubic spline regression model has the following advantages: (1) it provides linear statistical test; (2) it provides the graphical method to identify the deviation from linearity and determine which expression produces linearity; (3) it does not need to determine the function form but uses nonlinear modeling; (4) Be able to use current software to fit: (5) it can estimate coefficients and confidence intervals with standard methods; (6) Fit the model while adjusting other variables, including other spline functions; (7) The prediction equation can be generated.

When considering PH assumption, the time-dependent covariate interaction term of a variable x_1 can be used as the product term to fit the PH regression equation :

$$h(t, x) = h_0(t) \exp[\beta_1 x_1 + \beta_2 x_1 f(t)] \quad (25)$$

Where f (t) is a function of time. The corresponding LHRF(Log Hazards Ratio Function) of different individuals is LHRF= $\delta(\beta_1 + \beta_2 f(t))$, and $\delta = (x_{11} - x_{12})$ is the difference of x_1.

In order to generate a three-knot-time spline function in PH regression, a new time-dependent covariant $S_1(t)$ must be defined. Let t represent time, t_i represents the time at which the node is,

$$S_1(t) = (t - t_1)_+^3 - \frac{(t_3 - t_1)}{(t_3 - t_2)} (t - t_2)_+^3 + \frac{(t_2 - t_1)}{(t_3 - t_2)} (t - t_3)_+^3 \quad (26)$$

The three-knot time spline function generates a new hazard function expression

$$h(t, x) = h_0(t) \exp[\beta_1 x_1 + \beta_2 x_1 t + \beta_3 x_1 S_1(t)] \quad (27)$$

For the unit change of x_1, the corresponding LHRF is

$$LHRF = \beta_1 + \beta_2 t + \beta_3 S_1(t) \quad (28)$$

The test of $\beta_3 = 0$ is a nonlinear time-dependent test, while the test of $\beta_2 = \beta_3 = 0$ is the test of PH assumption. The LHRF estimate and its corresponding confidence interval can be illustrated to visually measure the time-dependent property[11].

The most obvious limitation of cubic spline regression is to select the number and location of knots. With regard to the selection of the number and location of knots, 3-5 knots can generally meet the needs, and three-knot splines can also be directly extended to four-knot, five-knot or even more knots. Some authors suggest that the knot position should be selected at the fixed quantile of time. For example, select the 10th, 50th, 90th (95th) quantile for three knots, and

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

select the 5th, 25th, 75th, and 95th quantile for four knots. Using the quantile as the position of the knot allows us to use the information of the survival time distribution rather than the real value, so as to reduce the subjectivity of knot selection. Hess (1994) put forward suggestions on the number and location of knots: (1) The quantiles of follow-up time (i.e., time to censoring and time to death); (2) Near but not on the extreme value; (3) Roughly evenly distributed on the quantiles are roughly uniform; (4) Limit quantity.

4. Conclusion

This paper introduces several methods to test the proportional hazard hypothesis from two aspects: graphical method and hypothesis test method. Among them, several common graphical methods have their advantages and disadvantages. Cox&K-M comparison method is difficult to determine whether the difference between curves is due to sampling error or real trend and the cumulative hazard function method cannot provide correction suggestions when the data does not meet the PH assumption. In addition, they need to partition covariates when analyzing quantitative data. The Schoenfeld residual graphical method provides the time-dependent information of the covariates, which is helpful to modify the model without partitioning the covariates, but it is difficult to evaluate the trend of the scattered points. In general, the graphical method is simple, clear, and does not require complex calculations, and thus plays an important role in statistical diagnosis. Because of its intuitive nature, it is easier to be accepted by the majority of practical workers. However, the graphical method depends largely on the subjective judgment of researchers.

Compared with the graphical method, the hypothesis test method is clear and easy to interpret, and can give more objective and accurate convincing conclusions. Among them, Nicholas (1997) showed that compared with other hypothesis testing methods, the time-dependent covariate method, the linear correlation test and the weighted residual score test have higher accuracy in testing the PH assumption, which are also emphasized in this paper. In addition, this paper also elaborates the method of using cubic spline function to estimate covariates and then test PH assumption. This method does not need to determine the specific form of the time function, it can avoid the result deviation caused by the model selection error and improve the testing efficiency.

Although the hypothesis test method is more formal and reliable, the test results are affected by the sample size. When the sample size is small, the test sensitivity decreases. When the sample size is too large, the original hypothesis may be rejected due to probability reasons. Therefore, in practical application, the graphical method is more popular in the existing research. Generally, it is only necessary to judge that the established model approximately meets the PH assumption. When by graphical method it is difficult to determine the degree of deviation from PH assumption or whether the original model should be rejected and replaced by another model, proper judgment can be made in combination with the formal statistical test method of PH assumption and the actual situation of the data. If the Cox PH assumption is not satisfied, it can be modified by fitting the stratified Cox model or the Cox model with time-dependent covariate interaction term on the basis of full analysis of the original data.

References

[1] Cox DR (1972). Regression models and life-tables(with discussion). Journal of the Royal Statistical Society, Series B: Methodology, vol.34, p.187-220.

[2] Harrel FE, Lee KL (1986). Verifying assumptions of the Cox proportional hazards model. Proceedings of the Eleventh Annual SAS Users Group International Conference, p.823-828.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0004

[3] Hess KR (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. Statistics in medicine, vol.14, p.1707-1723.

[4] Schoenfeld D (1982). Partial residuals for the proportional hazards regression model. Biornetrika, vol.69, p.239-241.

[5] Grambsch PM, Therneau TM (1994). Proportional hazard tests and diagnostics based on weighted residuals. Biometrika, vol.81, p.515-526.

[6] Therneau TM, Grambsch PM (1990). Martingale-based residuals for survival models. Biomatrika, vol.77, p.147-160.

[7] Moreau T, O'Quigley J, Mesbah M (1985). A global goodness-of-fit statistic for the proportional hazards model. Applied Statistics, vol.34, p.212-218.

[8] O' Quigley J, Pessione F (1989). Score tests for homogeneity of regression effect in the proportional hazards model. Biometrics, vol.45, p.135-144.

[9] Robert J. Gray (1992). Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. JASA, vol.87, p.942-951.

[10] Ruohua Y, Wei L (2016). Design, development, and deployment of a wireless sensor network for detection of landslides. Health statistics in China, vol.33, no.2, p.345-349.

[11] Hongmei Y, Yongyong X, Dawei H (2002). Using Cubic Spline Function to Examine the Proportional hazard Hypothesis of Cox Model. Health statistics in China, vol.19, no.1, p.20-22.

[12] Nicholas HN (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Statistics in Medicine, vol.16, p.611-626.