ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0006

Prediction of Lung Dataset Based on Cox Proportional Hazards Model

Yue Fu*

School of Economics, Jinan University, Guangzhou, 510632, China

Abstract

The proportional hazards regression model (Cox model) is a semiparametric regression model. It is the most widely used multivariate analysis method in survival analysis. The PH assumption is the most important assumption of the Cox model. In this paper, we use the schoenfeld residual method to test whether the PH assumption is satisfied. Breslow method and the partial likelihood function are used to estimate the unknown part of the model. Finally, the lung dataset in the R survival package is used to showed the methods mentioned in this article.

Keywords

Cox model; PH Assumption; Schoenfeld residual; Breslow method; Partial likelihood function.

1. Introduction

Let t be time, x_{ij}, $1 \le i \le n$, $1 \le j \le p$ is a variable that affects survival time T, $X_i = (x_{i1}, x_{i2}, \dots x_{ip})$ is a p-dimensional covariate, the proportional hazards regression model (Cox model) is expressed as follows:

$$h(t, X_i) = h_0(t) \sum_{j=1}^{p} \beta_j x_{ij} \quad (1)$$

where $h_0(t)$ is the baseline hazard, which is related to time t and not to covariate X_i. $\beta = (\beta_1, \beta_2, \dots \beta_p)$ is the coefficient of the covariate X_i, $\sum_{j=1}^{p} \beta_j x_{ij}$ is related to the covariateX_i, not to time t.

The Cox model is a semiparametric regression model proposed by the British statistician D.R. Cox (1972) [1]. The model uses the final outcome and survival time as dependent variables to analyze the influence of many factors on survival at the same time. Since its inception, the Cox model has been widely used in medical follow-up studies. It is the most widely used multivariate analysis method in survival analysis. However, the Cox model need to meet the proportional hazards assumption (PH assumption) [1].

The PH assumption means that the ratio of the risk function of two covariates does not change over time, which is the most important assumption of the Cox model [2]. Therefore, before using the Cox model, it is necessary to determine whether the PH assumption is true. At present, there are three commonly used judgment methods: time covariate method, Schoenfeld residual method and Kaplan-Meier (K-M) curve method.

D.R. Cox (1972) [1] proposed the time covariate method, which introduces a time-dependent covariate g(t) in the Cox model for regression analysis. If the coefficient of the covariate is 0, the PH assumption is satisfied, otherwise it is not satisfied. Schoenfeld (1982) [4] defines the residuals. The Schoenfeld residual method tests whether the residuals are related to survival time. Grambsch and Therneau (1994) [5] scaled the schoenfeld residuals and proposed

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0006

weighted schoenfeld residuals. Hess KR (1995) [3] proposes that the PH assumption is satisfied according to whether the K-M curve is similar to the survival curve under the Cox model, and if the K-M curve of multiple classes is close to the survival curve under the Cox model, the PH assumption is satisfied. According to the results of Nicholas (1997) [6], the time covariate method has high accuracy and similar efficiency to the weighted Schoenfeld residual method.

This article focuses on the steps and methods in the Cox model. The methods of PH assumption testing and the estimation methods of the coefficients of the parametric part and the non-parametric part of the Cox model are described in detail. We demonstrate the feasibility of the method through the lung dataset in the R survival package.

2. Method

The observed object with X_i has the survival function

$$S(t, X_i) = P(T > t, X_i), \tag{2}$$

it is the probability the survival time T is greater than a certain time t. When T>t, the probability of the observed object dying at time t is

$$h(t, X) = \lim_{\Delta t \to 0} \frac{P(t \le T \le t + \Delta t | T \ge t, X)}{\Delta t}, \tag{3}$$

it is the risk function.

In the model (1.1), the Hazard Ratio is

$$HR = \frac{h(t, X_1)}{h(t, X_2)} = exp[\sum_{i=1}^{p} \beta_i (x_{1i} - x_{2i})], \tag{4}$$

where X_1 and X_2 are any two covariates.

According to (2.1) and (2.2), we can derive the survival function

$$S(t, x_i) = S_0(t)^{exp \sum_{j=1}^{p} \beta_j x_{ij}} \tag{5}$$

where $S_0(t)$ is the baseline survival rate.

In this paper, the schoenfeld residual method is used to test the PH assumption. Schoenfel defined the residuals in 1982.$\{X_i, i \in S_{1k}\}$ are covariates of the observed object whose event occurs at time t_k with indices in S_{1k}. $\{X_i, i \in S_{2k}\}$ are covariates of the observed object that is still at risk t time t_k with indices in S_{2k},

$$R_{ik} = (r_{ik1}, \ldots, r_{ikj}, \ldots, r_{ikp})$$
$$\hat{r}_{ikj} = x_{ij} - E(x_{ij} | i \in S_{2k}), i \in S_{1k}$$

It is the residual of the j-th covariate of $X_i, i \in S_{1k}$ at time t_k. $\hat{r}_{kj}$ is the residual of the j-th covariate at time t_k. It can be demonstrated,E$(\hat{r}_{kj}) = 0$, and $\hat{r}_{kj}$ is approximately uncorrelated, then the image of $\hat{r}_{kj}$ to t_k should fluctuate with 0 as the center, and if there is a trend, it violates the PH assumption.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0006

The unknown part of the model (1.1) is solved by the Breslow method and the partial likelihood function. The baseline cumulative hazard rate function expression in the Breslow method is:

$$H_0(t_i) = \sum_{t_j<t_i} \frac{n_i}{\sum_{j\in R_i} exp(\beta' X_j)} \tag{6}$$

According to the relationship between cumulative hazard rate and survival rate: $S(t) = exp(-H(t))$,the baseline survival rate $S_0(t)$ can be obtained.

The Cox proportional hazards model solves for the parameters in the model by constructing a partial likelihood function, considering only the samples which the event occurred. First, the survival time of n samples is arranged from smallest to largest and obtain an ordered time series:$t_1 \le t_2 \le \cdots \le t_n$.Taking time t_i as an example, the set of all samples whose survival time is greater than t_i is called the danger set R_i. The conditional likelihood function of a dead individual at time t_i is:

$$L_i = \frac{h_0(t) exp \sum_{k=1}^{p} \beta_k x_{ik}}{\sum_{m\in R_i} h_0(t) exp \sum_{k=1}^{p} \beta_k x_{mk}} \tag{7}$$

the expression of the partial likelihood function is:

$$L = \prod_{i=1}^{n} L_i \tag{8}$$

First, take the logarithm on both sides of the above formula, then derive the β, and make its derivative 0, finally solve to obtain a maximum likelihood estimate of the β.

3. Empirical analysis

In this paper, the lung dataset in the R survival package is selected as a sample. We select the first 200 samples as the training set, the remaining samples as the test set, take time as the time variable, take status as the state variable, and construct a Cox proportional survival model with indicators that can significantly distinguish different states as covariates.

3.1. Significance test

The Mann-Whitney U test was used to test the significance of all variables, and preliminarily determine whether these indicators showed obvious differences in different states. After passing SPSS, the final test results are as follows:

Null assumption	Sig.	Outcome
The distribution of age is the same on the category of status	0.053	Preserve
The distribution of sex is the same on the category of status	0.000	Rejection
The distribution of ph.ecog is the same on the category of status	0.000	Rejection
The distribution of ph.karno is the same on the category of status	0.005	Rejection
The distribution of pat..karno is the same on the category of status	0.004	Rejection
The distribution of meal.cal is the same on the category of status	0.427	Preserve
The distribution of wt.loss is the same on the category of status	0.317	Preserve

So we chose the sex, ph.ecog, ph.karno, pat.karno to create the Cox model.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0006

The significance test of the Cox model is mainly done by the basic assumption that the null assumption is H0: all regression coefficients are 0, and the alternative assumption H1: at least one coefficient is not equal to 0. The basic assumption is used to test whether the Cox is reasonable.

Log-likelihood value	chi-square	Df	sig
1408.129	45.553	16	0.000

sig=0.00<0.05 rejects the null assumption, we believe that there are factors with a non-zero biased regression coefficient, which deserve further analysis. Significance tests were performed on four indicators, and the results were as follows:

index	Sex	Ph.ecog	Ph.karno	Pat.karno
p	0.0021	0.00302	0.08989	0.08200

Sex and ph.ecog are closely related to status, while Ph.karno and Pat.karnoa have little influence on status.

3.2. Building Cox model

	coef	exp(coef)	se(coef)	z	p
Sex	-0.5181	0.5956	0.1712	-3.026	0.00248
Ph.ecog	0.4577	1.5804	0.1145	3.997	6.42e-05

The second column is the regression coefficient of the variable, the third column is the index of the regression coefficient, the fourth column is the standard deviation of the regression coefficient, the fifth column is the statistical value, and the sixth column is the P value of the variable significance test. Since the P of both variables is less than 0.05, the null assumption is accepted. Sex, ph.ecog are closely related to status.

the risk function can be written as:

$$\mathrm{h(t,X)} = h_0(t)exp(-0.518sex + 0.458ph.ecog)$$

3.3. Judge the PH assumption

The PH assumption test uses the Schoenfeld residuals for statistical testing. This method compares the P value of each covariate and the P value of the whole model with a set significance level, and the specific results are shown in the following table:

	chisq	df	p
sex	1.72	1	0.19
Ph.ecog	2.02	1	0.15
global	3.78	2	0.15

The second column is the statistical value, the third column is the degree of freedom, and the fourth column is the P value of the test. Since the P value of each variable and the P value of whole model are greater than 0.05, the null assumption is accepted and the null assumption is satisfied.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0006

3.4. Baseline risk ratio

In the above, a maximum likelihood estimate of the parameters has been given, but estimate of the baseline risk ratio $h_0(t)$ has been not given. We use the basehaz function in R to find the baseline cumulative risk rate and calculate the baseline survival rate.

time	H(t)	S(t)	time	H(t)	S(t)
5	0.004612951	0.99539767	296	0.634432375	0.53023638
11	0.018568107	0.98160322	300	0.634432375	0.53023638
12	0.023278085	0.97699076	301	0.64596729	0.52415529
13	0.028032083	0.97235717	303	0.657843001	0.51796739
15	0.032830932	0.96770215	305	0.670012226	0.51170232
26	0.037647975	0.9630519	306	0.682292176	0.50545707
30	0.042511071	0.95837986	310	0.707423698	0.49291245
31	0.04742111	0.95368571	315	0.707423698	0.49291245
53	0.057312958	0.9442985	320	0.720359322	0.48657739
54	0.062309853	0.93959171	329	0.733428001	0.48025983
60	0.072428193	0.93013253	332	0.733428001	0.48025983
61	0.077525722	0.92540322	337	0.74692618	0.47382076
62	0.082653862	0.92066977	340	0.760569303	0.46740026
65	0.092985998	0.91120626	345	0.77430025	0.46102627
81	0.103462803	0.90170956	348	0.788172329	0.45467503
88	0.114075712	0.8921904	350	0.802135215	0.44837057
92	0.119435173	0.88742153	351	0.816190105	0.44211285
93	0.124830604	0.88264641	353	0.844932822	0.42958622
95	0.135773299	0.87304053	356	0.844932822	0.42958622
107	0.146862517	0.86341267	361	0.859711832	0.42328404
110	0.152463308	0.8585904	363	0.890018273	0.41064825
118	0.158103393	0.8537615	364	0.905520307	0.40433145
122	0.163844079	0.84887437	371	0.937706495	0.39152477
131	0.169623618	0.84398242	376	0.937706495	0.39152477
132	0.1813213	0.8341673	382	0.937706495	0.39152477
135	0.187253539	0.82923347	384	0.937706495	0.39152477
142	0.193229878	0.82429247	387	0.954565287	0.38497946
144	0.199250979	0.81934423	390	0.972001991	0.37832488
145	0.211419443	0.80943448	394	0.98982528	0.37164162
147	0.217557171	0.80448161	404	0.98982528	0.37164162
153	0.223724694	0.79953522	413	0.98982528	0.37164162
156	0.236212039	0.78961323	426	1.008753393	0.3646733
163	0.255410224	0.77459867	428	1.02795074	0.35773931
166	0.268508326	0.76451906	429	1.047442602	0.35083382
167	0.275136539	0.75946842	433	1.067418812	0.34389503
170	0.281797464	0.75442647	442	1.087583872	0.33702982
175	0.28851404	0.74937628	444	1.108074137	0.33019426
176	0.295287205	0.74431778	450	1.129420867	0.32322039
177	0.302096672	0.7392666	455	1.151110657	0.31628529
179	0.315960162	0.72908849	457	1.17317715	0.30938242
180	0.322982604	0.72398645	458	1.17317715	0.30938242
181	0.337287738	0.71370345	460	1.196525411	0.30224256

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0006

time	H(t)	S(t)	time	H(t)	S(t)
182	0.344536908	0.7085484	473	1.220572066	0.29506132
186	0.351825224	0.70340305	477	1.245054935	0.28792509
189	0.359153113	0.69826743	511	1.245054935	0.28792509
194	0.36652351	0.69313984	519	1.271587863	0.28038605
196	0.36652351	0.69313984	520	1.299026393	0.27279726
197	0.374003332	0.68797462	524	1.356145803	0.25765191
199	0.381553401	0.68279992	529	1.356145803	0.25765191
201	0.396836108	0.67244422	533	1.387147459	0.24978682
202	0.40454988	0.6672771	543	1.387147459	0.24978682
207	0.412338384	0.66210019	550	1.420680222	0.24154965
208	0.420203083	0.6569134	551	1.420680222	0.24154965
210	0.428140887	0.65171959	558	1.456510151	0.23304816
212	0.43615785	0.64651567	559	1.456510151	0.23304816
218	0.444303192	0.64127096	567	1.49402833	0.22446661
222	0.452531907	0.63601578	574	1.53338318	0.21580433
223	0.460895929	0.63071831	583	1.573996008	0.20721549
225	0.460895929	0.63071831	588	1.573996008	0.20721549
226	0.469495505	0.62531766	613	1.617645015	0.1983653
229	0.478150227	0.61992906	624	1.663800044	0.18941781
230	0.486899138	0.61452901	641	1.712766378	0.18036614
239	0.495744309	0.60911737	643	1.763576144	0.17143071
243	0.495744309	0.60911737	654	1.816502469	0.16259343
245	0.504747545	0.60365795	655	1.872925766	0.15367339
246	0.513811244	0.5982113	687	1.931971207	0.14486236
259	0.513811244	0.5982113	689	1.99371801	0.13618813
266	0.513811244	0.5982113	705	2.060602036	0.12737726
267	0.523045938	0.59271243	707	2.129651187	0.11887875
268	0.532348245	0.5872244	728	2.209545773	0.10975049
269	0.541715112	0.58174963	731	2.294467465	0.10081507
270	0.551234702	0.57623789	735	2.385091371	0.09208057
272	0.551234702	0.57623789	740	2.385091371	0.09208057
276	0.551234702	0.57623789	765	2.489773386	0.08292876
279	0.551234702	0.57623789	791	2.603257565	0.07403202
283	0.561101601	0.57058017	806	2.603257565	0.07403202
284	0.571091108	0.56490873	814	2.750732422	0.06388106
285	0.591586674	0.55344845	821	2.750732422	0.06388106
286	0.601990771	0.54772016	840	2.750732422	0.06388106
288	0.612480768	0.54200461	883	3.031711061	0.04823304
291	0.623191654	0.53623024	965	3.031711061	0.04823304
292	0.623191654	0.53623024	1010	3.031711061	0.04823304
293	0.634432375	0.53023638	1022	3.031711061	0.04823304

3.5. Analysis of the results

In this paper, the ratio of the censored sample to the population sample is used as the best decision point. If the survival prediction value is greater than or equal to the decision point, it is determined to be censored, otherwise it is dead. Since the population sample was 228 and

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202304_6(4).0006

the dead sample was 63, the best decision point was 0.28. The test set is used to test the prediction accuracy of the established Cox model. The results are shown in the following table:

forecast / actual	Dead=2	Censored=1
Dead=2	7	5
Censored=1	2	14

There are 28 samples in the test set, of which 9 are dead samples and 19 are censored samples. The overall prediction accuracy of the model was 75%, with 77.8% probability predicting dead for the dead sample and 73.7% of the probability predicting as censored for the censored sample.

4. Conclusion

In this paper, we discuss two problems in the Cox model: PH Assumption testing, estimating baseline survival rate and linearity coefficients. There are three commonly used PH assumption testing methods: time covariate method, Schoenfeld residual method and Kaplan-Meier (K-M) curve method. We chose the Schoenfeld residual method. We estimated the baseline survival rate using Breslow method and the coefficients using partial likelihood function. Finally, the lung dataset is used to illustrate the modeling process. The scope of this discussion is limited, there are many questions about the Cox model that are worth investigating.

References

[1] DR. Cox (1972). Regression models and life-tables. J Roy Stat Soc. Series B (Methodological). vol.34, no.2,p.187–220.

[2] Z. Zhang, J. Reinikainen ,KA. Adeleke, ME. Pieterse , CGM. Groothuis-Oudshoorn (2018). Time-varying covariates and coefcients in Cox regression models. Annals of Translational Medicine. vol.6,no.7,p.121.

[3] KR. Hess (1995). Graphical methods for assessing violations of the proportional hazards assumption in cox regression. Statistics in Medicine, vol.14, no.15, p.1707-1723.

[4] D.Schoenfeld(1982). Partial residuals for the proportional hazands model. Biometrika, vol.69, p.52-55.

[5] PM. Grambsch, TM. Therneau (1994), Proportional hazards test and diagnostics based on weighted residuals. Biometrika. Vol.81, p.515−526.

[6] NH. Ng'andu (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model statistics in Medicine.vol.16, no.6, p.611-26.