Semi-parametric regression model for survival data: graphical visualization with R

Zhongheng Zhang

Department of Emergency Medicine, Sir Run-Run Shaw Hospital, Zhejiang University, School of Medicine, Hangzhou 310016, China
Correspondence to: Zhongheng Zhang, MMed. 351#, Mingyue Road, Jinhua 321000, China. Email: zh_zhang1984@hotmail.com.

Author's introduction: Dr. Zhongheng Zhang is a fellow physician working at Sir Run Run Shaw Hospital. He graduated from School of Medicine, Zhejiang University in 2009, receiving Master Degree. His major research interests include hemodynamic monitoring in sepsis and septic shock, delirium, and outcome study for critically ill patients. He is experienced in data management and statistical analysis by using R and STATA, big data exploration, systematic review and meta-analysis. He has published more than 50 academic papers (science citation indexed) that have been cited for over 700 times. He has been appointed as reviewer for 10 journals, including *Journal of Cardiovascular Medicine*, *Hemodialysis International*, *Journal of Translational Medicine*, *Critical Care*, *International Journal of Clinical Practice*, *Journal of Critical Care*.

Zhongheng Zhang, MMed.

Abstract: Cox proportional hazards model is a semi-parametric model that leaves its baseline hazard function unspecified. The rationale to use Cox proportional hazards model is that (I) the underlying form of hazard function is stringent and unrealistic, and (II) researchers are only interested in estimation of how the hazard changes with covariate (relative hazard). Cox regression model can be easily fit with coxph() function in survival package. Stratified Cox model may be used for covariate that violates the proportional hazards assumption. The relative importance of covariates in population can be examined with the rankhazard package in R. Hazard ratio curves for continuous covariates can be visualized using smoothHR package. This curve helps to better understand the effects that each continuous covariate has on the outcome. Population attributable fraction is a classic quantity in epidemiology to evaluate the impact of risk factor on the occurrence of event in the population. In survival analysis, the adjusted/unadjusted attributable fraction can be plotted against survival time to obtain attributable fraction function.

Keywords: Survival analysis; nonparametric model; Cox proportional hazards; population attributable fraction; relative hazard

Submitted May 06, 2016. Accepted for publication Jun 20, 2016.
doi: 10.21037/atm.2016.08.61
View this article at: http://dx.doi.org/10.21037/atm.2016.08.61

Introduction

A fully parametric hazard function describes the basic underlying distribution of survival time and how that distribution changes as a function of covariates. If we want to describe the circuit life span in continuous renal replacement therapy as a function of serum ionized calcium and pH value, a fully parametric hazard function is required. However, if we want to see whether circuit life span is longer under acidosis (pH<7.35) when compared with that under normal condition, a complete description of survival time is of secondary importance to a description of how serum pH value modifies the survival time (1). A full description of survival time requires assumption of underlying mathematical model, which may be unnecessarily stringent and unrealistic. Survival time model that leaves its dependence on time unspecified but has a fully parametric regression structure is called semi-parametric regression.

The general form of hazard function is written as:

$$h(t,x,\beta)=h_0 \bullet r(x,\beta) \quad [1]$$

where h_0 reflects how hazard function changes with survival time, and $r(x,\beta)$ characterizes how hazard function changes with covariates. Cox has proposed exponential function for r(), and the hazard function is written as:

$$h(t,x,\beta)=h_0 \bullet e^{x\beta} \quad [2]$$

when x changed from x_0 to x_1, the hazard ratio is

$$HR(t,x_1,x_0)=\frac{h(t,x_1,\beta)}{h(t,x_0,\beta)}=\frac{h_0(t)\bullet e^{x_1\beta}}{h_0(t)\bullet e^{x_0\beta}}=e^{\beta(x_1-x_0)} \quad [3]$$

In the literature, the model is termed Cox proportional hazard model (2). Researchers are interested in the parameter β, which is interpreted as changing rate of hazard when the covariate changed by (x_1-x_0) unit. Note that the baseline hazard function $h_0(t)$ remains unknown, that is why the model is called semi-parametric model (3).

Cox proportional hazard model

Ovarian cancer survival data (ovarian) is used to illustrate fitting a Cox proportional hazard model. The study investigated survival in a randomized trial comparing two treatments for ovarian cancer (4).

```
> library(survival)
> cph.ovarian<-coxph(Surv(futime, fustat)~rx+age , ovarian)
> summary(cph.ovarian)
Call:
coxph(formula = Surv(futime, fustat) ~ rx + age, data = ovarian)

 n= 26, number of events= 12

       coef  exp(coef)  se(coef)  z       Pr(>|z|)
rx    -0.80397 0.44755   0.63205  -1.272  0.20337
age   0.14733  1.15873   0.04615  3.193   0.00141**
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

          exp(coef)  exp(-coef)  lower .95  upper .95
rx        0.4475     2.234       0.1297     1.545
age       1.1587     0.863       1.0585     1.268

Concordance= 0.798 (se = 0.091 )
Rsquare= 0.457 (max possible= 0.932 )
Likelihood ratio test= 15.89 on 2 df, p=0.0003551
Wald test     = 13.47 on 2 df, p=0.00119
Score (logrank) test = 18.56 on 2 df, p=9.341e-05
```

The fitting Cox regression model appears pretty easy

with only one line of R code. The function Surv() creates a survival object. Right side of "~" symbol lists covariates and they are connected with "+" operator. The last argument specifies the dataset containing covariates. Coxph() returns an object of class *coxph*, containing parameters and statistics we are interested in. the summary() function gives a general glimpse of the content of *coxph* object. There are 26 observations and 12 of them have the event of interest. The next table shows the coefficients of corresponding covariates. Exponentiation of coefficient gives the hazard ratio. Age is significantly associated with hazard and hazard increases by 1.16 times with each year increase in age. Variable age is denoted by double "*" symbols, suggesting a statistical significance with P<0.01.

The index of concordance (Concordance =0.798 in our example) is a "global" index for validating the predictive ability of a survival model. It is the fraction of pairs in the data, where the observation with the higher survival time has the higher probability of survival predicted by your model. Concordance is equivalent to the area under the ROC curve in logistic regression model. A value of 1 indicates perfect agreement, and a value of 0.5 is an agreement that the model is no better than chance. In other generalized linear models, R-square represents the proportion of variation that can be explained by the model (5). However, the R-square in Cox model also depends on the proportion of censored values. In other words, a perfectly adequate model may have a low R-square due to high large of censored data. Mathematical details of R-square in Cox regression model have been well described (6-10). The likelihood ratio test explores the difference between model with and without covariates. In the example, this statistic follows a Chi-square distribution with 2 degrees of freedom and thus can be used to obtain P value. Score test can be interpreted in a similar way that the model containing variables *rx* and *age* is significantly better than null model.

Stratification

The stratified Cox proportional hazard model allows the forms of underlying hazard function to vary across levels of stratification variable. The general form of stratified Cox model is written as:

$$h(t|X, Z = j) = h_j(t) \bullet e^{x\beta} \quad [4]$$

where j is the number of levels in Z. The covariate Z is adjusted for without estimating its effect. Someone may ask the question: why not incorporate Z as a covariate instead of using it as a stratification factor? The reason is that the predictor may not satisfy proportional hazards assumption and it can be very complex to model hazard ratio for that predictor as a function of time. Furthermore, stratification allows for graphical checks of the proportional hazards assumption. A drawback of stratification is that stratifying unnecessarily (proportional hazard assumption is met) reduces estimation efficacy.

```
> cph.ovarian.str<-coxph(Surv(futime,
fustat)~rx+strata(age>60), ovarian)
> summary(cph.ovarian.str)
Call:
coxph(formula = Surv(futime, fustat) ~ rx + strata(age >
60),
 data = ovarian)

 n= 26, number of events= 12

       coef     exp(coef) se(coef)  z        Pr(>|z|)
rx     -0.4300  0.6505    0.6003   -0.716   0.474

       exp(coef)  exp(-coef)  lower .95  upper .95
rx     0.6505     1.537       0.2006     2.11

Concordance= 0.585 (se = 0.115 )
Rsquare= 0.02 (max possible= 0.846 )
Likelihood ratio test= 0.52 on 1 df, p=0.4707
Wald test     = 0.51 on 1 df, p=0.4738
Score (logrank) test = 0.52 on 1 df, p=0.4709
```

The output of stratified Cox model is similar to the previous one. However, coefficient for *age* is not estimated because it is now used as stratification factor. Instead, two coefficients are estimated for variable *rx*, showing that the treatment effects are beneficial for young women and it is harmful for those aged over 60 years. One may wish to graphically examine the fitted model. Survival curves of patients with and without treatment stratified by age are depicted. In the survfit() function, a data frame with the same variable names as those that appear in the coxph formula is specified. The curve(s) produced corresponds to a cohort whose covariates equal to the values in *newdata*. If a covariate is not specified, the mean of the covariate used

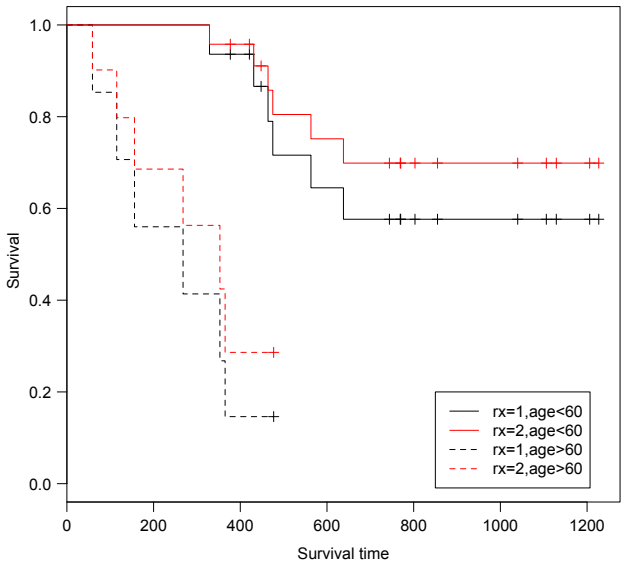

Figure 1 Comparison of survival probability of treated and untreated patients, stratified by age group.

in the *coxph* fit is used. The following example specifies patients with and without treatment (rx=1 and 2).

```
> strata.fit<-survfit(cph.ovarian.str,newdata=data.
frame(rx=c(1,2)))
> summary(strata.fit)
Call: survfit(formula = cph.ovarian.str, newdata = data.
frame(rx = c(1,
 2)))
```

age > 60=FALSE

time	n.risk	n.event	survival1	survival2
329	19	1	0.936	0.958
431	16	1	0.866	0.911
464	14	1	0.790	0.858
475	13	1	0.716	0.805
563	12	1	0.645	0.752
638	11	1	0.576	0.699

age > 60=TRUE

time	n.risk	n.event	survival1	survival2
59	7	1	0.853	0.902
115	6	1	0.707	0.798
156	5	1	0.560	0.686
268	4	1	0.414	0.563
353	3	1	0.268	0.424
365	2	1	0.146	0.286

The output shows survival curves of treated and untreated patients, stratified by age group. The first column is survival time. The last two columns display estimated survival probabilities stratified by treatment group. The result can also be visualized with the following syntax (*Figure 1*).

```
> plot(strata.fit,xlab="Survival time",ylab="Survival",lty=
c(1,1,2,2),col=c(1,2,1,2))
> legend(850,0.2,legend=c("rx=1,age<60","rx=2,age
<60","rx=1,age>60","rx=2,age>60"),col=c(1,2,1,2),lty
=c(1,1,2,2))
```

Visualization of relative importance of covariates

The interpretation of fitted Cox proportional hazards model depends on regression coefficients, significance level and prevalence of covariate patterns. Also, subject-matter audience may be interested in the importance of covariates in study population. In other words, the importance of a covariate is determined not only by coefficient but also by its distribution in a population. Karvanen and Harrell proposed the relative-hazard plot to visualize relative importance of covariates in proportional hazards model (11). The basic idea is to rank the covariate values and plot relative hazard against ranks. The relative hazard is scaled to the reference hazard. Reference hazard is related to the median of covariates. From this definition, it is obvious that the relative hazard can vary depending on the distribution of covariate values in a population. Rank-hazard plot can be created using rankhazardplot() function. Let's first install the package and load it onto the workspace. In order to make comparisons of relative importance for all covariates, a full model including all available covariates is built.

```
> library(rankhazard)
> install.packages("rankhazard")
> cph.full<-coxph(Surv(futime, fustat)~rx+age+resid.
ds+ecog.ps , ovarian,x=TRUE)
> rankhazardplot(cph.full,data=ovarian)
Y-axis range: 0.106 9.06

Relative hazards for each covariate:
          Min.      1st Qu.   Median   3rd Qu.   Max.
```

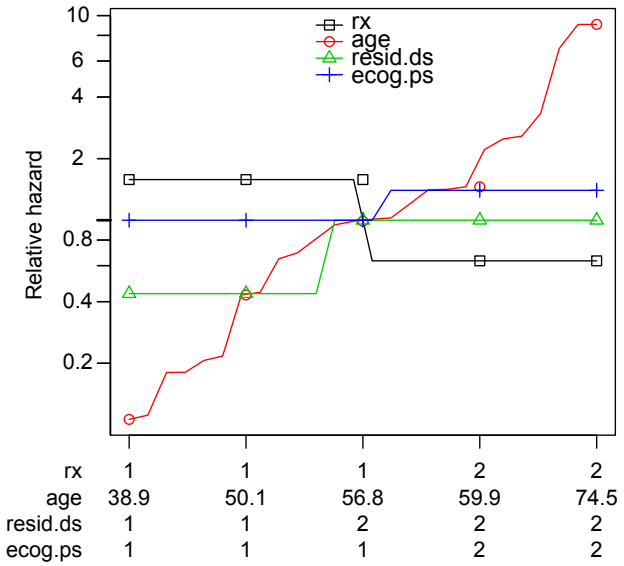

Figure 2 Rank-hazard plot of proportional hazards model where the hazard of ovarian cancer is explained by age (age), residual disease (resid.ds), treatment (rx) and ECOG performance status (ecog.ps). At the horizontal axis, the minimum, first quartile, median, third quartile and maximum values of each covariate are reported.

rx	0.633	0.633	1.11	1.58	1.58
age	0.106	0.434	1.00	2.03	9.06
resid.ds	0.438	0.438	1.00	1.00	1.00
ecog.ps	1.000	1.000	1.00	1.40	1.40

The output of function rankhazardplot() shows the Y-axis range. Covariates are scaled to an interval of [0,1]. By default, the minimum, first quartile, median, third quartile and maximum values of each covariate are reported. *Figure 2* shows that treatment is the most important determinant of survival at young age, but becomes less important for old age. Similar to the concept of population attributable fraction that the importance of a risk factor is influenced by its prevalence (12), the X-axis of relative-hazard plot displays scaled covariates for the ranking of their relative importance.

Test the proportional hazards assumption

Interpretation and use of Cox proportional hazards model depends on the proportional hazards assumption. Log-hazard function of proportional hazards model takes the form

$$In\left[h\left(t,x,\beta\right)\right]=In\left[h_0\left(t\right)\right]+x\bullet\beta \quad [5]$$

This function contains log of the baseline hazard function $In\left[h_0\left(t\right)\right]$ and linear predictor $x\bullet\beta$. x and β are highlighted in bold to represent vectors. The proportional hazard assumption dictates the baseline model as a function of time, not of the covariates. Suppose the covariate x has two levels. A plot of log-hazard over time will produce two continuous curve, one for x=0, $In\left[h_0\left(t\right)\right]$, and the other for x=1, $In\left[h_0\left(t\right)\right]+x\bullet\beta$. The difference between the two curves at any time points is β. If log-hazard functions produced by different levels of a given covariate are not equidistant over time, the proportional hazards assumption is violated. Grambsch and Therneau proposed one easily performed test and an associated graph to examine the critical assumption (13).

```
> cox.zph(cph.ovarian)
```

	rho	chisq	P
rx	0.2072	0.518	0.472
age	-0.0918	0.113	0.736
GLOBAL	NA	0.729	0.695

The cox.zph() function directly performed test for proportional hazards assumption. The output is a table containing rho, Chi-square and P value. Rho is the correlation coefficient between transformed survival time and the scaled Schoenfeld residuals. The correlation coefficient follows a chi-square distribution and the statistic is present in the second column. A P value is given for each covariate. A significant P value indicates that the proportional hazards assumption is violated for that covariate. For the global test there is no correlation and NA is entered into the cell.

```
> par(mfrow=c(2,1))
> plot(cox.zph(cph.ovarian))
```

Before drawing plots, rows and columns of graphical layout should be specified. The number of panels is determined by the number of covariates. In the example, there are two covariates and thus two panels are defined. If there is only one pannel in the graphical device, only one covariate can be displayed. The Y-axis of the plot is scaled Schoenfeld residuals and X-axis is survival time (*Figure 3*). The proportional hazards

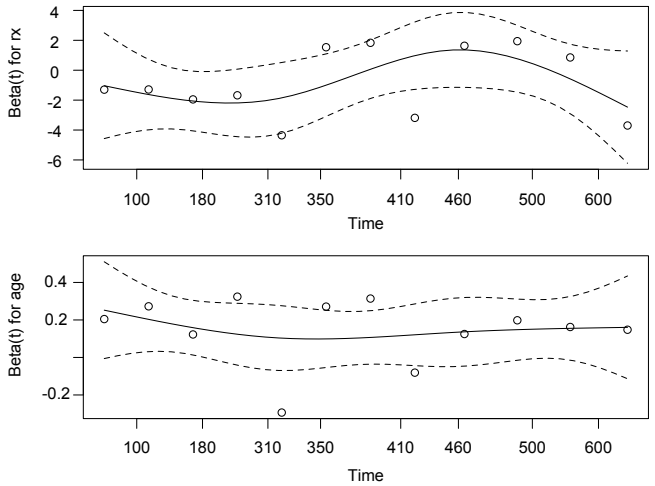

Figure 3 Plotting scaled Schoenfeld residuals against survival time to examine the proportional hazards assumption. The smoothed curve is a flat one for variable age, but there are small curvatures for the treatment (rx).

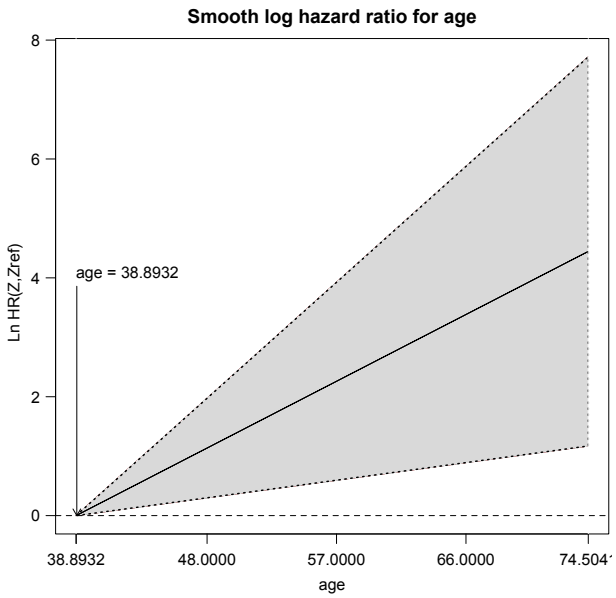

Figure 4 The hazard ratio curve shows that the age is in linear relation to the log-hazard, but the confidence interval is wide due to small number of observations.

assumption dictates that the residual should not change with survival time. Thus, the smoothed curve should be a flat one. Due to limited number of observations in the example, the sensitivity of the test can be quite low.

Hazard ratio curves for continuous predictors

The Cox proportional hazards model assumes that the continuous predictors are in linear association with log-hazard. However, this assumption may not true in reality. Visualization of the relationship between continuous predictor and hazard ratio helps investigators to check for the linearity assumption. Flexible hazard ratio curve can be plotted using *smoothHR* package.

```
> install.packages("smoothHR")
> library(splines)
> library(smoothHR)
> hr.plot<- smoothHR(data=ovarian, coxfit=cph.full)
> plot(hr.plot, predictor="age", prob=0, conf.level=0.95)
```

The smoothHR() function requires data frame containing covariates and the fitted Cox regression model. The prob argument specifies the reference value of the covariate for hazard ratio. The reference value is the minimum of the hazard ratio curve for prob=0. The hazard ratio curve shows that the age is in linear relation to the log-hazard, but the confidence interval is wide due to small number of observations (*Figure 4*).

Attributable fraction function

Population attributable fraction (PAF) is defined as

$$A = \frac{P(D=1) - P(D=1|Z=0)}{P(D=1)} \quad [6]$$

where D is the binary disease status and Z is the binary exposure indicator (14). PAF is defined as "the reduction in incidence that would be achieved if the population had been entirely unexposed, compared with its current (actual) exposure pattern", and it aims to evaluate the impact of risk factor on the occurrence of event in the population (15). Unlike relative risk, PAF considers the prevalence of risk factors in the population and thus quantifies the population impact of risk factors. PAF can be extended to adjusted attributable fraction (AF) which is defined as the reduction in incidence if some risk factors are eliminated from the population while other risk factors retain their actual levels. When AF is expressed as a function of survival time, it is called attributable fraction function (16). The package *paf* in R is composed to do the task.

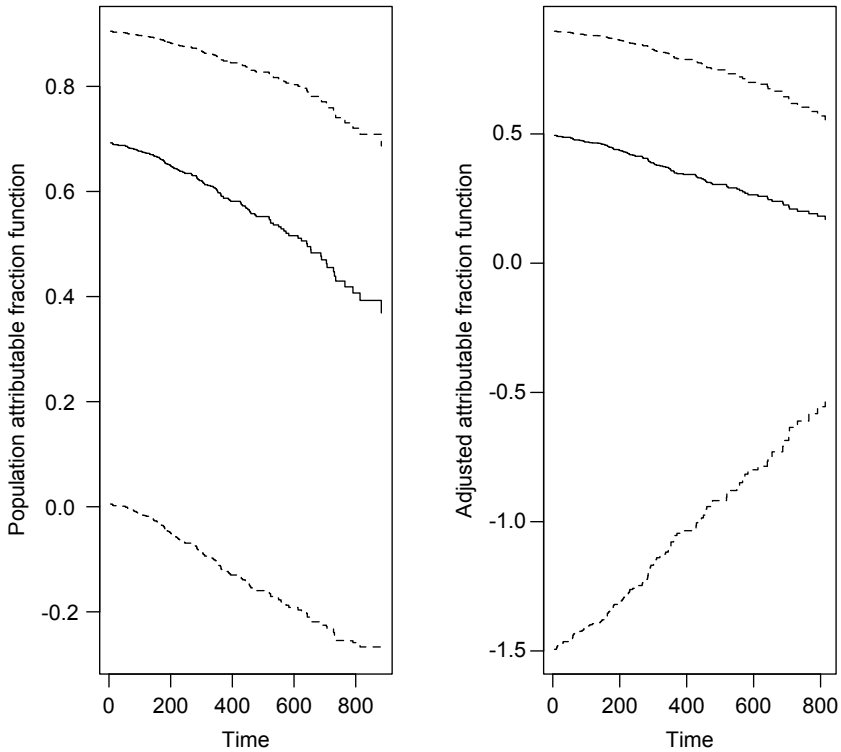

Figure 5 Estimation of attributable fraction functions for NCCTG Lung Cancer Data. The left panel shows the estimates of the population attributable fraction function: the solid curves pertain to the point estimates; and the dashed curves show the corresponding 95% confidence limits; the right panel shows the point estimate of the adjusted attributable fraction function and the corresponding 95% confidence limits.

```
> install.packages("paf")
> library(paf)
> par(mfrow=c(1,2))
> paf.adj<-paf(Surv(time, status)~sex+age+ph.
ecog+ph.karno+pat.karno+meal.cal+wt.loss,
data=lung, cov=c('age'))
> paf.pop<-paf(Surv(time, status)~age, data=lung,
cov=c('age'))
> plot(paf.pop,ylab="Population attributable fraction
function")
> plot(paf.adj,ylab="Adjusted attributable fraction
function")
```

The lung dataset (NCCTG Lung Cancer Data) is employed as worked example (17). Adjusted/unadjusted AF functions of a set of covariates are computed based on the Cox model. The first argument of paf() function is a formula object for the Cox model. Covariates of interest are specified in the *cov* argument and their AF functions are to be plotted against survival time. In the example, the variable age is investigated for its adjusted and unadjusted (population) attributable fraction function. The solid lines pertain to the point estimates of attributable fraction function and the dashed lines show the 95% confidence limits (*Figure 5*). It is seen from the figure that after adjustment for covariates the attributable fraction of age is decreased.

Acknowledgements

None.

Footnote

Conflicts of Interest: The author has no conflicts of interest to declare.

References

1. Zhang Z, Ni H, Lu B. Variables associated with circuit life span in critically ill patients undergoing continuous renal replacement therapy: a prospective observational study.

ASAIO J 2012;58:46-50.
2. Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society 1972;34:187-220.
3. Hosmer DW Jr, Lemeshow S, May S. Applied survival analysis: regression modeling of time to event data, second edition. Hoboken: Wiley-Interscience; 2008.
4. Edmonson JH, Fleming TR, Decker DG, et al. Different chemotherapeutic sensitivities and host factors affecting prognosis in advanced ovarian carcinoma versus minimal residual disease. Cancer Treat Rep 1979;63:241-7.
5. Mittlböck M, Schemper M. Explained variation for logistic regression. Stat Med 1996;15:1987-97.
6. Kejžar N, Maucort-Boulch D, Stare J. A note on bias of measures of explained variation for survival data. Stat Med 2016;35: 877-82.
7. Schemper M, Stare J. Explained variation in survival analysis. Stat Med 1996;15:1999-2012.
8. O'Quigley J, Xu R, Stare J. Explained randomness in proportional hazards models. Stat Med 2005;24:479-89.
9. Xu R, O'Quigley J. A measure of dependence for proportional hazards models. Journal of Nonparametric Statistics 1999;12:83-107.
10. Royston P. Explained variation for survival models. Stata Journal 2006;6:83-96.
11. Karvanen J, Harrell FE Jr. Visualizing covariates in proportional hazards model. Stat Med 2009;28:1957-66.
12. Deubner DC, Tyroler HA, Cassel JC, et al. Attributable risk, population attributable risk, and population attributable fraction of death associated with hypertension in a biracial population. Circulation 1975;52:901-8.
13. Grambsch PM, Therneau TM. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 1994;81:515-26.
14. Levin ML. The occurrence of lung cancer in man. Acta Unio Int Contra Cancrum 1953;9:531-41.
15. Rothman KJ, Greenland S, Lash TL. Modern Epidemiology. 3rd ed. Philadelphia: LWW; 2012.
16. Chen L, Lin DY, Zeng D. Attributable fraction functions for censored event times. Biometrika 2010;97:713-26.
17. Loprinzi CL, Laurie JA, Wieand HS, et al. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. J Clin Oncol 1994;12:601-7.

Cite this article as: Zhang Z. Semi-parametric regression model for survival data: graphical visualization with R. Ann Transl Med 2016;4(23):461. doi: 10.21037/atm.2016.08.61

