Applied Statistical Modeling 2

Lorenzoni Valentina
valentina.lorenzoni@santannapisa.it

Institute of Management - Scuola Superiore Sant' Anna, Pisa

12nd June 2025

Extension of the Cox proportional hazard model

The Cox proportional hazard model allows modelling survival data efficiently and easily, specifying a model in which it is assumed a common unspecified baseline hazard function that does not depend on time and the effect of covariates is assumed to be constant over time

$$h(t, x, \beta) = h_0(t)e^{x\beta}$$

The Cox model also assumes that the observation is continuous over time and that subjects could be right censored

Extension of Cox proportional hazard model

However, there are circumstances in which the assumptions of the Cox model do not hold because

- the effect of one or more covariates is not constant over time
- there are competing events or multiple events that occur on the same subject
- data are interval censored

in all these cases, we need to extend the Cox proportional hazard model or even use different regression models

Extension of Cox proportional hazard model

Extending the Cox proportional hazard model is generally easier than using alternative regression models, and this is sometimes feasible when the assumptions underlying the traditional Cox proportional hazard model are not verified

In particular, when some deviation from the proportional hazard assumption is detected, it is possible to accommodate the Cox proportional hazard model for the inclusion of the so-called *time-dependent* or *time-varying* covariates

Extension of Cox proportional hazard model

Time-dependent or *time-varying* covariates are variables for which

- a deviation from the proportional hazard assumption has been detected by the inspection of Schoenfeld residuals or other approaches discussed
- it is well known from previous studies that they could not be considered constant over time
- data are collected continuously over time and not just at baseline

Extension of Cox proportional hazard model

In the case of covariates collected just at baseline and for which deviation from the proportional hazard assumption has been detected or it is well-known from previous studies, possible extensions of the Cox proportional hazard model include

- the inclusion of *time-dependent* covariates
- the use of the stratified proportional hazard model

"Cox proportional hazard model" with time-varying covariate

When a covariate measured just at baseline does not satisfy the proportional hazard assumption, we have already discussed the possibility of eventually including it in a proportional hazard model by modelling its interaction with a function (i.e., linear or whatever is appropriate)

However, both in this case and in the case of variables that are recorded on multiple occasions over time (or could be derived, for example, age that could be derived by adding observation time to baseline age), it is possible to "appropriately" extend the Cox proportional model to account for time-varying covariate

"Cox proportional hazard model" with time-varying covariate

Indicating with $x(t)$ a covariate whose values depend on time, so they are determined by time t, the proportional hazard function could be generalized by the inclusion of time-varying covariate as follows

$$h(t, x(t), \beta) = h_0(t)e^{x(t)\beta}$$

which is no longer proportional, even it may also include fixed covariates, those for which $x(t) = x(t = 0) = x$

"Cox proportional hazard model" with time-varying covariate

In such a model, the interpretation of fixed covariates (where present) is the same as in the traditional Cox proportional hazard model, similar is also the derivation of the associated standard error even if, obviously, the partial likelihood function is again modified for the inclusion of time-varying covariates as follows

$$l_p(\beta) = \prod_{i=1}^{n} [\frac{e^{x_i(t_i)\beta}}{\sum_{l \in R(t_i)} e^{x_i(t_i)\beta}}]^{c_i}$$

"Cox proportional hazard model" with time-varying covariate

The inclusion of time-covariate in a regression model is both conceptually and practically hard, indeed the interpretation of model coefficients for time-varying covariates is not straightforward, and different approaches could be considered to include time-varying covariates in a regression model

Stratified Cox proportional hazard model

Stratified Cox proportional hazard model could be used when the effect of one (or more) covariate is not constant over time and the estimation of its effect is of secondary importance as compared to other covariates

The stratified model extends the traditional Cox model by introducing a stratum-specific baseline hazard function (as opposed to a common baseline hazard function) where strata are determined by the values of the covariates showing or known to do not satisfy the proportional hazard assumption

Stratified Cox proportional hazard model

The stratified Cox proportional hazard model could be specified using the following

$$h_s(t, x, \beta) = h_{s0}(t)e^{x\beta}$$

where $s = 1, 2, ..., S$ are the strata, and the model could be specified assuming a common slope (over strata) for all the other covariates, or even accommodating the possibility of having different slopes by adding the interaction with the covariate determining strata

Stratified Cox proportional hazard model

Even if the partial log-likelihood has a different form, as it is derived considering the product of the contributions of each stratum

$$l_{sp}(\beta) = \prod_{i=1}^{S} l_{sp}(\beta)$$

where $l_{Sp}(\beta)$ represents the contribution of the single stratum that is

$$l_{sp}(\beta) = \prod_{i=1}^{n_s} [\frac{e^{x_{si}\beta}}{\sum_{j \in R(t_{si})} e^{x_{si}\beta}}]^{c_{si}}$$

with n_s being the number of observations in the s_{th} stratum, t_{si} is the i_{th} obserevd value of time in the s_{th} stratum, c_{si} is the value of the censoring status at t_{si}, $R(t_{si})$ are subjects in stratum s in the risk set at time t_{si}, while x_{si} is the vector of p covariates

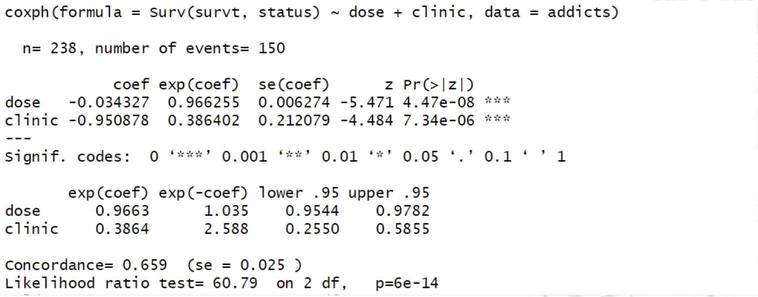
Stratified Cox proportional hazard model

Interpretation of model coefficients, their significance, and all the other approaches discussed to assess model assumptions remain the same as discussed for the traditional Cox proportional hazard model

Only in the case of a model specified allowing the possibility to have different slopes for the different strata, there is the need to pay attention to the interpretation of model coefficients taking into account the interaction between the covariate of interest and the covariate identifying strata

Stratified Cox proportional hazard model

Consider the proportional hazard Cox model adapted to the *addicts* data and aiming at assessing the effect of *methadone dose* and *clinic* on treatment withdrawal



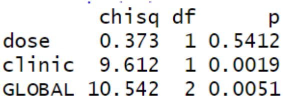
```
coxph(formula = Surv(survt, status) ~ dose + clinic, data = addicts)

  n= 238, number of events= 150

            coef exp(coef)  se(coef)      z Pr(>|z|)
dose   -0.034327  0.966255  0.006274 -5.471 4.47e-08 ***
clinic -0.950878  0.386402  0.212079 -4.484 7.34e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

       exp(coef) exp(-coef) lower .95 upper .95
dose      0.9663      1.035    0.9544    0.9782
clinic    0.3864      2.588    0.2550    0.5855

Concordance= 0.659  (se = 0.025 )
Likelihood ratio test= 60.79  on 2 df,   p=6e-14
```

Stratified Cox proportional hazard model

Results suggest a statistically significant effect

- for both the dose, with $HR = 0.965\ P - value < 0.001$, and
- clinic, $HR = 0.386\ P - value < 0.001$

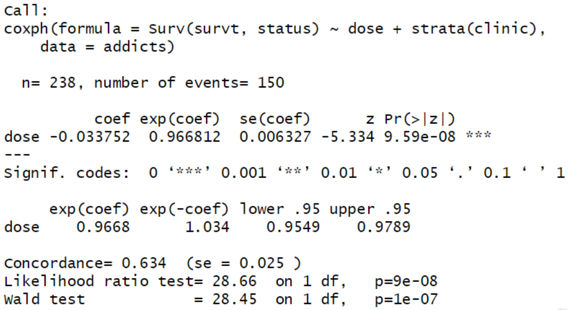
Anyway, the inspection of Schoenfeld residuals and the hypothesis testing for proportional hazard assumption suggested a deviation from the PH assumption for the overall model and for the variable *clinic*



```
        chisq df      p
dose    0.373  1 0.5412
clinic  9.612  1 0.0019
GLOBAL 10.542  2 0.0051
```

Stratified Cox proportional hazard model

If the effect of the variable *clinic* is of secondary importance, as compared to *methadone dose*, it is possible to adapt a Cox proportional hazard model including *methadone dose* as covariate and the variable *clinic* as strata

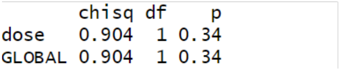


```
Call:
coxph(formula = Surv(survt, status) ~ dose + strata(clinic), 
    data = addicts)

  n= 238, number of events= 150 

          coef exp(coef)  se(coef)      z Pr(>|z|)    
dose -0.033752  0.966812  0.006327 -5.334 9.59e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

     exp(coef) exp(-coef) lower .95 upper .95
dose    0.9668      1.034    0.9549    0.9789

Concordance= 0.634  (se = 0.025 )
Likelihood ratio test= 28.66  on 1 df,   p=9e-08
Wald test            = 28.45  on 1 df,   p=1e-07
```

Stratified Cox proportional hazard model

The HR for dose is slightly different from that one obtained in the traditional proportional hazard model, but is still statistically significant

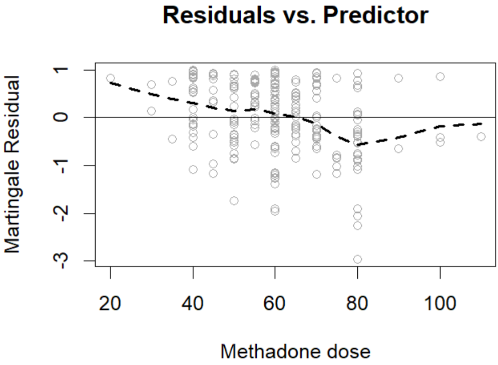
Moreover, stratifying there is no problem of deviation from the proportional hazard assumption for the overall model



```
         chisq df    p
dose     0.904  1 0.34
GLOBAL   0.904  1 0.34
```

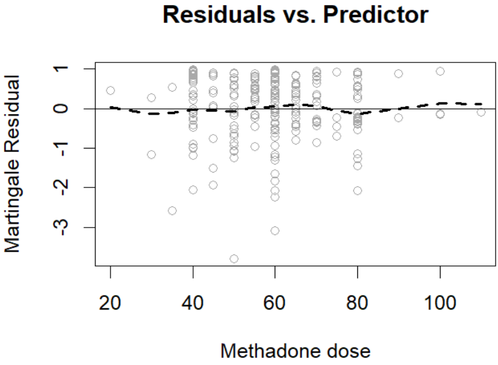
Stratified Cox proportional hazard model

In addition, while in the traditional model the plot of Martigale residuals showed a deviation from linearity of *methadone dose*



Stratified Cox proportional hazard model

With the stratified model also the linearity assumption seems to be met for *methadone dose*



Competing risk models

A situation in which the Cox proportional hazard model could not be used occurs when the event of interest is not a composite event, or even it is a specific type of event and there are other events that "compete" with the event of interest because the could preclude the possibility of experiencing the event of interest

Competing risk models

The typical situation where there is the need to take into account competing events is, for example, the one occurring in a clinical study where the event of interest is death from a specific cause (i.e., cardiovascular disease) and death from other causes represents a competing event

Competing risk models

When competing risks are present, the proper description of the survival experience is obtained considering the **cumulative incidence function (CIF)**, not the survival function

Competing risk models

Assume dealing with $j = 1, 2, ..., k$ types of events, the cause specific hazard function $h_j(t)$ describes the rate of failure at time t for the event j, given that subjects survived till at least time t

Competing risk models

If the k types of events are mutually exclusive, the overall hazard function could be obtained as

$$h(t) = \sum_{j=1}^{k} h_j(t)$$

thus the overall event-free survival function is

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

where $H(t) = \sum_{j=1}^{k} H_j(t)$ is the overall cumulative hazard function and $H_j(t)$ the cause specific culmulative hazard

$$H_j(t) = \int_0^t h_j(u)du$$

Competing risk models

The event-specific distribution function, called *subdistribution* or *cumulative incidence function (CIF)*, is then

$$F_j(t) = P(T \leq t, C = j) = \int_0^t h_j(u)S(u)du$$

and the overall distribution function is

$$F(t) = \sum_{j=1}^{k} F_j(t)$$

and $S(t) = 1 - F(t)$

Competing risk models

Suppose having a sample of *n* observation of follow-up time and event type (t_i, c_i), the CIF is the cumulative proportion of failure due to cause *j* at time *t* and could be estimated by summing the cause-specific failure rates conditional on being at risk of failure

$$\hat{F}_j(t) = \sum_{t_{(i)} \leq t} \hat{h}_j(t)\hat{S}(t_{i-1})$$

where t_i is the largest event time and $\hat{S}(t_{i-1})$ is the value of the Kaplan-Meier estimator that has already been discussed, and

$$\hat{h}_j(t) = \frac{d_{ij}}{n_i}$$

is the event-specific hazard estimator

Competing risk models

Consider a simple example of ten subjects for which we recorded the occurrence of two different competing events and related follow-up time

Table: Hypothetical data

Id	Time	Event type
1	9	2
2	27	0
3	18	0
4	30	1
5	10	1
6	33	2
7	13	1
8	19	2
9	24	2
10	4	1

Competing risk models

The estimated cumulative survival and the cause-specific hazard could be obtained by considering the calculus below

Table: Kaplan-Meier estimate of the overall survival probability and cause-specific hazard for the hypothetical data

Time	Event type	$\hat{S}(t_i)$	n_i	d_{i1}	d_{i2}	$h_1(t_i)$	$h_2(t_i)$
0	-	1	10	0	0	0	0
4	1	0.9	10	1	0	0.1	0
9	2	0.8	9	0	1	0	0.111
10	1	0.7	8	1	0	0.125	0
13	1	0.6	7	1	0	0.143	0
18	0	0.6	6	0	0	0	0
19	2	0.48	5	0	1	0	0.2
24	2	0.36	4	0	1	0	0.25
27	0	0.36	3	0	0	0	0
30	1	0.18	2	1	0	0.5	0
33	2	0	1	0	1	0	1

Competing risk models

Then the cause-specific cumulative incidence function could be obtained as shown below

Time	$\hat{S}(t_i)$	$h_1(t_i)$	$h_2(t_i)$	I_{1t_i}	I_{2t_i}	CIF_1	CIF_2
0	1	0	0	0	0	0	0
4	0.9	0.1	0	0.1	0	0.1	0
9	0.8	0	0.111	0	0.1	0.1	0.1
10	0.7	0.125	0	0.1	0	0.2	0.1
13	0.6	0.143	0	0.1	0	0.3	0.1
18	0.6	0	0	0	0	0.3	0.1
19	0.48	0	0.2	0	0.12	0.3	0.22
24	0.36	0	0.25	0	0.12	0.3	0.34
27	0.36	0	0	0	0	0.3	0.34
30	0.18	0.5	0	0.18	0	0.48	0.34
33	0	0	1	0	0.18	0.48	0.52

where $I_{ji} = h_j(t_i) * S(t-1)$

Competing risk models

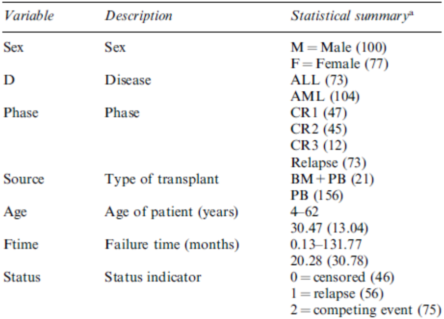
The simplest model that could be used to model those data is a proportional hazard model with cause-specific hazard function

$$h_{0j}(t)e^{x\beta}$$

which also allows for cause-specific estimates of effects, other common approach is to use the Fine-Gray subdistribution hazard model that allows for modeling the effect of covariates on the CIF

Competing risk models

Consider some examples of using the *bmtcrr* data



Variable	*Description*	*Statistical summary*[a]

Sex	Sex	M = Male (100)
		F = Female (77)

D	Disease	ALL (73) AML (104)
Phase	Phase	CR1 (47) CR2 (45) CR3 (12) Relapse (73)
Source	Type of transplant	BM + PB (21) PB (156)
Age	Age of patient (years)	4–62 30.47 (13.04)
Ftime	Failure time (months)	0.13–131.77 20.28 (30.78)
Status	Status indicator	0 = censored (46) 1 = relapse (56) 2 = competing event (75)