Exercise 8: Semiparametric Survival Analysis
2018 Spring

---

1 Cox proportional hazards model

A Cox proportional hazards model specifies the relationship between the hazards and the covariates:

$$h(t, \mathbf{X}) = h_0(t) \exp\left(\sum_{i=1}^{p} x_i \beta_i\right)$$

- $h_0(t)$: **baseline hazard rate**, time-dependent.
- $\mathbf{X}$: vector of explanatory variables, covariates.
- $\exp(\beta_i)$: **hazard ratio** for the coefficient β_i.
- The **ratio** between the predicted **hazard rate** of two individuals that differ by 1 unit of x_i.
- $\mathbf{X}\beta$: **Prognostic index (预后指数)**

There is a parameter β, but no parametric form for $h_0(t)$ in the model. That's the reason why we call it **semi-parametric model**.
Thus:

$$\begin{aligned} \log h(t, \mathbf{X}) &= \log h_0(t) + \mathbf{X}\beta \\ \log \tfrac{h(t,X)}{h_0(t)} &= \mathbf{X}\beta \end{aligned}$$

2 Accerlated Failure Time (AFT) Model

3 Cox's Proportional Hazards Model

In this section we will introduce the Cox's proportional hazards model, give a heuristic development of the **partial likelihood function**, and discuss adaptions to accomodate **tied observations**. We then explore some specific tests that arise from likelihood-based inferences based on the partial likelihood. Asymptotic properties of the resulting estimators and tests will also be covered.

3.1 PH Model

For n subjects with covariate vector Z and the survival outcome (U, δ) representing noninformatively right-censored values of a survival time T. That is, for subject i,

- Z_i: the covariate vector;
- T_i: the underlying survival time;
- C_i: the potential censoring time;
- $U_i = \min(T_i, C_i)$;
- $\delta_i = I(T_i \leq C_i)$;

- $T_i \perp C_i | Z_i$

One way to model a relationship between Z and T is by assuming $h(\cdot)$ **is functionally related to** Z:

$$T \sim \exp(\lambda_Z)$$

where $h(t) = \lambda_Z = \exp(\alpha + \beta Z) = \lambda_0 \exp(\beta Z)$ $(\lambda_0 = \exp(\alpha))$.
Thus, we might assume that $T_i \sim \exp(\lambda_0 \exp(\beta Z_i))$. Therefore, $\beta = 0$ means that λ_Z does not depend on Z, and also Z is not assocated with T.

3.1.1 Generalization

Let $h(t|Z)$ denote the hazard function for a subject with covariate Z. Suppose that

$$h(t|Z) = h_0(t) \times g(Z)$$

where $h_0(t)$ is a function of t without Z, and $g(Z)$ is a function of Z without t.
This can be called **multiplicative hazards model** or **proportional hazards model**. This factorization implies that

$$\frac{h(t|Z = Z_1)}{h(t|Z = Z_2)} = \frac{g(Z_1)}{g(Z_2)}$$

which is independent of t. That is the reason why it is called the **proportional hazards (PH) model**.

Example 1 *(***Cox's proportional hazards (Cox's PH) model***)*

$$g(Z) = \exp(\beta Z) \Rightarrow h(t|Z) = h_0(t) \times \exp(\beta Z)$$

where

$$\frac{h(t|Z = Z_1)}{h(t|Z = Z_2)} = \exp(\beta(Z_1 - Z_2))$$

For a scalar Z, $\exp(\beta)$ = hazard ratio corresponding to a unit change in Z.

Example 2 *(***Categorical** *Z)*

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix}, Z_1 = \begin{cases} 0 & R_x(treatment) = 0 \\ 1 & R_x(treatment) = 1 \end{cases}, Z_2 = \begin{cases} 0 & female \\ 1 & male \end{cases}$$

and $\beta = (\beta_1, \beta_2)$, then

$$h(t|Z) = \begin{cases} h_0(t) & R_x = 0, female \\ h_0(t) \exp(\beta_1) & R_x = 1, female \\ h_0(t) \exp(\beta_2) & R_x = 0, male \\ h_0(t) \exp(\beta_1 + \beta_2) & R_x = 1, male \end{cases}$$

3.2 Inference

Our inferential problems include:

- Estimate β and derives its statistical properties,

- Testing hypothesis $H_0 : \beta = 0$ or for part of β,
- Diagnostics of the assumptions

3.2.1 Estimation

How can we infer the coefficients β?

- Assumption 1: parametric form of $h_0(t)$, conduct parametric analysis (e.g., $h_0(t) = \lambda_0$)
- Assumption 2: arbitrary $h_0(t)$

The survival function:

$$S(u|Z) = (S_0(u))^{\exp(\beta Z)}$$

where

$$\begin{aligned} S_0(u) &= \exp(-\textstyle\int_0^u h_0(t)dt) \\ &= \text{survival function for someone with } Z = 0 \\ &= S(u|0) \end{aligned}$$

Also, we have $f(u|Z) = h(u|Z)S(u|Z)$.
Therefore , for n independent observations $(u_i, \delta_i, z_i), i = 1, \ldots, n$, the likelihood function is

$$\begin{aligned} L(\beta, h_0(\cdot)) &= \textstyle\prod_{i=1}^n f(u_i|z_i)^{\delta_i} S(u_i|z_i)^{1-\delta_i} \\ &= \textstyle\prod_{i=1}^n h(u_i|z_i)^{\delta_i} S(u_i|z_i) \\ &= \textstyle\prod (h_0(u_i)e^{\beta z_i})^{\delta_i} (e^{-\int_0^{u_i} h_0(t)dt})^{e^{\beta z_i}} \\ &= f(\text{data}, \beta, h_0(\cdot)) \end{aligned}$$

If we allow arbitrary $h_0(\cdot)$, then the **parameter space** is

$$\mathcal{H} \times \mathbb{R}^p = \left\{ (h_0(\cdot), \beta) : h_0(u) \geq 0 \text{ for all } u, \int_0^\infty h_0(u) = \infty \text{ and } \beta \in \mathbb{R}^p \right\}$$

where p is the dimension of the vector β. The condition

$$\int_0^\infty h_0(u)du = \infty$$

ensures that $S_0(\infty) = 0$.
In many application, the main goal is to make an inference about β and the underlying hazard $h_0(\cdot)$ is a nuisance function. Inference in such a settings are commonly called **semi-parametric**. Therefore, the standard likelihood theory, based on Euclidean parameter spaces, does not apply here.

3.2.2 Cox's partial likelihood estimator

Try to factorize $L(\beta, h_0(\cdot))$ into

$$L(\beta, h_0(\cdot)) = L_1(\beta) \times L_2(\beta, h_0)$$

where

- $L_1(\beta)$ is a function of β, whose maximum estimate $\hat{\beta}$ enjoys nice properties ($\hat{\beta} \xrightarrow{P} \beta$) and ($\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{\mathcal{L}} N$) although perhaps inefficient.

- $L_2(\beta, h_0(\cdot))$ is a function of $h_0(\cdot)$ and β which contains relatively little information about β.

Then, Cox recommends to infer β on the partial likelihood function $L_1(\beta)$.

3.3 Partial Likelihood Function, L_p

$$\begin{aligned} L_p = \prod_{i=1}^{d} q_i &= \prod_{i=1}^{d} \frac{h_0(t_i)\exp(\mathbf{X}_i\beta)}{\sum_{j\in R_i} h_0(t_i)\exp(\mathbf{X}_j\beta)} \\ &= \prod_{i=1}^{d} \frac{\exp(\mathbf{X}_i\beta)}{\sum_{j\in R_i}\exp(\mathbf{X}_j\beta)} \end{aligned}$$

where

- d: number of non-censored time points.
- R_i: risket set at time t_i.
- $q_i = \frac{h_i(t)}{\sum_{j\in R_i} h_j(t)}$ is the probability of failure at time t_i
- Only the non-censored subjects are included in the analysis for numerator.
- In the denominator term, all the subject in risk (including the censored) are included into the computation.

This function can also be rewritten as (for computational convenience):

$$L_p = \prod_{i=1}^{n} \left(\frac{\exp(\mathbf{X}_i\beta)}{\sum_{j\in R_i}\exp(\mathbf{X}_j\beta)} \right)^{\delta_i}$$

where

$$\delta_i = \begin{cases} 1 & \text{subject } i\text{failed} \\ 0 & \text{otherwise} \end{cases}$$

3.3.1 Log partial likelihood function

$$\begin{aligned} l(\beta) &= \log L_p \\ &= \sum_{i=1}^{d} \mathbf{X}_i\beta - \sum_{i=1}^{d} \log\left(\sum_{j\in R_i} \mathbf{X}_j\beta\right) \end{aligned}$$

Let $\frac{\partial l}{\partial \beta} = 0$, we can obtain the regression coefficients β through **Newton-Raphson** iterative approach.

3.4 Estimate of Coefficients and Hypothesis Tests

Example 3 *(**Survival analysis of nasal lymphoma patients**) Here is the follow-up data of 16 nasal lymphoma patients in a hospital:*

id	*gender*	*age*	*stage*	*bleed*	*rdx*	*chmx*	*days*	*status*
1	*1*	*45*	*2*	*2*	*0*	*1*	*578*	*1*
2	*0*	*36*	*2*	*2*	*0*	*1*	*1549*	*1*
3	*1*	*57*	*2*	*2*	*1*	*0*	*938*	*1*
4	*0*	*45*	*2*	*0*	*1*	*0*	*4717*	*0*
5	*0*	*42*	*2*	*0*	*1*	*1*	*4111*	*1*
6	*0*	*39*	*2*	*1*	*0*	*1*	*1245*	*1*
7	*1*	*38*	*2*	*1*	*1*	*1*	*4435*	*1*
8	*1*	*45*	*2*	*2*	*1*	*0*	*3750*	*1*
9	*1*	*30*	*2*	*0*	*1*	*0*	*3958*	*1*
10	*0*	*45*	*2*	*1*	*0*	*1*	*2581*	*1*
11	*0*	*45*	*3*	*1*	*0*	*1*	*3572*	*1*
12	*1*	*57*	*2*	*1*	*1*	*0*	*2938*	*1*
13	*0*	*57*	*2*	*2*	*0*	*1*	*1932*	*1*
14	*1*	*49*	*2*	*2*	*1*	*1*	*3205*	*1*
15	*1*	*33*	*2*	*1*	*0*	*1*	*3451*	*1*
16	*0*	*51*	*2*	*2*	*1*	*0*	*2363*	*1*

Analyze the data using the Cox's proportional hazards model, using the other six metrics as covariates:

```
library(survival)
lympho <- read.table("data/nasallym.dat", header=T)
for (i in c(2,4:7)){
        lympho[,i] <- factor(lympho[,i])
}
cox.mod <- coxph(Surv(days, status) ~ gender + age + stage + bleed +
   rdx + chmx,
                 data = lympho)
summary(cox.mod)
```

3.5 Cox's PH Models with Time-dependent Covariates

Since survival data is a time-series data, some covariates may also change over time, which we refer to as **time-dependent covariates**.
Here are some examples:

- Cumulative exposure to some risk factor,
- Smoking status,
- Heart (kidney) transplant status,
- Blood pressure

We might have more than one such covariate. For the i-th participants, we denote such covariates as:

$$Z_i(t) = (Z_{i1}(t), \ldots, Z_{iq}(t))^T$$

If the j-th covariate is time-independent, then $Z_{ij}(t)$ is constant over time.

Let $Z_i^H(t)$ denote the history of the time-dependent covariates up to time t,

$$Z_i^H(t) = \{Z_i(u), 0 \leq u \leq t\},$$

then we can define the hazard rate at time t conditional on this history:

$$\lambda(t|Z_i^H(t)) = \lim_{\Delta t \to 0} \frac{P[t \leq T_i \leq t + \Delta t | T_i \geq t, Z_i^H(t)]}{\Delta t}$$

This is the instantaneous rate of failure at time t, given that the individual was at risk at time t with a history of $Z_i^H(t)$. For such a conditional hazard rate, we may consider a proportional hazards model:

$$\lambda(t|Z_i^H(t)) = \lambda_0(t) \exp(\beta^T g(Z_i^H(t))),$$

More often we will choose $g(\cdot)$ as $g(Z_i^H(t)) = z_i(t)$, then

$$\lambda(t|Z_i^H(t)) = \lambda_0(t) \exp(\beta^T Z_i(t))$$

if we implicitly assume that the hazard rate at time t given the entire history of the covariates up to time t is only affected by the current values of the covariates at time t.

Example 4 *(***The effect of exposure to asbestos over time on mortality***) A sample of workers in a factory where asbestos is made were monitored for a period of time and data were collected on survival and asbestos exposure.*
For the i-th individuals, the data could be summarized as

$$(X_i, \delta_i, Z_i^H(X_i))$$

where

- $X_i = \min(T_i, C_i)$ *is the observed survival time or censoring time,*
- $\delta_i = I(T_i \leq C_i)$ *is the failure indicator,*
- $Z_i^H(X_i)$ *is the history of asbestos exposure up to time* X_i

Suppose we wish to use the above proportional hazards model with time-dependent covariates, then what should we use for the function $g(Z_i^H(t))$?

- Use the cumulative exposure (need extrapolation), i.e.

$$g(Z_i^H(t)) = \sum_j Z_i(u_{ij})(u_{ij} - u_{i(j-1)}),$$

- Use average exposure up to time t

$$g(Z_i^H(t)) = \frac{\sum_{u_{ij} < t} Z_i(u_{ij})}{\#\text{ of measurements up to } t}$$

- Use maximum exposure up to time t

$$g(Z_i^H(t)) = \max\{Z_i(u_{ij}) : u_{ij} < t\}$$

- Use the compound exposure term.

If we consider the model

$$\lambda(t|Z^H(t)) = \lambda_0(t)\exp(\beta^T Z(t))$$

then the partial likelihood function of β for this model is given by

$$PL(\beta) = \prod_u \left[\frac{\exp(\beta^T Z_{I(u)}(u))}{\sum_{l=1}^n \exp(\beta^T Z_l(u)) Y_(u)} \right]^{dN(u)}$$

where $I(u)$ is the indicator variable that identifies the individual label for the individuals who fail at time u.

3.6 Write a report on survival analysis

- Describe the event of interest (e.g., failure time)
- The start time and end time for the follow-up
- Type of censoring and the possible reasons for censoring
- The method for computing survival rate (Kaplan-Meier, or life table)
  - median survival time, or 5-year surval rate (estimate and confidence interval)
  - The statistical methods for comparing the survival rates (Logrank or Breslow) and also the p-value.
- Cox proportional hazards model for the relation between explantory variables and hazard:
  - hazard ratios and the corresponding confidence interval
  - hypothesis testing of the assumption of proportional hazards

3.7 Competing Risk Model

4 Frailty model

Exercises

1. (40 points) A clinical trial is intended to test a new treatment for malignant melanoma vs. current standard care. The main outcome is disease-free survival. Each patient has a time t in days after start of therapy that represents either the time of recurrence or death (if `status = 1`) or the end of the study (if `status = 0`). Each patient has two covariates we can use: `Tx = 1` if they are on the new therapy or `Tx = 0` if they are on standard of care. There are four subtypes of malignant melanoma, which we will characterize as `Type = A, B, C, or D`. The effect of the new therapy may differ among the four subtypes.

   (1) We can estimate the survival curve for patients on the new and standard therapy including all four subtypes using the `Kaplan-Meier product limit estimator`. Suppose at a given time t (in days after starting therapy), and for a particular subset of the patients, that there are 194 patients whose survival or censoring time is at least t. Suppose that 3 patients died or relapsed on day t and 2 patients had a censoring time of t. By what fraction does

the estimated survival curve drop at time t? How many patients are in the risk set just before t and just after t?

(2) Write down the Cox model for predicting survival from `Tx, Type, and the Tx by Type interaction` including a definition of the coefficients and their relationship to survival. State the important assumptions.

(3) If there are 20 patients at risk at a given time and 1 of them fails, write down the contribution (factor) to the partial likelihood from that failure time in terms of the model specification. Does it depend on the base hazard?

(4) Which will generate more accurate coefficient estimates, a study with 1000 patients of whom 900 survive to the end of the study without recurrence, or a study with 500 patients 300 of whom survive? Why?

(5) Describe the most appropriate hypothesis test for whether the interaction term is required in the model. How the test statistic be calculated? To what specific statistical distribution would the test statistic be compared?

(6) Suppose that the reference levels of the covariates are `Tx = 0` and `Type = A`. List all the coefficients in the model (symbolically). In terms of those coefficients, what would be the estimated log hazard ratio of a patient with `Type B` melanoma on `Tx = 1` to a patient with `Type C` melanoma on `Tx = 0`? What would be the estimated hazard ratio?

(7) How would you examine the proportionality assumption, graphically and/or with a statistical test?

(8) If it appears that the different subtypes have non-proportional hazards, how would you change the model so that this could be accommodated?
Solution
One way would be to use a strata term for a grouping variable that was non- proportional. The alternative way is to include the time-dependent covariate.

2. (55 points) Consider the data from Prevention of Events with Angiotensin Converting Enzyme Inhibition (PEACE) Trial. The goal of the study was to test whether ACE-inhibitor therapy, when added to modern conventional therapy, would reduce the rate of nonfatal myocardial infarction, death from cardiovascular causes, or revascularization in low-risk patients with stable coronary artery disease and normal or slightly reduced left ventricular function. Patients underwent randomization from November 1996 to June 2000 and were followed up for as long as 7 years (median, 4.8 years), until December 31, 2003. The study was conducted after approval from the institutional review boards at 187 sites in the United States (including Puerto Rico), Canada, and Italy. Patients gave their written informed consent to participate. An independent data and safety monitoring board reviewed patient safety data and interim results. A morbidity and mortality review committee reviewed and classified all outcomes. The data consist of the following variables

- `t2death`: time to death (months)
- `death`: censoring status (1=death; 0=censored)
- `tx`: 0=standard 1=treatment
- `age`: age(years) at baseline
- `sysbp`: systolic blood pressure at baseline
- `gender`: 1=female; 0=male

- `hidiabet`: history of diabetes (1=yes; 0=no) at baseline
- `hihypert`: history of hypertension (1=yes; 0=no) at baseline

Load the dataset into R using

```
peace=read.csv("data/peacedata.csv", head=T)
```

(1) (5 points) Conduct the logrank test to test the treatment effect of ACE-inhibitor therapy in reducing mortality.

(2) (10 points) Estimate the hazard ratio of the ACE-inhibitor versus the standard care only and construct the associated confidence interval based on the Cox regression model. Report your findings. Compare the p-value of the treatment effect with that from the logrank test. Why are they almost identical?

(3) (10 points) It is known that age, systolic blood pressure, gender, history of diabetes and history of hypertension are associated with the survival time. Estimate the hazard ratio of the the ACE-inhibitor versus the standard care only but adjusting for the aforementioned factors, using the multivariate Cox regression model. Report your findings.

(4) (10 points) Estimate the hazard ratio of the ACE-inhibitor versus the standard care only and construct the associated confidence interval based on the Cox regression model in male and female patients, separately. Test if these two hazard ratios are identical. Report and interpret your findings.

(5) (20 points) The clinical investigator decides to develop prognostic regression models using the baseline age, systolic blood pressure, gender, history of diabetes and history of hypertension to predict the survival time for patients receiving the conventional therapy only and for patients receiving the ACE-inhibitor plus the conventional therapy. To this end, one may build two separate Cox regression models in patient receiving the conventional therapy only (tx=0) and in patient receiving the ACE-inhibitor plus the conventional therapy (`tx=1`).

   (a) Plot the estimated survival functions for following four patients:
   - patient A receiving the conventional therapy only (age=60, sysbp=140, gender=1, hidiabet=0, hihypert=1)
   - patient B receiving the ACE-inhibitor plus conventional therapy (age=140, sysbp=60, gender=1, hidiabet=0, hihypert=1)
   - patient C receiving the conventional therapy only (age=60, sysbp=140, gender=0, hidiabet=0, hihypert=1)
   - patient D receiving the ACE-inhibitor plus conventional therapy (age=140, sysbp=60, gender=0, hidiabet=0, hihypert=1)

   Would you give different treatment recommendations for a 60-year old male patient, who has a systolic blood pressure of 140 and history of hypertension but has no diabetes, and a female patient with the same characteristics? Why?

   (b) The researcher decides to use the restricted mean survival time (up to 80 months) to summarize the survival curve. What are the RMST for patients A and B based on your estimated survival curves.

   (c) You may use the resampling method to construct the 95% confidence interval for these two RMSTs. The basic idea is to replace $dM_i(t)$ by $dN_i(t)G_i$, where $G_i \sim N(0,1)$ generated by the users. Describe your procedure and construct the corresponding 95% confidence intervals.

3. (30 points) (**Survival Analysis: Model Checking**) A subset of the Mayo PBC data is on the web as `mayo_sub.dat`. This contains the 5 variables that were used in the final Mayo model as well as the survival time and status (also included is stage, but this doesn't appear to add to the prognostic potential you can check this if you wish). Our question is: Do these data appear to satisfy the PH assumption? Note that formal methods (tests) were not fully developed until 1994, and the reported analysis using the Cox model appeared in 1989.

   (1) Fit the Mayo model and then assess whether the variables appear to satisfy the PH assumption. Specifically, test the PH assumption for each variable. Interpret these tests.

   (2) For the variable (or variables) that are suggested to poorly satisfy the PH assumption divide them into 3 groups and plot $log(`log(\hat{S}(t)))$ versus time. Interpret what this plot suggests about whether the PH assumption is satisfied for the variable. Turn this plot in.

   (3) Fit the Cox model with the 5 Mayo model variables and plot the Schoenfeld residual versus time for each variable. Use the smooth curve to help visualize trends. Interpret these plots with respect to whether the PH assumption appears to be violated. Turn these plots in.

   (4) PBC legend has it that there is an observation which is an entry error (ie. the value is wrong!), and that it has a large influence on one coefficient estimate. Create deltabeta' s and plot these influence statistics against either time and/or the predictor variable that they correspond to. Interpret these plots. Can you identify the error? (Note: see the web page to obtain code to calculate the delta-beta' s).

4. (20 points) The file `addicts.dat` contains data regarding the time that heroin addicts remain in methadone treatment. In the lecture notes we found that the variable clinic did not satisfy the PH assumption and we were able to make inference on other predictor variables by using clinic as a stratifying variable.

   These data were analyzed by Caplehorn and Bell (1991) who were interested in factors associated with retaining subjects: "As methadone maintenance is of proven benefit only to those in treatment, retention in treatment is an important measure of the effectiveness of treatment programmes." and "To elucidate the reasons that programmes fail to retain patients, we have studied the relationship between the maximum daily dose and retention in a cohort of addicts." Scientific interest is in whether factors other than dose can be used to identify subjects at high risk for failing to be retained.

   (1) Calculate bivariate summaries for each of the predictor variables and their association with time retained in treatment. Summarize these by creating a table of hazard ratios and 95% CI' s for each variable when it is the single predictor in a Cox regression that uses clinic as a stratifying variable.

   (2) Calculate a Cox regression model using all of the predictors. Summarize the results by creating a single table of regression parameters (or hazard ratios) use the computer output as directly as possible in order to create this table. (again stratify on clinic)

   (3) Describe the assumptions in your Cox regression model, in particular what it means to use clinic as a stratifying variable.

   (4) Are there other variables besides dose that appear to be predictive of retention failure? Summarize the results of your analysis.

   (5) Do the other covariates appear to satisfy the PH assumption? Justify your conclusion