3. Regression models for survival data

Introduction to proportional hazard regression modles
Partial likelihoods and estimation procedure
Testing procedures
Estimation of the survival function of the regression model

§3.1. Introduction to hazard regression models

Let $\boldsymbol{x} = (x_1, \ldots, x_p)$ be a vector of covariates which affect the survival function.

$S(t|\boldsymbol{x})$: survival function given $\boldsymbol{x}$.

$h(t|\boldsymbol{x})$: hazard function given $\boldsymbol{x}$.

A hazard regression model postulates a relationship between $h(t|\boldsymbol{x})$ and $\boldsymbol{x}$. If the form of this relationship is completely specified, the model is called a parametric model. It it is only partially specified, the model is called a semi-parametric model.

- **Proportional hazard model**

  General model:

  $$h(t|\boldsymbol{x}) = h_0(t)c(\boldsymbol{x}^t\boldsymbol{\beta}),$$

  where $h_0(t)$ is the baseline hazard function when $\boldsymbol{x}$ takes value $\boldsymbol{x}_0$ such that $c(\boldsymbol{x}_0^t\boldsymbol{\beta}) = 1$, $c(\cdot)$ is a positive function.

Cox proportional hazard model:

$$h(t|\boldsymbol{x}) = h_0(t)\exp\{\boldsymbol{x}^t\boldsymbol{\beta}\}.$$

The survival function given $\boldsymbol{x}$ is of the form:

$$S(t|\boldsymbol{x}) = S_0(t)^{\exp\{\boldsymbol{x}^t\boldsymbol{\beta}\}}$$

by the relationship among the survival function and cumulative hazard function, where $S_0(t)$ is the survival function corresponding to the baseline hazard function $h_0(t)$.

- **Coding of covariates**

  The covariates enter the regression model through the linear function

$$\boldsymbol{x}^t\boldsymbol{\beta} = \sum_{j=1}^{p} \beta_j x_j.$$

  For a continuous covariate, its main effect is represented by one x_j, which is a function of the covariate, in the linear form.

  For a categorical covariate with K categories, $K-1$ dummy variables are used to represent the main effect of the covariate in the linear form, e.g.,

$$x_k = \begin{cases} 1, & \text{if in category k} \\ 0, & \text{otherwise,} \end{cases}$$

$$k = 1, \ldots, K-1.$$

Interactions between covariates are modeled by including in the linear function the product terms of the covariates.

E.g., if x_1, x_2 represent two continous covariates, an additional term $x_3 = x_1 x_2$ represents the interaction between x_1 and x_2.

The interaction between a continuous covariate and a categorical covariate and interaction between two categorical covariates are similarly represented.

In general, there are as many terms of products as the number of dummy variables for the interaction between a continuous covariate and a categorical covariate, and there are as many terms as the number of cross products of the dummy variables for interaction between two categorical covariates.

§3.2. Partial likelihood and estimation procedure for Cox proportional hazard model

Data:

$$(T_i, c_i, \boldsymbol{x}_i), i = 1, \ldots, n.$$

T_i: time on study (survival time or censoring time).
c_i: censoring indicator.
$\boldsymbol{x}_i$: vector of covariates.
$t_1 < \cdots < t_D$: ordered distinct survival times.

- **The full likelihood function**

  Suppose only right censoring presents. The full likelihood function is given by

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n} h(T_i|\boldsymbol{x}_i)^{c_i} S(T_i|\boldsymbol{x}_i) \\ &= \prod_{i=1}^{n} h_0^{c_i}(T_i) \exp\{c_i \boldsymbol{x}_i^t \boldsymbol{\beta}\} S_0(T_i)^{\exp\{\boldsymbol{x}_i^t \boldsymbol{\beta}\}} \\ &= \prod_{i=1}^{n} h_0^{c_i}(T_i) \exp\{c_i \boldsymbol{x}_i^t \boldsymbol{\beta}\} \exp\{-H_0(T_i) \exp\{\boldsymbol{x}_i^t \boldsymbol{\beta}\}\} \\ &= \left[\prod_{i=1}^{D} h_0(t_i) \exp\{\boldsymbol{x}_i^t \boldsymbol{\beta}\}\right] \exp\left[-\sum_{i=1}^{n} H_0(T_i) \exp\{\boldsymbol{x}_i^t \boldsymbol{\beta}\}\right]. \end{aligned}$$

  The full likelihood involves $h_0(t)$ which is unknown. Thus the full likelihood can not be directly used for the estimation of $\boldsymbol{\beta}$.

- **Partial likelihood — Case I:** $d_i = 1, i = 1, \ldots, D$

  $\boldsymbol{x}_{(i)} = (x_{(i)1}, \ldots, x_{(i)p})$: vector of covariates associated with the individual whose survival time is t_i.

  $\mathcal{R}(t_i)$: risk set of individuals at time t_i.

  The partial likelihood:

$$L_p(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp\{\boldsymbol{x}_{(i)}^t \boldsymbol{\beta}\}}{\sum_{j \in \mathcal{R}(t_i)} \exp\{\boldsymbol{x}_j^t \boldsymbol{\beta}\}}.$$

The log partial likelihood:

$$l_p(\boldsymbol{\beta}) = \sum_{i=1}^{D} \left[\boldsymbol{x}_{(i)}^t \boldsymbol{\beta} - \ln \left(\sum_{j \in \mathcal{R}(t_i)} \exp\{\boldsymbol{x}_j^t \boldsymbol{\beta}\} \right) \right].$$

Remark: The partial likelihood can be explained from two points of view:

(i) The ith factor in $L(\boldsymbol{\beta})$ can be interpreted as the conditional probability that an individual fails at time t_i with covariate $\boldsymbol{x}_{(i)}$, given that one of the individuals in $\mathcal{R}(t_i)$ fails at this time, i.e.,

$$P(\text{individual}(\boldsymbol{x}_{(i)})\text{fails}|\text{one failure at } t_i).$$

(ii) The partial likelihood is the profile likelihood obtained by maximizing the full likelihood with respect to the nuisance factor $h_0(t_i), i = 1, \ldots, D$.

- **Partial likelihood — Case II: $d_i > 1$ for at least one i**

$\mathcal{D}(t_i)$: set of individuals who fail at t_i.

d_i: number of failures at t_i.

$\boldsymbol{s}_i = \sum_{j \in \mathcal{D}(t_i)} \boldsymbol{x}_j$.

Breslow's partial likelihood:

$$L_1(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{s}_i)}{\left[\sum_{j \in \mathcal{R}(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_j) \right]^{d_i}}.$$

Efron's partial likelihood:

$$L_2(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{s}_i)}{\prod_{j=1}^{d_i} \left[\sum_{k \in \mathcal{R}(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k) - \frac{j-1}{d_i} \sum_{k \in \mathcal{D}(t_i)} \exp(\boldsymbol{\beta}' \boldsymbol{x}_k) \right]}.$$

Cox's partial likelihood:

$$L_3(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{\exp(\boldsymbol{\beta}' \boldsymbol{s}_i)}{\sum_{q \in Q_i} \exp(\boldsymbol{\beta}' \boldsymbol{s}_q^*)},$$

where Q_i is the set of all d_i-tuples which could be selected from $\mathcal{R}(t_i)$, and $\boldsymbol{s}_q^* = \sum_{j=1}^{d_i} \boldsymbol{x}_{q_j}$ with $q = (q_1, \ldots, q_{d_i}) \in Q_i$.

- **MLE by Newton-Raphson procedure**

  Let $l_p(\boldsymbol{\beta})$ denote the log partial likelihood in all cases. Let

$$\begin{aligned} \boldsymbol{U}(\boldsymbol{\beta}) &= \frac{\partial l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \begin{pmatrix} \frac{\partial l_p(\boldsymbol{\beta})}{\partial \beta_1} \\ \vdots \\ \frac{\partial l_p(\boldsymbol{\beta})}{\partial \beta_p} \end{pmatrix}, \\ I(\boldsymbol{\beta}) &= -\frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} \\ &= -\begin{pmatrix} \frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \beta_1^2} & \cdots & \frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & \cdots & \frac{\partial^2 l_p(\boldsymbol{\beta})}{\partial \beta_p^2} \end{pmatrix} \end{aligned}$$

  $\boldsymbol{U}(\boldsymbol{\beta})$: score vector.

  $I(\boldsymbol{\beta})$: the estimated information matrix.

The Newton-Raphson procedure:

(a) Set initial value $\boldsymbol{\beta}^{(0)}$ and solve for $\boldsymbol{\beta}^{(1)}$:

$$I(\boldsymbol{\beta}^{(0)})(\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}^{(0)}) = \boldsymbol{U}(\boldsymbol{\beta}^{(0)}).$$

(b) Assign $\boldsymbol{\beta}^{(0)} = \boldsymbol{\beta}^{(1)}$, repeat (a).

(c) Repeat (a) and (b) until convergence.

- **The asymptotic distribution of the MLE**

Let $\hat{\boldsymbol{\beta}}$ denote the MLE of $\boldsymbol{\beta}$. Then, asymptotically,

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, [I(\hat{\boldsymbol{\beta}})]^{-1}).$$

Remark: The asymptotic distribution can be used to construct confidence intervals for the components or linear combinations of of $\boldsymbol{\beta}$.

§3.3. Testing procedures

- **Three general tests**

Likelihood Ratio Test

Wald's Test

Rao's Score Test.

- **Testing global hypothesis**

  Global hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$.

  Wald test:

  $$X_W^2 = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)' I(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0).$$

  Likelihood ratio test:

  $$X_{\mathrm{LR}}^2 = 2[l_p(\hat{\boldsymbol{\beta}}) - l_p(\boldsymbol{\beta}_0)].$$

  Score test:

  $$X_{\mathrm{SC}}^2 = \boldsymbol{U}(\boldsymbol{\beta}_0)' I^{-1}(\boldsymbol{\beta}_0)\boldsymbol{U}(\boldsymbol{\beta}_0).$$

  The three tests are asymptotically equivalent in the sense that, under H_0, the distribution of all the three test statistics converge to the χ^2-distribution with d.f. p.

  The tests reject H_0 at level α, if the test statistics exceeds $\chi_\alpha^2(p)$.

- **Testing local Hypothesis**

  Local hypothesis concerns with part of the components of the parameter vector $\boldsymbol{\beta}$. Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^t, \boldsymbol{\beta}_2^t)^t$, where $\boldsymbol{\beta}_1$ is of dimension q and $\boldsymbol{\beta}_2$ is of dimension $p - q$. A local hypothesis takes the form:

  $$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_{1_0}.$$

Wald test

Let $\hat{\boldsymbol{\beta}}$ and the estimated variance matrix of $\Sigma(\hat{\boldsymbol{\beta}}) = I^{-1}(\hat{\boldsymbol{\beta}})$ be partitioned as

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \end{pmatrix}, \ \Sigma(\hat{\boldsymbol{\beta}}) = \begin{pmatrix} \Sigma_{11}(\hat{\boldsymbol{\beta}}) & \Sigma_{12}(\hat{\boldsymbol{\beta}}) \\ \Sigma_{21}(\hat{\boldsymbol{\beta}}) & \Sigma_{22}(\hat{\boldsymbol{\beta}}) \end{pmatrix}.$$

The test statistic is given by

$$X^2_W = (\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1_0})^t \Sigma^{-1}_{11}(\hat{\boldsymbol{\beta}})(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_{1_0}).$$

Under H_0, X^2_W follows an asymptotic χ^2-distribution with d.f. q. H_0 is rejected at level α, if $X^2_W > \ \chi_\alpha(q)$.

Likelihood Ratio test

$$X^2_{\mathrm{LR}} = 2[l_p(\hat{\boldsymbol{\beta}}) - l_p(\boldsymbol{\beta}_{1_0}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{1_0 0})],$$

$l_p(\hat{\boldsymbol{\beta}})$: maximum log likelihood without restriction.

$l_p(\boldsymbol{\beta}_{1_0}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{1_0}))$: maximum log likelihood under H_0.

Under H_0, X^2_{LR} follows an asymptotic χ^2-distribution with d.f. q. H_0 is rejected at level α, if $X^2_{LR} > \ \chi_\alpha(q)$.

§3.4. Hazard regression analysis with Splus

- **The Splus function** `coxph`

```
coxph(formula, data, init,  method)
```

formula: a formula object of the form: `Surv( , ) ~ V1+V2`, which specifies the regression function $\boldsymbol{x}^t\boldsymbol{\beta}$.

data: a data frame containing the observations on the variables named in the formula, subset, and weights arguments.

init: a vector of initial values of $\boldsymbol{\beta}$ for the iteration. The default is zero for all components of $\boldsymbol{\beta}$.

method: specifies the method for tie handling. The choices are `"efron", "breslow", "exact"`.

For more arguments of `coxph`, see Splus help file for `coxph`

Output items from `coxph`:

coefficients: the MLE $\hat{\boldsymbol{\beta}}$.

var: the variance matrix of $\hat{\boldsymbol{\beta}}$.

loglik: the vector $(l_p(\hat{\boldsymbol{\beta}}), l_p(\boldsymbol{\beta}_0))$ where $\boldsymbol{\beta}_0$ is the initial value.

score: value of the score test statistic, at the initial value of the coefficients.

wald.test: value of the Wald statistic for testing of whether the coefficients differ from the initial values.

For more output items, see Splus help file for `coxph.object`

- **Example: Clinical trial on laryngeal cancer**

  90 males were diagnosed with cancer of the larynx durig the period 1970-1978 at a Dutch hospital. Time from first treatment to either death or the end of study for each patient was recorded. The age and desease stage (I, II, III or IV) of each patient at the first treatment were also recorded. In the following, a regression analysis is carried out to see how the age and desease stage affect the survival time of the patients. The data set `larynx.txt` can be downloaded from `www.stat.nus.edu.sg/~stachenz`

  The variables:

  $V1$: Stage of disease (1, 2, 3, 4)
  $V2$: Time to death or on-study time, months
  $V3$: Age at diagnosis of larynx cancer
  $V4$:Year of diagnosis of larynx cancer
  $V5$: Death indicator (0=alive, 1=dead)

  Splus procedures:

  Preliminary procedures:

```
larynx_read.table("larynx.txt")
attach(larynx)
```

```
# Create dummy variables for disease stage
  x1_V1
  x1[x1!=2]_0
  x1[x1==2]_1
  x2_V1
  x2[x2!=3]_0
  x2[x2==3]_1
  x3_V1
  x3[x3!=4]_0
  x3[x3==4]_1
```

Estimation and global tests:

```
  larynx.fit_coxph(Surv(V2,V5)~x1+x2+x3+V3,data=larynx,
                 method="breslow")
  list(larynx.fit, Log.likelihood=larynx.fit$loglik,
     SC.test=larynx.fit$score,
     Wald.test=larynx.fit$wald.test)

     coef exp(coef) se(coef)     z       p
  x1 0.1384      1.15   0.4623 0.299 0.76000
  x2 0.6381      1.89   0.3561 1.792 0.07300
  x3 1.6933      5.44   0.4222 4.011 0.00006
  V3 0.0189      1.02   0.0143 1.326 0.18000
  Likelihood ratio test=18.1  on 4 df, p=0.0012  n= 90

  $Log.likelihood:
```

```
[1] -197.2129 -188.1794
$SC.test:
[1] 24.32745
$Wald.test:
[1] 20.82556
```

Local tests:

$\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)$

$(\beta_1, \beta_2, \beta_3)$: effect of stages.

β_4: effect of age.

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0.$

Wald test:

Extract the $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ and its variance matrix from the object `larynx`, and compute Wald statistic.

```
b_larynx.fit$coef[-4]
A_larynx.fit$var[-4,-4]
X_t(b)%*%solve(A)%*%b

c(X, 1-pchisq(X,3))
(17.637, 0.0005)
```

Likelihood Ratio test:

a) From the object obtained from fitting the full model, get $l_p(\hat{\boldsymbol{\beta}})$:

```
larynx.fit$loglik
-197.2129 -188.1794
```

$l_p(\boldsymbol{b}) = -188.1794$.

b) Fit null model to get $l_p(\boldsymbol{\beta}_{1_0}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{1_0})$:

```
larynx.0_coxph(Surv(V2,V5)~V3,
          data=larynx, method="breslow")
larynx.0$loglik
-197.2129 -195.9059
```

$l_p(\boldsymbol{\beta}_{1_0}, \hat{\boldsymbol{\beta}}_2(\boldsymbol{\beta}_{1_0}) = -195.9059$.
$X^2_{\text{LR}} = 2[-188.1794 - (-195.9059)] = 15.453$.
p-value= 0.00147.

§3.5. Estimation of survival function

The survival function of an individual with covariate vector $\boldsymbol{x}$ is given by

$$S(t|\boldsymbol{x}) = S_0(t)^{\exp\{\boldsymbol{x}^t\boldsymbol{\beta}\}}.$$

The estimate of $S(t|\boldsymbol{x})$ is given by

$$\hat{S}(t|\boldsymbol{x}) = \hat{S}_0(t)^{\exp\{\boldsymbol{x}^t\hat{\boldsymbol{\beta}}\}},$$

where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ and $\hat{S}_0(t)$ is an estimate of $S_0(t)$.

The estimation of $S_0(t)$

Let

$$\hat{\alpha}_i = \exp\left\{-\frac{d_i}{\sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{x}_j^t\hat{\boldsymbol{\beta}}\}}\right\}.$$

The MLE of $S_0(t)$ is given by

$$\hat{S}_0(t) = \prod_{t_i\le t} \hat{\alpha}_i.$$

Remark:

(i) The negative of the exponent of $\hat{\alpha}_i$ is the estimate of the baseline hazard at t_i, i.e.,

$$\hat{h}_0(t_i) = \frac{d_i}{\sum_{j\in\mathcal{R}(t_i)} \exp\{\boldsymbol{x}_j^t\hat{\boldsymbol{\beta}}\}}.$$

(ii) The estimate is indeed the MLE of $S_0(t)$ obtained by maximizing the full likelihood in the case $d_i = 1$.

(iii) The computaion of $\hat{S}_0$ by Splus: The $\hat{h}_0(t_i)$ are provided by `coxph.detail(fit.object)`, which is list containing the fitting details of the object `fit.object`. The $\hat{h}_0(t_i)$ are contained in the component `coxph.detail(fit.object)$hazard`.

Example: Clinical trial on laryngeal cancer (cont.)

```
larynx.fit_coxph(Surv(V2,V5)~x1+x2+x3+V3,data=larynx,
                 method="breslow")
h0_coxph.detail(larynx.fit)$hazard
H0_0
S0_NULL
for (i in 1:length(h0)) {
   H0_H0+h0[i]
   S0[i]_exp(-H0)
}
S0.fit_cbind(coxph.detail(larynx.fit)$time, S0)
```