Chapter 8

Survival Data Models

8.1 Proportional Hazards Regression Models

8.1.1 Hazards and Survival Functions

– Let there be binary random variable Y_i which has outcomes as following:

$$Y_i = \begin{cases} 1, \text{when considered event happens to observation } i \text{ latest at time } t_s, \\ 0, \text{when considered event does not happen to observation } i \text{ latest at time } t_s. \end{cases}$$

– Let T_i be a random variable which measures the time when the random variable Y_i has the realization $y_i = 1$.

– If the random variable Y_i does not have the realization $y_i = 1$ before the *censored* time t_s, then the random variable T_i is considered to have realization $t_i = t_s$, and the random variable Y_i is marked to have the realization $y_i = 0$.

– In survival analysis, interest is to model the probability of surviving past the value t_i, i.e., interest is to model how probability $P(T_i \geq t_i)$ depends on the explanatory variables $X_1, X_2, \ldots, X_p$, when it is possible that $t_i = t_s$.

– The survival function, with given values of the explanatory variables $X_1, X_2, \ldots, X_p$, is defined as

$$S(t_i|\mathbf{x}_i) = P(T_i \geq t_i|\mathbf{x}_i) = 1 - F(t_i|\mathbf{x}_i), \; t_i > 0, \tag{8.1}$$

where $F(t_i|\mathbf{x}_i)$ cumulative distribution function.

– Thus the density function of the random variable T_i is

$$f(t_i|\mathbf{x}_i) = -\frac{\partial S(t_i|\mathbf{x}_i)}{\partial t_i} = -\frac{\partial\left[1 - F(t_i|\mathbf{x}_i)\right]}{\partial t_i},\ t_i > 0. \tag{8.2}$$

– In survival analysis, the hazard function $h(t_i|\mathbf{x}_i)$ is defined as

$$\begin{aligned}
h(t_i|\mathbf{x}_i) &= \lim_{\Delta t \to 0} \frac{P(t_i \le T_i < t_i + \Delta t | T_i \ge t_i)}{\Delta t} = \lim_{\Delta t \to 0} \frac{P(t_i \le T_i < t_i + \Delta t)}{P(T_i \ge t_i)\Delta t} \\
&= \lim_{\Delta t \to 0} \frac{P(t_i \le T_i < t_i + \Delta t)}{S(t_i|\mathbf{x}_i)\Delta t} = \frac{f(t_i|\mathbf{x}_i)}{S(t_i|\mathbf{x}_i)} = \frac{f(t_i)}{1 - F(t_i|\mathbf{x}_i)} \\
&= -\frac{\partial \log\left[1 - F(t_i|\mathbf{x}_i)\right]}{\partial t_i} = -\frac{\partial \log\left[S(t_i|\mathbf{x}_i)\right]}{\partial t_i}
\end{aligned} \tag{8.3}$$

– Intuitively, the hazard function $h(t_i|\mathbf{x}_i)$ measures the risk of event happening in a short interval $(t_i, t_i + \Delta t)$ immediately after t_i.

– The cumulative hazard function is

$$H(t_i|\mathbf{x}_i) = \int_0^{t_i} h(s_i|\mathbf{x}_i)\, ds_i = -\log\left[1 - F(t_i|\mathbf{x}_i)\right] = -\log\left[S(t_i|\mathbf{x}_i)\right]. \tag{8.4}$$

– Then also

$$S(t_i|\mathbf{x}_i) = e^{-H(t_i|\mathbf{x}_i)}, \tag{8.5}$$

$$f(t_i|\mathbf{x}_i) = h(t_i|\mathbf{x}_i)e^{-H(t_i|\mathbf{x}_i)}. \tag{8.6}$$

8.1.2 Cox Regression Model

– In survival analysis, the relationship of the hazard function to the explanatory variables is typically examined.

– The Cox regression model is log-linear model for the hazard function

$$\begin{aligned}\log\left[h(t_i|\mathbf{x}_i)\right] &= \alpha_0(t) + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \\ &= \log\left[h_0(t_i)\right] + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip},\end{aligned} \tag{8.7}$$

where $h_0(t_i)$ is called baseline hazard.

– The Cox model is often written as

$$h(t_i|\mathbf{x}_i) = h_0(t_i)e^{\mathbf{x}_i'\boldsymbol{\beta}}, \tag{8.8}$$

where the baseline hazard $h_0(t_i)$ is left completely unspecified.

– The Cox model is semi-parametric model because while the baseline hazard can take any form, the explanatory enter the model linearly.

– The Cox model is a proportional-hazards model, since for any different values $\mathbf{x}_i$ and $\mathbf{x}_{i_*}$, the following ratio has a form

$$\frac{h(t_i|\mathbf{x}_i)}{h(t_{i_*}|\mathbf{x}_{i_*})} = \frac{h_0(t_i)e^{\mathbf{x}_i'\boldsymbol{\beta}}}{h_0(t_{i_*})e^{\mathbf{x}_{i_*}'\boldsymbol{\beta}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{e^{\mathbf{x}_{i_*}'\boldsymbol{\beta}}} = \exp\left((\mathbf{x}_i' - \mathbf{x}_{i_*}')\boldsymbol{\beta}\right). \tag{8.9}$$

Example 8.1.

Consider the data set lung:

```
> library(survival)
> data(lung)
> lung
    inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1      3  306      2  74   1       1       90       100     1175      NA
2      3  455      2  68   1       0       90        90     1225      15
3      3 1010      1  56   1       0       90        90       NA      15
4      5  210      2  57   1       1       90        60     1150      11
.
.
227    6  174      1  66   1       1       90       100     1075       1
228   22  177      1  58   2       1       80        90     1060       0
```

(a) Let $T = \mathsf{time}$. Model the hazard function by the model

$$h_i(t|x_i) = h_0(t)e^{\beta x_i},$$

where $X = \mathsf{age}$. Estimate the value of the survival function $S(t|x_i) = P(T \geq t|x_i)$ at the time point $t = 800$ when $x_i = 70$. Create 95% confidence intervals too.

(b) Consider the model

$$h_i(t|x_i) = h_0(t)e^{\beta x_i},$$

where $X = \mathsf{age}$. Estimate the hazard ratio

$$\frac{h_i(t|x_i = 80)}{h_i(t|x_{i_*} = 40)}.$$

(c) Consider the following Cox proportional hazards regression model

$$h_i(t|\mathbf{x}_i) = h_0(t) \cdot \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \alpha_j),$$

where $X_1 = \mathsf{age}$, $X_2 = \mathsf{wt.loss}$ and $X_3 = \mathsf{sex}$. Test at 5% significance level, is the explanatory variable $X_2 = \mathsf{wt.loss}$ statistically significant variable.

8.1.3 Weibull Model

– If the random variable T_i follows the Weibull distribution $T_i \sim Wei(p, \lambda)$, then

$$
\begin{aligned}
f(t_i) &= \frac{p}{\lambda}\left(\frac{t_i}{\lambda}\right)^{p-1} \cdot \exp\left[-\left(\frac{t_i}{\lambda}\right)^{p}\right], \qquad t_i \geq 0, && \text{(8.10a)}\\
F(t_i) &= 1 - \exp\left[-\left(\frac{t_i}{\lambda}\right)^{p}\right], && \text{(8.10b)}\\
S(t_i) &= \exp\left[-\left(\frac{t_i}{\lambda}\right)^{p}\right], && \text{(8.10c)}\\
h(t_i) &= \frac{p}{\lambda}\left(\frac{t_i}{\lambda}\right)^{p-1}, && \text{(8.10d)}\\
H(t_i) &= \left(\frac{t_i}{\lambda}\right)^{p}. && \text{(8.10e)}
\end{aligned}
$$

– Also if $T_i \sim Wei(p, \lambda)$, then the expected value is $\mathrm{E}(T_i) = \lambda \cdot \Gamma\left(1 + \frac{1}{p}\right)$, where the gamma function $\Gamma(z) = \int_0^\infty x^{z-1} e^{-x}\, dx$.

– The Weibull proportional hazards model is written as

$$h(t_i|\mathbf{x}_i) = h_0(t_i)e^{\mathbf{x}_i'\boldsymbol{\beta}} = \frac{p}{\lambda}\left(\frac{t_i}{\lambda}\right)^{p-1} \cdot e^{\mathbf{x}_i'\boldsymbol{\beta}}, \tag{8.11}$$

where the baseline hazard $h_0(t_i) = \frac{p}{\lambda}\left(\frac{t_i}{\lambda}\right)^{p-1}$ is the Weibull hazard function.

– Note that $h(t_i|\mathbf{x}_i)$ can be written as

$$\begin{aligned} h(t_i|\mathbf{x}_i) &= \frac{p}{\lambda}\left(\frac{t_i}{\lambda}\right)^{p-1} \cdot e^{\mathbf{x}_i'\boldsymbol{\beta}} = \frac{p}{\lambda^p} t_i^{p-1} \cdot e^{\mathbf{x}_i'\boldsymbol{\beta}} \\ &= \frac{p}{\lambda^p} t_i^{p-1} \cdot (\exp(\mathbf{x}_i'\boldsymbol{\beta}/p))^p = \frac{p}{\lambda^p \cdot \left(\frac{1}{\exp(\mathbf{x}_i'\boldsymbol{\beta}/p)}\right)^p} t_i^{p-1} \\ &= \frac{p}{\left(\frac{\lambda}{\exp(\mathbf{x}_i'\boldsymbol{\beta}/p)}\right)^p} t_i^{p-1} = \frac{p}{\lambda_*^p} t_i^{p-1}, \end{aligned} \tag{8.12}$$

where

$$\lambda_* = \frac{\lambda}{\exp(\mathbf{x}_i'\boldsymbol{\beta}/p)}. \tag{8.13}$$

– Hence under The Weibull proportional hazards model $T_i|\mathbf{x}_i \sim Wei(p, \lambda_*)$

– Prediction intervals can be created by the following parametric bootstrap method.

Parametric bootstrap based method - Prediction Interval

1. Find the estimate $\hat{p}, \hat{\lambda}, \hat{\boldsymbol{\beta}}$.
2. Calculate $\hat{\lambda}_* = \frac{\hat{\lambda}}{\exp(\mathbf{x}_i'\hat{\boldsymbol{\beta}}/\hat{p})}$.
3. Simulate t_{f_*} from the distribution $t_{f_*} \sim Wei(\hat{p}, \hat{\lambda}_*)$.
4. Repeat M times the step 3, and then determine $\alpha/2$ and $1-\alpha/2$ the quantiles of the simulated values t_{f_*} .

Example 8.2.

Consider the data set lung:

```
> library(survival)
> data(lung)
> lung
    inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1      3  306      2  74   1       1       90       100     1175      NA
2      3  455      2  68   1       0       90        90     1225      15
3      3 1010      1  56   1       0       90        90       NA      15
4      5  210      2  57   1       1       90        60     1150      11
.
.
227    6  174      1  66   1       1       90       100     1075       1
228   22  177      1  58   2       1       80        90     1060       0
```

Consider the Weibull proportional hazards regression model

$$h_i(t|x_i) = \frac{p}{\lambda}\left(\frac{t}{\lambda}\right)^{p-1} e^{\beta x_i},$$

where $X = \mathsf{age}$.

(a) Estimate the value of the survival function $S(t|x_i) = P(T \geq t|x_i)$ at the time point $t = 800$ when $x_i = 70$.

(b) Find the estimate for the expected value $\mathrm{E}(T_i)$, when $x_i = 70$.

(c) Create 80% prediction interval for new observation T_f, when $x_f = 70$.