Proportional Hazards Regression: Part Two[1]

STA312 Fall 2023

[1]See last slide for copyright information.

Proportional Hazards Regression Model

Based on the hazard function

$$h(t) = h_0(t)\, e^{\beta_0 + \mathbf{x}^\top \boldsymbol{\beta}}$$

Swallow e^{β_0} into the baseline hazard function and get

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

- The regression model has no intercept.
- It's common practice to center the explanatory variables (but not the dummy variables) by subtracting off the overall sample mean of the variable.
- Then, the baseline hazard function is the hazard function of an individual in the reference category, who is "average" on all the quantitive explanatory variables.
- It's quite meaningful.

Hazard Ratio

$$\begin{aligned}\frac{h_1(t)}{h_2(t)} &= \frac{h_0(t)\, e^{\mathbf{x}_1^\top \boldsymbol{\beta}}}{h_0(t)\, e^{\mathbf{x}_2^\top \boldsymbol{\beta}}} \\ &= \frac{e^{\mathbf{x}_1^\top \boldsymbol{\beta}}}{e^{\mathbf{x}_2^\top \boldsymbol{\beta}}}\end{aligned}$$

- Proportional hazards.
- If x_k is increased by one unit, the hazard function is multiplied by e^{β_k}.
- This is true for every time t (according to the model).
- So you can just say the "hazard" or "risk" or even "chances" of the event are twice as much.
- It's a good way to talk and think about the results.

Need to estimate the hazard and survival functions

- What we have so far is good for significance testing.
- Need to estimate the hazard and survival functions.

Estimating the baseline hazard

$h_0(t)$ in $h(t) = h_0(t)\, e^{\mathbf{x}^\top \boldsymbol{\beta}}$

Remember how partial likelihood started.

$$
\begin{aligned}
h_0(t) e^{\mathbf{x}_{(i)}^\top \boldsymbol{\beta}} &\approx \frac{h_0(t) e^{\mathbf{x}_{(i)}^\top \boldsymbol{\beta}}}{\sum_{j \in R_{(i)}} h_0(t) e^{\mathbf{x}_j^\top \boldsymbol{\beta}}} \\
&= \frac{e^{\mathbf{x}_{(i)}^\top \boldsymbol{\beta}}}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \boldsymbol{\beta}}} \\
&= \frac{1}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \boldsymbol{\beta}}} \times e^{\mathbf{x}_{(i)}^\top \boldsymbol{\beta}}
\end{aligned}
$$

A leap of intuition

Humm,

$$h_0(t_{(i)}) \times e^{\mathbf{x}_{(i)}^\top \boldsymbol{\beta}} \approx \frac{1}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \boldsymbol{\beta}}} \times e^{\mathbf{x}_{(i)}^\top \boldsymbol{\beta}}$$

So how about

$$\widehat{h}_0(t_{(i)}) = \frac{1}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}}}$$

Well, there could be ties in practice, so based on the Kaplan-Meier estimated hazard $\widehat{q}_{(i)} = \frac{d_{(i)}}{n_{(i)}}$,

$$\widehat{h}_0(t_{(i)}) = \frac{d_{(i)}}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}}}$$

Almost always, $d_{(i)} = 1$ anyway.

Estimated Hazard Function(s)

Based on $h(t) = h_0(t)\, e^{\mathbf{x}^\top \boldsymbol{\beta}}$

$$\widehat{h}(t_{(i)}) = \widehat{h}_0(t_{(i)})\, e^{\mathbf{x}^\top \widehat{\boldsymbol{\beta}}}$$

- Nice for display. Can plot D points.
- Notice it depends on $\mathbf{x}$.

Estimating the Survival Function: Background

Using $H(t) = \int_0^t h(y)\,dy$ and $S(t) = e^{-H(t)}$

- $H_0(t) = \int_0^t h_0(y)\,dy$ is the baseline cumulative hazard function.
- $S_0(t) = e^{-H_0(t)} = e^{-\int_0^t h_0(y)\,dy}$ is the baseline survival function.
- With a little work we can show $S(t) = S_0(t)^{\exp\{\mathbf{x}_i^\top \boldsymbol{\beta}\}}$.
- This could be written $S(t|\mathbf{x}_i)$.

Estimating the Survival Curve (Cox and Oakes, 1982)

Using $S_0(t) = e^{-H_0(t)}$ and $S(t) = S_0(t)^{\exp\{\mathbf{x}^\top \boldsymbol{\beta}\}}$

Want an estimate of $H_0(t) = \int_0^t h_0(y)\, dy$, but

$$\widehat{h}_0(t_{(i)}) = \frac{d_{(i)}}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}}}$$

is only defined for $t_{(1)}, \ldots, t_{(D)}$, the times where uncensored observations occurred.

Approximate the integral with a finite sum:

$$\widehat{H}_0(t) = \sum_{t_{(i)} \le t} \frac{d_{(i)}}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}}}$$

Cox and Oakes argument continued

Using $S_0(t) = e^{-H_0(t)}$ and $S(t|\mathbf{x}) = S_0(t)^{\exp\{\mathbf{x}^\top \boldsymbol{\beta}\}}$

Have

$$\widehat{H}_0(t) = \sum_{t_{(i)} \leq t} \frac{d_{(i)}}{\sum_{j \in R_{(i)}} e^{\mathbf{x}_j^\top \widehat{\boldsymbol{\beta}}}}$$

Then

$$\begin{aligned} \widehat{S}_0(t) &= e^{-\widehat{H}_0(t)} \\ \widehat{S}(t|\mathbf{x}) &= \widehat{S}_0(t)^{\exp\{\mathbf{x}^\top \widehat{\boldsymbol{\beta}}\}} \end{aligned}$$

It works

- As usual, later work clarified matters and eliminated most of the guesswork.
- Cox's estimate of $S(t)$ is shown to arise from Breslow's method of approximating the partial likelihood when there are ties.
- There are several other estimates, all yielding results that are pretty close.
- To me, the biggest payoff is that $\widehat{S}(t|\mathbf{x})$ allows estimation of the median for any particular set of explanatory variable values.

Copyright Information