6 Modeling Survival Data with Cox Regression Models

6.1 The Proportional Hazards Model

A proportional hazards model proposed by D.R. Cox (1972) assumes that

$$\lambda(t|z) = \lambda_0(t)\mathrm{e}^{z_1\beta_1+\cdots+z_p\beta_p} = \lambda_0(t)\mathrm{e}^{z^T\beta}, \tag{6.1}$$

where z is a $p \times 1$ vector of covariates such as treatment indicators, prognositc factors, etc., and β is a $p \times 1$ vector of regression coefficients. Note that there is no intercept β_0 in model (6.1).

Obviously,

$$\lambda(t|z = 0) = \lambda_0(t).$$

So $\lambda_0(t)$ is often called the baseline hazard function. It can be interpreted as the hazard function for the population of subjects with $z = 0$.

The baseline hazard function $\lambda_0(t)$ in model (6.1) can take any shape as a function of t. The only requirement is that $\lambda_0(t) > 0$. This is the nonparametric part of the model and $z^T\beta$ is the parametric part of the model. So Cox's proportional hazards model is a semiparametric model.

Interpretation of a proportional hazards model

1. It is easy to show that under model (6.1)

$$S(t|z) = [S_0(t)]^{\exp(z^T\beta)},$$

where $S(t|z)$ is the survival function of the subpopulation with covariate z and $S_0(t)$ is the survival function of baseline population ($z = 0$). That is

$$S_0(t) = \mathrm{e}^{-\int_0^t \lambda_0(u)du}.$$

2. For any two sets of covariates z_0 and z_1,

$$\frac{\lambda(t|z_1)}{\lambda(t|z_0)} = \frac{\lambda_0(t)e^{z_1^T\beta}}{\lambda_0(t)e^{z_0^T\beta}} = e^{(z_1-z_0)^T\beta}, \quad \text{for all } t \geq 0,$$

which is a constant over time (so the name of proportional hazards model). Equivalently,

$$\log\left[\frac{\lambda(t|z_1)}{\lambda(t|z_0)}\right] = (z_1 - z_0)^T\beta, \quad \text{for all } t \geq 0.$$

3. With one unit increase in z_k while other covariate values being held fixed, then

$$\log\left[\frac{\lambda(t|z_k+1)}{\lambda(t|z_k)}\right] = \log(\lambda(t|z_k+1)) - \log(\lambda(t|z_k)) = \beta_k.$$

Therefore, β_k is the increase in log hazard (*i.e.*, log hazard-ratio) at **any** time with unit increase in the kth covariate z_k. Equivalently,

$$\frac{\lambda(t|z_k+1)}{\lambda(t|z_k)} = e^{\beta_k}, \quad \text{for all } t \geq 0.$$

So $\exp(\beta_k)$ is the hazard ratio associated with one unit increase in z_k. Furthermore, since $P[t \leq T < t + \Delta t | T \geq t, z] \approx \lambda(t|z)\Delta t$, we have

$$\frac{P[t \leq T < t + \Delta t | T \geq t, z_k + 1]}{P[t \leq T < t + \Delta t | T \geq t, z_k]} \approx e^{\beta_k}, \quad \text{for all } t \geq 0.$$

so $\exp(\beta_k)$ can be loosely interpreted as the ratio of two conditional probabilities of dying in the near future given a subject is alive at any time t. Since

$$\frac{\lambda(t|z_k+1) - \lambda(t|z_k)}{\lambda(t|z_k)} = e^{\beta_k} - 1.$$

So $e^{\beta_k} - 1$ can be interpreted as the percentage change (increase or decrease) in hazard with one unit increase in z_k while adjusting for other covariates.

Inferential Problems

From the interpretation of the model, it is obvious that β characterizes the "effect" of z. So β should be the focus of our inference while $\lambda_0(t)$ is a nuisance "parameter". Given a sample of censored survival data, our inferential problems include:

1. Estimate β; derive its statistical properties.

2. Testing hypothesis $H_0 : \beta = 0$ or for part of β.

3. Diagnostics.

Estimation

Since the baseline hazard $\lambda_0(t)$ is left completely unspecified (infinite dimensional), ordinary likelihood methods can't be used to estimate β. Cox conceived of the idea of a partial likelihood to remove the nuisance parameter $\lambda_0(t)$ from the proposed estimating equation.

Historical Note: Cox described the proportional hazards model in JRSSB (1972), in what is now the most quoted statistical papers in history. He also outlined in this paper the method for estimation which he referred to as using conditional likelihood. It was pointed out to him in the literature that what he proposed was not a conditional likelihood and that there may be some flaws in his logic. Cox (1975) was able to recast his method of estimation through what he called "partial likelihood" and published this in *Biometrika*. This approach seemed to be based on sound inferential principles. Rigorous proofs showing the consistency and asymptotic normality were not published until 1981 when Tsiatis (*Annals of Statistics*) demonstrated these large sample properties. In 1982, Anderson and Gill (*Annals of Statistics*) simplified and generalized these results through the use of counting processes.

6.2 Estimation Using Partial Likelihood

Data and Model

1. Data: (X_i, Δ_i, z_i), $i = 1, \cdots, n$, where for the ith individual

   $X_i = \min(T_i, C_i)$.

   $\Delta_i = I(T_i \leq C_i)$.

   $z_i = (z_{i1}, z_{i2}, \cdots, z_{ip})^T$ is a vector of covariates.

2. Model: Proportional hazards model

$$\lambda(t|z_i) = \lambda_0(t)\mathrm{e}^{z_i^T\beta},$$

where

$$\lambda(t|z_i) = \lim_{h\to 0^+}\left\{\frac{P[t \le T_i < t+h|T_i \ge t, z_i]}{h}\right\}.$$

Assume that C_i and T_i are conditionally independent given z_i. Then the cause-specific hazard can be used to represent the hazard of interest. That is (in terms of conditional probabilities)

$$P[x \le X_i < x+\Delta x, \Delta_i = 1|X_i \ge x, z_i] = P[x \le T_i < x+\Delta x|T_i \ge x, z_i] \approx \lambda_{T_i}(x|z_i)\Delta x.$$

Similar to the case of log rank test, we need to define some notation. Let us break the time axis (patient time) into a grid of points. Assume the survival time is continuous. We hence can take the grid points dense enough so that at most one death can occur within any interval.

Let $dN_i(u)$ denote the indicator for the ith individual being observed to die in $[u, u+\Delta u)$. Namely,

$$dN_i(u) = I(X_i \in [u, u+\Delta u), \Delta_i = 1).$$

Let $Y_i(u)$ denote the indicator for whether or not the ith individual is at risk at time u. Namely,

$$Y_i(u) = I(X_i \ge u).$$

Let $dN(u) = \sum_{i=1}^n dN_i(u)$ denote the number of deaths for the whole sample occurring in $[u, u+\Delta u)$. Since we are assuming Δu is sufficiently small, so $dN(u)$ is either 1 or 0 at any time u.

Let $Y(u) = \sum_{i=1}^n Y_i(u)$ be the total number from the entire sample who are at risk at time u.

Let $\mathcal{F}(x)$ denote the information up to time x (one of the grid points)

$$\mathcal{F}(x) = \{(dN_i(u), Y_i(u), z_i), i = 1, \cdots, n; \text{for grid points } u < x \text{ and } dN(x)\}.$$

Note: Conditional on $\mathcal{F}(x)$, we know who has died or was censored prior to x, when they died or were censored, together with their covariate values. We know the individuals at risk at time x and their corresponding covariate value. In addition, we also know if a death occurs at interval $[x, x + \Delta x)$.

What we don't know is the individual who was observed to die among those at risk at time x if $dN(x) = 1$.

Let $I(x)$ denote the individual in the sample who died at time x if someone died. If no one dies at time x, then $I(x) = 0$

For example, if $I(x) = j$, then this means that the jth individual in the sample with covariate vector z_j died in $[x, x + \Delta x)$.

Let $\mathcal{F}(\infty)$ denote all the data in the sample. Namely

$$\mathcal{F}(\infty) = \{(X_i, \Delta_i, z_i), i = 1, \cdots, n\}.$$

If we let $u_1 < u_2 < \cdots$ denote the value of the grid points along the time axis, then the data (with redundancy) can be expressed as

$$(\mathcal{F}(u_1), I(u_1), \mathcal{F}(u_2), I(u_2), \cdots, \mathcal{F}(\infty)).$$

Denote the observed values of the above random variables by lower cases. Then the likelihood of the parameter $\lambda_0(t)$ and β can be written as

$$\begin{aligned}
&P[\mathcal{F}(u_1) = f(u_1); \lambda_0(\cdot), \beta] \times P[I(u_1) = i(u_1) | \mathcal{F}(u_1) = f(u_1); \lambda_0(\cdot), \beta] \\
&\times P[\mathcal{F}(u_2) = f(u_2) | \mathcal{F}(u_1) = f(u_1), I(u_1) = i(u_1); \lambda_0(\cdot), \beta] \\
&\times P[I(u_2) = i(u_2) | \mathcal{F}(u_1) = f(u_1), I(u_1) = i(u_1), \mathcal{F}(u_2) = f(u_2); \lambda_0(\cdot), \beta] \\
&\times \cdots
\end{aligned}$$

and the last term can be simplified as

$$P[I(u_2) = i(u_2) | \mathcal{F}(u_1) = f(u_1), I(u_1) = i(u_1), \mathcal{F}(u_2) = f(u_2); \lambda_0(\cdot), \beta]$$

$$= \quad P[I(u_2) = i(u_2)|\mathcal{F}(u_2) = f(u_2); \lambda_0(\cdot), \beta].$$

That is, the full likelihood can be written as the product of a series of conditional likelihoods.

The partial likelihood (as defined by D.R. Cox) consists of the product of every other conditional probabilities in the above presentation. That is

$$PL = \prod_{\{\text{all grid pt } u\}} P[I(u) = i(u)|\mathcal{F}(u) = f(u); \lambda_0(\cdot), \beta].$$

Suppose we have the following small data set, we will try to find find out this partial likelihood:

Patient ID	x	δ	z
1	2	1	2
2	2	0	2
3	3	1	1
4	4	1	3

It turns out that the partial likelihood is

$$PL(\beta) = \frac{e^{2\beta}}{e^{2\beta} + e^{2\beta} + e^{\beta} + e^{3\beta}} \times \frac{e^{\beta}}{e^{\beta} + e^{3\beta}} \times \frac{e^{3\beta}}{e^{3\beta}}. \tag{6.2}$$

In general, we have to consider two cases in calculating the above partial likelihood.

Case 1: Suppose conditional on $\mathcal{F}(u)$ we have $dN(u) = 0$. That is, no death is observed at time u. In such a case, $I(u) = 0$ with probability 1.

Hence for any grid point u where $dN(u) = 0$, we have

$$P[I(u) = 0|\mathcal{F}(u) = f(u)] = 1.$$

Therefore, the partial likelihood is not affected at any point u such that $dN(u) = 0$.

Case 2: $dN(u) = 1$. Conditional on $\mathcal{F}(u)$, if we know that one individual dies at time u, then it must be one of the individuals still at risk (alive and not censored) at time u; *i.e.*, among the following individuals

$$\{i : Y_i(u) = 1\}.$$

Also conditional on $\mathcal{F}(u)$, we know the covariate vector z_i associated to each individual i such that $Y_i(u) = 1$. Therefore, we ask the following question:

> *Among* $Y(u) = \sum_{i=1}^{n} Y_i(u)$ *individuals, what is the probability that the observed death happened to the* i*th subject (who is actually observed to die at* u*) rather than to the other patients?*

Unlike the null hypothesis case for the two-sample problem, the probabilities of choosing these subjects are not equally likely, but rather, they are **proportional to** their cause-specific hazard of dying at time u, which can be derived as follows:

Let A_i = the event that subject i is going to die in $[u, u + \Delta u)$ given that he/she is still alive at u. If a patient is not at risk at u (*i.e.*, $Y_i(u) = 0$), then $A_i = \phi$. Since we chose Δu to be so small that there is at most one death in $[u, u + \Delta u)$, so we know

$$A_1, A_2, \cdots, A_n \text{ are mutually exclusive.}$$

Because of the independence of survival times and censoring times, those $Y(u)$ patients who are at risk at u (not censored and still alive at u) make up a random sample of the subpopulation consisting of the patients who will survive up to u (and with the same covariate value). Under independent censoring assumption, we already showed in Chapter 3 that the cause-specific hazard is the same as the hazard of interest; *i.e.*,

$$\lambda(u, \delta_i = 1|z_i) = \lambda(u, |z_i).$$

Since Δu is chosen to be very small, so

$$\begin{aligned} P[A_i] &\approx Y_i(u)\lambda(u, \delta_i = 1|z_i)\Delta u \\ &= Y_i(u)\lambda(u, |z_i)\Delta u. \\ &= Y_i(u)\lambda_0(u)\exp(z_i^T\beta)\Delta u, \end{aligned}$$

where the last equation is due to the assumption of the cox model. Therefore

$$\begin{aligned} & P[I(u)=i(u)|\mathcal{F}(u)=f(u);\lambda_0(\cdot),\beta] \\ = & P[A_{i(u)}|A_1\cup\cdots\cup A_n] \\ = & \frac{P[A_{i(u)}]}{\sum_{l=1}^n P[A_l]} \\ \approx & \frac{\lambda_0(u)\exp(z_{i(u)}^T\beta)\Delta u}{\sum_{l=1}^n \lambda_0(u)exp(z_l^T\beta)Y_l(u)\Delta u} \\ = & \frac{\exp(z_{i(u)}^T\beta)}{\sum_{l=1}^n \exp(z_l^T\beta)Y_l(u)}. \end{aligned}$$

Here $Y_{i(u)}(u)=1$ since we know this patient had to be at risk at u (since we know that this patient died in $[u,u+\Delta u)$).

Combining these cases, the partial likelihood can be written as

$$PL(\beta)=\prod_{\{\text{all grid pt } u\}}\left[\frac{\exp(z_{i(u)}^T\beta)}{\sum_{l=1}^n \exp(z_l^T\beta)Y_l(u)}\right]^{dN(u)}.$$

Remark: To be formal, we need to define z_0 even though it is never used. We can, for example, take $z_0=0$.

Other equivalent ways of writing the partial likelihood include: Let $t_1,\cdots,t_d$ define the distinct death times, then

$$\begin{aligned} PL(\beta) &= \prod_{j=1}^d\left[\frac{\exp(z_{i(t_j)}^T\beta)}{\sum_{l=1}^n \exp(z_l^T\beta)Y_l(t_j)}\right]; \\ PL(\beta) &= \prod_{i=1}^n\prod_{\{\text{all grid pt } u\}}\left[\frac{\exp(z_i^T\beta)}{\sum_{l=1}^n \exp(z_l^T\beta)Y_l(u)}\right]^{dN_i(u)}; \\ PL(\beta) &= \prod_{i=1}^n\left[\frac{\exp(z_i^T\beta)}{\sum_{l=1}^n \exp(z_l^T\beta)Y_l(x_i)}\right]^{\delta_i}. \end{aligned}$$

Remark: Stare at these different representations for a while, you will convince yourself that they are all equivalent.

The importance of using the partial likelihood is that this function depends **only** on β, the parameter of interest, and is free of the baseline hazard $\lambda_0(t)$, which is infinite dimensional nuisance function.

Cox suggested treating PL as a regular likelihood function and making inference on β accordingly. For example, we maximize the partial likelihood to get the estimate of β, often called MPLE (maximum partial likelihood estimate), and use the minus of the second derivative of the log partial likelihood as the information matrix, etc.

Properties of the score of the partial likelihood

For ease of presentation, let us focus on one covariate case. The extension is straightforward.

Obviously, the log partial likelihood function of β is

$$\ell(\beta) = \sum_{\{\text{all grid pts } u\}} dN(u) \left[z_{I(u)}\beta - \log\left(\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u) \right) \right].$$

The score function is

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{\{\text{all grid pts } u\}} dN(u) \left[z_{I(u)} - \frac{\sum_{l=1}^{n} z_l \exp(z_l\beta) Y_l(u)}{\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u)} \right],$$

and the second derivative is

$$\frac{\partial^2 \ell(\beta)}{\partial \beta^2} = -\sum_u dN(u) \left[\frac{\sum_{l=1}^{n} z_l^2 \exp(z_l\beta) Y_l(u)}{\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u)} - \left(\frac{\sum_{l=1}^{n} z_l \exp(z_l\beta) Y_l(u)}{\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u)} \right)^2 \right].$$

Define

$$\bar{z}(u,\beta) = \frac{\sum_{l=1}^{n} z_l \exp(z_l\beta) Y_l(u)}{\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u)} = \sum_{l=1}^{n} z_l w_l,$$

where

$$w_l = \frac{\exp(z_l\beta) Y_l(u)}{\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u)}$$

is the weight that is proportional to the hazard of the individual failing. So $\bar{z}(u,\beta)$ can be interpreted as the weighted average of the covariate z among those individuals still at risk at time u with weights w_l.

Define

$$V_z(u,\beta) = \left[\frac{\sum_{l=1}^{n} z_l^2 \exp(z_l\beta) Y_l(u)}{\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u)} - \left(\frac{\sum_{l=1}^{n} z_l \exp(z_l\beta) Y_l(u)}{\sum_{l=1}^{n} \exp(z_l\beta) Y_l(u)} \right)^2 \right]$$

$$= \left[\frac{\sum_{l=1}^n z_l^2 \exp(z_l\beta)Y_l(u)}{\sum_{l=1}^n \exp(z_l\beta)Y_l(u)} - \left(\bar{z}(u,\beta)\right)^2\right]$$
$$= \sum_{l=1}^n z_l^2 w_l - \left(\bar{z}(u,\beta)\right)^2.$$

This can be shown to be equal to

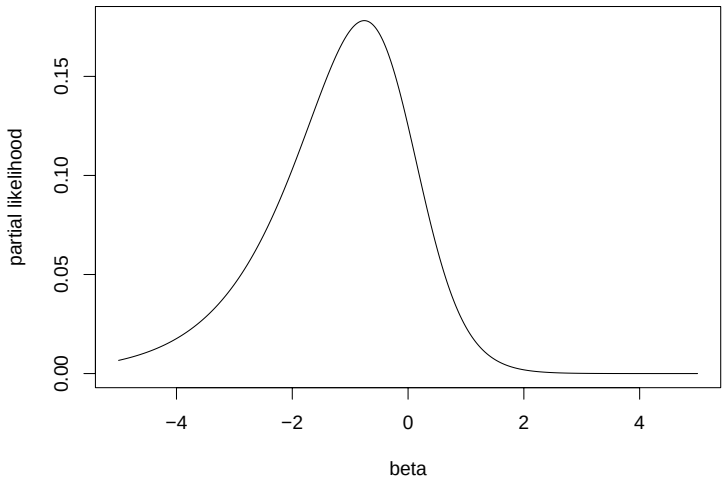
$$V_z(u,\beta) = \sum_{l=1}^n \left[\frac{(z_l - \bar{z}(u,\beta))^2 \exp(z_l\beta)Y_l(u)}{\sum_{l=1}^n \exp(z_l\beta)Y_l(u)}\right] = \sum_{l=1}^n \left(z_l - \bar{z}(u,\beta)\right)^2 w_l.$$

This last representation says that $V_z(u,\beta)$ can be interpreted as the weighted variance of the covariates among those individuals still at risk at u and hence $V_z(u,\beta) > 0$. Consequently,

$$\frac{\partial^2 \ell(\beta)}{\partial \beta^2} = -\sum_u dN(u) V_z(u,\beta) < 0.$$

The above property can also be displayed graphically. For example, the partial likelihood function (6.2) looks like:

Figure 6.1: *The partial likelihood (6.2)*



Therefore $\ell(\beta)$ has a unique maximizer and can be obtained uniquely by solving the following partial likelihood equation:

$$U(\beta) = \frac{\partial \ell(\beta)}{\partial \beta} = \sum_{\{\text{all grid pts } u\}} dN(u)\left[z_{I(u)} - \frac{\sum_{l=1}^n z_l \exp(z_l\beta)Y_l(u)}{\sum_{l=1}^n \exp(z_l\beta)Y_l(u)}\right] = 0.$$

This maximizer $\hat{\beta}$ defines the MPLE of β.

Terminology: The quantity

$$-\frac{\partial^2 \ell(\beta)}{\partial \beta^2} = \sum_u dN(u) V_z(u, \beta)$$

is defined as the partial likelihood observed information and is denoted by $J(\beta)$.

Ultimately, we want to show that the MPLE $\hat{\beta}$ has nice statistical properties. These include:

- Consistency: That is, $\hat{\beta}$ will converge to the true value of β which generated the data as the sample size gets larger. We call this true value β_0.
- Asymptotic Normality: $\hat{\beta}$ will be approximately normally distributed with mean β_0 and a variance which can be estimated from the data. This approximation will be better as the sample size gets larger. This result is useful in making inference for the true β.
- Efficiency: Among all other competing estimators for β, the MPLE has the smallest variance, at least, when the sample size gets larger.

In order to show the properties for $\hat{\beta}$, we expand $U(\hat{\beta})$ at the true value β_0 using Taylor expansion:

$$0 = U(\hat{\beta}) \approx U(\beta_0) + \frac{\partial U(\beta_0)}{\partial \beta}(\hat{\beta} - \beta_0).$$

Since

$$\frac{\partial U(\beta_0)}{\partial \beta} = \frac{\partial^2 \ell(\beta_0)}{\partial \beta^2} = -J(\beta_0),$$

therefore

$$(\hat{\beta} - \beta_0) \approx [J(\beta_0)]^{-1} U(\beta_0)$$

This expression indicates that we need to investigate the properties of the score function $U(\beta_0)$

$$U(\beta_0) = \sum_u dN(u) \left[z_{I(u)} - \bar{z}(u, \beta_0) \right].$$

Properties of the score:

(1) $\mathrm{E}[U(\beta_0)] = 0$.

Since

$$\begin{aligned}\mathrm{E}[U(\beta_0)] &= \mathrm{E}\left[\sum_u dN(u)\left(z_{I(u)} - \bar{z}(u,\beta_0)\right)\right] \\ &= \sum_u \mathrm{E}\left[dN(u)\left(z_{I(u)} - \bar{z}(u,\beta_0)\right)\right],\end{aligned}$$

and

$$\begin{aligned}&\mathrm{E}\left[dN(u)\left(z_{I(u)} - \bar{z}(u,\beta_0)\right)\right] \\ =\ &\mathrm{E}\left[\mathrm{E}\left[dN(u)\left(z_{I(u)} - \bar{z}(u,\beta_0)\right)\Big|\mathcal{F}(u)\right]\right]\end{aligned}$$

Conditional on $\mathcal{F}(u)$, $dN(u)$ and $\bar{z}(u,\beta_0)$ are both known. Consequently the inner expectation can be written as

$$dN(u)\left[\mathrm{E}[z_{I(u)}|\mathcal{F}(u)] - \bar{z}(u,\beta_0)\right].$$

Remember that $I(u)$ is the patient identifier for the individual that dies at time u and is set to zero if no one dies at u. If no one dies at u, then $dN(u) = 0$, and hence the above quantity is zero. If someone dies at u, then $dN(u) = 1$, and conditional on $\mathcal{F}(u)$, we know it has to be one of the $Y(u)$ people at risk at time u; *i.e.*, $I(u)$ must be one of the values $\{i : Y_i = 1\}$.

The conditional distribution of $z_{I(u)}$ given $\mathcal{F}(u)$ can be derived through the conditional distribution of $I(u)$ given $\mathcal{F}(u)$ as shown in Table 6.1.

Therefore

$$\mathrm{E}[z_{I(u)}|\mathcal{F}(u)] = \sum_{l=1}^{n} z_l w_l = \frac{\sum_{l=1}^{n} z_l \exp(z_l\beta_0)Y_l(u)}{\sum_{l=1}^{n}\exp(z_l\beta_0)Y_l(u)} = \bar{z}(u,\beta_0).$$

From this, we immediately get

$$\mathrm{E}[U(\beta_0)] = 0.$$

Table 6.1: *Conditional distribution of $z_{I(u)}$ given $\mathcal{F}(u)$*

Values of $I(u)$	Values of $z_{I(u)}$	Probability
1	z_1	$\exp(z_1\beta_0)Y_1(u)/\sum_{l=1}^n \exp(z_l\beta_0)Y_l(u) = w_1$
2	z_2	$\exp(z_2\beta_0)Y_2(u)/\sum_{l=1}^n \exp(z_l\beta_0)Y_l(u) = w_2$
$\vdots$	$\vdots$	$\vdots$
n	z_n	$\exp(z_n\beta_0)Y_n(u)/\sum_{l=1}^n \exp(z_l\beta_0)Y_l(u) = w_n$

Note: From the conditional distribution of $z_{I(u)}$ given $\mathcal{F}(u)$, it is easy to see the conditional variance of $z_{I(u)}$

$$\begin{aligned}
\text{Var}[z_{I(u)}|\mathcal{F}(u)] &= \sum_{l=1}^n \Big(z_l - \text{E}[z_{I(u)}|\mathcal{F}(u)]\Big)^2 w_l \\
&= \frac{\sum_{l=1}^n (z_l - \bar{z}(u,\beta_0))^2 \exp(z_l\beta_0)Y_l(u)}{\sum_{l=1}^n \exp(z_l\beta_0)Y_l(u)} \\
&= V_z(u,\beta_0).
\end{aligned}$$

(2) Finding an unbiased estimate for the variance of $U(\beta_0)$

Since $\text{E}[U(\beta_0)] = 0$, so

$$\begin{aligned}
\text{Var}[U(\beta_0)] &= \text{E}[U(\beta_0)]^2 \\
&= \text{E}\left[\sum_u dN(u)\left[z_{I(u)} - \bar{z}(u,\beta_0)\right]\right]^2 \\
&= \text{E}\left[\sum_u \left\{dN(u)\left[z_{I(u)} - \bar{z}(u,\beta_0)\right]\right\}^2\right] \\
&+ \text{E}\left[\sum_{u\neq u'} \left\{dN(u)\left[z_{I(u)} - \bar{z}(u,\beta_0)\right]\right\}\left\{dN(u')\left[z_{I(u')} - \bar{z}(u',\beta_0)\right]\right\}\right]
\end{aligned}$$

As usual, we will take an arbitrary cross-product and show it has zero expectation. Assume $u' > u$ and denote

$$A(u) = dN(u)\left[z_{I(u)} - \bar{z}(u, \beta_0)\right], \quad A(u') = dN(u')\left[z_{I(u)} - \bar{z}(u', \beta_0)\right].$$

Then the expectation of the cross-product is

$$\begin{aligned} & \mathrm{E}\left[A(u)A(u')\right] \\ = \; & \mathrm{E}\left[\mathrm{E}\left[A(u)A(u')\middle|\, \mathcal{F}(u')\right]\right]. \end{aligned}$$

Since $u' > u$, conditional on $\mathcal{F}(u')$, $A(u)$ is known. So

$$\mathrm{E}\left[A(u)A(u')\middle|\, \mathcal{F}(u')\right] = A(u)\mathrm{E}\left[A(u')\middle|\, \mathcal{F}(u')\right] = 0.$$

Therefore

$$\begin{aligned} \mathrm{Var}[U(\beta_0)] &= \mathrm{E}\sum_u \left[A^2(u)\right] \\ &= \sum_u \mathrm{E}\left[A^2(u)\right] \\ &= \sum_u \mathrm{E}\left[\mathrm{E}\left[A^2(u)\middle|\, \mathcal{F}(u)\right]\right] \end{aligned}$$

The inner conditional expectation is

$$\mathrm{E}\left[A^2(u)\middle|\, \mathcal{F}(u)\right] = \mathrm{E}\left[\left\{dN(u)\left[z_{I(u)} - \bar{z}(u, \beta_0)\right]\right\}^2\middle|\, \mathcal{F}(u)\right].$$

Since we pick the grid points in our partition of time fine enough so that $dN(u)$ is either 0 or 1, so $dN^2(u) = dN(u)$. Hence

$$\mathrm{E}\left[A^2(u)\middle|\, \mathcal{F}(u)\right] = \mathrm{E}\left[dN(u)\left[z_{I(u)} - \bar{z}(u, \beta_0)\right]^2\middle|\, \mathcal{F}(u)\right].$$

Conditional on $\mathcal{F}(u)$, $dN(u)$ is known, $\bar{z}(u, \beta_0)$ is also known and from Table 6.1

$$\bar{z}(u, \beta_0) = \mathrm{E}[z_{I(u)}|\mathcal{F}(u)].$$

Therefore

$$\begin{aligned} \mathrm{E}\left[A^2(u)\Big|\mathcal{F}(u)\right] &= dN(u)\mathrm{E}\left[\left[z_{I(u)}-\bar{z}(u,\beta_0)\right]^2\Big|\mathcal{F}(u)\right] \\ &= dN(u)\mathrm{Var}[z_{I(u)}|\mathcal{F}(u)] \\ &= dN(u)V_z(u,\beta_0). \end{aligned}$$

Consequently,

$$\begin{aligned} \mathrm{Var}\left[U(\beta_0)\right] &= \sum_u \mathrm{E}\left[dN(u)V_z(u,\beta_0)\right] \\ &= \mathrm{E}\left[\sum_u dN(u)V_z(u,\beta_0)\right]. \end{aligned}$$

Note that the quantity $\sum_u dN(u)V_z(u,\beta_0)$ is a statistic (can be calculated from the observed data), so $\sum_u dN(u)V_z(u,\beta_0)$ is an unbiased estimate of $\mathrm{Var}\left[U(\beta_0)\right]$. In fact, $\sum_u dN(u)V_z(u,\beta_0)$ is the partial likelihood observed information $J(\beta_0)$ we defined before.

Conclusion

The score $U(\beta_0)=\sum_u A(u)$ is a sum of conditionally uncorrelated mean zero random variables and its variance can be unbiasedly estimated by

$$J(\beta_0)=\sum_u dN(u)V_z(u,\beta_0).$$

By the martingale CLT, we have:

$$U(\beta_0)\overset{a}{\sim}\mathrm{N}(0,J(\beta_0)).$$

Previously, we have shown that

$$(\hat{\beta}-\beta_0)\approx[J(\beta_0)]^{-1}U(\beta_0).$$

Treating $J(\beta_0)$ as a constant, we get the approximate distribution of $(\hat{\beta}-\beta_0)$

$$(\hat{\beta}-\beta_0)\overset{a}{\sim}\mathrm{N}(0,J^{-1}(\beta_0)).$$

Of course, in practice, β_0 is unknown. But we can substitute $\hat{\beta}$ for β_0 and use $J^{-1}(\hat{\beta})$ as the estimated variance of $\hat{\beta}$. That is, we use the following approximate distribution for $(\hat{\beta} - \beta_0)$

$$(\hat{\beta} - \beta_0) \overset{a}{\sim} \mathrm{N}(0, J^{-1}(\hat{\beta})),$$

where

$$J(\hat{\beta}) = \sum_u dN(u) \left[V_z(u, \hat{\beta}) \right],$$

and $\hat{\beta}$ is the MPLE of β solving the following equation

$$U(\hat{\beta}) = \sum_u dN(u) \left[z_{I(u)} - \bar{z}(u, \hat{\beta}) \right] = 0.$$

Inference with a Single Covariate

Assume a proportional hazards model with a single covariate z

$$\lambda(t) = \lambda_0(t) \mathrm{e}^{z\beta}.$$

After we get our data (x_i, δ_i, z_i), we can obtain the MPLE $\hat{\beta}$ by solving the partial likelihood equation; *i.e.*, setting the partial score to zero. Then asymptotically,

$$\hat{\beta} \overset{a}{\sim} \mathrm{N}(\beta_0, J^{-1}(\hat{\beta})).$$

We can use this fact to construct confidence interval for β and test the hypothesis $H_0 : \beta = \beta_0$, etc. For example, a $(1 - \alpha)$ CI of β is

$$\hat{\beta} \pm z_{\alpha/2} [J^{-1}(\hat{\beta})]^{1/2}.$$

Myelomatosis data revisited: We analyzed myelomatosis data and did not find statistically significant difference between treatments 1 and 2. We want to quantify the difference by assuming the hazards of these two treatments are proportional to each other. Define a treatment indicator `trt1` which takes value 0 for treatment 1 and takes value 1 for treatment 2. Then we can use `Proc Phreg` for this purpose.

```
proc phreg data=myel;
  model dur*status(0)=trt1;
run;
```

Part of the output is given as follows:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

                                                    16:43 Thursday, March 2, 2000  15

                                The PHREG Procedure

     Data Set: WORK.MYEL
     Dependent Variable: DUR
     Censoring Variable: STATUS
     Censoring Value(s): 0
     Ties Handling: BRESLOW


                              Summary of the Number of
                              Event and Censored Values

                                                                 Percent
                         Total       Event    Censored          Censored

                            25          17           8             32.00


                         Testing Global Null Hypothesis: BETA=0

                       Without        With
    Criterion        Covariates    Covariates    Model Chi-Square

    -2 LOG L           94.084        92.765       1.319 with 1 DF (p=0.2508)
    Score                .             .          1.297 with 1 DF (p=0.2547)
    Wald                 .             .          1.263 with 1 DF (p=0.2610)


                        Analysis of Maximum Likelihood Estimates

                        Parameter     Standard      Wald         Pr >          Risk
Variable    DF           Estimate       Error    Chi-Square   Chi-Square      Ratio

TRT1         1           0.572807      0.50960      1.26344       0.2610      1.773
```

So $\hat{\beta} = 0.5728$ with standard error 0.5096. This means that compared to treatment 1, treatment 2 will increase the hazard of dying at *any* time by 77% ($\exp(\hat{\beta}) - 1$). A 95% CI of β is

$$\hat{\beta} \pm 1.96 * \text{se}[\hat{\beta}] = 0.5728 \pm 1.96 * 0.5096 = [-0.426, 1.572].$$

And a 95% CI for the hazard ratio $\exp(\beta)$ is

$$[\text{e}^{-0.426}, \text{e}^{1.572}] = [0.653, 4.816].$$

Note: The output also gives three tests for $H_0 : \beta = 0$: likelihood ratio, score and Wald tests.

Comparison of score test and two-sample log rank test

Assume z is the dichotomous indicator for treatment; *i.e.*,

$$z = \begin{cases} 1 & \text{for treatment 1} \\ 0 & \text{for treatment 0} \end{cases},$$

and the proportional hazards model:

$$\lambda(t) = \lambda_0(t)\mathrm{e}^{z\beta}.$$

Score test: Under $H_0 : \beta = 0$, the score $U(0)$ (evaluated under H_0) has the distribution

$$U(0) \overset{a}{\sim} \mathrm{N}(0, J(0)).$$

Or equivalently,

$$\left[\frac{U(0)}{J^{1/2}(0)}\right]^2 \overset{a}{\sim} \chi^2_1.$$

Since the score $U(0)$ has the expression

$$U(0) = \sum_u dN(u)\left[z_{I(u)} - \bar{z}(u, 0)\right].$$

Then

1. If a death occurs at time u, then $dN(u) = 1$, in which case there will a contribution to $U(0)$ by adding $[z_{I(u)} - \bar{z}(u, 0)]$. Otherwise no contribution.

2. Since $z = 1$ for treatment 1 and $z = 0$ for treatment 0, $z_{I(u)}$ will then the number of deaths at time u from treatment 1.

3. Under $H_0 : \beta = 0$, $\bar{z}(u, 0)$ is simplified to be

$$\bar{z}(u, 0) = \frac{\sum_{l=1}^n z_l Y_l(u)}{\sum_{l=1}^n Y_l(u)},$$

which is the proportion of individuals in group 1 among those at risk at time u. Since we only assume one death at time u, this proportion is the expected number of death for treatment 1 among those at risk at time u, under the null hypothesis of no treatment difference.

4. Therefore, $U(0)$ is the sum over the death times of the observed number of deaths from treatment 1 minus the expected number of deaths under the null hypothesis. This was the numerator of the two-sample log rank test:

$$\sum_u \left[dN_1(u) - \frac{Y_1(u)}{Y(u)} dN(u) \right]$$

where $dN_1(u) = \#$ of observed deaths from treatment 1, $Y_1(u) = \#$ at risk at time u from treatment 1, $Y(u) =$ total $\#$ at risk at time u from 2 treatments, $dN(u) =$ total $\#$ of deaths from 2 treatments.

5. The denominator of the score test was computed as

$$J^{1/2}(0) = \left[\sum_u dN(u) V_z(u,0) \right]^{1/2},$$

where

$$V_z(u,0) = \frac{\sum_l [z_l - \bar{z}(u,0)]^2 Y_l(u)}{\sum_l Y_l(u)}.$$

Note: Among the $Y(u)$ individuals at risk at time u, there are $Y_1(u)$ individuals whose z_l value of $z_l = 1$ and $Y_0(u)$ individuals whose z_l value of $z_l = 0$. We already argued that

$$\bar{z}(u,0) = \frac{Y_1(u)}{Y(u)}.$$

Therefore,

$$\begin{aligned} V_z(u,0) &= \frac{\sum_l [z_l - \bar{z}(u,0)]^2 Y_l(u)}{\sum_l Y_l(u)} \\ &= \frac{\left[1 - \frac{Y_1(u)}{Y(u)}\right]^2 Y_1(u) + \left[0 - \frac{Y_1(u)}{Y(u)}\right]^2 Y_0(u)}{Y(u)} \quad (z_l(u) \text{ takes 1 or 0}) \\ &= \frac{\frac{Y_0^2(u) Y_1(u)}{Y^2(u)} + \frac{Y_1^2(u) Y_0(u)}{Y^2(u)}}{Y(u)} \qquad (Y_1(u) + Y_0(u) = Y(u)) \end{aligned}$$

$$= \frac{Y_0(u)Y_1(u)Y(u)}{Y^3(u)}$$
$$= \frac{Y_0(u)Y_1(u)}{Y^2(u)}.$$

Therefore,

$$J(0) = \sum_u dN(u)\frac{Y_0(u)Y_1(u)}{Y^2(u)}.$$

Let us contrast this with the variance used to compute the logrank test statistic:

$$\sum_u \left[\frac{Y_1(u)Y_0(u)dN(u)[Y(u) - dN(u)]}{Y^2(u)[Y(u) - 1]}\right].$$

Note: In the special case where $dN(u)$ can only be one or zero, then above expression reduces to

$$\sum_u \left[\frac{Y_1(u)Y_0(u)dN(u)[Y(u) - 1]}{Y^2(u)[Y(u) - 1]}\right] = \sum_u \left[\frac{Y_1(u)Y_0(u)dN(u)}{Y^2(u)}\right],$$

which is exactly equal to $J(0)$.

Therefore, we have demonstrated with continuous survival time data with no ties, the score test of the hypothesis $H_0 : \beta = 0$ in the proportional hazards model is exactly the same as the logrank test for dichotomous covariate z.

The score test

$$\left[\frac{U(0)}{J^{1/2}(0)}\right]^2$$

can be used to test the hypothesis $H_0 : \beta = 0$ for the model

$$\lambda(t|z) = \lambda_0(t)\mathrm{e}^{z\beta}$$

for any covariate value z, whether or not z is discrete or continuous. The null hypothesis $H_0 : \beta = 0$ implies that the hazard rate at any time t is unaffected by the covariate z. This also implies that the survival distribution does not depend on z. The alternative hypothesis $H_A : \beta \neq 0$ implies that hazard rate increases or decreases (depending on the sign of β) as z increases throughout all time. Therefore, belief in this alternative hypothesis would mean that individuals with a higher value of z would have stochastically larger (or smaller depending on the sign of

β) survival distribution than those individuals with a smaller values of z. The `test` command in `Proc Lifetest` computes the score test of the hypothesis $H_0 : \beta = 0$ for the proportional hazards model. Consequently, when using the `test` command, the covariate z is not limited to being dichotomous, nor discrete.

For example, we can test the treatment difference between treatments 1 and 2 for myelomatosis data using the following `SAS` command:

```
proc lifetest data=myel;
  time dur*status(0);
  test trt;
run;
```

and part of the output is presented in the following:

Univariate Chi-Squares for the LOG RANK Test

Variable	Test Statistic	Standard Deviation	Chi-Square	Pr > Chi-Square
TRT	-2.3376	2.0522	1.2975	0.2547

Covariance Matrix for the LOG RANK Statistics

Variable	TRT
TRT	4.21151

Forward Stepwise Sequence of Chi-Squares for the LOG RANK Test

Variable	DF	Chi-Square	Pr > Chi-Square	Chi-Square Increment	Pr > Increment
TRT	1	1.2975	0.2547	1.2975	0.2547

Likelihood Ratio Test

As in the ordinary likelihood theory, the (partial) likelihood ratio test can also be used to test the null hypothesis:

$$H_0 : \beta = \beta_0.$$

Recall that $\ell(\beta)$ is the log partial likelihood. Intuitively, if H_0 is true, then $\hat{\beta}$, the MPLE of β, should be close to β_0. Hence $\ell(\hat{\beta})$ should be close to $\ell(\beta_0)$. Since $\ell(\hat{\beta}) - \ell(\beta_0)$ is always non-negative, so we should reject H_0 when this difference is large.

The likelihood ratio test uses the fact that

$$2\left[\ell(\hat{\beta}) - \ell(\beta_0)\right] \overset{a}{\sim} \chi^2_1, \quad \text{under } H_0 : \beta = 0.$$

Therefore, for a given level of significance α, we reject $H_0 : \beta = \beta_0$ if

$$2\left[\ell(\hat{\beta}) - \ell(\beta_0)\right] \geq \chi^2_{1,\alpha}$$

where $\chi^2_{1,\alpha}$ is the value such that $P[\chi^2_1 > \chi^2_{1,\alpha}] = \alpha$.

Expanding $\ell(\beta_0)$ at the MPLE $\hat{\beta}$, we get

$$\ell(\beta_0) \approx \ell(\hat{\beta}) + \frac{d\ell(\hat{\beta})}{d\beta}(\beta_0 - \hat{\beta}) + \frac{1}{2!}\frac{d^2\ell(\hat{\beta})}{d^2\beta}(\beta_0 - \hat{\beta})^2.$$

Since MPLE $\hat{\beta}$ maximizes $\ell(\beta)$, *i.e.*,

$$U(\hat{\beta}) = \frac{d\ell(\hat{\beta})}{d\beta} = 0,$$

and

$$\frac{d^2\ell(\hat{\beta})}{d^2\beta} = -J(\hat{\beta}),$$

so

$$2\left[\ell(\hat{\beta}) - \ell(\beta_0)\right] \approx J(\hat{\beta})(\hat{\beta} - \beta_0)^2.$$

We already derived that

$$(\hat{\beta} - \beta_0) \overset{a}{\sim} \mathrm{N}(0, J^{-1}(\hat{\beta})).$$

Therefore,

$$\begin{aligned} 2\left[\ell(\hat{\beta}) - \ell(\beta_0)\right] &\approx J(\hat{\beta})(\hat{\beta} - \beta_0)^2 \\ &= \left[\frac{\hat{\beta} - \beta_0}{J^{-1/2}(\hat{\beta})}\right]^2 \overset{a}{\sim} \chi^2_1 \text{ under } H_0 : \beta = \beta_0. \end{aligned}$$

Note: The `SAS` procedure `Phreg` can ONLY handle right censored data.