6. Cox Regression Models

(Part I)

The Proportional Hazards Model

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

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

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 the model. 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 $z = 0$, i.e. $\lambda(t|z = 0) = \lambda_0(t)$.

The baseline hazard function $\lambda_0(t)$ 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 PH model

- It is easy to show that under PH model, $S(t|z) = [S_0(t)]^{\exp(z^T\beta)}$, 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

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

- For any two sets of covariates z_0 and z_1,

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

which is a constant over time. Equivalently, $\log\left[\frac{\lambda(t|z_1)}{\lambda(t|z_0)}\right] = (z_1 - z_0)^T\beta$.

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

$$\log\left[\frac{\lambda(t|z_k+1)}{\lambda(t|z_k)}\right] = \beta_k \qquad \text{or,} \qquad \frac{\lambda(t|z_k+1)}{\lambda(t|z_k)} = e^{\beta_k}$$

Therefore, β_k $\left(e^{\beta_k}\right)$ is the increase in log hazard-ratio (hazard-ratio) at **any** time with unit increase in the kth covariate z_k. Furthermore, since

$$\frac{\mathrm{P}[\mathrm{t} \le \mathrm{T} < \mathrm{t} + \Delta t|T \ge \mathrm{t}, \mathrm{z_k} + 1]}{\mathrm{P}[\mathrm{t} \le \mathrm{T} < \mathrm{t} + \Delta t|T \ge \mathrm{t}, \mathrm{z_k}]} \approx \frac{\lambda(t|z_k+1)\Delta t}{\lambda(t|z_k)\Delta t} = e^{\beta_k}$$

so e^{β_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.

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.

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.

Estimation Using Partial Likelihood

Data: $(X_i, \Delta_i, z_i), i = 1, \dots, n$, where for the ith individual,

$X_i = \min(T_i, C_i)\,; \Delta_i = I(T_i, \leq C_i); z_i = \left(z_{i1}, z_{i2}, \dots, z_{ip}\right)^T$ is a vector of covariates.

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

$$\lambda(t|z_i) = \lim_{h \to 0^+} \left\{ \frac{\mathrm{P}[\mathrm{t} \leq \mathrm{T} < \mathrm{t} + h | T \geq \mathrm{t}, \mathrm{z_k}]}{h} \right\}$$

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)

$$P[\mathrm{x} \leq \mathrm{X_i} < \mathrm{x} + \Delta x, \Delta_i = 1 | \mathrm{X_i} \geq \mathrm{x}, \mathrm{z_i}] = P[\mathrm{x} \leq T_\mathrm{i} < \mathrm{x} + \Delta x | \mathrm{T_i} \geq \mathrm{x}, \mathrm{z}_i] \approx \lambda_{T_i}(x|z_i)\Delta x$$

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.

Notations

- 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,
$$Y_i(u) = I(X_i \geq u)$$
- Let $dN(x) = \sum_{i=1}^{n} dN_i(x)$ 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 $Y(x) = \sum_{i=1}^{n} Y_i(x)$ be the total number from the entire sample who are at risk at time u.
- Let $F(x)$ denote the information up to time *x* (one of the grid points)
$F(x) = \{(dN_i(u), Y_i(u), z_i), i = 1, \dots, n \text{ , for all grid points } u < x \text{ and } dN(x)\}$
Also, $F(\infty) = \{(X_i, \Delta_i, z_i), i = 1, \dots, n\}$ denote all the data in the sample.
- 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, $I(x) = j$ means that the jth individual died in $[x, x+\Delta x)$.
- Then the data (with redundancy) can be expressed as
$[F(u_1), I(u_1), F(u_2), I(u_2), \dots, F(\infty)]$, where $u_1 < u_2 < \cdots$ are the grid points

The Partial Likelihood

The likelihood of the parameter $\lambda_0(t)$ and β can be written as

$$P[F(u_1) = f(u_1); \lambda_0(t), \beta] \times P[I(u_1) = i(u_1) | F(u_1) = f(u_1); \lambda_0(t), \beta]$$
$$\times P[F(u_2) = f(u_2) | F(u_1) = f(u_1), I(u_1) = i(u_1); \lambda_0(t), \beta]$$
$$\times P[I(u_2) = i(u_2) | F(u_2) = f(u_2), F(u_1) = f(u_1), I(u_1) = i(u_1); \lambda_0(t), \beta]$$
$$\times \cdots$$

and the last term can be simplified as

$$P[I(u_2) = i(u_2) | F(u_2) = f(u_2), F(u_1) = f(u_1), I(u_1) = i(u_1); \lambda_0(t), \beta]$$
$$= P[I(u_2) = i(u_2) | F(u_2) = f(u_2); \lambda_0(t), \beta]$$

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

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

Ex:

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

$$PL = \frac{e^{2\beta}}{e^{2\beta}+e^{2\beta}+e^{\beta}+e^{3\beta}} \times 1 \times \frac{e^{\beta}}{e^{\beta}+e^{3\beta}} \times \frac{e^{3\beta}}{e^{3\beta}}$$

Case 1: Suppose conditional on $F(u)$ 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

$$P[I(u) = 0 | F(u) = f(u)\,] = 1$$

Case 2: $dN(u) = 1$. Conditional on $F(u)$, 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 $F(u)$, 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?*

A: It is **proportional to** *the* i*th subject's* cause-specific hazard of dying at time u.

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 $A_1, A_2, \cdots, A_n$ are mutually exclusive.

Also, under independent censoring assumption, the cause-specific hazard is the same as the hazard of interest (Chapter 3), *i.e.*, $\lambda(u, \delta_i = 1|z_i) = \lambda(u|z_i)$.

Since Δu is chosen to be very small, so

$$P[A_i] \approx Y_i(u)\lambda(u, \delta_i = 1|z_i)\Delta u = Y_i(u)\lambda(u|z_i)\Delta u = Y_i(u)\lambda_0(u)e^{z^T\beta}\Delta u$$

assumption of the cox model

$$P[I(u) = i(u)|F(u) = f(u); \lambda_0(t), \beta] = P\left[A_{i(u)}|A_1 \cup \cdots \cup A_n\right] = \frac{P\left[A_{i(u)}\right]}{\sum_{l=1}^{n} P[A_l]}$$

$$\approx \frac{\lambda_0(u)e^{z_{i(u)}^T\beta}\Delta u}{\sum_{l=1}^{n} \lambda_0(u)e^{z_l^T\beta}\Delta u Y_l(u)} = \frac{e^{z_{i(u)}^T\beta}}{\sum_{l=1}^{n} e^{z_l^T\beta}Y_l(u)}$$

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

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

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

$$PL(\beta) = \prod_{j=1}^{d} \left[\frac{e^{z_{i(u)}^T \beta}}{\sum_{l=1}^{n} e^{z_l^T \beta} Y_l(t_j)} \right]$$

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

$$PL(\beta) = \prod_{i=1}^{n} \left[\frac{e^{z_{i(u)}^T \beta}}{\sum_{l=1}^{n} e^{z_l^T \beta} Y_l(t_j)} \right]^{\delta_i}$$

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 log partial likelihood function of β is

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

The score function is

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

And the second derivative is

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

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

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

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

$$= \sum_{l=1}^n z_l^2 w_l - \left(\bar{z}(u,\beta)\right)^2$$

$$= \sum_{l=1}^n (z_l - \bar{z}(u,\beta))^2 w_l$$

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,

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

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

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

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

Terminology: The quantity

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

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: $\hat{\beta}$ 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:

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

$$\left(\hat{\beta} - \beta_0\right) \approx -\left[\frac{\partial U(\beta_0)}{\partial \beta}\right]^{-1} U(\beta_0) = [J(\beta_0)]^{-1} U(\beta_0)$$

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

$$U(\beta_0) = \sum_u dN(u)\,[z_l - \bar{z}(u,\beta)]$$

Properties of the Score – I

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

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

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

Notice that: $E\left[z_{I(u)}|F(u)\right] = \dfrac{\sum_{l=1}^{n} z_l e^{z_l\beta_0} Y_l(u)}{\sum_{l=1}^{n} e^{z_l\beta_0} Y_l(u)} = \sum_{l=1}^{n} z_l w_l = \bar{z}(u,\beta_0)$

Therefore, $\boxed{\mathrm{E}[U(\beta_0)] = 0}$

Conditional distribution of $z_{I(u)}$ given $F(u)$

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

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

$$Var\left[z_{I(u)}|F(u)\right] = \sum_{l=1}^{n}\left(z_l - E\left[z_{I(u)}|F(u)\right]\right)^2 w_l = \frac{\sum_{l=1}^{n}\left(z_l - \bar{z}(u,\beta_0)\right)^2 e^{z_l\beta_0}Y_l(u)}{\sum_{l=1}^{n} e^{z_l\beta_0}Y_l(u)}$$

$$= V_z(u,\beta_0)$$

Properties of the Score – II

$$\mathrm{Var}[U(\beta_0)] = E[U(\beta_0)]^2 = \mathrm{E}\left[\sum_u dN(u)\left(z_{I(u)} - \bar{z}(u,\beta_0)\right)\right]^2$$

$$= \mathrm{E}\left[\sum_u \left\{dN(u)\left(z_{I(u)} - \bar{z}(u,\beta_0)\right)\right\}^2\right]$$

$$+ \mathrm{E}\left[\sum_{u \neq u'} \left\{dN(u)\left(z_{I(u)} - \bar{z}(u,\beta_0)\right)\right\}\left\{dN(u')\left(z_{I(u')} - \bar{z}(u',\beta_0)\right)\right\}\right]$$

As usual, WLOG, assume $u' > u$, and denote

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

Then $E[A(u)A(u')] = E\{\,E[A(u)A(u')|F(u')]\,\} = E\{A(u)E[A(u')|F(u')]\,\} = 0$

So, the expectation of the cross-product is zero. Then,

$$\mathrm{Var}[U(\beta_0)] = \sum_u E[A^2(u)] = \sum_u E\{E[A^2(u)|F(u)]\}$$

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

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

$[dN(u)]^2 = dN(u)$

Conditional on $F(u)$, $dN(u)$ is known, $\bar{z}(u,\beta_0)$ is also known, and $\bar{z}(u,\beta_0) = E\left[z_{I(u)}|F(u)\right]$

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

Consequently, $\mathrm{Var}[U(\beta_0)] = \sum_u E[dN(u)V_z(u,\beta_0)] = E\left[\sum_u dN(u)V_z(u,\beta_0)\right]$

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 $\mathrm{Var}[U(\beta_0)]$. In fact, $\sum_u dN(u)V_z(u,\beta_0)$ is the partial likelihood observed information $J(\beta_0)$ we defined before.

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 $J(\beta_0)$. By the martingale CLT, we have:

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

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

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

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

In practice, since β_0 is unknown, we can substitute $\hat{\beta}$ for β_0 for the variance of $U(\beta_0)$.

Note that: $J^{-1}(\hat{\beta}) = \sum_u dN(u)V_z(u,\hat{\beta})$

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), the MPLE $\hat{\beta}$ can be obtained by solving the partial likelihood equation; i.e., setting the partial score to zero. Then asymptotically,

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

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

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

Example: Myelomatosis data (SAS code)

Define a treatment indicator trt1 which takes value 0 for treatment 1 and takes value 1 for treatment 2.

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
trt1	1	0.57276	0.50960	1.2633	0.2610	1.773

A 95% CI of β is:

$0.5728 \pm 1.96 * 0.5096 = [-0.426, 1.572]$

A 95% CI of the hazard ratio e^{β} is: $\left[\mathrm{e}^{-0.426}, \mathrm{e}^{1.572}\right] = [0.653, 4.816]$

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

$$U(0) \overset{\text{a}}{\sim} N(0, J(0)) \qquad \text{or,} \qquad \left[\frac{U(0)}{J^{1/2}(0)}\right]^2 \overset{\text{a}}{\sim} \chi_1^2$$

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

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 $\left[z_{I(u)} - \bar{z}(u, 0\,)\right]$. Otherwise no contribution.
2. Since $z = 1$ for treatment 1 and $z = 0$ for treatment 0, $z_{I(u)}$ will then be 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 then numerator of the two-sample log rank test:

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

5. The denominator of the score test was computed as

$$J^{\frac{1}{2}}(0) = \left[\sum_u dN(u) V_z(u,0) \right]^{\frac{1}{2}} = \left[\sum_u dN(u) \frac{Y_0(u) Y_1(u)}{Y^2(u)} \right]^{\frac{1}{2}}$$

since $V_z(u,0) = \frac{\sum_l [z_l - \bar{z}(u,0)]^2 Y_l(u)}{\sum_l Y_l(u)} = \frac{\left[1 - \frac{Y_1(u)}{Y(u)}\right]^2 Y_1(u) + \left[0 - \frac{Y_1(u)}{Y(u)}\right]^2 Y_0(u)}{Y(u)}$

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

Note that In the special case where $dN(u)$can only be one or zero, the variance used to compute the logrank test statistic is:

$$\sum_u \left[\frac{Y_1(u) Y_0(u) dN(u) [Y(u) - dN(u)]}{Y^2(u)[Y(u) - 1]} \right] = \sum_u \left[\frac{Y_0(u) Y_1(u)}{Y^2(u)} \right], \text{ which is } J(0)$$

Therefore, for 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.

In fact, the score test $\left[\frac{U(0)}{J^{1/2}(0)}\right]^2$ can be used to test for any covariate value z,

whether or not z is discrete or continuous.

The hypothesis H_0: $\beta = 0$ implies that the hazard rate at any time t is unaffected by the covariate z, or that the survival distribution does not depend on z. The alternative hypothesis H_a: $\beta \neq 0$ 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. Note that when using the *test* command, the covariate z is not limited to being dichotomous, nor discrete.

Example: Myelomatosis data revisited (SAS code)

Likelihood Ratio Test

As in the ordinary likelihood theory, the (partial) likelihood ratio test can also be used to test the null hypothesis: $\mathrm{H}_0\colon \beta = \beta_0$

The likelihood ratio test uses the fact that under$\mathrm{H}_0\colon \beta = \beta_0$,

$$2\left[l\left(\hat{\beta}\right) - l(\beta_0)\right] \overset{\text{a}}{\sim} \chi_1^2$$

Therefore, for a given level of significance α, we reject $\mathrm{H}_0\colon \beta = \beta_0$ if

$$2\left[l\left(\hat{\beta}\right) - l(\beta_0)\right] \geq \chi_{1,\alpha}^2 \qquad \left(P\left[\chi_1 > \chi_{1,\alpha}^2\right] = \alpha\right)$$

A Heuristic Proof: Expanding $l(\beta_0)$ at the MPLE $\hat{\beta}$, we get

$$l(\beta_0) \approx l\left(\hat{\beta}\right) + \frac{dl\left(\hat{\beta}\right)}{d\beta}\left(\hat{\beta} - \beta_0\right) + \frac{1}{2!}\frac{d^2 l\left(\hat{\beta}\right)}{d^2\beta}\left(\hat{\beta} - \beta_0\right)^2$$

Since $\hat{\beta}$ maximize $l(\beta)$, i. e.

$$U\left(\hat{\beta}\right) = \frac{dl\left(\hat{\beta}\right)}{d\beta} = 0 \quad \text{and} \quad \frac{d^2 l\left(\hat{\beta}\right)}{d^2\beta} = -J\left(\hat{\beta}\right)$$

So, $\mathrm{H}_0\colon \beta = \beta_0$,

$$2\left[l\left(\hat{\beta}\right) - l(\beta_0)\right] \approx J\left(\hat{\beta}\right)\left(\hat{\beta} - \beta_0\right)^2 = \left[\frac{\hat{\beta} - \beta_0}{J^{-1/2}(0)}\right]^2 \overset{\text{a}}{\sim} \chi_1^2 \qquad \text{By } \left(\hat{\beta} - \beta_0\right) \overset{\text{a}}{\sim} N\left(0, J^{-1}\left(\hat{\beta}\right)\right)$$

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