Emory University
Anna Guo

Applied survival analysis
2021 FALL

2023-03-16

Contents

1 introduction

Survival analysis Survival analysis is a kind of data analysis tricks for which the outcome variable of interest is time until an event occurs. In order to calculate time to event, you must clearly define the time origin and the event of interest.
Types of censoring

- Right censoring (most common): true survival time is more than the observed survival time
- Left censoring: true survival time is less than the observed survival time
- Interval censoring: survival time is only known to be between two value

Survival function probability that a person survives longer than some specified time t

$$S(t) = \mathrm{P}[\text{ surviving beyond time } t] = \mathrm{P}[T > t]$$

Complement of survival function probability that survival time is at most some specified time t

$$F(t) = 1 - S(t) = 1 - \mathrm{P}[T > t] = \mathrm{P}[T \leq t]$$

Rate

$$\text{derivative of } S(t) = \frac{d}{dt}S(t) = S'(t) = \lim_{\Delta t \to 0} \frac{S(t+\Delta t) - S(t)}{\Delta t}$$

$$\begin{aligned} \frac{d}{dt}S(t) &\approx \frac{S(t+\Delta t) - S(t)}{\Delta t} \\ &= \text{ slope of a straight line between } S(t) \text{ and } S(t+\Delta t) \end{aligned}$$

$$\text{rate } = \frac{\text{change in } S(t)}{\text{change in time}} = \frac{S(t+\Delta t) - S(t)}{(t+\Delta t) - t} = \frac{S(t+\Delta t) - S(t)}{\Delta t}$$

Hazard rate add ”-” here to make this number positive

$$\text{hazard rate } = h(t) = -\frac{S'(t)}{S(t)} = -\frac{\partial \log S(t)}{\partial t}$$

Interpretation: the probability that people who lived to t died within the next interval is the hazard rate

Cumulative hazard function

$$H(t) = \int_0^t h(s)ds$$

Connections

$$H(t) = -\log S(t), \quad S(t) = e^{-H(t)} = e^{-\int_0^t h(s)ds}$$

Odds ratio & relative risk Under the case

	Diseased	Healthy
Exposed	20	380
Not exposed	6	594

we have

$$\text{Relative Risk } = \frac{D_E/V_E}{D_N/V_N} = \frac{20/400}{6/600} = \frac{.05}{.01} = 5$$

$$\text{Odds Ratio } = \frac{D_E/H_E}{D_N/H_N} = \frac{20/380}{6/594} \approx \frac{.052}{.010} = 5.2$$

1.1 Average rate as estimate of hazard rate

$$\begin{aligned}
&\text{average (mortality) rate} \\
&\text{during the interval } (t, t+\Delta t) = \frac{\text{number of deaths}}{\text{total time-at-risk}} = \frac{N(S(t)-S(t+\Delta t))}{N(\int_t^{t+\Delta t} S(u)du)} \\
&= \frac{\text{mean number of deaths}}{\text{mean survival time}} = \frac{S(t)-S(t+\Delta t)}{\int_t^{t+\Delta t} S(u)du} \\
&\approx \frac{S(t)-S(t+\Delta t)}{\frac{1}{2}\Delta t[S(t)+S(t+\Delta t)]}
\end{aligned}$$

1.2 Probability q and rate R

$$\begin{aligned}
\text{probability of death } &= q = P[\text{ death between } t \text{ and } t+\Delta t \mid \text{ alive at time } t] \\
&= \frac{P[\text{ dead between } t \text{ and } t+\Delta t]}{P[\text{ alive at time } t]} = \frac{S(t)-S(t+\Delta t)}{S(t)} = \frac{d(t)}{S(t)}
\end{aligned}$$

Let R be the average approximate rate. The complementary probability of survival is $1-q=p=\frac{S(t+\Delta t)}{S(t)}$

$$R = \frac{S(t)-S(t+\Delta t)}{\Delta t\left[S(t)-\frac{1}{2}d(t)\right]} = \frac{S(t)/S(t)-S(t+\Delta t)/S(t)}{\Delta t\left[S(t)/S(t)-\frac{1}{2}d(t)/S(t)\right]} = \frac{q}{\Delta t\left(1-\frac{1}{2}q\right)}$$

solve for q

$$q = \frac{\Delta t R}{1+\frac{1}{2}\Delta t R}$$

the probability of death or disease in human population is almost always small ($p \approx 1$ or $q \approx 0$), under this circumstance

$$R \approx q/\Delta t$$

2 Life Tables

2.1 Types of life tables

Cohort life tables the cohort life table presents the mortality experience of a particular birth cohort.For example, reflects the mortality experience of an actual cohort from birth until no lives remain in the group.

Current life tables the period life table presents what would happen to a hypothetical cohort.The period life table may thus be characterized as rendering a ‘snapshot’ of current mortality experience and shows the long-range implications of a set of age-specific death rates that prevailed in a given year.

Abridged life tables an abridged life table is based on a sequence of age intervals of any chosen length, typically five years.

Complete life tables a complete life table contains data for every single year of age.

Our focus of the following with be on the current and complete life table.

2.2 Life table construction

We need seven elements in this process, which are as follows:
Age interval (x to x+1) the symbol x represents the age of the individuals described by the life table. Each age interval is one year except the last, which is open-ended

Number alive (l_x) represents the number of individuals alive (at-risk) at age x. It is the size of the life table population-at-risk at the beginning of the interval x. l_0, the number alive at age x=0 is set as some arbitrary number, say 100,000. l_0 is called the *radix*.

Deaths (d_x) represents the number of deaths between ages x and x + 1 (i.e., within this particular year).

Probability of death (q_x) represents the *conditional* probability that a member of the life table cohort who is alive at age x dies before age x + 1. In symbols,

$$\text{Conditional probability of death } \; q_x = P[\text{ death before age } x+1 \mid \text{ alive at age } x)] = \frac{d_x}{l_x}$$

Probability of alive (p_x) represents the *conditional* probability that a member of the life table cohort who is alive at age x is still alive at age x + 1. In symbols, $p_x = 1 - q_x$
Note: make sure you distinguish between the conditional survival probability p_x (conditional on a specific age) and the unconditional survival probability P_x. To calculate P_x, we can use

$$\begin{aligned}\text{life table probability of surviving beyond age } x = P_x & \\ &= \prod(1-q_i) = \prod p_i \\ &= \frac{l_1}{l_0}\cdot\frac{l_2}{l_1}\cdot\frac{l_3}{l_2}\cdots\frac{l_x}{l_{x-1}} \\ &= \frac{l_x}{l_0}\end{aligned}$$

Years lived (L_x) represents cumulative time lived by the entire cohort between the ages x and x + 1. Each individual alive at age x contributes to the total time lived during the next year: either *one year* if an individual lives the entire year or the proportion of the year lived when the individual dies within the one-year interval, usually taken as *1/2*.

$$L_x = \overbrace{(l_x - d_x)}^{\text{persons}} \underbrace{(1)}_{\text{year}} + \overbrace{\bar{a}_x}^{\text{year}} \underbrace{d_x}_{\text{persons}}$$

Total time lived (T_x) represents the total time lived beyond age x by all individuals who are age x. In symbols, $T_x = L_x + L_{x+1} + L_{x+2} + \cdots$

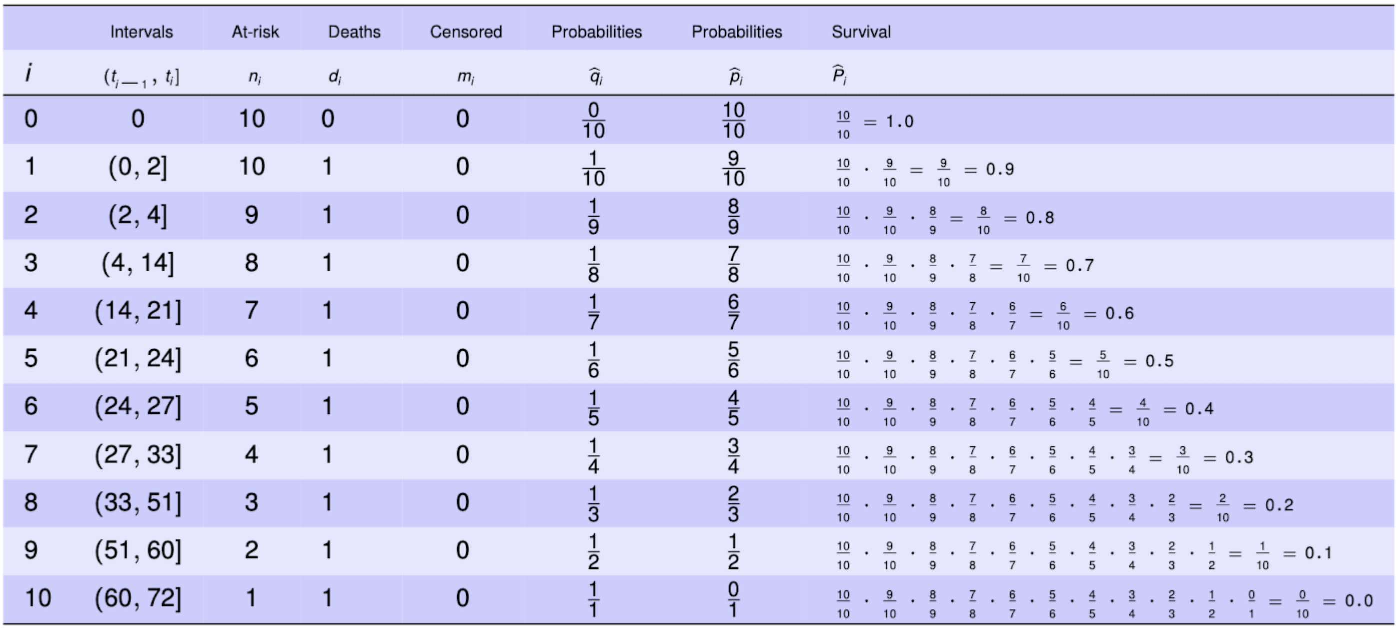
Expectation of life (e_x) represents the mean number of additional years lived by those members of the life table cohort who are age x. In symbols, $e_x = \frac{T_x}{l_x}$

$$\text{crude mortality rate } = \frac{\text{total deaths}}{\text{total person-years}} = \frac{l_0}{T_0} = \frac{1}{e_0}$$

3 Descriptive Methods for Survival Data

3.1 Unique & complete survival times

Unique means that all sampled survival times are different. *Complete* means that all survival times end in an observed outcome such as death. See the following example: Consider the unique and complete survival times of 10 (n = 10) hypothetical COVID-19 patients in the ICU: survival times (in days): 2, 72, 51, 60, 33, 27, 14, 24, 4, and 21. Unlike constructing a life table, $t_i - t_{i-1} \neq 1$, that is, the interval lengths vary and are determined by observed values.



	Intervals	At-risk	Deaths	Censored	Probabilities	Probabilities	Survival
i	$(t_{i-1}, t_i]$	n_i	d_i	m_i	$\hat{q}_i$	$\hat{p}_i$	$\hat{P}_i$
0	0	10	0	0	$\frac{0}{10}$	$\frac{10}{10}$	$\frac{10}{10} = 1.0$
1	$(0, 2]$	10	1	0	$\frac{1}{10}$	$\frac{9}{10}$	$\frac{10}{10}\cdot\frac{9}{10} = \frac{9}{10} = 0.9$
2	$(2, 4]$	9	1	0	$\frac{1}{9}$	$\frac{8}{9}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9} = \frac{8}{10} = 0.8$
3	$(4, 14]$	8	1	0	$\frac{1}{8}$	$\frac{7}{8}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8} = \frac{7}{10} = 0.7$
4	$(14, 21]$	7	1	0	$\frac{1}{7}$	$\frac{6}{7}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{6}{7} = \frac{6}{10} = 0.6$
5	$(21, 24]$	6	1	0	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6} = \frac{5}{10} = 0.5$
6	$(24, 27]$	5	1	0	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{4}{5} = \frac{4}{10} = 0.4$
7	$(27, 33]$	4	1	0	$\frac{1}{4}$	$\frac{3}{4}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{3}{4} = \frac{3}{10} = 0.3$
8	$(33, 51]$	3	1	0	$\frac{1}{3}$	$\frac{2}{3}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{3}{4}\cdot\frac{2}{3} = \frac{2}{10} = 0.2$
9	$(51, 60]$	2	1	0	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{3}{4}\cdot\frac{2}{3}\cdot\frac{1}{2} = \frac{1}{10} = 0.1$
10	$(60, 72]$	1	1	0	$\frac{1}{1}$	$\frac{0}{1}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{6}{7}\cdot\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{3}{4}\cdot\frac{2}{3}\cdot\frac{1}{2}\cdot\frac{0}{1} = \frac{0}{10} = 0.0$

Under the *unique* and *complete* assumption, we have

$$\widehat{q}_i = \mathrm{P}\,[\,\text{death at or before time } t_i \mid \text{ alive at } t_{i-1}] = \frac{\text{number of deaths}}{\text{number at risk}} = \frac{1}{n-i+1}$$

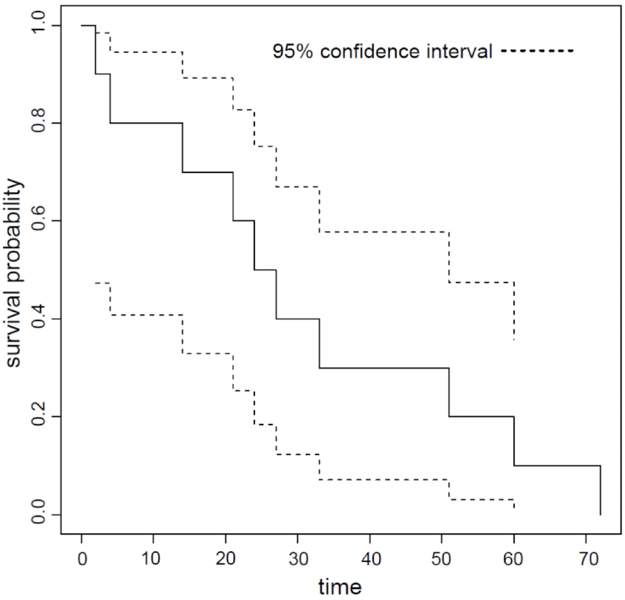
$$\widehat{p}_i = \mathrm{P}\,[\,\text{alive after time } t_i \mid \text{ alive at } t_{i-1}] = \frac{n-i}{n-i+1}$$

3.2 Kaplan-Meier estimator

Kaplan-Meier estimator the estimated probability of surviving beyond a specific time t_k which is $\hat{P}_k$ estimated via

$$\begin{aligned}\widehat{P}_k &= \widehat{p}_1 \cdot \widehat{p}_2 \cdots \widehat{p}_k \\ &= \prod \hat{p}_i = \prod \frac{n-i}{n-i+1} \\ &= 1 - \frac{k}{n} = \frac{n-k}{n},\ i = 1, 2, \ldots, k\end{aligned}$$

is called the *Kaplan-Meier estimate*
Plot the estimated survival probability against time we get the Kaplan-Meier curve from the step function.



Variance: the estimate of $\hat{P}_k$ from complete data is a typical estimate of a binomial probability. The variance is given by

$$\operatorname{Var}\left[\widehat{P}_k\right] = \widehat{P}_k\left(1 - \widehat{P}_k\right)/n$$

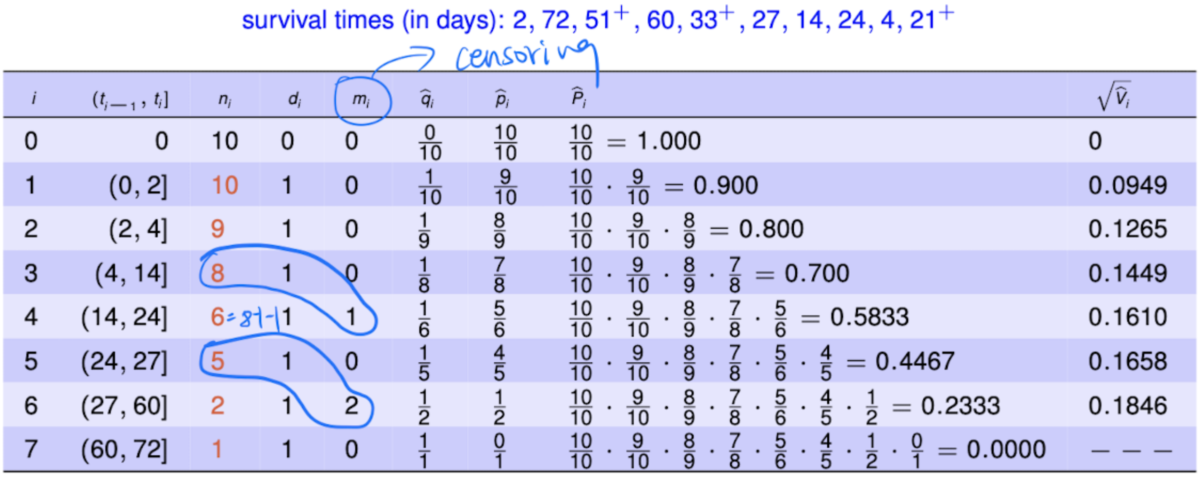
Mean survival time 1) mean survival time $= \bar{t} = \frac{1}{n}\sum t_i, i = 1, 2, \ldots, n$ 2) or based on the total area enclosed by the estimated survival function

$$\text{mean survival time } = \text{ area } = \widehat{\mu} = \sum \widehat{P}_{i-1}\left(t_i - t_{i-1}\right)$$

Median survival time When an estimated survival probability $\hat{P}_i$ does not exactly equal 0.5, an estimate of the median value is the upper bound of the interval containing the survival probability $\hat{P} = 0.5$

3.3 Incomplete survival times

Assumption The following analysis takes care of the case when the censoring is **non-informative**, that is, the reason that the time of death is not observed is entirely unrelated to the outcome under study.



survival times (in days): 2, 72, 51^+, 60, 33^+, 27, 14, 24, 4, 21^+

censoring

i	$(t_{i-1}, t_i]$	n_i	d_i	m_i	$\widehat{q}_i$	$\widehat{p}_i$	$\widehat{P}_i$	$\sqrt{\widehat{V}_i}$
0	0	10	0	0	$\frac{0}{10}$	$\frac{10}{10}$	$\frac{10}{10} = 1.000$	0
1	(0, 2]	10	1	0	$\frac{1}{10}$	$\frac{9}{10}$	$\frac{10}{10}\cdot\frac{9}{10} = 0.900$	0.0949
2	(2, 4]	9	1	0	$\frac{1}{9}$	$\frac{8}{9}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9} = 0.800$	0.1265
3	(4, 14]	8	1	0	$\frac{1}{8}$	$\frac{7}{8}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8} = 0.700$	0.1449
4	(14, 24]	6=8-1-1	1	1	$\frac{1}{6}$	$\frac{5}{6}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{5}{6} = 0.5833$	0.1610
5	(24, 27]	5	1	0	$\frac{1}{5}$	$\frac{4}{5}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{5}{6}\cdot\frac{4}{5} = 0.4467$	0.1658
6	(27, 60]	2	1	2	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{1}{2} = 0.2333$	0.1846
7	(60, 72]	1	1	0	$\frac{1}{1}$	$\frac{0}{1}$	$\frac{10}{10}\cdot\frac{9}{10}\cdot\frac{8}{9}\cdot\frac{7}{8}\cdot\frac{5}{6}\cdot\frac{4}{5}\cdot\frac{1}{2}\cdot\frac{0}{1} = 0.0000$	— — —

Adjustments the adjustments made considering censoring is that number of individuals at risk in each interval is **# of people at risk at beginning - # of people censored**.

Therefore, in each interval only two kinds of individuals enter into the calculation of the conditional survival probability: (1) those who died (creating the endpoint of the interval) and (2) those who survived the entire interval.

Greenwood's variance The variance of the distribution of an estimated survival probability is calculated from Greenwood's variance.

$$\widehat{V}_k = \operatorname{Var}\left[\widehat{P}_k\right] = \widehat{P}_k^2 \sum \frac{\widehat{q}_i}{n_i \hat{p}_i}, \quad i = 1, 2, \ldots, k$$

Confidence interval The construction of CI needs special care, i.e, use transformation to construct a more normally distributed variable

$$\widehat{s}_k = \log\left[-\log\left(\widehat{P}_k\right)\right]$$

then we have

$$\operatorname{Var}\left[\widehat{s}_k\right] = \operatorname{Var}\left[\log\left[-\log\left(\widehat{P}_k\right)\right]\right] = \frac{\widehat{V}_k}{\left[\widehat{P}_k \log\left(\widehat{P}_k\right)\right]^2}$$

then CI for $\hat{s_k}$ is upper-bound

$$A_k = \widehat{s}_k - 1.960\sqrt{\operatorname{Var}\left[\widehat{s}_k\right]}$$

and lower-bound

$$B_k = \widehat{s}_k + 1.960\sqrt{\operatorname{Var}\left[\widehat{s}_k\right]}$$

then $\hat{P}_k$ has CI

$$\left[e^{-\exp(B_k)}, e^{-\exp(A_k)}\right]$$

Variance of mean survival time define

$$A_k = \sum \text{ area }_i = \sum \widehat{P}_{i-1}\left(t_i - t_{i-1}\right), \quad i = k+1, k+2, \ldots, d$$

note that $A_0 = \hat{\mu}$ then we have

$$\mathrm{Var}\left[A_k\right] = \sum \frac{A_k^2}{n_k\left(n_k - 1\right)}, \quad k = 1, 2, \ldots, d$$

Variance of median survival time let $\hat{t_m}$ be the estimated median survival time. Then

$$\mathrm{Var}\left[\widehat{t}_m\right] = \left[\frac{t_l - t_u}{\widehat{P}_l - \widehat{P}_u}\right]^2 \widehat{V}_m$$

where t_l and t_u are the lower and upper bound that $\hat{t_m}$ located and $\hat{P_l}$ and $\hat{P_u}$ are the corresponding survival probability. $\hat{V_m}$ represents the Greenwood estimated variance of the estimated value $\hat{P_m}$ (e.g. if 27 is the median survival time in the above table, then $t_l = 24, \quad t_u = 60$)

3.4 Nelson-Aalen estimator

Cumulative hazard function cumulative hazard function is defined as

$$H(t) = \int_0^t h(u)du$$

and we have

$$\frac{d}{dt}H(t) = h(t)$$
$$S(t) = \exp[-H(t)]$$

then

$$\widehat{H}(t) = -\log \widehat{P}_k = -\log \prod \widehat{p}_i = -\sum \log \widehat{p}_i, \quad i = 1, 2, \ldots k$$

we have $\log p = \log(1 - q) \approx -q$ thus

$$\widetilde{H}(t) = \sum \hat{q}_i = \sum \frac{d_i}{n_i}, \quad i = 1, 2, \ldots k$$

the estimator of the variance of the Nelson-Aalen estimator is

$$\widehat{\mathrm{Var}}[\widetilde{H}(t)] = \sum \frac{d_i}{n_i^2}, \quad i = 1, 2, \ldots k$$

and it follows that the endpoints of the 95% confidence interval estimator are

$$\widetilde{H}(t) \quad \pm \quad 1.96\sqrt{\widehat{\mathrm{Var}}[\widetilde{H}(t)]}$$

thus the Nelson-Aalen estimator of the survival function is

$$\widehat{S}^*(t) = e^{-\widetilde{H}(t)}$$

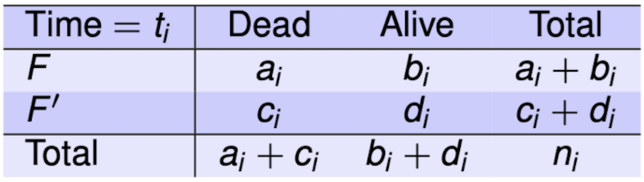
and the associated confidence interval is obtained by exponentiating the negative of the endpoints

$$\exp\{-\widetilde{H}(t) \pm 1.96\sqrt{\widehat{\mathrm{Var}}[\widetilde{H}(t)]}\}$$

The Nelson-Aalen estimator of the survival function is always $\geq$ the Kaplan-Meier estimator

3.5 Comparison of Survival Functions

Log-rank test it begins with drawing a 2*2 table at each time interval $(t_{i-1}, t_i]$, and the two compared groups are referred to as risk factor present (F) or absent ($F\prime$)



Time $= t_i$	Dead	Alive	Total
F	a_i	b_i	$a_i + b_i$
F'	c_i	d_i	$c_i + d_i$
Total	$a_i + c_i$	$b_i + d_i$	n_i

our null assumption is that there is no difference in survival probability between the case and control groups. Under this assumption

$$\text{expected number of deaths when risk factor is present: } A_i = \left[\tfrac{a_i+c_i}{n_i}\right](a_i+b_i)$$
$$\text{expected number of deaths when risk factor is absent: } C_i = \left[\tfrac{a_i+c_i}{n_i}\right](c_i+d_i)$$

and the corresponding observed numbers are a_i and c_i , respectively.
in filling the tables, we simply remove the censoring observations at each time interval, just like what we did in the survival table
the variance of the distribution of the a_i -counts is estimated with the expression

$$\operatorname{Var}[a_i] = \widehat{v}_i = \frac{(a_i+b_i)(a_i+c_i)(c_i+d_i)(b_i+d_i)}{n_i^2(n_i-1)}$$

when $a_i + c_i = 1$, then $b_i + d_i = n_i - 1$, i.e. no identical survival times occur, the expression for the same variance estimate is

$$\operatorname{Var}[a_i] = \widehat{V}_i = \frac{(a_i+b_i)(c_i+d_i)}{n_i^2}$$

now we have

- the total number of death among individuals with the presence of risk factor, represented by $\sum a_i$
- the total number of death among individuals with the risk factor estimated as if the risk factor and survival status were unrelated, represented by $\sum A_i$
- the variance of the summary $\sum a_i$ represented by $\sum \hat{V}_i$

then a formal test statistic measuring the overall strength of the association is

$$X^2 = \left[\frac{\sum(a_i - A_i)}{\sqrt{\operatorname{Var}[\sum a_i]}}\right]^2 = \frac{[\sum(a_i-A_i)]^2}{\operatorname{Var}[\sum a_i]} = \frac{[\sum a_i - \sum A_i]^2}{\sum \hat{v}_i}$$

the test statistic X^2 then has an approximate chi-square distribution with 1 degree of freedom when a_i and A_i differ by chance alone.
Note: The application of a chi-square distribution is not strictly correct. The estimated variance of a sum (i.e., $\operatorname{Var}(\sum a_i)$) is the sum of the estimated variances ($\sum \operatorname{Var}(a_i)$) only when the values a_i are uncorrelated. This is not the case among the 2 × 2 log-rank tables. All but the first table contain participants from earlier tables, introducing a table-to-table association. This lack of independence, however, has only a minor influence on the accuracy of the summary chi-square test statistic.

3.5.1 log-rank test for more than two groups

$$\mathrm{H}_0 : S_1(t) = S_2(t) = \cdots = S_G(t) \text{ for all } t$$

for more than two groups, the log-rank statistic has approximately a large sample chi-square distribution with df = G-1, where G is the number of groups

3.5.2 comparison between different tests

Wilcoxon test the generalized Wilcoxon test uses weights equal to the number at risk, it puts relatively more weight on differences between the survival functions at smaller values of time

log-rank test the log-rank test, because it uses weights equal to 1, places more emphasis than does the generalized Wilcoxon test on differences between the functions at larger values of time. It is a crude test; no measure of association is reported, only a p-value; the exposure can- not be continuous. That is, continuous exposures need to be converted into categorical exposures.

Tarone-Ware test Tarone-Ware uses weights equal to the square root of the number of subjects at risk at each survival time.

Peto-Prentice test Peto-Prentice suggested using a weight function that depends more explicitly on the observed survival experience of the combined sample.

4 Cox Proportional Hazards Regression

4.1 Cox PH model

Cox PH model

$$h(t, \mathbf{x}, \boldsymbol{\beta}) = h_0(t) \exp\left(\mathbf{x}'\boldsymbol{\beta}\right) = h_0(t) \exp\left(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p\right)$$

where

- $h(t)$ is the hazard rate at time t
- $h_0(t)$ is the baseline hazard
- $\exp\left(\sum_{j=1}^{p} \beta_j X_j\right)$ is the exponentiated linear function of a set of p fixed explanatory variables

under the *binary* case

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

the hazard ratio is

$$\mathrm{HR}(t, 1, 0) = \frac{h_0(t) e^{\beta(1)}}{h_0(t) e^{\beta(0)}} = e^{\beta}$$

Interpretation If the value of the coefficient is $\beta = \log 2 = 0.6931 > 0$ then the female are dying at twice (i.e. $e^{\beta} = 2$) the rate of males (if 1=female, 0=male)

Cumulative hazard function

$$\begin{aligned} H(t, \mathbf{x}, \boldsymbol{\beta}) &= \int_0^t h(u, \mathbf{x}, \boldsymbol{\beta}) du = \int_0^t h_0(u) \exp\left(\mathbf{x}'\boldsymbol{\beta}\right) du \\ &= \exp\left(\mathbf{x}'\boldsymbol{\beta}\right) \int_0^t h_0(u) du = \exp\left(\mathbf{x}'\boldsymbol{\beta}\right) H_0(t) \end{aligned}$$

Survival function

$$S(t, \mathbf{x}, \boldsymbol{\beta}) = \exp[-H(t, \mathbf{x}, \boldsymbol{\beta})] = e^{-\exp\left(\mathbf{x}'\boldsymbol{\beta}\right) H_0(t)} = \left[e^{-H_0(t)}\right]^{\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)} = \left[S_0(t)\right]^{\exp\left(\mathbf{x}'\boldsymbol{\beta}\right)} = \left[S_0(t)\right]^{\exp\left(\beta_1 x_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right)}$$

4.2 Partial likelihood

Triplets (t_i, x_i, c_i)

- t_i is the length of time a subject was observed
- x_i is the exposure variable whose value is determined at baseline and remains unchanged throughout the follow-up of the subject
- c_i indicator variable of censoring with 1=observed, 0=right censoring

then

- when $c_i = 1$ we know that the survival time was exactly t . Thus, the contribution to the likelihood for this triplet is the given by $f(t, \beta, x)$
- when $c_i = 0$ we know that the survival time was at least t. Thus the contribution to the likelihood function of this triplet is given by $S(t, \beta, x)$

Likelihood

$$I(\beta) = \prod_{i=1}^{n} \left[f\left(t_i, \beta, x_i\right)\right]^{c_i} \left[S\left(t_i, \beta, x_i\right)\right]^{1-c_i}$$

Log-likelihood

$$L(\beta) = \sum_{i=1}^{n} \left\{c_i \log f\left(t_i, \beta, x_i\right) + \left(1 - c_i\right) \log S\left(t_i, \beta, x_i\right)\right\}$$

given that $f(t, x, \beta) = h(t, x, \beta) S(t, x, \beta)$

$$L(\beta) = \sum_{i=1}^{n} \left\{c_i \log h_0\left(t_i\right) + \beta c_i x_i + e^{\beta x_i} \log S_0\left(t_i\right)\right\}$$

Partial likelihood

$$I_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\beta x_i}}{\sum_{j \in R(t_i)} e^{\beta x_j}} \right]^{c_i}$$

where the summation in the denominator is over all subjects in the risk set at time t_i and this expression assumes that there are no tied times and it is modified to exclude terms when $c_i = 0$, yielding

$$I_p(\beta) = \prod_{i=1}^{m} \left[\frac{e^{\beta x_i}}{\sum_{j \in R\left(t_{(i)}\right)} e^{\beta x_j}} \right]^{c_i}$$

the log partial likelihood is

$$L_p(\beta) = \sum_{i=1}^{m} \left[\beta x_{(i)} - \log \sum_{j \in R\left(t_{(i)}\right)} e^{\beta x_j} \right]$$

Information

$$\mathbf{I}(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2}$$

and variance

$$\widehat{\text{Var}}[\widehat{\beta}] = \mathbf{I}(\widehat{\beta})^{-1}$$

4.3 Assess the significance of the estimated coefficient

4.3.1 partial likelihood ratio test

The partial likelihood ratio test, denoted G, is calculated as twice the difference between the log partial likelihood of the model containing the exposure variable and the log partial likelihood for the model not containing the exposure. Specifically

$$G = 2\left[L_p(\widehat{\beta}) - L_p(0) \right]$$

where $L_p(0) = -\sum_{i=1}^{m} \log n_i$. Then under $H_0 : \beta = 0,\ G \sim \chi^2_{\text{df}=1}$

4.3.2 Wald test

$$z = \frac{\widehat{\beta}}{\widehat{\text{SE}}(\widehat{\beta})}$$

Then under $H_0 : \beta = 0,\ z^2 \sim \chi^2_{\text{df}=1}$

4.3.3 Score test

$$z^* = \left. \frac{\frac{\partial L_p}{\partial \beta}}{\sqrt{I(\beta)}} \right|_{\beta=0}$$

Then under $H_0 : \beta = 0,\ z^* \sim N(0,1)$

4.4 Interpretation

These equations will be useful later.

$$g(t, x, \beta) = \log\left[h_0(t)\right] + \beta x$$

$$\begin{aligned} g(t, x=a, \beta) - g(t, x=b, \beta) &= \{\log\left[h_0(t)\right] + \beta a\} - \{\log\left[h_0(t)\right] + \beta b\} \\ &= (a-b)\beta \end{aligned}$$

$$\begin{aligned} \text{HR}(t, x=a \text{ vs } x=b, \beta) &= \frac{h(t, a, \beta)}{h(t, b, \beta)} \\ &= \exp[g(t, x=a, \beta) - g(t, x=b, \beta)] \\ &= e^{(a-b)\beta} \end{aligned}$$

4.4.1 Nominal exposure

The idea for interpretation is that the Cox model does not contain intercept term. This is the price you pay for choosing the semiparametric proportional hazards model. *The implication of this in practice is that you cannot, from the regression output of a proportional hazards model, reconstruct group-specific hazard rates. Only hazard ratios can be estimated.* Thus to interpret we have to transfer the *hazard rate* to *hazard ratio* by

$$g(t, x = 1, \beta) - g(t, x = 0, \beta) = (1 - 0)\beta = \beta$$

$$\mathrm{HR}(t, x = 1 \text{ vs } x = 0, \beta) = e^{\beta}$$

95% confidence interval for HR follows

$$\exp[\widehat{\beta} \pm 1.96\widehat{\mathrm{SE}}(\widehat{\beta})]$$

then the interpretation of the HR is

1. the comparison group die at about HR times the rate of reference group, throughout the study period. The confidence interval suggests that ratios as low as *lower CI* or as high as *upper CI* are consistent with the observed data at the 95% level of confidence
   or
2. the death rate among comparison group is (HR-1)*100 percent larger than among reference group throughout the study period, and it could be as little as *(lower CI-1)*100%* smaller as much as *(upper CI-1)*100%* percent larger with 95% confidence

More than two groups
If a nominal scale covariate has more than two levels, denoted in general by K , model the variable using a collection of K-1 design (dummy or indicator) variables.

- *Reference cell coding* is the most frequent method of coding these design variables. With this method, you choose one level of the variable to be the reference level, against which all other levels are compared. *The resulting hazard ratios compare the hazard rate of each group to that of the referent group*.
  To contrast to groups rather than the reference group we can do the following

$$g(t, \text{ four age groups}, \boldsymbol{\beta}) = \log[h_0(t)] + \beta_1 \mathrm{AGECAT}_1 + \beta_2 \mathrm{AGECAT}_2 + \beta_3 \mathrm{AGECAT}_3$$

$$\begin{aligned} & g(t, \text{ age group } 3, \boldsymbol{\beta}) - g(t, \text{ age group } 2, \boldsymbol{\beta}) \\ & = \{\log[h_0(t)] + \beta_1(0) + \beta_2(0) + \beta_3(1)\} - \{\log[h_0(t)] + \beta_1(0) + \beta_2(1) + \beta_3(0)\} \\ & = \beta_3 - \beta_2 \end{aligned}$$

  The estimator of the hazard ratio is

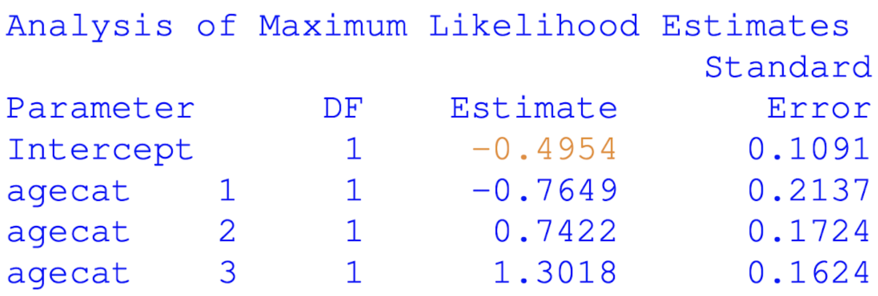
$$\widehat{\mathrm{HR}}(\text{ age group 3vs2}) = e^{\widehat{\beta}_3 - \widehat{\beta}_2}$$

$$\widehat{\mathrm{Var}}\left[\widehat{\beta}_3 - \widehat{\beta}_2\right] = \widehat{\mathrm{Var}}\left[\widehat{\beta}_2\right] + \widehat{\mathrm{Var}}\left[\widehat{\beta}_3\right] - 2\widehat{\mathrm{Cov}}\left(\widehat{\beta}_2, \widehat{\beta}_3\right)$$

- *Deviation from means coding (effect coding)* in effect coding, the comparison group is identified by the symbol -1. E.g. if we have age group 0-3 indicated by AGECAT0-AGECAT3. By setting AGECAT0 as the reference group, we have

$$AGECAT_i = \begin{cases} 1, & \text{if age category is i} \\ 0, & \text{if age categrory is other than i and 0} \\ -1, & \text{if age categroy is 0} \end{cases}$$

  The resulting estimated coefficient for an age group estimates the difference between the log hazard of the group and the *arithmetic mean* of the log hazards of all groups. The exponentiated estimated coefficient provides the ratio of the hazard rate of the particular group to the *geometric mean* of the hazard rates of all groups.



```
Analysis of Maximum Likelihood Estimates
                                     Standard
Parameter        DF     Estimate        Error
Intercept         1      -0.4954       0.1091
agecat      1     1      -0.7649       0.2137
agecat      2     1       0.7422       0.1724
agecat      3     1       1.3018       0.1624
```

For example
The log odds of death for age group 1 is $-0.4954 - 0.7649(1) + 0.7422(0) + 1.3018(0) = -1.2603 = g1$
The log odds of death for age group 2 is $-0.4954 - 0.7649(0) + 0.7422(1) + 1.3018(0) = 0.2468 = g2$
The log odds of death for age group 3 is $-0.4954 - 0.7649(0) + 0.7422(0) + 1.3018(1) = 0.8064 = g3$
The log odds of death for age group 0 is $-0.4954 - 0.7649(-1) + 0.7422(-1) + 1.3018(-1) = -1.7745 = g0$
the intercept coefficient is $-0.4954 = (-1.2603 + 0.2468 + 0.8064 - 1.7745)/4$
$exp(-0.4954) = exp((-1.2603 + 0.2468 + 0.8064 - 1.7745)/4) = \sqrt[4]{g1 \cdot g2 \cdot g3 \cdot g0}$

4.4.2 Continuous exposure

For continuous exposure variables, the estimated coefficient represents the rate of change of a function of the dependent variable **per-unit** change in the explanatory variable. We must decide what a clinically meaningful unit of change in the continuous exposure. If we want to get the estimate of hazard ratio comparing two groups that differ in *c* times the default unit. We do the following

$$\widehat{\mathrm{HR}}(c) = e^{c\widehat{\beta}}$$

with CI

$$\exp[c\widehat{\beta} \pm 1.96|c|\widehat{\mathrm{SE}}(\widehat{\beta})]$$

4.4.3 Multivariable models

Suppose that the primary risk factor, d , has two levels (coded 0 = absent and 1 = present)
crude model

$$g\left(t, d, \theta_1\right) = \log\left[h_0(t)\right] + d\theta_1$$

adjusted model by x

$$g(t, d, x, \boldsymbol{\beta}) = \log\left[h_0(t)\right] + d\beta_1 + x\beta_2$$

The magnitude of the confounding by x is on the scale of the coefficients or difference in the log-hazard. The measure of difference in the two coefficients is the percentage change

$$\Delta\widehat{\beta}\% = 100\left(\frac{\widehat{\theta} - \widehat{\beta}}{\widehat{\beta}}\right)$$

where $\hat{\theta}$:the estimator from the model that does not contain the potential confounder and $\hat{\beta}$: the estimator from the model that does include the potential confounder.

$$\widehat{\theta}_1 \approx \widehat{\beta}_1 + \widehat{\beta}_2\left(\bar{x}_1 - \bar{x}_0\right)$$

where $\bar{x_1}$ is the average value of x among subjects with d = 1, and $\bar{x_0}$ is the average value of x among subjects with d = 0.
the crude estimator will be approximately equal to the adjusted estimator if the difference in the mean of x of the two groups defined by d is 0 or if the coefficient for x is 0. The two estimators will differ if at least one of the two is large or both are moderate in size.
interactive model

$$g(t, d, x, \beta) = \ln\left[h_0(t)\right] + \beta_1 d + \beta_2 x + \beta_3 xd$$

then

$$\widehat{\mathrm{HR}}(t, d = 1 \text{ vs } d = 0, x) = \exp\left(\widehat{\beta}_1 + \widehat{\beta}_3 x\right)$$

$$\widehat{\mathrm{SE}}\left(\widehat{\beta}_1 + \widehat{\beta}_3 x\right) = \sqrt{\widehat{\mathrm{Var}}\left[\widehat{\beta}_1\right] + x^2\widehat{\mathrm{Var}}\left[\widehat{\beta}_1\right] + 2x\widehat{\mathrm{Cov}}\left(\widehat{\beta}_1, \widehat{\beta}_3\right)}$$

thus CI is

$$\exp\left\{\left(\widehat{\beta}_1 + \widehat{\beta}_3 x\right) \pm 1.96\widehat{\mathrm{SE}}\left(\widehat{\beta}_1 + \widehat{\beta}_3 x\right)\right\}$$

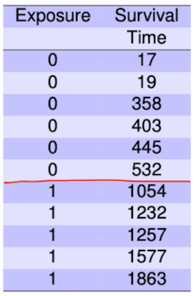
5 Model selection

5.1 Numerical problems that may occur in fitting the model

Complete separation occurs when there is a category or range of an explanatory variable with only one value of the response. This ideal state of affairs is not desirable; variation in the response is necessary to estimate the model parameters. Mathematically, the maximum likelihood estimate for the perfect prediction variable does not exist. The larger the coefficient for the perfect prediction variable, the larger the likelihood. In other words, the coefficient for the perfect prediction variable should be as large as it can be, which would be infinity! In this case, the model won't converge. *We shall be aware of the possibility of complete separation when the estimation of the coefficients and standard errors are very large*.

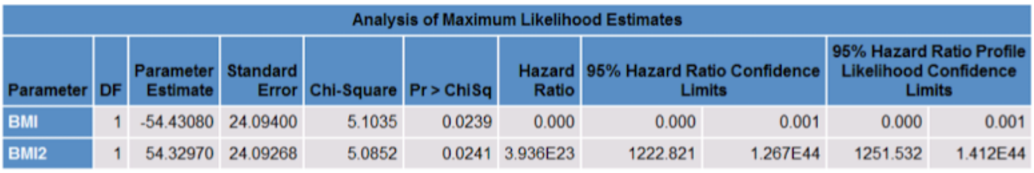
Solutions One approach to handling the problem of complete separation is to employ exact logistic regression, but exact logistic regression is complex and may require prohibitive computational resources. Another option is to use a Bayesian approach using penalized likelihood originally proposed by Firth (1993) and described fully in this setting by Heinze (2002, 2006), which is called **Firth's penalized likelihood**.

Monotone likelihood The table below reflects the problem of non-overlapping survival times, which can also leads to model divergence. It may not produce warnings in SAS, but the estimations can be very large. *Firth's correction* can also be applied in this case.



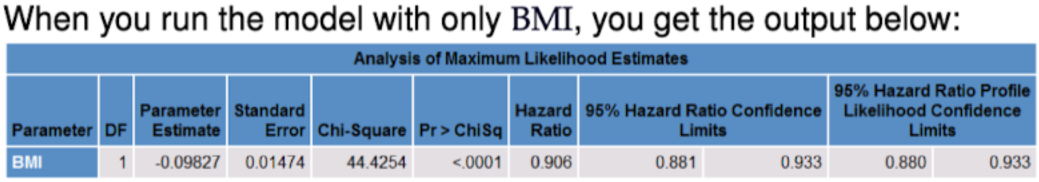
Exposure	Survival Time
0	17
0	19
0	358
0	403
0	445
0	532
1	1054
1	1232
1	1257
1	1577
1	1863

Multicollinearity can cause model divergence or bias the estimation. The results of fitting a proportional hazards model when the relationship between the two covariates is $\mathrm{BMI}_2 = \mathrm{BMI} + u$, where u is the value of a uniformly distributed random variable on the interval (0, 0.01). The correlation between the covariates is effectively 1.0. Tables below shows how the multicollinearity bias the estimation. *Variance inflation factor (VIF)* can be employed to detect multicollinearity.



Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		95% Hazard Ratio Profile Likelihood Confidence Limits	
BMI	1	-54.43080	24.09400	5.1035	0.0239	0.000	0.000	0.001	0.000	0.001
BMI2	1	54.32970	24.09268	5.0852	0.0241	3.936E23	1222.821	1.267E44	1251.532	1.412E44



When you run the model with only BMI, you get the output below:

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits		95% Hazard Ratio Profile Likelihood Confidence Limits	
BMI	1	-0.09827	0.01474	44.4254	<.0001	0.906	0.881	0.933	0.880	0.933

5.2 Model selection

5.3 Stepwise selection

The stepwise selection process consists of a series of alternating forward selection and backward elimination steps. The statistical test used as a criterion is most often the *partial-likelihood ratio test*. However, the score test and Wald test are often used by software packages. SAS, for example, uses the score test. The partial-likelihood ratio test has been shown to have the best statistical testing properties of the three and should be used when there is a choice.

- Step0: Assume that there are p possible variables, denoted $x_j, \quad j = 1, 2, \ldots, p$. Fitting the single variable model p times. For each model, the significance of the variable is derived using the partial likelihood ratio test by comparing with the null model (the model with no variables). The test statistics is

  $$G^{(0)}(j) = -2\left[L^{(0)}(j) - L(0)\right], \quad j = 1, 2, \ldots, p$$

  The significance level of the test is

  $$p^{(0)}(j) = \mathrm{P}\left[\chi^2(1) \geq G^{(0)}(j)\right]$$

The candidate for entry into the model at step 1 is the most significant variable and is denoted by x_{e_1}, where

$$p^{(0)}\left(e_1\right)=\min_j\left[p^{(0)}(j)\right]$$

For variable x_{e_1} to be entered into the model, its p-value *MUST be smaller than some prechosen criterion for significance, denoted by p_E*, otherwise, the process stops.

- Step1: In this step, we fit p-1 two-variable models, which contain x_{e_1}, and they are compared to the model containing only x_{e_1}. The test statistic is

  $$G^{(1)}(j)=-2\left[L^{(1)}(j)-L\left(x_{e_1}\right)\right], \quad j=1,2,\ldots,p, j\neq e_1$$

  The p-value is

  $$p^{(1)}(j)=\mathrm{P}\left[\chi^2(1)\geq G^{(1)}(j)\right]$$

  then the variable selected as the candidate for entry at step 2 is x_{e_2} where

  $$p^{(1)}\left(e_2\right)=\min_{j\neq e_1}\left[p^{(1)}(j)\right]$$

  . The p-value for the selected variable also needs to satisfy $p^{(1)}\left(e_2\right)<p_E$, otherwise, the process stops.

- Step2: this step begins with both x_{e_1} and x_{e_2} in the model. *During this step, two different evaluations occur.* The step begins with a *backward elimination* check for the continued contribution of x_{e_1}. That is, does x_{e_1} still contribute to the model after x_{e_2} has been added? A different significance criterion for this check may be used, denoted p_R. Choose this value such that $p_R>p_E$ to eliminate the possibility of entering and removing the same variable in an endless number of successive steps.

- Step3, if reached, is similar to step 2 in that the elimination process determines whether all variables entered into the model at earlier steps are still significant. The selection process then followed is identical to the selection part of earlier steps. This procedure is followed until the last step, step S.

- StepS: at this step, one of the two things happens: (1) all the variables are in the model and none may be removed, or (2) each variable not in the model has $p^{(S)}(j)>p_E$

Examine the variables picked by stepwise selection Research in linear regression by Bendel and Afifi (1977) and in discriminant analysis by Co-stanza and Afifi (1979) indicates that use of $p_E=0.05$ excludes too many important variables and one should choose a level of significance of 15%. In many applications, it may make sense to use 25%-50% to allow more variables to enter than will ultimately be used and then narrow the field of selected variables using p-value<0.15 to obtain a multivariable model for further analysis.
An unavoidable problem with any stepwise selection procedure is the potential for the inclusion of noise variables and the exclusion of important variables. One must always examine the variables selected and excluded for *basic scientific plausibility.*

5.3.1 Forward selection and backward elimination

Forward selection takes only the forward steps in the stepwise selection mentioned above, and thus once a variable is entered in the model, it is never removed from the model.
Backward elimination takes only the elimination steps in the stepwise selection mentioned above, and thus once a variable is removed from the model, it remains excluded.

5.4 Best-subsets Selection

In summary, best-subsets selection fits all the possible subset of the covariates to the model and select the best one based on certain criterion. There are a number of available criteria, such as R-Square, adjusted R-Square, Mallow's Cp , the PRESS statistic. Given the criterion, the software screens all models containing q variables and reports the variables in the best, say 5, models for q = 1, 2, . . . , p, where p denotes the total number of variables. Of course, the larger the number of candidate variables, the larger the number of possible regression models. For example, if there are 11 candidate variables, then there are $2^{11}=2048$ possible regression models. For example, we can use the *score test*, the larger the score is, the better the model. Like R-Square, the score test statistic tends to increase with the number of covariates in the model. The most frequently used criterion to compare normal errors linear regression models containing different numbers of variables is *Mallow's C.*

Good models are those with small values of Mallow's C. In the context of the proportional hazards model Mallow's C is defined as

$$C = W_q + (p - 2q)$$

where p is the number of variables under consideration and q denotes the number of variables not included in the subset model. The quantity W_q is the value of the multivariable Wald statistic testing that the coefficients for the q variables are simultaneously equal to 0 and is obtained from a fit of the model containing all p variables. The *score test* is used to approximate the value of W_q. Let the score test for the model containing all p variables be denoted S_p and the score test for the model containing a particular set of k (= p-q) variables be denoted S_k. The value of the score test for the exclusion of the q variables from the full p-variable model is approximately $S_q = S_p - S_k$. Because the Wald and score tests are asymptotically equivalent, this suggests that an approximation to Mallow's C for a fitted model containing p-q covariates is

$$C = S_q + (p - 2q)$$

Stepwise vs best-subsets selection: With stepwise selection, you are able to examine only progressively larger models and not ones with the same number of variables. With best-subsets selection, you are able to examine different models with the same number of variables. Note that both approaches, using different criteria, recommended the same set of variables. Complete agreement in variable selection may not always occur.

5.5 Purposeful Selection

- Fit a multivariable model containing all variables significant in the univariable analysis at the 20-25 percent level, as well as any other variables not selected with this criterion but judged to be of clinical importance. (add)
- Use the p-values from the Wald tests of the individual coefficients to identify explanatory variables that might be deleted from the model. Some caution should be taken at this point not to reduce the size of the model by deleting too many seemingly non-significant variables at one time. The p-value of the partial likelihood ratio test should confirm that the deleted explanatory variable is not significant. (delete)
- Assess whether removal of the explanatory variable has produced an important change in the coefficients of the variables remaining in the model. In general, use a value of about 20% as an indicator of an important change in a coefficient. If the variable excluded is an important confounder, it should be added back into the model. This process continues until no explanatory variables can be deleted from the model. (assess)
- Add to the model, one at a time, all variables excluded from the initial multivariable model to confirm that they are neither statistically significant nor an important confounder. It is possible that a variable that had a univariable test p-value which exceeded 0.8 became highly significant when added to the multivariable model obtained at step 3. Refer to the model at the conclusion of this step as the *preliminary main effects model*.
- Examine the scale of the continuous covariates in the preliminary main effects model. A number of techniques are available, all of which are designed to determine whether the data support the hypothesis that the effect of the explanatory variable is *linear* in the log hazard and, if not, what transformation of the variable is linear in the log hazard. Refer to the model at the end of step 5 as the *main effects model*. (check assumption)
- Determine whether interactions are needed in the model. Special considerations may dictate that a particular interaction term or terms be included in the model, regardless of the statistical significance of the coefficient(s). However, in most settings, there will be insufficient scientific theory to justify automatic inclusion of interactions. All interactions significant at the 5%, perhaps as low as 1% in some settings, level are added jointly to the main effects model. Wald statistic p-values are used as a guide to selecting interactions that may be eliminated from the model, but significance should be checked by the partial likelihood ratio test. Often when an interaction term enters a model, the coefficient of one of its component main effects may have a non-significant Wald statistic. All main effects of significant interactions should remain in the model because estimates of effect require both main effect and interaction coefficients. Refer to the model at the conclusion of step 6 as the *preliminary model*. It does not become the final model until it is thoroughly evaluated. (add interaction)
- Evaluate the model: check for adherence to key model assumptions using casewise diagnostic statistics to check for influential observations and testing for overall goodness-of-fit. This step is mandatory for any model building strategy, not just purposeful selection. (check assumptions)

6 Model Assessment

Like model development, model assessment involves a number of steps.

- examine and test the proportional hazards assumption
- evaluating subject-specific diagnostic statistics that measure leverage and influence on the fit of the proportional hazards model
- compute summary measures of goodness-of-fit

6.1 Residuals

Unfortunately, when fitting a proportional hazards model to censored survival data, there is no obvious analog to the usual ‘observed minus predicted’ residual used with other regression models. SAS and most software packages provide access to these residuals:

- **Schoenfeld and scaled Schoenfeld residuals**: calculated for a given subject with respect to an explanatory variable. It is the difference between the actual value of an explanatory variable for a subject and the expected value of the explanatory variable in the risk set.
- **martingale residuals**: calculated for a subject at time t. It is the difference between actual and expected number of events to time t.
- **score and scaled score residuals**: calculated for a subject with respect to an explanatory variable. It is a weighted difference between the value of the explanatory variable for a subject and the average value of the explanatory variable in the risk set.

6.1.1 Schoenfeld residuals

[Schoenfeld residuals are used to assess the proportional hazards assumption]
Assume that there are p explanatory variables and that the n independent observations of time, explanatory variables and censoring indicator are denoted by the triplet $(t_i, x_i, c_i), i = 1, 2, \cdots, n,$ where $c_i = 1$ for uncensored observations and $c_i = 0$ otherwise.
Schoenfeld residuals are based on the individual contributions to the derivative of the log partial likelihood. Recall that

$$\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} c_i \left[x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{\mathbf{x}'_j \beta}}{\sum_{j \in R(t_j)} e^{\mathbf{x}'_j \beta}} \right] = \sum_{i=1}^{n} c_i \left[x_{ik} - \bar{x}_{w_i k} \right]$$

is the derivative for the kth covariate, where $\bar{x}_{w_i k} = \sum_{j \in R(t_i)} x_{jk} e^{\mathbf{x}'_j \boldsymbol{\beta}} / \sum_{j \in R(t_i)} e^{\mathbf{x}'_j \boldsymbol{\beta}}$. The estimator of the Schoenfeld residual for the ith subject on the kth covariate is obtained by substituting the partial likelihood estimator of the coefficient, β, and is

$$\widehat{r}_{ik} = c_i \left[x_{ik} - \widehat{\bar{x}}_{w_i k} \right]$$

where $\widehat{\bar{x}}_{w_i k} = \sum_{j \in R(t_j)} x_{jk} e^{\mathbf{x}'_j \widehat{\boldsymbol{\beta}}} / \sum_{j \in R(t_i)} e^{\mathbf{x}'_j \widehat{\boldsymbol{\beta}}}$ is the estimator of the risk set conditional mean of the explanatory variable.
Schoenfeld residuals are based on the individual contributions to the derivative of the log partial likelihood. Because the partial likelihood estimator of the coefficient, $\hat{\beta}$, is the solution to the equations obtained by setting the partial derivatives equal to 0, *the sum of the Schoenfeld residuals is 0*. The Schoenfeld residuals are *equal to 0 for all censored subjects* and thus contain no information about the fit of the model.
Scaled Schoenfeld residuals Denote the vector of p Schoenfeld residuals for the ith subject as

$$\widehat{\mathbf{r}}_i = (\widehat{r}_{i1}, \widehat{r}_{i2}, \ldots, \widehat{r}_{ip})$$

Let the estimator of the p × p covariance matrix of the vector of residuals for the ith subject be denoted by $\widehat{\mathrm{Var}}\left[\widehat{\mathbf{r}}_i\right]$, and the estimator is missing if $c_i = 0$. The vector of scaled Schoenfeld residuals is the product of the inverse of the covariance matrix times the vector of residuals, namely

$$\widehat{\mathbf{r}}_i^* = \left(\widehat{\mathrm{Var}}\left[\widehat{\mathbf{r}}_i\right] \right)^{-1} \widehat{\mathbf{r}}_i$$

, where $\widehat{\mathrm{Var}}\left[\widehat{\mathbf{r}}_i\right]_{kk} = \sum_{j \in R(t_i)} \widehat{w}_{ij} \left(x_{jk} - \widehat{\bar{x}}_{w_i k} \right)^2$ and $\widehat{\mathrm{Var}}\left[\widehat{\mathbf{r}}_i\right]_{kl} = \sum_{j \in R(t_i)} \widehat{w}_{ij} \left(x_{jk} - \widehat{\bar{x}}_{w_i k} \right) \left(x_{jl} - \widehat{\bar{x}}_{w_i} \right)$, with $\widehat{w}_{ij} = \frac{e^{\mathbf{x}'_j \widehat{\boldsymbol{\beta}}}}{\sum_{l \in R(t_i)} e^{\mathbf{x}'_1, \widehat{\boldsymbol{\beta}}}}$. Given that $\left(\widehat{\mathrm{Var}}\left[\widehat{\mathbf{r}}_i\right] \right)^{-1} \approx m \widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}]$, where m is the observed number of uncensored survival times. We have

$$\widehat{\mathbf{r}}_i^* = \left(\widehat{\mathrm{Var}}\left[\widehat{r}_i\right] \right)^{-1} \widehat{\mathbf{r}}_i \approx m \widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}] \widehat{\mathbf{r}}_i$$

6.1.2 Martingale residuals (Cox-Snell residual)

[Cumulative martingale residuals can be used to assess PH assumption, and plot likelihood displacement against martingale residual can be used to find influential point]

Counting process representation

The counting process representation of the proportional hazards model is a linear-like model that counts whether the event occurs (e.g., the subject dies) at time t . The basic model is

$$N(t) = \Lambda(t, \mathbf{x}, \boldsymbol{\beta}) + M(t)$$

where

- $N(t)$ is the count that represents the observed part of the model and is defined to be equal to 0 until the exact time the event occurs and is equal to 1 thereafter. If the subject does not die during the entire follow-up, then $N(t) = 0$
- $\Lambda(t, \mathbf{x}, \boldsymbol{\beta}) = H(t, \mathbf{x}, \boldsymbol{\beta}) = H_0(t)e^{\mathbf{x}'\boldsymbol{\beta}}$. At the end of follow-up you will find the maximum value for the systematic component, regardless of whether the event occurred.
- $M(t) = N(t) - \Lambda(t, \mathbf{x}, \boldsymbol{\beta})$ is a martingale and plays the role of the error component. Under correct model, $E\left(M(t)\right) = 0$. In theory, it has a value at each time t , but the most useful choice of time at which to compute the residual is *the end of follow-up*. Given that $N(t_i) = c_i$, $\widehat{M}_i = c_i - \widehat{H}\left(t_i, \mathbf{x}, \widehat{\boldsymbol{\beta}}\right)$

6.1.3 Score residual

[Score residual is used to measure the leverage and influence, respectively, of particular subjects]
$\frac{\partial L_p(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{n} L_{ik}$, the score residual is

$$L_{ik} = \sum_{j=1}^{n} \left(x_{ik} - \bar{x}_{w_j k}\right) dM_i\left(t_j\right)$$

where $dM_i\left(t_j\right) = dN_i\left(t_j\right) - Y_i\left(t_j\right) e^{\mathbf{x}_i'\beta} h_0\left(t_j\right)$,
and $Y_i\left(t_j\right) = \begin{cases} 1 & \text{if } t_i \geq t_j \\ 0 & \text{if } t_i < t_j \end{cases}$, $dN_i(t_i) = \begin{cases} 1 & \text{at the actual observed survival time} \\ 0 & \text{otherwise} \end{cases}$
An expanded computational formula yields the estimator

$$\widehat{L}_{ik} = c_i\left(x_{ik} - \widehat{x}_{w_i k}\right) - x_{ik}\widehat{H}\left(t_i, \mathbf{x}, \widehat{\boldsymbol{\beta}}\right) + e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}} \sum_{t_j \leq t_i} \widehat{\bar{x}}_{w_j k} \frac{c_j}{\sum_{l \in R_j} e^{\mathbf{x}_l'\hat{\beta}}}$$

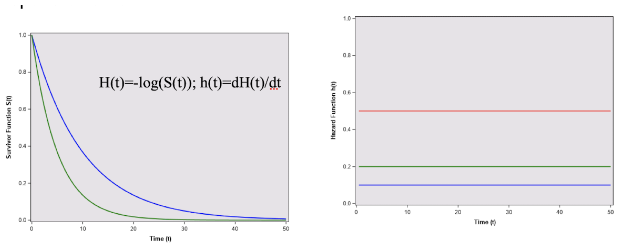
Scaled score residual
let $\widehat{\mathbf{L}}_{ik} = \left(\hat{L}_{i1}, \widehat{L}_{i2}, \ldots, \widehat{L}_{ip}\right)$ then scaled score residual is

$$\widehat{\mathbf{L}}_i^{*'} = \widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}]\widehat{\mathbf{L}}_i$$

Note:

- A positive Schoenfeld residual indicates that the value of the explanatory variable is higher than expected at that death time.
- If the exposure is binary with values 1 or 0, then the Schoenfeld residuals for this variable will be between –1 and 1.
- The range of martingale residuals is from $-\infty$ to 1.
- If the martingale residual for a subject is close to 1, this indicates the subject died too soon. On the other hand, if the martingale residual for a subject is negative and its absolute value is large, this indicates the subject lived too long.
- The mean of the martingale residuals is 0.
- A plot of the martingale residuals vs continuous explanatory variables may suggest continuous variables that are not properly fit.
- Statistics such as DFBETA and LD are typically used for smaller data sets because the influence of any single observation diminishes as the sample size increases.

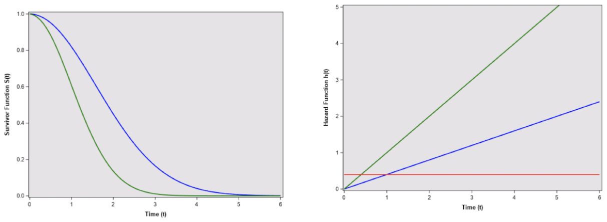
6.2 Proportional Hazards Assumption



$$S_1(t) = \exp(-0.1t) \quad \text{(blue graph)}$$
$$S_2(t) = \exp(-0.2t) \quad \text{(green graph)}$$

$$h_1(t) = 0.1 \quad \text{(blue graph)}$$
$$h_2(t) = 0.2 \quad \text{(green graph)}$$
$$\text{HR}(t) = \frac{h_1(t)}{h_2(t)} = \frac{1}{2} \quad \text{(red graph)}$$

constant hazards, constant hazard ratio



$$S_1(t) = \exp(-0.2t^2) \quad \text{(blue graph)}$$
$$S_2(t) = \exp(-0.5t^2) \quad \text{(green graph)}$$

$$h_1(t) = 0.4t \quad \text{(blue graph)}$$
$$h_2(t) = 1.0t \quad \text{(green graph)}$$
$$\text{HR}(t) = \frac{h_1(t)}{h_2(t)} = 0.4 \quad \text{(red graph)}$$

variable hazards, constant hazard ratio

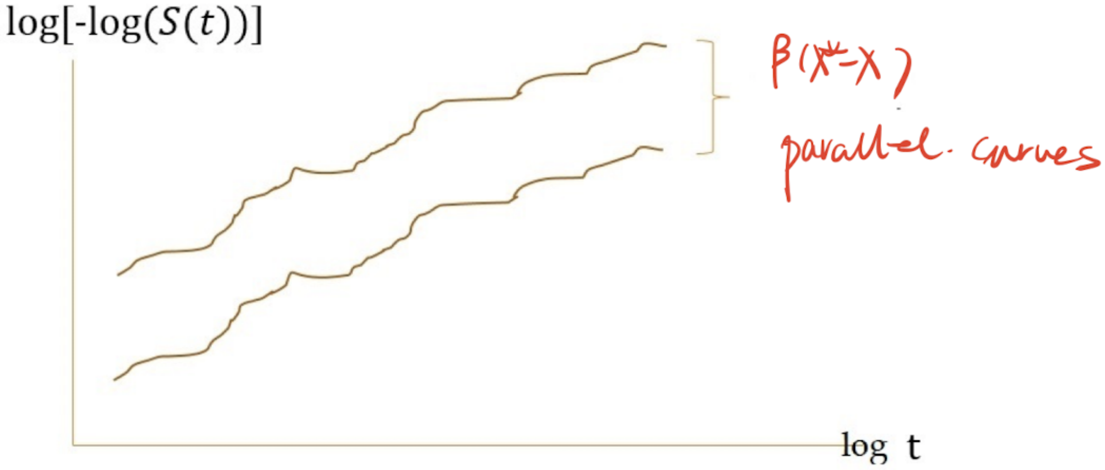
6.2.1 log-log plot

Model with single predictor

$$\begin{aligned} S(t, \mathbf{x}, \boldsymbol{\beta})) &= [S_0(t)]^{\exp(\mathbf{x}'\boldsymbol{\beta})} \\ \log(-\log S(t, \mathbf{x}, \boldsymbol{\beta}))) &= \mathbf{x}'\boldsymbol{\beta} + \log(-\log S_0(t)) \end{aligned}$$

Now consider comparing two groups of people with different values of x , denoted by $x^\star$ and x. If you subtract the log-log survival function for the second group from that for the first group you get

$$\begin{aligned} &\log(-\log S(t, x^*, \beta)) - \log(-\log S(t, x, \beta)) \\ =&[\beta x^* + \log(-\log S_0(t))] - [\beta x + \log(-\log S_0(t))] \\ =&\beta x^* - \beta x \\ =&\beta(x^* - x) \text{ independent of time} \end{aligned}$$



parallel curve indicates PH assumption satisfied

with only one covariate, we divide the data by the covariate and fit the KM model, in this way, we get empirical plots of log log survival curves, which means unadjusted curves.

Note: we need to categorize continuous variables for log-log plots. Categorization into two groups may give a different graphical picture from a categorization into three groups. The recommendation is that a small number of categories be chosen, that the choice of categories be as meaningful as possible and provide reasonable balance of category size.

Model with several predictors

To evaluate the proportional hazards assumption for several variables simultaneously. The strategy is to assess the assumption for one predictor adjusted for the other predictors that are assumed to satisfy the proportional hazards assumption.

1. stratify the data by the levels of the target covarite
2. fit a proportional hazards model containing the explanatory variable that assumed to satisfy the PH assumption in each stratum, and then
3. obtain adjusted survival probabilities using the mean of the explanatory in each stratum in the estimated survival curve formula for each stratum.

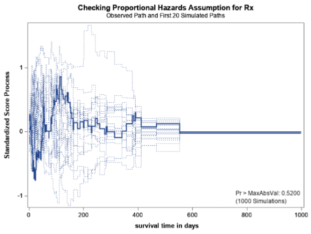
6.2.2 Using Schoenfeld residuals

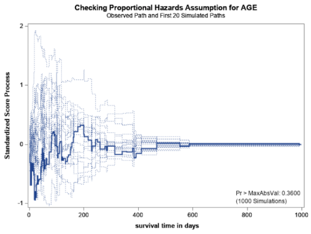
Suppose subject i has an event at time t. Then her Schoenfeld residual for Var is her observed value of Var minus a weighted average of the Var for the other subjects still at risk at time t. The weights are each subject's hazard. *The idea behind the statistical test is that if the proportional hazards assumption holds for a particular variable, then the Schoenfeld residuals for that covariate will not be related to survival time.*

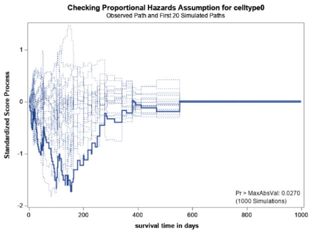
1. Run a Cox proportional hazards model and obtain Schoenfeld residuals for each explanatory variable
2. Create a variable that ranks the order of failures. The subject who has the earliest event gets a value of 1, the next gets a value of 2, and so on
3. Test the correlation between the variables created in the first and second steps. The null hypothesis is that the correlation between the Schoenfeld residuals and ranked failure time is 0. Thus, a nonzero slope is an indication of a violation of the proportional hazards assumption.

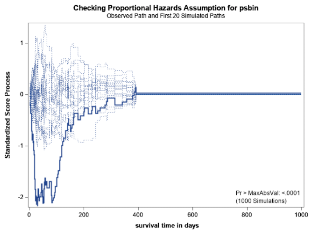
6.2.3 Cumulative martingale residuals

The solid lines represent the observed cumulative residuals, while dotted lines represent 20 simulated sets of residuals expected under the null hypothesis that the proportional hazards assumption is satisfied. A solid line that falls significantly outside the boundaries set up collectively by the dotted lines suggest that the proportional hazards assumption is not satisfied.









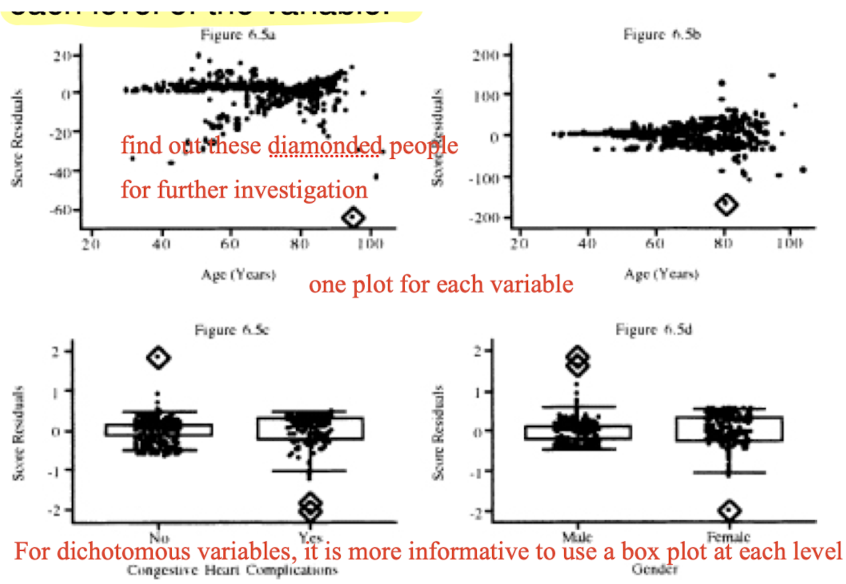
6.3 Influential & Poorly-fit Observations

6.3.1 Leverage

Leverage is a diagnostic statistic that measures how unusual the values of the explanatory variables are for an individual. In linear and logistic regression, leverage is calculated as the distance of the value of the explanatory variable for a subject to the overall mean of the explanatory variables. Leverage is not quite so easily defined nor does it have the same nice properties in proportional hazards regression.

6.3.2 Score residuals

[useful for identifying subjects with high leverage or who influence the value of a single coefficient.]
The score residual for the ith subject on the kth covariate is a weighted average of the distance of the value, x_{ik} , to the risk set means, $\bar{x}_{w_j k}$, where the weights are the change in the martingale residual, $dM_i(t_j)$. The net effect is that, for continuous explanatory variables, the score residuals have the linear regression leverage property that the farther the value is from the mean, the larger the score residual is.



find out these diamonded people for further investigation

one plot for each variable

For dichotomous variables, it is more informative to use a box plot at each level

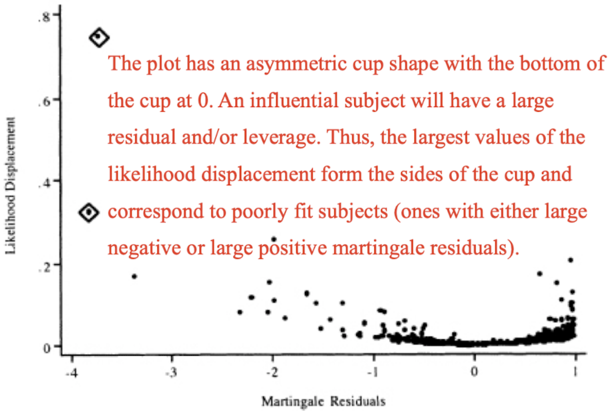
Scaled score residuals The purpose of Cook's distance is to obtain an easily computed statistic that approximates the change in the value of the estimated coefficients if a subject is deleted from the data. This is denoted as $\Delta\widehat{\beta}_{ki} = \widehat{\beta}_k - \widehat{\beta}_{k(-i)} \approx (\widehat{\mathrm{Var}}[\widehat{\beta}]\widehat{L}_i)_k$ = scale score residual, where k means the k-th subscript. Then we can plot the scaled score residual against the value of the covariate, like what we did in the plot above.

6.3.3 Likelihood displacement statistics

[provides useful information for assessing influence on the vector of coefficients.]
The overall measure of influence is

$$\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}\right)' (\widehat{\mathrm{Var}}[\widehat{\boldsymbol{\beta}}])^{-1} \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(-i)}\right) \approx Id_i = \widehat{L}_i'(\widehat{\mathrm{Var}}[\widehat{\beta}])\hat{L}_i \approx 2\left[L_p(\widehat{\boldsymbol{\beta}}) - L_p\left(\widehat{\boldsymbol{\beta}}_{(-i)}\right)\right]$$

This statistic has been shown by Pettitt and Bin Daud (1989) to be an approximation to the amount of change in the log partial likelihood when the i th subject is deleted. In this context, the statistic is called the **likelihood displacement statistic**. It makes the most sense to plot ld$_i$ *versus the martingale residuals*.



The plot has an asymmetric cup shape with the bottom of the cup at 0. An influential subject will have a large residual and/or leverage. Thus, the largest values of the likelihood displacement form the sides of the cup and correspond to poorly fit subjects (ones with either large negative or large positive martingale residuals).

After identify explicitly the subjects with the extreme values, refit the model deleting these subjects, and calculate the change in the individual coefficients. The final decision will depend on the observed percentage change in the coefficients that results from deleting the subject's data and, more importantly, the *clinical plausibility* of that subject's data.

6.4 Overall goodness-of-fit

Quintile	Point	Count	Observed	Expected	z	p
1	1.37	100	6	8.17	-0.76	0.448
2	2.17	100	15	22.47	-1.58	0.115
3	2.80	100	41	37.55	0.56	0.574
4	3.52	100	66	53.39	1.73	0.084
5	5.66	99	86	92.41	-0.67	0.505

The table above presents the observed and estimated expected numbers of events, the z -score and two-tailed p-value within each quintile of risk. The results in the observed and expected columns are obtained as follows:

1. Following the fit of the model, save the martingale residuals and risk score.
2. Sort the risk score and create a grouping variable with values 1 5 based on the quintiles of the risk score.
3. Calculate the observed number of events in each quintile by summing the censoring variable over the subjects in each quintile.
4. Create the model cumulative hazard by subtracting the martingale residual from the follow up status (censoring) variable.
5. Calculate the expected number of events in each quintile by summing the cumulative hazard over the subjects in each quintile.

7 Extensions of the Cox model

7.1 Stratified Cox Model

7.1.1 No-interaction assumption

The population under study can consist of a number of subpopulations, each of which has its own baseline hazard function. The stratified Cox model is a modification of the Cox proportional (PH) model that allows for stratification of a predictor that does not satisfy the proportional hazards assumption. Explanatory variables that are assumed to satisfy the proportional hazards assumption are included in the model, whereas the variable being stratified is not included. Under the stratified model, the hazard function in the s-th stratum is expressed as

$$h_s(t, \mathbf{x}, \boldsymbol{\beta}) = h_{s0}(t)e^{\mathbf{x}'\boldsymbol{\beta}}$$

where there are $s = 1, 2, \cdots, S$ strata.
Partial likelihood The contribution to the partial likelihood for the s-th stratum is

$$I_{sp}(\boldsymbol{\beta}) = \prod_{i=1}^{n_s}\left[\frac{e^{\mathbf{x}'_{si}\boldsymbol{\beta}}}{\sum_{j\in R(t_{si})} e^{\mathbf{x}'_{sj}\boldsymbol{\beta}}}\right]^{c_{si}}, \qquad I_{Sp}(\boldsymbol{\beta}) = \prod_{s=1}^{s} I_{sp}(\boldsymbol{\beta})$$

where $n_s =$ the number of obs in the s-th stratum, $t_{si} =$ i-th observed value of time in the s-th stratum, $c_{si} =$ value of the 0/1 censoring variable associated with t_{si}, $R(t_{si}) =$ the subject in the stratum s in the risk set at time t_{si}, $X_{st} =$ vector of p explanatory variables, and $I_{Sp} =$ full likelihood.
Note:

1. Watch for small numbers within any stratum. Small numbers within any stratum will result in an estimated baseline survival function with greater variance than the estimates from strata with more data.
2. Since the strata variable is not included in the model, it is not possible to obtain HR for it.

No-interaction assumption: For different strata, the model provides the same coefficients for β.
Multiple variables violate the PH assumption: Assume that k variables do not satisfy the proportional hazards assumption $Z_1, Z_2, \ldots, Z_k$. To perform the stratified Cox procedure, define a new variable, call it $Z^\star$, from the Z's by forming combinations of categories of Z's.
An example is as follows: Stratify on 2 variables: cell type and performance status. Cell type has 4 levels: large, adeno, small, squamous. The range of performance status is between 0 and 100. Create a new binary variable PSBIN, defined as PSBIN $= \begin{cases} 1 & \text{if performance } \geq 60 \\ 0 & \text{otherwise} \end{cases}$, $Z^\star$ has 4*2=8 categories and thus create 7 dummy variables when writing down the model.

$Z_1^* =$ large cell
$Z_2^* =$ adeno cell
$Z_3^* =$ small cell
$Z_4^* =$ PSbin
Squamous cell is the reference category for cell type.

,

$$Z_5^* = Z_1^* \times Z_4^*$$
$$Z_6^* = Z_2^* \times Z_4^*$$
$$Z_7^* = Z_3^* \times Z_4^*$$

PSBIN =0 is the reference category for PSBIN.

$$\begin{aligned} h_s(t,\mathbf{X}) =& h_{s0}(t)\exp[\beta_1\text{RX} + \beta_2\text{AGE} + \beta_{11}\text{RX}\times Z_1^* + \beta_{12}\text{RX}\times Z_2^* + \beta_{13}\text{RX}\times Z_3^* + \\ & \beta_{14}\text{RX}\times Z_4^* + \beta_{15}\text{RX}\times Z_5^* + \beta_{16}\text{RX}\times Z_6^* + \beta_{17}\text{RX}\times Z_7^* + \\ & \beta_{21}\text{AGE}\times Z_1^* + \beta_{22}\text{AGE}\times Z_2^* + \beta_{23}\text{AGE}\times Z_3^* + \beta_{24}\text{AGE}\times Z_4^* + \\ & \beta_{25}\text{AGE}\times Z_5^* + \beta_{26}\text{AGE}\times Z_6^* + \beta_{27}\text{AGE}\times Z_7^*] \end{aligned}$$

Model Term	Parameter(s) Estimated
RX	β_1
AGE	β_2
RX*CELLTYPE	$\beta_{11}, \beta_{12}, \beta_{13}$
RX*PSBIN	β_{14}
RX*CELLTYPE*PSBIN	$\beta_{15}, \beta_{16}, \beta_{17}$
AGE*CELLTYPE	$\beta_{21}, \beta_{22}, \beta_{23}$
AGE*PSBIN	β_{24}
AGE*CELLTYPE*PSBIN	$\beta_{25}, \beta_{26}, \beta_{27}$

stratum 1: $Z_1^* = Z_2^* = Z_3^* = Z_4^* = 0$. This stratum is defined by the combination of squamous cell type and a binary performance status value of 0. In this case, all the product terms are equal to 0, and the regression model contains only the main effect terms RX and AGE. $h_1(t,\mathbf{X}) = h_{10}(t)\exp[\beta_1\text{RX} + \beta_2\text{AGE}]$

$$\vdots$$

stratum 6: $Z_1^* = 1, Z_2^* = Z_3^* = 0, Z_4^* = 1,$ $\begin{aligned} h_6(t,\mathbf{X}) =& h_{60}(t)\exp[\beta_1\text{RX} + \beta_2\text{AGE} + \beta_{11}\text{RX} + \beta_{14}\text{RX} + \beta_{15}\text{RX} + \beta_{21}\text{AGE} + \\ & \beta_{24}\text{AGE} + \beta_{25}\text{AGE}] \end{aligned}$.

There are in total 8 strata.

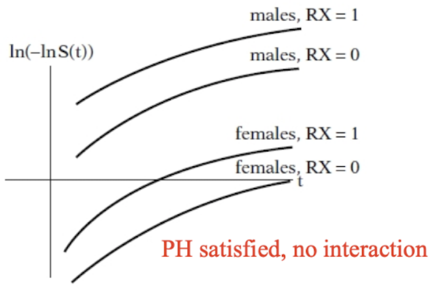
7.1.2 Interaction model

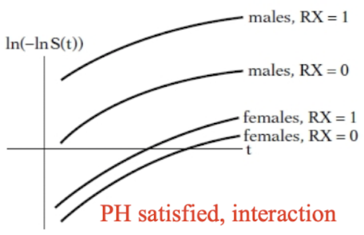
$$\begin{aligned} &\text{No-Interaction model:} && h_s(t,\mathbf{X}) = h_{s0}(t)\exp[\beta_1 RX + \beta_2\log WBC], s = 1,2 \\ &\text{Interaction model:} && h_s(t,\mathbf{X}) = h_{s0}(t)\exp[\beta_{1s} RX + \beta_{2s}\log WBC], s = 1,2 \end{aligned}$$

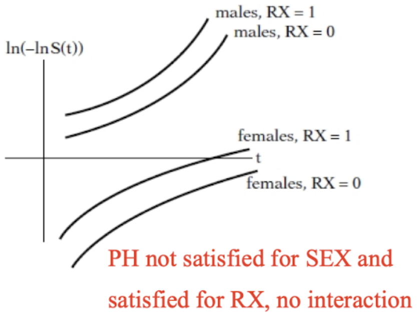
Note that each variable in the interaction model has a coefficient for females that is different from males, as indicated by the subscript s in $\beta s1$ and β_{s2}.
Interaction model can be written in the following two ways

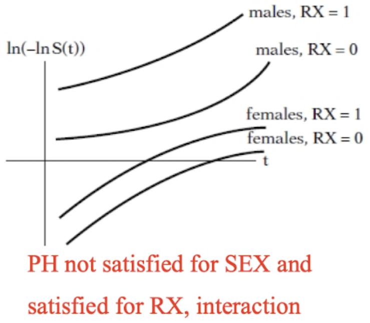
$$\begin{aligned} &\text{Interaction model A:} && h_s(t,\mathbf{X}) = h_{s0}(t)\exp[\beta_{1s}\text{RX} + \beta_{2s}\log\text{WBC}], s = 1,2 \\ &\text{Interaction model B:} && h_{s0}(t)\exp[\beta_1^* RX + \beta_2^*\log WBC + \beta_3^*\text{RX}\times\text{SEX} + \beta_4^*\log\text{WBC}\times\text{SEX}], s = 1,2 \end{aligned}$$

Note the SEX main effect is noticeably missing. We can perform **partial LR test** to decide whether to include interaction term in the model.









7.2 Extended Cox Model

There may be situations where one or more of the variables are measured during the period of follow up and their values change. Note: 1. one should give serious consideration to the nature of any time-varying variable before including it in the model. The value of any time-varying variable must depend only on study time, not on calendar time; 2. another concern is the potential to overfit a model when using time-varying variables. There are two types of time-varying variables

- **internal time-varying variable** is one whose value is subject-specific and requires that the subject be under periodic observation
- **external time-varying variable** is one whose value at a particular time does not require subjects to be under direct observation. Typically, these covariates are study or environmental factors that apply to all subjects under observation. Eg: subject's age; time itself.

However, neither AGE nor analysis TIME were modeled as a time-varying covariate in previous chapters because AGE and analysis TIME advance in parallel if linear functions of AGE or analysis TIME are considered. If you were to include AGE as a time-varying covariate, the estimate of its effect would not change because any effects relating to the advancement of AGE would be absorbed into the baseline hazard function.
The form of the extended Cox model can be expressed as

$$h(t, \mathbf{x}(t), \boldsymbol{\beta}) = h_0(t) \exp\left[\mathbf{x}'(t)\boldsymbol{\beta}\right]$$

and the generalized partial likelihood is

$$I_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i'\left(t_{(i)}\right)\beta}}{\sum_{l\in R\left(t_{(i)}\right)} e^{\mathbf{x}_l'\left(t_{(l)}\right)\beta}} \right]^{c_i}$$

A special form of the extend model is

$$h(t, \mathbf{x}(t), \boldsymbol{\beta}) = h_0(t) \exp\left[\mathbf{x}'(t)\boldsymbol{\beta}\right] = h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} \delta_k x_k g_k(t)\right]$$

with different choice of $g_k(t)$

- **Choice 1**: $g_k(t) = 0$. This is another way of stating the original proportional hazards model.
- **Choice 2**: $g_k(t) = t$. Then we get $h(t, \mathbf{x}(t), \boldsymbol{\beta}) = h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} \delta_k \left(t \cdot x_k\right)\right]$.
- **Choice 3**: $g_k(t) = \begin{cases} t & \text{for } k = L \\ 0 & \text{for other } k \end{cases}$. This allows you to focus on a particular time-varying variable, X_L.
- **Choice 4**: $g_k(t) = \log t$. Then $h(t, \mathbf{x}(t), \boldsymbol{\beta}) = h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k x_k + \sum_{k=1}^{p} \delta_k (\log t) x_k\right]$
- **Choice 5**$\star$ (covered in more detail below): Let $g_k(t)$ be a *heaviside function* of the form $g_k(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$

By comparing with the time-invariant model $h(t, \mathbf{x}, \boldsymbol{\beta}) = h_0(t) \exp\left[\mathbf{x}'\boldsymbol{\beta}\right] = h_0(t) \exp\left[\sum_{k=1}^{p} \beta_k x_k\right]$ using *LR test*, we can test the PH assumption of the variables. The null hypothesis is $H_0 : \delta_1 = \delta_2 = \cdots = \delta_p = 0$
Note:

- An important assumption is that the time-varying covariate effect, as measured by its coefficient, does not depend on time.
- In most settings where time-varying variables are included, the model will also contain time-invariant variables. For example, $h(t, \mathbf{x}(t), \boldsymbol{\beta}) = h_0(t) \exp\left[\beta_1 x_1(t) + \beta_2 x_2\right]$
- A serious bias can occur when you include a time-varying variable in the model, and the effect of a treatment on the outcome is mediated by this time-varying variable.

7.2.1 Heaviside function

- **single heaviside function**:

$$h(t, \mathbf{x}(t)) = h_0(t)\exp[\beta E + \delta E g(t)], \quad \text{where } g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

  then $\text{HR} = \begin{cases} e^{\beta+\delta} & t \geq t_0 \\ e^{\beta} & t < t_0 \end{cases}$

- **two heaviside functions**:

$$h(t, \mathbf{x}(t)) = h_0(t)\exp\left[\delta_1 E g_1(t) + \delta_2 E g_2(t)\right], \quad \text{where } g_1(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases} \quad g_2(t) = \begin{cases} 0 & \text{if } t \geq t_0 \\ 1 & \text{if } t < t_0 \end{cases}$$

  then $\text{HR} = \begin{cases} e^{\delta_1} & t \geq t_0 \\ e^{\delta_2} & t < t_0 \end{cases}$

  Note:

  1. The two heaviside model does not contain a main effect term for exposure
  2. The single heaviside model allows you to **test** whether the two HRs, $HR = e^{\beta+\delta}$ and $HR = e^{\beta}$, are the same, but it does not directly give you a p-value or a confidence interval for the hazard ratio when $t \geq t_0$.
  3. The two heaviside model does not allow you to test whether the two HRs are the same, but it does give you a p-value and a confidence interval for the hazard ratio when $t < t_0$ or when $t \geq t_0$.

- **Multiple heaviside functions**:

$$h(t, \mathbf{x}(t)) = h_0(t)\exp\left[\delta_1 E g_1(t) + \delta_2 E g_2(t) + \delta_3 E g_3(t) + \delta_4 E g_4(t)\right]$$

  where $g_1(t) = \begin{cases} 1 & \text{if } 0 \leq t < 0.5 \text{ year} \\ 0 & \text{otherwise} \end{cases}$ $g_3(t)$ $\begin{cases} 1 & \text{if 1.0 year } \leq t < 1.5 \text{ years} \\ 0 & \text{otherwise} \end{cases}$

  $g_2(t) = \begin{cases} 1 & \text{if 0.5 year } \leq t < 1.0 \text{ year} \\ 0 & \text{otherwise} \end{cases}$ $g_4(t) = \begin{cases} 1 & \text{if } t \geq 1.5 \text{ years} \\ 0 & \text{otherwise} \end{cases}$ then

	Time Interval	Hazard Ratio
1	$0 \leq t < 0.5$	$\text{HR}_1 = e^{\delta_1}$
2	$0.5 \leq t < 1.0$	$\text{HR}_2 = e^{\delta_2}$
3	$1.0 \leq t < 1.5$	$\text{HR}_3 = e^{\delta_3}$
4	$t \geq 1.5$	$\text{HR}_4 = e^{\delta_4}$

Summary: If we wish to separate the data into N separate time intervals, and for each interval, you wish to obtain a different HR estimate. You can obtain N different hazard ratios using

1. an extended Cox model containing a *main effect* of exposure and N-1 heaviside functions in the model as products with exposure, or
2. an extended Cox model containing no main effect exposure term, but with product terms involving exposure with N heaviside functions
3. there are actually other approaches, and let's illustrate it in the following example.

Example: if the CLINIC variable does not satisfy the PH assumption then we can do

1. **One heaviside function**:

$$h(t, \mathbf{x}(t)) = h_0(t)\exp\left[\beta_1 \text{ CLINIC-NEW } + \beta_2 \text{ PRISON } + \beta_3 \text{ DOSE } + \delta \text{ CLINIC-NEW } g(t)\right]$$

   where $g(t) = \begin{cases} 1 & \text{if } t \geq 365 \text{ days} \\ 0 & \text{if } t < 365 \text{ days} \end{cases}$ and CLINIC-NEW $= \begin{cases} 1 & \text{if clinic 1} \\ 0 & \text{if clinic 2} \end{cases}$

2. **Two heaviside functions**:

$$h(t, \mathbf{x}(t)) = h_0(t)\exp\left[\beta_2 \text{ PRISON } + \beta_3 \text{ DOSE } + \delta_1 \text{ CLINIC-NEW } g_1(t) + \delta_2 \text{ CLINIC-NEW } g_2(t)\right]$$

   where $g_1(t) = \begin{cases} 1 & \text{if } t \geq 365 \text{ days} \\ 0 & \text{if } t < 365 \text{ days} \end{cases}$ $g_2(t) = \begin{cases} 0 & \text{if } t \geq 365 \text{ days} \\ 1 & \text{if } t < 365 \text{ days} \end{cases}$ and CLINIC-NEW $= \begin{cases} 1 & \text{if clinic 1} \\ 0 & \text{if clinic 2} \end{cases}$

3. **Two heaviside functions**:

$$h(t, \mathbf{x}(t)) = h_0(t) \exp\left[\beta_2 \text{ PRISON } + \beta_3 \text{ DOSE } + \delta_1 g_1(t) + \delta_2 g_2(t)\right]$$

where $g_1(t) = \begin{cases} \text{CLINIC-NEW} & t \geq 365 \text{ days} \\ 0 & t < 365 \text{ days} \end{cases}$ $g_2(t) = \begin{cases} 0 & t \geq 365 \text{ days} \\ \text{CLINIC-NEW} & t < 365 \text{ days} \end{cases}$, and $\text{CLINIC-NEW } = \begin{cases} 1 & \text{if clinic 1} \\ 0 & \text{if clinic 2} \end{cases}$

4. **Time itself**:

$$h(t, \mathbf{x}(t)) = h_0(t) \exp\left[\beta_1 \text{ CLINIC-NEW } + \beta_2 \text{ PRISON } + \beta_3 \text{ DOSE } + \delta(\text{ CLINIC-NEW } \cdot t)\right]$$

with this model, we are able to estimate the effect of CLINIC-NEW on survival time, and thus HR, for any specified t.

$$\begin{aligned} HR_{\text{Clinic 1 vs Clinic 2}} &= \frac{h_0(t) \exp\left[\beta_1(1) + \beta_2 \text{ PRISON } + \beta_3 \text{ DOSE } + (\delta)(1)(t)\right]}{h_0(t) \exp\left[\beta_1(0) + \beta_2 \text{ PRISON } + \beta_3 \text{ DOSE } + (\delta)(0)(t)\right]} \\ &= \exp\left[\beta_1 + \delta t\right] \end{aligned}$$

8 Appendix

- censoring vs truncation
  The biggest difference between censoring and truncation is that when individuals are censored, they are included in the data set, even as their survival time is not completely known and when a data set is truncated, certain individuals are excluded in the analysis. Notice that while we refer to censored individuals we don't refer to truncated individuals. We say the data set is truncated, which is due to a selection process that may be inherent in the study design.

- Suppose $S(t)$ is the survival function for rats with total number of rats$=N$. Then
  $N \int_{30}^{45} S(t) dt,$ is the total survival time for all rats at risk during the interval 30 and 45 days
  $N(45-30) S(45),$ is the time at risk between 30 and 45 days among rats that remain tumor-free during this entire interval of time
  $N \int_{30}^{45} S(t) dt - N(45-30) S(45),$ is the time at risk between 30 and 45 days among rats that develop tumor during this interval of time

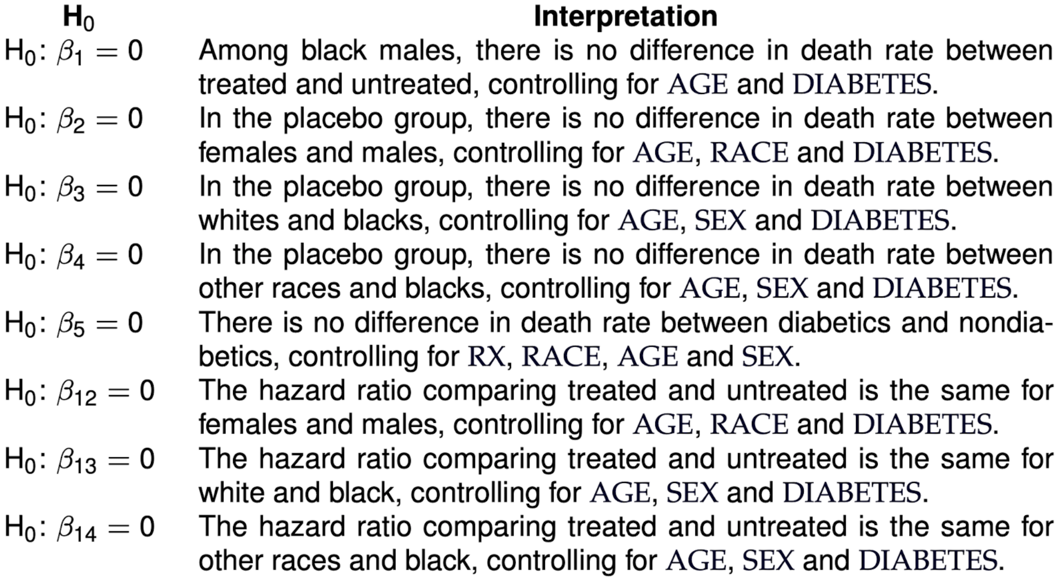
- Consider the following model, $h_s(t, \mathbf{X}) = h_{s0}(t) \exp(\beta_1 \mathrm{RX})$, where s = 1 for diabetics, s = 2 for nondiabetics, RX = 1 if receiving treatment, RX = 0 if receiving placebo.

  1. hazard rate for diabetics receiving treatment: $h_{10}(t) e^{\beta_1}$
  2. hazard ratio comparing diabetics receiving treatment to diabetics receiving placebo: $\frac{h_{10}(t) e^{\beta_1(1)}}{h_{10}(t) e^{\beta_1(0)}} = e^{\beta_1}$
  3. hazard rate for non-diabetics receiving placebo: $h_{20}(t) e^{\beta_1(0)} = h_{20}(t)$
  4. hazard ratio comparing a non-diabetic receiving treatment to a non-diabetic receiving placebo: $\frac{h_{20}(t) e^{\beta_1(1)}}{h_{20}(t) e^{\beta_1(0)}} = e^{\beta_1}$

- In a certain prospective cohort study, the outcome is time to all-cause death. The exposure is (RX): 1 for treated and 0 for untreated. Researchers wish to take into account the following variables: AGE (which is continuous), FEMALE (1 for females, 0 for males), RACE which has three levels: black, white and others, and DIABETES (1 for yes, 0 for no). Race is coded as two dummy variables, WHITE (which is 1 for white, 0 otherwise) and OTHERS (which is 1 for nonblack and nonwhite, 0 otherwise). AGE does not satisfy the proportional hazards assumption, and researchers create three AGE categories. Consider the following stratified no-interaction Cox model:

$$\begin{aligned} h_s(t, \mathbf{X}) = & h_{s0}(t) \exp\left[\beta_1 \mathrm{RX} + \beta_2 \mathrm{FEMALE} + \beta_3 \text{ WHITE } + \beta_4 \mathrm{OTHERS} + \beta_5 \text{ DIABETES } + \right. \\ & \left. \beta_{12}(\mathrm{RX})(\text{ FEMALE }) + \beta_{13}(\mathrm{RX})(\mathrm{WHITE}) + \beta_{14}(\mathrm{RX})(\text{ OTHERS })\right] \end{aligned}$$

  where where s = 0, 1, 2 represents the three AGE categories. Then



H_0	Interpretation
H_0: $\beta_1 = 0$	Among black males, there is no difference in death rate between treated and untreated, controlling for AGE and DIABETES.
H_0: $\beta_2 = 0$	In the placebo group, there is no difference in death rate between females and males, controlling for AGE, RACE and DIABETES.
H_0: $\beta_3 = 0$	In the placebo group, there is no difference in death rate between whites and blacks, controlling for AGE, SEX and DIABETES.
H_0: $\beta_4 = 0$	In the placebo group, there is no difference in death rate between other races and blacks, controlling for AGE, SEX and DIABETES.
H_0: $\beta_5 = 0$	There is no difference in death rate between diabetics and nondiabetics, controlling for RX, RACE, AGE and SEX.
H_0: $\beta_{12} = 0$	The hazard ratio comparing treated and untreated is the same for females and males, controlling for AGE, RACE and DIABETES.
H_0: $\beta_{13} = 0$	The hazard ratio comparing treated and untreated is the same for white and black, controlling for AGE, SEX and DIABETES.
H_0: $\beta_{14} = 0$	The hazard ratio comparing treated and untreated is the same for other races and black, controlling for AGE, SEX and DIABETES.

- Statistics that examine the influence of deleting an observation when performing linear regression: difference in fits, cook’s distance, deleted residual (studentized residuals are deleted residuals that are standardized)