Chapter 6

More on the Cox PH model

I. Confidence intervals and hypothesis tests
   - Two methods for confidence intervals
   - Wald tests and likelihood ratio tests
   - Interpretation of parameter estimates
   - An example with real data from an AIDS clinical trial

II. Predicted survival under proportional hazards

III. Predicted medians and P-year survival

6.1 Constructing Confidence intervals and tests for the Hazard Ratio

See Collett 3.4.

Many software packages provide estimates of β, but the hazard ratio (i.e., $\exp(\beta)$) is usually the parameter of interest.

We can use the delta method to get standard errors for $\exp(\hat{\beta})$:

$$Var(\exp(\hat{\beta})) = \exp(2\hat{\beta})Var(\hat{\beta})$$

6.1.1 Constructing confidence intervals for $\exp(\beta)$

Two options: (assuming that β is a scalar)

I. Using $se(\exp\hat{\beta})$ obtained above via the delta method as $se(\exp\hat{\beta}) = \sqrt{[Var(\exp(\hat{\beta}))]}$, calculate the endpoints as:

$$[L, U] \quad = \quad [e^{\hat{\beta}} - 1.96\, se(e^{\hat{\beta}}), e^{\hat{\beta}} + 1.96\, se(e^{\hat{\beta}})]$$

II. Form a confidence interval for $\hat{\beta}$, and then exponentiate the endpoints.

$$[L, U] \quad = \quad [e^{\hat{\beta} - 1.96 se(\hat{\beta})}, e^{\hat{\beta} + 1.96 se(\hat{\beta})}]$$

Method II is preferable since $\hat{\beta}$ converges to a normal distribution more quickly than $\exp(\hat{\beta})$.

6.2 Hypothesis Tests

For each covariate of interest, the null hypothesis is

$$H_o : \beta_j = 0$$

6.2.1 The Wald test

The Wald test[1] of the above hypothesis is constructed as:

$$Z = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \qquad \text{or} \qquad \chi^2 = \left(\frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \right)^2$$

The test for $\beta_j = 0$ assumes that all other terms in the model are fixed. If we have a factor A with a levels, then we would need to construct a χ^2 test with $(a-1)$ df, using a test statistic based on a quadratic form:

$$\chi^2_{(a-1)} \quad = \quad \widehat{\boldsymbol{\beta}}'_A Var(\widehat{\boldsymbol{\beta}}_A)^{-1} \widehat{\boldsymbol{\beta}}_A$$

where $\boldsymbol{\beta}_A = (\beta_2, ..., \beta_a)'$ are the $(a-1)$ coefficients corresponding to $Z_2, ..., Z_a$ (or $Z_1, ..., Z_{a-1}$, depending on the reference group).

6.2.2 Comparing nested models $\Rightarrow$ Likelihood Ratio Tests

Suppose there are $(p+q)$ explanatory variables measured:

$$Z_1, \ldots, Z_p, Z_{p+1}, \ldots, Z_{p+q}$$

and proportional hazards are assumed.

[1] The first follows a normal distribution, and the second follows a χ^2 with 1 df. STATA gives the Z statistic, while SAS gives the χ^2_1 test statistic (the p-values are also given, and don't depend on which form, Z or χ^2, is provided)

Consider the following models:

- **Model 1:** (contains only the first p covariates)

$$\frac{\lambda_i(t, \mathbf{Z})}{\lambda_0(t)} = \exp(\beta_1 Z_1 + \cdots + \beta_p Z_p)$$

- **Model 2:** (contains all $(p + q)$ covariates)

$$\frac{\lambda_i(t, \mathbf{Z})}{\lambda_0(t)} = \exp(\beta_1 Z_1 + \cdots + \beta_{p+q} Z_{p+q})$$

These are *nested* models. For such nested models, we can construct a **likelihood ratio** test of

$$H_0 : \beta_{p+1} = \cdots = \beta_{p+q} = 0$$

as:

$$\chi^2_{LR} \quad = \quad -2 \left[\log(\hat{L}(1)) - \log(\hat{L}(2)) \right]$$

Under H_0, this test statistic is approximately distributed as χ^2 with q df.

The likelihood-ratio test is implemented as follows:

Model 1:

```
coxph(formula = Surv(mactime, macstat) ~ karnof + rif + clari, data = mac)

  n= 1177, number of events= 121

           coef exp(coef) se(coef)      z Pr(>|z|)
karnof -0.04485   0.95614  0.01064 -4.217 2.48e-05 ***
rif     0.87197   2.39161  0.23694  3.680 0.000233 ***
clari   0.27557   1.31728  0.25801  1.068 0.285509
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
karnof    0.9561     1.0459    0.9364    0.9763
rif       2.3916     0.4181    1.5032    3.8051
clari     1.3173     0.7591    0.7944    2.1842

Concordance= 0.649  (se = 0.028 )
Rsquare= 0.027   (max possible= 0.73 )
Likelihood ratio test= 32.02  on 3 df,   p=5.193e-07
Wald test            = 32.29  on 3 df,   p=4.548e-07
Score (logrank) test = 33.16  on 3 df,   p=2.977e-07
```

Model 2:

```
coxph(formula = Surv(mactime, macstat) ~ karnof + rif + clari
> + cd4, data = mac)

  n= 1177, number of events= 121

            coef exp(coef)  se(coef)      z Pr(>|z|)
karnof -0.036874  0.963798  0.010665 -3.457 0.000546 ***
rif     0.879749  2.410294  0.237092  3.711 0.000207 ***
clari   0.252345  1.287041  0.258337  0.977 0.328664
cd4    -0.018360  0.981807  0.003684 -4.984 6.23e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

       exp(coef) exp(-coef) lower .95 upper .95
karnof    0.9638     1.0376    0.9439    0.9842
rif       2.4103     0.4149    1.5145    3.8360
clari     1.2870     0.7770    0.7757    2.1354
cd4       0.9818     1.0185    0.9747    0.9889

Concordance= 0.716  (se = 0.028 )
Rsquare= 0.053   (max possible= 0.73 )
Likelihood ratio test= 63.77  on 4 df,   p=4.682e-13
Wald test            = 55.59  on 4 df,   p=2.449e-11
Score (logrank) test = 56.22  on 4 df,   p=1.806e-11
```

The likelihood ratio test for the effect of CD4 is twice the difference in minus log-likelihoods between the two models:

$$\chi^2_{LR} \quad = \quad 2 * (754.533 - (738.66)) = 31.74$$

How does this test statistic compare to the Wald χ^2 test? The likelihood ratio is distributed according to a chi-square statistic with 1 degree of freedom, resulting in a p-value which is virtually zero.

The R code and results are as follows:

```
Analysis of Deviance Table
 Cox model: response is  Surv(mactime, macstat)
 Model 1: ~ karnof + rif + clari
 Model 2: ~ karnof + rif + clari + cd4
   loglik Chisq Df P(>|Chi|)
1 -754.49
2 -738.62 31.75  1 1.754e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

6.2.3 Estimating the hazard ratio by the Cox model

In the MAC study, there were three treatment arms (`rif`, `clari`, and the `rif+clari` combination). Because we have only included the `rif` and `clari` effects in the model,

the combination therapy is the "reference" group.

From **Model 2** above, the hazard ratio (results under column header `exp(coef)` in the output) associated with Rifabutin is $HR_{rif} = 2.41$. The hazard ratio associated with Clarithromycin is $HR_{clari} = 1.28$.

The interpretation of these hazard ratios has to do with whether patients on either of the individual treatments have higher (if $HR > 1$) or lower (if $HR < 1$) instantaneous risk of developing the even as compared to those who receive the combination treatment (i.e., on both Rifabutin and Clarithromycin). Thus, we see that those on Rifabutin-only therapy have almost 2.5-fold higher hazard to develop the event compared to the combination therapy (p-value=0.0002), while those who are on Clarithromycin-only therapy have 28% higher hazard (p-value=0.3287). We can say then that most of the benefit of the combination therapy comes from inclusion of Rifabutin in the treatment regimen (and that Clarithromycin conveys a weak, if any, benefit).

We can conduct an overall test of treatment via the `wald.test` command in R (part of the `aod` package):

```
Wald test:
----------

Chi-squared test:
X2 = 17.0, df = 2, P(> X2) = 2e-04
```

for a 2 df Wald chi-square test of whether both treatment coefficients are equal to 0. This `wald.test` command can be used to conduct an overall test for any number of effects as we will see below.

Now let's see how we would carry out a test for the *difference* between the two treatments. One way to do this is to make either treatment the reference group. This is done as follows:

```
Call:
coxph(formula = Surv(mactime, macstat) ~ karnof + rif + I(rif + clari)
> + cd4, data = mac)

  n= 1177, number of events= 121

                    coef exp(coef)  se(coef)      z Pr(>|z|)
karnof         -0.036874  0.963798  0.010665 -3.457 0.000546 ***
rif             0.627403  1.872742  0.211970  2.960 0.003078 **
I(rif + clari)  0.252345  1.287041  0.258337  0.977 0.328664
cd4            -0.018360  0.981807  0.003684 -4.984 6.23e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Note here the use of the identity function `I()`, which tells R to interpret `rif + clari` as is and not as two additional factors in the model statement (since "+" literally adds factors to the model).

The output is as follows:

```
Call:
coxph(formula = Surv(mactime, macstat) ~ karnof + rif + I(rif + clari)
> + cd4, data = mac)

  n= 1177, number of events= 121

                    coef exp(coef)  se(coef)      z Pr(>|z|)
karnof         -0.036874  0.963798  0.010665 -3.457 0.000546 ***
rif             0.627403  1.872742  0.211970  2.960 0.003078 **
I(rif + clari)  0.252345  1.287041  0.258337  0.977 0.328664
cd4            -0.018360  0.981807  0.003684 -4.984 6.23e-07 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

This means that, compared to Clarithromycin (the reference group), treatment with Rifabutin results in 87% *higher* hazard for the event. This explicitly shows that treatment with Clarithromycin alone is inferior to treatment with Rifabutin alone. The same customized test can be carried out explicitly via the `wald.test` command as follows:

- Define a contrast matrix $L = \{0, -1, 1, 0\}$ (since the coefficients for `rif` and `clar` are, respectively, second and third in the coefficient matrix b)
- Carry out the Wald test based on $\chi^2_{Wald} = (L\hat{\beta}) \left\{ L' \text{Var}(\hat{\beta}) L \right\}^{-1} (L\hat{\beta})'$.

The results are as follows:

```
Wald test:
----------

Chi-squared test:
X2 = 8.8, df = 1, P(> X2) = 0.0031
```

6.3 Predicted Survival using PH

The Cox PH model says that $\lambda_i(t, \mathbf{Z}) = \lambda_0(t) \exp(\boldsymbol{\beta}\mathbf{Z})$. What does this imply about the survival function, $S_z(t)$, for the i-th individual with covariates $\mathbf{Z}_i$?

For the baseline (reference) group, we have:

$$S_0(t) \quad = \quad e^{-\int_0^t \lambda_0(u)du} = e^{-\Lambda_0(t)}$$

This is by definition of a survival function (see intro notes). For the i-th patient with covariates $\mathbf{Z}_i$, we have:

$$\begin{aligned} S_i(t) &= e^{-\int_0^t \lambda_i(u)du} = e^{-\Lambda_i(t)} \\ &= e^{-\int_0^t \lambda_0(u)\exp(\boldsymbol{\beta}\mathbf{Z}_i)du} \\ &= e^{-\exp(\boldsymbol{\beta}\mathbf{Z}_i)\int_0^t \lambda_0(u)du} \\ &= \left[e^{-\int_0^t \lambda_0(u)du}\right]^{\exp(\boldsymbol{\beta}\mathbf{Z}_i)} \\ &= [S_0(t)]^{\exp(\boldsymbol{\beta}\mathbf{Z}_i)} \end{aligned}$$

This uses the mathematical relationship $[e^b]^a = e^{ab}$.

Say we are interested in the survival pattern for single males in the nursing home study. Based on the previous formula, if we had an estimate for the survival function in the reference group, i.e., $\hat{S}_0(t)$, we could get estimates of the survival function for any set of covariates $\mathbf{Z}_i$.

6.3.1 Estimating the baseline survival function, $S_0(t)$?

We could use the KM estimator, but there are a few disadvantages of that approach:

- It would only use the survival times for observations contained in the reference group, and not all the rest of the survival times.
- It would tend to be somewhat choppy, since it would reflect the smaller sample size of the reference group.
- It's possible that there are no subjects in the dataset who are in the "reference" group (ex. say covariates are age and sex; there is no one of age=0 in our dataset).

Instead, we will use a baseline hazard estimator which takes advantage of the proportional hazards assumption to get a smoother estimate.

$$\hat{S}_i(t) = [\hat{S}_0(t)]^{\exp(\widehat{\boldsymbol{\beta}}\mathbf{Z}_i)}$$

Using the above formula, we substitute $\widehat{\boldsymbol{\beta}}$ based on fitting the Cox PH model, and calculate $\hat{S}_0(t)$ by one of the following approaches:

- Breslow estimator (Stata)
- Kalbfleisch/Prentice estimator (SAS)

(1) **Breslow Estimator:**

$$\hat{S}_0(t) = \exp^{-\hat{\Lambda}_0(t)}$$

where $\hat{\Lambda}_0(t)$ is the estimated cumulative baseline hazard:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \left(\frac{d_j}{\sum_{k\in\mathcal{R}(\tau_j)} \exp(\beta_1 Z_{1k} + \ldots \beta_p Z_{pk})} \right)$$

(2) **Kalbfleisch/Prentice Estimator**

$$\hat{S}_0(t) = \prod_{j:\tau_j<t} \hat{\alpha}_j$$

where $\hat{\alpha}_j, j = 1, ...d$ are the MLE's obtained by assuming that $S(t; Z)$ satisfies

$$S(t; Z) = [S_0(t)]^{e^{\beta Z}} = \left[\prod_{j:\tau_j<t} \alpha_j \right]^{e^{\beta Z}} = \prod_{j:\tau_j<t} \alpha_j^{e^{\beta Z}}$$

Breslow Estimator: further motivation

The Breslow estimator is based on extending the concept of the Nelson-Aalen estimator to the proportional hazards model.

Recall that for a single sample with no covariates, the **Nelson-Aalen Estimator** of the cumulative hazard is:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \frac{d_j}{r_j}$$

where d_j and r_j are the number of deaths and the number at risk, respectively, at the j-th death time.

When there are covariates and assuming the PH model above, one can generalize this to estimate the cumulative baseline hazard by adjusting the denominator:

$$\hat{\Lambda}(t) = \sum_{j:\tau_j<t} \left(\frac{d_j}{\sum_{k\in\mathcal{R}(\tau_j)} \exp(\beta_1 Z_{1k} + \ldots \beta_p Z_{pk})} \right)$$

Heuristic: The expected number of failures in $(t, t + \delta t)$ is

$$d_j \approx \delta t \times \sum_{k\in\mathcal{R}(t)} \lambda_0(t) exp(z_k \hat{\beta})$$

Hence,

$$\delta t \times \lambda_0(t_j) \approx \frac{d_j}{\sum_{k\in\mathcal{R}(t)} exp(z_k \hat{\beta})}$$

Kalbfleisch/Prentice Estimator: further motivation

This method is analogous to the Kaplan-Meier Estimator. Consider a discrete time model with hazard $(1-\alpha_j)$ at the j-th observed death time.

(Note: we use $\alpha_j = (1-\lambda_j)$ to simplify the algebra!)

Thus, for someone with z=0, the survivorship function is

$$S_0(t) = \prod_{j:\tau_j<t} \alpha_j$$

and for someone with $Z \neq 0$, it is:

$$S(t;Z) = S_0(t)^{e^{\beta Z}} = \left[\prod_{j:\tau_j<t} \alpha_j\right]^{e^{\beta Z}} = \prod_{j:\tau_j<t} \alpha_j^{e^{\beta Z}}$$

The likelihood contributions under this model are:

- for someone censored at t: $S(t;Z)$
- for someone who fails at t_j:

$$S(t_{(j-1)};Z) - S(t_j;Z) = \left[\prod_{k<j} \alpha_j\right]^{e^{\beta z}} [1-\alpha_j^{e^{\beta Z}}]$$

The solution for α_j satisfies:

$$\sum_{k\in\mathcal{D}_j} \frac{\exp(Z_k\beta)}{1-\alpha_j^{\exp(Z_k\beta)}} = \sum_{k\in\mathcal{R}_j} \exp(Z_k\beta)$$

Note what happens when $Z=0$

6.3.2 Obtaining $\hat{S}_0(t)$ from software packages

Stata and R provide the Breslow estimator of $S_0(t;Z)$, but not predicted survivals at specified covariate values..... you have to construct these yourself

SAS uses the Kalbfleisch/Prentice estimator of the baseline hazard, and can provide estimates of survival at arbitrary values of the covariates with a little bit of programming.

In practice, they are **incredibly** close! (see Fleming and Harrington 1984, *Communications in Statistics*)

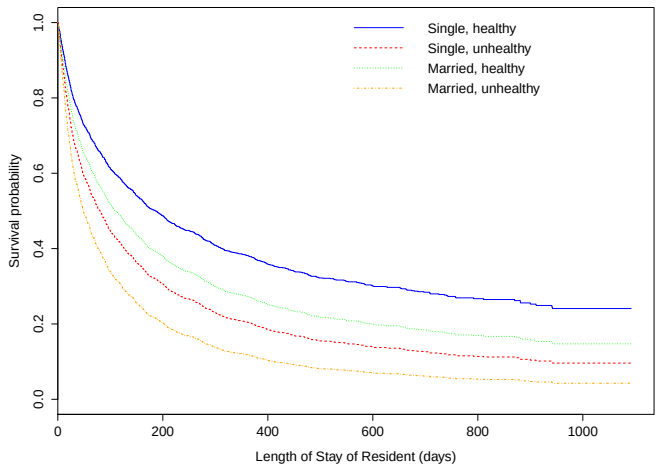
Using R to predict survival

Consider predicted survival for the groups of men and women with the best (`health`==2) and worst health (`health`==5). The R command `survfit` calculates the predicted survival values.

```
              1          2         3          4
  [1,] 0.9896288 0.98298849 0.9860573 0.97715684    1
  [2,] 0.9815838 0.96987163 0.9752767 0.95963670    2
  [3,] 0.9773101 0.96293165 0.9695622 0.95040011    3
  [4,] 0.9692168 0.94984305 0.9587642 0.93304299    4
  [5,] 0.9587077 0.93295246 0.9447896 0.91076637    5
  [6,] 0.9515173 0.92146452 0.9352587 0.89569475    6
  [7,] 0.9428590 0.90770541 0.9238150 0.87772909    7
   .    .         .          .         .             .
   .    .         .          .         .             .
   .    .         .          .         .             .
[592,] 0.2408962 0.09607878 0.1470458 0.04263921 1088
[593,] 0.2408962 0.09607878 0.1470458 0.04263921 1091
[594,] 0.2408962 0.09607878 0.1470458 0.04263921 1092
```

In the output above, 1-4 represent the four groups and the last column the 594 distinct failure times when the survival is estimated. We can get a visual picture of what the proportional hazards assumption implies by looking at these four subgroups as seen in Figure 6.1. Stata creates a predicted baseline survival estimate for every observed event time in the dataset, even if there are duplicates.

Figure 6.1: Estimated survival by the Cox model



Predicted Survival for Subgroups

We can estimate survival (i.e., $\hat{S}(t;\mathbf{Z})$) even for groups which are not included in the data.

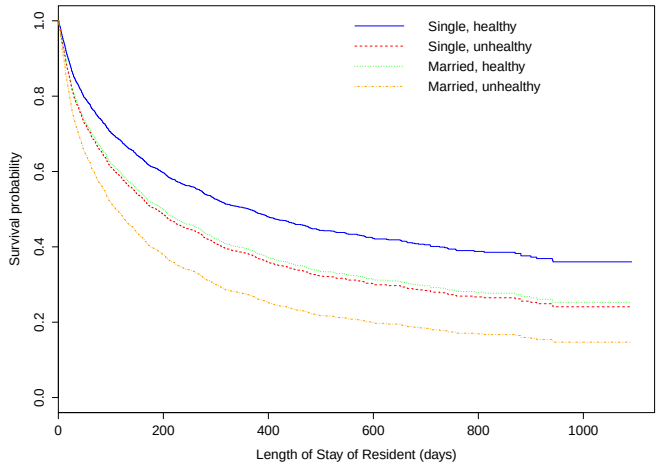
This is possible because, once we estimate $\hat{S}_0(t)$, the baseline survival, the survival of a group with measurements $\mathbf{Z}$ is equal to $\hat{S}(t;\mathbf{Z}) = \left[\hat{S}_0(t)\right]^{\exp(\boldsymbol{\beta}'\mathbf{Z})}$.

This is accomplished in R by creating a new data set with the values of $\mathbf{Z}$ as follows:

```
newdata2 = data.frame(married = c(0,0,1,1),health = c(0, 2, 0 ,2))
surv.cox2 = survfit(fit.cox, newdata = newdata2)
```

We can get a visual picture of what the proportional hazards assumption implies for these four subgroups as seen in Figure 6.2.

Figure 6.2: Estimated survival for subgroups outside the analysis data



6.3.3 Predicted medians and P-year survival

Predicted Medians

Suppose we want to find the predicted median survival for an individual with a specified combination of covariates (e.g., a single male with health status 0).

Three possible approaches:

(1) Calculate the median from the subset of individuals with the specified covariate combination (using KM approach)

(2) Generate predicted survival curves for each combination of covariates, and obtain the medians directly

(3) Generate the predicted survival curve from the estimated baseline hazard, for individuals with covariates $\mathbf{Z}_i$. This can be accomplished since we know

$$S(M; Z) = [S_0(M)]^{e^{\beta Z_i}} = 0.5$$

Hence, M satisfies (multiplying both sides by $e^{-\beta Z_i}$):

$$S_0(M) = [0.5]^{e^{-\beta Z_i}}$$

Example: Nursing home data

Predicting median survival by Kaplan Meier

Consider the following output:

```
            [,1] [,2]
  [1,] 0.97658863    1
  [2,] 0.96655518    2
     .   .           .
     .   .           .
     .   .           .
 [89,] 0.50167224  149
 [90,] 0.49832776  151
     .   .           .
     .   .           .
     .   .           .
[265,] 0.50370370   61
[266,] 0.48888889   62
     .   .           .
     .   .           .
     .   .           .
[347,] 0.52380952   90
[348,] 0.50000000   95
[349,] 0.47619048  113
     .   .           .
     .   .           .
     .   .           .
[382,] 0.51515152   22
[383,] 0.48484848   23
     .   .           .
     .   .           .
     .   .           .
```

Or using R ...

```
                          n events median 0.95LCL 0.95UCL
group=Single, healthy    299    227    151     126     199
group=Single, unhealthy  135    121     62      44      81
group=Married, healthy    42     35    104      64     195 <===!
group=Married, unhealthy  33     30     23      17      68
```

Notice that R appears to be interpolating between 95 days (when the survival probability was exactly 50%) and 113 days (when the probability was $< 50\%$) and sets the median at 104 days!

Predicting survival for various subgroups

Suppose we want to estimate the median survival for a single unhealthy subject from the nursing home data. The reciprocal of the hazard ratio for unhealthy (health=5) is: $e^{-0.165*5} = 0.4373$, (where $\hat{\beta} = 0.165$ for health status)

So, we want M such that $S_0(M) = (0.5)^{0.4373} = 0.7385$ and we can use the estimate of the baseline survival estimate through the Cox model to obtain this.

Consider the following output:

```
             1          2         3          4
 [47,] 0.7345588 0.60187880 0.6600326 0.50471265   47
 [48,] 0.7303122 0.59616290 0.6548988 0.49826795   48
   .    .         .          .         .            .
   .    .         .          .         .            .
   .    .         .          .         .            .
 [78,] 0.6583264 0.50256716 0.5694792 0.39588599   78
 [79,] 0.6561186 0.49979634 0.5669086 0.39294923   80
   .    .         .          .         .            .
   .    .         .          .         .            .
   .    .         .          .         .            .
[107,] 0.5977249 0.42871652 0.5000270 0.31960180  109
[108,] 0.5965826 0.42736892 0.4987404 0.31824952  110
   .    .         .          .         .            .
   .    .         .          .         .            .
   .    .         .          .         .            .
[171,] 0.5003829 0.31997626 0.3935705 0.21552407  185
[172,] 0.4991766 0.31870771 0.3922932 0.21437409  187
   .    .         .          .         .            .
   .    .         .          .         .            .
   .    .         .          .         .            .
```

This means that the medians are 187, 80, 110, and 48 days for single healthy, single unhealthy, married healthy, and married unhealthy, respectively[2].

[2]Notice how married people tend to remain in nursing homes for a shorter time; is this because they can be cared for at home or because of financial reasons?

The following R output summarizes the previous output as follows:

```
     n events median 0.95LCL 0.95UCL
1 1591   1269    187     155     227
2 1591   1269     80      65      97
3 1591   1269    110      86     149
4 1591   1269     48      39      66
```

Note: a similar logic can be followed to estimate other quantiles besides the median.

6.3.4 Estimating P-year survival

Suppose we want to find the P-year survival rate for an individual with a specified combination of covariates, $\hat{S}(P; \mathbf{Z}_i)$

For an individual with $\mathbf{Z}_i = 0$, the P-year survival can be obtained from the baseline survivorship function, $\hat{S}_0(P)$

For individuals with $\mathbf{Z}_i \neq 0$, it can be obtained as:

$$\hat{S}(P; \mathbf{Z}_i) = [\hat{S}_0(P)]^{e^{\widehat{\boldsymbol{\beta}}\mathbf{Z}_i}}$$

Notes:

- Although I say "P-year" survival, the units of time in a particular dataset may be days, weeks, or months. The answer here will be in the same units of time as the original data.
- If $\widehat{\boldsymbol{\beta}}\mathbf{Z}_i$ is positive, then the P-year survival rate for the i-th individual will be lower than for a baseline individual.

Why is this true?

Estimating P-year survival with R

R has the command `predict`, which has the option `"expected"`, which lists the expected number of events by time t for a given set of covariates. Note that this is the *estimated cumulative hazard* $\hat{\Lambda}(t; \mathbf{Z})$!

Thus, the estimated survival for time t (or, more relevantly for P-year survival) is,

$$\hat{S}(P; \mathbf{Z}) = \exp\left[-\hat{\Lambda}(t; \mathbf{Z})\right]$$

One and two-year survival in the nursing home example To estimate P-year survival for the four groups in the nursing home example is given as follows:

```
              1          2         3          4
  [1,] 0.9896288 0.98298849 0.9860573 0.97715684    1
  [2,] 0.9815838 0.96987163 0.9752767 0.95963670    2
    .   .         .          .         .            .
    .   .         .          .         .            .
    .   .         .          .         .            .
[286,] 0.3796947 0.20316145 0.2713845 0.11689747  364
[287,] 0.3790374 0.20258294 0.2707520 0.11644939  365
[288,] 0.3783757 0.20200128 0.2701156 0.11599931  366
    .   .         .          .         .            .
    .   .         .          .         .            .
    .   .         .          .         .            .
```

So that the one-year "survival" (remaining in the nursing home) is 37.9%, 20.2%, 27.0% and 11.6% for single healthy, single unhealthy, married healthy and married unhealthy individuals respectively. We can use R and the command `predict`[3] to calculate the estimated cumulative hazard at $t = P$ and the estimated survival $\hat{S}(P; \mathbf{Z})$:

```
newdata3 <- data.frame(married = c(0,0,1,1),
                       health = c(2,5,2,5),
                       fail= c(1,1,1,1), los=365)

predict(fit.cox, newdata=newdata3, type="expected")
[1] 0.9701205 1.5966059 1.3065521 2.1502986

exp(-predict(fit.cox, newdata=newdata3, type="expected"))
[1] 0.3790374 0.2025829 0.2707520 0.1164494
```

[3] In the R code above, note the addition of the time (`los=365`) and the censoring indicator (`fail= c(1,1,1,1)`) for the four groups in the definition of the new data set, based on which the predictions are to be made.