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

I. 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 $\exp(\hat{\beta})$:

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

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 $\exp(\hat{\beta})$.

Hypothesis Tests:

For each covariate of interest, the null hypothesis is

$$H_o : \beta_j = 0$$

A Wald test[a] 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).

[a]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)

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.

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

Some examples using the Stata `stcox` command:

Model 1:

```
. use mac
. stset mactime macstat
. stcox karnof rif clari, nohr

        failure _d:  macstat
  analysis time _t:  mactime
Cox regression -- Breslow method for ties

No. of subjects =         1151                 Number of obs   =      1151
No. of failures =          121
Time at risk    =       489509
                                               LR chi2(3)      =     32.01
Log likelihood  =   -754.52813                 Prob > chi2     =    0.0000
------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  karnof |  -.0448295   .0106355    -4.215  0.000    -.0656747   -.0239843
     rif |   .8723819   .2369497     3.682  0.000     .4079691    1.336795
   clari |   .2760775   .2580215     1.070  0.285    -.2296354    .7817903
------------------------------------------------------------------------------
```

Model 2:

```
. stcox karnof rif clari cd4, nohr

        failure _d:  macstat
  analysis time _t:  mactime


Cox regression -- Breslow method for ties

No. of subjects =         1151                 Number of obs   =      1151
No. of failures =          121
Time at risk    =       489509
                                               LR chi2(4)      =     63.74
Log likelihood  =   -738.66225                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
      _t |
      _d |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  karnof |  -.0368538   .0106652    -3.456   0.001    -.0577572   -.0159503
     rif |    .880338   .2371111     3.713   0.000     .4156089    1.345067
   clari |   .2530205   .2583478     0.979   0.327     -.253332    .7593729
     cd4 |  -.0183553   .0036839    -4.983   0.000    -.0255757   -.0111349
------------------------------------------------------------------------------
```

Notes:

- If we omit the **nohr** option, we will get the estimated hazard ratio along with 95% confidence intervals using Method II (i.e., forming a CI for the log HR (beta), and then exponentiating the bounds)

```
------------------------------------------------------------------------------
      _t |
      _d | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
---------+--------------------------------------------------------------------
  karnof |   .9638171   .0102793   -3.456   0.001     .9438791    .9841762
     rif |   2.411715   .5718442    3.713   0.000     1.515293    3.838444
   clari |    1.28791   .3327287    0.979   0.327     .7762102    2.136936
     cd4 |   .9818121   .0036169   -4.983   0.000     .9747486    .9889269
------------------------------------------------------------------------------
```

- We can also compute the hazard ratio ourselves, by exponentiating the coefficients:

$$HR_{cd4} \quad = \quad \exp(-0.01835) = 0.98$$

Why is this HR so close to 1, and yet still significant?

What is the interpretation of this HR?

- 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.
- We can conduct an overall test of treatment using the `test` command in Stata:

```
. test rif clari

 ( 1)  rif = 0.0
 ( 2)  clari = 0.0

           chi2(  2) =   17.01
         Prob > chi2 =    0.0002
```

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

- The `test` command can also be used to test whether there is a difference between the `rif` and `clari` treatment arms:

```
. test rif=clari

 ( 1)  rif - clari = 0.0

           chi2(  1) =    8.76
         Prob > chi2 =    0.0031
```

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

$$\chi^2_{LR} \quad = \quad 2 * (754.53 - (738.66)) = 31.74$$

How does this test statistic compare to the Wald χ^2 test?

II. 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 $\mathbf{Z}_i$.

How can we estimate the 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.

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

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 $Z = 0$)

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

- Stata provides 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 Stata to Predict Survival

The Stata command `basesurv` calculates the predicted survival values for the reference group, i.e., those subjects with all covariates=0.

(1) **Baseline Survival:**

To obtain the estimated baseline survival $\hat{S}_0(t)$, follow the example below (for the nursing home data):

```
. use nurshome

. stset los fail

. stcox married health, basesurv(prsurv)

. sort los

. list los prsurv
```

Estimating the Baseline Survival with Stata

```
            los      prsurv

   1.         1   .99252899
   2.         1   .99252899
   3.         1   .99252899
   4.         1   .99252899
   5.         1   .99252899
.
.
.
  37.         2   .98671824
  38.         2   .98671824
  39.         2   .98671824
  40.         3   .98362595
  41.         3   .98362595
.
.
.
```

Stata creates a predicted baseline survival estimate for every observed event time in the dataset, even if there are duplicates.

(2) **Predicted Survival for Subgroups**

To obtain the estimated survival $\hat{S}_i(t)$ for any other subgroup (i.e., not the reference or baseline group), follow the Stata commands below:

```
. predict betaz, xb

. gen newterm=exp(betaz)

. gen predsurv=prsurv^newterm

. sort married health los

. list married health los predsurv
```

Predicting Survival for Subgroups with Stata

	married	health	los	predsurv
1.	0	2	1	.9896138
8.	0	2	2	.981557
11.	0	2	3	.9772769
.....				
300.	0	3	1	.9877566
302.	0	3	2	.9782748
304.	0	3	3	.9732435
.....				
768.	0	4	1	.9855696
777.	0	4	2	.9744162
779.	0	4	3	.9685058
.				
.				
1468.	1	4	1	.9806339
1469.	1	4	2	.9657326
1472.	1	4	3	.9578599
.....				
1559.	1	5	1	.9771894
1560.	1	5	2	.9596928
1562.	1	5	3	.9504684

III. 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

OBS	MARRIED	HEALTH	LOS	PREDSURV
171	0	2	184	0.50104
172	0	2	185	0.49984
474	0	5	78	0.50268
475	0	5	80	0.49991

897	1	2	108	0.50114
898	1	2	109	0.49986
1233	1	5	47	0.50519
1234	1	5	48	0.49875

Recall that previously we defined the median as the *smallest* value of t for which $\hat{S}(t) \leq 0.5$, so the medians from above would be 185, 80, 109, and 48 days for single healthy, single unhealthy, married healthy, and married unhealthy, respectively.

(3) Generate the predicted survival curve from the estimated baseline hazard, as follows:

We want the estimated median (M) for an individual with covariates $\mathbf{Z}_i$. 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: 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 $S_0(M) = (0.5)^{0.4373} = 0.7385$

From the estimated *baseline* survival curve (this is tricky!... we might be tempted to look at the survival estimates for single unhealthy, but we actually need to look at those for single, health=0):

OBS	MARRIED	HEALTH	LOS	PREDSURV
79	0	0	78	0.74028
80	0	0	80	0.73849
81	0	0	81	0.73670

So the estimated median would still be 80 days.

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

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?