STAT 6390: Analysis of Survival Data

Textbook coverage: Chapter 3

Steven Chiou

Department of Mathematical Sciences,
University of Texas at Dallas

Cox proportional hazards model

- The Cox model is expressed by the hazard function.
- The hazard function can be (loosely) interpreted as the risk of dying at time t.
- The Cox model has the form:

$$h(t) = h_0(t) \cdot \exp\{\beta_1 x_1 + \beta_2 x_2 + \dots \beta_p x_p\},$$

  where

  - t is the survival time.
  - $\{x_1, \dots, x_p\}$ is a set of p covariates.
  - $\{\beta_1, \dots, \beta_p\}$ is the regression parameters; effect of covariates.
  - $h_0(t)$ is the baseline hazard. It is the value of the hazard when all x's are 0.
  - No need to specify an "intercept" term as it gets absorb to $h_0(t)$.

Cox proportional hazards model

- The quantity e^{β_i} is interpreted as the hazard ratio (HR).
  - $\beta_i > 0 \rightarrow$ HR $> 1 \rightarrow$ hazard increases $\rightarrow$ survival time decreases.
  - $\beta_i = 0 \rightarrow$ HR $= 1 \rightarrow$ no change in hazard $\rightarrow$ no change in survival time.
  - $\beta_i < 0 \rightarrow$ HR $< 1 \rightarrow$ hazard decreases $\rightarrow$ survival time increases.
- HR (and hazard) is negatively associated with the length of survival.
- The Cox model assumes the hazard curves among different patients should be proportional and cannot cross.

Fitting the Cox model in R

- We have used `coxph` to compute the Nelson–Aalen estimator.
- The usage of `coxph` is similar to that of `survreg`.
- We will start with one covariate, `gender`.

```
> fm <- Surv(lenfol, fstat) ~ gender
> fit.cox <- coxph(fm, data = whas100)
> fit.aft <- survreg(fm, data = whas100)
```

- The coefficients are in opposite directions.

```
> coef(fit.cox)
   gender
0.5548116
> coef(fit.aft)
(Intercept)      gender
   8.463727   -0.790436
```

- `fit.cox` does not have an intercept term.
- The two parameter estimates have opposite signs.

Fitting the Cox model in R

- The `summary` gives:

```
> summary(fit.cox)
Call:
coxph(formula = fm, data = whas100)

  n= 100, number of events= 51

         coef exp(coef) se(coef)     z Pr(>|z|)
gender 0.5548    1.7416   0.2824 1.965   0.0494 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
gender     1.742     0.5742     1.001     3.029

Concordance= 0.565  (se = 0.035 )
Rsquare= 0.037   (max possible= 0.985 )
Likelihood ratio test= 3.75  on 1 df,   p=0.05
Wald test            = 3.86  on 1 df,   p=0.05
Score (logrank) test = 3.96  on 1 df,   p=0.05
```

- The $\hat{\beta}$ is positive indicating that male patients ($\texttt{gender} = 1$) have higher risk of death.

Fitting the Cox model in R

- The hazard ratio in this example is $e^{0.5548} \approx 1.7416$.
- This implies males (`gender = 1`) die at about 1.74 times (74% higher) the rate of females.
- Like in interpreting the odds-ratio estimator in logistic regression, the end-points of a 95% confidence interval for the hazard ratio is

$$\exp\left[\hat{\beta} \pm 1.96(\hat{\mathrm{SE}})\hat{\beta}\right] = \exp\left[0.5548 \pm 1.96 \cdot 0.2824\right] \approx [1.001, 3.029].$$

- The 95% confidence interval lies entirely above 1, echoing the significance of $\hat{\beta}$ at $\alpha = 0.05$.

Fitting the Cox model in R

- Three related tests to assess the significance of the model.
  - Partial likelihood ratio test
  - Wald test
  - score test
- These three tests are also indicated in the bottom of the `summary`.

Partial likelihood ratio test

- The partial likelihood ratio test is calculated as twice the difference between the log-partial likelihood of the
    - "full model" , denoted by $\ell_p(\hat{\beta})$.
    - "reduced model", denoted by $\ell_p(0)$.
- The log-partial likelihood ratio is then defined as

$$G = 2 \cdot \{\ell_p(\hat{\beta}) - \ell_p(0)\},$$

where

$$\ell_p(\hat{\beta}) = \sum_{i=1}^{n} \Delta_i \left[X_i\hat{\beta} - \log \left\{ \sum_{j \in R(t_i)} e^{X_j\hat{\beta}} \right\} \right],$$

$$\ell_p(0) = -\sum_{i=1}^{n} \Delta_i \log(n_i), \text{ and } n_i \text{ is the number of risk at } t_i.$$

- The test statistic, G, follows approximately a chi-square distribution with *one* degree of freedom under null.

Partial likelihood ratio test

- The log-partial likelihood are stored in the `coxph` object as `loglik`:

```
> fit.cox$loglik
[1] -209.0977 -207.2247
```

- The two log-partial likelihoods are $\ell_p(0)$ and $\ell_p(\hat{\beta})$ respectively.
- $\ell_p(0)$ can be computed only with `survfit`, e.g.,

```
> fit.surv <- survfit(Surv(lenfol, fstat) ~ 1, data = whas100)
> with(fit.surv, -sum(n.event * log(n.risk)))
[1] -209.1197
```

- The test statistic, G, and its p-value can be computed as follow

```
> 2 * diff(fit.cox$loglik)
[1] 3.746038
> 1 - pchisq(2 * diff(fit.cox$loglik), 1)
[1] 0.05293283
```

Partial likelihood ratio test

- Two useful statistics are by-product of the log-partial likelihood.
- The `Rsquare` is the (generalized) R^2 statistic, defined as

$$R^2 = 1 - \left\{ \frac{L_p(0)}{L_p(\hat{\beta})} \right\}^{2/n} = 1 - \left\{ e^{\ell_p(0) - \ell_p(\hat{\beta})} \right\}^{2/n}.$$

- The following code verifies this in R.

```
> 1 - exp(-diff(fit.cox$loglik))^.02
[1] 0.03676742
```

- Note that the message `max possible` gives the most extreme possible value the R^2 can be achieved given the observed data.
- In this case `max possible = 0.985`, so 0.985 represents the “prefect fit” for the dataset.

Partial likelihood ratio test

- The other useful measure is the *concordance*.
- The concordance gives the fraction of pairs in the sample (`gender = 1` and `gender = 0`), where the observations with the higher survival time has the higher probability of survival predicted by the model.
- The concordance is robust to monotone transformation of the predictor.

Wald test

- The ratio of the estimated coefficient to its estimated standard error is commonly referred to as a *Wald statistic*.
- The Wald statistic (`z`) and its *p*-value are reported in `summary`.
- The Wald statistic follows a standard normal distribution under null.
- The Wald statistic and the two-sided *p*-value can be computed as follow:

```
> z <- coef(fit.cox) / sqrt(vcov(fit.cox))
> 2 - 2 * pnorm(abs(z))
           gender
gender 0.04944828
```

- The `Wald statistic` displayed below is the chi-square version of it.

```
> z^2
         gender
gender 3.860069
> 1 - pchisq(z^2, 1)
           gender
gender 0.04944828
```

Score test

- The test statistic for the score test is the ratio of the derivative of the log-partial likelihood to the square root of the observed information at $\beta = 0$.
- The test statistic, $(z^*)^2$, is evaluated based on the score function

$$z^* = \frac{\mathrm{d}\ell_p(\beta)}{\mathrm{d}\beta} \cdot \frac{1}{\sqrt{I(\beta)}}\bigg|_{\beta=0},$$

  where $I(\cdot)$ is the observed information.
- The score test is equivalent to the log-rank test.

Score test

- Recall that the score function has the form

$$\frac{\mathrm{d}\ell_p}{\mathrm{d}\beta} = \sum_{i=1}^{n} \Delta_i \left[X_i - \left\{ \frac{\sum_{j \in R(t_i)} X_j e^{X_j \beta}}{\sum_{j \in R(t_i)} e^{X_j \beta}} \right\} \right].$$

- Suppose X_i is a categorical variable that takes values 0 and 1.
- When $\beta = 0$,
  - $\sum_{j \in R(t_i)} e^{X_j \beta}$ is the total number in the risk set at t_i.
  - $\sum_{j \in R(t_i)} X_j e^{X_j \beta}$ is the number of individual from group 1 in the risk set at t_i.

Score test

- Recall the 2 by 2 table we used to construct the log-rank statistic:

	Group 1	Group 0	Total
Failure	d_{1i}	d_{0i}	d_i
Non-failure	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
At risk	n_{1i}	n_{0i}	n_i

- Using the table notations,

$$\sum_{j \in R(t_i)} e^{X_j \beta} = n_i \text{ and } \sum_{j \in R(t_i)} X_j e^{X_j \beta} = n_{1i}.$$

- We also used $\mathrm{E}(d_{1i}) = \frac{n_{1i} \cdot d_i}{n_i}$.
- The score function reduced to

$$\frac{\mathrm{d}\ell_p}{\mathrm{d}\beta} = \sum_{i=1}^{n} \Delta_i \left[X_i - \frac{n_{1i}}{n_i} \right] = \sum_{i=1}^{D} \{ d_{1i} - \mathrm{E}(d_{1i}) \}$$

under the assumption of no ties.

Score test

- Now recall the observed information based on the score function is

$$I(\beta) = -\frac{\mathrm{d}^2 \ell_p}{\mathrm{d}\beta^2} = \sum_{i=1}^{n} \Delta_i \sum_{j \in R(t_i)} (\beta) \cdot \left(X_j - \bar{X}\right)^2$$

$$= \sum_{i=1}^{n} \Delta_i \sum_{j \in R(t_i)} \left(X_j^2 - 2X_j\bar{X} + \bar{X}^2\right)$$

- With the table notations,
  - $\sum_{j \in R(t_i)} X_j^2 = \sum_{j \in R(t_i)} X_j = n_{1i}$
  - $\sum_{j \in R(t_i)} 2X_j\bar{X} = 2 \cdot n_{1i} \cdot \frac{n_{1i}}{n_i}$
  - $\sum_{j \in R(t_i)} \bar{X}^2 = n_i \cdot \left(\frac{n_{1i}}{n_i}\right)^2$
- Simple algebra shows

$$I(\beta) = \sum_{i=1}^{m} \mathrm{Var}(d_{1i}).$$

Score test

- The following confirms the relationship between `survdiff` and the score test.

```
> survdiff(fm, data = whas100)
Call:
survdiff(formula = fm, data = whas100)

          N Observed Expected (O-E)^2/E (O-E)^2/V
gender=0 65       28     34.6      1.27      3.97
gender=1 35       23     16.4      2.68      3.97

 Chisq= 4  on 1 degrees of freedom, p= 0.05
> fit.cox$score
[1] 3.958093
```

Fitting the Cox model in R

- In general, the numeric values of the three test statistics are usually similar, and thus lead to the same conclusion.
- In situations where there is disagreement, the partial likelihood ratio test is the preferred test*.
- The partial likelihood ratio test is also useful when comparing nested models.

Model selection with partial likelihood ratio tests

- Suppose we want to test whether the addition of `bmi` is statistically significant, we can fit a new Cox model with both `bmi` and `gender`:

```
> fit.cox2 <- update(fit.cox, ~. + bmi)
> summary(fit.cox2)
Call:
coxph(formula = Surv(lenfol, fstat) ~ gender + bmi, data = whas100)

  n= 100, number of events= 51

           coef exp(coef) se(coef)      z Pr(>|z|)
gender  0.53794   1.71248  0.28257  1.904  0.05694 .
bmi    -0.09460   0.90974  0.03391 -2.790  0.00527 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
gender    1.7125     0.5839    0.9842    2.9795
bmi       0.9097     1.0992    0.8512    0.9723

Concordance= 0.632  (se = 0.043 )
Rsquare= 0.113   (max possible= 0.985 )
Likelihood ratio test= 11.94  on 2 df,   p=0.003
Wald test            = 11.92  on 2 df,   p=0.003
Score (logrank) test = 12.26  on 2 df,   p=0.002
```

Model selection with partial likelihood ratio tests

- The corresponding partial likelihoods are.

```
> fit.cox$loglik
[1] -209.0977 -207.2247
> fit.cox2$loglik
[1] -209.0977 -203.1269
```

- Note that they share the same $\ell_p(0)$.
- The partial likelihood ratio test can be performed as follow:

```
> G <- 2 * sum(fit.cox2$loglik - fit.cox$loglik)
> 1 - pchisq(G, 1)
[1] 0.004199217
```

Baseline cumulative hazard

- The function `basehaz` function computes the cumulative baseline hazard function, $H_0(t)$.
- The only argument besides the `fit` is `centered`, which specifies the covariate value used in the plot.

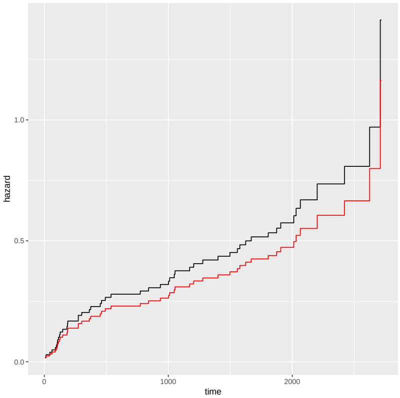
```
> args(basehaz)
function (fit, centered = TRUE)
NULL
```

- The default is `centered = TRUE` estimate at the mean of the covariates.
- Setting `centered = TRUE` does not always make sense, particularly for categorical covariates.

Baseline cumulative hazard

- The red curve gives the baseline cumulative hazard for `gender = 0`.
- The black curve gives the cumulative hazard at `gender = 0.5` (mean).
- The two curves are parallel to each other.

```
> ggplot(data = basehaz(fit.cox), aes(x = time, y = hazard)) + geom_step() +
+     geom_step(data = basehaz(fit.cox, center = FALSE), aes(x = time, y = hazar
```



Multiple covariates model

- The inference remains the same when multiple covariates are involved in the model.
- Recall the Cox model with both `gender` and `bmi`:

```
> summary(fit.cox2)
Call:
coxph(formula = Surv(lenfol, fstat) ~ gender + bmi, data = whas100)

  n= 100, number of events= 51

           coef exp(coef) se(coef)      z Pr(>|z|)
gender  0.53794   1.71248  0.28257  1.904  0.05694 .
bmi    -0.09460   0.90974  0.03391 -2.790  0.00527 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
gender    1.7125     0.5839    0.9842    2.9795
bmi       0.9097     1.0992    0.8512    0.9723

Concordance= 0.632  (se = 0.043 )
Rsquare= 0.113   (max possible= 0.985 )
Likelihood ratio test= 11.94  on 2 df,   p=0.003
Wald test            = 11.92  on 2 df,   p=0.003
Score (logrank) test = 12.26  on 2 df,   p=0.002
```

Multiple covariates model

- The hazard ratio can be interpreted jointly or separately while holding the other covariate constant.
- The regression paramter for `bmi` is negative.
- This indicates patients with higher `bmi` have lower risk of death.
- For every one unit increase in `bmi` the risk of death decrease by 10%.
- This result does not make (medical) sense and is possible due to non-linear effect in `bmi`

```
> fit.cox3 <- update(fit.cox2, ~ . + I(bmi^2))
```

Multiple covariates model

- The summary with $\texttt{bmi}^2$ is

```
> summary(fit.cox3)
Call:
coxph(formula = Surv(lenfol, fstat) ~ gender + bmi + I(bmi^2),
    data = whas100)

  n= 100, number of events= 51

              coef exp(coef)  se(coef)      z Pr(>|z|)
gender    0.407036  1.502358  0.293804  1.385  0.16593
bmi      -0.757980  0.468612  0.230462 -3.289  0.00101 **
I(bmi^2)  0.012260  1.012335  0.004165  2.944  0.00324 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
gender      1.5024     0.6656    0.8447    2.6721
bmi         0.4686     2.1340    0.2983    0.7362
I(bmi^2)    1.0123     0.9878    1.0041    1.0206

Concordance= 0.652  (se = 0.043 )
Rsquare= 0.174   (max possible= 0.985 )
Likelihood ratio test= 19.08  on 3 df,   p=3e-04
Wald test            = 23.11  on 3 df,   p=4e-05
Score (logrank) test = 25.05  on 3 df,   p=2e-05
```

- Note that the inclusion of `bmi` (and $\texttt{bmi}^2$) makes `gender` less significant.

Multiple covariates model

- The partial likelihood ratio test statistic can be computed similarly, but the Wald test statistics needs to be computed accounting for covariances.

```
> coef(fit.cox2) %*% solve(vcov(fit.cox2)) %*% coef(fit.cox2)
         [,1]
[1,] 11.91955
```

- The degrees of freedom is 2 because we now have 2 covariates in the model.

```
> 1 - pchisq(coef(fit.cox2) %*% solve(vcov(fit.cox2)) %*% coef(fit.cox2), 2)
            [,1]
[1,] 0.002580495
```