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:

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

  where

  - t is the survival time.
  - $\{x_1, \ldots, x_p\}$ is a set of p covariates.
  - $\{\beta_1, \ldots, \beta_p\}$ 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

- Three related tests to assess the significance of the coefficient, e.g., testing $H_0 : \beta = 0$.
  - 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
```