8.4 **Regression: Cox Proportional Hazards Model**

Suppose we wish to model the hazard function \( h(t) \) for a population, in terms of explanatory variables – or covariates – \( X_1, X_2, X_3, \ldots, X_m \). That is,

\[
h(t) = h(t; X_1, X_2, X_3, \ldots, X_m),
\]

so that all the individuals corresponding to one set of covariate values have a different hazard function from all the individuals corresponding to some other set of covariate values.

Assume initially that \( h \) has the general form \( h(t) = h_0(t) \cdot C(X_1, X_2, X_3, \ldots, X_m) \).

**Example**: In a population of 50-year-old males, \( X_1 = \) smoking status (0 = No, 1 = Yes), \( X_2 = \# \) pounds overweight, \( X_3 = \# \) hours of exercise per week. Consider

\[
h(t) = .02t^{e^{X_1 + 0.3X_2 - 0.5X_3}}.
\]

If \( X_1 = 0, \ X_2 = 0, \ X_3 = 0 \), then \( h_0(t) = .02t \). This is the baseline hazard. (Therefore, the corresponding survival function is \( S_0(t) = e^{-0.01t^2} \). Why?)

If \( X_1 = 1, \ X_2 = 10 \text{ lbs}, \ X_3 = 2 \text{ hrs/wk} \), then \( h(t) = .02t^{e^3} = .02t \cdot (20.1) = .402t \). (Therefore, the corresponding survival function is \( S(t) = e^{-201t^2} \). Why?)

Thus, the proportion of hazards \( \frac{h(t)}{h_0(t)} = e^{3} (= 20.1) \), i.e., constant for all time \( t \).
Furthermore, notice that this hazard function can be written as…

\[ h(t) = .02 \ t \ (e^{X_1}) (e^{0.3X_2}) (e^{-0.5X_3}). \]

Hence, with all other covariates being equal, we have the following properties.

- If \( X_1 \) is changed from 0 to 1, then the net effect is that of multiplying the hazard function by a constant factor of \( e^{1} \approx 2.72 \). Similarly,
- If \( X_2 \) is increased to \( X_2 + 1 \), then the net effect is that of multiplying the hazard function by a constant factor of \( e^{0.3} \approx 1.35 \). And finally,
- If \( X_3 \) is increased to \( X_3 + 1 \), then the net effect is that of multiplying the hazard function by a constant factor of \( e^{-0.5} \approx 0.61 \). (Note that this is less than 1, i.e., beneficial to survival.)

In general, the hazard function given by the form

\[ h(t) = h_0(t) \ e^{\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m} \]

where \( h_0(t) \) is the baseline hazard function, is called the Cox Proportional Hazards Model, and can be rewritten as the equivalent linear regression problem:

\[
\ln \left( \frac{h(t)}{h_0(t)} \right) = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m
\]

The “constant proportions” assumption is empirically verifiable. Once again, the regression coefficients are computationally intensive, and best left to a computer.

**Comment:** There are many practical extensions of the methods in this section, including techniques for hazards modeling when the “constant proportions” assumption is violated, when the covariates \( X_1, X_2, X_3, \ldots, X_m \) are time-dependent, i.e.,

\[
\ln \left( \frac{h(t)}{h_0(t)} \right) = \beta_1 X_1(t) + \beta_2 X_2(t) + \ldots + \beta_m X_m(t),
\]

when patients continue to be recruited after the study begins, etc. Survival Analysis remains a very open area of active research.