CPH 931 — Fall 2009 — Dr. Charnigo

Lecture 8

Parametric models in survival analysis

A familiar semiparametric model. The proportional hazards regression model, with which you became acquainted in CPH 930, is an example of a semiparametric model. Although

$$h_{x_1,\ldots,x_k}(t) = h_0(t)\exp[\beta_1 x_1 + \cdots + \beta_k x_k] \tag{1}$$

implies that all hazard ratios can be expressed in terms of finitely many parameters (namely, β_1 through β_k), the hazard function itself cannot. This is because $h_0(t)$ effectively contains infinitely many parameters, one for each value of t.

A useful analogy. We can draw an analogy between $h_0(t)$ and an entity that we encountered in Lecture 4.

Recall that one of the Lecture 4 topics was nonparametric regression, in which we represented the mean of a continuous response variable Y as a smooth function $m(x)$ of the value x assumed by a continuous explanatory variable X. Not requiring $m(x)$ to have the form $\alpha + \beta x$ protected us from putting a straight line through data points for which a straight line was inappropriate.

Yet, there are many applications in which putting a straight line through data points is reasonable. For such applications we usually eschew nonparametric regression in favor of the parametric model $\alpha + \beta x$ because the latter is convenient for hypothesis testing and highly amenable to interpretation. Moreover, the parametric model easily generalizes to accommodate multiple explanatory variables.

Viewing $h_0(t)$ as analogous to $m(x)$, then, we wonder whether there is any value in expressing $h_0(t)$ as a function of t that is known up to finitely many parameters.

Debating the merits of a parametric model. From the viewpoint of assessing the effects of X_1 through X_k on the hazard function $h_{x_1,\ldots,x_k}(t)$, there seems to be little value in specifying a functional form for $h_0(t)$ since $h_0(t)$ cancels out of any hazard ratio in which we may be interested.

However, if we wish to estimate the survival function

$$S_{x_1,\ldots,x_k}(t) := P(T > t \mid X_1 = x_1, \ldots, X_k = x_k) \tag{2}$$

for a subject with values x_1 through x_k on the explanatory variables, a parametric model may be useful. This is because a parametric model will admit not only a graphical representation of $S_{x_1,\ldots,x_k}(t)$ but also a tractable mathematical expression.

Survival functions and PROC PHREG. Most students leave CPH 930 without having estimated survival functions when using PROC PHREG. Survival function estimates are not part of the usual output from PROC PHREG, although with some extra coding you can request them. Have a look at page 9 of {LARYNXExamples.rtf}.

I will provide the scientific background for this example later, but for now note that estimates of four survival functions are displayed. (Actually, I have displayed estimates of cumulative distribution functions, which are basically "upside-down" survival functions.)

These estimates exhibit some irregularities. For instance, the plateaus from 2.5 to 3.1 suggest that none of the subjects in this particular sample expired between 2.5 and 3.1, but there is little reason to believe that the population survival functions should have plateaus from 2.5 to 3.1.

A parametric model can avoid such irregularities.

Discussion question. Explain why a parametric model will not be quite as simple as, say,

$$S_{x_1,\ldots,x_k}(t) = (\alpha_0 + \alpha_1 x_1 + \cdots + \alpha_k x_k) + (\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)t.$$

The Weibull model

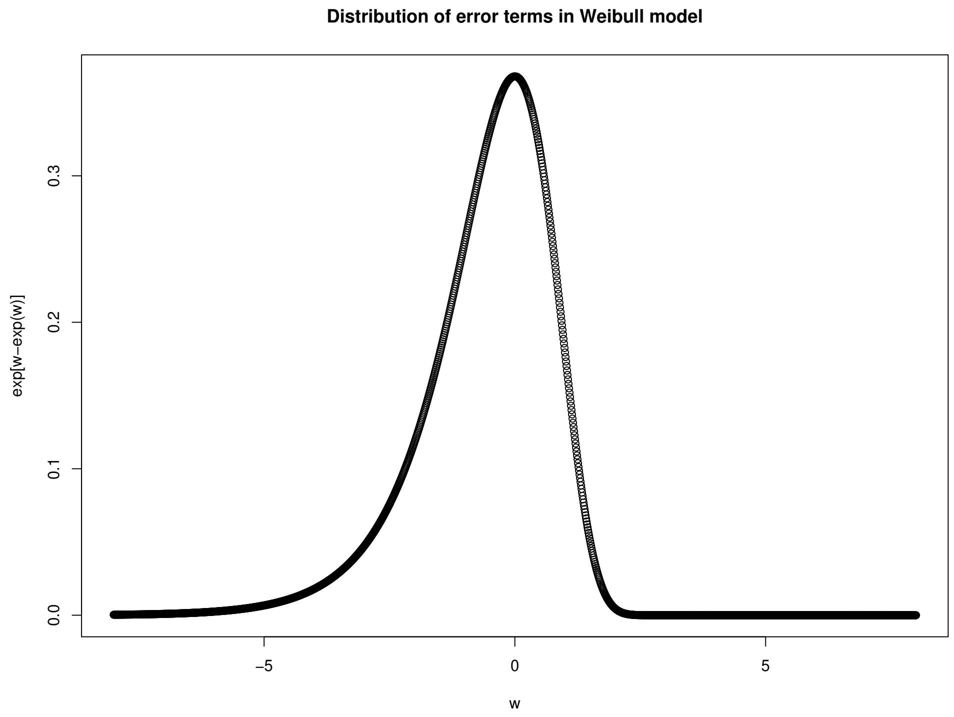
Model formulation. Let Y denote the natural logarithm of the time-to-event response variable T. The Weibull model has the form

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + \sigma W_i, \tag{3}$$

where σ is an estimable positive constant and W_1 through W_n are independent and identically distributed error terms.

Note that W_1 through W_n are <u>not</u> normally distributed. Actually, their distribution is described by the density function $\exp[w - \exp(w)]$, which I have displayed in the figure on the next page. The density function is vaguely bell shaped but, on close inspection, not quite symmetric.

Right censoring of T (and, hence, of Y) can be accommodated by PROC LIFEREG, the procedure with which SAS fits the Weibull model.



Implied survival function. Recalling that probabilities for continuous random variables are given by areas under their density functions, and that the formal mathematical tool for evaluating areas under curves is integral calculus, we find that

$$P(W > w) = \int_w^\infty \exp[u - \exp(u)]\, du = \exp[-\exp(w)]. \tag{4}$$

Let $\boldsymbol{\beta} \cdot \mathbf{x}_i$ be shorthand for $\beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}$. Then (3) yields

$$W_i = \sigma^{-1}\{Y_i - \boldsymbol{\beta} \cdot \mathbf{x}_i\}, \tag{5}$$

whence

$$\begin{aligned}
S_{x_{1,i},\ldots,x_{k,i}}(t) &:= P(T_i > t \mid X_1 = x_{1,i}, \ldots, X_k = x_{k,i}) \\
&= P(Y_i > \log t \mid X_1 = x_{1,i}, \ldots, X_k = x_{k,i}) \\
&= P(W_i > \sigma^{-1}\{\log t - \boldsymbol{\beta} \cdot \mathbf{x}_i\} \mid X_1 = x_{1,i}, \ldots, X_k = x_{k,i})
\end{aligned}$$

$$
\begin{aligned}
&= \exp[-\exp(\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\})] && (6)\\
&= \exp[-\exp(\sigma^{-1}\{\log t - \log[\exp\{\boldsymbol{\beta}\cdot\mathbf{x}_i\}]\})]\\
&= \exp[-\exp(\sigma^{-1}\{\log[t\exp\{-\boldsymbol{\beta}\cdot\mathbf{x}_i\}]\})]\\
&= S_0(t\exp\{-\boldsymbol{\beta}\cdot\mathbf{x}_i\}), && (7)
\end{aligned}
$$

where

$$S_0(t) := P(T_i > t \mid \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = 0) = \exp[-\exp(\sigma^{-1}\log t)].$$

Line (6) provides a tractable mathematical expression for the survival function $S_{x_{1,i},\ldots,x_{k,i}}(t)$, while line (7) displays the "accelerated failure time" property.

The accelerated failure time property. If having $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = 0$ is possible, then the survival function at time t for a person with $X_1 = x_{1,i}, \ldots, X_k = x_{k,i}$ equals the survival function at time $t\exp\{-\boldsymbol{\beta}\cdot\mathbf{x}_i\}$ for a person with $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = 0$. In effect, time passes more quickly for a person with $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k < 0$ than for a person with $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = 0$, while time passes less quickly for a person with $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k > 0$ than for a person with $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k = 0$.

To illustrate, suppose that $k = 1$ and that X_1 is defined as an indicator for smoking. Line (7) implies that a smoker's survival function is

$$S_0(t\exp[-\beta_0 - \beta_1]) = S_0(t\exp[-\beta_0]\exp[-\beta_1]),$$

while a nonsmoker's survival function is

$$S_0(t\exp[-\beta_0]).$$

Let t^* be an arbitrary time. If $\beta_1 = -0.693$, then the smoker's probability of surviving past time $t^*\exp[\beta_0]\exp[\beta_1] = t^*\exp[\beta_0]/2$ equals $S_0(t^*)$. This is also the nonsmoker's probability of surviving past time $t^* \exp[\beta_0]$. Putting $\tilde{t} := t^* \exp[\beta_0]/2$, we see that the smoker's probability of surviving past time $\tilde{t}$ equals the nonsmoker's probability of surviving past time $2\tilde{t}$. Thus, time passes twice as quickly for the smoker.

Implied hazard function. Starting from line (6), we can use differential calculus to recover the hazard function for a person with $X_1 = x_{1,i}, \ldots, X_k = x_{k,i}$ as

$$
\begin{aligned}
h_{x_{1,i},\ldots,x_{k,i}}(t) &= -\frac{d}{dt} \log S_{x_{1,i},\ldots,x_{k,i}}(t) \\
&= \frac{d}{dt} \exp(\sigma^{-1}\{\log t - \boldsymbol{\beta} \cdot \mathbf{x}_i\}) \\
&= \sigma^{-1} t^{-1} \exp(\sigma^{-1}\{\log t - \boldsymbol{\beta} \cdot \mathbf{x}_i\}) \\
&= \sigma^{-1} t^{-1} \exp(\log t^{\sigma^{-1}} - \sigma^{-1} \boldsymbol{\beta} \cdot \mathbf{x}_i) \\
&= \sigma^{-1} t^{\sigma^{-1}-1} \exp(-\sigma^{-1} \boldsymbol{\beta} \cdot \mathbf{x}_i). \qquad (8)
\end{aligned}
$$

Line (8) provides a tractable mathematical expression for the hazard function $h_{x_{1,i},\ldots,x_{k,i}}(t)$.

Discussion question. Use line (8) to determine the hazard ratio associated with a one-unit increase in X_1 when X_2 through X_k are fixed. What feature of the hazard ratio is remarkable?

Example with larynx cancer data. The file {LARYNXdata.xls} contains data on 90 males diagnosed with larynx cancer in the 1970's. The response variable TimetoDeath is time-to-event and represents the number of years lived after the diagnosis of larynx cancer. Right censoring of TimetoDeath is noted with a 0 on the indicator variable DiedinStudy. The explanatory variable Stage has four categories and represents the severity of the cancer; category 1 is least severe, while category 4 is most severe. I did not use any other variables from {LARYNXdata.xls} in this example.

Pages 1 and 2 of {LARYNXExamples.rtf} show SAS output for the Weibull model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \sigma W_i, \tag{9}$$

where Y is the natural logarithm of TimetoDeath and X_1 through X_3 are indicators for categories 1 to 3 of Stage.

The Type III Analysis of Effects box shows that $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ may be rejected on the basis of a 0.0001 p-value. The Analysis of Parameter Estimates box shows that all three individual null hypotheses $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$, and $H_0 : \beta_3 = 0$ may also be rejected with respective p-values < 0.0001, 0.0007, and 0.0030. The estimate of σ is 0.8846, so the estimate of σ^{-1} is $1/0.8846 = 1.1305$.

Page 3 shows a plot of the estimated survival functions for all four categories of Stage. We estimate that more than 80% of Stage 4 patients die within four years of diagnosis, while less than 50% of Stage 3 patients and about 30% of Stage 1 or Stage 2 patients die within four years of diagnosis.

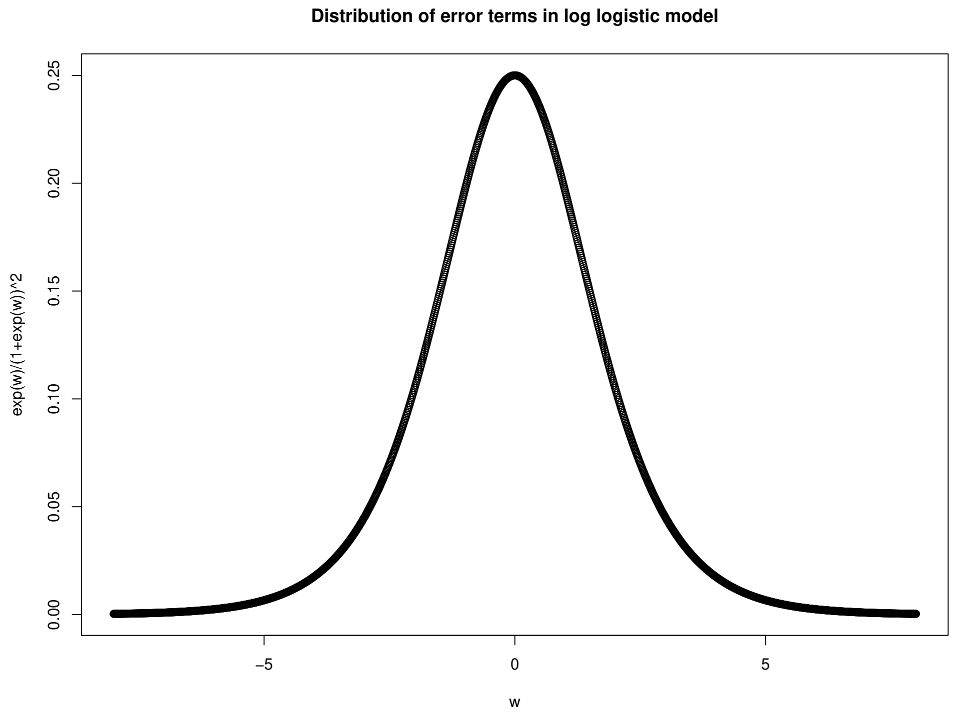
The log logistic model

Model formulation. Let Y denote the natural logarithm of the time-to-event response variable T. The log logistic model has the form

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i} + \sigma W_i, \qquad (10)$$

where σ is an estimable positive constant and W_1 through W_n are independent and identically distributed error terms.

The difference between the Weibull model and the log logistic model is that now the distribution of W_1 through W_n is described by the density function $\exp[w]/(1+\exp[w])^2$, which I have displayed in the figure below. This density function is more symmetric than the one for the Weibull model.



Implied survival function. We have

$$P(W > w) = \int_w^\infty \exp[u]/(1+\exp[u])^2\, du = 1/(1+\exp[w]), \tag{11}$$

whence

$$\begin{aligned}
S_{x_{1,i},\ldots,x_{k,i}}(t) &:= P(T_i > t \mid X_1 = x_{1,i}, \ldots, X_k = x_{k,i}) \\
&= P(Y_i > \log t \mid X_1 = x_{1,i}, \ldots, X_k = x_{k,i}) \\
&= P(W_i > \sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\} \mid X_1 = x_{1,i}, \ldots, X_k = x_{k,i}) \\
&= 1/(1+\exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]) \qquad (12) \\
&= 1/(1+\exp[\sigma^{-1}\{\log t - \log[\exp\{\boldsymbol{\beta}\cdot\mathbf{x}_i\}]\}]) \\
&= 1/(1+\exp[\sigma^{-1}\{\log[t\exp\{-\boldsymbol{\beta}\cdot\mathbf{x}_i\}]\}]) \\
&= S_0(t\exp\{-\boldsymbol{\beta}\cdot\mathbf{x}_i\}). \qquad (13)
\end{aligned}$$

Line (12) provides a tractable mathematical expression for the survival function $S_{x_{1,i},\ldots,x_{k,i}}(t)$, while line (13) displays the accelerated failure time property.

Implied hazard function. Starting from line (12), we can use differential calculus to recover the hazard function for a person with $X_1 = x_{1,i}, \ldots, X_k = x_{k,i}$ as

$$\begin{aligned}
h_{x_{1,i},\ldots,x_{k,i}}(t) &= -\frac{d}{dt}\log S_{x_{1,i},\ldots,x_{k,i}}(t) \\
&= \frac{d}{dt}\log\{1+\exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]\} \\
&= \frac{\sigma^{-1}t^{-1}\exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]}{1+\exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]} \\
&= \frac{\sigma^{-1}t^{\sigma^{-1}-1}\exp[-\sigma^{-1}\boldsymbol{\beta}\cdot\mathbf{x}_i]}{1+t^{\sigma^{-1}}\exp[-\sigma^{-1}\boldsymbol{\beta}\cdot\mathbf{x}_i]}. \qquad (14)
\end{aligned}$$

Line (14) provides a tractable mathematical expression for the hazard function $h_{x_{1,i},\dots,x_{k,i}}(t)$. Although line (14) is only slightly different from line (8), the difference is enough that the proportional hazards property is lost.

However, while the log logistic model lacks the proportional hazards property, it does have a "proportional odds" property.

The proportional odds property. Consider the quantity

$$\frac{S_{x_{1,i},\dots,x_{k,i}}(t)}{1 - S_{x_{1,i},\dots,x_{k,i}}(t)}, \tag{15}$$

which represents the odds of survival past time t. By line (12) we have

$$\begin{aligned} 1 - S_{x_{1,i},\dots,x_{k,i}}(t) &= 1 - \frac{1}{1 + \exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]} \\ &= \frac{\exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]}{1 + \exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]}. \end{aligned} \tag{16}$$

Hence, expression (15) can be written as

$$\begin{aligned} \frac{S_{x_{1,i},\dots,x_{k,i}}(t)}{1 - S_{x_{1,i},\dots,x_{k,i}}(t)} &= \frac{1}{1 + \exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]} \Big/ \frac{\exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]}{1 + \exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}]} \\ &= 1/\exp[\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}] \\ &= \exp[-\sigma^{-1}\{\log t - \boldsymbol{\beta}\cdot\mathbf{x}_i\}] \\ &= t^{-\sigma^{-1}}\exp[\sigma^{-1}\boldsymbol{\beta}\cdot\mathbf{x}_i] \\ &= \frac{S_0(t)}{1 - S_0(t)}\exp[\sigma^{-1}\boldsymbol{\beta}\cdot\mathbf{x}_i]. \end{aligned} \tag{17}$$

Example with larynx cancer data. Pages 4 and 5 of {LARYNXExamples.rtf} show SAS output for the log logistic model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \beta_3 x_{3,i} + \sigma W_i, \tag{18}$$

where Y is the natural logarithm of TimetoDeath and X_1 through X_3 are indicators for categories 1 to 3 of Stage.

The Type III Analysis of Effects box shows that $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ may be rejected on the basis of a 0.0002 p-value. The Analysis of Parameter Estimates box shows that all three individual null hypotheses $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$, and $H_0 : \beta_3 = 0$ may also be rejected with respective p-values < 0.0001, 0.0007, and 0.0218. The estimate of σ is 0.7141.

Page 6 shows a plot of the estimated survival functions for all four categories of Stage. This plot is similar to the one on page 3, although the prospects for Stage 4 patients appear slightly less grim with the log logistic model.

In fact, the estimated probability of surviving past four years is

$$1/(1 + \exp[0.7141^{-1}(\log 4 - 0.3275)]) = 0.185$$

with the log logistic model, compared to

$$\exp[-\exp[0.8846^{-1}(\log 4 - 0.7906)]] = 0.141$$

with the Weibull model.

Note that the ease with which such estimated probabilities are calculated is a consequence of using a parametric rather than a semiparametric model.

Finally, we can ask which parametric model is better for the larynx cancer data set: the Weibull model or the log logistic model? The Weibull model yields a log likelihood of -109.01, while the log logistic model yields a log likelihood of -108.81. With the number of parameters fixed a higher likelihood is better, so a higher log likelihood is also better. However, -108.81 does not seem importantly different from -109.01.