Regression Models for Survival Analysis

When considering several distributions that are commonly used for time-to-event data, we saw that log transformations often lead to location-scale family distributions. We saw that, for a continuous random variable T, we could often write

$$Y = \log T = \mu + \sigma W,$$

where W was a mean-zero, unit-variance error distribution, and μ represented the mean of the transformed variable. Depending on the distribution imposed on T, we would see different error distributions.

A good motivating example is taking T to be log-normal. This means that $Y = \log T \sim N(\mu, \sigma^2)$, and be extension in this case we would have $W \sim N(0, 1)$. For a specific observation, T_i, we could write this more succinctly as

$$Y_i = \log T_i = \eta_i + W_i,$$

where $W_i \sim N(0, \sigma^2)$ and η_i represents the mean of Y_i. This framing is reminiscent of a standard linear regression models, where we may set $\eta_i = x_i'\beta$, for some parameter vector β, taking it to be a standard linear predictor. In this sense we have specified a regression model for the log transform of our time-to-event data that arises naturally from distributions that are sensible for time-to-event data. This gives rise to **accelerated failure time models**.

Accelerated Failure Time (AFT) Models

An AFT model is characterized by suggesting that $Y_i = \log T_i = \eta_i + W_i$, where η_i is taken to be a linear predictor, and W_i are iid according to some mean-zero error distribution, with constant variance. This represents a class of regression models that naturally arises from our previous consideration of location-scale families. In the event that data are not censored, an AFT model could be fit using standard OLS regression techniques. What's more, while we motivated the use of these methods partly through a consideration of normal errors (log-normal survival times), this assumption tends to be appropriate in only very specific situations.

With any AFT, we can consider transforming back to the original time scale. Doing this, we see that

$$\begin{aligned} \log T &= \eta + W \\ \implies T &= e^{\eta} e^{W} \\ &= e^{\eta} T_0, \end{aligned}$$

where $T_0 = e^W$. This suggests the motivation for naming the models accelerated failure time. Covariates in an AFT act multiplicatively on the time. If $e^{\eta} = 0.5$, then this is like saying that the subject ages twice as fast as normal, where if $e^{\eta} = 2$ they age half as fast as normal.

If we take $f_0(t)$, $h_0(t)$, and $S_0(t)$ to be the density, hazard, and survivor functions for $T_0 = e^W$, then we can work out the implied density, survival, and hazard functions for T. In particular, by noting that

$$\begin{aligned} T_0 &= e^{-\eta}T \\ \frac{\partial}{\partial T}T_0 &= e^{-\eta} \\ \implies f(t) &== f_0(e^{-\eta}t)e^{-\eta} \\ S(t) &= S_0(e^{-\eta}t) \\ h(t) &= h_0(e^{-\eta}t)e^{-\eta}. \end{aligned}$$

This becomes quite a useful setup, particular because of its direct relation to regression setups, and the interpretability of the regression parameters. Thinking of covariates as scaling the survival time from some baseline time is a fairly useful formulation in terms of scientific inquiry. While AFT models are, I think, a fairly intuitive approach to survival analysis, they are not nearly as common as the use of **proportional hazards models**.

Proportional Hazards (PH) Models

A PH model treats the hazard as the key quantity to model. In particular, we state that for a particular individual i, the hazard for this individual is given by

$$h_i(t) = h_0(t)\exp(x_i'\beta).$$

Here $h_0(t)$ is a baseline hazard, equal to the hazard for an individual with $x_i = 0$, and note that x_i will not contain an intercept in this framing (if it did the term would get absorbed into the baseline). In this sense, instead of covariates multiplying or scaling the time, they multiply or scale the hazard itself.

If we take two individuals who are identical except for covariate j, then we can consider their hazard ratio, under the assumption of a proportional hazards model.

$$\frac{h_i(t|x_{ij} = x+1)}{h_{i'}(t|x_{i'j} = x)} = \frac{h_0(t)\exp(x_i'\beta)}{h_0(t)\exp(x_{i'}'\beta)} = \exp(\beta_j).$$

Generally, any relevant distribution can be selected for the baseline hazard function; most commonly this will be either an exponential, Weibull, or log-logistic hazard. Once this has been specified, we have a completely parametric model that can be fit using standard likelihood theory (under assumptions of independent and uninformative censoring). Recall

$$\begin{aligned} L \propto \prod_{i=1}^n h_i(t_i)^{\delta_i}S(t_i) \implies \ell &= \sum_{i=1}^n \{\delta_i \log h_i(t_i) + \log S(t_i)\} \\ &= \sum_{i=1}^n \left\{\delta_i \log h_i(t_i) + \log\left(\exp\left[-\int_0^t h(s)ds\right]\right)\right\} \\ &= \sum_{i=1}^n \left\{\delta_i \log h_i(t_i) - \int_0^{t_i} h_i(s)ds\right\}. \end{aligned}$$

From this relation, inserting $h_i(t) = h_0(t)\exp(x_i'\beta)$ provides an expression that can be optimized numerically for β. These optimization procedures have been programmed in most common statistical languages, and are widely used. It is worth noting that this imposes a fairly rigid form on the hazard which may not always be appropriate. It is sometimes helpful to specify a weakly parametric baseline hazard, to add flexibility to our modelling. An easy way to do this is to suppose that the baseline hazard is piecewise constant. Recall that we have seen that an exponential distribution has a constant hazard function, and so to specify something as a piecewise constant hazard function implies that we specify separate exponential hazards over different ranges.

Formally, define $0 = a_0 < a_1 < a_2 < \cdots < a_{K-1} < a_K = \infty$ as cut points on the real line. Then, we can specify a hazard function of the form

$$h_0(t) = h_{0k} \quad a_{k-1} \leq t < a_k,$$

for $k = 1, \dots, K$. This is unlikely to capture the true hazard function – jumps are not typically present in real world data – but it presents a very flexible model that can accommodate some fairly irregular shapes.

Weibull Regression

In general, AFT and PH models are irreconcilable. That is, you can make the assumption that your data follow the PH assumption, which rules out AFT models, or you can make the assumption that AFT models capture your data well, which rules out PH models. The exception to this rule is Weibull regression.

Suppose that you specify $h_0(t) = \kappa\rho^{-\kappa}t^{\kappa-1}$, which is the hazard function for a Weibull(ρ, κ) distribution, in a PH model. This renders

$$\begin{aligned} h_i(t) &= h_0(t)\exp(x_i'\beta) \\ &= \kappa\rho^{-\kappa}t^{\kappa-1}\exp(\eta_i) \\ &= \kappa\left(\rho\exp\left\{-\frac{\eta_i}{\kappa}\right\}\right)^{-\kappa} t^{\kappa-1} \\ &= \kappa\lambda^{-\kappa}t^{\kappa-1}. \end{aligned}$$

This is the hazard for a Weibull distribution with scale parameter $\lambda = \rho\exp\left\{-\frac{\eta_i}{\kappa}\right\}$ and shape parameter κ. If we then consider this Weibull distribution as the distribution for $T_i|x_i$, we can translate this to the corresponding accelerated failure time model. In particular

$$\begin{aligned} Y_i = \log T_i &= \mu_i + \sigma W_i \\ &= \log\lambda + \frac{1}{\kappa}W_i \\ &= \log\left(\rho\exp\left\{-\frac{\eta_i}{\kappa}\right\}\right) + \frac{1}{\kappa}W_i \\ &= -\frac{\eta_i}{\kappa} + \log\rho + \frac{1}{\kappa}W_i \\ &= \beta_0 + x_i'\beta^* + W_i^*, \end{aligned}$$

where W_i^* absorbs the constant variance term. As a result, if we were to fit this AFT style model, we would find that $\beta_0 = \log\rho$, and then $x_i'\beta^*$ will be related to the coefficients from the PH model (β) through

$$\beta = -\kappa\beta^*.$$

Put differently, taking a Weibull distribution in a PH model will have coefficients estimated as β; using a Weibull distribution for an AFT model will produce coefficients estimated as $-\kappa\beta$.