Weibull Regression[1]

STA312 Spring 2019

[1] See last slide for copyright information.

Background Reading

Section 10.6 in the text, but it refers to a lot of things we have not covered yet.

Overview

1. Exponential

2. Weibull

A multiplicative regression model

Exponential model, just one explanatory variable

Independently for $i = 1, \ldots n$,

$$t_i = e^{\beta_0 + \beta_1 x_i} \times \epsilon_i$$

where

β_0 and β_1 are unknown constants (parameters).

$x_1, \ldots, x_n$ are known, observed constants.

$\epsilon_1, \ldots, \epsilon_n$ are independent exponential(1) random variables.

$t_1, \ldots, t_n$ are observed failure times.

$\delta_1, \ldots, \delta_n$ are indicators for uncensored.

- These are sometimes called *accelerated failure time* models.
- Because the effect of $x \neq 0$ is to *multiply* the failure time by a constant.

Distribution of $t_i = e^{\beta_0 + \beta_1 x_i} \times \epsilon_i$, with ϵ_i exponential(1)

- If $\epsilon \sim \exp(1)$ and $a > 0$, $x = a\epsilon$ is also exponential.
- Expected value a (or $\lambda = 1/a$).
- Thus, $E(t_i) = e^{\beta_0 + \beta_1 x_i} \Leftrightarrow \log E(t_i) = \beta_0 + \beta_1 x_i$.
- We are adopting a linear model for the log of the expected value.
- Or, we can transform the failure times by taking logs.

$$\begin{aligned} \log t_i &= \beta_0 + \beta_1 x_i + \log \epsilon_i \\ &= \beta_0 + \beta_1 x_i + \epsilon_i^* \end{aligned}$$

where $\epsilon_i^* = \log \epsilon_i \sim G(0, 1)$.

Meaning of β_1

With $E(t_i) = e^{\beta_0+\beta_1 x_i}$

- Increase x_i by one unit.
- The effect is to multiply $E(t_i)$ by a constant.

$$
\begin{aligned}
e^{\beta_0+\beta_1(x_i+1)} &= c\, e^{\beta_0+\beta_1 x_i} \\
\Leftrightarrow c &= \frac{e^{\beta_0+\beta_1(x_i+1)}}{e^{\beta_0+\beta_1 x_i}} \\
&= \frac{e^{\beta_0+\beta_1 x_i+\beta_1}}{e^{\beta_0+\beta_1 x_i}} \\
&= e^{\beta_1}
\end{aligned}
$$

- So when x_i is increased by one unit, $E(t_i)$ is multiplied by e^{β_1}.
- If $\beta_1 > 0$, $E(t_i)$ goes up.
- If $\beta_1 < 0$, $E(t_i)$ goes down.

Natural extensions

- More than one explanatory variable.
- Centering the quantitative explanatory variables.

$$t_i = \exp\{\beta_0 + \beta_1(x_{i,1} - \bar{x}_1) + \ldots + \beta_{p-1}(x_{i,p-1} - \bar{x}_{p-1})\} \cdot \epsilon_i$$

- In this case, e^{β_0} is the expected failure time for average values of all the explanatory variables.
- If there are dummy variables, center only the quantitative variables (covariates).

Equivalent model on the log scale

Starting with $t_i = \exp\{\beta_0 + \beta_1(x_{i,1} - \bar{x}_1) + \ldots + \beta_{p-1}(x_{i,p-1} - \bar{x}_{p-1})\} \cdot \epsilon_i$

$$\begin{aligned} \log t_i &= \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \log \epsilon_i \\ &= \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i^* \\ &= \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i^*, \end{aligned}$$

where $\epsilon_i^* \sim G(0,1)$.

- Recall, if $Z \sim G(0,1)$, then $\sigma Z + \mu \sim G(\mu, \sigma)$.
- So the model says $\log t_i \sim G(\mathbf{x}_i^\top \boldsymbol{\beta}, 1)$
- Why should the variance of log survival time be $\frac{\pi^2}{6}$?
- Much more reasonable is
  $\log t_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \sigma \epsilon_i^*$
- In this case, $\log t_i \sim G(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma)$.

Switching back to the time scale

From the log time scale

$$
\begin{aligned}
& \log t_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \sigma \epsilon_i^* \\
\Leftrightarrow \quad & t_i = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \, e^{\sigma \epsilon_i^*} = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \, e^{\sigma \log \epsilon_i} = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \, e^{\log(\epsilon_i^\sigma)} \\
\Leftrightarrow \quad & t_i = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \, \epsilon_i^\sigma
\end{aligned}
$$

We have arrived at the multiplicative regression model:

$$
t_i = \exp\{\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}\} \cdot \epsilon_i^\sigma
$$

$t_i = \exp\{\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1}\} \cdot \epsilon_i^\sigma$

- It's an accelerated failure time model. Changing one of the x values multiplies t_i by something.
- In particular, increase $x_{i,k}$ by one unit while holding all other $x_{i,j}$ values constant.
- Then t_i is multiplied by e^{β_k}.
- Holding $x_{i,j}$ values constant is the meaning of "controlling" for explanatory variables in Weibull regression.
- Note that if β_k is negative, $e^{\beta_k} < 1$ and t_i goes down.
- Call it a "negative relationship" (controlling for the other variables).
- If β_k is positive, $e^{\beta_k} > 1$ and t_i goes up.
- Call this a "positive relationship" (controlling for the other variables).

Distribution of t_i

Recall

- We have established that $\log t_i \sim G(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma)$.
- Exponential function of Gumbel(μ, σ) is Weibull(α, λ) with $\lambda = e^{-\mu}$ and $\alpha = 1/\sigma$.
- Note that here, $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta}$.
- So, t_i is Weibull, with $\lambda_i = e^{-\mathbf{x}_i^\top \boldsymbol{\beta}}$ and $\alpha = 1/\sigma$.
- This means

$$
\begin{aligned}
E(T_i) &= \frac{\Gamma(1+\frac{1}{\alpha})}{\lambda} = e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\Gamma(1+\sigma) \\
\text{Median}(T_i) &= \frac{[\log(2)]^{1/\alpha}}{\lambda} = e^{\mathbf{x}_i^\top \boldsymbol{\beta}} \log(2)^\sigma \\
h(t) &= \alpha\lambda^\alpha t^{\alpha-1} = \frac{1}{\sigma}\exp\{-\frac{1}{\sigma}\mathbf{x}^\top \boldsymbol{\beta}\} t^{\frac{1}{\sigma}-1}
\end{aligned}
$$

Conclusions

Following from $\log t_i \sim G(\mathbf{x}_i^\top \boldsymbol{\beta}, \sigma)$

$$
\begin{aligned}
E(T_i) &= e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\,\Gamma(1+\sigma) \\
\text{Median}(T_i) &= e^{\mathbf{x}_i^\top \boldsymbol{\beta}}\,\log(2)^{\sigma} \\
h(t) &= \frac{1}{\sigma}\exp\{-\frac{1}{\sigma}\mathbf{x}^\top \boldsymbol{\beta}\} t^{\frac{1}{\sigma}-1}
\end{aligned}
$$

- Increasing value of x_j by c units multiplies the mean and median by $e^{c\beta_j}$.
- Same effect on the hazard function.
- Remarkable because the hazard function is a function of time t.
- And the effect is the same for every value of t.

Proportional Hazards

$h(t) = \frac{1}{\sigma} \exp\{-\frac{1}{\sigma}\mathbf{x}^\top \boldsymbol{\beta}\} t^{\frac{1}{\sigma}-1}$

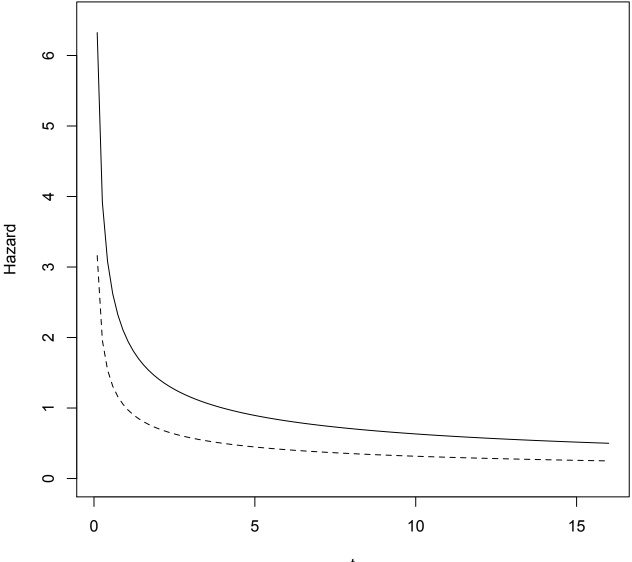
- Suppose two individuals have different $\mathbf{x}$ vectors of explanatory variable values.
- They have different hazard functions because their λ values are different.
- Look at the *ratio*:

$$\begin{aligned} \frac{h_1(t)}{h_2(t)} &= \frac{\frac{1}{\sigma} \exp\{-\frac{1}{\sigma}\mathbf{x}_1^\top \boldsymbol{\beta}\} t^{\frac{1}{\sigma}-1}}{\frac{1}{\sigma} \exp\{-\frac{1}{\sigma}\mathbf{x}_2^\top \boldsymbol{\beta}\} t^{\frac{1}{\sigma}-1}} \\ &= \frac{\exp\{-\frac{1}{\sigma}\mathbf{x}_1^\top \boldsymbol{\beta}\}}{\exp\{-\frac{1}{\sigma}\mathbf{x}_2^\top \boldsymbol{\beta}\}} \\ &= \exp\{\frac{1}{\sigma}(\mathbf{x}_2 - \mathbf{x}_1)^\top \boldsymbol{\beta}\} \end{aligned}$$

The point is that $h_1(t)$ and $h_2(t)$ are always in the same proportion for every value of t.

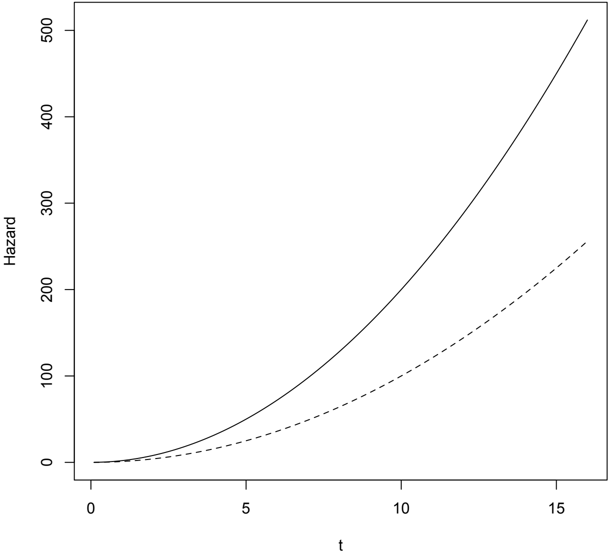
Proportional Hazards

$h_1(t) = 2\,h_2(t)$ with $\sigma = 2$



Proportional Hazards

$h_1(t) = 2\,h_2(t)$ with $\sigma = 1/3$



Copyright Information