Chapter 6

(AST405) Lifetime data analysis

Md Rasel Biswas

Lecture Outline

1. 6. Parametric Regression Models

Section 1

6. Parametric Regression Models

Subsection 1

6.1 Log-location-scale (Accelerated Failure Time) Regression Models

Linear regression model

- Distributional assumption for the response

$$(Y \mid x) = Y(x) \sim \mathcal{N}(\mu(x), \sigma^2)$$

- Regression model for the parameters

$$\begin{aligned} \mu(x) &= \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}'\beta \\ \mathrm{var}(Y \mid x) &= \sigma^2 \end{aligned}$$

Linear regression model

- Instead of the parameters, linear regression model can be defined in terms of other functions, such as survivor function

$$
\begin{aligned}
S_Y(y) &= Pr(Y > y) \\
&= 1 - \Phi\left(\frac{y - \mu(x)}{\sigma}\right)
\end{aligned}
$$

Regression models for lifetimes

- Similar to continuous and binary responses, regression analysis of lifetimes involves specifications for the distribution of a lifetime (T) given a vector of p-dimensional (say) covariate $\mathbf{x}$

$$(T \mid \mathbf{x}) = T(\mathbf{x})$$

Regression models for lifetimes

- For parametric regression models for lifetimes T, parameters (e.g. scale and shape parameters) need to be defined as a function of measured covariates (linear predictors)
- It requires selecting a link function (e.g. identity, log, logit, etc.) for relating model parameters with linear predictors
- Similar to linear and logistic regression models, maximum likelihood method of estimation is used to estimate parameters of the model

Log-location-scale AFT model

- For a lifetime that follows a distribution of the log-location-scale family of distributions, the survivor function of lifetime T for a given covariate vector $\mathbf{x}$ is defined as

$$S(t \mid \mathbf{x}) = S_0^\star\Big([t/\alpha(\mathbf{x})]^\delta\Big) \tag{1}$$

  - Scale parameter $\alpha(\mathbf{x})$ is defined as a function of covariate vector $\mathbf{x}$
  - Shape parameter δ does not depend on $\mathbf{x}$
  - Survivor function of the corresponding standardized distribution $S_0^\star(x) = S_0(\log x)$ is defined earlier

Log-location-scale AFT model

- For a log-lifetime that follows a distribution of the location-scale family of distribution, the survivor function of log-lifetime Y for a given covariate vector $\mathbf{x}$ is defined as

$$S(y \,|\, \mathbf{x}) = S_0\left(\frac{y - u(\mathbf{x})}{b}\right) \tag{2}$$

  - Location parameter $u(\mathbf{x})$ is defined as a function covariate vector $\mathbf{x}$
  - Scale parameter b does not depend on $\mathbf{x}$

- The model (Equation 2) for log-lifetime is similar to the linear regression model (Equation 1) with

$$\mu(\mathbf{x}) = u(\mathbf{x}), \sigma = b, \text{ and } \Phi(x) = 1 - S_0(x)$$

Log-location-scale AFT model

- The model for lifetime (Equation 1) or log-lifetime (Equation 2) is known as **accelerated failure time (AFT) model**

$$S(t \mid \mathbf{x}) = S_0^\star\left([t/\alpha(\mathbf{x})]^\delta\right)$$

$$S(y \mid \mathbf{x}) = S_0\left(\frac{y - u(\mathbf{x})}{b}\right)$$

- Models for the parameters $\alpha(\mathbf{x})$ and $u(\mathbf{x})$ are defined so that associated parametric restrictions are satisfied, $\alpha(\mathbf{x}) > 0$ and $-\infty < u(\mathbf{x}) < \infty$, e.g.

$$u(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}'\beta$$

$$\alpha(\mathbf{x}) = \exp\left(\mathbf{x}'\beta\right)$$

Log-location-scale AFT model

- AFT model can also be expressed as

$$\frac{Y - u(\mathbf{x})}{b} = Z \quad \Rightarrow \quad Y = u(\mathbf{x}) + bZ \tag{3}$$

  - $Z \sim S_0(z)$, i.e. Z follows a standardized log-location-scale distribution, e.g. standard normal or extreme-value distributions with location 0 and scale 1, etc.

- Linear regression model (Equation 1) can also be expressed as Equation 3:

$$Y = \mu(\mathbf{x}) + \sigma Z, \quad Z \sim \mathcal{N}(0, 1)$$

Log-location-scale AFT model

- In AFT model defined in terms of the distribution of lifetime T, covariates alter the time scale
  - If $\alpha(\mathbf{x}) = \exp\left(\mathbf{x}'\beta\right) > 1$, the effect of covariate vector is to increase time (decelerate time)
  - If $\alpha(\mathbf{x}) = \exp\left(\mathbf{x}'\beta\right) < 1$, the effect of covariate vector is to shorten time (accelerate time)

Log-location-scale AFT model

- The accelerated failure time model is a general model for survival data, in which explanatory variables measured on an individual are assumed to act multiplicatively on the time-scale
- **Log-location-scale AFT** models are a special case of AFT models where the log of survival time follows a location-scale distribution.
- AFT models assume that covariates accelerate or decelerate the time to event.

Log-location-scale AFT model

- The following example is described in Collett (2015)
- Suppose patients are randomized to receive one of the two treatments A (standard) and B (new)
- Under an accelerated failure time model, the survival time of an individual on the new treatment is taken to be a multiple of the survival time for an individual on the standard treatment.
- Thus, the effect of the new treatment is to “speed up” or “slow down” the passage of time

Log-location-scale AFT model

- For a specific time t

$$S(t \mid \text{trt} = B) = S(t\alpha \mid \text{trt} = A)$$

- One interpretation of this model is that the lifetime of an individual on the new treatment (B) is α times the lifetime that the individual would have experienced under the standard treatment (A)
- When the end-point of concern is the *death of a patient*
  - $\alpha > 1$ new treatment is promoting longevity
  - $\alpha < 1$ new treatment is worse (accelerating death)
- The quantity α is therefore termed the **acceleration factor**

Log-location-scale AFT model

- The acceleration factor can also be interpreted in terms of the **median survival times** of patients on the new and standard treatments, $t_A(50)$ and $t_B(50)$

$$S_B\{t_B(50)\} = S_A\{t_A(50)\} = 0.50$$

- Under AFT model

$$S_B\{t_B(50)\} = S_A\{t_B(50)/\alpha\} \;\Rightarrow\; t_B(50) = \alpha t_A(50)$$

- Under the AFT model, the median survival time of a patient on the new treatment is α times that of a patient on the standard treatment

Log-location-scale AFT model

- Under AFT model, the survivor functions with covariate vectors $\mathbf{x}_1$ and $\mathbf{x}_2$ can be compared as

$$S(t \,|\, \mathbf{x}_1) = S(c\, t \,|\, \mathbf{x}_2)$$

  - If $c > 1$, subjects with covariate $\mathbf{x}_2$ survives *longer* compared to subjects with covariate vector $\mathbf{x}_1$
  - If $c < 1$ subjects with covariate $\mathbf{x}_2$ survives *shorter* compared to subjects with covariate vector $\mathbf{x}_1$

Log-location-scale AFT model

- Under AFT model, $S_1(t) = S_2(ct)$ for $c > 0$, we can express the mean survival time μ_2 of Population 2 can be expressed in terms of μ_1, mean survival time of Population 1 as

$$\begin{aligned}
\mu_2 &= \int_0^\infty S_2(t)\, dt \\
&= c\int_0^\infty S_2(cu)\, du \\
&= c\int_0^\infty S_1(u)\, du \\
&= c\mu_1
\end{aligned}$$

Log-location-scale AFT model

- In general, let φ is a population quantity such that $S(\varphi) = \theta$ for some $\theta \in (0, 1)$ and

$$S_2(\varphi_2) = \theta = S_1(\varphi_1) = S_2(c\varphi_1)$$

- Then $\varphi_2 = c\varphi_1$, i.e., under the AFT model, the expected survival time, median survival time of population 2 all are c times as much as those of population 1

Log-location-scale AFT model

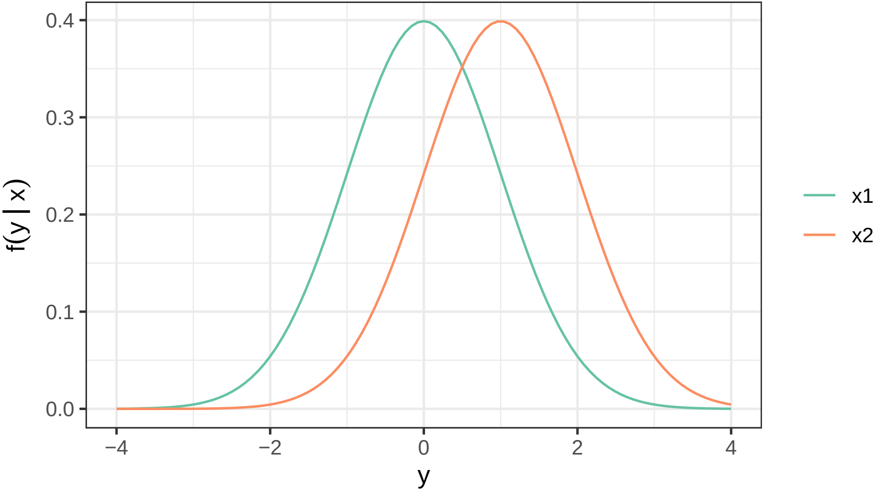


Figure 1: Comparison between two log-location density functions with covariate vectors x_1 and x_2, where $u(x_2) > u(x_1)$

Log-location-scale AFT model

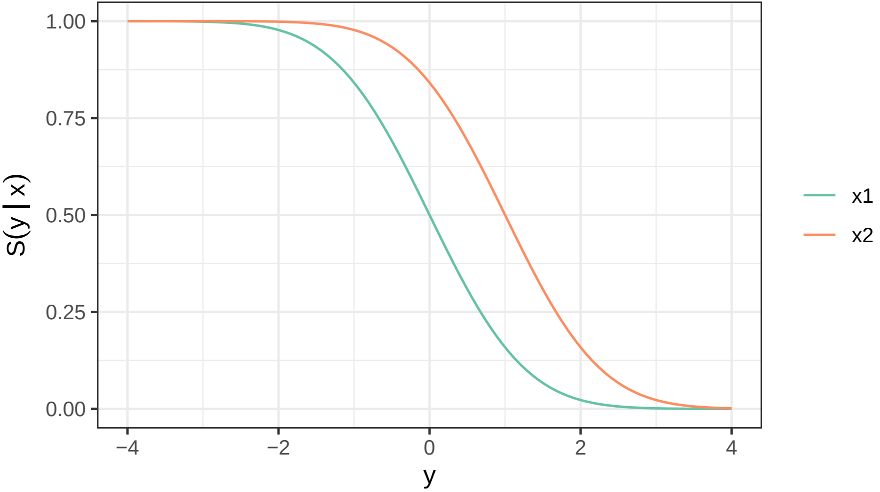


Figure 2: Comparison between two log-location survival functions with covariate vectors x_1 and x_2, where $u(x_2) > u(x_1)$

Log-location-scale AFT model

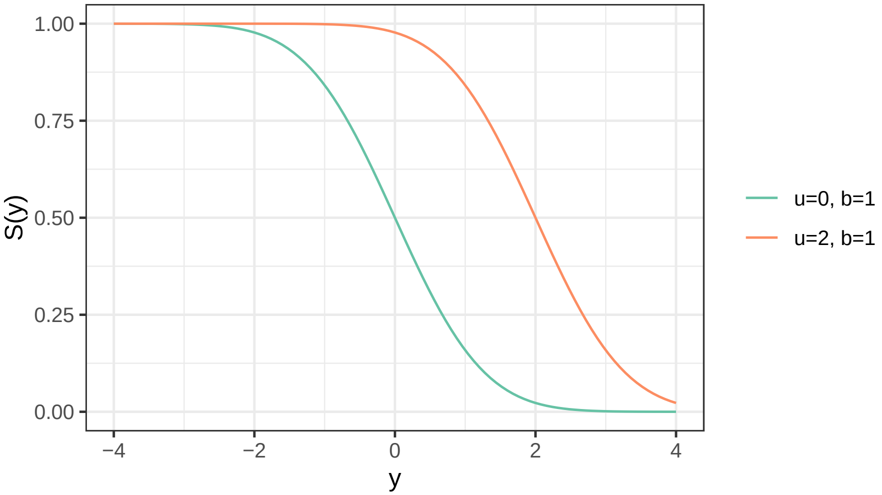


Figure 3: Comparison between two log-location survival functions with the same scale parameters, but different location parameters

Log-location-scale AFT model

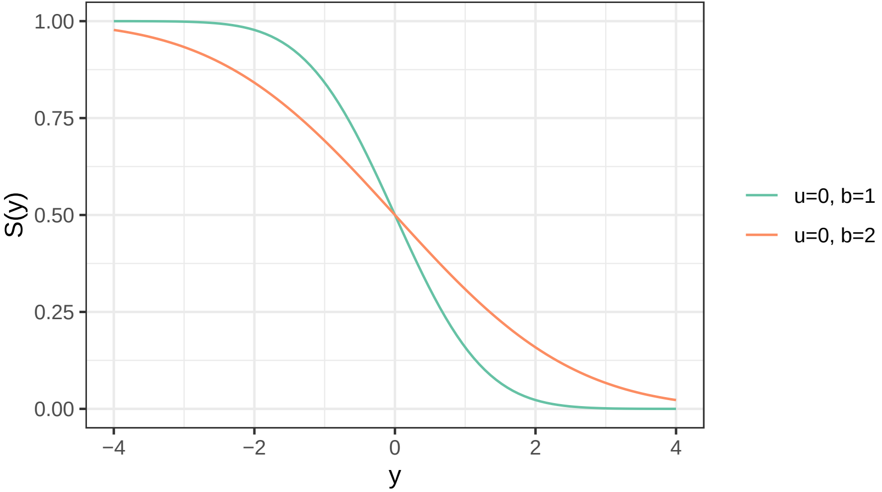


Figure 4: Comparison between two log-location survival functions with the same location parameters, but different scale parameters

Proportional hazards model

- There are two approaches to regression modeling for lifetimes
  1. AFT model, where the effects of covariates are assessed by comparing corresponsing time scales
  2. Hazards model, where effects of covariates on the hazard function are studied

Proportional hazards model

- The most common hazards model is the *proportional hazards model* (Cox 1972), where hazard function for lifetime T given $\mathbf{x}$ is defined as

$$h(t \mid \mathbf{x}) = h_0(t)\, r(\mathbf{x})$$

  - $r(\mathbf{x}) \rightarrow$ a positive-valued function of linear predictor, e.g. $r(\mathbf{x}) = \exp(\mathbf{x}'\beta)$, which does not include the intercept term
  - $h_0(t) \rightarrow$ a positive-valued function, which is known as **baseline hazards function**, i.e. $h(t \mid \mathbf{x} = \mathbf{0}) = h_0(t)$
  - $h_0(t)$ could be either fully parametric or unspecified

Proportional hazards model

If you take two individuals with covariates x_1 and x_2:

$$\frac{h(t|x_1)}{h(t|x_2)} = \frac{h_0(t)e^{\beta x_1}}{h_0(t)e^{\beta x_2}} = e^{\beta(x_1 - x_2)}$$

This ratio **does not depend on time (t)**, this is exactly the **proportional hazards property**.

Proportional hazards model

$$h(t \,|\, x) = h_0(t)\; e^{x\beta}$$

- For a binary predictor x (1=male, 0=female), the hazard ratio can be defined as

$$\frac{h(t \,|\, x=1)}{h(t \,|\, x=0)} = \frac{h_0(t)e^{\beta}}{h_0(t)} = e^{\beta}$$
$$h(t \,|\, x=1) = h(t \,|\, x=0)e^{\beta}$$

- $\beta > 0 \;\Rightarrow$ Hazard of the event is higher for male compared to female

Proportional hazards model

- Under proportional hazards model, the cumulative hazard function is defined as

$$
\begin{aligned}
H(t \,|\, \mathbf{x}) &= \int_0^t h(u \,|\, \mathbf{x}) du \\
&= r(\mathbf{x}) \int_0^t h_0(u) du \\
&= r(\mathbf{x}) H_0(t)
\end{aligned}
$$

Proportional hazards model

- Under proportional hazards model, the survivor function is defined as

$$S(t \mid \mathbf{x}) = e^{-H(t \mid \mathbf{x})} = e^{-r(\mathbf{x})H_0(t)} = \Big[S_0(t)\Big]^{r(\mathbf{x})}$$

  - $S_0(t) \rightarrow$ baseline survivor function and $r(\mathbf{x}) > 0$
  - Interpret the survival probabilities for the following cases

$$(a)\ r(\mathbf{x}) > 1 \quad \text{and} \quad (b)\ r(\mathbf{x}) < 1$$

Proportional hazards model

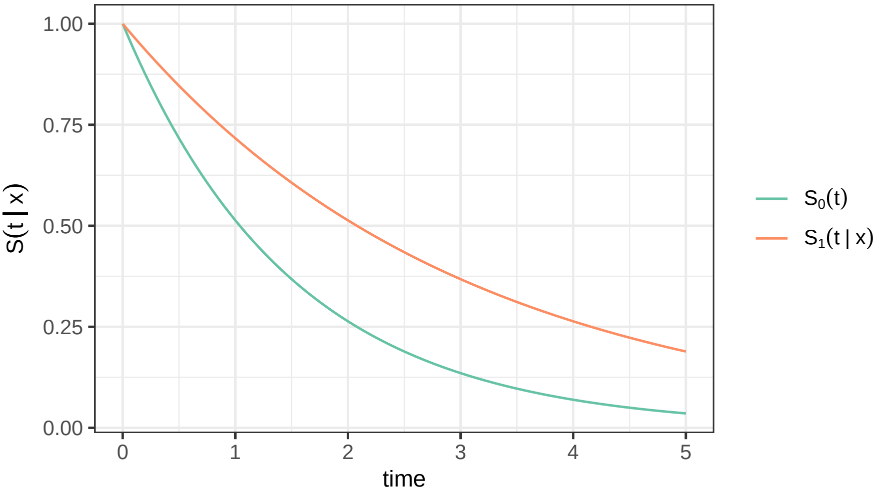


Figure 5: Under proporitonal hazards model, comparison between baseline survivor function $S_0(t)$ and $S_1(t \mid x) = [S_0(t)]^{0.5}$

Proportional hazards model

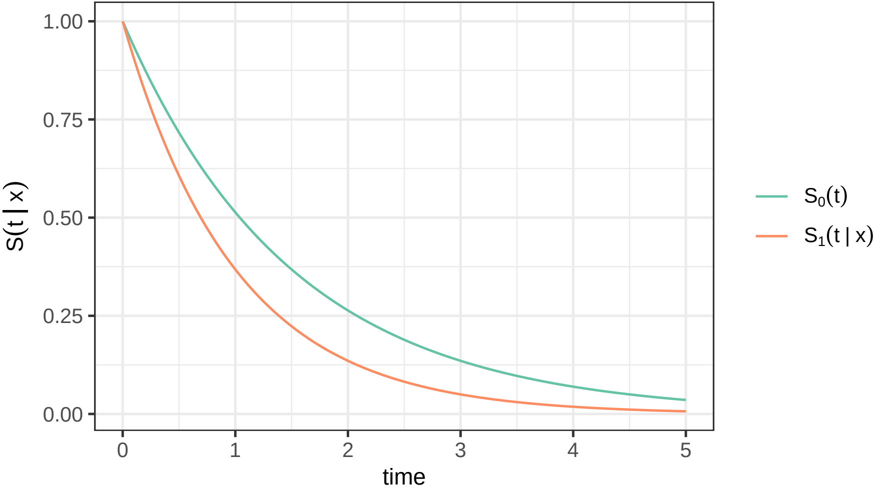


Figure 6: Under proporitonal hazards model, comparison between baseline survivor function $S_0(t)$ and $S_1(t \mid x) = [S_0(t)]^{1.5}$

Proportional hazards model

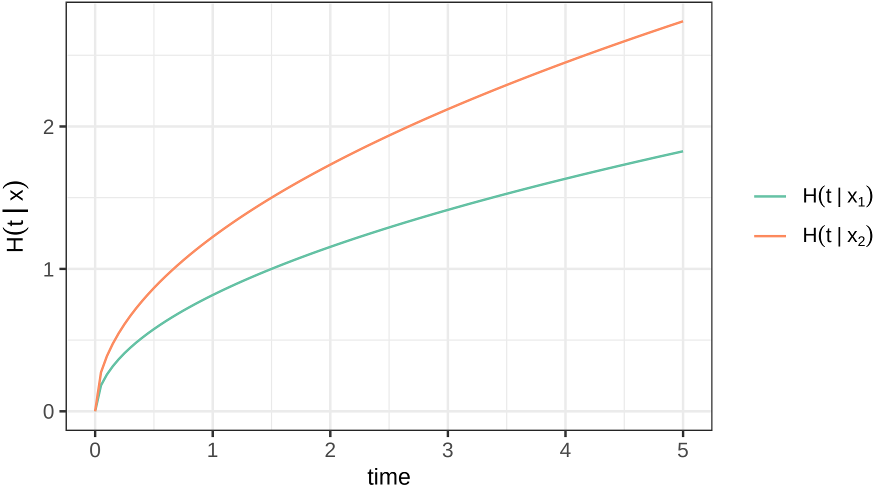


Figure 7: Under proporitonal hazards model, comparison between hazard functions $H(t \,|\, x_1)$ and $H(t \,|\, x_2) = 1.5H(t \,|\, x_1)$

Parametric proportional hazards model

- Depending on whether the baseline hazard function $h_0(t)$ is fully parametric or not, a PH model $h(t \,|\, \mathbf{x}) = h_0(t) r(\mathbf{x})$ could be either parametric or semi-parametric
  - PH model is parametric if $h_0(t) = h_1(\alpha, t)$ for some parameter vector α
  - PH model is semi-parametric if $h_0(t)$ is unspecified

Parametric proportional hazards model

- Weibull model can be defined as both AFT and PH model

Weibull regression model

- Weibull as an AFT model

$$S(t \mid \mathbf{x}) = \exp\Big(-[t/\alpha(\mathbf{x})]^{\delta}\Big)$$

where

$$\alpha(\mathbf{x}) = \exp\left(\mathbf{x}'\beta_{\mathsf{AFT}}\right) \tag{4}$$

Weibull regression model

- Weibull as a PH model

$$h(t \,|\, \mathbf{x}) = \frac{\delta}{\alpha(\mathbf{x})}\left[\frac{t}{\alpha(\mathbf{x})}\right]^{\delta-1}$$
$$= (\delta t^{\delta-1})[\alpha(\mathbf{x})]^{-\delta}$$
$$= h_1(\delta, t)\, r(\mathbf{x})$$

  - Assume

$$\begin{aligned} r(\mathbf{x}) &= \exp\left(\mathbf{x}'\beta_{\mathsf{PH}}\right) = [\alpha(\mathbf{x})]^{-\delta} \\ \Rightarrow\ \exp\left(-\mathbf{x}'\beta_{\mathsf{PH}}/\delta\right) &= \alpha(\mathbf{x}) \end{aligned} \tag{5}$$

Weibull regression model

- Equating the expression of $\alpha(\mathbf{x})$ from the AFT (Equation 4) and PH (Equation 5) Weibull model, we can show

$$\exp\left(-\mathbf{x}'\beta_{\mathsf{PH}}/\delta\right) = \alpha(\mathbf{x}) = \exp\left(\mathbf{x}'\beta_{\mathsf{AFT}}\right)$$
$$\Rightarrow \quad \beta_{\mathsf{PH}} = -\delta\beta_{\mathsf{AFT}} = -\frac{1}{b}\beta_{\mathsf{AFT}}$$

AFT and PH model

- Survivor function for some constants $c > 0$ and $r(\mathbf{x}) > 0$

$$S(t \,|\, \mathbf{x}_2) = S(c\,t \,|\, \mathbf{x}_1)$$
$$S(t \,|\, \mathbf{x}_2) = \left[S(t \,|\, \mathbf{x}_1)\right]^{r(\mathbf{x}_1)/r(\mathbf{x}_2)}$$
$$H(t \,|\, \mathbf{x}_2) = \left[r(\mathbf{x}_2)/r(\mathbf{x}_1)\right] H(t \,|\, \mathbf{x}_1)$$

Subsection 2

6.2 Inference for Log-location-scale AFT Models

Likelihood methods

- Data

$$\{(y_i, \delta_i, \mathbf{x}_i), i = 1, \ldots, n\}$$

  - Log-lifetime or log-censoring $y_i = \log t_i$
  - Censoring indicator $\delta_i = I($ith observation is a failure$)$
  - $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})'$ is a vecor of covariates

Likelihood methods

- Assume Y_i follows a location-scale distribution with location parameter $u(\mathbf{x}_i; \beta)$ and scale parameter b
- Regression model

$$u(\mathbf{x}_i; \beta) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} = \mathbf{x}_i' \beta$$

  - Vector of regression parameters

$$\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$$

  - Covariate vector $\mathbf{x}_i$ contains both categorical and quantitative variables, and for accurate computation, quantitative variables are centered

Likelihood methods

The log-likelihood function

$$\ell(\beta, b) = -r \log b \sum_{i=1}^{n} \left[\delta_i \log f_0(z_i) + (1 - \delta_i) \log S_0(z_i)\right] \qquad (6)$$

- $r = \sum_{i=1}^{n} \delta_i$
- $z_i = \frac{y_i - u(\mathbf{x}_i;\beta)}{b}$
- $u(\mathbf{x}_i; \beta) = \mathbf{x}_i' \beta$

Likelihood methods

Score functions

Elements of $(p+2)$-dimensional vector of score function

$$U_j(\beta, b) = \frac{\partial \ell(\beta, b)}{\partial \beta_j},\ j = 0, 1, \dots, p$$

$$U_b(\beta, b) = \frac{\partial \ell(\beta, b)}{\partial b}$$

- *Homework*: Obtain the expressions of score function (Eq. 6.3.3 and 6.3.4 of textbook)

Likelihood methods

Information matrix

Elements of observed information matrix

$$I(\beta, b) = -\begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta \partial \beta'} & \frac{\partial^2 \ell}{\partial \beta \partial b} \\ \frac{\partial^2 \ell}{\partial b \partial \beta'} & \frac{\partial^2 \ell}{\partial b^2} \end{bmatrix}$$

- *Homework*: Obtain the expressions of information matrix (Eq. 6.3.5, 6.3.6 and 6.3.7 of textbook)

Likelihood methods

MLEs

$$(\hat{\beta}^{'}, \hat{b})' = \arg\max_{(\beta', b)' \in \Theta} \ell(\beta, b)$$

- Iterative procedures (e.g. Newton-Raphson method) is used obtain MLE for β and b
- MLEs $(\hat{\beta}^{'}, \hat{b})'$ follow a $(p+2)$-variate normal distribution with mean $(\beta^{'}, b)'$ and variance matrix

$$\hat{V} = \left[I(\hat{\beta}, \hat{b})\right]^{-1}$$

- Large sample based tests and confidence intervals can be obtained using the sampling distribution of $(\hat{\beta}^{'}, \hat{b})'$

Test of hypothesis

Let $\beta' = (\beta_1', \beta_2')$, where β_1 is a k-dimensional vector of regression parameters, where $k < p$

$$H_0 : \beta_1 = \beta_1^0$$

- Likelihood ratio test statistic

$$\Lambda_1 = 2\ell(\hat{\beta}_1, \hat{\beta}_2, \hat{b}) - 2\ell(\beta_1^0, \tilde{\beta}_2, \tilde{b})$$

  - $(\tilde{\beta}', \tilde{b})' = \arg\max_{(\beta_2', b)' \in \Theta} \ell(\beta_1^0, \beta_2, b)$
  - Under H_0, $\Lambda_1 \sim \chi^2_{(k)}$

Test of hypothesis

Let $\beta' = (\beta_1', \beta_2')$, where β_1 is a k-dimensional vector of regression parameters, where $k < p$

$$H_0 : \beta_1 = \beta_1^0$$

- Wald statistic

$$\Lambda_2 = (\beta_1 - \beta_1^0)' V_{11}^{-1} (\beta_1 - \beta_1^0)$$

  - $V_{11} = var(\hat{\beta}_1)$ is a $k \times k$ matrix and

$$\hat{V} = \left[I(\hat{\beta}, \hat{b})\right]^{-1} = \begin{bmatrix} V_{11} & V_{12} \\ V_{12}' & V_{22} \end{bmatrix}$$

Under H_0, $\Lambda_1 \sim \chi^2_{(k)}$

Test of hypothesis

Null hypothesis

$$H_0 : \beta_j = 0$$

- Test statistic

$$Z_j = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)}$$

- $100(1-\alpha)\%$ confidence interval for β_j

$$\hat{\beta}_j \pm z_{1-\alpha/2}\, se(\hat{\beta}_j)$$

- For a small sample, LRT statistic can be used to test the hypothesis and to obtain confidence interval

Quantiles

- The pth quantile of Y given $\mathbf{x}$

$$y_p(\mathbf{x}) = \mathbf{x}'\beta + bw_p$$

- Estimate and corresponding SEs of pth quantile

$$\hat{y}_p(\mathbf{x}) = \mathbf{x}'\hat{\beta} + \hat{b}w_p$$

$$\text{and using delta method,} \quad se(\hat{y}_p(\mathbf{x})) = \mathbf{a}'V\mathbf{a}$$

  - $\mathbf{a} = (\mathbf{x}', w_p)$ and $w_p = S_0^{-1}(1-p)$

- $100(1-\alpha)\%$ confidence interval for y_p

$$\hat{y}_p \pm z_{1-\alpha/2}\, se(\hat{y}_p(\mathbf{x}))$$

Survival probability

- We are interest to obtain confidence interval for $S(y_0)$, which can be expressed in terms of the parameters of location-scale distribution as

$$S(y_0) = S_0\left(\frac{y_0 - \mathbf{x}'\beta}{b}\right)$$

$$S_0^{-1}(S(y_0)) = \frac{y_0 - \mathbf{x}'\beta}{b} = \psi(\mathbf{x})$$

Survival probability

- Estimate and the corresponding SE of $\psi(\mathbf{x})$

$$\hat{\psi}(\mathbf{x}) = \frac{y_0 - \mathbf{x}'\hat{\beta}}{\hat{b}} \text{ and using delta method, } se(\hat{\psi}(\mathbf{x})) = [\mathbf{a}'V\mathbf{a}]^{1/2}$$

  - $\mathbf{a}' = (-1/\hat{b})(\mathbf{x}', \hat{\psi}(\mathbf{x}))$

- $(1-\alpha)100\%$ confidence interval for $\psi(\mathbf{x})$

$$\hat{\psi}(\mathbf{x}) \pm z_{1-\alpha/2}\, se(\hat{\psi}(\mathbf{x}))$$

Survival probability

- Wald-type $(1-\alpha)100\%$ confidence interval for $S(y_0)$

$$\hat{\psi}(\mathbf{x}) - z_{1-\alpha/2}\, se(\hat{\psi}(\mathbf{x})) < \psi(\mathbf{x}) < \hat{\psi}(\mathbf{x}) + z_{1-\alpha/2}\, se(\hat{\psi}(\mathbf{x}))$$
$$L < \psi(\mathbf{x}) < U$$
$$L < S_0^{-1}(S(y_0)) < U$$
$$S_0(L) < S(y_0) < S_0(U)$$

Subsection 3

6.3 Weibull AFT

6.3 Weibull AFT

- Distributional assumption

$$T(\mathbf{x}) = (T \mid \mathbf{x}) \sim \mathsf{Weib}(\alpha(\mathbf{x}), \delta)$$
$$Y(\mathbf{x}) = (Y \mid \mathbf{x}) = (\log T \mid \mathbf{x}) \sim \mathsf{EV}(u(\mathbf{x}), b)$$

- Regression model for the parameters

$$u(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}'\beta$$
$$\alpha(\mathbf{x}) = \exp\left(\mathbf{x}'\beta\right)$$

  - $\mathbf{x} = (1, x_1, \ldots, x_p)'$
  - $\beta = (\beta_0, \beta_1, \ldots, \beta_p)'$

6.3 Weibull AFT

- Regression model for the response

$$Y(\mathbf{x}) = \mathbf{x}'\beta + bZ$$

  - $Z \sim \mathsf{EV}(0, 1)$
  - $f_0(z) = \exp(z - e^z)$
  - $S_0(z) = \exp(-e^z)$

6.3 Weibull AFT

- Log-likelihood function

$$\ell(\beta, b) = -r \log b + \sum_{i=1}^{n} \Big[\delta_i \log f_0(z_i) + (1-\delta_i) \log S_0(z_i)\Big]$$
$$= -r \log b + \sum_{i=1}^{n} (\delta_i z_i - e^{z_i})$$

  - $z_i = (y_i - \mathbf{x}_i' \beta)/b$

- We can now obtain score functions, information matrix, and MLE's for β and b (according to Section **??**.)

6.3 Weibull AFT

- We've already seen that the Weibull model implies a proportional hazard model
- It is the only parametric model that is both an AFT model and a Proportional Hazards (PH) model at the same time

Leukemia survival times

- Data on survival times for 33 leukemia patients are available, where survival times are in weeks from diagnosis
- Data on two covariates are also available
  - White blood cell count (WBC) at diagnosis
  - Binary variable AG indicates a positive (AG=1) or negative (AG=0) test related to white blood cell characteristics

Leukemia survival times

Table 6.1. Leukemia Survival Data

t	AG	wbc	t	AG	wbc
65	1	2.3	56	0	4.4
140[a]	1	.75	65	0	3.0
100	1	4.3	17	0	4.0
134	1	2.6	7	0	1.5
16	1	6.0	16	0	9.0
106[a]	1	10.5	22	0	5.3
121	1	10.0	3	0	10.0
4	1	17.0	4	0	19.0
39	1	5.4	2	0	27.0
121[a]	1	7.0	3	0	28.0
56	1	9.4	8	0	31.0
26	1	32.0	4	0	26.0
22	1	35.0	3	0	21.0
1	1	100.0	30	0	79.0
1	1	100.0	4	0	100.0
5	1	52.0	43	0	100.0
65	1	100.0			

[a] Denotes a censoring time; wbc = WBC ÷ 1000.

Leukemia survival times

```
tab6_1
```

```
# A tibble: 33 x 5
    time   wbc    AG status   lwbc
   <dbl> <dbl> <int>  <dbl>  <dbl>
 1    65  2.3      1      1  0.833
 2   140  0.75     1      0 -0.288
 3   100  4.3      1      1  1.46
 4   134  2.6      1      1  0.956
 5    16  6        1      1  1.79
 6   106 10.5      1      0  2.35
 7   121 10        1      1  2.30
 8     4 17        1      1  2.83
 9    39  5.4      1      1  1.69
10   121  7        1      0  1.95
# i 23 more rows
```

Leukemia survival times

- Consider Weibull AFT model with covariates $x_1 = AG$ and $x_2 = \log(wbc)$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + bZ \tag{7}$$

  - $Z \sim \mathsf{EV}(0, 1)$

Leukemia survival times

Fit Weibull regression model Equation 7 using R

```
mod62 <- survreg(Surv(time, status) ~ AG + lwbc,
                 data = tab6_1, dist = "weibull")

mod62E <- survreg(Surv(log(time), status) ~ AG + lwbc,
                 data = tab6_1, dist = "extreme")
```

Leukemia survival times

MLEs of model parameters

```
tidy(mod62, conf.int = T) |>
  mutate(p.value = scales::pvalue(p.value))
```

```
# A tibble: 4 x 7
  term        estimate std.error statistic p.value conf.low conf.hig
  <chr>          <dbl>     <dbl>     <dbl> <chr>      <dbl>    <dbl
1 (Intercept)    3.84      0.534     7.19  <0.001     2.79     4.89
2 AG             1.18      0.427     2.76  0.006      0.340    2.01
3 lwbc          -0.366     0.150    -2.45  0.014     -0.660   -0.07
4 Log(scale)     0.112     0.147     0.765 0.444     NA       NA
```

Leukemia survival times

Fitted model with $x_1 = AG$ and $x_2 = \log(wbc)$

$$\hat{Y} = 3.841 + 1.177x_1 - 0.366x_2 + \exp(1.119)Z$$

Leukemia survival times

- Variance matrix of the estimated parameters

```
vcov(mod62) %>% round(3)
```

```
            (Intercept)     AG   lwbc Log(scale)
(Intercept)       0.286 -0.130 -0.067      0.003
AG               -0.130  0.182  0.016      0.005
lwbc             -0.067  0.016  0.022     -0.005
Log(scale)        0.003  0.005 -0.005      0.021
```

Leukemia survival times

Table 1: Summary of Weibull AFT model fit

term	estimate	std.error	statistic	p.value	conf.low	conf.high
β_0	3.841	0.534	7.188	<0.001	2.794	4.889
β_1	1.177	0.427	2.757	0.006	0.340	2.014
β_2	-0.366	0.150	-2.449	0.014	-0.660	-0.073
$\log b$	0.112	0.147	0.765	0.444	NA	NA

- AG and WBC have significant effects on leukemia survival times. Positive AG and low WBC count are associated with more prolonged survival
- Since $\log b$ is not significant, i.e. there is not enough evidence to reject $H_0 : \log b = 1$, exponential AFT model would be appropriate for analyzing this data

Leukemia survival times

Interpretations

$$\exp(\hat{\beta}_1) = \exp(1.177) = 3.246$$

- A specific quantile (say median) lifetime of a patient with a positive AG value (i.e. $x_1 = 1$) is 3.2 times that of a patient with a negative AG (i.e. $x_1 = 0$) value provided WBC value remains constant
- Note this interpretation is true for any quantile (Why?)

Leukemia survival times

$$\exp(\hat{\beta}_2) = \exp(-0.366) = 0.693$$

- A specific quantile (say median) lifetime of a patient decreases 30.7 percent with one unit increase of log(WBC) [or 2718 unit increase of true WBC count] provided AG value remains constant

Leukemia survival times

Fitted values

```
augment(mod62, type.predict = "response") |>
  select(1:4) |>
  slice(1:3)
```

```
# A tibble: 3 x 4
  `Surv(time, status)`    AG   lwbc .fitted
                <Surv> <int>  <dbl>   <dbl>
1                 65       1  0.833   111.
2                140+      1 -0.288   168.
3                100       1  1.46     88.6
```

Leukemia survival times

```
augment(mod62, type.predict = "link") |>
  mutate(.fittedE = exp(.fitted)) |>
  select(2:4, .fittedE) |>
  slice(1:3)
```

```
# A tibble: 3 x 4
     AG   lwbc .fitted .fittedE
  <int>  <dbl>   <dbl>    <dbl>
1     1  0.833    4.71    111.
2     1 -0.288    5.12    168.
3     1  1.46     4.48     88.6
```

- Estimate for a subject with $AG = 1$ and $\log(wbc) = .833$

$$\hat{u} = \hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(.833) = (3.841) + (1.177)(1) + (-0.366)(.833) = 4.713$$

Leukemia survival times

```
#predict(object = mod62, newdata = tibble(AG = 1, lwbc = .833),
#        predict = "response")
augment(x = mod62, newdata = tibble(AG = 1, lwbc = .833),
        type.predict = "response")
```

```
# A tibble: 1 x 4
     AG  lwbc .fitted .se.fit
  <dbl> <dbl>   <dbl>   <dbl>
1     1 0.833    111.    41.3
```

- Estimate for a subject with $AG = 1$ and $\log(wbc) = .833$

$$\hat{\alpha} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1(1) + \hat{\beta}_2(.833)\right) = \exp\left((3.841) + (1.177)(1) + (-0.366)(.833)\right)$$
$$= \exp\left(4.713\right) = 111.399$$

Leukemia survival times

LRT

- Likelihood ratio tests for $H_0 : \beta_1 = 0$

$$\Lambda_1(0) = 2\ell(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \log \hat{b}) - 2\ell(\tilde{\beta}_0, 0, \tilde{\beta}_2, \log \tilde{b})$$

  - $\Lambda_1(0) \sim \chi^2_{(1)}$

- The corresponding Z statistic

$$Z = \mathsf{sign}(\hat{\beta}_1)\Lambda_1^{1/2} \sim \mathcal{N}(0, 1)$$

Leukemia survival times

Estimate of model parameters under $H_0 : \beta_1 = 0$

```
# mod62a <- update(mod62, formula = . ~ . - AG)
mod62a <- survreg(Surv(time, status) ~  lwbc,
                  data = tab6_1, dist = "weibull")
tidy(mod62a)
```

```
# A tibble: 3 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)    4.85      0.500      9.71 2.67e-22
2 lwbc          -0.500     0.165     -3.03 2.41e- 3
3 Log(scale)     0.222     0.146      1.52 1.28e- 1
```

Leukemia survival times

```r
LRTa <- anova(mod62a, mod62)
```

Terms	Resid. Df	-2*LL	Df	Deviance	Pr(>Chi)
lwbc	30	271.931	NA	NA	NA
AG + lwbc	29	265.013	1	6.918	0.009

- $\Lambda_1(0) = 6.918 \quad \Rightarrow \quad Z = 2.63$

Leukemia survival times

Table 3: Comparison between Wald- and LRT-type Z statistics

term	estimate	Wald	LRT
β_0	3.841	7.188	NA
β_1	1.177	2.757	2.63
β_2	-0.366	-2.449	-2.46
$\log b$	0.112	0.765	NA

Leukemia survival times

Quantiles

$$\hat{y}_p = \mathbf{x}'\hat{\beta} + \log(-\log(1-p))\,\hat{b}$$

- Consider a subject with covariate values $x_1 = 1$ and $x_2 = \log(10)$, the linear predictor $\mathbf{x}'\hat{\beta}$

$$\hat{u} = \boldsymbol{x}'\hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 + \log(10)\hat{\beta}_2 = 4.175$$

Leukemia survival times

- Median survival time of the patient with covariate values $x_1 = 1$ and $x_2 = \log(10)$

$$\hat{y}_{.50} = 4.175 + (-0.367)(1.119) = 3.765$$
$$\hat{t}_{.50} = \exp(3.765) = 43.163 \text{ weeks}$$

- *Homework*: Obtain a 95% confidence interval of the median survival time of a patient with covariate values $x_1 = 1$ and $x_2 = \log(10)$

Leukemia survival times

Survival probability

$$S(y_0) = \exp\left\{-\exp\left[(y_0 - \mathbf{x}'\hat{\beta})/\hat{b}\right]\right\}$$

For a patient with covariate values $x_1 = 1$ and $x_2 = \log(10)$, obtain $S(\log 10)$

$$S(\log 10) = \exp\left\{-\exp\left[\left(\log 10 - 4.175\right)/1.119\right]\right\} = 0.816$$

- *Homework*: Obtain the 95% CI for $S(\log 10)$

Subsection 4

6.4 Log-normal AFT

6.4 Log-normal AFT

- Distributional assumption

$$T(\mathbf{x}) = (T \mid \mathbf{x}) \sim \text{log-Norm}\big(\mu(\mathbf{x}), \sigma^2\big)$$
$$Y(\mathbf{x}) = (Y \mid \mathbf{x}) = (\log T \mid \mathbf{x}) \sim \mathcal{N}\big(\mu(\mathbf{x}), \sigma^2\big)$$

6.4 Log-normal AFT

- Regression model for the parameters

$$\mu(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}'\beta$$

- Regression model for the response

$$Y(\mathbf{x}) = \mathbf{x}'\beta + \sigma Z$$

  - $Z \sim \mathcal{N}(0, 1)$
  - $f_0(z) = \phi(z)$
  - $S_0(z) = 1 - \Phi(z)$

Times to pulmonary exacerbation

- Patients with *cystic fibrosis* are susceptible to an accumulation of mucus in the lungs, which leads to pulmonary exacerbation and deterioration of lung function
- A clinical trial was conducted to investigate the efficacy of the new drug DNase-1
  - Subjects are randomly assigned to a new treatment or a placebo
- Time of interest is the time to first exacerbation after randomization, and data on fev (forced expiratory volume at the time of randomization) are also measured

Times to pulmonary exacerbation

Table 1.4. Times to First Pulmonary Exacerbation for 10 Subjects

t (days)[a]	trt	fev[b]
168*	1	28.8
169*	1	64.0
65	0	67.2
168*	1	57.6
171*	0	57.6
166*	1	25.6
168*	0	86.4
90	0	32.0
169*	1	86.4
8	0	28.8

[a] Starred values are censoring times.
[b] fev measure is percent of predicted normal fev, based on sex, age, and height.

Times to pulmonary exacerbation

```
# A tibble: 761 x 13
        id   trt  time   fev  inst entry.dt   end.dt     ivstart ivsto
     <int> <int> <dbl> <dbl> <int> <date>     <date>       <dbl>  <dbl
 1       1     1   168  28.8     1 1992-03-20 1992-09-04      NA     N
 2       2     1   169  64       1 1992-03-24 1992-09-09      NA     N
 3       3     0    65  67.2     1 1992-03-24 1992-09-08      65     7
 4       4     1   168  57.6     1 1992-03-26 1992-09-10      NA     N
 5       5     0   171  57.6     1 1992-03-24 1992-09-11      NA     N
 6       6     1   166  25.6     1 1992-03-27 1992-09-09      NA     N
 7       7     0   168  86.4     1 1992-03-27 1992-09-11      NA     N
 8       8     0    90  32       1 1992-03-28 1992-09-10      90    10
 9       9     1   169  86.4     2 1992-02-27 1992-08-14      NA     N
10      10     0     8  28.8     2 1992-03-06 1992-08-22       8     2
# i 751 more rows
# i 3 more variables: status <dbl>, fevm <dbl>, visit <int>
```

Times to pulmonary exacerbation

- Assume survival time $T(\mathbf{x})$ follows a log-normal distribution with scale parameter $\alpha(\mathbf{x})$ and shape parameter δ
- Consider following AFT model for log survival time

$$Y(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \sigma Z$$

  - $Z \sim \mathcal{N}(0, 1)$
  - $x_1 = I(\text{trt} = 1)$
  - $x_2 = \text{fev} - \text{mean(fev)}$

Times to pulmonary exacerbation

- R codes for fitting the AFT model

```
mod63a <- survreg(Surv(log(time), status) ~ trt + fevm,
                  dist = "gaussian",
                  data = tab1_4)
```

Times to pulmonary exacerbation

```
tidy(mod63a)
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   5.09     0.0684      74.4  0
2 trt           0.336    0.0951       3.53 4.19e- 4
3 fevm          0.0159   0.00197      8.09 5.91e-16
4 Log(scale)    0.137    0.0408       3.36 7.84e- 4
```

Times to pulmonary exacerbation

- AFT model

$$Y = \mathbf{x}'\beta + bZ \;\;\Rightarrow\;\; T = \exp(\mathbf{x}'\beta)\exp(bZ)$$

  - $\mathbf{x}'\beta = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

- For a binary predictor x_j

$$T = \exp(\mathbf{x}'\beta)\exp(bZ) = \begin{cases} \exp(bZ) & \text{for control} \\ \exp(\beta_j)\exp(bZ) & \text{for treatment} \end{cases}$$

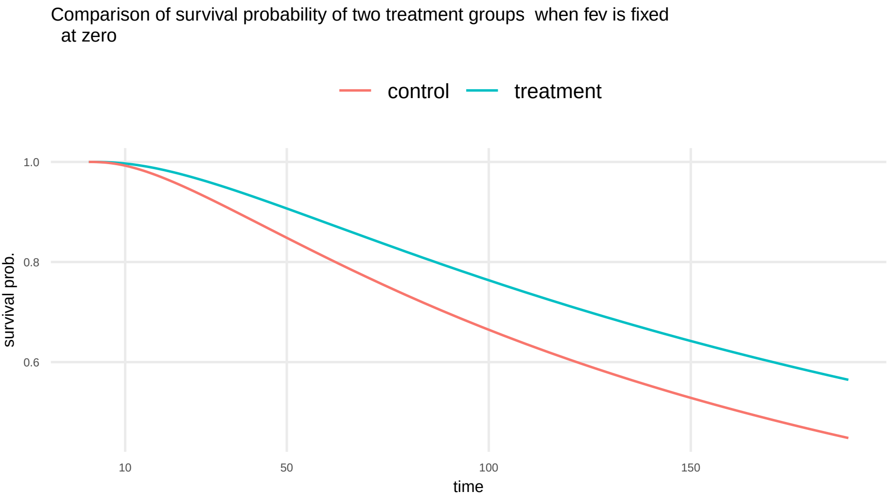
- It can be shown that

$$T_{trt} = \exp(\beta)\, T_{control}$$

Times to pulmonary exacerbation

- $\beta_{trt} = 0.336 \;\; \Rightarrow \;\; \exp(\beta_{trt}) = 1.399$
  - Treatment increases the time to first pulmonary exacerbation by about 40% compared to the control when `fev` is fixed
- $\beta_{fev} = 0.016 \;\; \Rightarrow \;\; \exp(\beta_{fev}) = 1.016$
  - One-unit increase in `fev` results about 2% increase in lifetime provided treatment is constant

Times to pulmonary exacerbation



Subsection 5

6.5 Log-logistic AFT

6.5 Log-logistic AFT

- Distributional assumptions

$$T(\mathbf{x}) = (T \mid \mathbf{x}) \sim \text{log-logistic}(\alpha(\mathbf{x}), \beta)$$
$$Y(\mathbf{x}) = (Y \mid \mathbf{x}) = (\log T \mid \mathbf{x}) \sim \text{logistic}(u(\mathbf{x}), b)$$

6.5 Log-logistic AFT

- Regression model for the parameters

$$u(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \mathbf{x}'\beta$$
$$\alpha(\mathbf{x}) = e^{u(\mathbf{x})}$$

- Regression model for the response

$$Y(\mathbf{x}) = \mathbf{x}'\beta + bZ$$

  - $Z \sim$ Logistic$(0, 1)$
  - $f_0(z) = e^z[1 + e^z]^{-2}$
  - $S_0(z) = [1 + e^z]^{-1}$

6.5 Log-logistic AFT

- Lifetime distribution

$$T(\boldsymbol{x}) \sim \text{Log-Logistic}\big(\alpha(\boldsymbol{x}), \delta\big)$$

- The survivor function

$$S(t \,|\, \boldsymbol{x}) = \frac{1}{1 + \big(t/\alpha(\boldsymbol{x})\big)^{\delta}} \;\Rightarrow\; \frac{1 - S(t \,|\, \boldsymbol{x})}{S(t \,|\, \boldsymbol{x})} = \big(t/\alpha(\boldsymbol{x})\big)^{\delta}$$

  - $\big(t/\alpha(\boldsymbol{x})\big)^{\delta} \rightarrow$ the odds of failure at time t for a subject with covariate vector $\boldsymbol{x}$

6.5 Log-logistic AFT

- For two subjects with covariate vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$

$$\frac{[1 - S(t \mid \boldsymbol{x}_2)]/S(t \mid \boldsymbol{x}_2)}{[1 - S(t \mid \boldsymbol{x}_1)]/S(t \mid \boldsymbol{x}_1)} = \left[\frac{\alpha(\boldsymbol{x}_1)}{\alpha(\boldsymbol{x}_2)}\right]^{\delta}, \text{ independent of } t$$

- A model of the form

$$\frac{1 - S(t \mid \boldsymbol{x})}{S(t \mid \boldsymbol{x})} = \big(t/\alpha(\boldsymbol{x})\big)^{\delta} \Rightarrow \log \frac{1 - S(t \mid \boldsymbol{x})}{S(t \mid \boldsymbol{x})} = \delta \log(t) - \delta \log \alpha(\boldsymbol{x})$$

is known as the proportional odds model

6.5 Log-logistic AFT

- Consider a model $\log \alpha(x) = \beta_0 + \beta_1 x$

$$\frac{[1 - S(t \,|\, x = 1)]/S(t \,|\, x = 1)}{[1 - S(t \,|\, x = 0)]/S(t \,|\, x = 0)} = e^{-\delta\beta_1} = e^{-\beta_1^\star}$$

  - The odds of failure at time t for a subject with $x = 1$ is $\exp(-\beta^\star)$ times that of the odds of failure for a subject with $x = 0$

Times to pulmonary exacerbation

- R codes for fitting AFT model

```
mod63b <- survreg(Surv(log(time), status) ~ trt + fevm,
                  dist = "logistic",
                  data = tab1_4)
```

Times to pulmonary exacerbation

```r
tidy(mod63b)
```

```
# A tibble: 4 x 5
  term        estimate std.error statistic  p.value
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>
1 (Intercept)   5.08     0.0600      84.6  0
2 trt           0.293    0.0861       3.41 6.55e- 4
3 fevm          0.0145   0.00181      8.00 1.20e-15
4 Log(scale)   -0.489    0.0466     -10.5  8.08e-26
```

Times to pulmonary exacerbation

- $\beta_{trt} = 0.293 \;\; \Rightarrow \;\; \exp(\beta_{trt}) = 1.341$
  - Treatment increases the time to first pulmonary exacerbation by about 34% compared to the control when `fev` is fixed
- $\beta_{fev} = 0.014 \;\; \Rightarrow \;\; \exp(\beta_{fev}) = 1.015$
  - One-unit increase in `fev` results in a 1.5% increase in lifetime provided treatment is constant

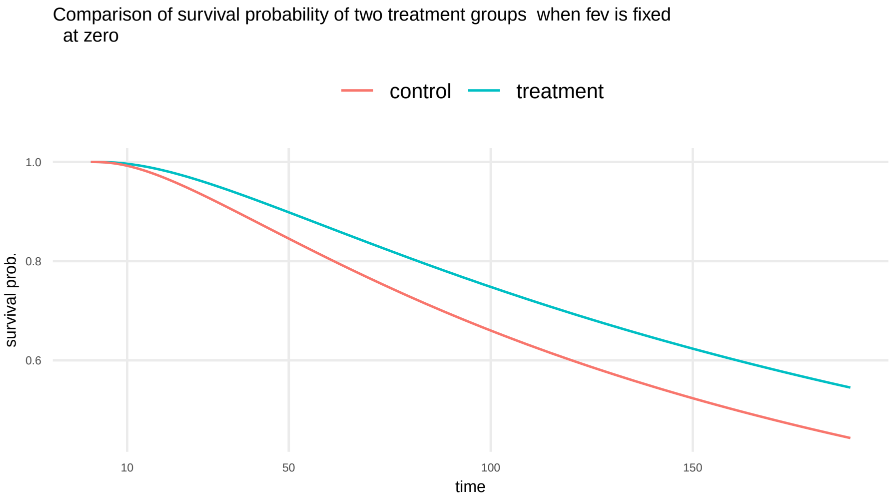
Times to pulmonary exacerbation

- Interpret the treatment effect in terms of odds of failure

$$\frac{[1 - S(t \,|\, trt = 1, fev = x)]/S(t \,|\, trt = 1, fev = x)}{[1 - S(t \,|\, trt = 0, fev = x)]/S(t \,|\, trt = 0, fev = x)} = \exp(-\widehat{\delta}\widehat{\beta}_1) = 0.62$$

  - $\widehat{\delta} = \exp(-\log \widehat{b}) = \exp(0.489) = 1.631$
  - The odds of failure is 38% lower in the treatment group compared to the control group provided `fev` value is fixed

Times to pulmonary exacerbation



Times to pulmonary exacerbation

Table 4: Comparison between normal and logistic regression models in analysing time to pulmonary exacerbation data

term	est	se	est	se
(Intercept)	5.093	0.068	5.078	0.060
trt	0.336	0.095	0.293	0.086
fevm	0.016	0.002	0.014	0.002
Log(scale)	0.137	0.041	-0.489	0.047

Other regression models

- Additive hazards model

$$h(t \,|\, \boldsymbol{x}) = h_0(t; \alpha) + r(\boldsymbol{x}; \boldsymbol{\beta})$$

Subsection 6

6.6 Graphical methods and model assessment

6.6 Graphical methods and model assessment

- Graphical methods are helpful in summarizing information and suggesting possible models
- These methods also provide ways to check assumptions concerning the form of a lifetime distribution and its relationship to covariates
- Exploratory analysis of a lifetime distribution given covariates would helpful to select the appropiate Model for the analysis

6.6 Graphical methods and model assessment

- For a single quantitative covariate, a plot of lifetime or log-lifetime against the covariate or a function of it could indicate the nature of the relationship between lifetime and the covariate
- If the proportion of censoring is small, such a plot would be helpful, different symbols can be used in those plots for censored and failure times
- When more than one quantitative covariate and light censoring, one can consider grouping individuals so that within a group, individuals will have similar values of important covariates
- Let there are J such groups and $\hat{S}_j$ is the Kaplan-Meier estimate for the group $j = 1, \dots, J$

6.6 Graphical methods and model assessment

AFT model

$$S(t \,|\, \mathbf{x}) = S_0\left[\frac{\log t - u(\mathbf{x})}{b}\right]$$

- If $u(\mathbf{x})$ is approximately constant for individuals within each group $j = 1, \dots, J$, and if an AFT model is appropriate, the plots of

$$\log[-\log S(t \,|\, \mathbf{x})] \quad \text{vs} \quad \log t$$

should be roughly parallel in horizontal direction $(\log t)$

6.6 Graphical methods and model assessment

Proportional hazards model

$$S(t \,|\, \mathbf{x}) = \left[S_0(t)\right]^{r(\mathbf{x})}$$

- If $r(\mathbf{x})$ is approximately constant for individuals within each group $j = 1, \ldots, J$, and if a proportional hazards model is appropriate, the plots of

  $$\log[-\log S(t \,|\, \mathbf{x})] \quad \text{vs} \quad \log t$$

  should be roughly parallel in vertical direction

6.6 Graphical methods and model assessment

- If the plots of $\log[-\log S(t \mid \mathbf{x})]$ vs $\log t$ is roughly linear then Weibull models are suggested
- In addition to linear, if the plots are parallel, then Weibull models with a constant shape parameter are suggested, in that case, both AFT and PH models can be considered
- Statistical analysis of data is an iterative process involving exploration, model fitting, and model assessment

Collett, David. 2015. *Modelling Survival Data in Medical Research*. Chapman; Hall/CRC. https://doi.org/10.1201/b18041.

Cox, David R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34 (2): 187–202.