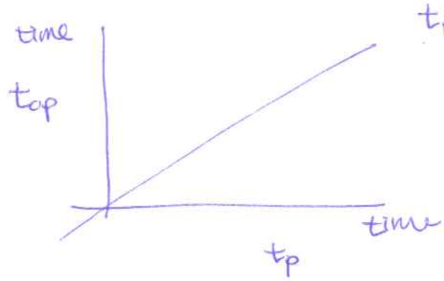
- Some comments on AFT models
- Is an AFT model appropriate?
  - Suppose we have two groups in data ($\Leftrightarrow$ one binary coviariate).
  - Recall that the p-th percentile under AFT models,

$$t_p \exp(\beta_1) = t_{0p}.$$

  - We can use the Kaplan-Meier (Nelson-Aalen) method for each group and get nonparametric estimates $\hat{t}_p$ and $\hat{t}_{0p}$.
  - Check if a plot of $\hat{t}_p$ versus $\hat{t}_{0p}$ goes through the origin with slope approximately equal to the accelerated factor $\exp(\beta_1)$.

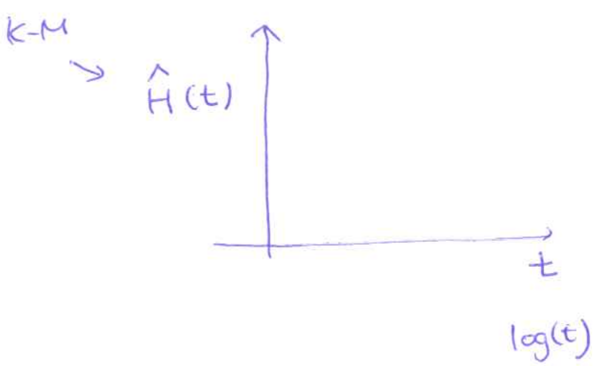
$\hat{t}_p \underbrace{e^{x\beta}}_{C} \approx \hat{t}_{op}$



- Is a particular F_W more appropriate?
  - We can use the Kaplan-Meier (Nelson-Aalen) method for each group and get a nonparametric estimate of cumulative hazard $\hat{H}$.
  - Recall cumulative hazard functions.

Models	$H(t)$
Exp	γt
Weibull	γt^{α} $\rightarrow$ $\log(\gamma t^{\alpha}) = \log(\gamma) + \alpha \log(t)$
Log-normal	$-\log(1 - \Phi((\log(t) - \mu)/\sigma)) = H$

  - Check which model is supported by $\hat{H}(t)$.



$$1 - \Phi\left(\frac{\log t - \mu}{\sigma}\right) = e^{-H}$$

$$\frac{\log(t) - \mu}{\sigma} = \Phi^{-1}(1 - e^{-H})$$

Diagnostics for AFT

- Define

$$r_i = \frac{Y_i + \beta_0 + \boldsymbol{\beta}'\mathbf{X}_i}{\sigma}$$

- If the model fits well, we expect r_i should have the same distribution as W $\{r_1, \ldots, r_n\}$ a right censored ~~rando~~ random sample from F_W

- Further, we can use the Cox-Snell residuals. Please read K-M for more

- Any drawback?

$$\log(T) = Y = -\beta_0 - \boldsymbol{\beta}'\mathbf{X} + \sigma W$$

F_W

Standard EV
= Standard N
Log

★★ A direction extension of the classical linear model's construction for conventional data

★★ Restricted by the error distribution:

  - If a correct model is specified, gives more precise estimate of parameters
  - If the model is incorrectly specified, provides inconsistent estimates.

★★ Popular choices for a distribution of W: Standard extreme value distribution, standard logistic distribution, normal distribution. Then **which model** is better? Use model comparison criteria such as DIC and AIC.

• Deviance Information Criterion (DIC) – Bayesian Data Analysis Chapter 6

poor fit

⋆⋆ [Definition] Deviance: $D(y, \theta) = -2 \log p(y \mid \theta)$. → $p(y|\theta)$ small
Note: It is a function of *both*, θ and y. → $-2 \cdot \log p(y|\theta)$ big

⋆⋆ Consider two quantities,

$$D_{\widehat{\theta}}(y) = D(y, \widehat{\theta}), \quad \text{and} \quad D_{\text{avg}}(y) = \mathrm{E}(D(y, \theta) \mid y) \approx \frac{1}{L} \sum_{\ell=1}^{L} D(y, \theta^{(\ell)}),$$

$= \hat{D}_{avg}(y)$

where $\widehat{\theta}$: a point estimate of θ, $\theta^{(\ell)}$: posterior simulations.

⋆⋆ DIC is defined as

$$\begin{aligned} \text{DIC} &= \boxed{2\widehat{D}_{\text{avg}}(y) - D_{\hat{\theta}}(y)} = 2\widehat{D}_{\text{avg}}(y) - D(y, \hat{\theta}) \\ &= D(y, \hat{\theta}) + \underbrace{2(\widehat{D}_{\text{avg}}(y) - D(y, \hat{\theta}))}_{\text{effective number of parameters}} \end{aligned}$$

small DIC wins

more complex model ⇒ large

poor fit ⇒ large

$*$ [Example: Male Laryngeal Cancer Patients] We use the accelerated failure-time model using the main effects of age and stage for this data;

$$Y = \log(T) = -\beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_3 X_3 - \beta_4 X_4 + \sigma W,$$

where X_k, $k = 1, 2, 3$ are the indicators of stage II, III and IV disease, respectively, and X_4 is the age of the patient.

F_W	DIC
Extreme Value	232.7812
Logistic	224.5492
Normal	832.1809

- Normal for F_W looks the far worst.
- The logistic distribution performs the best, followed by the extreme value distribution.

♣ Nonparametric Bayesian Accelerated Failure-Time Model

- We have

$$\log(T) = Y = -\beta_0 - \boldsymbol{\beta}'\mathbf{X} + \sigma W$$

$\Rightarrow$ We placed priors on unknown parameters, β_0, $\boldsymbol{\beta}$ and σ.

- More elaborate priors? See ICS Chapter 10.2.

  ⋆⋆ $W \sim G$ and $G \sim$ DP $\Rightarrow$ Done in Kuo and Mallick (1997)

  ⋆⋆ $W \sim G$ and $G \sim$ Pólya Tree $\Rightarrow$ Done in Walker and Mallick (1999)

AMS 276
Lecture 4: Proportional Hazards Regression

Fall 2016

♣ Regression Models for Survival Data

⋆⋆ Often interested in studying the relationship between the failure time (T) and covariates ($\mathbf{X}$: $p \times 1$ associated with T).

e.g. Predict the distribution of the failure time from a set of covaraites.

⋆⋆ Adjust the survival function to account for covariates.

- Two Common Approaches:

⋆⋆ Accelerated Failure-Time Model (cleared!)

⋆⋆ **Proportional Hazards Model (Multiplicative Hazards Model - Cox-type model).**

♣ Approach 2: Proportional Hazards Regression Model

- KM Chapters 8 & 9, ICS Chapter 1.4.1 & 1.4.3

- Recall that the survival time t has
  - ⋆⋆ density function $f(t)$
  - ⋆⋆ distribution function $F(t)$
  - ⋆⋆ hazard function $h(t) = \frac{f(t)}{S(t)} > 0$, where $S(t) = 1 - F(t)$.

- Proposed by Cox (1972), primarily to *model the relationship between hazard function and covariates.*

• Proportional Hazards Regression Models: The hazard function depends on both time (t) and a set of covariates ($\mathbf{X}$).

• Express the conditional hazard rate for an individual with $\mathbf{X}$ as

$$\underbrace{h(t \mid \mathbf{X})}_{>0} = \underbrace{h_0(t)}_{>0}\underbrace{c(\beta'\mathbf{X})}_{>0}.$$

$S(t) = e^{-\int_0^t h(u)du}$

(called, Cox model, proportional hazard model)

⋆⋆ A baseline hazard rate $h_0(t) = h(t \mid \mathbf{X} = \mathbf{0})$: arbitrary β_0

⋆⋆ $\beta = (\beta_1, \ldots, \beta_p)'$: a parameter vector (no intercept!)

⋆⋆ A **nonnegative function** can be used for the link function $c(\cdot)$

⋆⋆ Multiplicative model: covariates are assumed to affect survival probability by multiplying the baseline hazard.

- Consider two individuals with $\mathbf{X}_1$ and $\mathbf{X}_2$ (all the covariates are fixed at time 0).
- The ratio of their hazard rates is

$$\frac{h(t \mid \mathbf{X}_1)}{h(t \mid \mathbf{X}_2)} = \frac{\cancel{h_0(t)}c(\beta'\mathbf{X}_1)}{\cancel{h_0(t)}c(\beta'\mathbf{X}_2)} = \frac{c(\beta'\mathbf{X}_1)}{c(\beta'\mathbf{X}_2)}.$$

⋆⋆ the **relative risk (hazard ratio)** of an individual with $\mathbf{X}_1$ having the event as compared to an individual with $\mathbf{X}_2$.

⋆⋆ Constant over time (independent of time) provided that $\mathbf{X}$ does not change over time.

⋆⋆ The ratio of hazards for two individuals depends on the difference between their $\mathbf{X}$ at any time.

- Express the conditional hazard rate for an individual with $\mathbf{X}$ as

$$h(t \mid \mathbf{X}) = h_0(t)c(\beta'\mathbf{X}).$$

⋆⋆ This is called a semiparametric proportional hazards regression model. *Why?*

- A known parametric form is assumed for $c(\cdot)$.
- $h_0(t)$ is unspecified and it will be treated nonparametrically.

- Common choice: $c(\beta'\mathbf{X}) = \exp(\beta'\mathbf{X}) > 0$

○ Assume to involve **X** through a log-linear model.

$$c(\beta'\mathbf{X}) = \exp(\beta'\mathbf{X}) = \exp\left(\sum_{k=1}^{p} \beta_k X_k\right).$$

$c(\beta'x)$

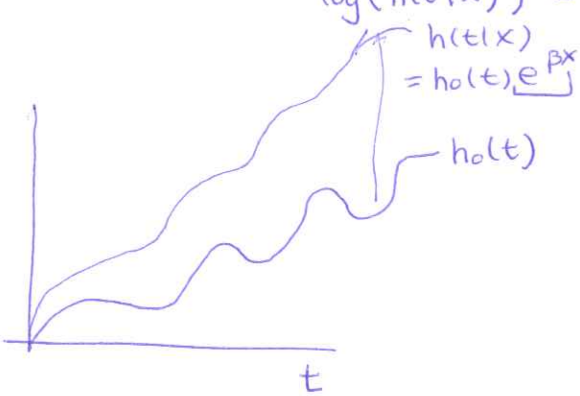
$$\Rightarrow \quad h(t \mid \mathbf{X}) = h_0(t)\exp(\beta'\mathbf{X}) = h_0(t)\exp\left(\sum_{k=1}^{p} \beta_k X_k\right)$$

$$\Rightarrow \quad \log\left(\frac{h(t \mid \mathbf{X})}{h_0(t)}\right) = \beta'\mathbf{X} = \sum_{k=1}^{p} \beta_k X_k$$

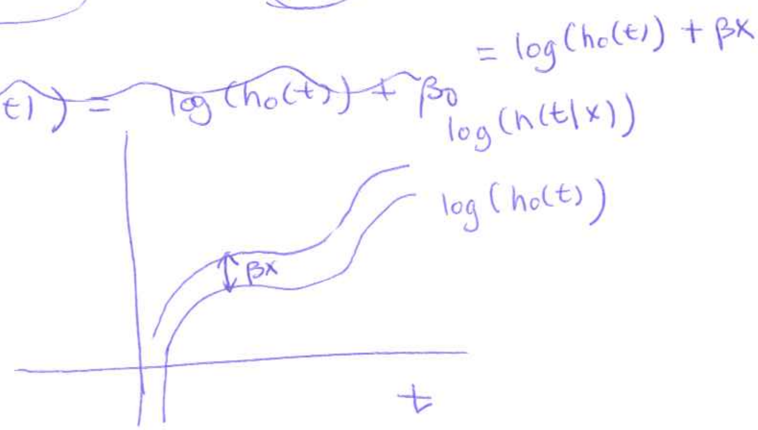
i.e., Similar to the usual linear models for formulation for the effects of covariates .

$\Rightarrow \quad h(t|x) = h_0(t)e^{\beta x}$

$\log(h(t|x)) = \log(h_0(t)) + \beta x \; \cancel{+\beta_0}$



$\log(h_0'(t)) = \log(h_0(t)) + \beta_0$



- How does **X** affect the hazard function under the proportional hazards regression model?

○ Interpretation of $\boldsymbol{\beta}$

e.g. Let X_1 be the treatment ($X_1 = 1$ for female, $X_1 = 0$ for male).

$$\frac{h(t \mid X_1 = 1)}{h(t \mid X_1 = 0)} = \exp(\beta_1)$$

$\Rightarrow$ The risk of having the event for fe males~~s~~ relative to the risk for ~~fe~~males is $\exp(\beta_1)$.

$\Leftrightarrow$ $\exp(\beta_1)$ is the ratio of hazards (assumed constant for all t)

$\Leftrightarrow$ β_1 is the difference in log-hazard at any time for a female subject.

$\Leftrightarrow$ If $\beta_1 > 0$, $h(t \mid X_1 = 1) \uparrow$ and $S(t \mid X_1 = 1) \downarrow$.

- What is the goal in general? **Inference for β !!!**

  ⋆⋆ β characterizes the effect of $\mathbf{X}$.

  ⋆⋆ Treat the baseline hazard, $h_0(t)$ as a nuisance parameter function. Don't even estimate $h_0(t)$. S

  ⋆⋆ Use a partial or conditional likelihood rather than a full likelihood approach to estimate β (via the Newton-Raphson).

  ⋆⋆ Then we can do tests, $H_0 : \beta_1 = \beta_{10}$ and variable selection (model selection) based on the test or some other criteria (AIC...)

  ⋆⋆ (Yep!) Sometimes we are interested in estimating the survival function for a patient with a certain set of conditions and characteristics. ⇒ We need to model $h_0(t)$ as well (will discuss later).

$\hat{h}(t|x) = h_0(t) e^{x\hat{\beta}} \Rightarrow \hat{S}(t|x)$

- We have

$$h(t \mid \mathbf{X}) = h_0(t)\exp(\beta'\mathbf{X}).$$

- $h_0(t)$ is left completely unspecified (nuisance parameter).

$\Rightarrow$ Can't use standard maximum likelihood methods to estimate β.

- Cox proposed the idea of a **partial likelihood** to remove $h_0(t)$ from the proposed estimating equation.

- The proportional hazards regression model is also called the Cox model (Cox, 1972, JRSS-B).

Regression Models and Life-Tables - JStor
https://www.jstor.org/stable/2985181 JSTOR
by DR Cox - 1972 - Cited by 42144 - Related articles
1972] 187. Regression Models and Life-Tables. BY D. R. Cox. Imperial College, London. [Read before the ROYAL STATISTICAL SOCIETY, at a meeting ...

- Likelihood: conditional, marginal and partial likelihood.
- Consider a general case. Suppose

  ⋆⋆ $\mathbf{X} = (\mathbf{V}, \mathbf{W})$: data (observations) $f'(\underline{x}' \mid \beta)$

  ⋆⋆ $\boldsymbol{\theta} = (\boldsymbol{\beta}, \phi)$: parameters

  ⋆⋆ $\boldsymbol{\beta}$: parameters of interest, ϕ: nuisance parameter

  ⋆⋆ density of $\mathbf{X}$: $f(\mathbf{X} \mid \theta)$
- Goal: inference on $\boldsymbol{\beta}$ (part of the parameter)
- We modify the likelihood function to extract the evidence in data concerning a parameter of interest $\boldsymbol{\beta}$ (construct a likelihood-like function using the density of just part of the data, pseudo-likelihood)

i.e., conditional likelihood, marginal likelihood, partial likelihood....

$*$ Profile ~~likehood~~ likelihood

$\mathcal{L}(\beta, \phi)$

For a fixed β, find $\hat{\phi}_\beta = \underset{\phi}{\operatorname{argmax}} \ \mathcal{L}(\beta, \phi)$

$\Rightarrow$ Find $\hat{\beta} = \underset{\beta}{\operatorname{argmax}} \ \mathcal{L}(\beta, \hat{\phi}_\beta)$

- Likelihood: $\mathcal{L}(\boldsymbol{\theta}) = f(\mathbf{X} \mid \boldsymbol{\theta}) = f(\mathbf{W} \mid \mathbf{V}, \boldsymbol{\theta}) f(\mathbf{V} \mid \boldsymbol{\theta})$.

$f(x \mid \theta) = f(w \mid v, \beta) \cdot \cancel{f(v \mid \beta, \phi)}$ ignore

  ⋆⋆ $f(\mathbf{W} \mid \mathbf{V}, \boldsymbol{\theta})$ does not involve $\boldsymbol{\phi}$

  $\Rightarrow$ Use $f(\mathbf{W} \mid \mathbf{V}, \boldsymbol{\beta})$ (*conditional likelihood*): Case 1

  ⋆⋆ $f(\mathbf{V} \mid \boldsymbol{\theta})$ does not involve $\boldsymbol{\phi}$ $f(x \mid \theta) = \cancel{f(w \mid v, \beta, \phi)} f(v \mid \beta)$

  $\Rightarrow$ Use $f(\mathbf{V} \mid \boldsymbol{\beta})$ (*marginal likelihood*): Case 2

- Side note: Possible loss of useful information about $\boldsymbol{\beta}$.

  ⋆⋆ Case 1: Ignore $f(\mathbf{V} \mid \boldsymbol{\theta})$ and use $f(\mathbf{W} \mid \mathbf{V}, \boldsymbol{\theta})$.

  $\Rightarrow$ ignoring their variability by conditioning

  ⋆⋆ Case 2: Ignore $f(\mathbf{W} \mid \mathbf{V}, \boldsymbol{\theta})$ and use $f(\mathbf{V} \mid \boldsymbol{\theta})$.

  $\Rightarrow$ ignoring some of the data by marginalization

$*$ Read "Integrated Likelihood Methods for Eliminating Nuisance Parameters" by Berger et al.

$$f(x \mid \beta) = \int f(x \mid \beta, \phi)\, \pi(\phi \mid \beta)\, d\phi$$

$$\Rightarrow \pi(\beta \mid x) \propto \pi(\beta)\, f(x \mid \beta)$$

$\Leftrightarrow$

$$\pi(\beta, \phi \mid x) \propto f(x \mid \beta, \phi)\, \pi(\phi \mid \beta)\, \pi(\beta)$$

$$\Rightarrow \pi(\beta \mid x) = \int \pi(\beta, \phi \mid x)\, d\phi$$

- Represent $\mathbf{X} = (V_1, W_1, V_2, W_2, \ldots, V_K, W_K)$.
- Write the likelihood,

$$
\begin{aligned}
f(\mathbf{X} \mid \theta) &= f(V_1, W_1, V_2, W_2, \ldots, V_K, W_K \mid \theta) \\
&= f(V_1 \mid \theta) \cdot f(W_1 \mid V_1, \theta) \cdot f(V_2 \mid V_1, W_1, \theta) \cdot f(W_2 \mid V_1, W_1, V_2, \theta) \ldots \\
&= \left\{ \prod_{i=1}^{K} f(W_i \mid Q_i, \theta) \right\} \left\{ \prod_{i=1}^{K} f(V_i \mid P_i, \theta) \right\}.
\end{aligned}
$$

⋆⋆ $P_1 = \phi,\ P_i = (V_1, W_1, \ldots, V_{i-1}, W_{i-1})$

⋆⋆ $Q_1 = V_1,\ Q_i = (V_1, W_1, \ldots, W_{i-1}, V_i)$

⋆⋆ If $\prod_{i=1}^{K} f(W_i \mid Q_i, \theta)$ is free of ϕ, then use $\prod_{i=1}^{K} f(W_i \mid Q_i, \beta)$ (*partial likelihood*).

- *side note*: Marginal and conditional likelihoods are special cases of the more general partial likelihood (Cox, 1975).

observed survival times are distinct

- Partial likelihoods for distinct-event time data

- We will express the data in V and W to find the partial likelihood.

- Set-up

  ⋆⋆ Data: $(y_i, \nu_i, \mathbf{X}_i)$, $i = 1, \ldots, n$ (n individuals)

  ⋆⋆ Absolutely continuous failure time distribution

  ⋆⋆ Assume noninformative censoring

  ⋆⋆ d distinct event times (d observed failures) and $n - d$ right censored survival times.

  ⋆⋆ $t_0(= 0) < t_1 < t_2 < \ldots < t_d < t_{d+1}(= \infty)$: the distinct ordered event times (no ties between the event times)

  ⋆⋆ Let (j) be the label for individual failing at t_j. Note that $y_{(j)} = t_j$.

- Set-up (contd)

  ⋆⋆ Covariates for d failures, $\mathbf{X}_{(j)}$, $j = 1, \ldots, d$

  ⋆⋆ Censorship times in $[t_j, t_{j+1})$: $(t_{j1}, \ldots, t_{jm_j})$ with corresponding covariates $\mathbf{X}_{j1}, \ldots, \mathbf{X}_{jm_j}$.

- Now we divide the data into sets

$$(V_1, W_1, V_2, W_2, \ldots, V_{d+1}, W_{d+1}),$$

where t_{j-1}

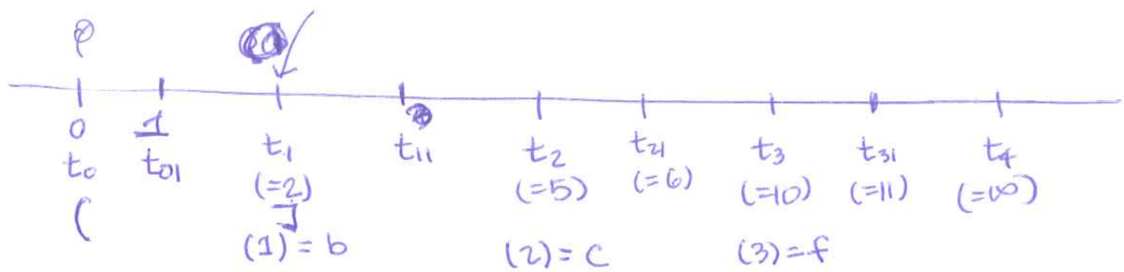
  ⋆⋆ $V_j = \{t_{j-1,1}, \ldots, t_{j-1,m_{j-1}}, t_j\}$: tells us who has died or was censored in $(t_{j-1}, t_j]$.

  ⋆⋆ $W_j = \{(j)\}$: tells us who died at time t_j in the sample.

- Example:

id	a	b	c	d	e	f	g
y_i	1	2	5	3	11	10	6
ν_i	0	1	1	0	0	1	0

censored   observed



$V_1 = \{t_{01}, t_1\}$ $V_2 = \{t_{11}, t_2\}$ $V_3 = \{t_{21}, t_3\}$

$V_4 = \{t_{31}, t_4\}$

$W_1 = \{b\}$ $W_2 = \{c\}$ $W_3 = \{f\}$