Duration Models: Modeling Strategies

Brad Jones\textsuperscript{1}

\textsuperscript{1}Department of Political Science
University of California, Davis

February 21, 2007
Today: Different Modeling Approaches
Parametrics: Motivation

- Some Motivation for Parametrics
Parametrics: Motivation

- Some Motivation for Parametrics
- Consider the hazard rate:

\[ \frac{dh(t)}{dt} > 0, \]

Hazard increasing wrt time.

\[ \frac{dh(t)}{dt} < 0, \]

Hazard decreasing wrt time.

\[ \frac{dh(t)}{dt} = 0, \]

Hazard “flat” wrt time.
Premise

- Parametric models give structure (shape) to the hazard function.
Premise

- Parametric models give structure (shape) to the hazard function.
- N.B.: the structure is a function of the c.d.f., not necessarily of the “real world.”
Premise

- Parametric models give structure (shape) to the hazard function.
- N.B.: the structure is a function of the c.d.f., not necessarily of the “real world.”
- . . . though some c.d.f.s do a good job of approximating some failure-time processes.
Premise

- Parametric models give structure (shape) to the hazard function.
- N.B.: the structure is a function of the c.d.f., not necessarily of the “real world.”
- ...though some c.d.f.s do a good job of approximating some failure-time processes.
- Any c.d.f. with positive support on the real number line will work.
Premise

- Parametric models give structure (shape) to the hazard function.
- N.B.: the structure is a function of the c.d.f., not necessarily of the “real world.”
- ...though some c.d.f.s do a good job of approximating some failure-time processes.
- Any c.d.f. with positive support on the real number line will work.
- Lots of choices: exponential, Weibull, gamma, Gompertz, log-normal, log-logistic ... etc.
Artwork

Figure: This figure graphs typical functional forms for several common parametric distribution functions.
Some Models

▶ For parametrics, we work with standard likelihood methods.
Some Models

- For parametrics, we work with standard likelihood methods.
- Specify a distribution function and write out the log-likelihood for the data.
Some Models

- For parametrics, we work with standard likelihood methods.
- Specify a distribution function and write out the log-likelihood for the data.
- The question is, which distribution function?
Some Models

- For parametrics, we work with standard likelihood methods.
- Specify a distribution function and write out the log-likelihood for the data.
- The question is, which distribution function?
- In all software programs/computing environments, you’re given a menu.
Some Models

- For parametrics, we work with standard likelihood methods.
- Specify a distribution function and write out the log-likelihood for the data.
- The question is, which distribution function?
- In all software programs/computing environments, you’re given a menu.
- Stata: streg, R: survreg, eha
Some Models

- For parametrics, we work with standard likelihood methods.
- Specify a distribution function and write out the log-likelihood for the data.
- The question is, which distribution function?
- In all software programs/computing environments, you’re given a menu.
- Stata: streg, R: survreg, eha
- It is easy (maybe too easy?)
Exponential

- A very simple distribution is the exponential.
A very simple distribution is the exponential.
Probably unrealistic for many settings (though apparently fits well in industrial engineering)
Exponential

▶ A very simple distribution is the exponential.
▶ Probably unrealistic for many settings (though apparently fits well in industrial engineering)
▶ Easy to see why it’s simple; consider hazard function:

\[ h(t) = \lambda \quad t > 0, \lambda > 0 \]  

(1)
A very simple distribution is the exponential.

Probably unrealistic for many settings (though apparently fits well in industrial engineering)

Easy to see why it’s simple; consider hazard function:

\[
h(t) = \lambda \quad t > 0, \lambda > 0 \quad (1)
\]

\(h(t)\) is a constant: flat wrt time.
Exponential

- A very simple distribution is the exponential.
- Probably unrealistic for many settings (though apparently fits well in industrial engineering)
- Easy to see why it’s simple; consider hazard function:

\[ h(t) = \lambda \quad t > 0, \lambda > 0 \] (1)

- \( h(t) \) is a constant: flat wrt time.
- Other functions (really simple!)

\[ S(t) = \exp -\lambda(t) \] (2)

\[ f(t) = \lambda(t) \exp -\lambda(t) \] (3)
Figure: This figure graphs a typical example of the exponential hazard rate.
A more flexible distribution function is given by the Weibull.
Weibull

- A more flexible distribution function is given by the Weibull.
- Named for Waloddi Weibull, who “discovered” it (1939, 1951)
Weibull

- A more flexible distribution function is given by the Weibull.
- Named for Waloddi Weibull, who “discovered” it (1939, 1951)
- Two-parameter distribution; \( h(t) \):

\[
h(t) = \lambda p (\lambda t)^{p-1} \quad t > 0, \lambda > 0, p > 0
\]

\( \lambda \) is positive scale parameter; \( p \) is shape parameter.
Weibull

- A more flexible distribution function is given by the Weibull.
- Named for Waloddi Weibull, who “discovered” it (1939, 1951)
- Two-parameter distribution; \( h(t) \):

\[
h(t) = \lambda p(\lambda t)^{p-1}, \quad t > 0, \lambda > 0, p > 0
\]  

\( \lambda \) is positive scale parameter; \( p \) is shape parameter.
- \( p > 1 \), the hazard rate is \textit{monotonically} increasing with time.
- \( p < 1 \), the hazard rate is \textit{monotonically} decreasing with time.
- \( p = 1 \), the hazard is flat, i.e. \textit{exponential}. 

A more flexible distribution function is given by the Weibull.

Named for Waloddi Weibull, who “discovered” it (1939, 1951)

Two-parameter distribution; $h(t)$:

$$h(t) = \lambda p(\lambda t)^{p-1} \quad t > 0, \lambda > 0, p > 0$$  \hspace{2cm} (4)

$\lambda$ is positive scale parameter; $p$ is shape parameter.

$p > 1$, the hazard rate is *monotonically* increasing with time.

$p < 1$, the hazard rate is *monotonically* decreasing with time.

$p = 1$, the hazard is flat, i.e. *exponential*.

Note that $\lambda$ corresponds to covariates ($\exp \beta_k x_i$)
Figure: This figure graphs three typically shaped Weibull hazard rates. Note the monotonicity of the Weibull hazard; note also that when the shape parameter is 1, the exponential hazard is obtained.
Weibull

- Survivor function

\[ S(t) = \exp - (\lambda t)^p \]  \hspace{1cm} (5)

- PDF

\[ f(t) = \lambda p(\lambda t)^{p-1} \exp - (\lambda t)^p \]  \hspace{1cm} (6)
Cautionary Note

- There are a couple of ways to express the Weibull (exponential)
Cautionary Note

- There are a couple of ways to express the Weibull (exponential)
- (1): Model $h(t)$; (2): Model $\log(T)$
Cautionary Note

- There are a couple of ways to express the Weibull (exponential)
- (1): Model $h(t)$; (2): Model $\log(T)$
- In (1), coefficients relate to the hazard function.
Cautionary Note

- There are a couple of ways to express the Weibull (exponential)
  - (1): Model $h(t)$; (2): Model $\log(T)$
  - In (1), coefficients relate to the hazard function.
  - In (2), coefficients relate to log of the failure time.
There are a couple of ways to express the Weibull (exponential)

(1): Model $h(t)$; (2): Model $\log(T)$

In (1), coefficients relate to the hazard function.

In (2), coefficients relate to log of the failure time.

Signs will differ depending on choice.
Cautionary Note

- There are a couple of ways to express the Weibull (exponential)
- (1): Model $h(t)$; (2): Model $\log(T)$
- In (1), coefficients relate to the hazard function.
- In (2), coefficients relate to log of the failure time.
- Signs will differ depending on choice.
- Stata defaults to (1); R (survreg) defaults to (2).
There are a couple of ways to express the Weibull (exponential)

1. Model $h(t)$;
2. Model $\log(T)$

In (1), coefficients relate to the hazard function.
In (2), coefficients relate to log of the failure time.
Signs will differ depending on choice.
Stata defaults to (1); R (survreg) defaults to (2).
(2) is sometimes called “accelerated failure time”
The Two “Different” Models

- Proportional Hazards:

\[
h(t \mid x) = h_0 t \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_j x_j),
\]  

(7)
The Two “Different” Models

▶ Proportional Hazards:

\[ h(t \mid x) = h_0 t \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_j x_j), \quad (7) \]

▶ Accelerated Failure Time:

\[ \log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \sigma \epsilon, \quad (8) \]

\( \epsilon \) is a stochastic disturbance term with type-1 extreme-value distribution scaled by \( \sigma \).
The Two “Different” Models

- Proportional Hazards:

\[ h(t \mid x) = h_0 t \exp(\alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_j x_j), \]  

(7)

- Accelerated Failure Time:

\[ \log(T) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_j x_j + \sigma \epsilon, \]  

(8)

\( \epsilon \) is a stochastic disturbance term with type-1 extreme-value distribution scaled by \( \sigma \).

- \( \sigma = 1/p. \)
The Two “Different” Models

- **Proportional Hazards:**
  \[
  h(t \mid x) = h_0t \exp(\alpha_1x_{i1} + \alpha_2x_{i2} + \ldots + \alpha_jx_{ij}),
  \]
  (7)

- **Accelerated Failure Time:**
  \[
  \log(T) = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \ldots + \beta_jx_{ij} + \sigma\epsilon,
  \]
  (8)
  
  $\epsilon$ is a stochastic disturbance term with *type-1 extreme-value* distribution scaled by $\sigma$.

- $\sigma = 1/p$.

- $F(\epsilon)$ is a type-1 extreme value distribution.
The Two “Different” Models

- Proportional Hazards:
  \[ h(t \mid x) = h_0t \exp(\alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_j x_{ij}), \quad (7) \]

- Accelerated Failure Time:
  \[ \log(T) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \sigma \epsilon, \quad (8) \]
  \( \epsilon \) is a stochastic disturbance term with type-1 extreme-value distribution scaled by \( \sigma \).
  \( \sigma = 1/p. \)
  \( F(\epsilon) \) is a type-1 extreme value distribution.
  Close connection to Weibull: the distribution of the log of a Weibull distributed random variable yields a type-1 extreme value distribution.
The Two “Different” Models

- Proportional Hazards:

\[ h(t | \mathbf{x}) = h_0(t) \exp(\alpha_1 x_{i1} + \alpha_2 x_{i2} + \ldots + \alpha_j x_{ij}), \]  

(7)

- Accelerated Failure Time:

\[ \log(T) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_j x_{ij} + \sigma \epsilon, \]  

(8)

\( \epsilon \) is a stochastic disturbance term with type-1 extreme-value distribution scaled by \( \sigma \).

- \( \sigma = 1/p \).

- \( F(\epsilon) \) is a type-1 extreme value distribution.

- Close connection to Weibull: the distribution of the log of a Weibull distributed random variable yields a type-1 extreme value distribution.

- Sometimes this parameterization is referred to as a log-Weibull distribution.
## Connection between Parameterizations

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\beta$</td>
<td>$\beta = \frac{-\alpha}{p}$</td>
<td>$+\alpha \equiv \uparrow h(t \mid x_{ij})$</td>
<td>$+\beta \equiv \uparrow \log(T)$</td>
</tr>
<tr>
<td>$\alpha = -\beta p$</td>
<td>$\alpha = -\beta p$</td>
<td>$-\alpha \equiv \downarrow h(t \mid x_{ij})$</td>
<td>$-\beta \equiv \downarrow \log(T)$</td>
<td></td>
</tr>
<tr>
<td>$p$</td>
<td>$\sigma$</td>
<td>$\sigma = \frac{1}{p}$</td>
<td>$p &gt; 1 \equiv \uparrow h(t \mid x_{ij})$</td>
<td>$\sigma &gt; 1 \equiv \downarrow h(t \mid x_{ij})$</td>
</tr>
<tr>
<td>$p = \frac{1}{\sigma}$</td>
<td>$p = \frac{1}{\sigma}$</td>
<td>$p &lt; 1 \equiv \downarrow h(t \mid x_{ij})$</td>
<td>$\sigma &lt; 1 \equiv \uparrow h(t \mid x_{ij})$</td>
<td></td>
</tr>
</tbody>
</table>
Other Distributions: log-logistic and log-normal

- Weibull is monotonic; others are not.
Other Distributions: log-logistic and log-normal

- Weibull is monotonic; others are not.
- Log-logistic and log-normal allow for nonmonotonic hazards:
Other Distributions: log-logistic and log-normal

- Weibull is monotonic; others are not.
- Log-logistic and log-normal allow for nonmonotonic hazards:
- Both estimated only as AFT models:

\[
\log(T) = \beta'_j x + \sigma \epsilon. \tag{9}
\]
Other Distributions: log-logistic and log-normal

- Weibull is monotonic; others are not.
- Log-logistic and log-normal allow for nonmonotonic hazards:
- Both estimated only as AFT models:
  \[
  \log(T) = \beta_j'x + \sigma \epsilon. \tag{9}
  \]
- Neither “more flexible” than Weibull, though (all two parm. distributions)
Log-Logistic

Hazard:

\[ h(t) = \frac{\lambda p(\lambda t)^{p-1}}{1 + (\lambda t)^p} \]  \hspace{1cm} (10)
Log-Logistic

- Hazard:
  \[ h(t) = \frac{\lambda p(\lambda t)^{p-1}}{1 + (\lambda t)^p} \]  
  (10)

- \( h(t) \) increases and then decreases if \( p > 1 \); monotonically decreasing when \( p \leq 1 \).
Figure: This figure graphs some typically shaped hazard rates for the log-logistic model.
Log-Logistic

Survivor function:

\[ S(t) = \frac{1}{1 + (\lambda t)^p}, \quad (11) \]
Log-Logistic

- Survivor function:

\[ S(t) = \frac{1}{1 + (\lambda t)^p}, \]  

(11)

- PDF:

\[ f(t) = \frac{\lambda p(\lambda t)^{p-1}}{(1 + (\lambda t)^p)^2}, \]  

(12)
Log-Normal

- Logistic density is similar to the normal density.
Log-Normal

- Logistic density is similar to the normal density.
- Therefore, log-logistic and log-normal models are often similar.
Log-Normal

- Logistic density is similar to the normal density.
- Therefore, log-logistic and log-normal models are often similar.
- Derivation of the log-normal hazard rate involves integrals of the standard normal distribution.
Log-Normal

- Logistic density is similar to the normal density.
- Therefore, log-logistic and log-normal models are often similar.
- Derivation of the log-normal hazard rate involves integrals of the standard normal distribution.
- Survivor function:

\[ S(t) = 1 - \Phi \left( \frac{\log(t) - \beta'x}{\sigma} \right), \quad (13) \]

where \( \Phi \) is the cumulative distribution function for the standard normal distribution and \( \beta'x \) are the covariates and parameter vector from (9).
Log-Normal

- Logistic density is similar to the normal density.
- Therefore, log-logistic and log-normal models are often similar.
- Derivation of the log-normal hazard rate involves integrals of the standard normal distribution.
- Survivor function:

\[
S(t) = 1 - \Phi \left( \frac{\log(t) - \beta'x}{\sigma} \right),
\]

where \( \Phi \) is the cumulative distribution function for the standard normal distribution and \( \beta'x \) are the covariates and parameter vector from (9)
- PDF:

\[
f(t) = \frac{1}{\sigma \sqrt{(2\pi)}} t^{-1} \exp \left[ - \frac{1}{2} \left( \frac{\log(t) - \beta'x}{\sigma} \right)^2 \right]
\]
Figure: This figure graphs some typically shaped hazard rates for the log-normal model.
Other distributions?

- Gompertz
Other distributions?

- Gompertz
- Rayleigh
Other distributions?

- Gompertz
- Rayleigh
- Gamma
Other distributions?

- Gompertz
- Rayleigh
- Gamma
- ... and others.
Other distributions?

- Gompertz
- Rayleigh
- Gamma
- ... and others.
- The four just considered are most commonly applied in Political Science.
Estimation

- Previous can be estimated through MLE
Estimation

- Previous can be estimated through MLE
- Imagine $n$ observations upon which $t_1, t_2, \ldots t_n$ duration times are measured.
Estimation

- Previous can be estimated through MLE
- Imagine $n$ observations upon which $t_1, t_2, \ldots t_n$ duration times are measured.
- Assume conditional independence of $t_i$ (may be herculean assumption; more later)
Estimation

- Previous can be estimated through MLE
- Imagine $n$ observations upon which $t_1, t_2, \ldots t_n$ duration times are measured.
- Assume conditional independence of $t_i$ (may be herculean assumption; more later)
- Specify a PDF (or CDF); if $f(t)$ is derived, $S(t)$ easily follows (see last week’s slides)
Estimation

- Previous can be estimated through MLE
- Imagine \( n \) observations upon which \( t_1, t_2, \ldots, t_n \) duration times are measured.
- Assume conditional independence of \( t_i \) (may be herculean assumption; more later)
- Specify a PDF (or CDF); if \( f(t) \) is derived, \( S(t) \) easily follows (see last week’s slides)
- Write out likelihood function and maximize (standard algorithm is Newton-Raphson)
Likelihood

- Generic Likelihood:

\[ L = \prod_{i=1}^{n} \left\{ f(t_i) \right\}^{\delta_i} \left\{ S(t_i) \right\}^{1-\delta_i} \]  

(15)

\( \delta \) is censoring indicator.
Likelihood

- **Generic Likelihood:**
  \[
  \mathcal{L} = \prod_{i=1}^{n} \left\{ f(t_i) \right\}^{\delta_i} \left\{ S(t_i) \right\}^{1-\delta_i}
  \]  
  (15)
  
  $\delta$ is censoring indicator.

- **Example: Weibull**
  \[
  f(t) = \lambda p(\lambda t)^{p-1} \exp{-(\lambda t)^p};
  \]
  survivor function
  \[
  S(t) = \exp{-(\lambda t)^p};
  \]
  likelihood of the $t$ duration times:
  \[
  \mathcal{L} = \prod_{i=1}^{n} \left\{ \lambda p(\lambda t)^{p-1} \exp{-(\lambda t)^p} \right\}^{\delta_i} \left\{ \exp{-(\lambda t)^p} \right\}^{1-\delta_i}
  \]  
  (16)
Likelihood

- **Generic Likelihood:**

\[
\mathcal{L} = \prod_{i=1}^{n} \{ f(t_i) \}^{\delta_i} \{ S(t_i) \}^{1-\delta_i}
\]

(15)

\(\delta\) is censoring indicator.

- **Example: Weibull**

\[
f(t) = \lambda p(\lambda t)^{p-1} \exp(-\lambda t)^p;
\]

survivor function

\[
S(t) = \exp(-\lambda t)^p;
\]

likelihood of the \(t\) duration times:

\[
\mathcal{L} = \prod_{i=1}^{n} \{ \lambda p(\lambda t)^{p-1} \exp(-\lambda t)^p \}^{\delta_i} \{ \exp(-\lambda t)^p \}^{1-\delta_i}
\]

(16)

- **All good statistical packages have these functions coded up.**
Adjudication

- Lots of Choices
Adjudication

- Lots of Choices
- Selection can be arbitrary
Adjudication

- Lots of Choices
- Selection can be arbitrary
- If parametrically nested, standard LR tests apply.
Adjudication

- Lots of Choices
- Selection can be arbitrary
- If parametrically nested, standard LR tests apply.
- Encompassing Distribution: generalized gamma:

\[ f(t) = \frac{\lambda p (\lambda t)^{p\kappa - 1} \exp[-(\lambda t)^p]}{\Gamma(\kappa)} \]  

(17)
Adjudication

- Lots of Choices
- Selection can be arbitrary
- If parametrically nested, standard LR tests apply.
- Encompassing Distribution: generalized gamma:

\[ f(t) = \frac{\lambda p(\lambda t)^{p\kappa-1}\exp[-(\lambda t)^p]}{\Gamma(\kappa)} \]  

(17)

- When \( \kappa = 1 \), the Weibull is implied; when \( \kappa = p = 1 \), the exponential distribution is implied; when \( \kappa = 0 \), the log-normal distribution is implied; and when \( p = 1 \), the gamma distribution is implied.
### Adjudication

- Lots of Choices
- Selection can be arbitrary
- If parametrically nested, standard LR tests apply.
- Encompassing Distribution: generalized gamma:

\[
f(t) = \frac{\lambda p(\lambda t)^{p\kappa-1}\exp[-(\lambda t)^p]}{\Gamma(\kappa)}
\]

When \(\kappa = 1\), the Weibull is implied; when \(\kappa = p = 1\), the exponential distribution is implied; when \(\kappa = 0\), the log-normal distribution is implied; and when \(p = 1\), the gamma distribution is implied.

That is, these distributions are encompassed within generalized gamma.
Adjudication

- Lots of Choices
- Selection can be arbitrary
- If parametrically nested, standard LR tests apply.
- Encompassing Distribution: generalized gamma:

\[
f(t) = \frac{\lambda p(\lambda t)^{p\kappa-1}\exp[-(\lambda t)^p]}{\Gamma(\kappa)}
\]  

\[(17)\]

- When \(\kappa = 1\), the Weibull is implied; when \(\kappa = p = 1\), the exponential distribution is implied; when \(\kappa = 0\), the log-normal distribution is implied; and when \(p = 1\), the gamma distribution is implied.
- That is, these distributions are encompassed within generalized gamma.
- Use of AIC: \(-2(\log L) + 2(c + p + 1)\),
where \(c\) denotes the number of covariates in the model and \(p\) denotes the number of structural parameters for the model.
Or you could just estimate a Cox Model

▸ Developed by Sir David Cox
Or you could just estimate a Cox Model

- Developed by Sir David Cox
- Fundamentally, an important achievement of 20c. statistics.
Or you could just estimate a Cox Model

- Developed by Sir David Cox
- Fundamentally, an important achievement of 20c. statistics.
- “Workhorse” model in many fields.
Or you could just estimate a Cox Model

- Developed by Sir David Cox
- Fundamentally, an important achievement of 20c. statistics.
- “Workhorse” model in many fields.
- Objective: estimate the impact of the covariates on the hazard rate, without specifying the distribution of the duration dependency.
Cox Model Moving Parts

- Hazard:

\[ h_i(t) = h_0(t) \exp(\beta'x) \]  \hspace{1cm} (18)

where \( h_0(t) \) is the baseline hazard function, and \( \beta'x \) are the covariates and regression parameters.
Cox Model Moving Parts

▶ Hazard:

\[ h_i(t) = h_0(t) \exp(\beta'x) \]  \hspace{1cm} (18)

where \( h_0(t) \) is the baseline hazard function, and \( \beta'x \) are the covariates and regression parameters.

▶ Proportional Hazards:

\[
\frac{h_i(t)}{h_0(t)} = \exp[\beta'(x_i - x_j)],
\]  \hspace{1cm} (19)
Cox Model Moving Parts

- Hazard:

\[ h_i(t) = h_0(t) \exp(\beta'x) \]  

(18)

where \( h_0(t) \) is the baseline hazard function, and \( \beta'x \) are the covariates and regression parameters.

- Proportional Hazards:

\[ \frac{h_i(t)}{h_0(t)} = \exp[\beta'(x_i - x_j)], \]  

(19)

- Look familiar? (Think: proportional odds)
Cox Model Moving Parts

- Hazard:
  \[ h_i(t) = h_0(t) \exp(\beta'x) \] (18)
  where \( h_0(t) \) is the baseline hazard function, and \( \beta'x \) are the covariates and regression parameters.

- Proportional Hazards:
  \[ \frac{h_i(t)}{h_0(t)} = \exp[\beta'(x_i - x_j)], \] (19)

- Look familiar? (Think: proportional odds)
- PH means the hazard ratio is a fixed proportion across time (i.e. \( \exp(\beta) \)).
Cox Model Moving Parts

- Hazard:
  \[ h_i(t) = h_0(t) \exp(\beta'x) \]  
  where \( h_0(t) \) is the baseline hazard function, and \( \beta'x \) are the covariates and regression parameters.

- Proportional Hazards:
  \[ \frac{h_i(t)}{h_0(t)} = \exp[\beta'(x_i - x_j)], \]  
  Look familiar? (Think: proportional odds)

- PH means the hazard ratio is a fixed proportion across time (i.e. \( \exp(\beta) \)).

- \( h_0(t) \) is assumed to be unknown and is left unparameterized. This differs considerably from the parametric case.
Cox Model

Cox regression models do not have an intercept term.

\[ h_i(t) = \exp(\beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}) h_0(t), \]  

(20)

Or as log of the hazard ratios:

\[ \log \left\{ \frac{h_i(t)}{h_0(t)} \right\} = \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \]  

(21)

Neither (20) nor (21) contains a constant term \( \beta_0 \). This term is "absorbed" into the baseline hazard function. This isn't a problem.
Partial Likelihood

To accomplish the goal of deriving parameter estimates without specifying $h_0(t)$, Cox developed “partial likelihood.”
Partial Likelihood

- To accomplish the goal of deriving parameter estimates without specifying $h_0(t)$, Cox developed “partial likelihood.”
- Partial likelihood assumes intervals between successive failure times contributes no information on relationship between covariates and hazard rate.
Partial Likelihood

- To accomplish the goal of deriving parameter estimates without specifying \( h_0(t) \), Cox developed “partial likelihood.”
- Partial likelihood assumes intervals between successive failure times contributes no information on relationship between covariates and hazard rate.
- This rate is assumed to have an arbitrary form and could actually be zero in the intervals between successive failures.
Partial Likelihood

- To accomplish the goal of deriving parameter estimates without specifying $h_0(t)$, Cox developed “partial likelihood.”
- Partial likelihood assumes intervals between successive failure times contributes no information on relationship between covariates and hazard rate.
- This rate is assumed to have an arbitrary form and could actually be zero in the intervals between successive failures.
- It is the *ordered failure times* rather than interval between failure times that contributes information to the partial likelihood function.
Partial Likelihood

- To accomplish the goal of deriving parameter estimates without specifying $h_0(t)$, Cox developed “partial likelihood.”
- Partial likelihood assumes intervals between successive failure times contributes no information on relationship between covariates and hazard rate.
- This rate is assumed to have an arbitrary form and could actually be zero in the intervals between successive failures.
- It is the *ordered failure times* rather than interval between failure times that contributes information to the partial likelihood function.
- Parametric methods use all information on $T$; Cox models use only a part of the information; hence “partial” likelihood methods.
The Logic (following D. Collett’s [2003] approach)

Table: Data Sorted by Ordered Failure Time

<table>
<thead>
<tr>
<th>Case Number</th>
<th>Duration Time</th>
<th>Censored Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>7</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>21</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>30</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>45</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>46</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>51</td>
<td>No</td>
</tr>
</tbody>
</table>

Data are sorted by the duration time. The duration time for censored cases denotes the time of last observation.
The Logic

Figure: Duration times for nine censored and uncensored (failed) cases.
The Logic

- What are the main features of these data?
  - Events can be ordered.
  - At $t_0$ all cases are at risk of failing.
  - After the first failure, the risk set decreases by 1.
  - The risk set successively dwindles as events occur.
The Logic

- What are the main features of these data?
  - Events can be ordered.
  - At $t_0$ all cases are at risk of failing.
  - After the first failure, the risk set decreases by 1.
  - The risk set successively dwindles as events occur.

- To motivate the partial likelihood estimator, let $\psi = \exp(\beta'x_i)$ (this notation is from Collett, 1994, p. 64).
The partial likelihood function for these data would be equivalent to:

\[
L_p = \frac{\psi(7)}{\psi(1) + \psi(2) + \psi(3) + \psi(4) + \psi(5) + \psi(6) + \psi(7) + \psi(8) + \psi(9)} \times \\
\frac{\psi(4)}{\psi(1) + \psi(2) + \psi(3) + \psi(4) + \psi(5) + \psi(6) + \psi(8) + \psi(9)} \times \\
\frac{\psi(5)}{\psi(1) + \psi(2) + \psi(3) + \psi(5) + \psi(6) + \psi(8) + \psi(9)} \times \\
\frac{\psi(3)}{\psi(1) + \psi(3) + \psi(6) + \psi(8)} \times \\
\frac{\psi(1)}{\psi(1) + \psi(6)} \times \\
\frac{\psi(6)}{\psi(6)}.
\]
More Formally

- Suppose we have $n$ observations and $k$ distinct failure times.
More Formally

- Suppose we have $n$ observations and $k$ distinct failure times.
- Cox estimation first proceeds by sorting the ordered failure times: $t_1 < t_2 < \ldots < t_k$, where $t_i$ denotes the failure time for the $i$th individual.
More Formally

► Suppose we have \( n \) observations and \( k \) distinct failure times.

► Cox estimation first proceeds by sorting the ordered failure times: \( t_1 < t_2 < \ldots < t_k \), where \( t_i \) denotes the failure time for the \( i \)th individual.

► Censoring: define \( \delta_i \) to be 0 if the case is right-censored, and 1 if the case is uncensored.
More Formally

- Suppose we have $n$ observations and $k$ distinct failure times.
- Cox estimation first proceeds by sorting the ordered failure times: $t_1 < t_2 < \ldots < t_k$, where $t_i$ denotes the failure time for the $i$th individual.
- Censoring: define $\delta_i$ to be 0 if the case is right-censored, and 1 if the case is uncensored.
- Ordered event times are modeled as a function of covariates: $x$
More Formally

- Suppose we have $n$ observations and $k$ distinct failure times.
- Cox estimation first proceeds by sorting the ordered failure times: $t_1 < t_2 < \ldots < t_k$, where $t_i$ denotes the failure time for the $i$th individual.
- Censoring: define $\delta_i$ to be 0 if the case is right-censored, and 1 if the case is uncensored.
- Ordered event times are modeled as a function of covariates: $x$
- Partial likelihood function is derived by taking the product of the conditional probability of a failure at time $t_i$, given the number of cases that are at risk of failing at time $t_i$. 
More Formally

- Suppose we have $n$ observations and $k$ distinct failure times.
- Cox estimation first proceeds by sorting the ordered failure times: $t_1 < t_2 < \ldots < t_k$, where $t_i$ denotes the failure time for the $i$th individual.
- Censoring: define $\delta_i$ to be 0 if the case is right-censored, and 1 if the case is uncensored.
- Ordered event times are modeled as a function of covariates: $x$
- Partial likelihood function is derived by taking the product of the conditional probability of a failure at time $t_i$, given the number of cases that are at risk of failing at time $t_i$.
- In words: given that some event has occurred, what is the probability the event occurred to the $i$th individual from a risk set of size $n$?
Define $R(t_i)$ to denote the number of cases that are at risk of experiencing an event at time $t_i$. 
Partial Likelihood

- Define $R(t_i)$ to denote the number of cases that are at risk of experiencing an event at time $t_i$.
- This is the “risk set.”
Partial Likelihood

- Define $R(t_i)$ to denote the number of cases that are at risk of experiencing an event at time $t_i$.
- This is the “risk set.”
- The probability that the $j$th case will fail at time $T_i$ is given by

$$\Pr(t_j = T_i \mid R(t_i)) = \frac{e^{\beta'x_i}}{\sum_{j \in R(t_i)} e^{\beta'x_j}} \quad (22)$$

The summation operator in the denominator is summing over all individuals in the risk set.
Taking the product of the conditional probabilities in (22) yields the partial likelihood function:

\[
L_p = \prod_{i=1}^{K} \left[ \frac{e^{\beta'x_i}}{\sum_{j \in R(t_i)} e^{\beta'x_j}} \right] ^{\delta_i},
\]  

(23)
Partial Likelihood

- Taking the product of the conditional probabilities in (22)
yields the partial likelihood function:

\[ L_p = \prod_{i=1}^{K} \left[ \frac{e^{\beta'x_i}}{\sum_{j \in R(t_i)} e^{\beta'x_j}} \right]^{\delta_i} \]  \hspace{1cm} (23)

- With corresponding log-likelihood function,

\[ \log L_p = \sum_{i=1}^{K} \delta_i \left[ \beta'x_i - \log \sum_{j \in R(t_i)} e^{\beta'x_j} \right]. \]  \hspace{1cm} (24)
Importance

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
Importance

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the PL function.
Importance

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the PL function.
- Censored cases contribute information only pertinent to the risk set (i.e. the denominator, not the numerator).
Importance

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the PL function.
- Censored cases contribute information only pertinent to the risk set (i.e. the denominator, not the numerator)
- The critical thing here is to note that no assumptions about the shape of the baseline hazard need to be made.
Importance

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the PL function.
- Censored cases contribute information only pertinent to the risk set (i.e. the denominator, not the numerator)
- The critical thing here is to note that no assumptions about the shape of the baseline hazard need to be made.
- Another way to see this is to think about the heuristic partial likelihood function above. All we need to know to compute a probability is $\psi$ (or $\exp \beta' x_i$).
Importance

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the PL function.
- Censored cases contribute information only pertinent to the risk set (i.e. the denominator, not the numerator)
- The critical thing here is to note that no assumptions about the shape of the baseline hazard need to be made.
- Another way to see this is to think about the heuristic partial likelihood function above. All we need to know to compute a probability is $\psi$ (or $\exp \beta' x_i$).
- Cox (1972, 1975) showed that maximum partial likelihood estimation produces parameter estimates that have the same properties as maximum likelihood estimates.
Importance

- Specifying the baseline hazard, $h_0(t)$ is unnecessary.
- The interval between events does not inform the PL function.
- Censored cases contribute information only pertinent to the risk set (i.e. the denominator, not the numerator).
- The critical thing here is to note that no assumptions about the shape of the baseline hazard need to be made.
- Another way to see this is to think about the heuristic partial likelihood function above. All we need to know to compute a probability is $\psi$ (or $\exp \beta' x_i$).
- Cox (1972, 1975) showed that maximum partial likelihood estimation produces parameter estimates that have the same properties as maximum likelihood estimates.
- Ties can be an issue: don’t use Stata defaults!
Discrete-Time Models

- Another approach entails use of models for discrete data.
Discrete-Time Models

- Another approach entails use of models for discrete data.
- Important to note that duration data can be recorded as binary sequence.
Another approach entails use of models for discrete data.

Important to note that duration data can be recorded as binary sequence.

Import of Beck, Katz, and Tucker (1998) was to show political scientists this fact.
Discrete-Time Models

- Another approach entails use of models for discrete data.
- Important to note that duration data can be recorded as binary sequence.
- Import of Beck, Katz, and Tucker (1998) was to show political scientists this fact.
- If so, logit, probit, cloglog, (generic binary link models) can be fit to duration data.
## Discrete Data

**Table: Example of Discrete-Time Event History Data**

<table>
<thead>
<tr>
<th>Case I.D.</th>
<th>Event Occurrence</th>
<th>Year</th>
<th>Elapsed</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1974</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1975</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1986</td>
<td>13</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>1987</td>
<td>14</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1974</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>1974</td>
<td>1</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>1975</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>1992</td>
<td>19</td>
</tr>
<tr>
<td>45</td>
<td>0</td>
<td>1993</td>
<td>20</td>
</tr>
</tbody>
</table>

These data are a portion of a data set originally analyzed in Brace, Hall, and Langer (1999). I thank Laura Langer for letting us use them.
Let the random variable $T$ denote a discrete random variable indicating the time of an event occurrence.
Let the random variable $T$ denote a discrete random variable indicating the time of an event occurrence.

PMF:

\[ f(t) = \Pr(T = t_i) \quad (25) \]

Gives the probability of an event occurring at time $t_i$. 
Discrete-Time Moving Parts

- Let the random variable $T$ denote a discrete random variable indicating the time of an event occurrence.
- PMF:
  \[ f(t) = \Pr(T = t_i) \quad (25) \]
  Gives the probability of an event occurring at time $t_i$.
- The survivor function for the discrete random variable $T$ is given by
  \[ S(t) = \Pr(T \geq t_i) = \sum_{j \geq i} f(t_j), \quad (26) \]
  where $j$ denotes a failure time.
Let the random variable $T$ denote a discrete random variable indicating the time of an event occurrence.

**PMF:**

$$f(t) = \Pr(T = t_i) \quad (25)$$

Gives the probability of an event occurring at time $t_i$.

The survivor function for the discrete random variable $T$ is given by

$$S(t) = \Pr(T \geq t_i) = \sum_{j \geq i} f(t_j), \quad (26)$$

where $j$ denotes a failure time.

Connection between $f(t)$ and $S(t)$:

$$h(t) = \frac{f(t)}{S(t)}, \quad (27)$$
Likelihood

Note $h(t)$ as conditional failure probability:

$$h(t) = \Pr(T = t_i \mid T \geq t_i, x).$$  (28)
Likelihood

- Note $h(t)$ as conditional failure probability:
  \[
  h(t) = \Pr(T = t_i \mid T \geq t_i, \mathbf{x}).
  \]

- Likelihood: $\exists n$ cases observed over $t$ periods. For each observation, the dependent variable is a binary indicator coded 1 if an event occurs and 0 if an event does not occur at time $t$. 

Where $T$ is the event time, $t_i$ is the observed event time, and $\mathbf{x}$ represents the vector of covariates.
Likelihood

- Note $h(t)$ as conditional failure probability:
\[
h(t) = \Pr(T = t_i \mid T \geq t_i, x). \tag{28}\]

- Likelihood: $\exists n$ cases observed over $t$ periods. For each observation, the dependent variable is a binary indicator coded 1 if an event occurs and 0 if an event does not occur at time $t$.

- The likelihood of the data set is:
\[
\mathcal{L} = \prod_{i}^{n} \left[ h(t_i) \prod_{i=1}^{t-1} (1 - h(t_i)) \right]^{y_{it}} \left[ \prod_{i=1}^{t} (1 - h(t_i)) \right]^{1-y_{it}} \tag{29}\]

(Derivation can be found in B-S and Jones (pp. 71–72).
Likelihood

- Note $h(t)$ as conditional failure probability:
  \[
  h(t) = \Pr(T = t_i \mid T \geq t_i, x) \quad (28)
  \]

- Likelihood: \( \exists n \) cases observed over \( t \) periods. For each observation, the dependent variable is a binary indicator coded 1 if an event occurs and 0 if an event does not occur at time \( t \).

- The likelihood of the data set is:
  \[
  \mathcal{L} = \prod_{i=1}^{n} \left[ h(t_i) \prod_{i=1}^{t-1} (1 - h(t_i)) \right]^{y_{it}} \left[ \prod_{i=1}^{t} (1 - h(t_i)) \right]^{1 - y_{it}} \quad (29)
  \]
  
  (Derivation can be found in B-S and Jones (pp. 71–72).

- This is equivalent to:
  \[
  \mathcal{L} = \prod_{i=1}^{n} \left\{ f(t) \right\}^{y_{it}} \left\{ S(t) \right\}^{1 - y_{it}} \quad (30)
  \]

  Which looks very similar to other likelihood functions.
Models

- Binary data that are event history data.
Models

- Binary data that are event history data.
- Of interest:

\[ h(t) = \Pr(T = t_i \mid T \geq t_i, x). \]  

(31)
Models

- Binary data that are event history data.
- Of interest:
  \[ h(t) = \Pr(T = t_i \mid T \geq t_i, x). \] (31)
- Let \( \Pr(y_{it} = 1) = \lambda_i \), and the probability of a nonoccurrence as \( \Pr(y_{it} = 0) = 1 - \lambda_i \).
Models

- Binary data that are event history data.
- Of interest:
  \[ h(t) = \Pr(T = t_i \mid T \geq t_i, x). \]  \hspace{1cm} (31)
- Let \( \Pr(y_{it} = 1) = \lambda_i \), and the probability of a nonoccurrence as \( \Pr(y_{it} = 0) = 1 - \lambda_i \).
- Specify:
  \[ \lambda_{it} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \]  \hspace{1cm} (32)
Models

- Binary data that are event history data.
- Of interest: 
  \[ h(t) = \Pr(T = t_i \mid T \geq t_i, x). \] (31)
- Let \( \Pr(y_{it} = 1) = \lambda_i \), and the probability of a nonoccurrence as \( \Pr(y_{it} = 0) = 1 - \lambda_i \).
- Specify: 
  \[ \lambda_{it} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \] (32)
- Now we’re cooking with gas.
Some Models

- Logit:

$$\log \left( \frac{\lambda_i}{1 - \pi} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}.$$  \hspace{1cm} (33)
Some Models

- **Logit:**

  \[
  \log \left( \frac{\lambda_i}{1 - \pi} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \tag{33}
  \]

- **Probit:**

  \[
  \Phi^{-1}[\lambda_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}, \tag{34}
  \]
Some Models

- **Logit:**

\[
\log \left( \frac{\lambda_i}{1 - \pi} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \tag{33}
\]

- **Probit:**

\[
\Phi^{-1} [\lambda_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}, \tag{34}
\]

- **Complementary log-log:**

\[
\log[- \log(1 - \lambda_i)] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki}. \tag{35}
\]
Some Models

- Logit:
  \[
  \log \left( \frac{\lambda_i}{1 - \pi} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} . \tag{33}
  \]

- Probit:
  \[
  \Phi^{-1} [\lambda_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} , \tag{34}
  \]

- Complementary log-log:
  \[
  \log [-\log (1 - \lambda_i)] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_k x_{ki} . \tag{35}
  \]

- Others?
An Issue: Time Dependency

- Discrete models have a touchstone with parametrics.
An Issue: Time Dependency

- Discrete models have a touchstone with parametrics.
- Basic Model:

\[
\log \left( \frac{\lambda_i}{1 - \lambda_i} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{36}
\]

where \( x_{ki} \) are two covariates of interest that have a mean of 0, and \( \beta_0 \) is the constant term.
An Issue: Time Dependency

- Discrete models have a touchstone with parametrics.

- Basic Model:
  \[
  \log \left( \frac{\lambda_i}{1 - \lambda_i} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{36}
  \]
  where \(x_{ki}\) are two covariates of interest that have a mean of 0, and \(\beta_0\) is the constant term.

- The “baseline” hazard:
  \[
  \hat{\lambda}_i = h_0(t) = e^{\beta_0}, \tag{37}
  \]
  which is a constant.
An Issue: Time Dependency

- Discrete models have a touchstone with parametrics.
- Basic Model:
  \[
  \log \left( \frac{\lambda_i}{1 - \lambda_i} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i},
  \]
  where \( x_{ki} \) are two covariates of interest that have a mean of 0, and \( \beta_0 \) is the constant term.
- The “baseline” hazard:
  \[
  \hat{\lambda}_i = h_0(t) = e^{\beta_0},
  \]
  which is a constant.
- We’ve seen this.
An Issue: Time Dependency

- Discrete models have a touchstone with parametrics.
- Basic Model:

\[
\log \left( \frac{\lambda_i}{1 - \lambda_i} \right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{36}
\]

where \( x_{ki} \) are two covariates of interest that have a mean of 0, and \( \beta_0 \) is the constant term.

- The “baseline” hazard:

\[
\hat{\lambda}_i = h_0(t) = e^{\beta_0}, \tag{37}
\]

which is a constant.

- We’ve seen this.
- Exponential
An Issue: Time Dependency

What do you do?
An Issue: Time Dependency

What do you do?

Ignore it? Bad strategy usually.
An Issue: Time Dependency

- What do you do?
- Ignore it? Bad strategy usually.
- Other choices?
  - Piecewise Functions
  - Transformations on $t$
  - Smoothing functions (like splines, lowess, etc.)
Next Time

- Applications of all this stuff.
- Diagnostics for Cox Model.
- Other issues.