Lecture 3: Duration Models and Maximum Likelihood

Chris Conlon

NYU Stern

February 24, 2019

Consider a linear regression with $\varepsilon_i | X_i \sim N(0, \sigma^2)$

$$Y_{it} = X_i' \beta_i + \varepsilon_i$$

We've discussed the least squares estimator:

$$\widehat{\beta}_{ols} = \arg\min_{\beta} \sum_{i=1}^{N} (Y_i - X_i' \beta)^2$$

$$\widehat{\beta}_{ols} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Review: What is a Likelihood?

Suppose we write down the joint distribution of our data (y_i, x_i) for $i = 1, \ldots, n$.

$$Pr(y_1, \ldots, y_n, x_1, \ldots, x_n|\theta)$$

If (y_i, x_i) are I.I.D then we can write this as:

$$Pr(y_1, \ldots, y_n, x_1, \ldots, x_n|\theta) = \prod_{i=1}^{N} Pr(y_i, x_i|\theta) \propto \prod_{i=1}^{N} Pr(y_i|x_i, \theta) = L(\mathbf{y}|\mathbf{x}, \theta)$$

We call this $L(\mathbf{y}|\mathbf{x}, \theta)$ the likelihood of the observed data.

If we know the distribution of ε_i we can construct a maximum likelihood estimator

$$(\widehat{\beta}_{MLE}, \widehat{\sigma}^2_{MLE}) = \arg\min_{\beta,\sigma^2} L(\beta, \sigma^2)$$

Where

$$L(\beta, \sigma^2) = \prod_{i=1}^{N} p(y_i | x_i, \beta, \sigma^2)$$

$$= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - X_i'\beta)^2\right]$$

$$l(\beta, \sigma^2) = \sum_{i=1}^{N} -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2$$

MLE: FOC's

Take the FOC's

$$l(\beta, \sigma^2) = -\frac{N}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2$$

Where

$$\frac{\partial l(\beta, \sigma^2)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{N}(Y_i - X_i'\beta) = 0 \rightarrow \widehat{\beta}_{MLE} = \widehat{\beta}_{OLS}$$

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma^2} = -N\frac{1}{2\sigma^2} - \frac{1}{2\sigma^4}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2 = 0$$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i'\beta)^2$$

Note: the unbiased estimator uses $\frac{1}{N-K-1}$.

MLE: General Case

1. Start with the joint density of the data $Z_1, \ldots, Z_N$ with density $f_Z(z, \theta)$
2. Construct the likelihood function of the sample $z_1, \ldots, z_n$

$$L(\mathbf{z}|\theta) = \prod_{i=1}^{N} f_Z(z_i, \theta)$$

3. Construct the log likelihood (this has the same $\arg\max$)

$$l(\mathbf{z}|\theta) = \sum_{i=1}^{N} \ln f_Z(z_i, \theta)$$

4. Take the FOC's to find $\widehat{\theta}_{MLE}$

$$\theta : \frac{\partial l(\theta)}{\partial \theta} = 0$$

Consider the following example:

- Unemployment durations from 479 unskilled workers
- Characteristics: [age, local unemp rate, replacement ratio]
- Economic theory of job-search
  - Receive offers arriving at some rate $\lambda(t)$ so that expected number of jobs is $\lambda(t)dt$.
  - Each offer is: wage $w \sim F_W(w)$.
  - Compare to reservation wage $w > \overline{w}(t) \rightarrow$ Accept (otherwise reject)
  - Probability of acceptance is $1 - F_W(\overline{w}(t))$.

Example: Constant Arrival Rate

Now we have that $\lambda(t)dt = \lambda dt$

- Optimal reservation wage is constant so that $\theta = \lambda(1 - F_W(\overline{w}))$
- Implied distribution for the duration of an unemployment spell is exponential

$$f_Y(y) = \theta \exp(-y\theta)$$

- Exponential distribution is common for waiting times (memorylessness property)

$$E[Y - c|Y > c] = \frac{1}{\theta}$$

- Distribution has mean $\frac{1}{\theta}$ and variance $\frac{1}{\theta^2}$

We have defined what is known as a (constant) Hazard Model

- Survivor Function: $S(y) = 1 - F_Y(y) = \exp(-y\theta)$
- Hazard Function: $\lim_{dy \to 0^+} \frac{Pr(y<Y<y+dy}{Pr(y<Y)} = \frac{f_Y(y)}{S(y)} = \theta$
- Exponential has constant hazard property (We will show this later).

Taking the Model to Data: Perfect Data

Suppose we have data on Exact Failure Times

- This is the easy one, we see the exact unemployment duration for everyone y_i.
- We can just write down the density of observing each duration for exactly y_i.

$$L(\theta) = \prod_{i=1}^{N} f(y_i|\theta) = \prod_{i=1}^{N} h(y_i|\theta)S(y_i|\theta)$$

Taking the Model to Data: Indicator

Suppose we have data on Indicator for Survival

- We see a group of people become unemployed, and we see which are still unemployed c time later.
- But we don't see anything else

$$L(\theta) = \prod_{i=1}^{N} F(c|\theta)^{d_i}(1 - F(c|\theta))^{1-d_i} = \prod_{i=1}^{N}(1 - S(c|\theta))^{d_i} S(c|\theta)^{1-d_i}$$

- This is exactly what the Survivor Function tells us about

Suppose we have data on Observation over Fixed Period of Time

- We see who is still unemployed after c amount of time (just an indicator)
- We see exact duration of unemployment if $y_i < c$.
- Our data are Right Censored

$$L(\theta) = \prod_{i=1}^{N} f\left(y_i|\theta\right)^{d_i} \cdot S(c|\theta)^{1-d_i} = \prod_{i=1}^{N} h\left(y_i|\theta\right)^{d_i} \cdot S\left(y_i|\theta\right)^{d_i} \cdot S(c|\theta)^{1-d_i}$$

- Helpful to define $t_i = \min(y_i, c) = d_i \cdot y_i + (1 - d_i) \cdot c$ is the minimum of the actual *duration* and the censoring time

$$L(\theta) = \prod_{i=1}^{N} h\left(t_i|\theta\right)^{d_i} \cdot S\left(t_i|\theta\right)$$

- Recall $f(y|\theta) = \theta \exp(-y\theta)$

$$L(\theta) = \theta^{\sum_{i=1}^{N} d_i} \exp\left(-\sum_{i=1}^{N} t_i\theta\right)$$

- the MLE is

$$\hat{\theta}_{mle} = \sum_{i=1}^{N} d_i / \sum_{i=1}^{N} t_i = 1/(\bar{t}/\bar{d}) = \bar{d}/\bar{t}$$

Given the MLE $\widehat{\theta}_{MLE} = \frac{\overline{d}}{\overline{t}}$.

Two Bad ideas:

- Pretend that $y_i = c$ for people still unemployed at c
  - Pretend Censored observations ($d_i = 0$) exited $\theta = \frac{1}{\overline{t}}$.
  - Overestimates θ because $\overline{d} \to 1$.
- Ignore individuals who did not exit before c
  - Ignore censored obervations and estimate $\theta = \frac{\sum d_i}{\sum d_i t_i}$.
  - Underestimates θ because $\sum_{i=1} t_i \to \sum_{i=1} t_i d_i$ in denominator.

Suppose individuals differ in censoring time c_i

- Assume $c_i \perp y_i$.

$$L(\theta) = \prod_{i=1}^{N} f\left(y_i|\theta\right)^{d_i} \cdot S\left(c_i|\theta\right)^{1-d_i} = \prod_{i=1}^{N} f\left(t_i|\theta\right)^{d_i} \cdot S\left(t_i|\theta\right)^{1-d_i} = \prod_{i=1}^{N} h\left(t_i|\theta\right)^{d_i} \cdot S\left(t_i|\theta\right)$$

A Different Sampling Method

- All methods assume we see individuals when they enter unemployment.
- Suppose we just sample individuals from stock of unemployed.
- Imagine we draw someone who has been unmployed for $s_i = 3$ weeks and finds a job after a duration of $s_i = 9$ weeks
- Let s_i be duration when we first observe them, this gives:

$$L(\theta) = \prod_{i=1}^{N} f\left(y_i|\theta\right)/S\left(s_i|\theta\right) = \prod_{i=1}^{N} h\left(y_i|\theta\right) \cdot \frac{S\left(y_i|\theta\right)}{S\left(s_i|\theta\right)}$$

- In general we need to know how long somone has been unemployed when we first see them.
- To deal with left censoring we probably need more assumptions.

Basic Setup: we know $F(z|\theta_0)$ but not θ_0. We know $\theta_0 \in \Theta \subset \mathbb{R}^K$.

- Begin with a sample of z_i from $i = 1, \ldots, N$ which are I.I.D. with CDF $F(z|\theta_0)$.
- The MLE chooses

$$\widehat{\theta}_{MLE} = \arg\max_{\theta} l(\theta) = \arg\max_{\theta} \sum_{i=1}^{N} \ln f_Z(z_i, \theta)$$

MLE: Technical Details

1. Consistency. When is it true that for $\epsilon > 0$?

$$\lim_{N\to\infty} \Pr\left(\left\|\hat{\theta}_{mle} - \theta_0\right\| > \varepsilon\right) = 0$$

2. Asymptotic Normality. What else do we need to show?

$$\sqrt{N}\left(\hat{\theta}_{mle} - \theta_0\right) \xrightarrow{d} \mathcal{N}\left(0, -\left[E\frac{\partial^2}{\partial\theta\partial\theta'}\left(Z_i, \theta_0\right)\right]^{-1}\right)$$

3. Optimization. How to we obtain $\widehat{\theta}_{MLE}$ anyway?

MLE: Example # 1

- $Z_i \sim N(\theta_0, 1)$ and $\Theta = (-\infty, \infty)$. In this case:

$$l(\theta) = -N \cdot \ln(2\pi) - \sum_{i=1}^{N} (z_i - \theta)^2 /2$$

- MLE is $\widehat{\theta}_{MLE} = \bar{z}$ which is consistent for $\theta_0 = E[Z_i]$
- Asymptotic distribution is $\sqrt{N}(\bar{z} - \theta_0) \sim N(0, 1)$.
- Calculating mean is easy!

MLE: Example # 2

- $Z_i = (Y_i, X_i)$— X_i has finite mean and variance (but arbitrary distribution)
- $(Y_i|X_i x) \sim N(x'\beta_0, \sigma_0^2)$

$$\widehat{\beta}_{MLE} = (X'X)^{-1}X'Y$$
$$\widehat{\sigma}^2_{MLE} = \frac{1}{N}\sum(y_i - x_i\widehat{\beta}_{MLE})^2$$

- We already have shown consistency and AN for linear regression with normally distributed errors...

- $Z_i = (Y_i, X_i)$— X_i has finite mean and variance (but arbitrary distribution)
- $Pr(Y_i = 1|X_i x) = \frac{e^{x'\theta_0}}{1+e^{x'\theta_0}}$
- Solution is the logit model.
- No simple MLE solution, establishing properties is not obvious...

Jensen's Inequality

Let $g(z)$ be a convex function. Then $\mathbb{E}[g(Z)] \geq g(\mathbb{E}[Z])$, with equality only in the case of a linear function.

Define Y as the ratio of the density at θ to the density at the true value θ_0 both evaluated at Z

$$Y = \frac{f_Z(Z;\theta)}{f_Z(Z;\theta_0)}$$

- Let $g(a) = -\ln(a)$ so that $g'(a) = \frac{-1}{a}$ and $g''(a) = \frac{1}{a^2}$.
- Then by Jensen's Inequality $\mathbb{E}[-\ln Y] \geq -\ln \mathbb{E}[Y]$.
- This gives us

$$\mathbb{E}_Z\left[-\ln\left(\frac{f_Z(Z;\theta)}{f_Z\,(Z;\theta_0)}\right)\right] \geq -\ln\left(\mathbb{E}_Z\left[\frac{f_Z(Z;\theta)}{f_Z\,(Z;\theta_0)}\right]\right)$$

- The RHS is

$$\mathbb{E}_Z\left[\frac{f_Z(Z;\theta)}{f_Z\,(Z;\theta_0)}\right] = \int \frac{f_Z(z;\theta)}{f_Z\,(z;\theta_0)} \cdot f_Z\,(z;\theta_0)\,dz = \int f_Z(z;\theta)dz = 1$$

Because $\log(1) = 0$ this implies:

$$\mathbb{E}_z\left[-\ln\left(\frac{f_Z(Z;\theta)}{f_Z\,(Z;\theta_0)}\right)\right] \geq 0$$

Therefore

$$\begin{aligned} -\mathbb{E}\left[\ln f_Z(Z;\theta)\right] + \mathbb{E}\left[\ln f_Z\,(Z;\theta_0)\right] &\geq 0 \\ \mathbb{E}\left[\ln f_Z\,(Z;\theta_0)\right] &\geq \mathbb{E}\left[\ln f_Z(Z;\theta)\right] \end{aligned}$$

- We maximize the expected value of the log likelihood at the true value of θ!
- Helpful to work with $E[\log f(z;\theta)]$ sometimes.

Information Matrix Equality

We can relate the Fisher Information to the Hessian of the log-likelihood

$$\mathcal{I}(\theta_0) = -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial\theta\partial\theta}(z;\theta_0)\right] = \mathbb{E}\left[\frac{\partial \ln f}{\partial\theta}(z;\theta_0) \times \frac{\partial \ln f}{\partial\theta}(z;\theta_0)'\right]$$

- This is sometimes known as the outer product of scores.
- This matrix is negative definite
- Recall that $\mathbb{E}\left[\frac{\partial \ln f}{\partial\theta}(z;\theta_0)\right] \approx 0$ at the maximum

Proof

$$1 = \int_z f_Z(z;\theta)dz \Rightarrow 0 = \frac{\partial}{\partial\theta}\int_z f_Z(z;\theta)dz$$

With some regularity conditions

$$0 = \int_z \frac{\partial f_Z}{\partial\theta}(z;\theta)dz = \underbrace{\int_z \frac{\partial \ln f_Z}{\partial\theta}(z;\theta)\cdot f_Z(z;\theta)dz}_{\mathbb{E}\left[\frac{\partial \ln f_Z}{\partial\theta}(z;\theta_0)\right]}$$

- This gives us the FOC we needed.
- Can get information identity with another set of derivatives.

We can relate the Fisher Information to the Hessian of the log-likelihood

$$\mathcal{I}(\theta) = -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial\theta\partial\theta'}(Z|\theta)\right]$$

It turns out this provides a bound on the variance

$$\mathrm{Var}(\hat{\theta}(Z)) \geq \mathcal{I}(\theta_0)^{-1}$$

Because we can't do better than Fisher Information we know that MLE is most efficient estimator!

MLE: Discussion

Tradeoffs

- How does this compare to GM Theorem?
- If MLE is most efficient estimate, why ever use something else?

Exponential Example

$$f_{Y|X}(y|x,\beta_o) = e^{x'\beta_o}\exp\left(-ye^{x'\beta_o}\right)$$

With log likelihood

$$l(\beta) = \sum_{i=1}^{N} \ln f_{Y|X}\left(y_i|x_i,\beta\right) = \sum_{i=1}^{N} X_i'\beta - y_i \cdot \exp\left(x_i'\beta\right)$$

And Score, Hessian, and Information Matrix:

$$\begin{aligned}
\mathcal{S}_i(y_i,x_i,\beta) &= x_i'\left(1 - y_i\exp\left(x_i'\beta\right)\right)\\
\mathcal{H}_i(y_i,x_i,\beta) &= -y_i x_i x_i' \exp\left(x_i'\beta\right)\\
\mathcal{I}\left(\beta_o\right) &= \mathbb{E}\left[YXX'\exp\left(X'\beta_o\right)\right] = \mathbb{E}\left[XX'\right]
\end{aligned}$$

Computing Maximum Likelihood Estimators

Consider the Taylor series for $f(x)$ approximated around $f(x_0)$:

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0) + f''(x_0) \cdot (x - x_0)^2 + o_p(3)$$

Suppose we wanted to find a root of the equation where $f(x^*) = 0$ and solve for x:

$$0 = f(x_0) + f'(x_0) \cdot (x - x_0)$$
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

This gives us an iterative scheme to find x^*:

1. Start with some x_k. Calculate $f(x_k), f'(x_k)$
2. Update using $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$
3. Stop when $|x_{k+1} - x_k| < \epsilon_{tol}$.

Newton-Raphson for Minimization

We can re-write optimization as root finding;

- We want to know $\hat{\theta} = \arg\max_\theta \ell(\theta)$.
- Construct the FOCs $\frac{\partial \ell}{\partial \theta} = 0 \rightarrow$ and find the zeros.
- How? using Newton's method! Set $f(\theta) = \frac{\partial \ell}{\partial \theta}$

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 \ell}{\partial \theta^2}(\theta_k)\right]^{-1} \cdot \frac{\partial \ell}{\partial \theta}(\theta_k)$$

The SOC is that $\frac{\partial^2 \ell}{\partial \theta^2} > 0$. Ideally at all θ_k.
This is all for a single variable but the multivariate version is basically the same.

Newton's Method: Multivariate

Start with the objective $Q(\theta) = -l(\theta)$:

- Approximate $Q(\theta)$ around some initial guess θ_0 with a quadratic function
- Minimize the quadratic function (because that is easy) call that θ_1
- Update the approximation and repeat.

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial Q}{\partial\theta}(\theta_k)$$

- The equivalent SOC is that the Hessian Matrix is positive semi-definite (ideally at all θ).
- In that case the problem is globally convex and has a unique maximum that is easy to find.

Back to Duration Example

Let $Z_i = (Y_i, X_i)$ and assume that $(Y_i|X_i = X)$ $Exp(\lambda)$ so that hazard rate is $exp[x'\beta_0]$ and $E[Y_i|X_i = x] = \exp(-x_i'\beta_0)$. This extends the exponential duration model to include covariates $x_i'\beta$

$$f(y|x, \beta_0) = e^{x'\beta_0} \exp\left(-ye^{x'\beta_0}\right)$$

This gives the log-likelihood

$$\ell(\beta) = \sum_{i=1}^{N} \ln f(y_i|x_i, \beta) = \sum_{i=1}^{N} x_i'\beta - y_i \cdot \exp\left(x_i'\beta\right)$$

With derivatives (No analytic solution!)

$$\frac{\partial \ell}{\partial \beta}(\beta) = \sum_{i=1}^{N} x_i \cdot \left(1 - y_i \cdot \exp\left(x_i'\beta\right)\right)$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'}(\beta) = -\sum_{i=1}^{N} x_i x_i' \cdot y_i \cdot \exp\left(x_i'\beta\right)$$

We can generalize to Quasi-Newton methods:

$$\theta_{k+1} = \theta_k - \lambda_k \underbrace{\left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1}}_{A_k} \frac{\partial Q}{\partial\theta}(\theta_k)$$

Two Choices:

- Step length λ_k
- Step direction $d_k = A_k \frac{\partial Q}{\partial \theta}(\theta_k)$
- Often rescale the direction to be unit length $\frac{d_k}{\|d_k\|}$.
- If we use A_k as the true Hessian and $\lambda_k = 1$ this is a full Newton step.

Newton's Method: Alternatives

Choices for A_k

- $A_k = I_k$ (Identity) is known as gradient descent or steepest descent
- BHHH. Specific to MLE. Exploits the Fisher Information.

$$A_k = \left[\frac{1}{N}\sum_{i=1}^{N}\frac{\partial \ln f}{\partial \theta}(\theta_k)\frac{\partial \ln f}{\partial \theta'}(\theta_k)\right]^{-1}$$
$$= -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta \partial \theta'}(Z,\theta^*)\right] = \mathbb{E}\left[\frac{\partial \ln f}{\partial \theta}(Z,\theta^*)\frac{\partial \ln f}{\partial \theta'}(Z,\theta^*)\right]$$

- Alternatives SR1 and DFP rely on an initial estimate of the Hessian matrix and then approximate an update to A_k.
- Usually updating the Hessian is the costly step.
- Non invertible Hessians are bad news.

Back to Duration Models

Simple cases:

- The simplest cases are single irreversible transitions
  - Alive → Dead
  - Working → Failure
- Other easy cases are “resetting” processes:
  - Employed → unemployed for zero weeks, one week, etc.
  - Healthy → Sick Day 1, Sick Day 2, etc.
  - Not on strike → Strike Day 1, Strike Day 2, etc.
- Let’s start with these before we worry about multivariate outcomes or more complicated cases.

Decisions

Have to make some decisions first

1. Do we model spell length directly or probability of transition?
   - Most of the time we want to work with probability of transition.
   - If we work with probability of transition, we have to pay attention to frequency
2. What outcomes do we measure: stocks? or flows?
   - Do we measure the number of people who lose/find jobs?
   - Do we measure the number of unemployed people each month?
3. Is the data truncated or censored?
   - People who are still alive are not in the dataset!

For now we will think about single-spells, and measure them using flow data.

Examples

There are lots of different names (depending on your discipline):

- Life table analysis
- Hazard Analysis
- transition analysis
- survival analysis
- failure time analysis

Examples:

- How long does a government last?
- How long does a part last?
- How long before a firm adopts a new technology?
- How long do marriages last?
- How long before criminals re-offend?

Start with a Graph!

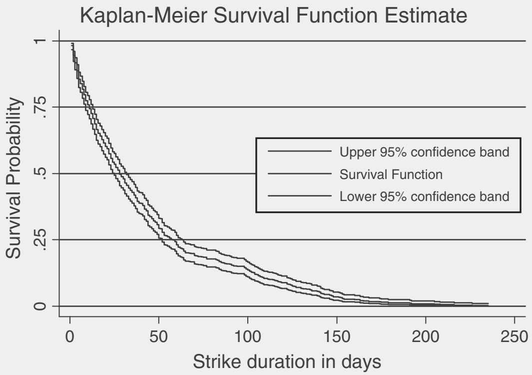

Figure 17.1: Strike duration: Kaplan-Meier estimate of survival function. Data on completed spells for 566 strikes in the U.S. during 1968–76.

What did we just plot?

The empirical survival function

- We ignored any covariates, including calendar time.
- The x-axis was the duration
- The y-axis was the fraction of observations still alive “alive” after x periods.
- If nothing is infinitely lived then the graph always starts at 1 and always ends at zero.
- If things are infinitely lived we call the duration distribution defective.

Let's start with some deeply parametric stuff

- density function: $f(t) = dF(t)/dt$: unconditional probability of instantaneous failure
- CDF: $F(t) = Pr(T \leq t) = \int_0^\infty f(s)ds$. (Probability that spell is less than length t).
- Survival Function: $S(t) = 1 - F(t) = Pr(T > t)$. This has the nice property that it integrates to expected duration $\int_0^\infty S(t)dt = E[T]$.
- Hazard Function: $\lambda(t) = \lim_{\Delta t \to 0} \frac{Pr[t \leq T < t + \Delta t | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)}$.
- All of these functions represent the same information!

More about hazard functions

- Hazard is conditional probability of leaving unemployment after being unemployed for t.
- Hazard is percentage change in survivor function $S(t)$
- Hazard also gives us the distribution of duration T:

$$\lambda(t) = -\frac{\partial \log S(t)}{\partial t}, \quad S(t) = \exp\left[-\int_0^{\infty} \lambda(u)du\right]$$

- Often we'd like to estimate $\lambda(t|x)$ instead of $E[T|x]$ especially since we often have censored data so that $\lambda(t|x)$ is still well defined but $E[T|x]$ is not.
- In practice $\lambda(t|x)$ can be tricky to estimate (especially since it may contain zeros at some t in finite sample. Solution: Cumulative Hazard Function.

$$\Lambda(t) = \int_0^{\infty} \lambda(s)ds = -\log S(t)$$

- Just like we preferred to estimate CDF instead of PDF!

Table 17.1. *Survival Analysis: Definitions of Key Concepts*

Function	Symbol	Definition	Relationships
Density	$f(t)$		$f(t) = dF(t)/dt$
Distribution	$F(t)$	$\Pr[T \leq t]$	$F(t) = \int_0^t f(s)ds$
Survivor	$S(t)$	$\Pr[T > t]$	$S(t) = 1 - F(t)$
Hazard	$\lambda(t)$	$\lim_{h \to 0} \frac{\Pr[t \leq T < t + h \mid T \geq t]}{h}$	$\lambda(t) = f(t)/S(t)$
Cumulative hazard	$\Lambda(t)$	$\int_0^t \lambda(s)ds$	$\Lambda(t) = -\ln S(t)$

What about Discrete Time?

- Maybe we only see survival annually/weekly/etc. not actual failure time.
- Basic idea is the same. Have to be careful about ties. Divide failures into t_j buckets

$$
\begin{aligned}
\lambda_j &= Pr[T = t_j | T \geq t_j] = f^d(t_j)/S^d(t_{j-}) \\
\Lambda^d(t) &= \sum_{j|t_j \leq t} \lambda_j \\
S^d &= Pr[T \geq t] = \prod_{j|t_j \leq t} (1 - \lambda_j)
\end{aligned}
$$

- Can define the product integral which is regular product in discrete case and exponential of integral in continuous case.

- Without censoring, things are easy: just let

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(T_i \geq t).$$

- if you want a smooth hazard function, take a smooth estimator, e.g. (with some "small" bandwidth $w > 0$)

$$\hat{S}(t) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{1 + \exp((t - T_i)/w)},$$

- and then take minus the derivative of the log of this estimate.

What if there is censoring? Kaplan-Meier!

- We define the ordered durations as

$$T_{(1)} < \ldots < T_{(n)},$$

- let d_j be the number of observations i for which $T_i = T_{(j)}$
- Let m_j number of spells censored in $[t_j, t_{j+1})$
- and r_j the cardinality of the risk set at duration t_{j-} $r_j = \sum_{l|l\geq j} d_l + m_l$
- Simple estimate of the hazard function $\hat{\lambda}_j = \frac{d_j}{r_j}$.
- Kaplan-Meier estimator of the survival function is the Product Limit Estimator

$$\hat{S}(t) = \Pi_{j|t_j\leq t}\left(1 - \frac{d_j}{r_j}\right) = \Pi_{j|t_j\leq t}\left(\frac{r_j - d_j}{r_j}\right)$$

- It is normally distributed (asymptotically), with (Greenwood) variance

$$\hat{V}[\hat{S}(t)] = (\hat{S}(t))^2 \cdot \sum_{j|t_j\leq t} \frac{d_j}{r_j(r_j - d_j)}.$$

Other stuff

Think about what happens when $m_j = 0$ (no censoring)

- $r_j = \sum_{l|l \geq j} d_l + m_l \rightarrow r_{j+1} = r_j - d_j$.
- $\hat{S}(t) = \prod_{j|t_j \leq t} \left(\frac{r_j - d_j}{r_j} \right) = \prod_{j|t_j \leq t} \frac{r_{j+1}}{r_j} = \frac{r_j}{N}$
- Again – exactly what we would expect – one minus the ECDF.

How do we deal with ties?

- Lots of ties can create problems. Implicitly we assume all deaths are at same time in period.
- Why does this matter– well how many are remaining in r_j?
- r_j is potentially biased if we have lots of ties.
- Can either try corrections or sample data at higher frequency

Exponential and Weibull

- The exponential is popular because it has a constant hazard rate $\lambda(t) = \gamma$ which does not depend on t.
- This is often referred to as the memorylessness property of the exponential.
- This is analytically convenient but it makes it hard to fit things in practice (you only have one parameter!)
- The Weibull is a generalization with $\lambda(t) = \gamma\alpha t^{\alpha-1}$. For $\alpha = 1$ we have exponential.
- For $\alpha > 1$ i is increasing and for $\alpha < 1$ it is decreasing (monotonically).
- Weibull used to be popular in econometrics for simple parametric analysis.

Table 17.4. *Exponential and Weibull Distributions: pdf, cdf, Survivor Function, Hazard, Cumulative Hazard, Mean, and Variance*

Function	**Exponential**	**Weibull**
$f(t)$	$\gamma \exp(-\gamma t)$	$\gamma \alpha t^{\alpha-1} \exp(-\gamma t^{\alpha})$
$F(t)$	$1 - \exp(-\gamma t)$	$1 - \exp(-\gamma t^{\alpha})$
$S(t)$	$\exp(-\gamma t)$	$\exp(-\gamma t^{\alpha})$
$\lambda(t)$	γ	$\gamma \alpha t^{\alpha-1}$
$\Lambda(t)$	γt	γt^{α}
$\mathrm{E}[T]$	γ^{-1}	$\gamma^{-1/\alpha}\Gamma(\alpha^{-1} + 1)$
$\mathrm{V}[T]$	γ^{-2}	$\gamma^{-2/\alpha}[\Gamma(2\alpha^{-1} + 1) - [\Gamma(\alpha^{-1} + 1)]^2]$
γ, α	$\gamma > 0$	$\gamma > 0, \alpha > 0$

Comparison of Parametric Models

Table 17.5. *Standard Parametric Models and Their Hazard and Survivor Functions*[a]

Parametric Model	Hazard Function	Survivor Function	Type
Exponential	γ	$\exp(-\gamma t)$	PH, AFT
Weibull	$\gamma\alpha t^{\alpha-1}$	$\exp(-\gamma t^{\alpha})$	PH, AFT
Generalized Weibull	$\gamma\alpha t^{\alpha-1} S(t)^{-\mu}$	$[1-\mu\gamma t^{\alpha}]^{1/\mu}$	PH
Gompertz	$\gamma \exp(\alpha t)$	$\exp(-(\gamma/\alpha)(e^{\alpha t}-1))$	PH
Log-normal	$\frac{\exp\left(-(\ln t-\mu)^2/2\sigma^2\right)}{t\sigma\sqrt{2\pi}[1-\Phi((\ln t-\mu)/\sigma)]}$	$1-\Phi\left((\ln t-\mu)/\sigma\right)$	AFT
Log-logistic	$\alpha\gamma^{\alpha}t^{\alpha-1}/\left[(1+(\gamma t)^{\alpha})\right]$	$1/\left[1+(\gamma t)^{\alpha}\right]$	AFT
Gamma	$\frac{\gamma(\gamma t)^{\alpha-1}\exp[-(\gamma t)]}{\Gamma(\alpha)[1-I(\alpha,\gamma t)]}$	$1-I(\alpha,\gamma t)$	AFT

[a] All the parameters are restricted to be positive, except that $-\infty < \alpha < \infty$ for the Gompertz model.

Adding Covariates

- We can also add covariates by letting $\gamma = \beta X$.
- Sometimes this is called link function or generalized linear models similar to what we saw with the logit or probit.
- It is usually a bad idea to link more than one nonlinear parameter this way.
- We would typically estimate via MLE. Writing down the full-data log-likelihood is straightforward.
- A frequently used special-case are proportional hazard models

The Proportional Hazard Model

With covariates x, the hazard function is $h(t|x)$; we specify

$$\lambda(t|x) = \lambda_0(t)\phi(x).$$

- λ_0 and ϕ are up to a positive multiplicative constant.)
- We call λ_0 the baseline hazard; every individual has a hazard that is just a proportional version of the baseline hazard.

The baseline hazard could be:

- constant: the survival function is exponential
- a power function $\lambda_0(t) = \gamma t^\alpha$; e.g. for $\alpha < 0$ we have negative duration dependence (the long-term unemployed...)
- more complicated (flexible) specifications.

Maximum likelihood: works for any parametric modelx $\lambda(t|x,\beta)$ of the full hazard function;
(here: w/o censoring, without corrleation across individuals):

$$\max_{\beta} \sum_{i=1}^{n} \ln f(T_i|x_i,\beta),$$

where $f(t|x,\beta)$ is the density of the duration T induced by λ:

$$f(t|x) = \lambda(t|x)S(t|x) = \lambda(t|x)\exp(-\Lambda(t|x)),$$

so the log-likelihood for i is just $\ln \lambda(T_i|x_i,\beta) - \Lambda(T_i|x_i,\beta)$.

What's the point?

- The (partial) additive separability of the log-likelihood in the PH model is designed to make our lives easier.
- Presumably, we specified λ so that its integral Λ is easy to compute.
- For PH: the log-likelihood for i is:
  $\ln \lambda_0(T_i, \beta) + \ln \phi(x_i, \beta) - \Lambda_0(T_i, \beta)\phi(x_i, \beta)$.
- The most common choice is $\phi(x_i, \beta) = \exp(x_i\beta)$ so that $\ln \phi(x_i, \beta) = x_i\beta$.
- In that case we have that $\partial\lambda/\partial x_j = \beta_j \cdot \lambda$.
- One remaing problem: what to do with the baseline hazard function (is that even identified?).

Cox's Partial Likelihood for the PH Model

- if we do not want to assume anything about the shape of the baseline hazard function
- but we are happy specifying $\phi(x, \beta)$
- then we will only look at the *order* of the durations: we reorder individuals so that $T_{i_1} < \ldots < T_{i_n}$
- ...and we forget about the durations! Then the partial likelihood is:

$$\sum_{j=1}^{n} \left(\ln \phi(x_{i_j}, \beta) - \ln \left(\sum_{l=j}^{n} \phi(x_{i_l}, \beta) \right) \right).$$

- This is a limited information maximum likelihood estimator. It is not fully efficient!
- But it may be robust to mis-specifying λ_0. Is it actually a valid likelihood? not sure!.

How did that work?

Once we have ordered everything:

- Let $R(t_j)$ be the set of spells at risk (still alive) at t_j
- d_j are the deaths at time t_j $\sum_l \mathbf{1}[t_l = t_j]$.
- Consider only at-risk spells ending a fixed t_j

$$
\begin{aligned}
Pr[T_j = t_j | R(t_j)] &= \frac{Pr[T_j = t_j | T_j \geq t_j]}{\sum_{l \in R(t_j)} Pr[T_l = t_l | T_l \geq t_j]} \\
&= \frac{\lambda_j(t_j | x_j, \beta)]}{\sum_{l \in R(t_j)} \lambda_l(t_j, x_l, \beta)} \\
&= \frac{\phi(x_j, \beta)}{\sum_{l \in R(t_j)} \phi(x_l, \beta)}
\end{aligned}
$$

- λ_0 drops out because of PH.

- *Intuition:* those individuals who exit first are (on average) those in the risk set whose covariates x give them the largest $\phi(x, \beta)$.
- After we have $\hat{\beta}$ we can estimate the baseline integrated hazard; denoting $N(t)$=number of individuals with $T = t$

$$\widehat{\Lambda_0}(T_{i_j}) = \sum_{m=1}^{j} \frac{N(T_{i_m})}{\sum_{l=m}^{n} \phi(x_{i_l}, \hat{\beta})}.$$

A simple way to test the model:

- just take two different groups of individuals, estimate PH on each, check whether the baseline hazards look proportional **NOT** equal

testing a parametric specification of the baseline hazard $\bar{\Lambda}_0$:

- define generalized residuals $\bar{u}_i = \bar{\Lambda}_0(T_i)$
- Under the true model, for any z

$$\Pr(\bar{u} < z) \simeq \Pr(T_i < \bar{\Lambda}_0^{-1}(z)) = 1 - S_0(\bar{\Lambda}_0^{-1}(z)).$$

- it should be $1 - \exp(-z)$ if $S_0 = \exp(-\bar{\Lambda}_0)$.
- So you can estimate the integrated hazard of $(\bar{u}_1, \ldots, \bar{u}_n)$; it should be $\Lambda_u(z) \equiv z$.

The PH Model is Usually too Restrictive

- **Fact:** the hazard rate of leaving unemployment decreases in time;
- It could be *skimming*: the more able, more willing, better connected find a job faster;
- or it could be "technological": skills deteriorate over time.
- Under the PH model it can only be the latter: negative duration dependence. $\rightarrow$ introduce unobserved heterogeneity:

$$\lambda(t|x, v) = \lambda_0(t)\phi(x)v.$$

- v is a "type" that is unobserved by the econometrician; we only assume that it is uncorrelated with x and independent of t.

Dynamic selection

- The model with v is called the **Mixed PH model** (MPH).
- In the unemployment story: the larger v's have a higher hazard rate, so they find a job faster
- Over time, the distribution of v moves (stochastically) to the left.
- This dynamic selection is a general phenomenon in the MPH model: $\lambda(t|x)$ has "more negative duration dependence" than $\lambda(t|x, v)$.
- Can we test dynamic selection vs true negative duration dependence (λ_0 decreasing)? $\rightarrow$ identification issues.
- This idea shows up in dynamic models of durable goods purchases as well.

We still can recover the aggregate survival function from the data, but now it is a mixture:

$$S^A(t|x) = \Pr(T \geq t|x) = \int \exp(-v\phi(x)\Lambda_0(t))dF(v).$$

- Can we recover ϕ and λ_0 without assuming anything on F?
- Almost … in theory: we just need to assume that $E(v)$ is finite.

A Constructive Proof, 1

- Normalize $Ev = 1$; and $\phi(x_0) = 1$ for some x_0.
- Then the aggregate hazard function is

$$\lambda^A(t|x) = -\frac{\partial \log S^A}{\partial t}(t|x)$$

that is

$$\frac{\int v\phi(x)\lambda_0(t)\exp(-v\phi(x)\Lambda_0(t))dF(v)}{S^A(t|x)}.$$

- Look at $x = x_0$ and $t = 0^+$: then $\Lambda_0(t) \simeq 0$, so

$$\lambda^A(0^+|x_0) = \frac{Ev \times k(x_0) \times \lambda_0(0)}{S^A(0|x_0)} = \lambda_0(0).$$

- and

$$\phi(x) = \frac{\lambda^A(0^+|x)}{\lambda^A(0^+|x_0)}.$$

A Constructive Proof, 2

- Now we can define

$$m^A(t|x) = -\frac{\partial \log S^A(t,x)}{\partial \phi(x)}$$

- and we get the baseline hazard from

$$\frac{\lambda_0(t)}{\Lambda_0(t)} = \frac{\lambda^A(t|x)}{m^A(t|x)};$$

- and we can also recover F.
- In practice we would specify functional forms of course.

Is that Practical?

- We are relying heavily on "identification at 0": that is where we get $\phi(x)$, the rest depends on it.
- Empirical researchers have found that it is often a slim basis (and a very slow-converging estimator)—but anything else will be parametric.
- The alternative is to use richer data: multiple durations/multiple spells.

E.g. Cahuc/Postel-Vinay-Robin, *Econometrica* 2006.

- Workers are heterogeneous, so are firms;
- a worker quits when he gets a better outside offer (exogenous Poisson(λ)).
- We observe (given matched employer-employee data):
  - job durations (how long each worker stays in a job)
  - and distributions of wages (mostly) across firms.

Bad luck

- The likelihood for the duration of job spells is independent of heterogeneity!

$$f(t) = \frac{\delta(\delta + \lambda)}{\lambda} \int_{\delta t}^{(\delta+\lambda)t} \frac{\exp(-x)}{x} dx.$$

- So we can identify λ and δ, and nothing about heterogeneity of firms and workers.
- (But the good thing is that we don't need to assume anything about it and we get δ and λ).

- Given bargaining on wages, outside options matter;
- and outside options generate option values, which increase with heterogeneity (volatility!).
- “So” by looking at the distribution of wages we can infer heterogeneity.

Application 2: moral hazard in insurance

Abbring-Chiappori-Pinquet, *JEEA* 2003.

- Insurees have exogenous types (risk) v that are unobserved; we call this adverse selection;
- they also decide to adopt a risky behavior or not: moral hazard.
- Data typically gives us a series of claims for each individual.
- A state could be: “I have had exactly p claims so far” and a spell is the time between two claims.

Duration dependence

- Adverse selection induces positive duration dependence: the time between claims is positively correlated.
- On the other hand, with experience rating a claim (at fault) increases premia and makes risky behavior more costly—typically
- so moral hazard induces negative duration dependence.
- How can we test for the latter while controlling for the former?

The Model

- The hazard function for claim $(p+1)$ at t, given state p, is (dropping x)

$$vh_0(t)A^{-p},$$

- with A and h_0 unknown.
- v models exogeneous unobserved risk,
- every time a claim occurs, the hazard for the next claim is divided by A: moral hazard.
- It is the MPH, with a twist: the p.

Estimating Finite Mixtures

- In practice estimating finite mixture models can be tricky.
- A simple example is the mixture of normals (incomplete data likelihood)

$$f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^{N} \sum_{k=1}^{K} \pi_k f(x_i|\mu_k, \sigma_k)$$

- We need to find both mixture weights $\pi_k = Pr(z_k)$ and the components (μ_k, σ_k) the weights define a valid probabiltiy measure $\sum_k \pi_k = 1$.
- Easy problem is label switching. Usually it helps to order the components by say decreasing $\pi_1 > \pi_2 > \ldots$ or $\mu_1 > \mu_2 > \ldots$
- The real problem is that which component you belong to is unobserved. We can add an extra indicator variable $z_{ik} \in \{0, 1\}$.
- We don't care about z_{ik} per-se so they are nuisance parameters.

- We can write the complete data log-likelihood (as if we observed z_{ik}):

$$l(x_1, \ldots, x_n|\theta) = \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} I[z_i = k]\pi_k f(x_i \mu_k, \sigma_k)\right)$$

- We can instead maximized the expected log-likelihood where we take the expectation $E_{z|\theta}$

$$\alpha_{ik}(\theta) = Pr(z_{ik} = 1|x_i, \theta) = \frac{f_k(x_i, z_k, \mu_k, \sigma_k)\pi_k}{\sum_{m=1}^{K} f_m(x_i, z_m, \mu_m, \sigma_m)\pi_m}$$

- Now we have a probability $\hat{\alpha}_{ik}$ that gives us the probability that i came from component k. We also compute $\hat{\pi}_k = \frac{1}{N}\sum_{i=1}^{N} \alpha_{ik}$

- Treat the $\hat{\alpha}_k(\theta^{(q)})$ as data and maximize to find μ_k, σ_k for each k

$$\hat{\theta}^{(q+1)} = \arg\max_{\theta} \sum_{i=1}^{N} \log\left(\sum_{k=1}^{K} \hat{\alpha}_k(\theta^{(q)}) f(x_i|z_{ik}, \theta)\right)$$

- We iterate between updating $\hat{\alpha}_k(\theta^{(q)})$ (E-step) and $\hat{\theta}^{(q+1)}$ (M-step)
- For the mixture of normals we can compute the M-step very easily:

$$\begin{aligned} \mu_k^{(q+1)} &= \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_k(\theta^{(q)}) x_i \\ \sigma_k^{(q+1)} &= \frac{1}{N} \sum_{i=1}^{N} \hat{\alpha}_k(\theta^{(q)}) (x_i - \overline{x})^2 \end{aligned}$$

- EM algorithm has the advantage that it avoids complicated integrals in computing the expected log-likelihood over the missing data.
- For a large set of families it is proven to converge to the MLE
- That convergence is monotonic and linear. (Newton's method is quadratic)
- This means it can be slow, but sometimes $\nabla_\theta f(\cdot)$ is really complicated.

Thanks!