[10] Duration Model

- Reference:
  - Chung, Schmidt and Witte (1991), Journal of Quantitative Criminology.
  - Kiefer (1988), Journal of Economic Literature.
  - LIMDEP Manual.

(1) Basic Model

- T = length of time until an event (E) occurs.
- Examples:
  - Recidivism in criminology:

    E = returning to prison.

    T = interval from release ro returning.
  - Unemployment Duration

    E = quit job search (get a job or give up)

    T = time length for a job search.
  - Time interval between equity trades or exchange rate quotes.

    [Engle and Russell, ECON, 1998]

- Assume that T is continuous:

  $T \sim$ pdf, $f(t|\theta)$; cdf $F(t|\theta)$.

- Some important concepts:

1) $F(t|\theta) = Pr(T < t)$.

  - Probability of failure within t.
  - Probability that E occurs within t.

    ex: probability of returning to prison within t;

    ex: Probability of quitting job search within t.

2) Probability of survival: $S(t|\theta) = 1 - F(t|\theta) = Pr(T \geq t)$.

  - Probability that E does not occurs with t.

    ex: probability that an ex-prisoner does not in prison until t.

    ex: probability that job search still goes on until t.

  - In general, $S(t|\theta) \rightarrow 0$ as $t \rightarrow \infty$.

3) Hazard function:

  - $h(t|\theta) = f(t|\theta)/Pr(T \geq t) = f(t|\theta)/S(t|\theta) = f(t|\theta)/[1-F(t|\theta)]$.
  - Probability that E occurs just after time t conditional on no failure prior to t.

    ex: probability of returning prison right after t.

    ex: probability of quitting search right after t.

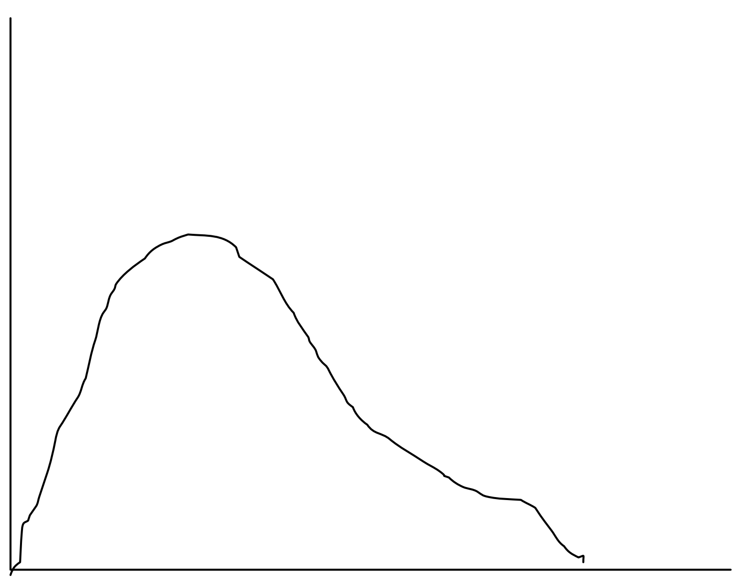
4) Intergrated hazard function.

- H(t|θ) = $\int_0^t h(x|\theta)dx$.
- S(t|θ) = exp[-H(t|θ)] or H(t|θ) = $-\ln S(t|\theta)$.
- h(t|θ) = $-\frac{d[\ln S(t|\theta)]}{dt}$.
- If you know s(t|θ), we can get h(t|θ), and vice versa.

5) State Dependence:

- $\frac{dh(t|\theta)}{dt} > 0$: positive state dependence.
- $\frac{dh(t|\theta)}{dt} < 0$: negative state dependence.

Example: Recidivism



(2) Maximum Likelihood Estimation

- t_i^* = the i'th individual's actual exit time
- c_i = follow-up time for i (time interval following i).
- $t_i = \begin{cases} t_i^* \; if \; t_i^* < c_i; \\ c_i \; if \; t_i^* \geq c_i. \end{cases} = \min(t_i^*, c_i).$
- $d_i = \begin{cases} 1 \; if \; t_i < c_i; \\ 0 \; if \; t_i = c_i. \end{cases}$
- $$\begin{aligned} l_N(\theta) &= \sum_{t_i < c_i} \ln f(t_i \mid \theta) + \sum_{t_i = c_i} \Pr(t_i^* \geq c_i) \\ &= \sum_{t_i < c_i} \ln f(t_i \mid \theta) + \sum_{t_i = c_i} \ln[1 - F(c_i \mid \theta)] \\ &= \sum_{t_i < c_i} \ln f(t_i \mid \theta) + \sum_{t_i = c_i} \ln[S(c_i \mid \theta)] \\ &= \Sigma_{i=1}^N \{ d_i \ln f(t_i \mid \theta) + (1 - d_i) \ln S(c_i \mid \theta) \} \end{aligned}$$

(3) Distributions

1) Exponential:

- For simplicity, suppress "i".
- $f(t|\theta) = \lambda\exp(-\lambda t)$, where $\theta = \lambda > 0$.
- $S(t|\theta) = \exp(-\lambda t)$
- $h(t|\theta) = \lambda$ (no state dependence).
- $E(t) = 1/\lambda$ [expected duration]; $var(t) = 1/\lambda$.

2) Weibull:

- $t^p \sim$ Exponential(λ).
- $f(t|\theta) = p\lambda^p t^{p-1}\exp[-(\lambda t)^p]$, where $\theta = (\lambda, p)'$.
- $S(t|\theta) = \exp[-(\lambda t)^p]$.
- $h(t|\theta) = \lambda p(\lambda t)^{p-1}$.

$$\rightarrow \frac{dh(t|\theta)}{dt} = (p-1)\lambda^p p t^{p-2} \begin{cases} > 0 \; if \;\; p > 1; \\ = 0 \; if \;\; p = 1; \\ < 0 \; if \;\; p < 1. \end{cases}$$

- $E(t) = \frac{1}{\lambda}\Gamma\left(\frac{1}{p}+1\right)$, where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1}e^{-y}dy$.

<Poof>

Let $y = (\lambda t)^p$; then, $t = \frac{1}{\lambda} y^{\frac{1}{p}}$ and $dt = \frac{1}{p\lambda} y^{\frac{1}{p}-1} dy$. Hence,

$$E(t) = \int_0^\infty t f(t,\theta) dt = \int_0^\infty \frac{1}{\lambda} y^{\left(\frac{1}{p}+1\right)-1} \exp(-y) dy$$

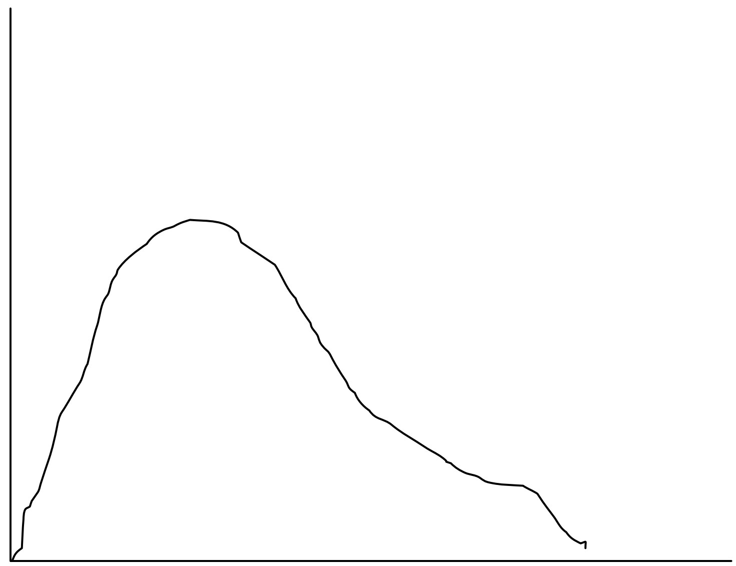
$$= \frac{1}{\lambda} \Gamma\left(\frac{1}{p} + 1\right),$$

where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} \exp(-y) dy$.

3) Log-normal:

- $\ln(t) \sim N(\mu, \sigma^2)$, where $\mu = -\ln(\lambda)$ and $\sigma = 1/p$.
- $f(t \mid \theta) = \frac{p}{\sqrt{2\pi t}} \exp\left(-\frac{p^2}{2}(\ln t + \ln \lambda)\right)$.
- $S(t|\theta) = 1 - \Phi\left(\frac{\ln t + \ln \lambda}{\sigma}\right)$.
- $E(t) = \frac{1}{\lambda} + \exp\left(\frac{1}{2p^2}\right)$; $E(\ln t) = -\ln \lambda$.

- $h(t|\theta)$ = complicated, but



4) Log-logistic

- $f(t\,|\,\theta) = \dfrac{\lambda p(\lambda t)^{p-1}}{\left(1+(\lambda t)^{2p}\right)^2}$, $\theta = (\lambda, p)'$.

  $\rightarrow$ f(w) = $\dfrac{\exp(-w)}{[1+\exp(-w)]^2}$, where $w = -p\ln(\lambda t)$.

- $S(t\,|\,\theta) = \dfrac{1}{1+(\lambda t)^p}$; $h(t\,|\,\theta) = \dfrac{\lambda p(\lambda t)^{p-1}}{1+(\lambda t)^p}$.

- $\dfrac{dh(t\,|\,\theta)}{dt} = \lambda^p p t^{p-2} \dfrac{(p-1)-(\lambda t)^p}{\left(1+(\lambda t)^p\right)^2}$.

  $\rightarrow$ Negative if p < 1; negative if t is large; positive if p > 1 and t is small.

5) Gamma:

- $f(t\,|\,\theta) = \dfrac{\lambda p(\lambda t)^{p\xi-1}}{\Gamma(\xi)}\exp\left(-(\lambda t)^p\right)$, where $\theta = (\lambda, p, \xi)'$.

- Weibull if $\xi = 1$ and exponential if $\xi = p = 1$.

(4) Estimation

- Set $\frac{1}{\lambda_i} = \exp(x_i'\beta)$ and $\sigma = \frac{1}{p}$ so that $\beta_j > 0$ means "x_{ji} prolongs duration."
- LIMDEP estimates β and σ.
- Caution!
  - We here assume that x_i does not change during the follow-up period, c_i.
  - This may cause dome misspecification problem. For the cases of time-varying x_i, see Heckman and Singer (Social Science Duration Analysis, 1985, Ch. 2).

(5) Checking Specification

- Consequences of distributional misspecification:
  - $\hat{\beta}_{ML}$ would be inconsistent.
  - Even if the ML estimator may be consistent for some special cases, it could be inconsistent. When it is consistent, use $(H_N)^{-1}B_N(H_N)^{-1}$ to estimate $Cov(\hat{\beta}_{ML})$. [See Gourieroux, Monfort and Trognon (ECON, 1984).]

- Vuong Test [Vuong (ECON, 1989)]
  - Wish to decide on which of two competing models would be more plausible.
  - Here, the goal is not to find the correct model, but to find a better model between $f(t_i|x_i,\theta)$ and $g(t_i|z_i,\gamma)$. Both models could be misspecified.

[CASE 1] Nonnested Models

- EX: Weibull Vs. Log-normal.
- Let θ_* and γ_* be the maximizers of $E[f(t|x,\theta)]$ and $E[g(t|z,\gamma)]$, respectively.
- $\omega_*^2 = \text{var}\left(\log\frac{f(y|x,\theta_*)}{g(y|z,\gamma_*)}\right)$.
- $H_o : E\left[\log\frac{f(y|x,\theta_*)}{g(y|z,\gamma_*)}\right] = 0$;

  $H_f : E\left[\log\frac{f(y|x,\theta_*)}{g(y|z,\gamma_*)}\right] > 0$; $H_g : E\left[\log\frac{f(y|x,\theta_*)}{g(y|z,\gamma_*)}\right] < 0$.
- Let $\hat{\theta}_N$ and $\hat{\gamma}_N$ be the ML estimators based on $f(t_i|x_i,\theta)$ and $g(t_i|z_i,\gamma)$, respectively.

- Define:

$$\hat{l}_i = \log\frac{f(t_i|x_i,\hat{\theta}_N)}{g(t_i|z_i,\hat{\gamma}_N)};\ \hat{l}_N = \Sigma_{i=1}^N \hat{l}_i;$$

$$\hat{\omega}_N^2 = \frac{1}{N}\Sigma_{i=1}^N\left[\hat{l}_i\right]^2 - \left[\frac{1}{N}\hat{l}_N\right]^2;\ \hat{\omega}_N = \sqrt{\hat{\omega}_N^2};$$

$$t_N = \sqrt{N}\frac{\hat{l}_N}{\hat{\omega}_N}.$$

- Under H_o, $t_N \to N(0,1)$; Under H_f, $t_N \to +\infty$; Under H_g, $t_N \to -\infty$.

[CASE 2] Nested Models

- Exponential vs. Weibull with same regressors $x_i = z_i$.
- Assume $g(y_i | z_i, \gamma_*) \subset f(y_i | x_i, \theta_*)$.
- $H_o : g(y_i | z_i, \gamma_*) = f(y_i | x_i, \theta_*)$;

  $H_a : g(y_i | z_i, \gamma_*) \neq f(y_i | x_i, \theta_*)$
- The usual LR statistic, $2\hat{l}_N$, is not χ^2 under H_o if both models are misspecificed [in fact, a weighted χ^2].
- If the general model [$f(y_i|x_i,\theta_*)$] is correctly specified, then ,the LR statistic is χ^2 under H_o.

- Rule of Thumb Methods:
  1) Compare the ML results to non-parametric estimates of h(t) or s(t) ["Kaplan-Meier" estimates].
  2) Consider whether estimated parameters are reasonably signed. [Compare to Cox' proportional hazard model (a semi-parametric model).]

(6) Heteroskedasticity

- Suppose that $\theta = \begin{pmatrix} \beta \\ \sigma \end{pmatrix}$ differs across different i (θ_i). If you estimate a model incorrectly assuming θ constant, the estimated hazard functions tend to be biased to negative duration dependence [Heckman and Singer, 1985, Ch. 2].

- Heckman and Singer's suggestion (ECON, 1984)
  - Assume that the θ_i are random variables with pdf $g(\theta_i)$.
  - $f(t_i, \theta_i \mid x_i) = f(t_i \mid x_i, \theta_i) g(\theta_i)$.
  - $f(t_i \mid x_i) = \int f(t_i \mid x_i, \theta_i) d\theta_i$.
  - Estimate $g(\theta_i)$ non-parametrically. Then, using the estimated $g(\theta_i)$, estimate $f(t_i|x_i)$.
  - How about specifying $g(\theta_i)$?

    → Results are too sensitive to specification of $g(\theta_i)$.

    → H-S develop a nonparametric method than can estimate $g(\theta_i)$.

(7) Proportional Hazards

- Cox (Biometrika, 1975).
- A partial solution to the problem of distributional misspecifications.
- $h(t|x_i,\beta) = h_o(t)\lambda_i$,

  where $\lambda_i = \exp(-x_i'\beta)$ and $h_o(t)$ is the common trend (baseline hazard) among i.

- Partial likelihood:
  - Suppose that n individuals (out of N) exit within the follow-up periods.
  - Order them by $0 = t_0 < t_1 < t_2 < ... < t_n$:
    - Don't use censored people.
    - Assume no tie (for convenience only. Can allow ties).
  - $R(t_j)$ = {i| not yet exit just prior ro t_j} where i is a person "at risk" at time t_j

    = {j, j+1, ..., n}.

Short Digression:

- Question:
  - Bob and Steve went fishing.
  - A fish on the pond.
  - Bob uses very good baits, so that the probability (P_B) of Bob's catching the fish (if he fishes alone) = 1.
  - Steve uses bad baits, so that the probability (P_S) of Steve's catching fish (if he fishes alone) = ¼.
  - What is the conditional probability that Steve caught the fish given that one of Bob and Steve caught that fish? [One of Steve and Bob caught a fish. What is the probability that it is Steve?]
- Answer:

$$\frac{P_S}{P_S + P_B} = \frac{1/4}{1/4+1} = \frac{1}{5}.$$

End of Digression

- Probability that "j" exits after time t_j given that one person in $R(t_j)$ fails just after t_j

$$= \frac{h(t_j \mid x_j, \beta)}{\sum_{i \in R(t_j)} h(x_i \mid x_i, \beta)} = \frac{h_o(t_j)\exp(-x_j'\beta)}{\sum_{i \in R(t_j)} h_o(t_j)\exp(-x_i'\beta)} = \frac{\exp(-x_j'\beta)}{\sum_{i \in R(t_j)} \exp(-x_i'\beta)}.$$

- If there are m_j ties at time t_j, let $x_{j,h}$ be the vector of regressors for the h'th person (h = 1, ... , m_j) in the group with t_j. Then use:

$$\prod_{h=1}^{m_j} \frac{\exp(-x_{j,h}'\beta)}{\sum_{i\in R(t_j)} \exp(-x_i'\beta)}.$$

All of the m_j persons should be in $R(t_j)$.

- If an individual's spell is censored between t_j and t_{j+1}, the person will be included in the denominators of $R(t_i)$ for i = 1, ... , j, but not for i = j + 1, ... The person does not influence the numerators.

- Partial ML estimation:
  - $l_N(\beta) = \Sigma_{j=1}^{n} \log\left(\frac{\exp(-x_j'\beta)}{\sum_{i\in R(t_j)} \exp(-x_i'\beta)}\right).$
  - Let m_j = # of exits at time t_j. Then,

$$\hat{h}_o(t_j) = \frac{m_j}{\sum_{i\in R(t_j)} \exp(-x_i'\beta)}.$$

  - $\hat{H}_o(t_j) = \Sigma_{i=1}^{j} \hat{h}_o(t_i)$.
  - $\hat{S}(t_j \mid x_i) = \exp[-\hat{H}_o(t_j)\exp(-x_i'\beta)]$.

- It can be shown (Kiefer, 1988, JEL) that

$$v_i = x_j{}'\beta - \ln[-\ln s(t_j \mid x_j)] \sim \exp(-v_i)\exp\left(-\exp(v_i)\right).$$

[pdf of the extreme-value distribution]

- So, the model specification of the proportional hazard can be tested by investigating the distribution of the $\hat{v}_i$.

(8) Non-Parametric Approach

- Let c_{max} = maximum follow-up periods.
- Divide this interval into k equal subintervals:

  $0 = t_0 < t_1 < ... < t_{j-1} < t_j < t_k = c_{max}$.

- Define:
  - r_j = # of people in (t_{j-1} t_j]

    = # of people "at risk"

    = # of all people - # of people who exited in (0,t_{j-1}]

    \- (# of people censored in (t_{j-1},t_j])/2
  - n_j = # of people who exit in (t_{j-1},t_j].
  - $\hat{h}(t_j) = \frac{n_j}{r_j}$ [Kaplan-Meier estimator].
  - $\hat{H}(t_j) = \Sigma_{a=1}^{j} \hat{h}(t_a)$.
  - $\hat{S}(t_j) = \exp\left(-\hat{H}(t_j)\right)$.

- Comments
  - No regressors.
  - Can determine what h(t) looks like.

(9) Split Population Model

- Schmidt and Witte (1989, JEC)
- So far, we have assumed that S(t|θ) = 1-F(t|θ) → 0 as t →∞. [This means that an ex-prisoner will return to prison surely some time in the future.] But, it might be the case that s(t|θ) = 1 for some individuals. [Some ex-prisoners will never return to prison.]
- Assumptions:
  - $y_i^* = z_i'\gamma + v_i$, where the v_i are N(0,1) (or logistic)

    → $\begin{cases} y_i = 1 \text{ } iff \text{ } y_i^* > 0 \Rightarrow no\ return; \\ y_i = 0 \text{ } iff \text{ } y_i^* < 0 \Rightarrow return\ some\ time \end{cases}$

    → Pr(i never returns) = $\Phi(z_i'\gamma)$;

    Pr(i returns some time) = $1-\Phi(z_i'\gamma)$.

  - For the people with $y_i^* < 0$, t_i^* is the length of actual exit time. [t_i^* is defined only for the people with $y_i^* > 0$.]

    → Let t_i be the observed length of exit time. Then,

    $$t_i = \begin{cases} t_i^* \text{ } if \text{ } t_i^* < c_i; \\ c_i \text{ } if \text{ } t_i^* \geq c_i \end{cases}$$

  - Let $g(t_i^* \mid y_i = 0)$ be the pdf of t_i^* conditional on $y_i = 0$. Assume:

    $g(t_i^* \mid y_i = 0) = f(t_i^* \mid x_i, \theta)$, where f is exponential, Weibull, etc.

- $$\begin{aligned}\Pr(t_i < c_i) &= \Pr(y_i = 0, t_i^* < 0) = \Pr(t_i^* < c_i \mid y_i = 0)\Pr(y_i = 0) \\ &= F(c_i \mid x_i, \theta)[1-\Phi(z_i'\gamma)] = [1 - S(c_i \mid x_i, \theta)][1-\Phi(z_i'\gamma)] \\ &= 1-\Phi(z_i'\gamma) - S(c_i \mid x_i, \theta)[1-\Phi(z_i'\gamma)].\end{aligned}$$

- $$\begin{aligned}\Pr(t_i = c_i) &= 1 - F(c_i \mid x_i, \theta)[1-\Phi(z_i'\gamma)] \\ &= \Phi(z_i'\gamma) + S(c_i \mid x_i, \theta)[1-\Phi(z_i'\gamma)]\end{aligned}.$$

- $$\begin{aligned}g(t_i^* \mid t_i < c_i) &= g(t_i^* \mid y_i = 0, t_i^* < c_i) = f(t_i^* \mid t_i^* < c_i) \\ &= \frac{f(t_i^*)}{\Pr(t_i^* < c_i)} = \frac{f(t_i^* \mid x_i, \theta)}{F(t_i^* \mid x_i, \theta)}\end{aligned}.$$

- Log-likelihood function:

$$\begin{aligned}l_N(\theta,\gamma) &= \sum_{t_i<c_i} \ln\{g(t_i \mid t_i < c_i)\Pr(t_i < c_i)\} + \sum_{t_i=c_i} \ln\{\Pr(t_i = c_i)\} \\ &= \sum_{t_i<c_i} \ln\left\{f(t_i \mid x_i, \theta)[1-\Phi(z_i'\gamma)]\right\} \\ &\quad + \sum_{t_i=c_i} \ln\left\{\Phi(z_i'\gamma) + S(c_i \mid x_i, \theta)[1-\Phi(z_i'\gamma)]\right\}\end{aligned}$$

- Comment:
  - If the Kaplan-Meier estimator shows positive duration dependence, it is very hard to get split-population MLE results.
  - If it shows rapid negative duration dependence, it is worth trying split MLE.

- Testing split-population model:
  - MLE without split,

    $$l_N^R(\theta,\gamma) = \sum_{t_i<c_i} \ln\{ f(t_i \mid x_i,\theta)\} + \sum_{t_i=c_i} \ln\{S(c_i \mid x_i,\theta)]\},$$

    which is the split log-likelihood function with the restriction

    $\Phi(z_i'\gamma) = 0.$
  - $LR_N = 2[l_N(\hat{\theta},\hat{\gamma}) - l_N^R(\tilde{\theta})] \to \chi^2(df = \dim(\gamma))$ under H$_o$: no split.

    → Really?

(10) Autoregressive Conditional Duration (ACD) Model

- Trading times: $t_0 < t_1 < t_2 < .... < t_N$.

- Duration of an interval (time between two trades): $x_i = t_i - t_{i-1}$.

- EACD(2,2) [Exponential ACD of orders 1 and 1]:
  - Assume x_i ~ exponential(λ_i), where $\psi_i = 1/\lambda_i$.
  - $E(x_i | \Omega_{t_{i-1}}) = \psi_i$.
  - Assume $\psi_i = \omega + \alpha_1 x_{i-1} + \alpha_2 x_{i-2} + \beta_1 \psi_{i-1} + \beta_2 \psi_{i-2}$.

- WACD(2,2) [Weibull ACD of orders 1 and 1]:
  - Assume $(x_i)^p$ ~ exponential(λ_i), where $\psi_i = 1/\lambda_i$.
  - Assume $\psi_i = \omega + \alpha_1 x_{i-1} + \alpha_2 x_{i-2} + \beta_1 \psi_{i-1} + \beta_2 \psi_{i-2}$.