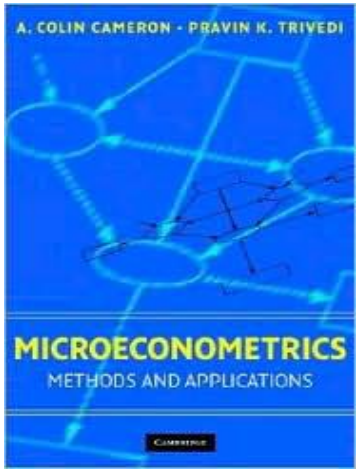
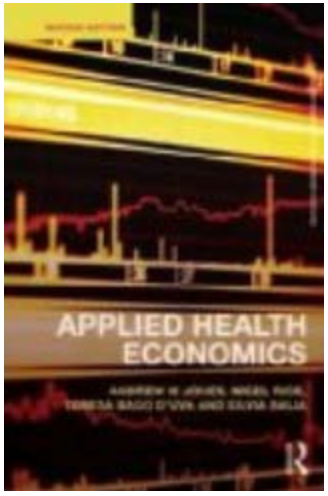
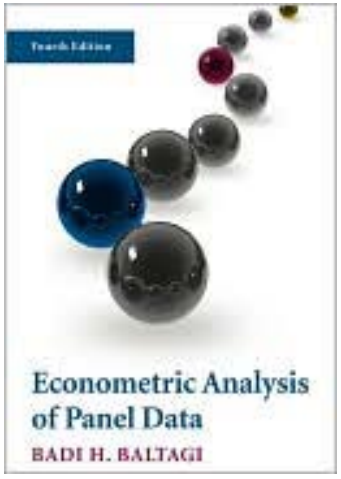
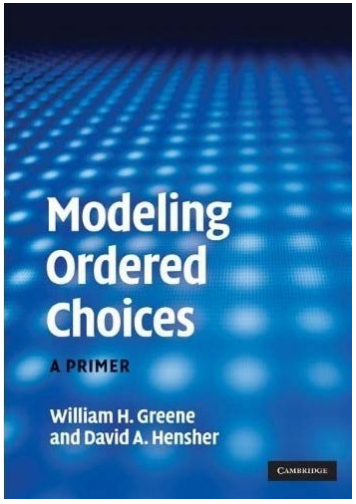
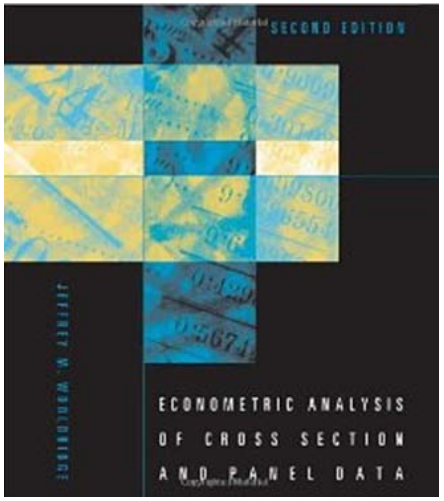


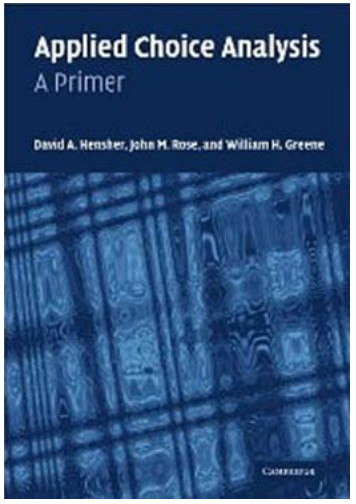




11. Duration Modeling







Modeling Duration

- Time until retirement
- Time until business failure
- Time until exercise of a warranty
- Length of an unemployment spell
- Length of time between children
- Time between business cycles
- Time between wars or civil insurrections
- Time between policy changes
- Etc.

The Hazard Function

For the random variable t = time until an event occurs, $t \geq 0$.

f(t) = density; F(t) = cdf = Prob[time $\leq$ t]= S(t) = 1-F(t) = survival

Probability of an event occurring at or before time t is F(t)

A conditional probability: for small $\Delta > 0$,

h(t)= Prob(event occurs in time t to t+Δ | has not already occurred)

h(t)= Prob(event occurs in time t to t+Δ | occurs after time t)

$$= \frac{F(t+\Delta)-F(t)}{1-F(t)}$$

Consider as $\Delta \rightarrow 0$, the function

$$\lambda(t) = \frac{F(t+\Delta)-F(t)}{\Delta \quad (1-F(t))} \rightarrow \frac{f(t)}{S(t)}$$

$$\lambda(t) \rightarrow \frac{f(t)}{S(t)} = \text{the "hazard function" and } \Delta\lambda(t) \approx \text{Prob[time} \leq t \leq \text{time}+\Delta \mid \text{time} \geq t]$$

$\lambda(t)$ is a characteristic of the distribution

Hazard Function

Since $\lambda(t) = f(t)/S(t) = -d\log S(t)/dt$,

$$F(t) = 1 - \exp\left[-\int_0^t \lambda(s)ds\right], t \geq 0.$$

$$dF(t) / dt = -\exp\left[-\int_0^t \lambda(s)ds\right](-1)\lambda(t)$$

$$= \lambda(t)\exp\left[-\int_0^t \lambda(s)ds\right] \text{(Leibnitz's Theorem)}$$

Thus, $F(t)$ is a function of the hazard;

$S(t) = 1 - F(t)$ is also,

and $f(t) = S(t)\lambda(t)$

A Simple Hazard Function

The Hazard function

Since f(t) = dF(t)/dt and S(t) = 1-F(t),

$$h(t)=\frac{f(t)}{S(t)}=-d\log S(t)/dt$$

Simplest Hazard Model - a function with no "memory"

$\lambda(t)$ = a constant, λ

$$\frac{f(t)}{S(t)}=\lambda=-d\log S(t)/dt.$$

The second simplest differential equation;

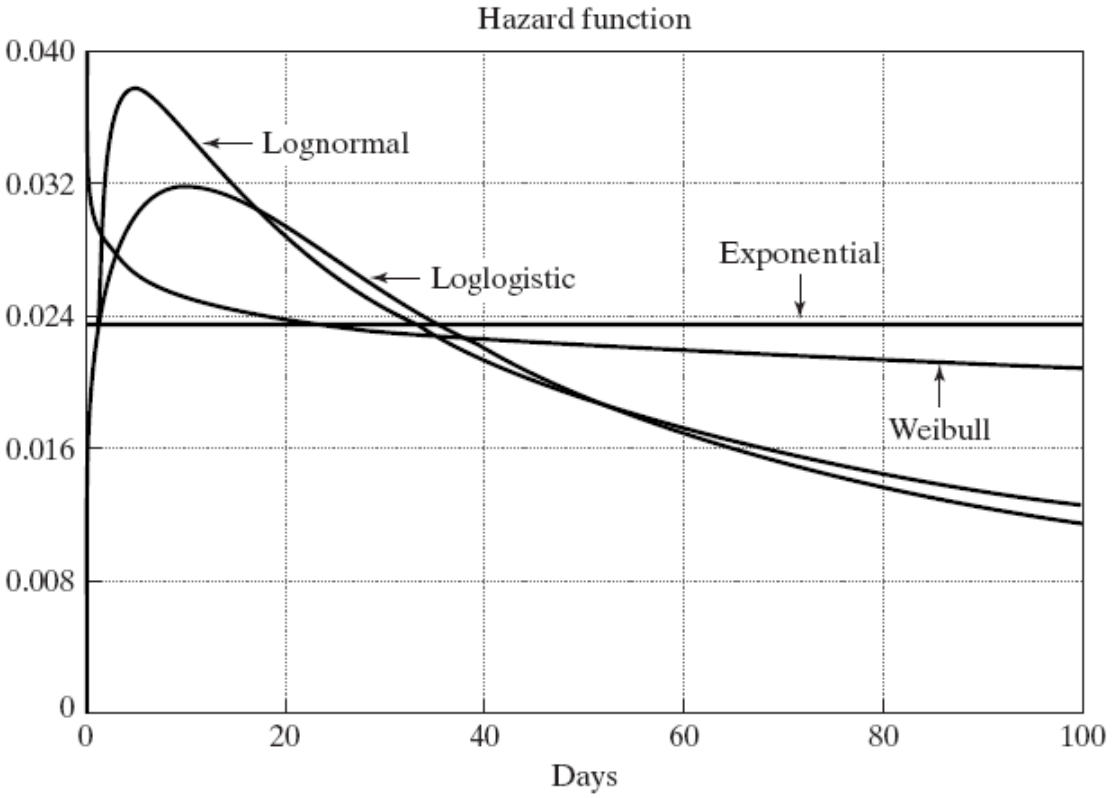
$d\log S(t)/dt=-\lambda \Rightarrow S(t)=K\exp(-\lambda t)$, K = constant of integration

Particular solution requires S(0)=1, so K=1 and $S(t)=\exp(-\lambda t)$

F(t) = 1-exp(-λt) or $f(t)=\lambda\exp(-\lambda t), t\geq 0$. Exponential model.

Duration Dependence

When $d\lambda(t)/dt \neq 0$, there is 'duration dependence'



Parametric Models of Duration

There is a large menu of parametric models for survival analysis:

Exponential: $\lambda(t)=\lambda$,

Weibull: $\lambda(t)=\lambda p(\lambda t)^{p-1}$; p=1 implies exponential,

Loglogistic: $\lambda(t)=\lambda p(\lambda t)^{p-1} / [1+(\lambda t)^{p}]$,

Lognormal: $\lambda(t)=\phi[-\text{plog}(\lambda t)]/\Phi[-\text{plog}(\lambda t)]$,

Gompertz: $\lambda(t)=p \exp(\lambda t)$,

Gamma: Hazard has no closed form and must be numerically integrated,

and so on.

Censoring

Most data sets have incomplete observations.

Observation is not t, but t* < t. I.e., it is known (expected) that failure takes place after t.

How to build censoring into a survival model?

Accelerated Failure Time Models

$\lambda(.)$ becomes a function of covariates.

$\mathbf{x}$=a set of covariates (characteristics) observed at baseline

Typically,

$\lambda(t|\mathbf{x})=h[\exp(\mathbf{x}'\beta),t]$

E.g., Weibull: $\lambda(t|\mathbf{x})=\exp(\mathbf{x}'\beta)p[\exp(\mathbf{x}'\beta)t]^{p-1}$

E.g., Exponential: $\lambda(t|\mathbf{x})=\exp(\mathbf{x}'\beta)[\exp(\mathbf{x}'\beta)t]$;

$f(t|\mathbf{x})=\exp(\mathbf{x}'\beta)\exp[-\exp(\mathbf{x}'\beta)t]$

Proportional Hazards Models

$\lambda(t \mid \mathbf{x}) = g(\mathbf{x})\lambda(t)$

$\lambda(t)$ = the 'baseline hazard function'

Weibull: $\lambda(t|\mathbf{x}) = p\exp(\mathbf{x}'\beta)^{p}(t)^{p-1}$

None of Loglogistic, F, gamma, lognormal, Gompertz, are proportional hazard models.

ML Estimation of Parametric Models

Maximum likelihood is essentially the same as for the tobit model

$f(t|\mathbf{x})$ = density

$S(t|\mathbf{x})$ = survival

For observed t, combined density is

$$g(t|\mathbf{x}) = [f(t|\mathbf{x})]^{d}[S(t|\mathbf{x})]^{(1-d)}$$

d = 1 if not censored, 0 if censored. Rearrange

$$g(t|\mathbf{x}) = \left[\frac{f(t|\mathbf{x})}{S(t|\mathbf{x})}\right]^{d}[S(t|\mathbf{x})] = [\lambda(t|\mathbf{x})]^{d}S(t|\mathbf{x})$$

$$\log L = \sum_{i=1}^{n} d_i \log\lambda(t_i|\mathbf{x}_i) + \log S(t_i|\mathbf{x}_i)$$

Time Varying Covariates

Hazard function must be defined as a function of the covariate path up to time t;

$\lambda(t,X(t)) = \ldots$

Not feasible to model a continuous path of the individual covariates. Data may be observed at specific intervals, $[0,t_1 \mid x(0)), [t_1, t_2 \mid x(1)), \ldots$

Treat observations as a sequence of observations.

Build up hazard path piecewise, with time invariant covariates in each segment. Treat each interval save for the last as a censored (at both ends) observation.

Last observation (interval) might be censored, or not.

Unobserved Heterogeneity

Typically multiplicative - **variable with mean 1.**

$\lambda(t|\mathbf{x},u)=u\lambda(t|\mathbf{x})$

Also typical:

$\lambda(t|\mathbf{x},u)=u\lambda[\exp(\mathbf{x}'\beta), t]$

In proportional hazards models like Weibull,

$\lambda(t|\mathbf{x},u)=u\exp(\mathbf{x}'\beta)\lambda[t] = \exp(\mathbf{x}'\beta+\varepsilon)\lambda[t]$

Approaches: Assume f(u), then integrate u out of f(t|**x**,u).

(1) (log)Normally distributed ε (u), amenable to quadrature (Butler/Moffitt) or simulation based estimation

(2) (very typical). Log-gamma u has $f(u)=\dfrac{\theta^{\theta}\exp(-\theta u)u^{\theta-1}}{\Gamma(\theta)}$

Produces $f(t|\mathbf{x})=\dfrac{[A(t)]^{\theta+1}\lambda p(\lambda t)^{P-1}}{[1+\theta(\lambda t)^{P}]^{1/\theta}}$,

A(t) = survival function without heterogeneity, for exponential or Weibull.

Interpretation

- What are the coefficients?
- Are there ‘marginal effects?’
- What quantities are of interest in the study?

Cox's Semiparametric Model

Cox Proportional Hazard Model

$$\lambda(t_i \mid \mathbf{x}_i) = \exp(-\mathbf{x}_i'\beta)\lambda_0(t_i)$$

Conditional probability of exit - with K distinct exit times in the sample:

$$\text{Prob}[t_i = T_k \mid \mathbf{X}_k] = \frac{\exp(\mathbf{x}_i'\beta)}{\sum_{\text{All individuals with } t_s \geq T_k} \exp(\mathbf{x}_s'\beta)}$$

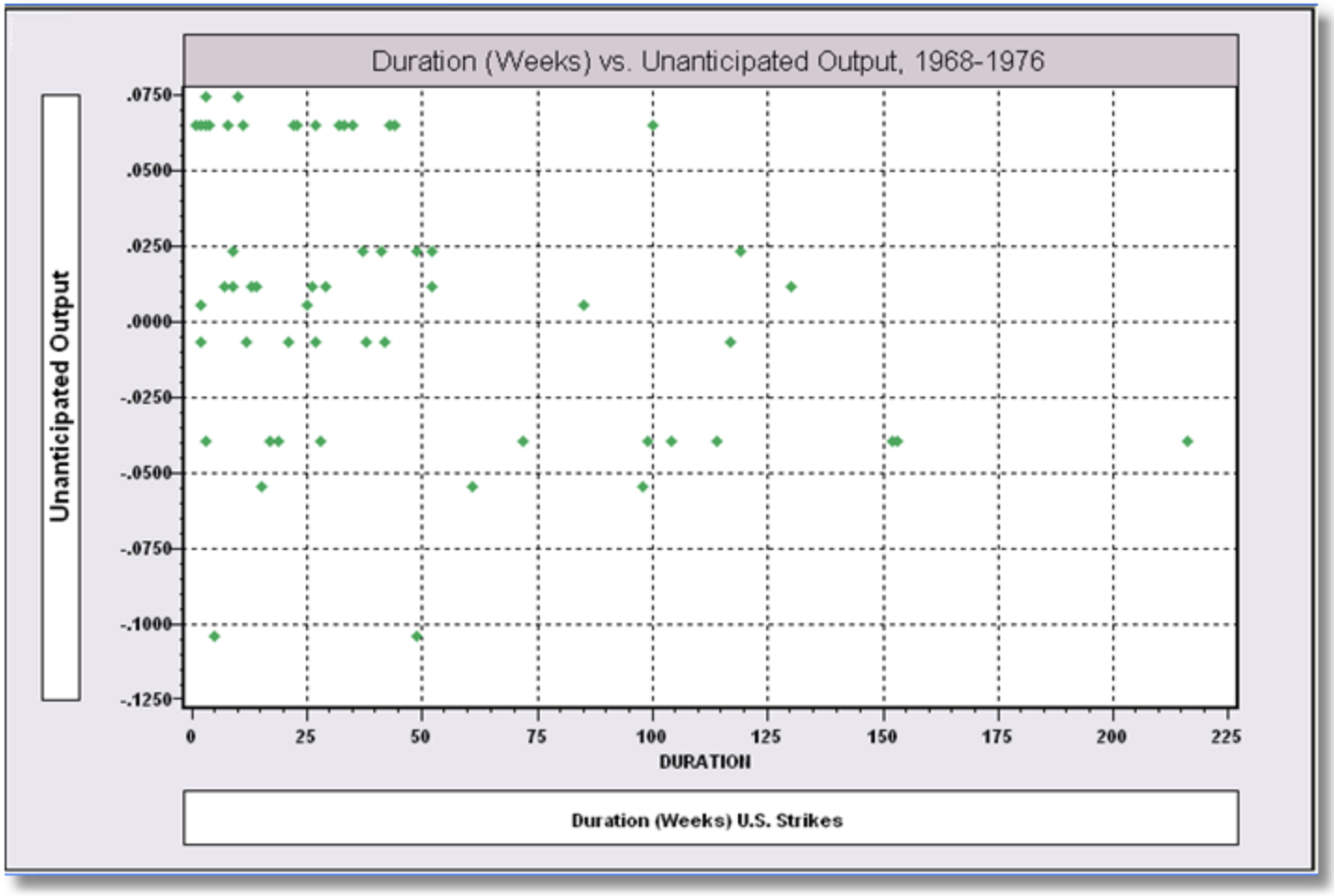
(The set of individuals with $t_s \geq T_k$ is the risk set.

Partial likelihood - simple to maximize.

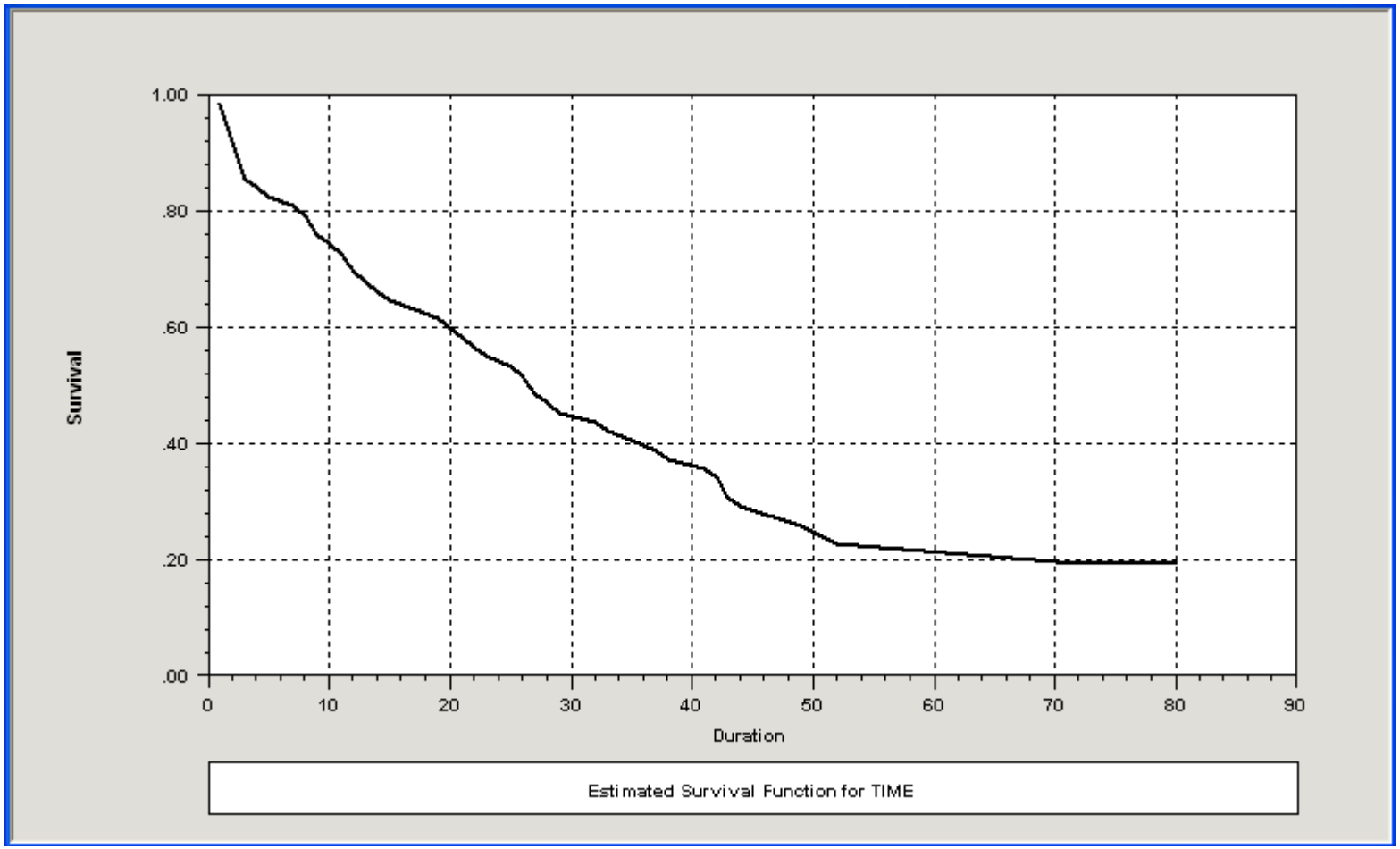
Nonparametric Approach

- Based simply on counting observations
  - K spells = ending times 1,...,K
  - d_j = # spells ending at time t_j
  - m_j = # spells censored in interval $[t_j , t_{j+1})$
  - r_j = # spells in the risk set at time $t_j = \Sigma (d_j+m_j)$
- Estimated hazard, $h(t_j) = d_j/r_j$
- Estimated survival = $\Pi_j [1 – h(t_j)]$
  (Kaplan-Meier “product limit” estimator)

Kennan's Strike Duration Data



Kaplan Meier Survival Function

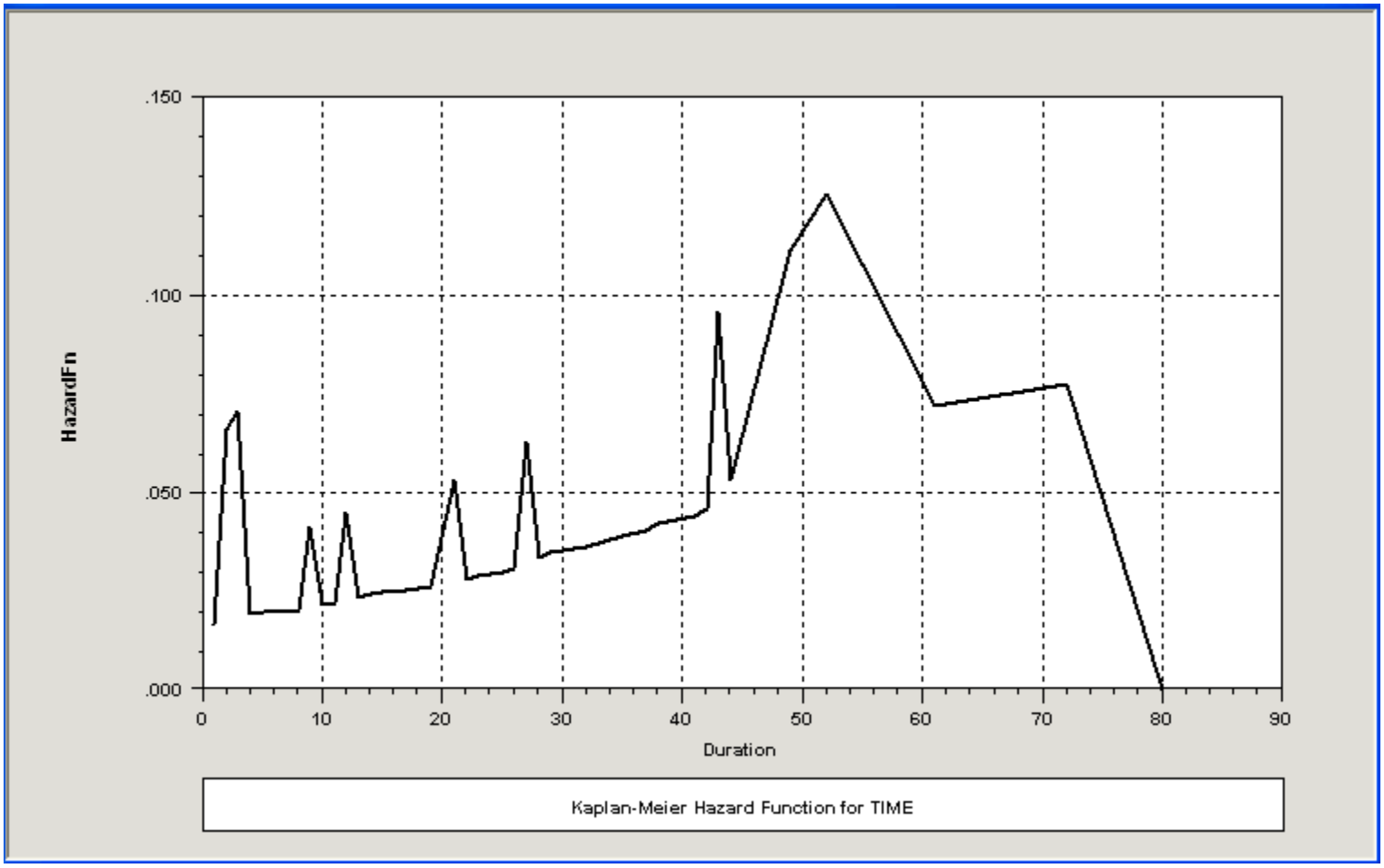


Hazard Rates

Number of observations in stratum = 62
Number of observations exiting = 50
Number of observations censored = 12

Survival		Enter	Cnsrd	At Risk	Exited	Survival Rate	Hazard Rate
.0-	4.0	62	0	62	9	1.0000 (.000)	.0391 (.013)
4.0-	8.0	53	0	53	3	.8548 (.045)	.0146 (.008)
8.0-	12.0	50	0	50	5	.8065 (.050)	.0263 (.012)
12.0-	16.0	45	0	45	5	.7258 (.057)	.0294 (.013)
16.0-	20.0	40	0	40	2	.6452 (.061)	.0128 (.009)
20.0-	24.0	38	0	38	4	.6129 (.062)	.0278 (.014)
24.0-	28.0	34	0	34	4	.5484 (.063)	.0313 (.016)
28.0-	32.0	30	0	30	2	.4839 (.063)	.0172 (.012)
32.0-	36.0	28	0	28	3	.4516 (.063)	.0283 (.016)
36.0-	40.0	25	0	25	2	.4032 (.062)	.0208 (.015)
40.0-	44.0	23	0	23	4	.3710 (.061)	.0476 (.024)
44.0-	48.0	19	0	19	1	.3065 (.059)	.0135 (.014)
48.0-	52.0	18	0	18	2	.2903 (.058)	.0294 (.021)
52.0-	56.0	16	0	16	2	.2581 (.056)	.0333 (.024)
56.0-	60.0	14	0	14	0	.2258 (.053)	.0000 (.000)
60.0-	64.0	14	0	14	1	.2258 (.053)	.0185 (.019)
64.0-	68.0	13	0	13	0	.2097 (.052)	.0000 (.000)
68.0-	72.0	13	0	13	0	.2097 (.052)	.0000 (.000)
72.0-	76.0	13	0	13	1	.2097 (.052)	.0200 (.020)
76.0-	80.0	12	12	6	0	.1935 (.050)	.0000 (.000)

Kaplan Meier Hazard Function



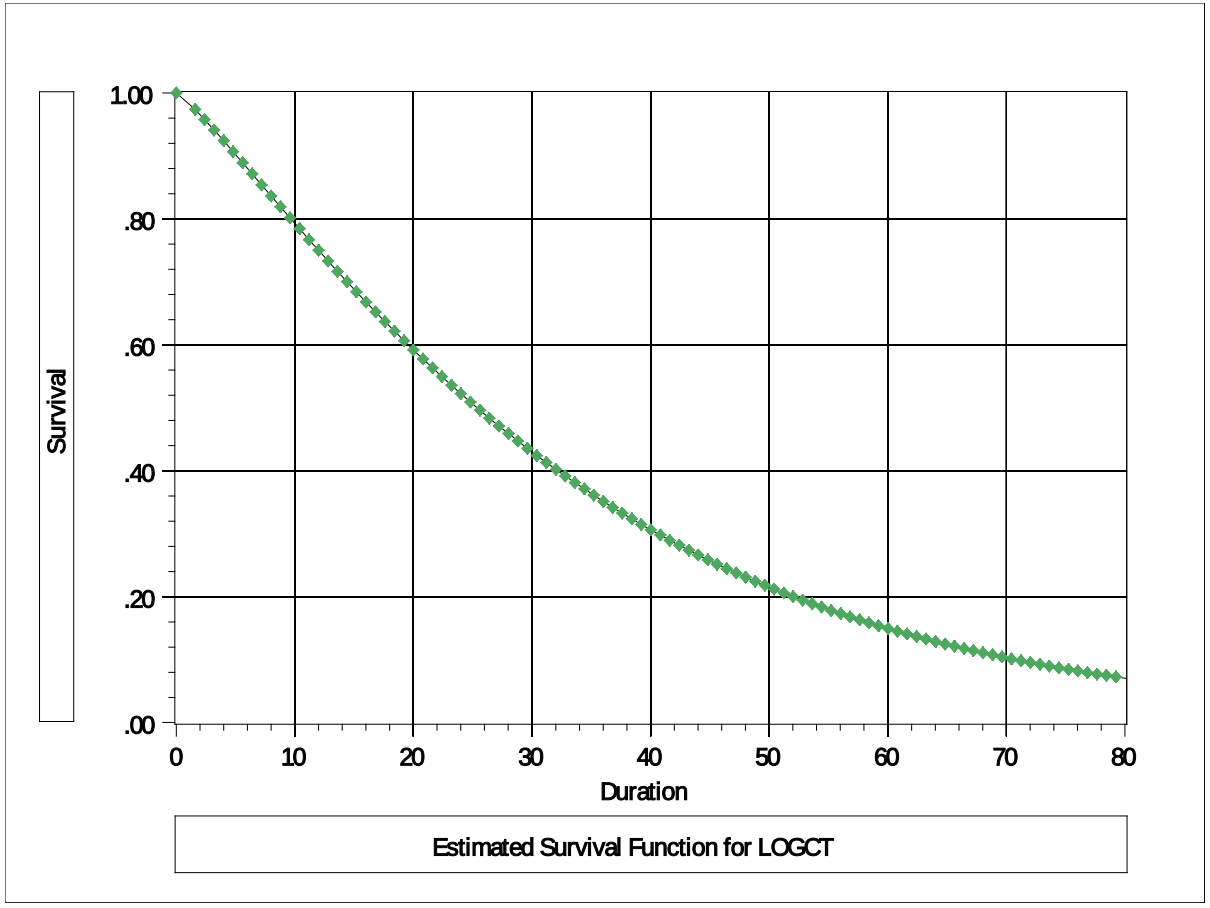
Weibull Accelerated Proportional Hazard Model

```
+--------------------------------------------+
| Loglinear survival model: WEIBULL          |
| Log likelihood function       -97.39018    |
| Number of parameters                3      |
| Akaike IC=  200.780  Bayes IC=  207.162    |
+--------------------------------------------+
+---------+-------------+----------------+--------+---------+----------+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+-------------+----------------+--------+---------+----------+
          RHS of hazard model
 Constant      3.82757279      .15286595    25.039   .0000
 PROD        -10.4301961     3.26398911    -3.196   .0014     .01102306
          Ancillary parameters for survival
 Sigma         1.05191710      .14062354     7.480   .0000
```

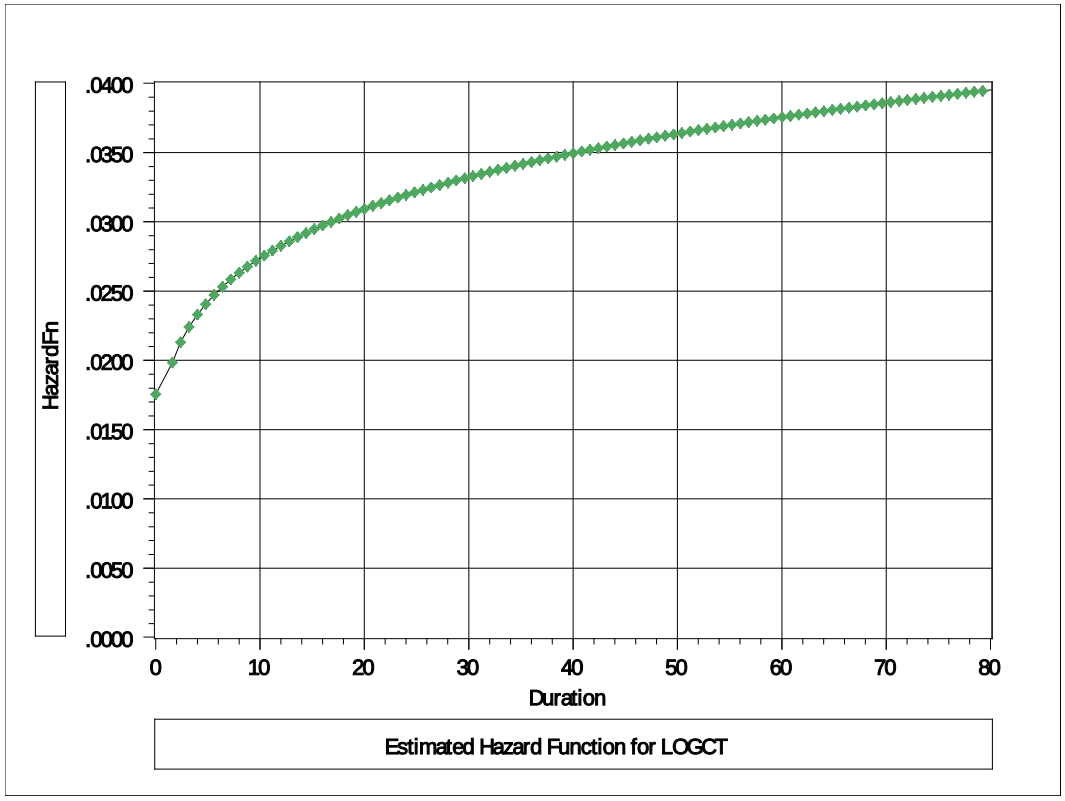
Weibull Model

Parameters of underlying density at data means:				
Parameter	Estimate	Std. Error	Confidence Interval	
Lambda	.02441	.00358	.0174 to	.0314
P	.95065	.12709	.7016 to	1.1997
Median	27.85629	4.09007	19.8398 to	35.8728
Percentiles of survival distribution:				
Survival	.25	.50	.75	.95
Time	57.75	27.86	11.05	1.80

Survival Function



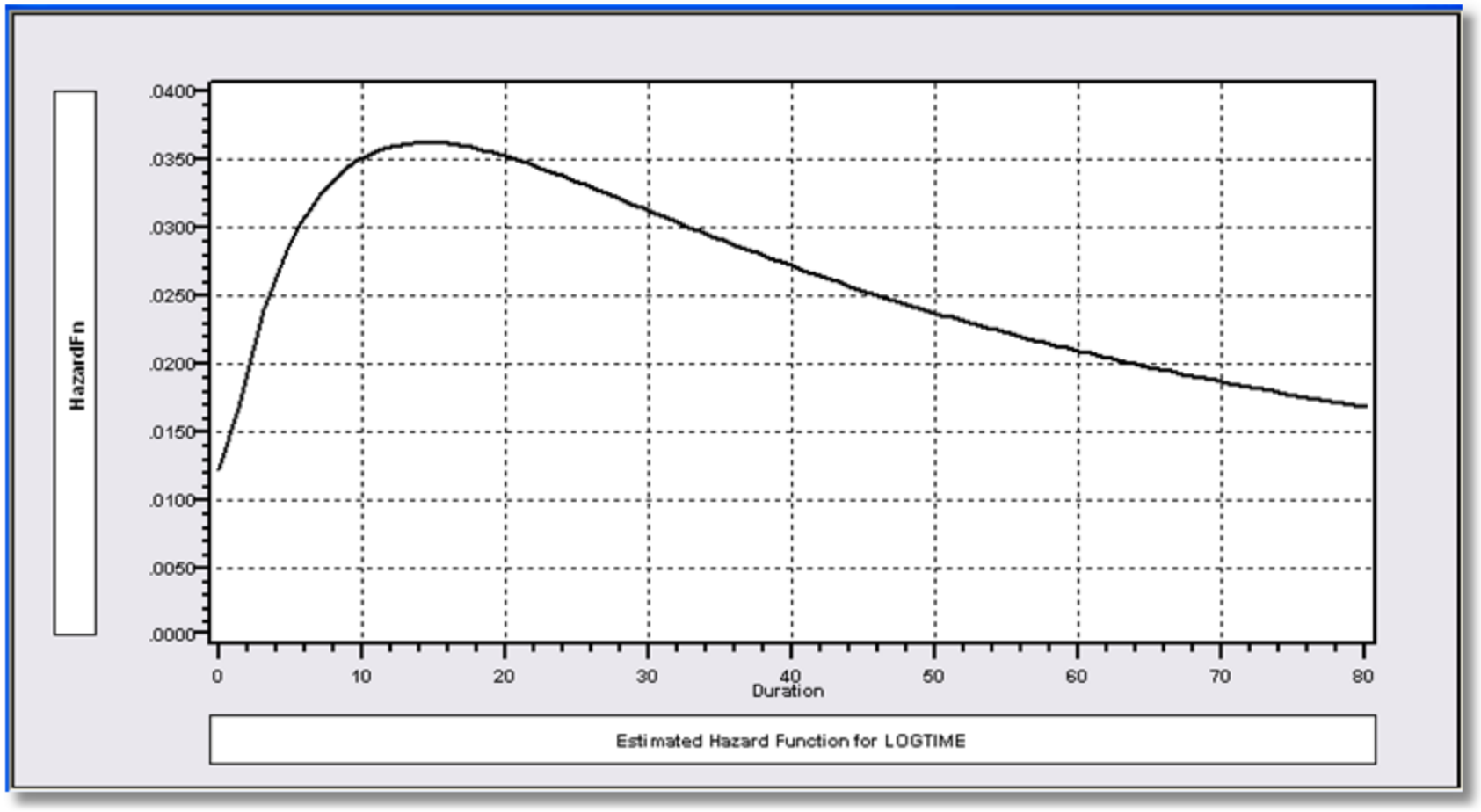
Hazard Function with Positive Duration Dependence for All t



Loglogistic Model

```
+--------------------------------------------+
| Loglinear survival model: LOGISTIC         |
| Dependent variable               LOGCT     |
| Log likelihood function       -97.53461    |
+--------------------------------------------+
+---------+-------------+----------------+--------+---------+----------+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+-------------+----------------+--------+---------+----------+
          RHS of hazard model
 Constant      3.33044203      .17629909    18.891   .0000
 PROD        -10.2462322     3.46610670    -2.956   .0031     .01102306
          Ancillary parameters for survival
 Sigma          .78385188      .10475829     7.482   .0000
+--------------------------------------------+
| Loglinear survival model: WEIBULL          |
| Log likelihood function       -97.39018    |
+--------------------------------------------+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+---------+-------------+----------------+--------+---------+----------+
          RHS of hazard model
 Constant      3.82757279      .15286595    25.039   .0000
 PROD        -10.4301961     3.26398911    -3.196   .0014     .01102306
          Ancillary parameters for survival
 Sigma         1.05191710      .14062354     7.480   .0000
```

Loglogistic Hazard Model



HEALTH ECONOMICS
Health Econ. **21**(Suppl. 1): 56–100 (2012)
Published online in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/hec.2811

THE IMPACT OF HEALTH CHANGES ON LABOR SUPPLY: EVIDENCE FROM MERGED DATA ON INDIVIDUAL OBJECTIVE MEDICAL DIAGNOSIS CODES AND EARLY RETIREMENT BEHAVIOR

BENT JESPER CHRISTENSEN* and MALENE KALLESTRUP-LAMB
Department of Economics and Business, Aarhus University, Aarhus, Denmark

2. DATA

The full database contains annual observations on all individuals in Denmark above 18 years of age for the period 1985–2001, with measurement in November each year. The data are based on administrative registers and contain no survey element. We have information on various individual demographic, financial, and socio-economic characteristics, health, and labor market status. This enables us to identify individual transitions between different labor market states and health events on an annual basis.

$$Duration = Retirement\ age - 50. \tag{1}$$

Right censoring occurs at age 67 or in the event of death. By definition, spells start at age 50, so there is no left censoring.

Table I. Number of spells by duration

Duration	Observed spells	Exits
1—age 51	164	123
2—age 52	206	162
3—age 53	227	184
4—age 54	242	198
5—age 55	295	249
6—age 56	326	287
7—age 57	492	455
8—age 58	603	556
9—age 59	695	637
10—age 60	1606	1548
11—age 61	968	925
12—age 62	578	540
13—age 63	675	628
14—age 64	514	479
15—age 65	332	314
16—age 66	211	187
17—age 67	1195	6
Total	9329	7478

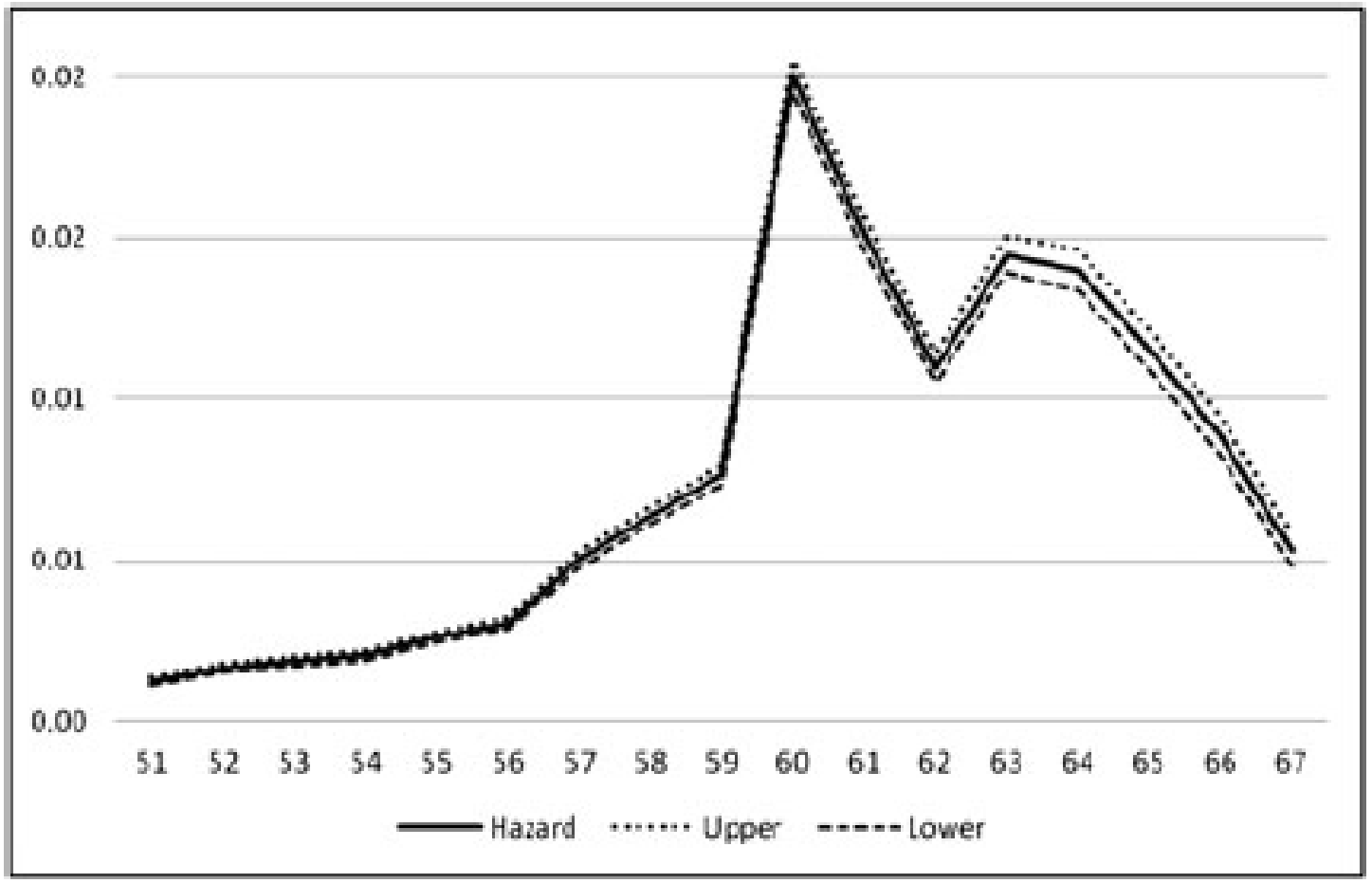


Figure 1. Kaplan–Meier hazard estimate and confidence band

where the t_k are known constants.[4] Any duration falling into the last interval, $[t_K, \infty)$, is censored at t_K. Given explanatory variables at level x_k over the course of the kth interval, the conditional probability that duration T is greater than t_k given that it is greater than t_{k-1} is

$$P(T > t_k | T > t_{k-1}, x_k) = \exp[-\int_{t_{k-1}}^{t_k} h(t|x_k)dt] = \alpha_k(x_k, \theta), \tag{8}$$

thus defining the probability $\alpha_k = \alpha_k(x_k, \theta)$ for given unknown parameter vector θ to be estimated.[5] Given T is not in one of the $k-1$ first intervals, the probability that it is in the kth interval is $1-\alpha_k$, and with probability α_k, it is in a later interval, so α_k and $1-\alpha_k$ are the discrete time conditional survivor and hazard, respectively. The individual contribution to the likelihood function for an individual with duration in the kth interval and observed regressors $x_j, j = 1, \ldots, k$, is

$$\begin{aligned} L(\theta, k, x) &= \left[(1 - \alpha_k(x_k, \theta))\prod_{j=1}^{k-1}\alpha_j(x_j, \theta)\right]^d \left[\prod_{j=1}^{k}\alpha_j(x_j, \theta)\right]^{1-d} \\ &= (1 - \alpha_k(x_k, \theta))^d \alpha_k(x_k, \theta)^{1-d}\prod_{l=1}^{k-1}\alpha_l(x_l, \theta), \end{aligned} \tag{9}$$

where $d = 1$ if the duration is uncensored and zero otherwise.

The log likelihood function for a sample of n individuals, with the ith retiring in the k_ith interval, for observed regressors $x_{i,j}$ is therefore

$$l(\theta) = \sum_{i=1}^{n}\left\{\begin{array}{c} d_i\left[\ln(1 - \alpha_{k_i}(x_{i,k_i}, \theta)) + \sum_{j=1}^{k_i-1}\ln(\alpha_j(x_{i,j}, \theta))\right] \\ +(1 - d_i)\left[\sum_{j=1}^{k_i}\ln(\alpha_j(x_{i,j}, \theta))\right] \end{array}\right\}. \tag{10}$$

$$h(t|x_t) = \lambda(t)\exp(x_t'\beta), \tag{11}$$

ensures the interpretation

$$\beta = \frac{\partial h(t|x_t)/\partial x_t}{h(t|x_t),} \tag{12}$$

that is, when shifting a given covariate, the corresponding coefficient in β gives the resulting proportional change in the hazard. We still allow the covariates x_t to vary between intervals k, so the hazard may vary over time because of changes in both factors in (11). In this case,

$$\alpha_k(x_{i,k},\theta) = \exp\left[-\int_{t_{k-1}}^{t_k} h(t|x_{it})dt\right] = \exp\left[-\int_{t_{k-1}}^{t_k} \lambda(t)\exp\left(x_{i,k}'\beta\right)dt\right] = \exp\left[-\exp\left(x_{i,k}'\beta\right)\cdot\Lambda_k\right], \tag{13}$$

where $\Lambda_k = \int_{t_{k-1}}^{t_k} \lambda(t)dt$ is the increment to the integrated baseline hazard over the k'th interval. This is inserted

3.2. Unobserved heterogeneity

The models, so far, only allow for observed heterogeneity, controlled for through the regressors x_t. We now consider, in addition, unobserved heterogeneity entering via an individual-specific term v, so that the conditional hazard is of the form $h(t|x_t, v)$ (see, e.g., Hougaard (1984)). Maintaining proportionality as in (11), this yields the mixed proportional hazard specification

$$h(t|x_t, v) = v \cdot \lambda(t) \exp(x_t' \beta). \tag{18}$$

Conditionally on the heterogeneity term, discrete time survivor (13) becomes

$$\alpha_k(x_{i,k}, v_i, \theta) = \exp\left[-v_i \cdot \exp(x_{i,k}' \beta) \cdot \Lambda_k\right]. \tag{19}$$

The likelihood function is formed by integrating out the unobserved v_i by using a parametric or nonparametric distributional assumption. Writing $G(v|x, \theta)$ for the conditional heterogeneity distribution, the general form of the individual contribution to likelihood is obtained by substituting (19) for α_k in (9) and integrating,

$$\begin{aligned} L(\theta, k, x) = \int & \left(1 - \exp\left[-v \cdot \exp(x_k' \beta) \cdot \Lambda_k\right]\right)^d \exp\left[-v \cdot \exp(x_k' \beta) \cdot \Lambda_k\right]^{1-d} \\ & \cdot \prod_{l=1}^{k-1} \exp\left[-v \cdot \exp(x_l' \beta) \cdot \Lambda_l\right] G(dv|x, \theta), \end{aligned} \tag{20}$$

Log Baseline Hazards

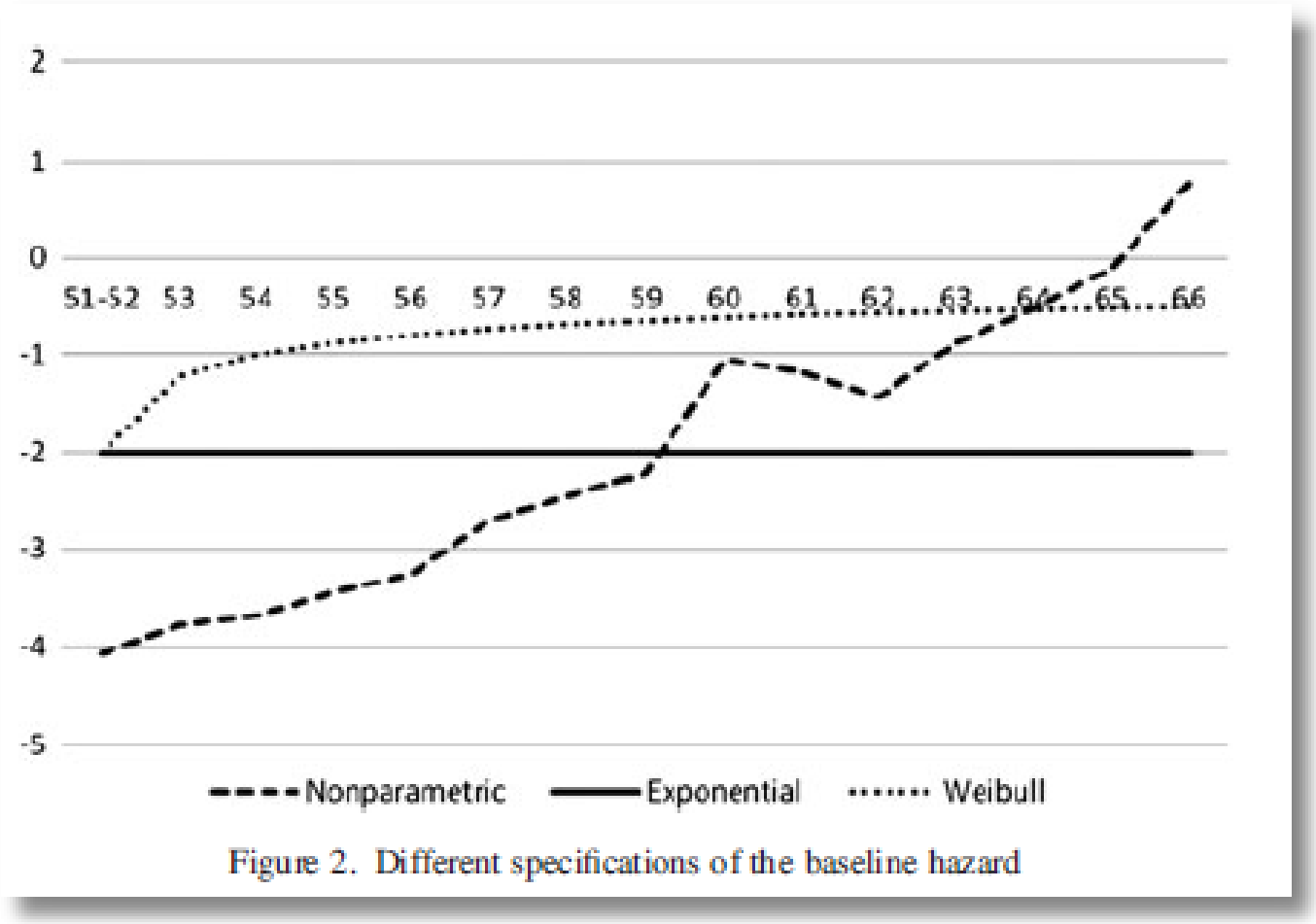


Figure 2. Different specifications of the baseline hazard

Log Baseline Hazards - Heterogeneity

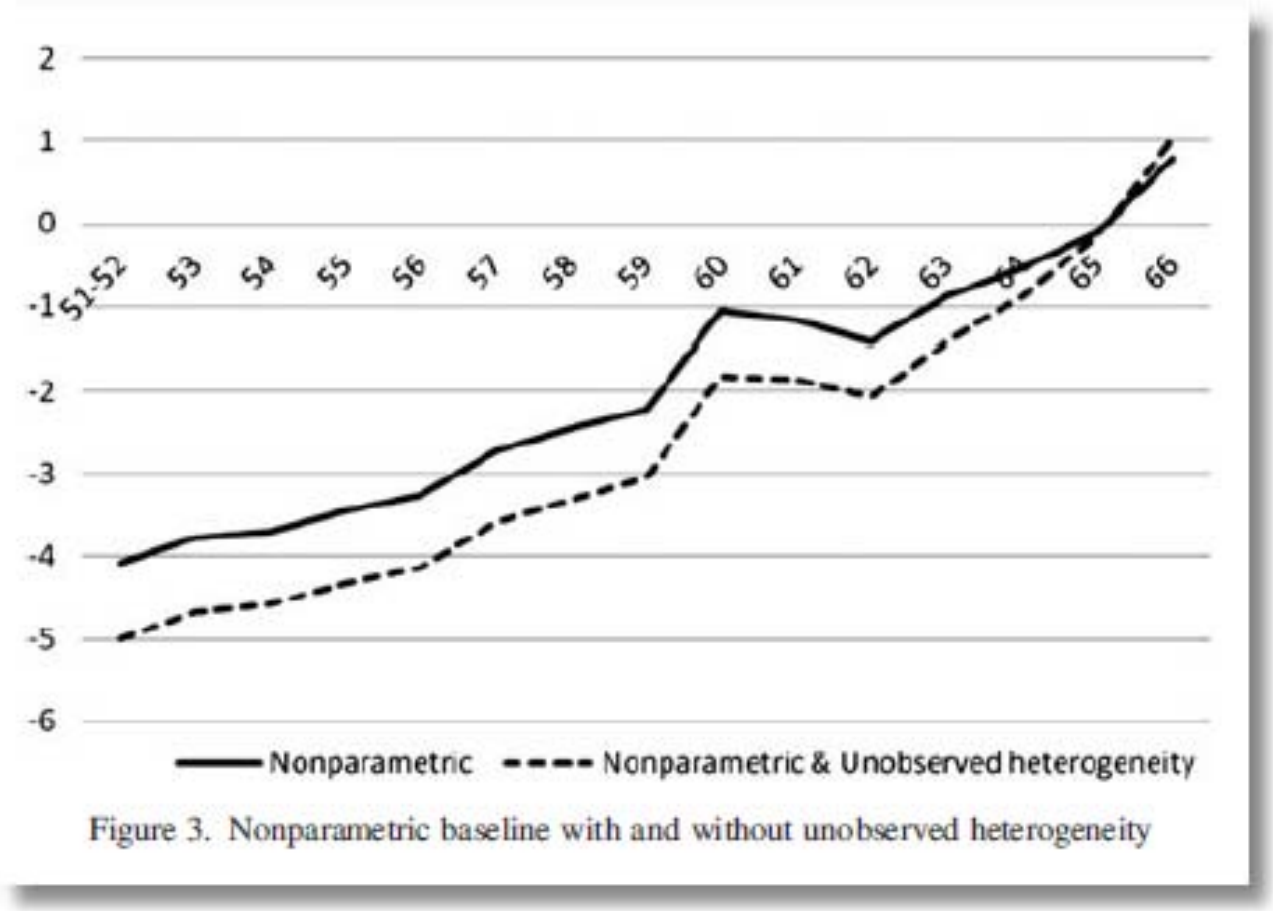


Figure 3. Nonparametric baseline with and without unobserved heterogeneity