[8] Davidoff, T., Brown, J.R. & Diamond, P.A. (2005). Annuities and individual welfare, *American Economic Review* **95**, 1573–1590.
[9] Brown, J.R. (2003). Redistribution and insurance: mandatory annuitization with mortality heterogeneity, *Journal of Risk and Insurance* **70**, 17–41.
[10] Financial Services Authority (2006). *2006 Survey of the Persistency of Life and Pension Policies*, http://www.fsa.gov.uk/pubs/other/Persistency_2006.pdf.
[11] Financial Services Authority (2000). *Persisting: why consumers stop paying into policies*, Consumer Research 6.
[12] Hendel, I. & Lizzeri, A. (2003). The role of commitment in dynamic contracts: evidence from life insurance, *Quarterly Journal of Economics* **118**, 299–327.
[13] Blake, D., Cairns, A.J. & Dowd, K. (2007). *Facing Up to the Uncertainty of Life: the Longevity of Fan Charts*, Pensions Institute, Discussion Paper No 0703.
[14] Finkelstein, A. & Poterba, J. (2004). Adverse selection in insurance markets: policy holder evidence from the UK annuity market, *Journal of Political Economy* **112**, 193–208.
[15] Brown, J.R. & Goolsbee, A. (2002). Does the Internet make markets more competitive? Evidence from the life insurance industry, *Journal of Political Economy* **110**, 481–507.

REBECCA DRIVER

Lifetime Models and Risk Assessment

Researchers in different fields have investigated risk assessment using methods for lifetime data. Lifetime data have been referred to as *survival time data* in biomedical research, as *reliability data* in engineering applications, and as *time-to-event data* in social sciences. In general, lifetime data can be defined as the times to the occurrence of a certain event. Statistical methods for analyzing lifetime data include the definition of survival time, censored data, prediction of probability of survival, prediction of instantaneous hazards, evaluation of cure rates, etc. We review different types of lifetime data, methods for estimating hazard and survival functions, and methods for risk assessment. Detailed information can be found in the following books: Kalbfleisch and Prentice [1], Nelson [2], Cox and Oakes [3], Fleming and Harrington [4], Andersen *et al.* [5], Marubini and Valsecchi [6], Klein and Moeschberger [7], Lawless [8], Lee and Wang [9], and others. Hougaard [10] gives good descriptions of analytical methods for multivariate survival data. Bayesian survival analysis can be found in Ibrahim *et al.* [11]. A Bayesian approach for reliability models and risk assessment is given by Singpurwalla [12]. Lifetime models, based on stochastic processes hitting a boundary, have been investigated by Whitmore [13, 14], Lu [15], Whitmore *et al.* [16], Onar and Padgett [17], Aalen and Gjessing [18], Padgett and Tomlinson [19, 20], and others. A review of threshold regression for risk assessment, based on the concept of a first hitting time (FHT), can be found in Lee and Whitmore [21].

Important Parametric Distributions for Lifetime Models

Several parametric distributions have been used in analyzing lifetime data, including the exponential, Weibull, inverse Gaussian, gamma and generalized gamma, lognormal, log-logistic, and various accelerated failure-time distributions.

Censored and Truncated Data

In practical situations, it often happens that we cannot observe all the subjects in the study for a complete lifetime data. For example, patients may be either lost to follow-up or still alive at the end of the study. These kinds of data are called *censored data*. There are several types of censoring mechanisms.

1. **Type I censoring**
Each individual has a fixed potential censoring time $C_i > 0$ such that lifetime T_i is observed if $T_i \leq C_i$; otherwise, we only know that $T_i > C_i$.

2. **Type II censoring**
In a random sample of size n, we only observe the r smallest lifetimes $t_{(1)} \leq t_{(2)} \cdots \leq t_{(r)}$, where $r < n$. Both type I and type II censored data, as defined above, are also referred to as *being right censored*.

3. **Independent random censoring**
It is common that the censoring mechanism is random and independent of the failure process. In this situation, the lifetimes T_i and censoring times C_i

are independent, for $i = 1, \ldots, n$. Letting I denote an indicator function and defining $\delta_i = I(T_i \leq C_i)$, the data observed from n individuals in a random censoring scheme are therefore pairs (t_i, δ_i), for $i = 1, \ldots, n$, where $t_i = \min(T_i, C_i)$.

4. **Interval censoring**

In longitudinal studies or clinical trials, interval censoring may occur when a patient has periodic follow-ups and the patient's event time is only known to fall in an interval $(L_i, R_i]$, for $i = 1, \ldots, n$.

5. **Current status data**

Interval censored data where the intervals are either (C_i, ∞) or $(0, C_i]$ are called *current status data*. These data arise when subject i is examined once at time C_i and the event occurrence is determined at the same time C_i. For example, the time to tumor occurrence of a mouse can only be checked at the time of autopsy.

6. **Truncation**

Truncation is a condition where only those individuals who experience certain events can be observed by the investigator. For example, left truncation occurs when individuals of different ages enter a retirement study (i.e., delayed entry) and are followed until either an event or right censoring occurs. Individuals who die before retirement cannot be included in the study. In AIDS studies, right truncation often occurs in that individuals who are yet to develop AIDS are neither observed nor are they included in the study sample.

Calendar or Running Time

In many applications, the natural time scale of the failure process is not calendar or clock time but rather some measure of use or cumulative exposure. Cox and Oakes [3, Section 1.2, pp. 3–4] point out that "often the *scale* for measuring time is clock time, although other possibilities certainly arise, such as the use of operating time of a system, mileage of a car, or some measure of cumulative load encountered" Lawless [8, p. 6] also states that "Miles driven might be used as a time scale for motor vehicles, and number of pages of output for a computer printer or photocopier." This kind of alternative time scale is sometimes referred to as an *operational time* or *running time*. Survival analysis can be done in terms of either calendar time or operational time.

Estimation of Survival and Hazard Functions

The following are the basic theoretical quantities in lifetime models and related risk assessment:

1. **Failure-time probability density function $f(t)$**

This function describes the rate at which items will fail at any given time $T = t$.

$$f(t)\,\mathrm{d}t = Pr(T \in [t, t + \mathrm{d}t)) \tag{1}$$

2. **Survival function $S(t)$**

This function is the probability that an item will fail after time t.

$$S(t) = Pr(T > t) \tag{2}$$

3. **Hazard function $h(t)$**

This function describes the rate of failure at time t among items that have survived to time t.

$$h(t) = -\frac{\mathrm{d}\ln S(t)}{\mathrm{d}t} = \frac{f(t)}{S(t)} \tag{3}$$

4. **Cumulative hazard function $H(t)$:**

This function is the aggregate hazard to which an item is exposed up to time t.

$$H(t) = \int_0^t h(u)\,\mathrm{d}u = -\ln[S(t)] \tag{4}$$

Estimating the hazard function and survival function is the major focus of lifetime data analysis. From a process point of view, Aalen and Gjessing [22] present a good understanding of the hazard function.

If the data have already been grouped into intervals, a life table can be constructed to assess survival probabilities and hazard rates at different intervals. For parametric models, likelihood-ratio testing procedures and parametric regression models for handling censored and/or truncated data have been well developed.

Kaplan – Meyer Nonparametric Estimator for Survival Function

When individual lifetimes can be observed, the product-limit nonparametric estimator of the survival function developed by Kaplan and Meyer [23] is the most widely used method for estimating survival functions. In the situation where n distinct lifetimes are observed without censoring such that

their ordered values are $t_{(1)} < t_{(2)} < \cdots < t_{(n)}$, the survival function can be estimated by

$$\hat{S}(t) = \prod_{t_{(i)} \le t} \frac{(n-i)}{(n-i+1)} \tag{5}$$

In general, suppose d_i subjects have lifetime $t_{(i)}$, $1 \le i \le n$. Let Y_i denote the stochastic process indicating the number of individuals at risk at time $t_{(i)}$. Then

$$\hat{S}(t) = \prod_{t_{(i)} \le t} \left(1 - \frac{d_i}{Y_i}\right) \tag{6}$$

Nelson – Aalen Estimator for Cumulative Hazard Function

In a reliability context, Nelson [24] introduced the following estimator for the cumulative hazard function

$$\hat{H}(t) = \sum_{t_{(i)} \le t} \left(1 - \frac{d_i}{Y_i}\right) \tag{7}$$

Using methods of counting processes, Aalen [25] also derived this estimator, which is now referred to as the *Nelson – Aalen estimator*.

Semiparametric Proportional Hazards Models and Cox Regression

Cox [26] introduced the proportional hazards (PH) model.Given a covariate vector $\mathbf{X} = \mathbf{x} = (x_1, \ldots, x_n)$, this model assumes that the hazard function for lifetime T is of the form

$$h(t|\mathbf{X} = \mathbf{x}) = h_0(t) \exp(\boldsymbol{\beta}'\mathbf{x}) \tag{8}$$

where $\boldsymbol{\beta}$ is a vector of regression coefficients and $h_0(t)$ denotes the baseline hazard function.

For any two individuals with covariate vectors $\mathbf{X}_1 = \mathbf{x}_1$ and $\mathbf{X}_2 = \mathbf{x}_2$, the ratio of their hazard rates is constant through time.

$$\frac{h(t|\mathbf{X}_2 = \mathbf{x}_2)}{h(t|\mathbf{X}_1 = \mathbf{x}_1)} = \frac{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_2)}{h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}_1)} = \exp[\boldsymbol{\beta}'(\mathbf{x}_2 - \mathbf{x}_1)] \tag{9}$$

This quantity is called the *hazard ratio* (or *relative risk*) and is very useful in applications. For PH models, it can be shown that

$$S(t|\mathbf{X} = x) = [S_0(t)]^{\exp(\beta' \mathbf{x})} \tag{10}$$

where $S_0(t)$ is the baseline survival function.

Additive Hazards Models

In a PH model, as mentioned above, the covariates act multiplicatively on some unknown baseline hazard function. In an additive hazards regression model, the hazard rate at time t is assumed to be a linear combination of the covariate vector $\mathbf{X}$.

$$h(t|\mathbf{X} = \mathbf{x}) = \beta_0(t) + \boldsymbol{\beta}'\mathbf{x} \tag{11}$$

Therefore, in this model, the covariates act in an additive manner on an unknown baseline hazard function $\beta_0(t)$.

Accelerated Failure-Time Model

Given any lifetime t, the accelerated failure-time model is defined by the equation

$$S(t|\mathbf{X} = \mathbf{x}) = S_0[t \exp(\boldsymbol{\theta}'\mathbf{x})] \tag{12}$$

The factor $\exp(\boldsymbol{\theta}'\mathbf{x})$ is called an *acceleration factor* and indicates how a change in a covariate value changes the baseline time scale. Specifically, the hazard rate for an individual with covariate vector $\mathbf{X} = \mathbf{x}$ can be written as

$$h(t|\mathbf{x}) = \exp(\boldsymbol{\theta}'\mathbf{x}) h_0[t \exp(\boldsymbol{\theta}'\mathbf{x})] \tag{13}$$

Frailty Models

In multivariate survival analysis, the frailty model has become popular. In practice, a frailty is an unobservable random effect shared by subjects within a subgroup. In one formulation of frailty, a common random effect acts multiplicatively on the hazard rates of all members in the subgroup. The shared frailty model extends the PH regression model as follows. The hazard rate for the jth individual in the ith group is of the form

$$h_{ij}(t|\mathbf{x}_{ij}) = h_0(t) \exp[\sigma w_i + \beta'\mathbf{x}_{ij}] \tag{14}$$

where w_i is the frailty shared by the ith group.

First Hitting Time Models and Threshold Regression

FHTs arise naturally in many types of stochastic processes. In a lifetime context, the state of the

underlying process represents the strength of an item or the health of an individual. The item fails or the individual experiences a clinical endpoint when the process reaches an adverse threshold state for the first time. In many applications, the process is latent (i.e., unobservable). *Threshold regression* refers to FHT models with regression structures that accommodate covariate data. The parameters of the process, threshold state, and time scale may depend on the covariates.

An FHT model has two basic components: (a) a *parent stochastic process* $\{X(t)\}$ and (b) a *threshold*. The FHT is the time when the stochastic process first crosses the *threshold*. Whether the sample path of the parent process is observable or latent (unobservable) is an important distinguishing characteristic of the FHT model. Latent processes are the most common by far. As an example, we consider a Wiener process $\{X(t), t \geq 0\}$ with mean parameter μ, and variance parameter σ^2 as the parent stochastic process, and initial value $X(0) = x_0 > 0$.

In threshold regression, parameters μ, σ^2, and initial value x_0 will be connected to linear combinations of covariates using suitable regression link functions, as illustrated below for some parameter, say θ.

$$g_\theta(\theta_i) = \mathbf{z}_i \boldsymbol{\beta} \tag{15}$$

Here g_θ is the link function, parameter θ_i is the value of parameter θ for individual i, $\mathbf{z}_i = (1, z_{i1}, \ldots, z_{ik})$ is the covariate vector of individual i (with a leading unit to include an intercept term), and $\boldsymbol{\beta}$ is the associated vector of regression coefficients.

An identity function of form

$$\mu = \mathbf{z}\boldsymbol{\beta} = \beta_0 + \beta_1 z_1 + \cdots + \beta_k z_k \tag{16}$$

might be used to link parameter μ to the covariates and a logarithmic function

$$\ln(x_0) = \mathbf{z}\boldsymbol{\gamma} = \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_k z_k \tag{17}$$

for the initial value parameter x_0. Parameter σ^2 can be given an arbitrary value (say 1) because the process $\{X(t)\}$ is latent.

A running-time scale transformation can be accommodated by the threshold regression model. If $r(t)$ denotes the transformation of calendar time t to running time r, with $r(0) = 0$, and $\{X(r)\}$ is the parent process defined in terms of running time r, then the resulting process expressed in terms of calendar time is the process $X^*(t) = X[r(t)]$. A *composite running time* might be defined by

$$r(t) = \sum_{j=1}^{J} \alpha_j r_j(t) \tag{18}$$

where the $r_j(t)$ are different accumulation measures that can advance degradation or disease progression and the α_j are positive parameters that weight the contributions of the different measures.

As an example, consider a Wiener process with $x_0 > 0$ and the boundary as the zero level. It is well known that the FHT has an inverse Gaussian distribution if $\mu < 0$. The corresponding survival function $S(r|\mu, \sigma^2, x_0)$ can be written as

$$\boldsymbol{\Phi}\left[\frac{\mu r + x_0}{\sqrt{\sigma^2 r}}\right] - \exp(-2x_0\mu/\sigma^2)\boldsymbol{\Phi}\left[\frac{\mu r - x_0}{\sqrt{\sigma^2 r}}\right] \tag{19}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution. Lee *et al.* [27] demonstrated how to use this model in analyzing AIDS data. Lee *et al.* [28] used the threshold regression model in assessing lung cancer risk to a large cohort of US railroad workers from diesel exhaust exposure. Bayesian methods for FHT models have been discussed by Pettit and Young [29] and Shubina [30, 31].

Acknowledgment

This project was supported in part by NIH Grant OH008649.

References

[1] Kalbfleisch, J.D. & Prentice, R.L. (1980). *The statistical Analysis of Failure Time Data*, John Wiley & Sons, New York.

[2] Nelson, W. (1982). *Applied Life Data Analysis*, John Wiley & Sons, New York.

[3] Cox, D.R. & Oakes, D. (1984). *Analysis of Survival Data*, Chapman & Hall, New York.

[4] Fleming, T.R. & Harrington, D.P. (1991). *Counting Processes and Survival Analysis*, John Wiley & Sons, New York.

[5] Andersen, P.K., Borgan, O., Gill, R.D. & Keiding, N. (1993). *Statistical Models Based on Counting Processes*, Springer-Verlag, New York.

[6] Marubini, E. & Valsecchi, M.G. (1995). *Analyzing Survival Data from Clinical Trials and Observational Studies*, John Wiley & Sons, New York.

[7] Klein, J.P. & Moeschberger, M.L. (1997). *Survival Analysis: Techniques for Censored and Truncated Data*, Springer.
[8] Lawless, J.F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd Edition, John Wiley & Sons.
[9] Lee, E.T. & Wang, J.W. (2003). *Statistical Methods for Survival Data Analysis*, John Wiley & Sons.
[10] Hougaard, P. (2000). *Analysis of Multivariate Survival Data*, Springer, New York.
[11] Ibrahim, J., Chen, M.-H. & Sinha, D. (2001). *Bayesian Survival Analysis*, Springer-Verlag.
[12] Singpurwalla, N.D. (2006). *Reliability and Risk: A Bayesian Perspective*, John Wiley & Sons.
[13] Whitmore, G.A. (1986). First passage time models for duration data regression structures and competing risks, *The Statistician* **35**, 207–219.
[14] Whitmore, G.A. (1995). Estimating degradation by a Wiener diffusion process subject to measurement error, *Lifetime Data Analysis* **1**, 307–319.
[15] Lu, J. (1995). *A reliability model based on degradation and lifetime data*, Ph.D. thesis, McGill University, Montreal.
[16] Whitmore, G.A., Crowder, M.J. & Lawless, J.F. (1998). Failure inference from a marker process based on a bivariate Wiener model, *Lifetime Data Analysis* **4**, 229–251.
[17] Onar, A. & Padgett, W.J. (2000). Inverse Gaussian accelerated test models based on cumulative damage, *Journal of Statistical Computation and Simulation* **66**, 233–247.
[18] Aalen, O.O. & Gjessing, H.K. (2004). Survival models based on the Ornstein-Uhlenbeck process, *Lifetime Data Analysis* **10**, 407–423.
[19] Padgett, W.J. & Tomlinson, M.A. (2004). Inference from accelerated degradation and failure data based on Gaussian process models, *Lifetime Data Analysis* **10**, 191–206.
[20] Padgett, W.J. & Tomlinson, M.A. (2005). Accelerated degradation models for failure based on geometric Brownian motion and Gamma processes, *Lifetime Data Analysis* **11**, 511–527.
[21] Lee, M.-L.T. & Whitmore, G.A. (2006). Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary, *Statistical Sciences* **21**, 501–513.
[22] Aalen, O.O. & Gjessing, H.K. (2001). Understanding the shape of the hazard rate: a process point of view, *Statistical Sciences* **16**, 1–22.
[23] Kaplan, E.L. & Meyer, P. (1958). Nonparametric estimation from incomplete observations, *Journal of the American Statistical Association* **53**, 457–481.
[24] Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data, *Technometrics* **14**, 945–965.
[25] Aalen, O.O. (1978). Nonparametric inference for a family of counting processes, *Annals of Statistics* **6**, 701–726.
[26] Cox, D.R. (1972). Regression models and lifetables (with discussion), *Journal of the Royal Statastical Society (JRSSB), Series B* **34**, 187–220.
[27] Lee, M.-L.T., DeGruttola, V. & Schoenfeld, D. (2000). A model for markers and latent health status, *Journal of the Royal Statistical Society, Series B* **62**, 747–762.
[28] Lee, M.-L.T., Garshick, E., Whitmore, G.A., Laden, F. & Hart, J. (2004). Assessing lung cancer risk to rail workers using a first hitting time regression model, *Environmetrics* **15**, 1–12.
[29] Pettit, L.I. & Young, K.D.S. (1999). Bayesian analysis for inverse Gaussian lifetime data with measures of degradation, *Journal of Statistical Computation and Simulation* **63**, 217–234.
[30] Shubina, M. (2005). *Bayesian analysis for markers and degradation*, Ph.D. Dissertation, Biostatistics Department, Harvard School of Public Health, Boston.
[31] Shubina, M. (2005). *Threshold models with markers measured before observed event times*, Ph.D. Dissertation, Biostatistics Department, Harvard School of Public Health, Boston.

MEI-LING TING LEE

Lifetime Survival Function *see* Non- and Semiparametric Models and Inference for Reliability Systems

Limit Reliability Function of a System *see* Reliability of Large Systems

Linkage Analysis

In biology, a *trait* or character is a feature of an organism. Most biological traits including human health and physical characteristics are determined primarily by the chemical composition of the proteins created in cells. The composition and formation

ENCYCLOPEDIA OF QUANTITATIVE RISK ANALYSIS AND ASSESSMENT

Volume 3

K–Q

Editors-in-Chief

EDWARD L. MELNICK

New York University, New York, NY, USA

BRIAN S. EVERITT

Institute of Psychiatry, King's College, London, UK

WILEY

The Atrium,
Southern Gate,
Chichester,
West Sussex,
PO19 8SQ, England

Telephone: (+44) 1243 779777
Email (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on www.wiley.com

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street,
Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street,
San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12,
D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 165 Cremorne Street,
Richmond Vic. 3121, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01,
Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 6045 Freemont Boulevard,
Mississauga, Ontario, Canada L5R 4J3

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Cataloging-in-Publication Data

Encyclopedia of quantitative risk analysis and assessment / editors-in-chief, Edward L. Melnick, Brian S. Everitt.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-03549-8 (cloth)
1. Risk assessment – Encyclopedias. 2. Mathematical statistics – Encyclopedias. I. Melnick, Edward L. II. Everitt, Brian.
QA297.E52 2008
363.1'02 – dc22

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library 2008012270

978-0-470-03549-8 (HB)

Typeset in 9½/11½ pt Times by Laserwords Private Limited, Chennai, India.
Printed and bound in Markono Group, Singapore.
This book is printed on acid-free paper responsibly manufactured from sustainable forestry in which at least two trees are planted for each one used for paper production.