1

Introduction to Lifetime Data and Regression Models

COPYRIGHTED MATERIAL

1.1. Basics

The principal aim of this book is to present a particular approach to the regression analysis of lifetime data in detail, and to discuss its features and advantages in comparison to older-established approaches. This objective will first involve undertaking a general review of the various regression methods that can be found in the literature.

This opening chapter provides a brief review of the basic features of lifetime data and its modeling. First of all, what is meant by lifetime data? A great deal of statisticians' activity goes into studying whether or not an event occurs and how various factors influence its occurrence. If we move on from the simple yes/no fact of occurrence to also examining how long it takes until the event occurs, we enter the realm of "time to event" or "lifetime" data. Basic examples include how long a machine operates until it breaks down (the event is the breakdown) and how long a patient lives after undergoing heart transplantation (the event is death). These examples show two major areas of application. One is in engineering and technology (where the subject is usually known as reliability modeling) and the other is in biomedical sciences (known as survival analysis). However, other areas of application include all those in which statistics is used - in other words, in virtually every science. As the two examples of a machine's breakdown and a patient's survival suggest, applications of lifetime data analysis can have immense practical importance. Well-known textbooks with wide coverage of lifetime data analysis include

Lawless [LAW 03], Collett [COL 14], Kalbfleisch and Prentice [KAL 02] and Klein and Moeschberger [KLE 03]. A brief review was given by Hougaard [HOU 99]. Other papers reviewing the analysis of lifetime data include Kiefer [KIE 88] in economics and Chung *et al.* [CHU 91] in criminology.

In mathematical terms, lifetime is denoted by T. No two machines are identical or operate under identical conditions; no two people are quite alike. Consequently, we treat T as a random variable, which follows some distribution in the relevant population of machines or people (in general, units). We note that T must be non-negative. Furthermore, in this book, we will follow the vast majority of the literature in treating the time scale as continuous. Consequently, we suppose that $T \sim f(t), t > 0$ for some probability density function (pdf) $f(.)$, and hence that $F(t) = P(T \leq t) = \int_0^t f(u)du$. The functions that present particular interest are the following:

– the survival function $S(t) = \bar{F}(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du$

– the hazard function $h(t) = f(t)/S(t)$.

The former is $P(T > t)$, the probability of survival for at least time t - the probability that the machine is still operating, or that the patient is still alive, after this time. In engineering and technological applications, this probability is called reliability and the notation $R(t)$ is usually used instead of $S(t)$. The term hazard function is replaced by failure rate. Other terms in use for the same function include force of mortality (in demography) and intensity.

The survival function or reliability $P(T > t)$ is a quantity of basic scientific and practical importance. For example, in medical settings, a patient's prognosis might be expressed as his or her five-year survival probability, and in manufacturing, reliability is obviously related to how long a guarantee period can be offered for a product. The hazard function can be interpreted as the instantaneous rate of failure at time t, given that the unit has survived that long, and hence the term failure rate. However, it is important to remember that the hazard refers to failure conditionally on survival to that time (the unconditional failure rate is of course given by the pdf of the lifetime distribution). More precisely, the hazard function gives the conditional probability of failure in the next short interval of time $(t, t + \delta t]$,

for a unit that is still functioning at time t:

$$P(t < T \leq t + \delta t | T > t) = h(t)\delta t. \quad [1.1]$$

Also useful and important is the cumulative hazard, $H(t) = \int_0^t h(u)du = -\ln S(t)$ as well as the mean residual life, given by

$$\mu(t) = \frac{\int_t^\infty S(u)du}{S(t)},$$

which is the expected lifetime still to come for a unit that has already survived until time t.

The functions f, F, S, h and H are all equivalent, in the sense that knowing any one of them enables all the others to be deduced. Complete details can be found in standard references (e.g. Lawless [LAW 03]), which also present detailed expressions for the more widely used parametric lifetime distributions $f(.)$, such as the exponential, Weibull, gamma, log-normal and others. In the following section, we present the best known parametric lifetime distribution as an example, the Weibull distribution. Some details of another parametric distribution - the inverse Gaussian distribution - can be found in section 2.5 and elsewhere in the text. Some general properties of lifetime distributions are presented briefly by Olkin [OLK 16] and at length by Marshall and Olkin [MAR 07].

Some aspects of these distributions that have major importance in the analysis of lifetime data, such as the hazard function, present little interest in other fields of statistics. Conversely, some properties of distributions that have great general importance do not concern us much in lifetime data analysis. The prime example is the mean of the distribution. This is because most lifetime distributions are highly skew, with a long tail to the right. For a distribution of this shape, the median is usually reported rather than the mean. Furthermore, in practice, the restricted duration of a study may make it difficult to estimate the mean accurately (see comments in section 1.9). However, sometimes a restricted mean survival time (RMST) can be used. By the definition of the mean μ of a distribution, and assuming that it exists, we have

$$\mu = E(T) = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt.$$

Adapting this, we define the RMST up to time t^* as

$$\mu(t^*) = \int_0^{t^*} S(t)dt.$$

This can also be written as $\mu(t^*) = E[\min(T, t^*)]$. The RMST thus represents the expected duration of survival up to time t^*: in other words, how much of the interval $(0, t^*)$ an individual will survive, on average. RMSTs can be computed for different values of t^* and compared between groups of subjects; this will be especially useful if the relation between survival in the groups is not simple (e.g. the first group does better than the second initially, but later on, the second group has the lower hazard). See for example A'hern [AHE 16] and references therein.

Two further general comments about lifetime data must be made. First, it is a characteristic feature of such data that not all the units under study will actually experience the event during the study. Some patients will still be alive when the medical researcher closes the data file for analysis; some machines will still be functioning when the time allotted to the study runs out. The lifetimes of these units are said to be right censored at the times when they were observed. They provide information that must be taken into account in the analysis even though this information takes the different form $T > t$ rather than $T = t$. This can only be done easily if the censoring process is uninformative about the lifetime (see section 3.1). Other types of censoring as well as the less common phenomenon of truncation are discussed in standard references (see, for example, Lawless [LAW 03, Chapter 2]).

The second additional comment is the observation that a "lifetime" need not correspond to clock time, or even be measured in units of time at all. For a machine, the relevant time may be the time for which it is actually operating, excluding periods when it is turned off or is idle. For a car, the operating "time" would probably be measured better by how many kilometers it has covered rather than by the calendar age of the vehicle, because this will be the more important factor as far as wear and tear is concerned. Sometimes there may be several alternatives: for an aircraft, for example, calendar age, flight hours and number of landings could all be relevant measures of lifetime (see Duchesne and Lawless [DUC 00]). The question of the appropriate time scale is also discussed by Farewell and Cox [FAR 79], Oakes [OAK 95] and Kordonsky and Gertsbakh [KOR 93, KOR 97] as well as others. Later on, we will see

cases where overall "time" is a weighted sum of the durations of the periods of time spent in different states (e.g. the movement of an employee through different jobs with varying exposure to health risks).

Finally, we observe that the concept of a non-negative random variable describing the point at which an event occurs can be adapted to cases where the variable is not a time at all, but, for example, the load placed on a structure. The load is increased until the structure fails.

1.2. The classic lifetime distribution: the Weibull distribution

Here, for the purpose of illustration, we provide details of the *Weibull distribution* (named after the Swede Waloddi Weibull), which is the most widely used parametric model for lifetime data. Empirically, it has been found to fit well to data of many kinds, and in fact, its use with lifetime data can be justified by theoretical arguments (see below).

The pdf of the Weibull distribution in one common parameterization is

$$f(t) = \exp\{-(t/\alpha)^\eta\}\eta t^{\eta-1}/\alpha^\eta, \;\; t > 0,$$

where $\alpha > 0$ is the scale parameter and $\eta > 0$ is the shape parameter. The special case $\eta = 1$ gives the exponential distribution. The survival or reliability function is

$$\begin{aligned} S(t) &= \int_t^\infty e^{-(u/\alpha)^\eta}\eta\alpha^{-1}(u/\alpha)^{\eta-1}du \\ &= \int_{(t/\alpha)^\eta}^\infty e^{-v}dv \quad [\text{substituting } v = (u/\alpha)^\eta] \\ &= \left[-e^{-v}\right]_{(t/\alpha)^\eta}^\infty \\ &= \exp\{-(t/\alpha)^\eta\} \end{aligned}$$

and therefore the hazard function is

$$h(t) = \eta\, t^{\eta-1}/\alpha^\eta, \;\; t > 0.$$

The behavior of the hazard function is as follows:

$$h(t) = \begin{cases} increasing, & \text{when } \eta > 1 \\ constant, & \text{when } \eta = 1 \\ decreasing, & \text{when } \eta < 1. \end{cases}$$

This means that the Weibull distribution is quite flexible when it comes to describing lifetime data. However, it is unable to capture various features that are sometimes observed in hazard functions in real life, such as when the hazard increases to a peak and then falls, or when it falls to a minimum and then increases.

Figure 1.1 presents examples of the shapes of the Weibull pdf, survival function and hazard function for various values of the parameters of the distribution. Note that the distribution is skewed to the right, which is a characteristic feature of lifetime distributions.

The expected value and the variance of the lifetime T can be found using the following expression for the r-th moment of the distribution:

$$\begin{aligned} E(T^r) &= \int_0^\infty t^r f(t) dt \\ &= \int_0^\infty \alpha^r u^{r/\eta} e^{-u} du \quad [\text{substituting } u = (t/\alpha)^\eta] \\ &= \alpha^r \int_0^\infty u^{r/\eta} e^{-u} du \\ &= \alpha^r \Gamma(1 + r/\eta), \end{aligned}$$

where $\Gamma(\cdot)$ is the gamma function. Setting $r = 1$ and 2 gives

$$E(T) = \alpha\Gamma(1 + \eta^{-1}) \quad \text{and} \quad E(T^2) = \alpha^2\Gamma(1 + 2\eta^{-1})$$

and hence, the variance of the lifetime T is

$$V(T) = \alpha^2 \left[\Gamma(1 + 2\eta^{-1}) - \{\Gamma(1 + \eta^{-1})\}^2\right].$$

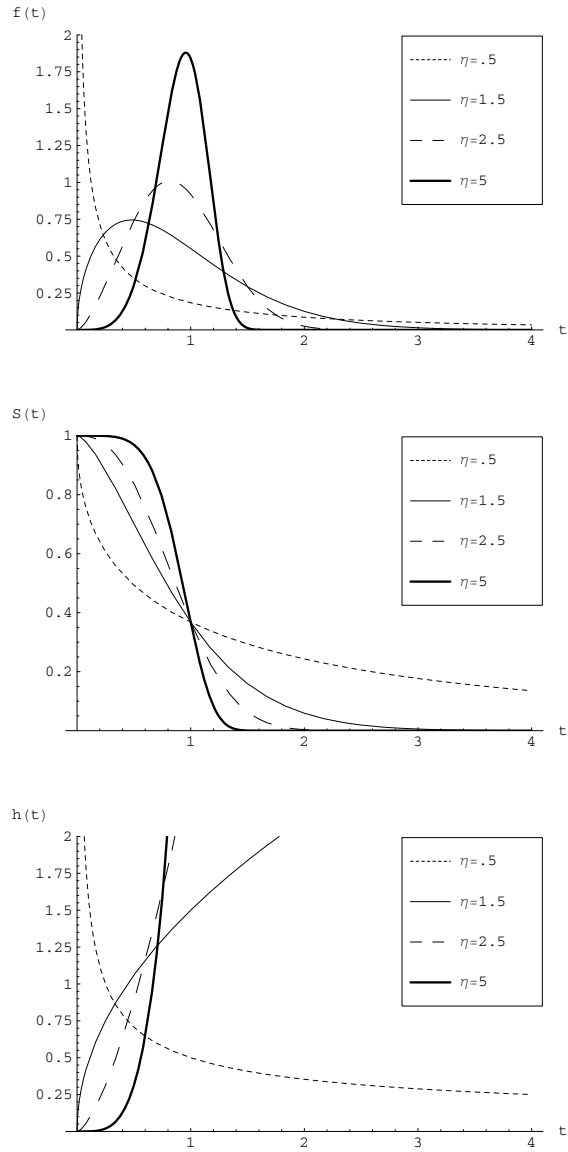

Figure 1.1. *Plots of the pdf (upper diagram), survival function (middle diagram) and hazard function (lower diagram) of the Weibull distribution for selected values of* η*, with* $\alpha = 1$

The following alternative parameterization of the Weibull distribution is also often seen in the literature:

$$f(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^{\beta}},\ t > 0,\ \alpha > 0,\ \beta > 0.$$

Countless examples of applications of the Weibull distribution can be found in the literature, which mostly concern the lifetimes or strengths of materials. One application of a different kind is by McDonald *et al.* [MCD 96] to the lifetimes of a species of bird. By fitting a Weibull distribution to lifetimes and finding that the shape parameter is greater than one, they concluded that mortality rates increase with age - so-called actuarial senescence. This was claimed to be the first demonstration of the phenomenon in an unmanipulated, natural population and thus constituted the first empirical evidence against a long-held assumption that mortality of birds is generally independent of age.

The theoretical argument for the Weibull distribution's widespread use in practical situations is the following "weakest link" argument. Many of the units that we study can be regarded as being made up of smaller components or parts, and it may be reasonable to suppose that the durability or strength of the whole is equal to the durability or strength of the weakest part, just as a chain is made up of links and the chain's strength is given by the strength of its weakest link. Given this structure, the distribution of the unit's lifetime is determined by the distribution of the *minimum* of the set of random variables that represent the lifetimes of the unit's components. Statistical theory demonstrates that only certain distributions have the necessary properties to represent such a minimum. One of these *extreme value distributions* is the Weibull.

The literature contains many lifetime distributions, most of which do not see any practical application. For example, many extensions of the Weibull distribution have been devised (see, for example, Caroni [CAR 14a]). In gaining extra flexibility, these extensions lose appealing properties of the Weibull distribution, such as the extreme value interpretation and the properties of the regression models that will be discussed later on in this chapter.

1.3. Regression models for lifetimes

Although the fact that no two units can be identical means that there will always be a random component in the lifetime, in part it may be possible to predict the lifetime from the factors or covariates that describe the unit or the conditions under which it has been operating. A patient's prognosis after an operation, for example, is likely to depend to some degree on his or her age, on medical history and on the variables that describe the state of health at the time of the operation. An older patient, in poor condition and with a long history of ill health will be expected to have a shorter time-to-event (death, relapse) than a younger patient who started out in better shape. Car tires would be expected to wear out quicker if the vehicle is often driven off-road.

The concept of introducing the dependence of an outcome variable such as the duration of a lifetime on the values of covariates is familiar from the multiple linear regression model

$$y = \boldsymbol{\beta}'\mathbf{x} + \epsilon,$$

where $\mathbf{x} = (x_0, x_1, \ldots, x_p)'$ is the vector of covariates, with $x_0 \equiv 1$. The standard model takes the distribution of the random error term as $\epsilon \sim N(0, \sigma^2)$, in which case the model for the dependent variable can be written as

$$y \sim N(\mu, \sigma^2) \text{ with } \mu = \mu(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x}. \qquad [1.2]$$

This expression suggests one way of extending a regression model to situations where it is not reasonable to assume a normally distributed dependent variable: select a more appropriate distribution (e.g. Poisson with parameter μ) and link its parameters in some way to the linear predictor $\boldsymbol{\beta}'\mathbf{x}$ formed from the covariates (e.g. $\ln \mu = \boldsymbol{\beta}'\mathbf{x}$ is often an appropriate choice in combination with the Poisson distribution). In this way, we obtain the class of generalized linear models (GLM) in which the mean parameter is related to the linear predictor (see McCullagh and Nelder [MCC 89]). For example, the GLM version of [1.2] when the dependent variable Y is a count of the number of events and therefore might follow the Poisson distribution is

$$y \sim \text{Poisson } (\mu) \text{ with } \ln \mu = \boldsymbol{\beta}'\mathbf{x}.$$

In the much wider class of generalized additive models for location, scale and shape (GAMLSS), as many as four parameters of the distribution can depend on covariates (see Rigby and Stasinopoulos [RIG 05]).

The same approach can be taken to distributions that are often used in modeling lifetime data. For example, the inverse Gaussian distribution and the gamma distribution both belong to the exponential family that is modeled in the standard framework of generalized linear models. The inverse Gaussian distribution will be mentioned in this context in section 3.2. However, there are other ways of approaching the matter in the context of lifetime data, which give rise to the general classes of models that will be considered in the following sections.

Parametric lifetime regression models are usually fitted by direct maximization of a likelihood function using numerical methods. Given a sample of n independent observations $\{(t_i, \mathbf{x}_i, \delta_i), i = 1, \ldots, n\}$, where unit i with covariates $\mathbf{x}_i$ has lifetime t_i and censoring indicator δ_i (=1 if t_i is an observed failure time, 0 if t_i is a right censored observation time), the likelihood is

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i|\mathbf{x}_i, \boldsymbol{\theta})^{\delta_i} S(t_i|\mathbf{x}_i, \boldsymbol{\theta})^{1-\delta_i}, \qquad [1.3]$$

where the parameter vector $\boldsymbol{\theta}$ includes the regression coefficients. Using the relationships between the probability density function, hazard function and survival function, this likelihood can be written in various alternative forms, if desired. For example, using $h(t) = f(t)/S(t)$ to substitute for $f(t)$, the likelihood can be written in terms of the hazard and survival functions as

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} h(t_i|\mathbf{x}_i, \boldsymbol{\theta})^{\delta_i} S(t_i|\mathbf{x}_i, \boldsymbol{\theta}),$$

which may sometimes be convenient.

1.4. Proportional hazards models

To illustrate one of the main approaches to the regression modeling of lifetime data, we begin with the widely used Weibull distribution. Note that

this does not fall within the framework of generalized linear models. Its survival function is

$$S(t) = \exp\{-(t/\alpha)^{\eta}\}, \;\; t > 0, \; \alpha > 0, \; \eta > 0.$$

We introduce the effect of covariates $\mathbf{x}$ on the parameters of the model, giving

$$S(t|\mathbf{x}) = \exp\{-(t/\alpha(\mathbf{x}))^{\eta}\}.$$

The scale parameter α now depends on $\mathbf{x}$. (This is the usual form of model, although it is possible to allow the parameter η to depend on $\mathbf{x}$ instead, or to let both the parameters depend on covariates at the same time. In the latter case, the covariates affecting α and η do not need to be the same. A recent paper by Burke and MacKenzie discusses the general approach where both the parameters depend on the covariates for the Weibull distribution and in general [BUR 16a]. See section 3.2.)

Let $\alpha(\mathbf{x}) = \alpha e^{\boldsymbol{\beta}'\mathbf{x}}$ or simply $\alpha(\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}}$ since the constant α can be absorbed into the exponent. (Once again, this is the usual form of the model, although not the only possibility.) Note that the function $e^{\boldsymbol{\beta}'\mathbf{x}}$ is positive, a restriction that is necessary here.

The hazard function is readily obtained from $h(t) = -\frac{d}{dt}\ln S(t)$ as

$$h(t|\mathbf{x}) = \eta t^{\eta-1} e^{\boldsymbol{\theta}'\mathbf{x}},$$

where $\boldsymbol{\theta} = -\eta\boldsymbol{\beta}$. Now compare the hazard functions of the two units with covariate vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. Their ratio is

$$\frac{h(t|\mathbf{x}_1)}{h(t|\mathbf{x}_2)} = \frac{\eta t^{\eta-1} e^{\boldsymbol{\theta}'\mathbf{x}_1}}{\eta t^{\eta-1} e^{\boldsymbol{\theta}'\mathbf{x}_2}} = e^{\boldsymbol{\theta}'(\mathbf{x}_1-\mathbf{x}_2)},$$

which does not depend on time. In other words, the hazard function of one unit remains in constant proportion to the hazard of the other. This is the proportional hazards (PH) model, which applies for any non-negative $\alpha(\mathbf{x})$, not just $e^{\boldsymbol{\beta}'\mathbf{x}}$. A particular version of the PH model - Cox's semi-parametric PH regression model - has virtually become the standard model for evaluating biomedical lifetime data. This model will be described in section 1.14.

Another example of a model that possesses the PH property is obtained from the Gompertz distribution, which is described most simply by its hazard function

$$h(t) = \mu\phi^t,\ \mu > 0,$$

which is decreasing in t for $\phi < 1$, increasing for $\phi > 1$ and constant for $\phi = 1$, in which case it reduces to the exponential distribution. The PH model modifies the distribution into another Gompertz distribution with a different value of μ but the same ϕ (see Hougaard [HOU 99]). The use of the Gompertz distribution is restricted mainly to demography and actuarial science, where it has a long history. It has been used, for example, to describe mortality among adults. In these contexts, $\phi > 1$ (increasing hazard - i.e. mortality - at older ages).

Now consider what happens if $\phi < 1$. In the survival function

$$S(t) = \exp[-\mu\left(\phi^t - 1\right) / \ln\phi\,]$$

the term ϕ^t tends to zero, therefore the limit of $S(t)$ as t tends to infinity is not zero. In fact, with $\mu > 0$ as before, the limit of $S(t)$ is e^{ξ} where $\xi = \mu/\ln\phi < 0$ and hence $0 < S(\infty) < 1$. This version of the Gompertz distribution is called the negative Gompertz distribution by Marshall and Olkin [MAR 07, Chapter 10]). Because $S(\infty) < 1$, it is an improper distribution or defective distribution. However, this feature is not necessarily a defect as far as using the distribution as a statistical model goes. The existence of a positive probability mass at infinity can be interpreted to mean that the corresponding proportion of the population will never die. This assumption is clearly meaningless in the actuarial study of human mortality, but could possibly be very realistic in a shorter-term study of mortality from a disease after a treatment. In the latter case, those who "never die" (at least, from the disease under study) are those who have been cured of the disease. Thus, the apparent defect becomes an asset of the model in its ability to model data. Examples of the application of the Gompertz distribution that exploit this characteristic include Cantor and Shuster [CAN 92] and Gieser *et al.* [GIE 98].

This feature will be mentioned (under the name of cured fraction or long-term survivors) quite often in this book, because it is shared by the inverse Gaussian distribution, which, as will appear in due course, is a central topic.

Both the basic examples of a distribution that possess the PH property, the Weibull and the Gompertz distributions, have hazard functions that are monotonic in t. This is not a necessary condition for a PH distribution. If we write the PH property in its general form

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x}),$$

where $h_0(t)$ is a baseline hazard function, it is obvious that if the baseline hazard function $h_0(t)$ has a maximum or a minimum at a value t_0, then the hazard functions $h(t|\mathbf{x})$ for every $\mathbf{x}$ likewise have maxima or minima as the case may be at this same value t_0. For example, if the hazard function falls to a minimum and thereafter increases - often claimed to be a realistic form in various situations - then that minimum would have to occur at the same time irrespective of the values of the covariates. This seems unlikely to be true in practice.

Bagdonavičius and Nikulin [BAG 99] proposed an extension of the PH model to the generalized PH model. The hazard function can be written as

$$h(t|\mathbf{x}) = r\left\{\mathbf{x}(t)\right\} q\left\{H(t|\mathbf{x})\right\} h_0(t),$$

where r and q are positive functions. Thus, the hazard rate at time t depends not only on the current values of the covariates (as in PH) but also on their history as expressed by the cumulative hazard $H(t)$. One special case is the generalized linear PH model, in which $r(\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}}$ as usual and

$$q\left\{H(t|\mathbf{x})\right\} = e^{\gamma\, H(t|\mathbf{x})},$$

so that

$$h(t|\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}+\gamma\, H(t|\mathbf{x})} h_0(t).$$

Thus, the cumulative hazard up to this moment in time is treated as an additional, unknown covariate. This model is examined further by Bagdonavičius *et al.* [BAG 05].

1.5. Checking the proportional hazards assumption

The theory that was outlined in the preceding paragraphs requires the assumption of PH. If this assumption is inappropriate for the data, then it is

meaningless to fit this particular regression model. How can we check that the assumption is appropriate?

The hazard function

$$h(t|\mathbf{x}) = h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}}$$

gives the survival function

$$S(t|\mathbf{x}) = \exp\{-H_0(t)e^{\boldsymbol{\beta}'\mathbf{x}}\},$$

where $H_0(t)$ is the cumulative hazard function corresponding to the baseline hazard function $h_0(t)$.

Consequently,

$$\ln\{-\ln S(t|\mathbf{x})\} - \ln H_0(t) = \boldsymbol{\beta}'\mathbf{x}$$

which means that the curves

$$\ln\{-\ln S(t|\mathbf{x})\}$$

for different values of $\mathbf{x}$ are simply the horizontally displaced versions of the curve $\ln H_0(t)$ when plotted against t. Consequently, all the curves $\ln\{-\ln S(t|\mathbf{x}_i)\}$ for different $\mathbf{x}_i$ are parallel to each other.

This observation suggests a simple way of checking for PH:

– compute non-parametric Kaplan-Meier estimates of the survivor function $\hat{S}(t|\mathbf{x})$ for selected $\mathbf{x}$;

– plot $\ln\{-\ln \hat{S}(t|\mathbf{x})\}$ against t for each selected $\mathbf{x}$.

If all the lines for the various $\mathbf{x}$ are indeed parallel to each other, then the assumption of the proportional hazards is correct. This idea applies to *any* PH model, but does not tell us which distribution is the appropriate one if we are to carry out a parametric regression. Also, it does not require that the proportionality be expressed by the multiplicative factor $g(\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}}$; any non-negative function $g(\mathbf{x})$ would do.

However, we could plot against the appropriate function of time to help determine the distribution. For example, if the lifetime distribution is thought to be Weibull, then from $S(t) = \exp\{-(t/\alpha)^{\eta}\}$, it follows that

$$\ln\{-\ln \hat{S}(t)\} = \eta \ln t - \eta \ln \alpha.$$

The plot of $\ln\{-\ln \hat{S}(t|\mathbf{x})\}$ against $\ln t$ should give a straight line. It is usually easier to see that straight lines are parallel rather than arbitrary curves.

The weakness of this procedure is that the estimates $\hat{S}(t|\mathbf{x})$ will only be satisfactory for this purpose if they are based on sufficiently large numbers of observations; otherwise, the sampling variability will be so large that it might be hard to say whether the curves are parallel or not. This means that there must be a reasonably large number of observations that share the same value of the covariates. For this reason, the method can only be applied if the covariates are few, or by carrying out suitable grouping of values of the covariates.

EXAMPLE 1.1.– *Table 1.1 provides McCool's data on hardened steel specimens tested until failure at four different levels of stress [MCC 80].*

Stress (10^6 psi)	**Ordered lifetimes**
.87 :	1.67 2.20 2.51 3.00 3.90 4.70
	7.53 14.70 27.80 37.40
.99 :	0.80 1.00 1.37 2.25 2.95 3.70
	6.07 6.65 7.05 7.37
1.09 :	0.012 0.18 0.20 0.24 0.26 0.32
	0.32 0.42 0.44 0.88
1.18 :	0.073 0.098 0.117 0.135 0.175 0.262
	0.270 0.350 0.386 0.456

Table 1.1. *McCool's data on hardened steel specimens tested until failure at four different levels of stress [MCC 80]*

Figure 1.2 shows the results of carrying out the above graphical procedure on these data. Remember that we are looking for parallel lines describing the sets of points corresponding to these four stress levels in order to confirm the PH assumption. If furthermore they are straight lines, then the Weibull distribution seems to apply. For easier comparison, we have superimposed on the diagram the lines obtained by fitting a Weibull distribution by maximum likelihood to each sample separately. At first sight, it seems very doubtful that the lines are parallel, although one could possibly say that, with only ten observations per group, there will be quite a large sampling variation in the four slopes. However, notice that there is one rather unusual data point, namely the value of 0.012 in the third group, which appears in the bottom left of the diagram. This very early failure seems to be an "outlier", that is, an

"observation that is not consistent with the model and the bulk of the data" (see Nelson [NEL 90]). If we omit this point from the fitting, as in Figure 1.3, then the line for the third group is very similar to that of the fourth. All the lines are straight; therefore, it seems that the Weibull model, and hence PH, are reasonable for these data, with the reservation that there is one outlying observation.

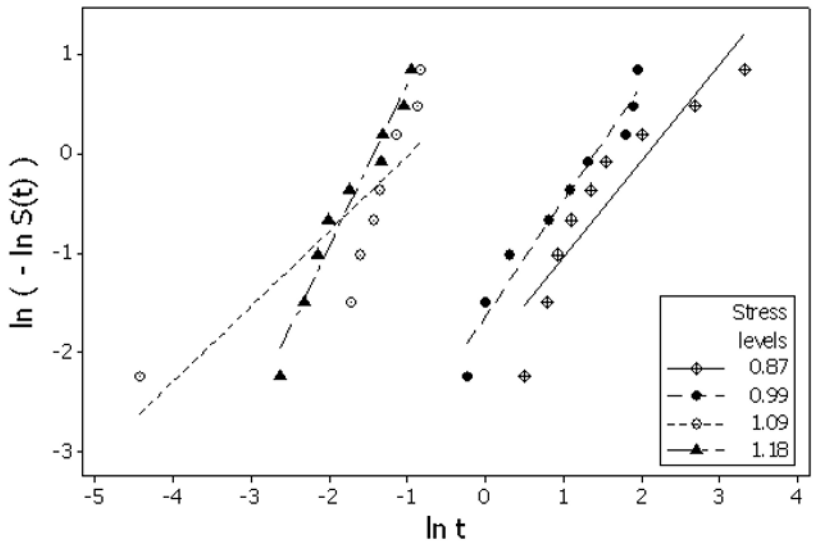

Figure 1.2. *Plot for checking the PH assumption for McCool's data*

We note here that outliers can have a major effect on the fit of statistical models to data, and there is an enormous amount of literature on their detection (see Barnett and Lewis [BAR 94]), although not much of it is applicable to lifetime data. The first step when faced by a possible outlier is to check that the value was recorded correctly and to try to find out if it was recorded under conditions that differed in any way from the rest of the data. If it is confirmed that the value is invalid, then it can be omitted. Otherwise, Nelson suggests that it may be wise to analyze the data with and without this point, to see whether it affects the results appreciably. However, he also points out that "in a sense, suspect data are always right; that is, they reflect something real happening. Only the model or our understanding is inadequate" [NEL 90, p. 209].

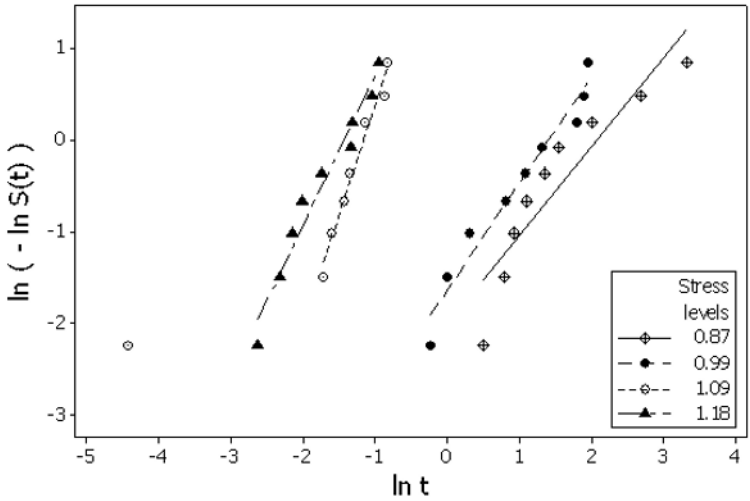

Figure 1.3. *Checking the proportional hazards assumption for McCool's data: one point at bottom left omitted from the fitting*

In section 1.16, we look at another way of checking whether a model provides an adequate description of the data, that is, by examining residuals. As the method is not restricted to PH, but applies equally well to other regression models, we will first look at the other prominent type, accelerated failure time models, and several other less widely used regression models.

1.6. Accelerated failure time models

In order to describe the second main way of introducing dependence on covariates into lifetime models, we begin by considering a model for lifetimes in the form "systematic component + random error" suggested by the general linear regression model given in section 1.3, with the dependent variable the logarithm of lifetime T (thus avoiding the problem of restriction to non-negative values):

$$\ln T_{\mathbf{x}} = \mu + \boldsymbol{\beta}'\mathbf{x} + \sigma\epsilon, \qquad [1.4]$$

where the error term ϵ has location parameter zero and scale parameter 1. The important case of the log-normal distribution arises from taking $\epsilon \sim N(0,1)$ in the regression model for $\ln T_{\mathbf{x}}$. The distribution of $T_{\mathbf{x}}$ is then log-normal

for any $\mathbf{x}$. Although the log-normal distribution is widely used because it often provides a good fit to lifetime data, it has a feature that may be unrealistic in many cases, namely that the hazard function increases to a peak and then declines as time increases. On the other hand, the PH property usually implies that the hazard function is monotonically increasing or decreasing (see section 1.4). Thus, we observe that the AFT model incorporates a much wider range of behavior than the PH model is capable of doing. Other distributions for T follow from other assumptions on ϵ. For example, T follows the Weibull distribution if ϵ follows the Gumbel distribution (see section 1.13).

From the above equation for $\ln T_{\mathbf{x}}$,

$$S(t|\mathbf{x}) = P(T_{\mathbf{x}} > t) = P(\mu + \boldsymbol{\beta}'\mathbf{x} + \sigma\epsilon > \ln t), \qquad [1.5]$$

hence,

$$\begin{aligned} S(t|\mathbf{x}) &= P\left(\ln T_0 + \boldsymbol{\beta}'\mathbf{x} > \ln t\right) \\ &= P\left(T_0 > te^{-\boldsymbol{\beta}'\mathbf{x}}\right) \\ &= S_0\left(te^{-\boldsymbol{\beta}'\mathbf{x}}\right), \end{aligned}$$

where S_0 is a baseline survival function. Thus, the effect of the covariates $\mathbf{x}$ is to change the time scale. The probability of survival beyond time t, given $\mathbf{x}$, is the same as the baseline probability of survival beyond time $te^{-\boldsymbol{\beta}'\mathbf{x}}$. If $\boldsymbol{\beta}'\mathbf{x} < 0$ this is a longer time than t, and therefore, the survival probability is smaller and the effect of $\mathbf{x}$ is to bring the event forward to shorten lifetimes; hence, the name accelerated failure time (AFT) model. On the other hand, if $\boldsymbol{\beta}'\mathbf{x} > 0$, the effect is to tend to lengthen lifetimes (a deceleration of the time scale).

The last expression above suggests an extended definition of an AFT model in a more general form as $S(t|\mathbf{x}) = S_0(tg(\mathbf{x}))$ for non-negative g, as in section 1.13. This form no longer corresponds to the familiar model [1.4].

Rewriting [1.5]

$$\begin{aligned} S(t|\mathbf{x}) &= P\left(\epsilon > \left(\ln t - \mu - \boldsymbol{\beta}'\mathbf{x}\right)/\sigma\right) \\ &= S_\epsilon\left(\left(\ln t - \mu - \boldsymbol{\beta}'\mathbf{x}\right)/\sigma\right), \end{aligned}$$

where S_ϵ denotes the survival function of an error term ϵ. It follows that

$$f(t|\mathbf{x}) = \frac{1}{\sigma t} f_\epsilon \left(\left(\ln t - \mu - \boldsymbol{\beta}' \mathbf{x} \right) / \sigma \right)$$

and hence the likelihood [1.3] becomes

$$L = \prod_{i=1}^{n} (\sigma t_i)^{-\delta_i} [f_\epsilon(\epsilon_i)]^{\delta_i} [S_\epsilon(\epsilon_i)]^{1-\delta_i}$$

where $\epsilon_i = \left(\ln t_i - \mu - \boldsymbol{\beta}' \mathbf{x}_i \right) / \sigma$.

The AFT formulation has great appeal in the field of reliability, that is, in the engineering and technological applications of lifetime data analysis. In biostatistics, lifetime data are generally obtained from observational studies and only rarely from experimental studies that involve the manipulation of conditions to which units are exposed, with randomized trials providing one notable exception. On the other hand, in reliability, where inanimate objects can be treated in a way that is not possible with human and animal subjects, there is a strong tradition of experimental work. This often involves operating the experimental units under conditions more extreme than will be encountered in normal usage, such as higher temperatures. The purpose is essentially to cause failures to happen quicker than they would be expected to under normal operating conditions. This means that data on a substantial number of failures can be acquired within a rather short space of time. Compare, for example, the lifetimes between the different stress levels in Table 1.1. A more extreme example can be found in Schmee and Hahn's early article about regression with censored data, in which there were no failures at all at the lowest test temperature [SCH 79]. Obtaining many failures at the standard level would either require a study of very long duration - which conflicts with the need to establish results before a product is put on the market or brought into service - or a study including an impractically large number of units.

The design of experiments is, of course, a major field of statistics with a vast literature. Although the general principles of experimental design certainly apply to reliability experimentation, many of the details cannot be carried over easily because of the presence of censoring in reliability data. For extensive material on the design of experiments in reliability, see the books by Meeker and Escobar [MEE 98a] and Nelson [NEL 90].

1.7. Checking the accelerated failure time assumption

A graphical test for the suitability of the accelerated failure time assumption can be derived as follows. The model supposes that

$$S(t|\mathbf{x}) = S_0(tg(\mathbf{x})),$$

where $g(\mathbf{x}) = e^{-\boldsymbol{\beta}'\mathbf{x}}$ in the basic theory and S_0 is the baseline survivor function,

$$\begin{aligned} &= P(T_0 \geq tg(\mathbf{x})) \\ &= P(\ln T_0 \geq \ln t + \ln g(\mathbf{x})) \\ &= S_0^*(y + \ln g(\mathbf{x})), \end{aligned}$$

where $y = \ln t$ and S_0^* is the survivor function of the random variable $Y = \ln T$. This result implies that a plot of $S(t|\mathbf{x})$ against $\ln t$ for particular $\mathbf{x}$ should be a horizontal displacement of S_0^* against $\ln t$. Consequently, the AFT assumption is verified if all the curves $S(t|\mathbf{x})$ for different $\mathbf{x}$ differ from each other only in horizontal displacement when plotted against $\ln t$. To construct these curves, it is necessary to have estimates $\hat{S}(t|\mathbf{x})$, usually Kaplan-Meier. As in the case of the similar graphical test for the PH assumption (see section 1.5), the method is feasible only if there are sufficient data for good estimation of S for each $\mathbf{x}$, or if the $\mathbf{x}$ can be grouped suitably so that each group contains enough observations.

EXAMPLE 1.2.– *Table 1.2 provides Nelson's data on the time to breakdown of an insulating fluid subjected to different voltages [NEL 72]. For simplicity of illustration, three further voltages with a small number of observations are excluded.*

Figure 1.4 shows the plot for checking the AFT assumption for the four groups of observations corresponding to the four voltages. We see that, instead of all the four lines being parallel, the line for the 32 kV level cuts across the others. Hence, the AFT assumption appears to be violated.

The functional form of $g(\mathbf{x})$ could also be investigated graphically. Although each unit's lifetime T follows a different distribution, depending on the value of $\mathbf{x}$, the model requires that the quantities $W = Tg(\mathbf{x})$ all have the

same survival function, S_0. Hence, the quantities

$$\ln W = \ln T + \ln g(\mathbf{x})$$

are identically distributed, and consequently, the terms $\ln W$ in the equation

$$\ln T = -\ln g(\mathbf{x}) + \ln W$$

behave like a residual or error term. This means that plotting $\ln t$ against a covariate could indicate the correct functional form of g. For example, if $g(\mathbf{x}) = e^{-\boldsymbol{\beta}'\mathbf{x}}$, then $\ln t$ against x_1 would be a straight line. On the other hand, if x_1 should be replaced by x_1^2, then $\ln t$ against x_1^2 would be a straight line.

Voltage	Failure times					
30kV :	7.74	17.05	20.46	21.02	22.66	43.40
	47.30	139.07	144.12	175.88	194.90	
32kV :	0.27	0.40	0.69	0.79	2.75	3.91
	9.88	13.95	15.93	27.80	53.24	82.85
	89.29	100.58	215.10			
34kV :	0.19	0.78	0.96	1.31	2.78	3.16
	4.15	4.67	4.85	6.50	7.35	8.01
	8.27	12.06	31.75	32.52	33.91	36.71
	72.89					
36kV :	0.35	0.59	0.96	0.99	1.69	1.97
	2.07	2.58	2.71	2.90	3.67	3.99
	5.35	13.77	25.50			

Table 1.2. *Part of Nelson's data on the time to breakdown of an insulating fluid subjected to different voltages [NEL 72]*

Actually, in the experiments in which AFT models are widely used, it is often the case that theory or experience shows the correct functional form for $g(\mathbf{x})$. For example, when the *accelerating factor* (covariate) is temperature, it is common to use the inverse of absolute temperature

$$1/(T + 273.16)$$

or the *Arrhenius transformation*

$$11604.83/(T + 273.16)$$

where the numerator is *Boltzmann's* constant. When the accelerating factor is a load or stress V, then an inverse power relationship $V^{-\alpha}$ is often assumed.

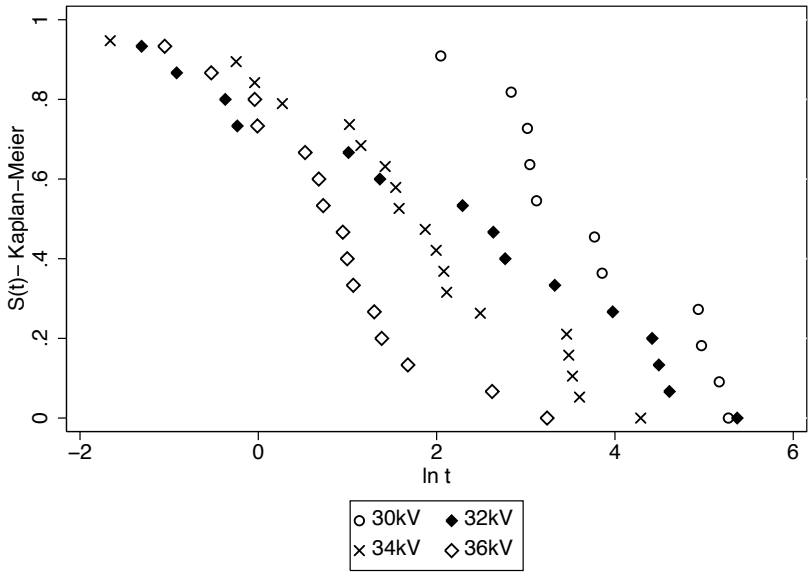

Figure 1.4. *Plot for checking the accelerated failure time assumption in Nelson's data*

1.8. Proportional odds models

Another well-known regression model for lifetime data is the proportional odds (PO) model (see Bennett [BEN 83a, BEN 83b]), which is based on the *odds* of the occurrence of the event by time t:

$$\theta(t) = \frac{F(t)}{1 - F(t)} = \frac{1 - S(t)}{S(t)}.$$

The PO model specifies that

$$\theta(t|\mathbf{x}) = \theta_0(t)g(\mathbf{x}),$$

where $\theta_0(t)$ is the baseline odds and $g(\mathbf{x})$ is a suitable non-negative function as before. The usual choice $g(\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}}$ gives the model

$$\ln \theta(t|\mathbf{x}) = \ln \theta_0(t) + \boldsymbol{\beta}'\mathbf{x},$$

which is a linear model for log odds, in other words, *logistic regression*. This is the most commonly used regression model for binary data. Despite this

appealing link, PO models have been used infrequently compared to others since they were introduced by Bennett [BEN 83a, BEN 83b]. With the exception of the regression model based on the log-logistic distribution, which is both a PO model and an AFT model (see below), they are mathematically and computationally more difficult to handle. Thus, they have not been included in some of the computing packages, and therefore, we will not be discussing them in detail here. Detailed discussions of PO models can be found in Collett [COL 14] and Dauxois and Kirmani [DAU 03], for example. The latter presents a graphical procedure for examining the hypothesis of PO between groups of respondents and a formal test for the case of two groups.

In order to see an important difference between PO and PH models, consider the hazard function of a PO model. Solving the previous equations for $S(t|\mathbf{x})$ gives

$$S(t|\mathbf{x}) = \{1 + \theta_0(t)g(\mathbf{x})\}^{-1}$$

from which the ratio of the hazard functions for two units with different covariate values is

$$\frac{h\left(t|\mathbf{x}_1\right)}{h\left(t|\mathbf{x}_2\right)} = \frac{g\left(\mathbf{x}_1\right)}{1 + \theta_0(t)g(\mathbf{x}_1)} \Big/ \frac{g\left(\mathbf{x}_2\right)}{1 + \theta_0(t)g(\mathbf{x}_2)}$$

$$\rightarrow 1 \quad (t \rightarrow \infty)$$

because $\theta_0(t) \rightarrow \infty$ as $t \rightarrow \infty$ (since $S_0(t)$ tends to zero), whereas $g(.)$ does not change with t. This fact - that the hazards for two different units tend to equalize over time under the PO model - stands in major contrast to the PH property that two units' hazards remain in the same ratio forever. On the one hand, the PO model says that initial differences disappear; on the other hand, the PH model says that they never change. Depending on the context, either postulate might be more appropriate. A badly made unit remains badly made; therefore, the implication of the PO model seems unreasonable in such situations. However, some treatments that a patient receives may only have a temporary effect that wears off with time, and in a case like that, it is the PH property - that initial differences in the covariates continue to have the same effect for ever and ever - that seems unrealistic.

All distributions of the Marshall-Olkin extended form [MAR 97] have the PO property (see also [ECO 07, SAN 08, CAR 10]). The Marshall-Olkin formula is

$$S^*(t|\alpha) = \frac{\alpha S(t)}{1 - \bar{\alpha} S(t)}, \qquad [1.6]$$

where S is the survival function of the original distribution and S^* is the survival function of the new, extended distribution; $\alpha > 0$ is a constant, and $\bar{\alpha} = 1 - \alpha$. Applying the formula again to [1.6] just results in a different value of α; therefore, the family of distributions is closed. However, the PO model states that

$$\frac{1 - S\left(t|\mathbf{x}\right)}{S\left(t|\mathbf{x}\right)} = g(\mathbf{x}) \frac{1 - S_0\left(t\right)}{S_0\left(t\right)},$$

which gives

$$S\left(t|\mathbf{x}\right) = \frac{\frac{1}{g(\mathbf{x})} S_0\left(t\right)}{1 - \left(1 - \frac{1}{g(\mathbf{x})}\right) S_0(t)},$$

and this is just the Marshall-Olkin form with $\alpha = 1/g\left(\mathbf{x}\right)$. Therefore, all Marshall-Olkin extended distributions are PO distributions. For example, the Marshall-Olkin extended Weibull distribution (see Ghitany *et al.* [GHI 05]; Caroni [CAR 10]) is a PO distribution, although the Weibull distribution itself is not a PO distribution.

Zucker and Yang [ZUC 06] note that the PO and PH models are both special cases of the general form

$$h(S(t|\mathbf{x})) = h(S_0(t)) \exp\left(\boldsymbol{\beta}'\mathbf{x}\right),$$

where h is a suitable monotonically decreasing function from $[0, 1]$ to $[0, \infty]$. For the PH model, $h(s) = -\ln(s)$, and for the PO model, $h(s) = (1 - s)/s$. These are both included within the Box-Cox family

$$h(s, \rho) = \rho^{-1}\left(s^{-\rho} - 1\right),$$

with the PH model arising when $\rho \to 0$ and the PO model when $\rho = 1$. Zucker and Yang give references to previous appearances of this family in

the survival analysis literature; their contribution consists of two estimation methods, although they observe that the model essentially cannot be fitted if the event rate is low.

Further work on the PO model is reviewed by Chen *et al.* [CHE 12], who extend the model by incorporating external time-varying covariates (see section 4.2).

1.9. Proportional mean residual life models

The mean residual life (MRL) of a unit is defined conditionally on the unit's present age t and represents its expected lifetime beyond this point,

$$\mu(t) = E[T - t|T \geq t],$$

which, if it exists, is equal to

$$\mu(t) = \frac{\int_t^\infty S(u)du}{S(t)},$$

as in section 1.1. The proportional mean residual life (PMRL) model proposes the relationship

$$\mu(t|\mathbf{x}) = g(\mathbf{x})\mu_0(t),$$

where $\mu_0(t)$ is the baseline MRL function.

Oakes and Dasu [OAK 90, OAK 03] suggest that the MRL function provides a more natural basis for modeling lifetime data than the hazard function because it summarizes the entire remaining life distribution and not just the immediate risk of failure. They claim that this is likely to be the more important information for the design of maintenance and repair strategies. The MRL is also used extensively in demography under the name life expectancy. On the other hand, Hougaard [HOU 99] states that, in contrast to these industrial and demographic applications, the evaluation of mean lifetime in biostatistics is considered unacceptable. He gives three reasons for this: one, the difficulty of estimating the right tail of the residual lifetime distribution (exacerbated by censoring), which - as acknowledged by Oakes and Dasu [OAK 03] - can have a strong influence on the mean; two, the

possible tendency for readers to think in terms of the normal distribution when they are presented with means, which could be very misleading; and three, the fact that for some types of events, there may be a proportion of the population that will never experience the event (see section 1.4 and elsewhere), which makes it impossible to calculate a mean. The last of these objections obviously does not apply in industrial and demographic applications, because units will always fail and individuals will always die eventually.

Further development of this model was taken up by Maguluri and Zhang [MAG 94] and subsequently by others; see Chen and Cheng [CHE 05]. The PMRL model does not seem to have entered general use at present. It is noticeable that an extensive review of statistical estimation of the remaining useful life of an item (Si *et al.* [SI 11]) does not mention the PMRL model at all.

1.10. Proportional reversed hazard rate models

Yet another "proportional" model is the proportional reversed hazard rate (PRHR) model (see Gupta and Gupta [GUP 07]), which is defined in a closely similar way to the PH model. The reversed hazard rate $r(t)$ is related to the conditional probability that an event occurred in the interval of length δt *before* time t, in contrast to the hazard that is related to the occurrence of the event in the interval of length δt *after* time t as in equation [1.1]. Thus,

$$r(t)\delta t = P(t - \delta t < T \leq t | T \leq t) = f(t)\delta t / F(t),$$

and therefore, $r(t) = (d/dt) \ln F(t)$. The PRHR model defines a multiplicative effect of covariates on the baseline function $r_0(t)$,

$$r(t|\mathbf{x}) = g(\mathbf{x}) r_0(t).$$

Equivalently, the model may also be defined by the relation $F(t|\mathbf{x}) = [F_0(t)]^{g(\mathbf{x})}$ similar to the definition $S(t|\mathbf{x}) = [S_0(t)]^{g(\mathbf{x})}$, which can be derived for the PH model.

An example of a family of distributions with the PRHR property is the exponentiated Weibull, which has baseline distribution function $F(t|\alpha) = [1 - \exp(-t^{\alpha})]^{\theta}$ [MUD 96].

A PRHR model was first suggested, although not studied, by Kalbfleisch and Lawless [KAL 89], who examined a problem in which occurrences were ascertained after the event, in which case the retrospective nature of the reversed hazard rate has a natural appeal.

1.11. The accelerated hazards model

The accelerated hazards (AccH) model incorporates ideas from both the PH and AFT models. As seen earlier, the effect of covariates in the PH model can be simply expressed by a multiplicative effect on the hazard function: in its usual form, the model is

$$h(t|\mathbf{x}) = h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}}.$$

In the AFT model, the effect of covariates is a shift of time scale, so that the survival function becomes

$$S(t|\mathbf{x}) = S_0\left(te^{\boldsymbol{\beta}'\mathbf{x}}\right).$$

Adapting these ideas, the AccH model takes the time scale shift and places it in the hazard function

$$h(t|\mathbf{x}) = h_0\left(te^{\boldsymbol{\beta}'\mathbf{x}}\right).$$

This model was introduced by Chen and Wang [CHE 00] for a two-group comparison and extended to the more general regression case by Chen [CHE 01a]. The motivation for the model was provided by a clinical trial, in which there appeared to be no immediate difference in hazard rates between the two treatment groups, whereas both the PH and AFT models imply that covariates have an effect even at time zero. This implication is somewhat problematic for randomized clinical trials, in which the groups should not differ at baseline and the treatment is unlikely to have an instant effect. As the trial progressed, an increasing difference between hazard rates was seen, which is again in conflict with the PH property. Furthermore, after some time, the two hazard functions crossed over, which is a feature that the PH model cannot reproduce, although the AFT model can; neither PH nor AFT allows survival functions to cross. The new AccH model, on the other hand, does

allow hazard functions and survival functions to cross over, depending on the form of the baseline hazard function. Conditions for crossovers in hazard functions in the various models are given by Zhang and Peng [ZHA 09].

In the AFT model, the hazard function is

$$h(t|\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}} h_0\left(te^{\boldsymbol{\beta}'\mathbf{x}}\right),$$

which suggests a generalized model encompassing all the three PH, AFT and AccH models:

$$h(t|\mathbf{x}) = h_0\left(te^{\boldsymbol{\beta}_1'\mathbf{x}}\right) e^{\boldsymbol{\beta}_2'\mathbf{x}}.$$

The PH model corresponds to $\boldsymbol{\beta}_1 = \mathbf{0}$, the AFT model to $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ and the AccH model to $\boldsymbol{\beta}_2 = \mathbf{0}$. This model and its semi-parametric estimation is studied by Chen and Jewell [CHE 01b], but in fact, the model had been introduced and analyzed several years earlier by Ciampi and Etezadi-Amoli [CIA 85, ETE 87] under the name extended hazard regression. Subsequently, Shyur *et al.* discussed its applicability in the field of reliability [SHY 99].

The two vectors of regression coefficients can be interpreted as measuring different impacts of the covariates on survival. While the appropriate component of $\boldsymbol{\beta}_1$ measures a specific covariate's contribution to the acceleration (or deceleration) factor, the corresponding component of $\boldsymbol{\beta}_2$ indicates its independent contribution to the relative hazards. In the example used for illustration of the method by Etezadi-Amol and Ciampi (survival of patients with ovarian cancer), two of the five covariates appeared to have a simple PH effect; one had an AFT effect, and the remaining two had both the effects. One of these, the patient's age, had opposite signs of its two coefficients, suggesting an increased hazard for older patients at any given age, but slower tumor growth in older patients than in younger ones. This structure, in which a covariate can affect lifetimes in two ways, has some similarity to the first hitting time regression model based on an underlying Wiener process, which is the main topic of this book and will be introduced in the next chapter. The issue of opposing effects indicated by the signs of the two coefficients associated with the same covariate will be mentioned in that context too.

Although the general model may give a better description of the data than any of the three separate models included in it, Chen and Jewell [CHE 01b]

suggest that its main value may lie in bringing out the differences between these three models. Earlier, Chen [CHE 01a] commented on the potential usefulness of fitting the general model as a guide to which of the separate models to fit. Of course, this will only become true in practice when the readily available software makes it easy to fit the more general model, which is not the case at the moment.

1.12. The additive hazards model

It can often be supposed that an organism or machine comprises many components or parts that must all be operating; otherwise the organism dies or the machine fails (a series system). If a particular component j has hazard rate $h_j(t)$ at time t, then the overall hazard $h(t)$ at this time is $h(t) = \sum_j h_j(t)$. Consequently, it may often seem natural to represent a hazard function in additive form (see Elandt-Johnson [ELA 80]).

In the additive hazards model, the effect of the covariates on the baseline hazard function $h_0(t)$ is additive

$$h(t|\mathbf{x}) = h_0(t) + \boldsymbol{\beta}'\mathbf{x}$$

instead of multiplicative as in the PH model. This model was first suggested by Aalen [AAL 78] and further developed by Aalen [AAL 89] and others. Lin and Ying [LIN 94] proposed a semi-parametric estimation method with $h_0(t)$ unspecified, along the lines of the semi-parametric PH model. They take the regression parameters $\boldsymbol{\beta}$ as fixed, whereas Aalen's formulation allowed for time-varying coefficients $\boldsymbol{\beta}(t)$.

For an example of the application of the additive hazards model in a biostatistical context, with comparison of results between its different versions and also with the Cox model, see Xie *et al.* [XIE 13]. They point out that the multiplicative and additive hazards model address different questions. While the PH model provides estimates of relative hazard, the additive hazards model estimates absolute differences in hazard. Therefore, in the latter case and assuming that the event rate is low, the differences between cumulative hazards give an approximation to differences in cumulative incidence. Thus, an estimate of attributable risk is obtained. This information could be important for the purpose of public health planning and intervention.

The application of the additive hazards model in the reliability context, especially concerning repairable systems, is considered by Pijnenburg [PIJ 91].

1.13. PH, AFT and PO distributions

Based on our choice of how to introduce covariates (PH, AFT or PO - we will not be considering other models any further), we will find that some lifetime distributions are much easier to use than others. Let us begin with the Weibull distribution in a PH model. Since the baseline hazard is

$$h_0(t) = \frac{\eta t^{\eta-1}}{\alpha^\eta},\ t > 0,$$

we have

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x}) = \eta t^{\eta-1}\alpha^{-\eta}e^{\boldsymbol{\beta}'\mathbf{x}} = \frac{\eta t^{\eta-1}}{\left(\alpha e^{-\boldsymbol{\beta}'\mathbf{x}/\eta}\right)^\eta},$$

which corresponds to another Weibull distribution. The shape parameter η is the same as that of the baseline hazard, but the scale parameter is different: α has become $\alpha e^{-\boldsymbol{\beta}'\mathbf{x}/\eta}$. Thus, PH implies that the lifetime distribution for any unit is always Weibull when the baseline hazard is Weibull, and in this sense, the Weibull is a "PH distribution", and is therefore a natural choice to use in the context of a PH model.

The AFT model in section 1.6 was formulated as a model for the logarithm of T. If $T \sim Weibull(\alpha, \eta)$, then $\ln T \sim Gumbel$ with

$$S(t) = \exp\left(-e^{(\ln t - \mu)/\sigma}\right),$$

where $\mu = \ln \alpha$, $\sigma = \eta^{-1}$. This implies that, if $\epsilon \sim Gumbel(0, 1)$ in the AFT model

$$\ln T_{\mathbf{x}} = \mu + \boldsymbol{\beta}'\mathbf{x} + \sigma\epsilon,$$

then $T_{\mathbf{x}} \sim Weibull$ for any $\mathbf{x}$. Therefore, the Weibull is also an "AFT distribution", in addition to being a PH distribution; in fact, no other distribution has this dual property (see below). The Weibull distribution is not a PO distribution. As noted by Hougaard, the PH model modifies the

Gompertz distribution with parameters μ and ϕ into another Gompertz distribution with the same ϕ but different μ. However, the AFT model changes both the parameters [HOU 99]. Thus, the PH and AFT models are not equivalent for the Gompertz distribution.

Other important AFT distributions, which are not also PH, include the log-normal and log-logistic distributions. T has a log-logistic distribution when $Y = \ln T$ follows a logistic distribution. The survival function of the log-logistic distribution is

$$S(t) = (1 + e^{\kappa} t^{\gamma})^{-1}, \ \gamma > 0, \ \kappa > 0,$$

and the hazard function is

$$h(t) = \frac{\gamma t^{\gamma-1} e^{\kappa}}{1 + e^{\kappa} t^{\gamma}}.$$

We take this as the baseline hazard function $h_0(t)$ in the following. Now for any AFT model, we have

$$\begin{aligned} h(t|\mathbf{x}) = \frac{f(t|\mathbf{x})}{S(t|\mathbf{x})} &= \frac{-d \ln S(t|\mathbf{x})}{dt} \\ &= \frac{-d \ln S_0(te^{-\boldsymbol{\beta}'\mathbf{x}})}{dt} \\ &= e^{-\boldsymbol{\beta}'\mathbf{x}} h_0(te^{-\boldsymbol{\beta}'\mathbf{x}}) ; \end{aligned}$$

therefore, in the case of the log-logistic distribution, we obtain

$$\begin{aligned} h(t|\mathbf{x}) &= e^{-\boldsymbol{\beta}'\mathbf{x}} h_0(te^{-\boldsymbol{\beta}'\mathbf{x}}) \\ &= \frac{e^{-\boldsymbol{\beta}'\mathbf{x}} \gamma (te^{-\boldsymbol{\beta}'\mathbf{x}})^{\gamma-1} e^{\kappa}}{1 + e^{\kappa} (te^{-\boldsymbol{\beta}'\mathbf{x}})^{\gamma}} \\ &= \frac{\gamma t^{\gamma-1} e^{\kappa - \gamma\boldsymbol{\beta}'\mathbf{x}}}{1 + (e^{\kappa - \gamma\boldsymbol{\beta}'\mathbf{x}}) t^{\gamma}}, \end{aligned}$$

which is still the hazard function of a log-logistic distribution, although the parameter κ has changed to $\kappa - \gamma\boldsymbol{\beta}'\mathbf{x}$, with the shape parameter γ remaining unchanged. Hence, the log-logistic is an AFT distribution.

In the case of the PO model, a PO distribution will be one for which the distribution of

$$\theta(t|\mathbf{x}) = g(\mathbf{x})\theta_0(t)$$

has the same functional form as the baseline odds function $\theta_0(t)$. The simplest example is the log-logistic distribution with survival function given above and odds $e^{\kappa}t^{\gamma}$. Then

$$\theta(t|\mathbf{x}) = g(\mathbf{x})e^{\kappa}t^{\gamma} = e^{\kappa+\ln g(\mathbf{x})}t^{\gamma},$$

which corresponds to a log-logistic distribution with parameter κ changed to $\kappa + \ln g(\mathbf{x})$ and shape γ unchanged. Thus, the log-logistic distribution is both PO and AFT; in fact, it is unique in this respect (see below). It is not, however, PH.

Proof of the unique status of the Weibull and log-logistic distributions in possessing dual properties (PH and AFT for the Weibull, PO and AFT for the log-logistic) is given by [LAW 86]. Define a general family of regression models by

$$\psi_1\{S_{\mathbf{x}}(t)\} = \psi_1\{S_0(t)\} + g_1(\mathbf{x}),\ t > 0,$$

where $S_{\mathbf{x}}(t)$ is a survival function for a unit with vector of covariates $\mathbf{x}$ and baseline $S_0(t)$, and $g_1(\mathbf{0}) = 0$. This family includes PH models, for which

$$\psi_1(u) = \ln(-\ln u) \qquad [1.7]$$

and also PO models, choosing

$$\psi_1(u) = \ln\{(1-u)/u\}. \qquad [1.8]$$

Furthermore, define a second family by

$$\psi_2\{Q_{\mathbf{x}}(p)\} = \psi_2\{Q_0(p)\} + g_2(\mathbf{x}),\ \ 0 < p < 1,$$

where $Q_{\mathbf{x}}(p)$ is the quantile function of T given $\mathbf{x}$, with baseline $Q_0(p)$, and $g_2(\mathbf{0}) = 0$. This family includes AFT models, for which $\psi_2(u) = \ln u$.

Lawless [LAW 86] shows that for given ψ_1 and ψ_2, the unique family satisfying both [1.7] and [1.8] simultaneously is

$$S_0(t) = \psi_1^{-1}\left\{a\psi_2(t) + b\right\}.$$

In particular, choosing $\psi_1(u) = \ln(-\ln u)$ (a PH model) and $\psi_2(u) = \ln u$ (an AFT model) gives

$$S_0(t) = \exp\left(-e^b t^a\right),$$

where $a \neq 0$ and b are constants, which is a Weibull model. Hence, the Weibull distribution, and no other, is both PH and AFT. The same ψ_2 and $\psi_1(u) = \ln\{(1-u)/u\}$ (a PO model) gives

$$S_0(t) = \left(1 + e^b t^a\right)^{-1},$$

which is a log-logistic model. Hence, the log-logistic distribution, and no other, is both AFT and PO.

1.14. Cox's semi-parametric PH regression model

We present in this section the widely used version of the PH regression model known as Cox's semi-parametric regression model. First, we recall from section 1.4 that the PH property requires that the hazard functions of two units, with covariate vectors $\mathbf{x}_1$ and $\mathbf{x}_2$, must be in constant ratio to each other over time. This is achieved if the hazard function takes the form

$$h(t|\mathbf{x}) = h_0(t)g(\mathbf{x}),$$

where $h_0(t)$ is a baseline hazard and $g(\mathbf{x})$ is a suitable non-negative function of $\mathbf{x}$. The baseline could be, for example, the hazard that applies to a unit with $\mathbf{x} = \mathbf{0}$ (although in many applications, this value cannot be realized). Cox [COX 72] proposed taking the form of $g(\mathbf{x})$ that we have already seen

$$g(\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}}$$

but - and this is the model's crucial feature - not specifying the functional form of h_0 at all. Only the part of the model that expresses dependence on the

covariates is expressed parametrically (hence the description of the model as semi-parametric) and in fact, only this part needs to be estimated.

Consider the set of units R_i that are at risk of failure at time $t_{(i)}$, where $t_{(1)} < t_{(2)} < \cdots < t_{(k)}$ are unique failure times recorded in a study of $n \geq k$ units. This risk set R_i excludes any units that have already failed or were censored earlier than $t_{(i)}$. From the definition of hazard rate, the probability that unit $j \in R_i$ with covariates $\mathbf{x}_j$ fails in the time interval $(t_{(i)}, t_{(i)} + \delta t)$ is

$$h_0(t_{(i)})\, e^{\boldsymbol{\beta}'\mathbf{x}_j} \delta t.$$

Hence, the conditional probability that it is unit j that fails at time $t_{(i)}$, given that we know that one unit does fail at this time, is

$$\frac{h_0(t_{(i)})\, e^{\boldsymbol{\beta}'\mathbf{x}_j} \delta t}{\sum_{\ell \in R_i} h_0(t_{(i)})\, e^{\boldsymbol{\beta}'\mathbf{x}_\ell} \delta t} = \frac{e^{\boldsymbol{\beta}'\mathbf{x}_j}}{\sum_{\ell \in R_i} e^{\boldsymbol{\beta}'\mathbf{x}_\ell}},$$

in which h_0 ultimately does not appear at all. Taking the product of these conditional probabilities over all the failure times $\left\{t_{(i)} : i = 1, \ldots, k\right\}$ gives the partial likelihood

$$\prod_{i=1}^{k} \left\{ \frac{e^{\boldsymbol{\beta}'\mathbf{x}_i}}{\sum_{\ell \in R_i} e^{\boldsymbol{\beta}'\mathbf{x}_\ell}} \right\}.$$

This is treated as a standard likelihood, which is maximized over $\boldsymbol{\beta}$ in order to obtain estimates of the effects of the covariates on the hazard rate, without needing to estimate the hazard rate itself (although a non-parametric estimate of $h_0(t)$ can be obtained if desired). The theory, which can be extended to allow for tied failure times, is presented in detail in Therneau and Grambsch [THE 00] and elsewhere. The justification for treating the partial likelihood as a standard likelihood was presented intuitively in Cox's initial paper on this regression method [COX 72] and was subsequently placed on a firmer foundation by Cox [COX 75] and others.

The remarkable simplicity that is made possible by using the partial likelihood method is one of the features behind the widespread adoption of Cox's semi-parametric PH regression model (often referred to simply as Cox

regression), at least in the biomedical sciences where it has become virtually the default approach toward analyzing lifetime data. It also has an advantage of apparently easy interpretability. From $h(t|\mathbf{x}) = h_0(t)e^{\boldsymbol{\beta}'\mathbf{x}}$, the effect of a unit increase in the value of the j^{th} covariate x_j (the familiar way of interpreting the size of a regression coefficient) is to multiply the hazard by $e^{\hat{\beta}_j}$. This is very similar to the interpretation of the coefficients of logistic regression, which is another almost-default method of analysis of biomedical data.

In sharp contrast to the remarkably high level of use of Cox's PH model in biostatistics, it is not often seen in the field of reliability. A few illustrative applications can be found in, for example, Bendell *et al.* [BEN 86, BEN 91], Dale [DAL 85], Elsayed and Chan [ELS 90], Krivtsov *et al.* [KRI 02] and Madeira *et al.* [MAD 13]. Some reviews warn against the unthinking transfer of biostatistical methods to reliability analysis (see Bendell *et al.* [BEN 91] and Kumar and Klefsjö [KUM 94]), particularly because of the preponderance of repairable systems in the latter field, and hence the need to find realistic models for recurrent events, that is, repeated events occurring in the same unit. The preference in the field of reliability for parametric AFT models is undoubtedly related to the importance of analysis of experiments in which one or more factors (such as the operating temperature) have been manipulated specifically in order to bring forward the times of failure. The concept of the AFT formulation is a natural fit to the nature of such data. Elsayed and Chan remark in their presentation of Cox modeling in a reliability problem in electronics that "The proportional hazards relation has not been used much for modeling the hazard (failure) rate of electronic devices because there appears to be no physical basis for hazard-rate scaling" [ELS 90, p.331], in contrast to the familiar concept of time scaling in the AFT model. However, they go on to interpret the PH relation in terms of one of the reliability models taken from the literature.

1.15. PH versus AFT

The dominance of PH models in biostatistics and AFT in reliability has led to much consideration of the differences between the two and recommendations for which model is preferable in the given circumstances. The rapid adoption of Cox's PH model in biostatistics to the relative neglect of the AFT model, and the converse situation in reliability, gave rise to a

number of articles in the literature, whose main purpose is to encourage readers who are used to one formulation not to ignore the other. The papers in the reliability literature on the Cox model - for example, those cited in section 1.14 - tend to be of this type. Conversely, papers appear in the biostatistical literature promoting the virtues of alternatives to the Cox model, notably AFT models. One example where an AFT model makes more sense than a PH model is in a trial of a drug for influenza (see Kay and Kinnersley [KAY 02] and Patel *et al.* [PAT 06]). The effect of the drug is to shorten the duration of the illness in treated patients compared to untreated patients, but almost all the patients would be expected to recover from their symptoms during the course of the trial, irrespective of the treatment. Therefore, a PH model - with constant hazards throughout - is an illogical structure, whereas the AFT framework describes the effect (acceleration of recovery) more appropriately. In another example, Argyropoulos *et al.* [ARG 09] discuss the survival of hemodialysis patients and argue against the PH model because it evaluates the effect of a covariate at particular time points rather than considering its history. If a covariate acts through "accumulated damage", then AFT rather than PH would be appropriate.

In situations where prior considerations lead to preference for neither model over the other, model-checking methods should be applied as in all the applications of statistical analyses. A wide range of diagnostics - some based on analogies with linear regression, others founded on the particular properties of the survival analysis context - exists for the Cox PH model (see, for example, Therneau and Grambsch [THE 00] and Caroni [CAR 04]). Nardi and Schemper [NAR 03] illustrate the use of residuals in examining Cox and parametric AFT models. However, most published applications do not present an investigation of the validity of the assumptions underlying Cox's model. For example, Altman *et al.* [ALT 95] found that the assumptions were checked in only 2/43 (5%) of the papers that they examined, and none assessed goodness of fit. Similarly, Ford *et al.* [FOR 95, p.745] stated that "model validation is an important prerequisite to the interpretation of parameter estimates. In this respect, the almost de facto assumption of the Cox model in the analysis of survival data is a cause for concern".

If the basic PH assumption appears to be violated, a number of solutions are available while remaining within a general PH framework (see Therneau and Grambsch [THE 00]). One popular method is stratification. If the PH assumption does not hold for a particular covariate, then the model can be

fitted (without any additional technical difficulty) stratifying by values of that covariate. The model in the m^{th} stratum becomes

$$h_m(t|\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}} h_{m0}(t),$$

where $h_{m0}(t)$ is the baseline hazard function in stratum m, thus allowing different baseline hazards in each stratum, but assuming common effects of the other covariates in every stratum. With obvious extension of the notation of section 1.14, the log partial likelihood for events in the m^{th} stratum is

$$\ell_m(\boldsymbol{\beta}) = \sum_{i=1}^{k_m} \boldsymbol{\beta}'\mathbf{x}_{mi} - \sum_{i=1}^{k_m} \ln \left\{ \sum_{\ell \in R_{mi}} e^{\boldsymbol{\beta}'\mathbf{x}_{m\ell}} \right\},$$

and the overall log partial likelihood for all s strata is

$$\ell(\boldsymbol{\beta}) = \sum_{m=1}^{s} \ell_m(\boldsymbol{\beta}) = \sum_{m=1}^{s} \sum_{i=1}^{k_m} \boldsymbol{\beta}'\mathbf{x}_{mi} - \sum_{m=1}^{s} \sum_{i=1}^{k_m} \ln \left\{ \sum_{\ell \in R_{mi}} e^{\boldsymbol{\beta}'\mathbf{x}_{m\ell}} \right\}.$$

This can be maximized with no more difficulty than for the single-stratum model. A drawback of this formulation is that it does not yield a direct estimate of the effect of the stratification factor on lifetimes.

Important general results on the properties of PH and AFT models have been obtained, particularly with regard to the robustness of estimates against misspecification of the model. It is known that omitting a relevant covariate from a model that is truly PH induces a model that is no longer PH. The estimated regression coefficients, their standard errors and the ratios of coefficients to standard errors all tend to be smaller than the corresponding quantities obtained by estimation under the true model. Unlike linear regression, this is true even if the omitted covariates are orthogonal to those that are included. References relevant to this topic include Gail *et al.* [GAI 84], Struthers and Kalbfleisch [STR 86], Schumacher *et al.* [SCH 87], Schmoor and Schumacher [SCH 97] and Gerds and Schumacher [GER 01]. An extensive study in the context of fully parametric PH and AFT models was reported by Hutton and Monaghan [HUT 02]. A key conclusion is that estimates from a misspecified PH model can be seriously biased, and the apparent shape of the hazard function can be misleading. Furthermore, the size of Wald tests is underestimated. On the other hand, AFT models are more

robust to misspecification, a property which is attributed to their log-linear form. Hougaard [HOU 99, p. 22] says "the accelerated failure parameter η is robust toward neglected covariates, whereas the proportional hazards parameter β is not .... It is a major drawback of the PH model...". As far as estimated survival is concerned, bias in the lower and upper percentiles can be substantial from a misspecified model, less so for the median.

There are however results that show that some similarity of results between PH and AFT regressions can be expected. Fitting a PH model when the true model is AFT leaves the relative importance of covariates unchanged to first order under conditions (see Solomon [SOL 84] and Struthers and Kalbfleisch [STR 86]). Under the AFT model, the hazard function given covariates $\mathbf{x}$ takes the form

$$h_{AFT}(t|\mathbf{x}) = e^{\boldsymbol{\beta}'\mathbf{x}} h_0(e^{\boldsymbol{\beta}'\mathbf{x}}\, t)$$

for baseline hazard h_0. Now, following Kwong and Hutton, take a Taylor series expansion in $e^{\boldsymbol{\beta}'\mathbf{x}}$ about $\boldsymbol{\beta} = \mathbf{0}$ to first order:

$$\begin{aligned} h_{AFT}(t|\mathbf{x}) &\approx e^{\boldsymbol{\beta}'\mathbf{x}} h_0 \left\{(1 + \boldsymbol{\beta}'\mathbf{x})\, t\right\} \\ &= e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t + t\,\boldsymbol{\beta}'\mathbf{x}) \\ &\approx e^{\boldsymbol{\beta}'\mathbf{x}} \left\{h_0(t) + \boldsymbol{\beta}'\mathbf{x}\, t\, h_0'(t)\right\}, \end{aligned}$$

where the second approximation is obtained from a first order Taylor series expansion about t. Consequently, we have

$$\begin{aligned} h_{AFT}(t|\mathbf{x}) &\approx e^{\boldsymbol{\beta}'\mathbf{x}} h_0(t) + e^{\boldsymbol{\beta}'\mathbf{x}} \boldsymbol{\beta}'\mathbf{x}\, t\, h_0'(t) \\ &= h_{PH}(t|\mathbf{x}) + e^{\boldsymbol{\beta}'\mathbf{x}} \boldsymbol{\beta}'\mathbf{x}\, t\, h_0'(t), \end{aligned}$$

where $h_{PH}(t|\mathbf{x})$ is the hazard function that holds for a unit with covariates $\mathbf{x}$ under the PH assumption. This expression implies that hazards derived under AFT and PH will not differ greatly so long as (a) covariate effects $\boldsymbol{\beta}$ are small, and (b) the hazard function h_0 varies slowly so that $t\, h_0'(t)$ is small [KWO 03].

As the Cox model is semi-parametric, it might be expected that it would yield less efficient parameter estimates than an appropriate fully parametric model. This has been investigated by Oakes [OAK 77] and Efron [EFR 77],

among others. Nardi and Schemper [NAR 03] summarize the conditions for this to be true as follows:

a) if parameter values are not close to zero;

b) if follow-up depends on the values of the covariates;

c) if the covariates show a strong time trend.

Furthermore, as expected, the loss in precision is greater for small samples.

It is often remarked that the PH model is not based on any persuasive rationale, and that its popularity is to a large extent due to its apparent simplicity. For example, according to Oakes [OAK 13, p.453]: "Cox (1972) emphasized that there is usually no simple physical or biological motivation for the assumption of PH. The appeal of this model arises rather from the intuitive interpretation of the hazard ratio in terms of conditional risks, and from the simplicity and numerical stability of the algorithms used to fit the model. However in particular situations other approaches may be preferable". Cox himself, in his original presentation, claimed that his model was "intended as a representation of the behaviour of failure-time that is convenient, flexible and yet entirely empirical" [COX 72, p.200], and concluded the paper with the claim that the model "as a basis for rather empirical data reduction ... seems flexible and satisfactory" [COX 72, p.201]. However, Freedman's objection seems entirely reasonable: "if the model is wrong, why are the parameter estimates a good summary of the data?" [FRE 08, p.117].

1.16. Residuals

One basic way of checking the suitability of a statistical model is to inspect the *residuals* after the model has been fitted. Examination of the residuals can show if the model's assumptions are satisfied and how well the model fits the data, not just overall but for each point individually. In the familiar case of linear regression, the residuals are

$$\begin{aligned}\hat{\epsilon}_i &= y_i - \hat{y}_i \\ &= y_i - \hat{\boldsymbol{\beta}}' \mathbf{x}_i,\end{aligned}$$

in other words, the difference between the observed value y_i and the predicted $\hat{y}_i$. These residuals can be examined in various ways, often graphically. For example, their distribution can be investigated, or possible outliers can be identified.

However, most statistical models do not give rise to residuals of this familiar form. Consequently, Cox and Snell [COX 68] proposed *generalized residuals*. Suppose that the random variable Z_i for unit i has a distribution that depends on covariates $\mathbf{x}_i$ and parameters $\boldsymbol{\theta}$. If there exist functions

$$w_i(Z_i|\mathbf{x}_i, \boldsymbol{\theta}),$$

independently and identically distributed, following a distribution that does not depend on unknown parameters, then

$$\hat{\epsilon}_i = w_i(Z_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})$$

can fulfill the role of residuals.

Suitable functions for this task can be found using the general result that, if Y is a random variable with distribution function $F(.)$, then the random variable

$$V = F(Y) \sim U(0, 1).$$

It follows that the random variable $U = -\ln(1 - F(Y))$ has pdf

$$g(u) = e^{-u},\ u > 0,$$

which is the exponential distribution with parameter 1. Since

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

the residuals in a lifetime data model could be the values

$$-\ln \hat{S}(t_i) = \hat{H}(t_i) = \hat{\epsilon}_i$$

where $\hat{S}(.)$ and $\hat{H}(.)$ are estimates of the survival function and the baseline cumulative hazard function, respectively. In the case of a parametric model such as Weibull regression, $\hat{H}(.)$ follows simply from the estimation of the

parameters of the model. These Cox-Snell residuals are regarded as very useful in parametric models.

When an observation is right-censored, then

$$1 - \ln \hat{S}(t_i)$$

is usually used as that observation's residual. The reason for this is as follows. Since the right-censored observation t_i is less than the unknown true value, $-\ln \hat{S}(t_i)$ is likewise less than it should be. The difference between $-\ln \hat{S}(t_i)$ and its true value is similar to a residual lifetime (see section 1.9) and is a random variable that follows the exponential distribution with parameter 1. Consequently, its expected value is 1, and we add on this value in order to estimate the residual that would have been obtained if the observation had not been censored. Making this adjustment for the censored observations, the set of residuals can be examined in a probability plot against the exponential distribution with parameter 1.

EXAMPLE 1.3.– *To illustrate the use of residuals, we fit a Weibull regression model to the set of experimental data given in Table 1.3. These are the failure times of glass capacitors in a* 4×2 *factorial experiment (four levels of voltage, two temperatures) with eight replications. Note that "Type II" censoring was applied: the experiment at each temperature/voltage combination ran until four of the eight units had failed. The remaining four were right censored at that time. The original analysis fitted exponential distributions with a guarantee parameter. (A guarantee parameter is in effect a minimum possible lifetime. The exponential distribution modified in this way has pdf* $f(t) = \lambda \exp(-\lambda(t - \tau))$ *with* $t \geq \tau$.)

We fit the Weibull regression model to the data including the right-censored observations, with covariates voltage V and temperature T. We treat V as a quantitative measurement without transformation. Since T takes only two values, it makes no real difference whether we treat it as quantitative or categorical. Fitting is by maximum likelihood, and likelihood ratio tests confirm that both V and T should be included in the model.

Figure 1.5 shows the probability plot, against the exponential (1) distribution, of the Cox-Snell residuals (corrected for right censoring where necessary) from the analysis that includes both covariates V and T. The plot is reasonably close to the expected straight line.

	Applied voltage			
Temperature	200	250	300	350
170^oC	439	572	315	258
	904	690	315	258
	1092	904	439	347
	1105	1090	628	588
	1105^*	1090^*	628^*	588^*
	1105^*	1090^*	628^*	588^*
	1105^*	1090^*	628^*	588^*
	1105^*	1090^*	628^*	588^*
180^oC	959	216	241	241
	1065	315	315	241
	1065	455	332	435
	1087	473	380	455
	1087^*	473^*	380^*	455^*
	1087^*	473^*	380^*	455^*
	1087^*	473^*	380^*	455^*
	1087^*	473^*	380^*	455^*

Table 1.3. *Zelen's data from life tests of capacitors (lifetimes in hours) [ZEL 59]. Asterisks denote right-censored observations*

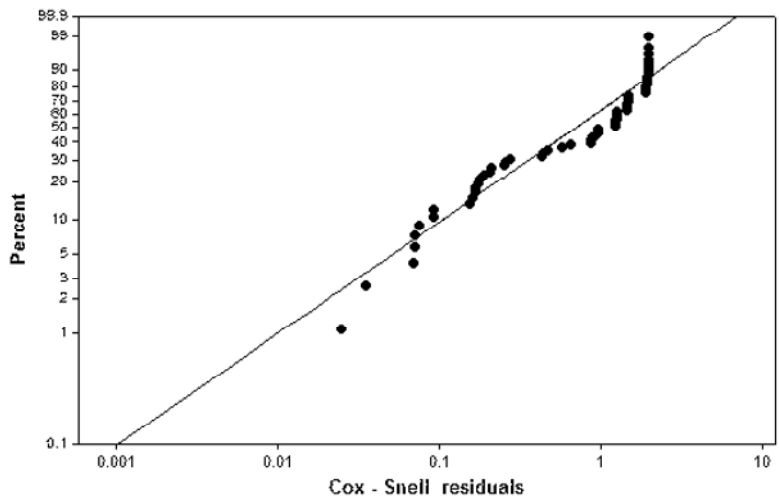

Figure 1.5. *Probability plot of corrected Cox-Snell residuals from Weibull regression model fitted to the data of Table 1.3*

1.17. Cured fraction or long-term survivors

As already noted in the discussion of the Gompertz distribution in section 1.4, it is a characteristic feature of lifetime data - present in most studies - that the data are incomplete, in the sense that some of the units under study have not experienced the event. Therefore, they contribute right-censored lifetimes. Very often, this happens because it is not feasible to allow data collection to continue until all units have failed, because that could take years in contexts such as studies of highly reliable machines or human survival. However, it is implied by the form of the basic models for lifetime data analysis that every unit is susceptible to failure. Therefore, the event would have been recorded for every unit if only the observation could have gone on long enough.

There are many contexts, however, in which it is possible that not every unit is in fact susceptible to failure. The obvious examples come from medical studies of the time from the end of treatment until relapse or death from the disease. If a complete cure is a possibility, then individuals who have been cured by the treatment are free of the disease and therefore will not relapse and will not die from the disease. Having been cured, they are no longer susceptible. The proportion of the population that is no longer susceptible after treatment is called the cured fraction. Other terminologies are immunes and - recognizing that, in practice, no lifetime is infinite - long-term survivors.

If there is a cured fraction, then it follows that the usual condition on the survival function $S(t) \to 0$ $(t \to \infty)$ does not hold: this was also remarked in the discussion of the Gompertz distribution. Sometimes, this can be seen in the appearance of the estimated survival function. Figure 1.6 shows an example from a study of the time to graduation of 15,541 undergraduate students in a Greek technical university (see Caroni [CAR 11b]). Starting from the minimum duration of studies of five years, the number of surviving students (i.e. those who have not graduated yet) falls steeply for about two years, but much more gradually thereafter. Rather than tending towards zero, it looks as if $S(t)$ tends towards a limit of the order of 10%. (In fact, the analysis gives an estimate of 12.4% with a standard error of 0.4.) This is a significant proportion of the student intake, and it should be represented somehow in the model that describes the data.

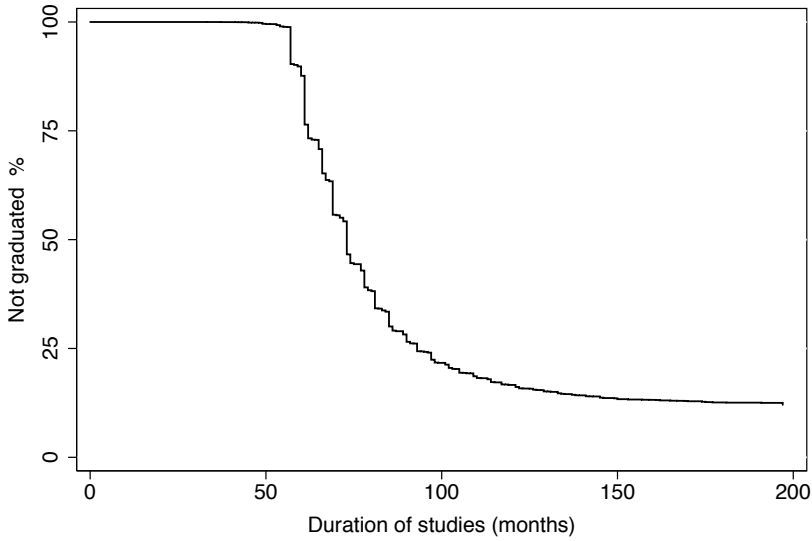

Figure 1.6. *Proportion of students who have not yet graduated, by time since commencement of studies (15541 students in a Greek technical university [CAR 11b])*

The credit for the first attempt to tackle the issue of the presence of a cured fraction is usually given to Boag [BOA 49] who proposed a mixture model with survival function:

$$S(t) = (1 - \pi) + \pi S_0(t),$$

where π is the proportion of the population that is susceptible and $1 - \pi$ the proportion that is non-susceptible. S_0 is the survival function among those who are susceptible, which, in Boag's application to data on breast cancer, is given by the log-normal distribution. Mixture models are also known as split-population models [CHU 91].

Many applications on these lines have appeared (see Maller and Zhou [MAL 96]). A large number of them concern the follow-up of disease, where the concepts of cure and immunity have clear meanings. There are also applications in the social sciences, for example, the study of recidivism (see for example, [CHU 91]): how long until someone released from prison re-offends? In that case, however, the mixture model is open to question. Is it

plausible to suppose that everyone can be firmly characterized at the moment of release as either susceptible or immune to re-offending? It seems more realistic to suppose that re-offending, or avoiding re-offending, is an outcome of what happens after the release from prison, rather than being a fate that is already fixed at the moment of release. Similarly, in the analysis of the time until a student's graduation, it is hard to accept a model that characterizes some students from the beginning of their studies as ones who will never graduate. It is much more believable that failure to complete studies arises because of what happens along the way: other life events (marriage, parenthood); fading interest in the subject of study; finding a job that leaves no time for studies, and so on. Farewell [FAR 82, FAR 86] cautioned against the use of mixture models unless there was a clear scientific basis for the existence of an immune proportion.

As already noted in the discussion of the Gompertz distribution, we will see in the following chapter how it is possible for long-term survival to arise as a feature of a model that describes lifetimes, without the need to split the population into groups as in the mixture model.

1.18. Frailty

Covariates are introduced into a statistical model in order to account for heterogeneity between units. However, it is often doubtful that the available covariates are sufficient to represent the heterogeneity completely. There may be other covariates that should ideally have been taken into account, but either were not recorded for some reason, or could not be recorded. An example that is often given of the latter is a genetic factor that is presumed to affect a patient's outcome but cannot be measured.

This unobserved heterogeneity may be introduced into the model by supposing that it can be represented by an individual random effect, specifically each individual's value λ of an unobserved non-negative random variable Λ, which in survival analysis is called the frailty (see Vaupel *et al.* [VAU 79]). The effect of frailty enters the model in a similar way to the effect of covariates. Thus, in a PH context, the hazard function for this individual becomes

$$h(t|\lambda) = \lambda h_0(t),$$

where h_0 is the baseline hazard function as usual. Observed covariates may be included in the same model. Distributions of Λ that are commonly assumed include the gamma distribution (as in the example below), the inverse Gaussian distribution and the positive stable distribution (see Hougaard [HOU 95]).

It is particularly important to allow for all sources of heterogeneity in hazard-based modeling. Without suitable adjustment, a strong selection effect operates: individuals with high frailty λ will tend to die first, leaving a population with relatively low frailty and therefore lower hazard. This gives the impression of a decreasing hazard rate over time; however, this describes the study cohort and should not be misinterpreted as a decline in the risk faced by an individual member of that cohort. The fact that the hazard rate depends on these selection effects as well as on any actual variation in individuals' risk means that, according to Aalen, "the hazard rate is a rather more obscure concept than one should wish, and must be interpreted with great caution" [AAL 94, p. 227]. Selection effects are discussed in detail by Vaupel *et al.* [VAU 79] and Vaupel and Yashin [VAU 85], among others.

However, it should be noted that the frailty term in a PH regression model cannot be used as a handy means of gathering up all the variation that is not accounted for by the measured covariates, including the heterogeneity that would have been accounted for by the unmeasured covariates if they had been available. This is because omitting relevant covariates always results in the attenuation of estimates for the covariates that have been included, as remarked in section 1.15. It has been argued by Hougaard *et al.* [HOU 94] and Keiding *et al.* [KEI 97] that from this point of view, it is preferable to use an AFT model rather than a PH frailty model.

Frailty can also be a very useful device in the analysis of multivariate survival data. Multivariate data can arise when units fall into groups, and common (although unmeasured) factors are expected to be affecting each member of the group: for example, a pair of human twins forms a group of this kind (see Hougaard *et al.* [HOU 92]). This correlation may be represented by each group member having the same value λ of the frailty, even though they possibly differ in values of the measured covariates. Correlation would also be expected between repeated times-to-event measured on the same unit, and the device of introducing an individual random effect may therefore also be useful for recurrent events data (see section 4.5).

EXAMPLE 1.4.– *Suppose that the basic lifetime distribution is Weibull with pdf*

$$S_0(t) = \exp\{-(t/\alpha)^\eta\}$$

and the frailty distribution is a gamma distribution

$$g(\lambda) = \frac{1}{\Gamma(\nu)}\lambda^{\nu-1}e^{-\lambda}$$

with shape parameter ν and scale parameter 1. Because $S(t|\lambda) = S_0(t)^\lambda$, the unconditional survival or reliability function of the lifetime is

$$\begin{aligned} S(t) &= \int_{\lambda=0}^{\infty} \frac{\lambda^{\nu-1}e^{-\lambda}e^{-\lambda u}}{\Gamma(\nu)}\, d\lambda \quad [where\ u = (t/\alpha)^\eta] \\ &= \int_{\lambda=0}^{\infty} \frac{\lambda^{\nu-1}e^{-\lambda(1+u)}}{\Gamma(\nu)}\, d\lambda \\ &= \frac{1}{(1+u)^\nu}\int_{\lambda=0}^{\infty} \frac{(1+u)^\nu\lambda^{\nu-1}e^{-\lambda(1+u)}}{\Gamma(\nu)}\, d\lambda \\ &= \frac{1}{(1+u)^\nu} \\ &= \frac{1}{\{1+(t/\alpha)^\eta\}^\nu} \end{aligned}$$

since the integral integrates the pdf of a gamma distribution with parameters $1+u$ and ν over its entire range, and therefore equals one. The distribution with this survival function is known as the Burr distribution.

1.19. Models for discrete lifetime data

The models that have been discussed so far in this Chapter assume that the time measurement T is a continuous random variable. This is usually true, but two other possibilities need to be at least mentioned, although their relative lack of practical importance can be judged by the small number of pages allotted to them even in such a comprehensive text as Lawless [LAW 03]. One of these possibilities is that the "time" variable is inherently discrete; the other is that it is a continuous measurement that has been grouped into categories.

Inherently discrete measurements of the time until an event occurs may correspond to an operational time such as how often a machine has been used. Another example is how many times something has been attempted before success, or for how many semesters a student has been enrolled until obtaining a degree. Even inherently discrete time variables, however, may often be satisfactorily treated as continuous in order to gain access to the rich array of models for continuous times, few of which have any counterpart in discrete time. The main exception is when the lifetime is short. For a simple analogy, a standard model for a count variable is the Poisson distribution, which can often be adequately approximated by a continuous distribution, but certainly not when the parameter value implies that the count variable typically takes small integer values.

Not many useful discrete distributions are available for modeling lifetimes. The simplest one is the familiar geometric distribution that can model the time to the first success in a sequence of independent Bernoulli trials with constant probabilities of success p and failure $q = 1 - p$:

$$P(T = t) = q^{t-1}p, \;\; t = 1, 2, 3, \ldots$$

For a discrete distribution, the hazard rate at time t is defined as the conditional probability of failure at this moment, given survival so far:

$$h_t = P(T = t | T \geq t) = p_t / S_t,$$

where p_t is the probability distribution and $S_t = P(T \geq t)$. For the geometric distribution, $S_t = q^{t-1}$, therefore $h_t = p$, which is constant for all t. Consequently, the geometric distribution can be regarded as the discrete equivalent of the continuous exponential distribution in that it preserves the property of constant hazard. There have been attempts to define equivalents of other well-known lifetime distributions, such as a discrete Weibull distribution. Since the exponential distribution's survival function is

$$S(t) = e^{-\lambda t} = (e^{-\lambda})^t$$

and the Weibull's is

$$S(t) = e^{-(t/a)^\eta} = (e^{-\lambda})^{t^\eta},$$

where $\lambda = \alpha^{-\eta}$, a "discrete Weibull distribution" could be defined by the survival function

$$S_t = q^{(t-1)^\beta},\ t = 1, 2, \ldots$$

(see Nakagawa and Osaki [NAK 75]).

Further extensions and applications of these ideas are so few that it is not worthwhile pursuing them here. In particular, adapting the regression models presented in the preceding sections from continuous time to discrete time is generally difficult. It is awkward to define a time-transformation method equivalent to AFT that observes the restriction of the time variable to a set of specific discrete values. PH models can be used, but need to be adapted so that the hazards, which are probabilities in the context of discrete time, correctly observe the restriction to the $(0, 1)$ range. (Continuous hazard rates can take any non-negative value.) On the other hand, a PO specification is quite natural for discrete times. In general, the possible models are rather close to generalized linear models (see Lawless [LAW 03]).

Grouped and interval censored data

The second possible form of discrete data arises when all the observations have been recorded in the same intervals. (Grouping may also be used for easier presentation of a large dataset, but if the original observations are available, they should be used in the analysis.) A familiar example of such grouped data is the life table, showing the declining size of a population or cohort year-by-year. Life tables have formed part of the statistical literature for 500 years. Discussion and references on their analysis can be found in Lawless [LAW 03]. An issue that has to be considered is the handling of censored data if the time of censoring is not known exactly, but only known to lie within an interval.

A related form of data arises from interval censoring [SUN 06]. This again means that observations of a continuous time variable have not been recorded exactly, but in intervals. However, the intervals are not necessarily the same throughout the dataset. (It is true of course that it can be claimed that all our apparently continuous measurements are in fact interval censored, with intervals corresponding to the accuracy of our measuring and recording - nearest minute, nearest hour, etc. However, this is rarely a problem; otherwise

we would never be able to use the basic tools of statistical analysis.) Grouping of times arises most commonly because individuals are not being observed continuously but only at specific times. Suppose in general that unit i is inspected at a prespecified sequence of times (e.g. a person in a study has monthly appointments at a clinic) until it is found at one of these inspections that the unit has failed (the event of interest has occurred) during the time that has elapsed since the previous inspection. The information available on the lifetime T_i of this unit is $U_i < T_i \leq V_i$, where $(U_i, V_i]$ denotes the interval in which the failure took place. For a unit that had already failed by the time of the first inspection after entering the study, $U_i = 0$ and the observation is left censored. If a unit is still operating at the final inspection before the study is terminated, then $V_i = \infty$ and the observation is right censored. Clearly, the probability of failure in the interval $(U_i, V_i]$ is $F(V_i) - F(U_i)$, where F is the distribution function of lifetimes, with $F(0) = 0$ and $F(\infty) = 1$ as usual. Therefore, the likelihood of the data is simply

$$L = \prod_i \left[F(V_i) - F(U_i)\right]. \qquad [1.9]$$

Note that for a fully parametric model - including regression specifications - this likelihood can be maximized with the same effort as is required for maximizing the likelihood [1.3]. In this respect, interval censoring presents no analytical difficulty whatsoever. As mentioned for grouped data above, however, care is required in formulating the problem if any observations were censored within intervals (e.g. after some point, a patient in the study failed to turn up for further appointments) or when the next inspection time is not prespecified or even determined independently of the lifetime process, but instead depends in some way on the unit's condition at the previous inspection. Further details can be found in Lawless [LAW 03].

However, it should be pointed out that, in contrast to fully parametric models, the maximization of [1.9] becomes notably difficult for the semi-parametric Cox PH model, because the remarkable simplification that is provided by the reduction to partial likelihood does not work for interval-censored data. (This is because the unspecified baseline hazard $h_0(t)$ does not cancel out.) Consequently, this basic tool of many medical investigators is unavailable for a research design that they quite commonly use. Some approaches to this problem are reviewed by Caroni [CAR 11a].

New methods have been proposed recently by Sun *et al.* [SUN 15] and Wang *et al.* [WAN 16].

One special case of interval-censored data that is, however, of major importance and consequently has largely developed its own literature is current status data. This arises when there is only one inspection of each unit. This means that all that is known for each unit is the inspection time and whether or not the unit had failed at this time. Every observation must be either left censored (if the unit has already failed) or right censored (not yet failed). For reviews of the topic, see, for example, Diamond and McDonald [DIA 92] and Jewell and van der Laan [JEW 03]. It is interesting to note that common practice in biostatistics involves some contradiction in handling these data. Suppose that all units are inspected after the same time, t_0. Given the binary dependent variable (failed/not failed) and a covariate vector $\mathbf{x}$, the routine methodology calls for fitting a logistic regression model, so that the Bernoulli probability $\pi_{\mathbf{x}}$ of failure is modeled as depending on $\mathbf{x}$ through the logit link function

$$\ln\left\{\frac{\pi_{\mathbf{x}}}{1-\pi_{\mathbf{x}}}\right\} = \boldsymbol{\beta}'\mathbf{x}.$$

However, the same practitioners might routinely use the Cox PH model in the same research area if they had lifetimes recorded exactly. Under PH,

$$\begin{aligned} 1-\pi_{\mathbf{x}} &= S(t_0|\mathbf{x}) \\ &= \exp\left\{-e^{\boldsymbol{\beta}'\mathbf{x}}H_0(t_0)\right\} \end{aligned}$$

for baseline cumulative hazard function $H_0(t)$. However, $H_0(t_0)$, being the same for all units, can be absorbed into the linear predictor $\boldsymbol{\beta}'\mathbf{x}$, giving

$$1-\pi_{\mathbf{x}} = \exp\left(-e^{\boldsymbol{\beta}'\mathbf{x}}\right)$$

which can be rewritten as

$$\ln\left(-\ln(1-\pi_{\mathbf{x}})\right) = \boldsymbol{\beta}'\mathbf{x}.$$

In other words, the generalized linear model for current status data that corresponds to the Cox PH model should not use the logit link function, but

the complementary log-log link function instead. Using the logit link (logistic regression) matches the use of the PO model for exactly observed lifetimes, not PH. Of course, as is well known, the logit and complementary log-log transformations are so similar over most of their range that the practical importance of the discrepancy that we have noted is minimal.

1.20. Conclusions

This introductory chapter has aimed to present, at least in outline, the main ways of approaching the regression analysis of lifetime data; that is, how covariates that influence lifetimes can be included in an empirical model. As seen, the most important ways (at least, in terms of the frequency of their application) are by means of the PH model - often equated with Cox's semi-parametric version of the model - and by the AFT model. In addition, the PO model is sometimes seen, and occasionally the additive hazards model, but the others are hardly ever seen. As noted, the great appeal of Cox's model is attributable to the apparent ease of interpretation of the regression coefficients and also to its semi-parametric nature, which avoids the need to specify the parametric form of the baseline hazard function. Although the PH model was initially motivated in section 1.4 by appeal to the device of making a Weibull distribution's parameters depend on the covariates - as is widely done in statistical modeling - the consequence that hazard functions of different units are proportional to each other is a very attractive property. Subsequently, as seen in several sections of this chapter, other "proportional" models have been defined in terms of other functions derived from the distribution of lifetimes, such as the MRL. Because of the need to compete with the predominant semi-parametric Cox model, the promoters of these alternatives have put much effort into their semi-parametric estimation. In fact, in all lifetime data models with non-informative right censoring, fully parametric modeling offers no difficulty in principle. All that is required is to state the model, write down the likelihood for the uncensored and right-censored cases and maximize it numerically.

The AFT model occupies a rather different position. First, it can be written in the form of a familiar regression model, with the logarithm of lifetime as dependent variable and with a suitable error distribution replacing the normal. Second, it is generally used in the fully parametric form, which seems appropriate for its association with experimental data and other reliability

data. However, as noted earlier in this chapter, because of its various desirable properties, it should be more widely used in biostatistical applications as various authors have pointed out.

Neither the AFT nor the PH model includes a representation of the process by which the covariates act upon lifetimes. The same applies to the other regression models that have been mentioned in this chapter. The major objective of this book is to present an alternative class of models that sets out to model the lifetime as the observed outcome of some underlying process. Modeling the mechanism may lead to a more satisfactory model offering greater scope for scientific insight than is possible from strictly empirical models.