Mathematical Survival Analysis

Dominic Owens

May 2, 2019

A project of 20 credit points at level 3.

Supervised by Dr. Skevi Michael.

Acknowledgement of Sources

For all ideas taken from other sources (books, articles, internet), the source of the ideas is mentioned in the main text and fully referenced at the end of the report.

All material which is quoted essentially word-for-word from other sources is given in quotation marks and referenced.

Pictures and diagrams copied from the internet or other sources are labelled with a reference to the web page or book, article etc.

Signed
Dominic Owens

Date
May 2, 2019

Abstract

This report outlines and compares a selection of models developed for survival data analysis. Emphasis is placed on the theoretical bases of the concepts presented. Inferential properties and estimation procedures are described for each model. The investigation is motivated by the limitations of the generalised linear model framework in describing time-to-event problems; consideration is given to amendments or replacements to the framework. Demonstrations of methods are provided in the statistical programming software **R**, recreating results from studies on unemployment benefit programmes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview and Background

Survival analysis is an area of statistics which aims to model *time-to-event* phenomena. Suppose we have observed, for a sample of individuals, both the time taken until the event occurs and a set of covariates containing supplementary information. How might this information be used to describe and infer further information about the entire population? Survival analysis endeavours to answer this question.

Survival analysis is further distinguished from other modelling problems by the presence of *censoring* (formalised in chapter 2), where information on the true survival time is obscured.

For illustration, consider a situation in which we are interested in the lifetime of a group of patients diagnosed with cancer, having recorded their age, sex, and whether the patient is a smoker; or alternatively, the time for which individuals claim unemployment benefits having recorded their previous employment contract (it is this scenario which we study within this report).

The discipline emerged in seventeenth-century London, when a merchant named John Graunt published tables of statistical inference based on mortality records [25]. De Moivre and Bernoulli formalised some key results while they studied actuarial science and demography, though the catalyst for much of the development of the parametric methodology (see chapter 5) in use today was the advent of the second world war, and the subsequent desire by engineers to model the reliability of machinery [2]. More recently, applications to medical statistics and clinical trials have given cause for development of much of the non-parametric methodology, as discussed here in chapter 3.

1.2 The Relationship between Unemployment Benefits and Re-employment Probabilities

Throughout this text, example graphics and calculations are given using data from a study by Jenkins and Garcia-Serrano (2004) [11]. Results from other research on the same topic are recreated (in particular, [18] and [19]).

This data records a sample of Spanish men aged 18-54 years who began to claim unemployment insurance in February 1987, and were followed until they left the scheme or exhausted their entitlement.

The dataset contains various pieces of information associated with each subject. The response we are interested in is the length of the spell they spend in the scheme, *conmths*. We also have continuous quantities including the net replacement rates for given intervals of time spent claiming (*rn1*, *rn2*, *rn3*) and *potmths*, the number of months the claimant is qualified to remain in the scheme for. The provided categorical measures include *famresp*, denoting whether has family responsibilities or not, and *tyentry*, which indicates whether the subject had a permanent or temporary contract prior to becoming unemployed. Information on censoring is given by *exit*.

1.3 Limitations of the Generalised Linear Model Framework

Linear and Generalised linear models (GLMs) are perhaps the most widely-used parametric methods for modelling numerical responses as a function of covariates. One might try to apply these tools to survival data, though as we shall see this quickly becomes problematic.

Consider the linear model. Let Y_i , $i = 1, ..., n$ denote the random variable we wish to model, $(y_i, \mathbf{x}_i)$ be observations of Y_i so that $\mathbf{x}_i$ is a p-vector of covariates associated with y_i. ϵ_i are identically distributed (not necessarily normal) errors with mean 0 and variance σ^2. Let $\boldsymbol{\beta}$ be the p-vector of regression coefficients. In the linear model, we have [4]

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i \tag{1.1}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1.2}$$

This will suffice when describing a situation in which the assumptions hold, and y_i is a real number. In survival analysis the response is a time, and so we require that y_i is positive.

To achieve this, we define further $\mu_i = E(Y_i)$ and the *link function* $g(\mu_i) = log(\mu_i)$, so we have the logistic linear model

$$\eta_i = log(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta} \tag{1.3}$$

This is a particular instance of the GLM, which gives us a response $\eta_i \geq 0$.

It remains to account for incomplete information about the observed survival time. Consider the setting of an observational study in which we collect data on the time individuals spend unemployed; incomplete information might arise for numerous reasons. For instance, a subject may enter the study some time after the start, or leave early for some reason other than finding a job. Even when we do have information on the time of the failure we may only have a range of times in which the event occurred; perhaps the respondents are surveyed at regular intervals regarding their employment status. These problems in combination motivate the concept of *censoring*, discussed in detail in chapter 2.

Moreover, some quantities (such as the age of the subject) vary over the duration of the study [24]. Here, it does not make sense to consider the responses distributed identically conditionally on the covariates.

In **Chapter 2** we define the key quantities and concepts used in studying survival analysis, including the survival function, the hazard function, and censoring.

Each of the next three chapters cover a particular model framework. At the end of each chapter, the methods discussed are demonstrated on a dataset.

In **Chapter 3** we study non-parametric models, including the Actuarial estimate and the Product-Limit estimate, going on to explain how these models can compare groups within data. We also examine a Bayesian non-parametric model.

Chapter 4 concerns semi-parametric models. We focus on the Cox Proportional Hazards model, fitting a model in the absence of a closed-form likelihood, and diagnostic techniques for model evaluation.

Chapter 5 covers the closest relatives of the generalised linear model in survival analysis, parametric models, focusing on accelerated lifetime models for known distributions. Again, diagnostic techniques are discussed.

Chapter 2

Fundamentals of Survival Analysis

In this chapter we introduce two functions fundamental to survival analysis, the *survival function* and the *hazard function*, which characterise a survival time. We examine how these relate to each other and to the probability density function (pdf) of the survival time [3]. Throughout this report, we let T be the random variable which our functions describe.

The second half of this chapter introduces forms of censoring and the likelihood function for censored observations.

2.1 The Survival and Hazard Functions

The *Survival Function*, $S(t)$, is the probability that T is larger than some time t. Equivalently, this is the probability of an individual surviving beyond time t:

$$S(t) = P(T > t) \tag{2.1}$$

We can simply deduce from the definition of $F(t)$, the cumulative density function of T, that

$$S(t) = 1 - P(T \leq t) = 1 - F(t) \tag{2.2}$$

$S(t)$ is thus a decreasing function of t, with $S(0) = 1$, and $\lim_{t\to\infty} S(t) = 0$.

The *Hazard Function* $h(t)$ is the limit of the probability of T failing in a time interval, given that T has not failed by time t:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \tag{2.3}$$

This provides a measure of the instantaneous chance of failure at a given time. Note that $h(t)$ is not a probability, it is in fact a rate [1], meaning it may take any positive value; this is true by virtue of the divisor δt being infinitesimally small.

Similarly to (2.2), we can derive

$$h(t) = f(t)/(1 - F(t)) \tag{2.4}$$

This expression enables us to say $h(t)$ is positive and unbounded above.

A natural quantity to identify is the *Cumulative Hazard Function*, the integral of the hazard function over the range of times up until t:

$$H(t) = \int_0^t h(u)du \tag{2.5}$$

This gives the chance that the individual has failed before time t.

Having either the survival function or the hazard function allows us to obtain the other. This fact can be shown through manipulation of our definitions [3].

Theorem 2.1.1. *S(t), h(t), H(t) and f(t) are equivalent; that is, knowing one of the four functions, it is possible to derive each of the other functions*

Proof. Firstly, by (2.2) and (2.5),

$$h(t) = f(t)/S(t) \tag{2.6}$$

Rearranging (2.2) gives

$$f(t) = \frac{d}{dt}[1 - S(t)] = -S'(t) \tag{2.7}$$

Substituting (2.7) into (2.6) gives

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} log S(t) \tag{2.8}$$

Then integrating (2.8) over the range $[0, t)$, and using the definition of $H(t)$ given in (2.5):

$$H(t) = -log S(t) \tag{2.9}$$

The two integrals are unique, as determined by $h(t)$ and $f(t)$ having integrals over the entire domain equal to 1. □

2.2 Censoring

2.2.1 Types of Censoring

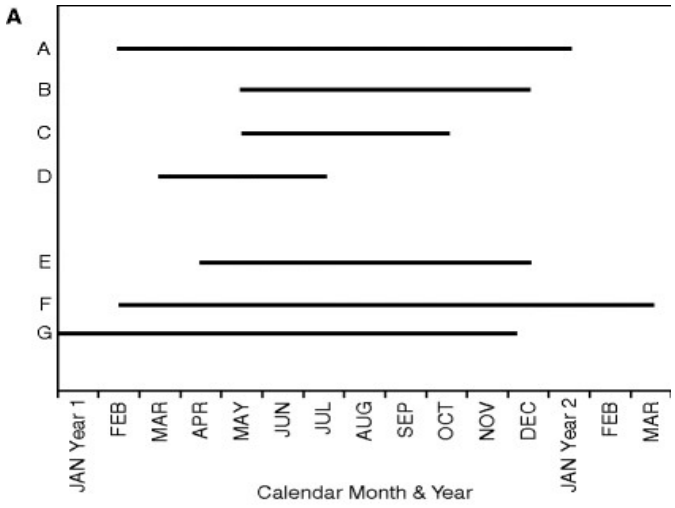

Figure 2.1: Censoring [6]

We consider an observation of T for the individual i to be *right-censored*, the first case of censoring, when information on the failure time T is obscured. We may know that the individual i did not fail during the period for which we have information, but we do not know if i failed at any time after. Observe G in figure 2.1. In practice, this might be an individual who keeps claiming the unemployment benefit until the end of our study, or who stops reporting for some unrelated reason.

The second case, *left-censoring*, occurs analogously at the beginning of the time interval. An individual entering the study late would be left-censored. Observe A or B, for instance, in figure 2.1.

We use the term *interval censoring* to describe a scenario where we may only be able to obtain observations of T within a certain time interval [2]. This is both left- and right-censoring; observe, say, C or D in figure 2.1. Right-censoring occurs more frequently in practice [1], so is of particular interest as a point of study.

A similar yet distinct feature of survival data is *truncation*, which occurs when observations outside a given interval are excluded from the study [5]. This is of less mathematical interest than censoring, as the unrecorded data has no effect on any measured quantity.

We are able to distinguish right-censoring into different cases. Suppose we are unable to observe T after some fixed time c, so that any time exceeding this must be recorded as c^+; this is *Type I* censoring, and in context corresponds to the end of the survey.

Suppose instead that we wish to collect a fixed number of observed failures, and thereafter we end the survey. We call this *Type II* censoring.

Finally, *Type III* or *random* censoring describes when instances of censoring occur randomly. Let the random variable C be the censoring time for T, so that if T exceeds C, we say that T has been censored at C, writing $t = c^+$ for the observed time [2]. In our example, this describes the possibility of a subject leaving the study.

In general, and indeed in this report, we make the important assumption that T and C are independent in their distribution, so that the censoring is *non-informative*. When we do not make this assumption, we are studying *competing risks* (see [2, p.177-179]).

2.2.2 Likelihood Under Right-Censoring

The presence of censoring alters the form of the likelihood used in fitting the generalised linear model. Let $t_i = min\{t_i', c_i\}$ be the observation we record, where t_i' is the realisation of T and c_i the realisation of C_i.

We define the *censoring indicator*, which records if a survival time is censored, to be

$$\delta_i = \begin{cases} 0 \text{ if } t_i' \leq c_i \\ 1 \text{ if } t_i' \geq c_i \end{cases} \tag{2.10}$$

and the densities $f(t)$, $f_c(t)$, survival functions $S(t)$, $S_c(t)$, and hazard functions $h(t)$, $h_c(t)$ respectively for T and C.

Then, using (2.6),

$$\begin{aligned} L &= \prod_{\delta_i=1} f(t_i')S_c(t_i') \prod_{\delta_i=0} f_c(c_i)S(c_i) \\ &= \prod_i h(t_i)^{\delta_i} S(t_i) \prod_i h_c(t_i)^{1-\delta_i} S_c(t_i) \end{aligned} \tag{2.11}$$

The second factorisation expresses the likelihood as a product of a distinct likelihood for T and C, which allows us to perform inference without referencing the distribution of C [21] meaning we need not estimate quantities for C or make further assumptions regarding its' distribution.

Chapter 3

Non-Parametric Models

We have seen that the assumptions of the generalised linear model are not sufficient for survival analysis. Rather than try to adapt what exists by making more assumptions, we could do away entirely with the idea that observations are being drawn from a predetermined distribution, and use solely the information communicated by the data. In this chapter, we see that non-parametric methods specify simple, elegant models which permit group comparison, but the inference which can be conducted is limited by not incorporating covariates.

3.1 Life Tables and The Actuarial Method

The first methods developed for estimating the survival function make use of *life tables*. In these, the total duration of the study is split into k time intervals I_i, usually of equal length, and we record the quantities n_i, d_i, w_i, p_i and h_i, defined below, for each. We tabulate the data in the format of Table 3.1.

Interval	Count at start of I_i	Failures during I_i	Withdrawals during I_i	Proportion failing during I_i	Cumulative Failure
I_1	n_1	d_1	w_1	p_1	h_1
I_2	n_2	d_2	w_2	p_2	h_2
I_3	n_3	d_3	w_3	p_3	h_3
I_4	n_4	d_4	w_4	p_4	h_4
I_5	n_5	d_5	w_5	p_5	h_5
I_k	n_k	d_k	w_k	p_k	h_k

Table 3.1: Life Table

We can easily deduce that $h_j = \sum_{i=1}^{j} p_i$, and that $n_{i+1} = n_i - d_i - w_i$, which is the number of elements in the *risk set*; equivalently, this is the number of individuals yet to fail at the start of time interval I_{i+1}.

The *reduced sample method* uses only observations in $[0, t_k)$ to estimate $S(t_k)$. Let $n = n_1 - \sum_{i=1}^{k} w_i$ and $d = \sum_{i=1}^{k} d_i$, then

$$\hat{S}(t_k) = 1 - d/n \tag{3.1}$$

By omitting the information contained in each w_i this estimator is biased downward and thus underestimates $S(t_k)$ [2].

A more sophisticated relative of this is the *actuarial method*. Making use the fact that we can decompose $S(t_k)$ into a product of probabilities as follows:

$$S(t_k) = P(T > t_k) = \prod_{i=1}^{k} p_i \tag{3.2}$$

where

$$p_i = P(T > t_i | T > t_{i-1})$$

we have an estimate for $S(t_k)$ by estimating each individual p_i and multiplying them,.

Using $1 - d_i/n_i$ to estimate p_i would not account for the censored data w_i. We assume withdrawals occur half way through the interval I_i, giving us the *effective sample size* [2],

$$n_i' = n_i - w_i/2 \tag{3.3}$$

and the estimate for p_i,

$$\hat{p}_i = 1 - d_i/n_i' \tag{3.4}$$

Note this does account for information in w_i. The *actuarial estimate* for $S(t_k)$ is

$$\hat{S}(t_k) = \prod_{i=1}^{k} \hat{p}_i \tag{3.5}$$

By the form of $S(t)$ given in (3.2) and the censored likelihood equation given in section 2.11, we have the likelihood function

$$L(\mathbf{h}|\mathbf{d}, \mathbf{n}) = \prod_{i=1}^{k} h_i^{d_i} (1 - h_i)^{n_i' - d_i} \tag{3.6}$$

which means the log-likelihood function is

$$\ell(\mathbf{h}|\mathbf{d},\mathbf{n}) = \sum_{i=1}^{k}(d_i log(h_i) + (n_i' - d_i)log(1-h_i)) \tag{3.7}$$

Maximising this with respect to h_i gives

$$\frac{\delta\ell}{\delta h_i} = \frac{d_i}{h_i} - \frac{n_i' - d_i}{1-h_i} = 0 \implies \hat{h_i} = d_i/n_i' \tag{3.8}$$

Comparing this to (3.5) tells us the actuarial estimate is the ML estimate.

The two above methods can give us a point estimate for $S(t_k)$, but when conducting inference we also require a form of the variance, either analytically or by estimation. Using the expressions

$$log\hat{S}(t_k) = \sum_{i=1}^{k} log\hat{p}_i \tag{3.9}$$

and

$$n_i'\hat{p}_i = n_i' - d_i \sim Binomial(n_i, p_i) \tag{3.10}$$

it follows from the delta method [5, p. 401] that

$$Var(log\hat{p}_i) \approx Var(\hat{p}_i)(\frac{d}{dp_i}logp_i)^2 = q_i/n_i'p_i \tag{3.11}$$

Assuming independence of $(p_i)_{i=1}^k$ implies

$$\hat{Var}[log\hat{S}(t_k)] = \sum_{i=1}^{k}\hat{q}_i/n_i'\hat{q}_i = \sum_{i=1}^{k}\frac{d_i}{n_i'(n_i'-d_i)} \tag{3.12}$$

and again using the delta method, we derive *Greenwood's formula*

$$\hat{V}(\hat{S}(t_k)) = \hat{S}(t_k)^2\sum_{i=1}^{k}\frac{d_i}{n_i'(n_i'-d_i)} \tag{3.13}$$

as an estimate for the variance of $\hat{S}(t)$.

3.2 The Kaplan-Meier Method

By taking the actuarial estimator and allowing the end points of the intervals I_i to vary, we can construct the *Product-Limit (PL)* estimator, first proposed by Kaplan and Meier (1958). The order statistics $\{t_{(j)}|j = 1, ..., k\}$, so that $t_{(1)} < t_{(2)} < ... < t_{(k)}$, give us natural end points for each interval. Formally, the k-many intervals can be expressed as

$$I_j = (t_{(j-1)}, t_{(j)}] \tag{3.14}$$

where $t_{(0)}$ is set to 0 or to the lower bound of the entire observation interval.

At time t_j there is an incremental decrease in $\hat{S}(t)$. We define

$$S(t_j-) = P(T \geq t_j) \tag{3.15}$$

so that the size of the decrease at t_j is

$$S(t_j-) - S(t_j) = P(T = t_j) \tag{3.16}$$

Rearranging this gives

$$S(t_j) = S(t_j-)\left(1 - \frac{P(T = t_j)}{S(t_j-)}\right) = S(t_j-)\big(1 - P(T = t_j|T \geq t_j)\big) \tag{3.17}$$

Define $h_j = P(T = t_j|T \geq t_j)$ be the *discrete hazard* from (3.17). We can estimate this with the maximum likelihood (ML) estimate $\hat{h_j} = d_j/n'_j$.

The Product-Limit estimator is the step function defined as

$$\hat{S}(t_k) = \prod_{j=1}^{k}(1 - \hat{h_j}) \tag{3.18}$$

We can derive an estimate for the variance of the PL estimator identical in form to (3.13), again using Greenwood's Formula.

As previously noted, the actuarial estimate is the ML estimate on a given set of intervals, so the PL estimate is the ML estimate on the highest resolution of intervals.

Interestingly, the product limit has some desirable estimator properties. Peterson (1977) demonstrated that the PL estimate is consistent - that

is, $\hat{S}(t)$ converges almost surely to $S(t)$ [2, p. 61-63]. To do so, we are required to define the *subsurvival functions*

$$S_u(t) = P(Y > t, \delta = 1) = \int_t^\infty [1 - G(u)]dF(u) \tag{3.19}$$

$$S_c(t) = P(Y > t, \delta = 0) = \int_t^\infty [1 - F(u)]dG(u) \tag{3.20}$$

Note, by the law of total probability, $S^*(t) = S_u^*(t) + S_c^*(t)$. Analogously, define the *empirical subsurvival functions*

$$\hat{S}_u(t) = 1/n \sum_{i=1}^{n} I(T > t, \delta = 1) \tag{3.21}$$

$$\hat{S}_c(t) = 1/n \sum_{i=1}^{n} I(T > t, \delta = 0) \tag{3.22}$$

Theorem 3.2.1. *Let $\hat{S}(t)$ be the product limit estimator of the survival function $S(t)$. Then, as $n \to \infty$, $\hat{S}(t) \xrightarrow{a.s.} S(t)$.*

Proof. According to *Peterson's representation* [2] of the survival function as a function of the subsurvival functions

$$S^*(t) = \Psi(S_u^*; S_c^*; t) \tag{3.23}$$

we can express the PL estimator as

$$\hat{S}(t) = \Psi(\hat{S}_u; \hat{S}_c; t) \tag{3.24}$$

By the Glivenko-Cantelli theorem, both of the empirical functions converge almost surely to their respective subsurvival functions in t. Moreover, Ψ is a continuous function of S_u and S_c in the sup norm, so we conclude

$$\hat{S}(t) = \Psi(\hat{S}_u; \hat{S}_c; t) \xrightarrow{a.s.} \Psi(S_u; S_c; t) = S(t) \tag{3.25}$$

□

Moreover, the PL estimator is asymptotically normal.

Theorem 3.2.2. *$\hat{S}(t)$ converges in distribution to X, where*

$$X \sim N\Big(S(t), \hat{V}(\hat{S}(t))\Big)$$

and $\hat{V}$ is Greenwood's variance.

Proof. See [2, p. 63-65]. □

This permits confidence intervals at level α for $S(t)$ of the form

$$CI_\alpha\big(S(t)\big) = \left[\hat{S}(t) \pm z_{1-\alpha/2}\sqrt{\hat{V}(t)}\right] \tag{3.26}$$

3.3 Hazard Function Estimators

Alternatively, we may estimate the hazard function and obtain the survival function using the equivalence established in 2.1.1.

Intuitively corresponding to the definition of $H(t)$ as an integral of the hazard function, the *Nelson-Aalen (NA)* estimator is

$$\hat{H}(t_k) = \sum_{i=1}^{k} \frac{d_i}{n_i} = \sum_{i=1}^{k} \hat{h}_i \tag{3.27}$$

The *Fleming-Harrington (FH)* estimator is

$$\hat{S}(t_k) = exp\big(-\hat{H}(t_k)\big) \tag{3.28}$$

Compare the functions $1 - x$ and $exp(-x)$ on the interval $[0, 1]$; we can see that $1 - x \leq exp(-x)$, so the FH estimate is greater than or equal to the KM estimate for any given sample.

Similarly to the proof of theorem 3.2.2, we can see that the FH estimator is asymptotically normal, and construct identical confidence intervals. Because the FH estimator also has these desirable properties, both methods can be used interchangeably or as a compliment to each other.

3.4 Group Comparison in Non-Parametric Models

A common problem which arises when conducting practical survival analysis is that of comparing survival times for different groups of individuals. Consider, say, a clinical trial in which we want to test the effect of a treatment on recidivism rates on tumors; one group is given a treatment, and the other is not. Non-parametric methods accomodate for these problems well and allow formal hypothesis testing. Here we consider the case for two groups, though arguments can easily be generalised for a larger number.

One approach is to compare estimates of the hazard functions of the two groups. We use the order statistics to denote the distinct times at which observations are recorded - note that we may have more than one observation recorded at a given time $t_{(i)}$; we denote the *multiplicity* as $d_{(i)}$. Using a second subscript j we denote the multiplicity for group j as d_{ij}, and the number in the risk set of group j by n_{ij}.

Without prior information to suggest we act otherwise, we wish to test the null hypothesis H_0, that the two groups have the same hazard rate, against the two-sided alternative hypothesis H_1, that the two groups differ. Treating $t_{(i)}$ as an independent observation sampled from the risk set, the probability of failure given $t_{(i)}$ is in group j is n_{ij}/n_i.

We can hence say that the total number of observations in group j at time i are akin to a sample without replacement of size d_i from the risk set. d_{ij} hence has a hypergeometric distribution with expectation and variance, respectively,

$$\begin{aligned} E(d_{ij}) &= d_i n_{ij}/n_i \\ Var(d_{ij}) &= \sigma^2 = \frac{n_{i1}n_{i2}(n_i - d_i)d_i}{n_i^2(n_i - 1)} \end{aligned} \tag{3.29}$$

Conditioning on the filtration up to t_i, the random variable $d_{ij} - n_{ij}d_i/n_i$ has expectation 0 and variance σ^2. We can multiply this by a weight $W(t_i)$ to obtain a random variable with expectation 0 and conditional variance $W(t_i)^2\sigma^2$, so that for any $k = 1, ..., m$ the sum

$$M_k = \sum_{i=1}^{k} W(t_i)(d_{ij} - n_{ij}d_j/n_j) \tag{3.30}$$

has expectation 0 and variance $\sum_{i=1}^{k} W(t_i)^2\sigma_i^2$ [21, p. 91].

Here, we invoke the Martingale Central Limit theorem [5, p. 85] to show that M_k is approximately normal, and that the test statistic

$$Z = \frac{\sum_{i=1}^{k} W(t_i)(d_{ij} - n_{ij}d_j/n_j)}{\sqrt{\sum_{i=1}^{k} W(t_i)^2 \frac{n_{i1}n_{i2}(n_i - d_i)d_i}{n_i^2(n_i - 1)}}} \tag{3.31}$$

has the standard normal distribution under H_0.

We are able to adapt Z to suit different testing conditions via manipulation of the weight function W. Setting $W(t_i) = 1 \forall i$ gives the *log-rank test*. This has an equivalent in the Cox regression testing setting, and is demonstrated further on.

The *Gehan test* uses $W(t_i) = n_i$, which reduces the magnitude of the contribution of later observations. Other cases such as the Peto-Prentice and Tarone-Ware statistics exist; see [3, p. 113]. Any such test statistic can be expressed in the form

$$X = \sum_{i=1}^{m} \sum_{j=i}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{3.32}$$

where O_i and E_i denote the number of observed and expected events respectively. This amplifies the effect of the magnitude of the differences in comparison to Z from 3.31. This is the sum of $k(m-1)$-many squared standard normal variables, so X has the asymptotic distribution $\chi^2_{k(m-1)}$, allowing us to conduct a (one-sided) hypothesis test and construct confidence intervals.

3.5 Bayesian Non-Parametric Models

Bayesian models are a specific category of non-parametric model. Like the models discussed previously in this chapter, they attempt to minimise the assumptions required in constructing a workable model. However, Bayesian models are built on entirely different interpretations of the nature of probability, and hence these lead to considerably different results and inferences.

In general, the Bayesian approach to modelling entails selecting a prior distribution for the data, perhaps via experience or expert belief, and then updating this with observations to create a more accurate posterior distribution for the data. Parameters in the model are realisations of random variables, as opposed to unobserved parameters used in the frequentist models in this report. For a detailed description of the differences, see Klein [5, p. 187].

We will consider briefly how Bayesian methods may be used to estimate quantities in survival analysis by adapting the Kaplan-Meier estimator.

Susarla and Van Ryzin developed an adaption of the PL estimator in the Bayesian setting [15]. This uses a Dirichlet process prior for the survival function with a parameter function $\boldsymbol{\alpha}$ and squared-error loss (for definitions of both, see [5, p. 188]). The posterior distribution is conjugate, with the parameter $\boldsymbol{\alpha}^*$ equal to the original parameter plus a point mass

wherever failures occur; that is, on the interval (a, b)

$$\boldsymbol{\alpha}^*((a,b)) = \boldsymbol{\alpha}((a,b)) + \sum_{j=1}^{n} I[\delta_j > 0, a < T_j < b] \tag{3.33}$$

We have the following quantities at time t_i: Y_i, the number of individuals in the risk set; d_i, the cumulative failures; λ_i, the number of censored individuals.

The Bayes estimator derived from this posterior is

$$\hat{S}_B(t) = \frac{\boldsymbol{\alpha}(t,\infty) + Y_{i+1}}{\boldsymbol{\alpha}(0,t) + n} \prod_{k=1}^{i} \frac{\boldsymbol{\alpha}(t,\infty) + Y_{i+1} + \lambda_k}{\boldsymbol{\alpha}(t_k,\infty) + Y_{k+1}} \tag{3.34}$$

The prior is strong, and dominates the influence of observations for small values for n; for large n, this converges to the Kaplan-Meier estimate [5, p. 190]. This estimator often gives a smaller mean square error than $\hat{S}(t)$, even when the prior is inaccurate [2, p. 79]

3.6 Application of Non-Parametric Methods

Using the unemployment study data outlined in chapter 1, we can find point estimates and associated confidence intervals with a PL model. The first six of these are given below in Table 3.1.

time	n.risk	n.event	survival	std.err	lower 95%CI	upper 95%CI
1	1507	64	0.958	0.00519	0.947	0.968
2	1443	111	0.884	0.00825	0.868	0.900
3	1332	106	0.814	0.01003	0.794	0.833
4	717	62	0.743	0.01253	0.719	0.768
5	655	33	0.706	0.01349	0.680	0.733
6	622	30	0.672	0.01419	0.644	0.700

Table 3.1: Kaplan-Meier Table

When plotted, the KM estimates appear as a step function (Figure 3.1).

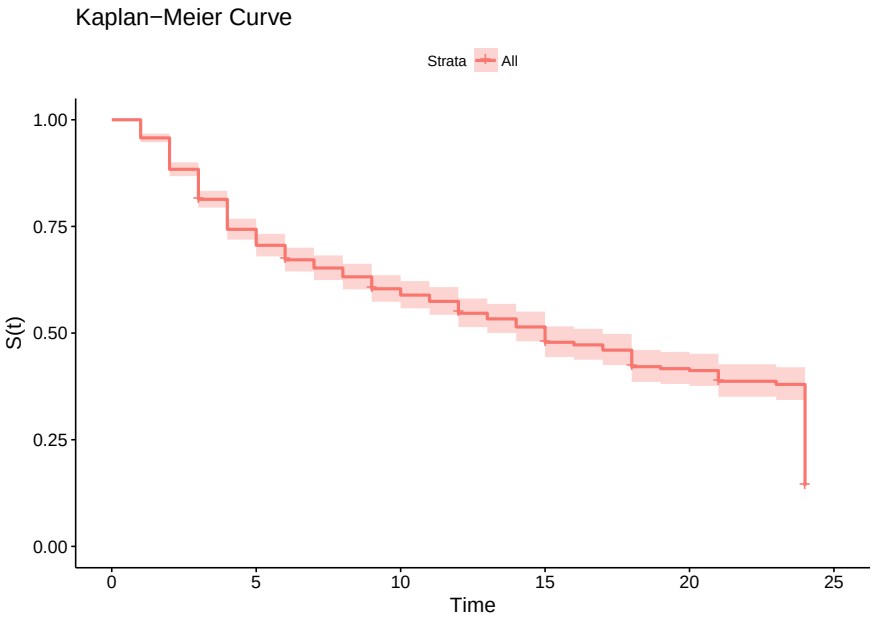

Figure 3.1: Kaplan-Meier Curve

We see, plotting the Fleming-Harrington and Kaplan-Meier estimates simultaneously (Figure 3.2), that the former is biased below the latter.

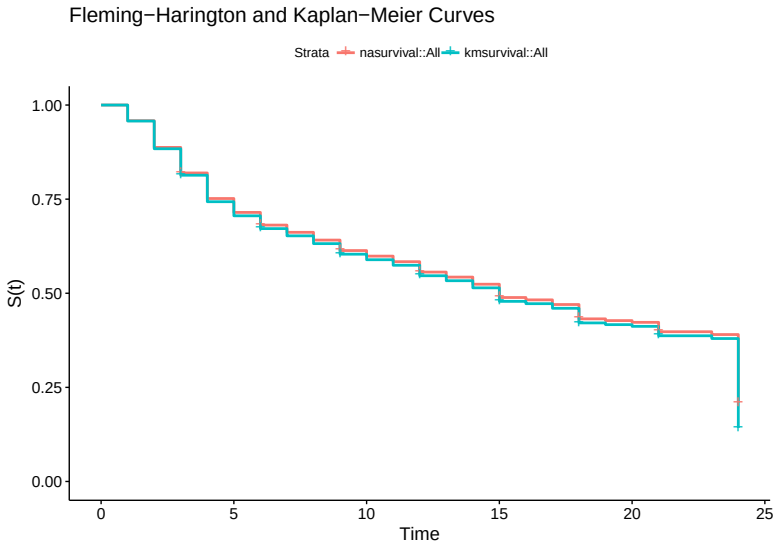

Figure 3.2: Fleming-Harrington and Kaplan-Meier Curves

Perhaps of most interest is comparing survival functions for two groups present in the study. Stratifying the subjects by their type of entry into the

study (*tyentry*), the type of contract the person held before claiming the benefit, gives us the two KM curves as seen below in Figure 3.3.

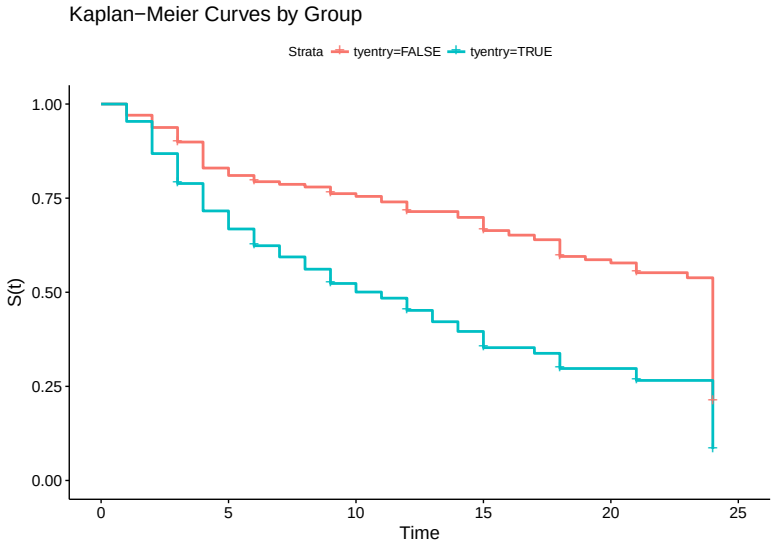

Figure 3.3: Group Comparison by Type of Entry

Note the visible vertical difference between the two curves, where the blue curve ("TRUE") denotes the group previously in temporary employment. We can infer that the temporarily-employed group leave the scheme at a faster rate; this is consistent with the findings of Alba et. al., that "Individuals with temporary contracts in their previous employment have a higher escape rate from unemployment compared to those with permanent contracts."[18].

We can formally test the hypothesis that the two groups have the same survival function, and the output is given below in Table 3.2.

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
tyentry=FALSE	337	204	288	24.4	60.5
tyentry=TRUE	1170	462	378	18.6	60.5

Table 3.2: Log-Rank Test

We have calculations of the statistics in 3.32 for each group. Testing these against the null distribution χ^2_1, we reject the null hypothesis at the 0.05 significance level, indicating a difference in the survival functions depending on the previous contract of the subject in our study.

Chapter 4

Semi-Parametric Models

The results in Chapter 3 demonstrate that, when ignoring the parametric assumptions of the generalised linear model, it is possible to estimate quantities for the whole population, and to compare groups stratified by a categorical covariate. This does not, however, make use of the information encoded in the many other covariates, particularly those which take a continuous value. Semi-parametric models make use of this information to refine inference and estimation procedures, while still making few assumptions on the properties of the model. In this chapter, we focus on one particular semi-parametric model, the Cox proportional hazards model.

For an individual i we have a p-vector of covariates $\mathbf{x}_i = (x_1, x_2, ..., x_p)^T$. The *baseline hazard function* $h_0(t)$ exists for observations with every covariate equal to 0; for our purposes, this need not be specified.

The *proportional hazards (PH) assumption* states

$$h(t; \mathbf{x}) = \rho(\mathbf{x}) h_0(t) \tag{4.1}$$

that is, the *conditional hazard function* $h(t; \mathbf{x})$ is proportional to the baseline hazard by some positive function ρ of the covariates.

Equivalently by (2.1.1) there exists the *conditional survival function*

$$S(t; \mathbf{x}) = S_0(t)^{\rho(\mathbf{x})} \tag{4.2}$$

Cox proposed the model with $\rho(\mathbf{x}) = exp(\boldsymbol{\beta}^T \mathbf{x})$, so that the p-vector of coefficients $\boldsymbol{\beta}$ captures the covariate effects. A categorical (here, binary) covariate x_i with levels $x_i = 0$ and $x_i = 1$ has a multiplicative effect $exp(\beta_i)$ on the hazard function when $x_i = 1$. For a continuous covariate x_j, a unit change also gives a multiplicative change in response by $exp(\beta_i)$. These are identical to the effects seen in the logistic regression model (1.3).

The primary strength of the model Cox proposed is that the *hazard ratio* of two individuals i and j, defined as

$$\frac{h(t;\mathbf{x}_i)}{h(t;\mathbf{x}_j)} = \frac{h_0(t;\mathbf{x}_i)exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{h_0(t;\mathbf{x}_j)exp(\boldsymbol{\beta}^T\mathbf{x}_j)} = exp(\boldsymbol{\beta}^T(\mathbf{x}_i - \mathbf{x_j})) \tag{4.3}$$

where $\mathbf{x}_i$ and $\mathbf{x}_j$ are the observed covariates, is independent of $h_0(t)$. This has a closed form, permitting a heuristic for comparing the survival times of individuals.

4.1 Model Fitting by Partial Likelihood

How might we fit a model, given that the likelihood function is also unspecified? For simplicity, we consider the scenario in which there are no tied survival times. Recalling that n_i is the number in the risk set R_i at time t_i, we have the probabilities

$$P(\text{Any failure in}[t_i, t_i + \delta t)|R_i) = \sum_{j \in R_i} exp(\boldsymbol{\beta}^T\mathbf{x}_j)h_0(t_i)\delta t \tag{4.4}$$

so that

$$P(\text{Failure of element i at time } t_i|\text{A failure at time } t_i) = \frac{exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{\sum_{j \in R_i} exp(\boldsymbol{\beta}^T\mathbf{x}_j)} \tag{4.5}$$

The product of the distinct probabilities (4.5) for each element i leads to the *partial likelihood function*

$$L_p(\boldsymbol{\beta}) = \prod_i \frac{exp(\boldsymbol{\beta}^T\mathbf{x}_i)}{\sum_{j \in R_i} exp(\boldsymbol{\beta}^T\mathbf{x}_j)} \tag{4.6}$$

We call this the "partial" likelihood since this only considers the likelihood for uncensored observations which do indeed fail. Censored observations do implicitly influence L_p by being present in the risk set. This extends to a conditional log-likelihood function ℓ, a score function ℓ', and the information matrix $I(.)$ of second derivatives in much the same way as the full likelihood does. Maximising the log-likelihood function with respect to $\boldsymbol{\beta}$ (which often requires numerical solutions due to the form of the score function - see [5, Appendix A]) gives the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$.

4.1.1 Tied Survival Times

Consider the situation in which tied survival times occur. If we concern ourselves with a dataset from a real-world setting, it is possible that the observed survival times for tied events are in fact different, but recorded as equal due to the limitations of our perception (for instance, the resolution of recording equipment). We could then treat these tied times as distinct. (This supposition is interesting to consider, and potentially problematic; two or more observations on a continuous scale may actually be equal).

We proceed by generalising the argument for two tied events given in Zhang [22, 7.1], recalling that d_i is the event multiplicity for a given time. The observations recorded at t_i may occur in any order, so we consider the $d_i!$-many possible orderings of the events occurring. Let A_r denote one such ordering of events; for example, we have

$$\begin{aligned} P(A_1) = \frac{exp(\boldsymbol{\beta}^T \mathbf{x}_1)}{\sum_{j \in R_i} exp(\boldsymbol{\beta}^T \mathbf{x}_j)} \times \frac{exp(\boldsymbol{\beta}^T \mathbf{x}_2)}{\sum_{j \in R_i} exp(\boldsymbol{\beta}^T \mathbf{x}_j) - exp(\boldsymbol{\beta}^T \mathbf{x}_1)} \times ... \\ \times \frac{exp(\boldsymbol{\beta}^T \mathbf{x}_{d_i})}{\sum_{j \in R_i} exp(\boldsymbol{\beta}^T \mathbf{x}_j) - \sum_{d=1}^{d_i - 1} exp(\boldsymbol{\beta}^T \mathbf{x}_d)} \end{aligned} \tag{4.7}$$

So by the law of total probability,

$$P[\text{Observing } d_i \text{ failures at } t_i] = P(\bigcup_{r=1}^{d_i!} A_r) = \sum_{r=1}^{d_i!} P(A_r) \tag{4.8}$$

This probability will have $d_i! \times d_i$ many terms to calculate, making exact computation very time-consuming and expensive. To overcome this, numerical approximations suffice.

Multiple approximations of (4.8) have been proposed for data with ties ([5, p. 259] outlines some of these in detail), but we focus here on perhaps the simplest, *Breslow's approximation*. This follows from the profile likelihood [5, p. 258].

Define $\mathbf{s}_i$ to be the sum of all the covariate vectors $\mathbf{x}_i$ with observed survival time t_i, and D to be the number of distinct observed survival times, which are contained in the set $\mathbb{D}$. Then we have

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{D} \frac{exp(\boldsymbol{\beta}^T \mathbf{s}_i)}{\prod_{j=1}^{d_i} \left[\sum_{j \in R_i} exp(\boldsymbol{\beta}^T \mathbf{x}_j) - \frac{j-1}{d_i} \sum_{j \in \mathbb{D}} exp(\boldsymbol{\beta}^T \mathbf{x}_j) \right]} \tag{4.9}$$

For low numbers of ties, this approximates L_p accurately, and indeed when there are no ties at any time this is equal to the partial likelihood.

4.1.2 Estimation

Having obtained an estimate for the parameter vector, we have all the information necessary to estimate the hazards ratio for two covariate vectors. By entering the estimate into (4.3), we have the ratio estimate

$$\frac{\hat{h}(t;\mathbf{x}_i)}{\hat{h}(t;\mathbf{x}_j)} = exp(\hat{\boldsymbol{\beta}}^T(\mathbf{x}_i - \mathbf{x}_j)) \tag{4.10}$$

For purposes beyond comparison, we might wish to derive an estimate for the survival function. The natural estimate would be

$$\hat{S}(t;\mathbf{x}) = S_0(t)^{exp(\hat{\boldsymbol{\beta}}^T\mathbf{x})} \tag{4.11}$$

Since the baseline function is unspecified, however, we are required to estimate h_0 or S_0.

Again, multiple solutions are proposed [2, p.134], though here we consider the estimator given by Tsiatis [10]; this permits an estimate for the variance within the likelihood framework. Following Tsiatis' derivation [10, C. 2] we obtain the cumulative hazard estimate

$$\hat{H}_0(t_k) = \sum_{i=1}^{k} \frac{1}{\sum_{j\in R_i} exp(\hat{\boldsymbol{\beta}}^T\mathbf{x}_j)} \tag{4.12}$$

and the survival function estimate

$$\hat{S_0}(t) = exp(-\hat{H}_0(t)) \tag{4.13}$$

An estimator for the asymptotic variance of the hazard function can be found in the same paper [10, p. 33]

4.1.3 Inference

Using the asymptotic normality of any ML estimator, we can test the hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ with, for instance, the Wald statistic

$$X = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T I(\boldsymbol{\beta}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \tag{4.14}$$

This is asymptotically distributed as χ^2_p under H_0. The estimate $\hat{\boldsymbol{\beta}}$ can be derived from the partial likelihood [2, p.132].

4.2 Model Diagnosis

Having fitted a model, we should want to assess the suitability of our model to the observations. In particular, this means evaluating the proportional hazards assumption and identifying problematic data points - those with undue influence, or those which are outliers. Three approaches to this problem are considered here: graphical methods, residual analysis, and time-dependent variable analysis.

4.2.1 Graphical Methods

By plotting certain quantities against each other, we might examine the relationship they hold. This provides a reasonable heuristic for assessing the proportional hazards assumption.

Log-Log plots overlay the function $-ln(-ln(\hat{S}(t;\mathbf{x})))$ evaluated for different covariate vectors $\mathbf{x}$. If the proportional hazards assumption is valid, these curves will be parallel.

Suppose the proportional hazards assumption from section 4.1 holds for some vector $\mathbf{x}_i$, then $-ln(-ln(\hat{S}(t;\mathbf{x}_i))) = -\boldsymbol{\beta}^T\mathbf{x}_i - ln(-ln(\hat{S}_0(t)))$. [1] Hence for two such vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, the difference in log-log functions is equal to $\boldsymbol{\beta}^T(\mathbf{x}_i - \mathbf{x}_j)$. This is constant, inducing the parallel property in the plot.

We can also use the *Arjas plot* to test the PH assumption for one categorical covariate, even when other categorical covariates are present in the data. Suppose we have fitted a model with a hazard function for the individual i as $h(t;\mathbf{x_i})$, and we want to decide whether to include the binary categorical variable z_i [21]. Define the *weighted time on test* for i at time t_j as $\hat{H}(t_j \wedge T_i|\mathbf{x_i})$, and the *total time on test* at level g (taken as 0 or 1) to be

$$TOT_g(t_j) = \sum_{z_i=g} \hat{H}(t_j \wedge T_i|\mathbf{x_i}) \tag{4.15}$$

The number of failures at level g is

$$N_g(t_j) = \sum_{z_i=g} \delta_i \mathbf{1}(T_i \leq t_j) \tag{4.16}$$

Under the null hypothesis, when the total time on test and number of failures are plotted against each other a line of gradient 1 will form. Levels with proportional effects show tilted slopes, while non-proportional effects will not be lines [21, p. 98].

4.2.2 Residuals

Of course, as mathematicians we want to provide a more rigorous analysis than that given by commenting on plots; conclusions drawn in this manner are subjective and non-systematic. Perhaps a more appealing alternative is to perform testing on *residuals*, the differences between observed survival times and those predicted by the model.

Raw residuals for a linear model are well-known to be $e_i = y_i - \hat{y}_i$. In the generalised case, of course, these are less easily calculated due to the presence of link functions and heteroscedastic data. The situation under censoring is further obfuscated, giving rise to various proposed residuals. Each of these is useful for different purposes.

The simplest residuals are the *Cox-Snell residuals*. Under the Cox model, $H(T)$ is distributed as $Exp(1)$[21, p. 103], allowing us to adapt the residuals found under the GLM with an exponential link function.

Calculating $\hat{H}(t_i; \mathbf{x}_i) = -log(\hat{S}(t_i; \mathbf{x}_i))$ for each observation $(t_i, \mathbf{x}_i)$ gives us an observational distribution. We can plot these against the observed outcomes and conduct graphical analysis as before, or conduct hypothesis testing for example on the sample mean of our observed survival times - if this differs significantly from 1, we have grounds to say the exponential model does not fit.

Conveniently, we can adapt the Cox-Snell residuals to find the *Martingale residuals*, which are roughly analogous to the residuals found for a linear model in terms of distribution. Loosely, these are the difference between the observed number of failures for an individual and the expected number of failures from the model. More formally, denoting for individual i an indicator taking value 1 when the individual has failed, we have

$$\delta_i - H_i(t) = \delta_i - \int_0^{T_i} h_0(s) exp(\boldsymbol{\beta}^T \mathbf{x}_i) ds \tag{4.17}$$

We estimate this quantity using (4.12) for the estimated cumulative hazard

$$M_i(t) = \delta_i - \hat{H}_i(T_i) = \delta_i - \sum_{t_j \leq T_i} exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_i) \frac{1}{\sum_{l \in R_j} exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_l)} \tag{4.18}$$

When no ties are present, these sum to 0, as is the case for linear model residuals.

If true values rather than parameter estimates were used in (4.18), this would be a Martingale process [5] (the theory of which we will not cover

in this report), giving us a framework to analyse our quantity which describes the number of events not captured by the model. While we cannot compare these to a known distribution as we could with linear model residuals, we can use these to find an optimal transformation to apply to a covariate to preserve the proportional hazard form.

For a particular covariate z_i in the vector $\mathbf{x}$, instead of the relative risk $exp(\beta z_i)$ the actual relationship may be $exp(\beta f(z_i))$ for some function f. By expressing our vector of covariates as $(\mathbf{x}'^T, z)^T$, and the log-risk as

$$log\rho(\boldsymbol{\beta}, \mathbf{x}, z) = \boldsymbol{\beta}^T \mathbf{x}' + f(z) \tag{4.19}$$

we have a form which permits testing of the PH assumption for only the covariate z. From Fact 7.2 in Steinsaltz [21, p. 106]

$$E[M|z] \approx \frac{\sum \delta_i}{n}(f(z) - log\big(E[P\{h(z)|\text{at risk time } \infty; \mathbf{x}\}]\big) \tag{4.20}$$

in combination with the calculated martingale residuals, we can estimate $f(z)$. Individuals with large, positive values for $f(z)$ are indicated by $E[M|z]$ being much larger than the calculated residual, while large negative values for $f(z)$ are indicated by having less calculated failures than expected.

The final form of residuals we consider are the *deviance residuals*, which are particularly suited to identifying outlying results. We might hope the martingale residuals would indicate outliers by exhibiting large values, however they are restricted to the range $(-\infty, 1]$. Moreover, we don't know the form of the distribution these should follow under a null hypothesis, meaning we cannot elucidate exactly what constitutes "large".

Using the definition of deviance for any model, and in particular a linear model [4, p. 57], as

$$D = 2[\ell(\text{saturated model}) - \ell(\text{fitted model})] \tag{4.21}$$

we can find a form of the deviance for the Cox model [5] and thus derive the deviance residuals

$$d_i = sgn(M_i)\Big\{ -2[M_i + \delta_i log(\delta_i - M_i)] \Big\}^{1/2} \tag{4.22}$$

We know the distribution of the deviance follows χ^2_{p-q}, where p and q are the respective number of parameters fitted. Thus, each deviance residual is drawn from the standard normal distribution under the null hypothesis, so values of magnitude greater than 2.5 can be considered outliers (this is analogous to a confidence interval at the .05 significance level).

4.2.3 Time-Dependent Variables for Diagnosis

Throughout this chapter we have studied the Cox model under the assumption that the covariates contained in $\mathbf{x}$ are constant with respect to time. By instead considering each as a function of time, we can test if the PH assumption holds.

We test the time-dependence of a single covariate z within the vector $\mathbf{x}$ by defining a further covariate

$$z'(t) = z \times g(t) \tag{4.23}$$

where $g(t)$ is a function of time. Fitting a single model with covariates z_i and z'_i for every observation i gives coefficient estimates $\hat{\beta}$ and $\hat{\beta}'$ respectively. The hazard function, with all other covariates set to 0, is

$$h(t; z, z') = exp(\hat{\beta}z + \hat{\beta}'(z \times g(t)))h_0(t) \tag{4.24}$$

and our test statistic, the hazards ratio for individuals i and j, is

$$\frac{h(t_i; z_i, z'_i)}{h(t_j; z_j, z'_j)} = exp(\hat{\beta}(z_i - z_j) + \hat{\beta}'g(t)(z_i - z_j)) \tag{4.25}$$

(Note our test is equivalent to testing the null hypothesis $\hat{\beta}' = 0$).

How to determine the form of g is a pertinent question here - we may simply plot the covariate against time and guess an appropriate g to use by eye, or repeat the testing procedure with different functions and select that which has the greatest p-value.

Common functional forms for g include $g(t) = t$, $g(t) = log(t)$, and the heaviside function incremented at a chosen time t_0 [1, p. 225].

With time-dependent covariates we have the *extended Cox model*. The hazard function is then

$$h(t; \mathbf{x}(t)) = exp(\boldsymbol{\beta}^T \mathbf{x}(t))h_0(t) \tag{4.26}$$

and the partial likelihood is

$$L_p(\boldsymbol{\beta}) = \prod_i \frac{exp(\boldsymbol{\beta}^T \mathbf{x}(t)_i)}{\sum_{j \in R_i} exp(\boldsymbol{\beta}^T \mathbf{x}(t)_j)} \tag{4.27}$$

This permits parameter estimation in the same way as the time-invariant case. However, inference cannot be conducted in the same manner, as no proof exists of the asymptotic normality of the ML estimator $\hat{\boldsymbol{\beta}}$ [2, p. 140].

4.3 The Additive Hazards Model

Throughout this chapter, we have assumed proportional hazards; recall this means that conditioning on an individual's observed covariates gives the hazard function as a multiple of the baseline hazard function. While proportional hazards models are elegant in their calculations, the PH assumption often does not hold in practice. The additive model offers a more general alternative, which is both easier to fit to time-varying covariates and permits a simpler interpretation of coefficients [21, p. 91] as *excess mortality*, the change in lifetime associated with an increase in one particular covariate.

Express the conditional hazard function as a linear combination of the covariates, so that

$$h(t;\mathbf{x}) = h_0(t) + \sum_{k=1}^{p} \beta_k x_k = h_0(t) + \boldsymbol{\beta}^T \mathbf{x} \tag{4.28}$$

The model with constant coefficients in (4.28) is that proposed by Lin and Ying [12], which we will focus on. An alternative model in which the coefficients are functions of time was proposed by Aalen [5, p. 328].

This model permits easier estimation of $\boldsymbol{\beta}$, based on counting process theory [12, Section 2]. We do so following the procedure given by Klein and Moeschberger [5, 10.3]. Recall that δ_i is the censoring indicator. Define an indicator $Y_j(t)$ which is equal to 1 when the individual j is yet to fail at time t, the vector $\bar{\mathbf{x}}(t)$, which is the covariate average at time t,

$$\bar{\mathbf{x}}(t) = \frac{\sum_{i=1}^{n} \mathbf{x}_i Y_i(t)}{\sum_{i=1}^{n} Y_i(t)} \tag{4.29}$$

the matrix $\mathbf{A}$ such that

$$\mathbf{A} = \sum_{i=1}^{n} \sum_{j=1}^{i} (T_j - T_{j-1})[\mathbf{x}_i - \bar{\mathbf{x}}(T_j)]^T [\mathbf{x}_i - \bar{\mathbf{x}}(T_j)] \tag{4.30}$$

the vector $\mathbf{b}$ such that

$$\mathbf{b} = \sum_{i=1}^{n} \delta_i [\mathbf{x}_i - \bar{\mathbf{x}}(T_i)] \tag{4.31}$$

and the matrix $\mathbf{C}$ such that

$$\mathbf{C} = \sum_{i=1}^{n} \delta_i [\mathbf{x}_i - \bar{\mathbf{x}}(T_j)]^T [\mathbf{x}_i - \bar{\mathbf{x}}(T_j)] \tag{4.32}$$

giving the estimator

$$\hat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\mathbf{b}^T \tag{4.33}$$

which maximises the partial likelihood, and a corresponding variance estimate

$$\hat{\mathbf{V}} = \hat{Var}(\hat{\boldsymbol{\beta}}) = \mathbf{A}^{-1}\mathbf{C}\mathbf{A}^{-1} \tag{4.34}$$

These extend to a test statistic for the hypothesis $H_0 : \boldsymbol{\beta}_j = 0$

$$\frac{\hat{\boldsymbol{\beta}}_j}{\sqrt{\hat{\mathbf{V}}_{jj}}} \tag{4.35}$$

which is asymptotically distributed as the standard normal.

4.4 Application of Semi-Parametric Methods

Here we will build and evaluate a Cox model for the study data, providing contrast to the PL model fitted in chapter 3. Fitting a full Cox model for the data (output for which is given in Figure 4.1), which includes all the covariates, yields only two covariates deemed significant at the $p = 0.05$ level. These are *potmths* and *tyentry*. This is not in agreement with the single-hazard model fitted by van Soest and Nagore Garcia [19] in a separate study, which found that the region in which the individual lived was significant (note, however, their data includes the unemployment rate of the region as opposed to only the level of the region itself).

Table 4.1: Full Cox Model

	Dependent variable:
	sfit
age	-0.0001
	(0.005)
famrespNo responsibilities	-0.101
	(0.084)
groupregIslands	-0.068
	(0.199)
groupregNorth	0.044
	(0.099)
groupregNorth-Ea	-0.149
	(0.127)
groupregSouth	-0.132
	(0.113)
potmths	-0.042^{***}
	(0.008)
rn1	0.956^{*}
	(0.575)
rn2	0.254
	(0.813)
rn3	-0.044
	(0.629)
tyentry	0.519^{***}
	(0.103)
Note:	*p<0.1; **p<0.05; ***p<0.01

Fitting a new model with only the selected explanatory variables included gives the following survival plot (Figure 4.1).

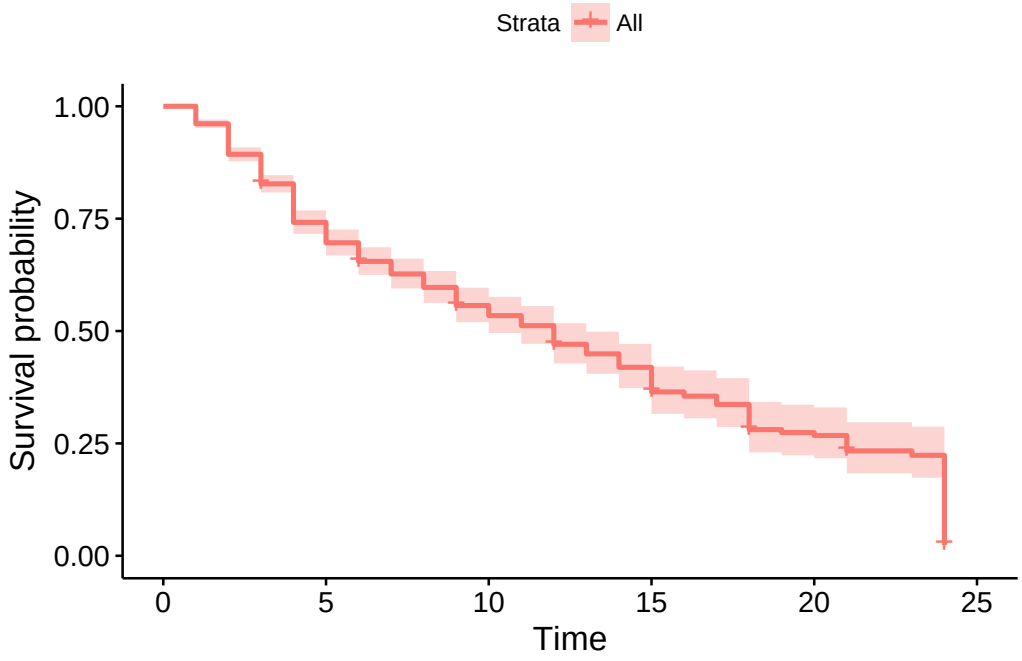

Figure 4.1: Cox Model Plot

Note this fit is calculated using Efron's approach to tied survival times [2, p. 134], as opposed to Breslow's method described in this chapter.

Performing the Wald test described in section 4.1.3 returns a test statistic of 79.3 on 2 degrees of freedom, indicating the parameters are significantly different from 0.

We are able to compare log-log plots for the model stratified by *tyentry*, allowing evaluation of the PH assumption. The plot below (Figure 4.2) appears to show the permanent and temporary groups' curves as vertical translations of each other, indicating they share proportional hazards.

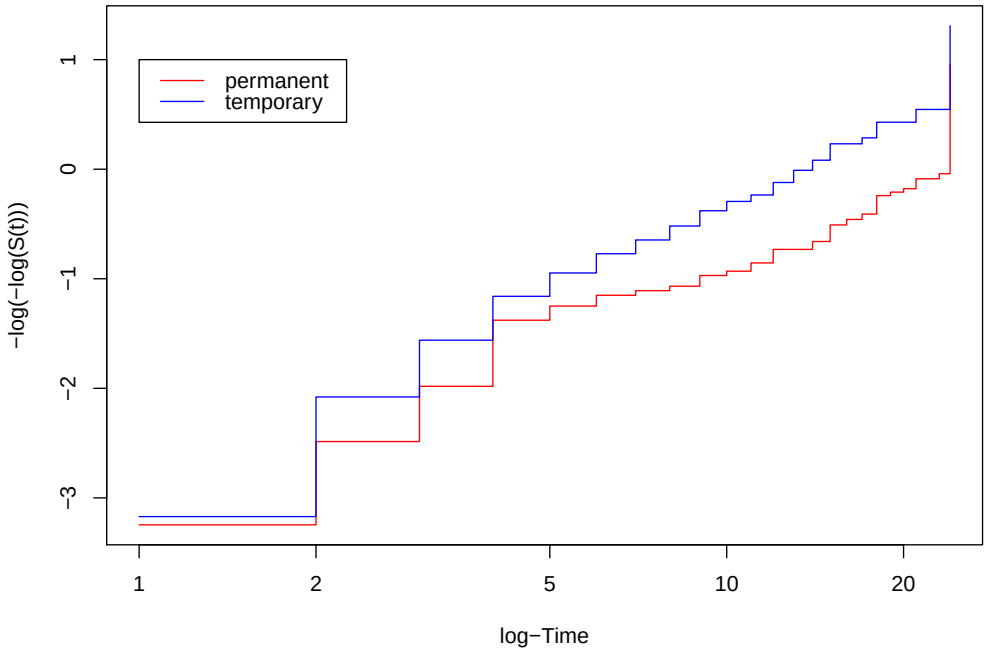

Figure 4.2: Log-Log Plot

To judge the goodness-of-fit of the model, deviance plots allow us to see how close fitted values are to observed values. The martingale (Figure 4.3) and deviance residuals (Figure 4.4) for the reduced model are shown below; note some observations have deviance residuals greater than 2.5, suggesting they are outliers in the data. Moreover, there appears to be a decreasing trend in both residuals, suggesting the model is misspecified.

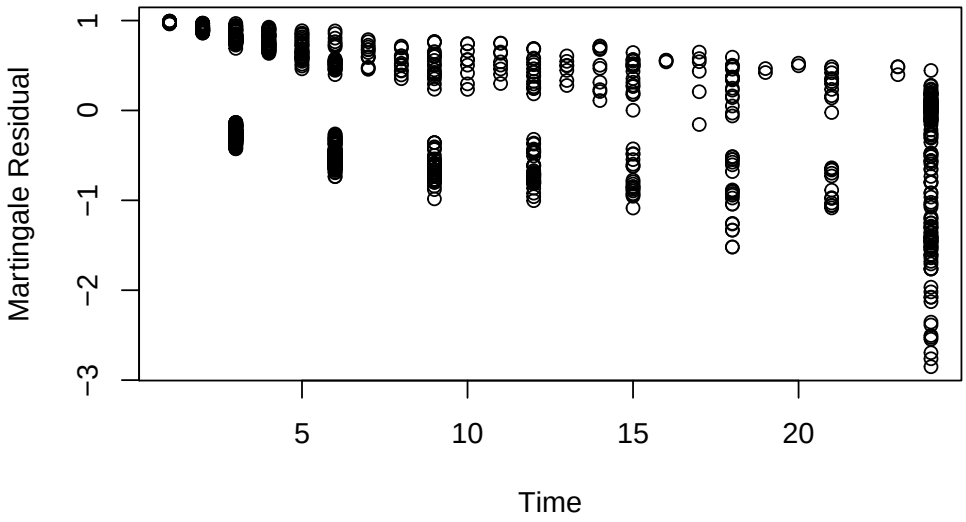

Figure 4.3: Martingale Residual Plot

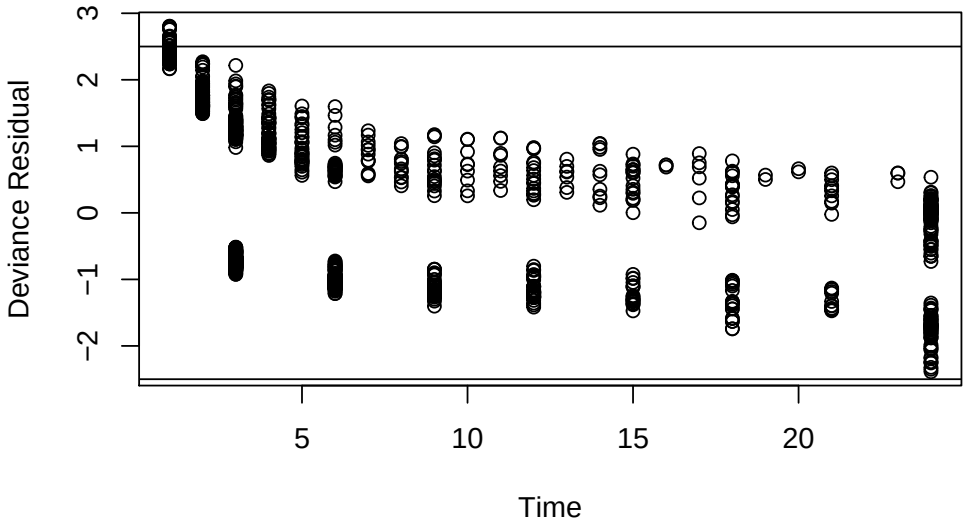

Figure 4.4: Deviance Residual Plot

Chapter 5

Parametric Models

Perhaps the most obvious response to the shortcomings of the generalised linear model is to use another parametric survival model. These models assume that an individual's survival time follows a known probability distribution, which can be altered by conditioning on the observed covariates, just the same as a GLM would. Three main forms are proposed in the literature for the dependence relationship with the covariates, two of which we considered in the previous chapter: additive hazards in section 4.3, proportional hazards 4.1, and *accelerated lifetime* (AL) models (also known as accelerated failure time (AFT) models). We concentrate on the latter form.

5.1 Accelerated Lifetime Models

Suppose we have a baseline survival function $S_0(t)$ which all individuals adhere to, but each experiences it at a different rate conditionally depending on their covariates. We say the individual i has the *acceleration parameter* ρ_i, which captures the proportional speed at which i passes through the baseline function. The most commonly used parameter is $\rho_i = exp(\boldsymbol{\beta}^T \mathbf{x}_i)$, which is strictly positive and therefore appropriate for describing a "stretch" factor.

This induces a survival function for i of the form

$$S_i(t) = S_0(\rho_i t) = S_0(exp(\boldsymbol{\beta}^T \mathbf{x}_i)t) \tag{5.1}$$

Following from the functional equivalence established in chapter 2, we have the cumulative hazard $H_i(t) = H_0(\rho_i t)$ and the hazard function $h_i(t) = \rho_i h_0(\rho_i t)$. Evidently, covariate changes have a multiplicative effect

on the survival time, whereas in PH models covariates have a multiplicative effect on the hazard [1, p. 266].

Both other models can be expressed as AL models by substituting t with $\rho_i t$ [21, p. 76]. As such, we will only consider AL models during this chapter.

As with the GLM, we specify the AL model in matrix form for multiple observations. Given the positivity constraint imposed on the response, it is natural to choose the link function $y_i = log(t_i)$, so that [2, p. 143]

$$E(Y_i) = \boldsymbol{\beta}^T \mathbf{x}_i + E(log(h_0(\boldsymbol{\beta}^T \mathbf{x}_i t))) \tag{5.2}$$

and

$$\mathbf{y} = log(\mathbf{t}) = \boldsymbol{\mu} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{5.3}$$

where $\boldsymbol{\varepsilon}$ is the error vector. This model can alternatively be described as a linear model represented in log time [5, p. 393].

How we choose to define the distribution of the errors is perhaps the most important part of parametric modelling; some potential choices are discussed further on in this chapter.

5.2 Maximum Likelihood Estimation

We saw the derivation of the likelihood under right-censoring in section 2.2.2. How might we obtain a maximum-likelihood or least-squares parameter estimate from this function? As is so often the case, multiple solutions have been proposed to this problem, and we will focus on one: the *Miller estimators*.

We follow the argument given by Miller [2, p. 146-150] when $p = 1$; the generalisation to the multidimensional ($p > 1$) case is intuitive.

Without censoring, we find $\hat{\alpha}$ and $\hat{\beta}$ which solve the least-squares problem

$$\sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 = n \int_{-\infty}^{\infty} z^2 dF_n(z) \tag{5.4}$$

where the right side is a stochastic integral with respect to F_n, the empirical distribution of the errors $z_i = y_i - \alpha_i - \beta_i$.

In our case, where censoring is present, Miller suggests to minimise

$$n \int_{-\infty}^{\infty} z^2 d\hat{F}(z) = \sum_{i=1}^{n} \hat{w}_i(\beta)(y_i - \alpha - \beta x_i)^2 \tag{5.5}$$

where $\hat{F}$ is the product limit estimator as defined in (3.18), and the weights $\hat{w}_i(\beta)$ are the changes in the estimator (these appear as vertical distances when plotted). These weights are dependent on censoring; although if $\delta_i = 0$ (i.e. i is censored) then $\hat{w}_i(\beta) = 0$, the PL estimator of the survival function for i does depend on censoring. We hence obtain a weighted squares formula with dependency on censoring, allowing us to further obtain parameter estimates informed by censoring.

Differentiating (5.5) by α, we find $\hat{\alpha}$ as a function of β:

$$\hat{\alpha} = \sum_{i=1}^{n} \hat{w}_i(\beta)(y_i - \beta x_i) \tag{5.6}$$

Substituting this form for $\hat{\alpha}$ into (5.5) gives the weighted squares quantity as a function of β alone

$$f(\beta) = \sum_{i=1}^{n} \hat{w}_i(\beta)(y_i - \hat{\alpha} - \beta x_i)^2 \tag{5.7}$$

This permits numerical estimation akin to the Newton-Raphson method; for more on this applied to survival analysis, see [5, Appendix A].

5.3 Defined Distributions

In some situations we might have reason to believe that the survival times follow a particular known distribution. As is the nature of statistical modelling, we cannot observe or prove that the data follows a distribution, but as we discuss later we can assess the fit of a chosen parametric distribution.

One distribution the errors may follow is the *Weibull distribution*. This occurs frequently, for instance, in engineering applications when modelling machine lifespans [21, p. 10]. $Weibull(\alpha, \rho)$ is defined by the probability density function

$$f(t) = \rho\alpha(\rho t)^{\alpha-1} exp\big(-(\rho t)^{\alpha}\big) \tag{5.8}$$

or equivalently the survival function

$$S(t) = exp(\big(-(\rho t)^{\alpha}\big);\ \rho = exp(\boldsymbol{\beta}^T \mathbf{x}) \tag{5.9}$$

or hazard function

$$h(t) = \rho\alpha(\rho t)^{\alpha-1} \tag{5.10}$$

By inspection, the hazard function is monotone with respect to t; whether this is increasing, decreasing or constant depends on the shape parameter α and scale parameter ρ. This is the only model which has both a PH and AL representation [5, p. 395].

Here, the likelihood function takes the form

$$L(\alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\alpha \rho_i^{\alpha} t_j^{\alpha - 1} \right)^{\delta_i} exp\big(- (\rho_i t_i)^{\alpha} \big) \tag{5.11}$$

Maximum likelihood estimates $\hat{\alpha}$ and $\hat{\boldsymbol{\beta}}$ may only be obtained numerically [21, p. 77].

Instead, the errors might follow the *exponential distribution*, a particular case of the Weibull distribution with shape parameter $\alpha = 1$. This has the likelihood

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \rho_i^{\delta_i} exp\big(- \rho_i t_i \big) \tag{5.12}$$

We can test the null hypothesis that $\alpha = 1$ using the fact

$$2log\hat{L}_{Weib} - 2log\hat{L}_{Exp} \sim \chi_1^2 \tag{5.13}$$

[21, p. 88], which is an application of the test statistic from (3.32).

Alternatively, the errors might draw from the *log-normal distribution*. As suggested by the name, this distribution corresponds to a random variable whose logarithm follows a normal $N(\mu, \sigma^2)$ distribution. $Lognormal(\mu,\sigma^2)$ is characterised by the functions

$$f(t) = (t\sigma\sqrt{2\pi})^{-1} exp\left(\frac{(lnt - \mu)^2}{2\sigma^2} \right) \tag{5.14}$$

$$S(t) = 1 - \Phi\left(logt - (\mu + \boldsymbol{\beta}^T \mathbf{x})/\sigma \right) \tag{5.15}$$

We can generalise all of the aforementioned distributions as the *generalised gamma distribution*, characterised by the pdf

$$f(t) = \frac{|\theta|(exp(\theta t)/\theta^2)^{1/\theta^2} exp(-exp(\theta t)/\theta^2}{\Gamma(1/\theta^2)} \tag{5.16}$$

The particular cases of interest are $\theta = 1$, which is the Weibull model; $\theta = 0$, the log-normal model; and the case when $\theta = 1$ and $\alpha = 1$, the exponential model [5, p. 406].

We are able to test hypotheses regarding these parameters by considering the generalised gamma distribution within the exponential dispersion family, details of which were described by Jorgensen [13, p. 12]. We should note here that there is not necessarily correspondence here between the generalised least squares estimate and the MLE [14], as is the case for an exponential family distribution.

5.4 Model Diagnosis

The natural progression, having fitted one of the models discussed above, is to assess the fit and evaluate the assumptions of the model. One might prefer graphical methods similar to those shown in section 4.2.1, given that analytical methods often suffer from low power in small-sample scenarios or bias in large-sample scenarios [5, p. 409]. Both are presented for consideration.

5.4.1 Graphical Methods

An effective way to check the fit of a model is to plot a function of the cumulative hazard which, under a correct model, will be linear against some function of time [5, p. 410]. For example, the exponential model has cumulative hazard $H(t) = \rho t$, so we expect the Nelson-Aalen estimator (3.27) to be linear against t. The following table (Table 5.1), adapted from Klein [5, p. 410], gives the plots for the distributions we focus on.

Model	Cumulative Hazard	Plot
Exponential	ρt	$\hat{H}$ vs t
Weibull	ρt^{α}	$log\hat{H}$ vs $logt$
Log-Normal	$-log(1-\Phi(log(t)-\mu))/\sigma)$	$\Phi^{-1}(1-exp(-\hat{H}))$ vs $logt$

Table 5.1: Parametric Distribution Plots for Goodness-of-fit

5.4.2 Analytical Methods

Given the form of the parametric (AL) model, we can translate many of the analytical methods for diagnosis used when fitting GL models.

We can discriminate between the many possibly-fitted models by selecting that which minimises the *Akaike Information Criterion* (AIC). This

rewards increases in likelihood while penalising the amount of parameters specified, and is derived by estimating the likelihood equation for the model fitted to the true parameters. In context, this is

$$AIC(p,k) = -2\ell(.) + 2(p+k) \tag{5.17}$$

where p is the number of parameters supplied by the covariate vector, and k is the number of parameters specified by the distribution of the errors; hence, $k = 1$ for the exponential model, $k = 2$ for the Weibull and log-normal models, and $k = 3$ for the generalised gamma [5, p. 406].

When evaluating a GL model, we may compare the sum of the standardised residuals with a χ^2_p variable, where p is the number of parameters fitted [4, p. 133]. We do the same in the survival analysis setting with standardised residuals of the form

$$s_j = \frac{logT_j - \hat{\mu} - \hat{\boldsymbol{\beta}}^T \mathbf{x}_j}{\hat{\sigma}} \tag{5.18}$$

We know that the Weibull and exponential distributions have errors following the standard extreme value distribution by $f(w) = exp(w - exp(w))$ [5, p. 396], while the log-normal distribution has errors following a censored sample from the standard normal distribution [5, p. 406]. Hypothesis testing with each of these as a null hypothesis permits a goodness-of-fit test.

This is equivalent to testing on the Cox-Snell residuals, as defined for the Cox model in section 4.2.2. These are given for parametric distributions in Table 5.2. Under the null hypothesis, these should follow a standard exponential distribution [5, p. 414], so we can test against the sample mean not being equal to 1.

Model	Cox-Snell Residual
exponential	$r_j = \hat{\rho} t exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_j)$
Weibull	$r_j = \hat{\rho} exp(\hat{\boldsymbol{\beta}}^T \mathbf{x}_j) t^{\hat{\alpha}}$
log-normal	$r_j = log\left(1 - \Phi\left(\frac{logT_j - \hat{\mu}\ \hat{\boldsymbol{\beta}}^T \mathbf{x}_j}{\hat{\sigma}}\right)\right)$

Table 5.2: Parametric Cox-Snell Residuals

See also the Kolmogorov-Smirnoff goodness-of-fit test for distributions for a stronger test on scale, location and shape parameters [17].

5.5 Application of Parametric Methods

We'll fit the three parametric models discussed in this chapter - the exponential, Weibull, and log-normal - to our study data and decide which fits best. Based on previous analysis, we'll fit a model with *potmths* and *tyentry* as covariates. Below, in table 5.3, we have parameter estimates with corresponding p-values for each model coefficient.

Table 5.3: Parametric Model Results

	Dependent variable: Survival Time		
	Weibull	*Exponential*	*Lognormal*
potmths	0.044***	0.017***	0.025***
tyentry	-0.252***	-0.374***	-0.416***
Constant	2.156***	2.807***	2.281***
Observations	1,507	1,507	1,507
Log Likelihood	-2,441.332	-2,520.635	-2,469.079
χ^2 (df = 2)	181.395***	54.177***	96.663***
Note:			*p<0.1; **p<0.05; ***p<0.01

We obtain AIC values, respectively, of 5047.269, 4890.664, and 4946.158, weakly suggesting that the Weibull model is the best of the models fitted. This is visualised in figure 5.1, split by *tyentry* and with survival curves drawn for different values of *potmths*; the differences by category and by covariate value can be identified.

Separately, performing the test described in (5.13) for an exponential model, given we have a Weibull model, returns a test statistic of 158.6 and a p-value approximately equal to 1. This strongly indicaties the Weibull is a better model for the data.

We might test our model against a larger model to see if including further covariates gives a better model fit. Trying the initial Weibull model against a larger model including the responsibilities the individual has to their family, *famresp*, we can perform the one-way ANOVA test. This returns a p-value of 0.17, and so we do not reject the initial model at the .05 significance level.

It remains to assess the fit of our model. We conduct a graphical test of fit by plotting the Nelson-Aalen estimator against log-time (Figure 5.2). This appears to be linear, suggesting the model fits satisfactorily, though the final observation appears anomalous.

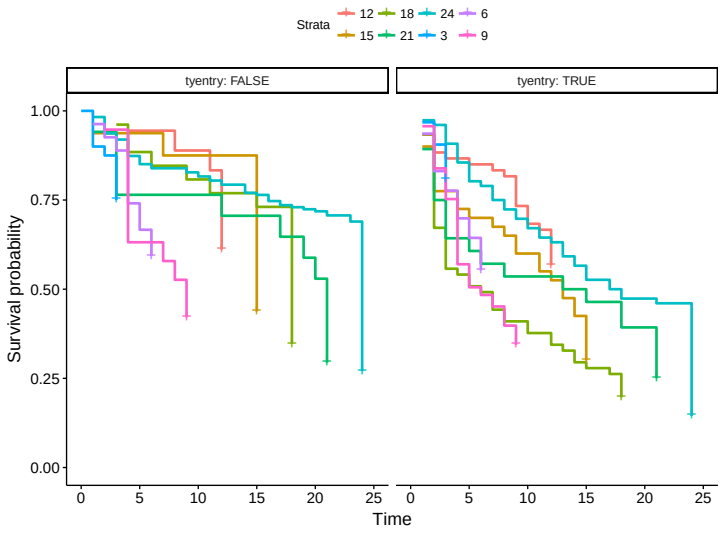

Figure 5.1: Weibull Model

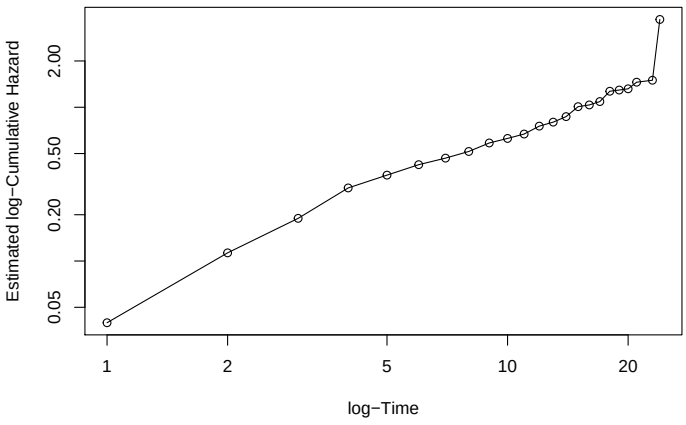

Figure 5.2: Diagnostic Plot for Weibull Model

Conclusion

This report set out to investigate how we might adapt, or indeed replace, well-established methods from generalised linear models when studying survival analysis. We have seen a range of contrasting approaches.

We started in chapter 2 by defining the key quantities and problems involved in survival analysis.

Over the next three chapters, we considered the three major families of model commonly used for survival analysis. Some of the techniques proposed were demonstrated with the dataset from the unemployment benefit study, with results from a few such studies recreated.

In chapter 3, we examined non-parametric models, focusing on the Kaplan-Meier product limit estimator. We proved that this has some desirable asymptotic properties, and saw how this permits comparison of groups within the data. An introduction to Bayesian methods showed an alternative to the PL estimator.

In chapter 4, we considered semi-parametric models, which comparatively relied on a small but strong set of assumptions. In particular, we studied the Cox proportional hazards model, reliant on the PH assumption, and briefly the additive model. We looked at conducting inference and estimating parameters in this setting, and how diagnosis takes place using different residuals and graphical methods. We found that making use of covariate data can improve both the fit of the model and our ability to conduct inference, at the expense of making more assumptions regarding the model.

The final major branch of models, proposed in chapter 5, were parametric models. We saw how these relate to GLMs through the accelerated lifetime structure. We considered estimation and diagnosis for models with errors following three parametric distributions, and saw these could be generalised as one distribution. Again, by supplying more information to the model we conducted better inference, but the assumptions we made to do so were very strong, perhaps too much so in situations with poor information.

In all cases, we observed methods both analytical and graphical for assessing goodness-of-fit. These can be used to the end of validating the assumptions made in each case, giving a coherent procedure for practical survival analysis.

Topics for Further Research

While this report has provided an overview of many of the major techniques and models for studying survival analysis, it is far from comprehensive. There are many more interesting topics which extend the theory to different problems, or answer existing questions in an alternative fashion.

Some such topics are: Modelling survival times as Markov (counting) processes, results relating to which have been shown [5, p. 80]; Multivariate survival analysis for simultaneous variables and large dimensions [5, p. 425]; The Buckley-James least-squares estimator; Random forest models; Recurrent event analysis [1, p. 332]; Modelling under dependent censoring or competing risks [2, p.177-179].

Moreover, survival analysis has been used in a wide range of applications. While these are of less explicitly mathematical, and hence not studied here in detail, they often motivate development and refinement of existing techniques, or synthesise ideas from other branches of mathematics. For one such example, in which the authors construct a market-beating survival analysis model for financial derivatives, see [7].

Bibliography

[1] Kleinbaum, D.G. and Klein, M. (1996). *Survival Analysis: A Self-Learning Text*, Springer; New York, NY.

[2] Miller, R.G. Jr. (1986), *Survival Analysis*, John Wiley and Sons; New York, NY.

[3] Lee, E.T. (1992), *Statistical Methods for Survival Data Analysis*, John Wiley and Sons; New York, NY.

[4] Dobson, A.J. (1990), *An Introduction to Generalized Linear Models*, Chapman and Hall

[5] Klein, J.P., Moeschberger M.L. (2003), *Survival analysis : Techniques for Censored and Truncated Data*, Springer-Verlag

[6] Guo, S. (2009), *Survival Analysis*, Oxford University Press

[7] Gao, G., Bu, Z., Liu, L., Cao, J., Wu, Z., *A Survival Analysis Method for Stock Market Prediction*, 2015 International Conference on Behavioral, Economic and Socio-cultural Computing, 2015

[8] Kaplan, E.L. and Meier, P., *Nonparametric Estimation from Incomplete Observations*, Journal of the American Statistical Association, Vol. 53, No. 282 (Jun., 1958)

[9] Peterson,*Expressing the Kaplan-Meier Estimator as a Function of Empirical Subsurvival Functions* Journal of the American Statistical Association, Vol. 72, No. 360 (Dec., 1977)

[10] Tsiatis, A. *A Large Sample Study of the Estimate for the Integrated Hazard Function in Cox's Regression Model for Survival Data* University of Winsconsin-Madison

[11] Jenkins, S., Garcia-Serrano, C., *The Relationship between Unemployment Benefits and Re-employment Probabilities: Evidence from Spain* Oxford Bulletin of Economics and Statistics, 2004, vol. 66, issue 2, 239-260

[12] Lin, D.Y., Ying, Z. *Semiparametric Analysis of the Additive Risk Model* Biometrika vol.81, No.1.

[13] Jorgensen, B., *Exponential Dispersion Models and Extensions: A Review* International Statistical Review, Vol. 60, No. 1 (Apr., 1992), pp. 5-20

[14] Engeman, Richard and Keefe, Thomas, *On Generalized Least Squares Estimation of the Weibull Distribution* Communications in Statistics-theory and Methods, volume 11

[15] Susarla, V., and Van Ryzin, J., *Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations* Journal of the American Statistical Association, Vol. 71, No. 356 (Dec., 1976), pp. 897-902

[16] John D. Kalbfleisch, *Non-Parametric Bayesian Analysis of Survival Time Data* Journal of the Royal Statistical Society. Series B (Methodological) Vol. 40, No. 2 (1978), pp. 214-221

[17] David K. Neal, *Goodness of Fit Tests for Normality* Mathematica in Education and Research Year: 1996 Volume: 5 Issue: 2 Page range: 23-30

[18] Muoz-Bulln, Fernando, Arranz, Jose and Alba, Alfonso, (2006), *Unemployment duration, unemployment benefits and recalls*, DEE - Working Papers. Business Economics. WB, Universidad Carlos III de Madrid. Departamento de Economa de la Empresa

[19] Nagore Garcia, A. and van Soest, Arthur, (2014), *Unemployment Transitions to Stable and Unstable Jobs Before and During the Crisis*, No 2014-026, Discussion Paper, Tilburg University, Center for Economic Research.

[20] Kim, Y., Lee, J., Bayesian Analysis of Proportional Hazard Models The Annals of Statistics 2003, Vol. 31, No. 2, 493511

[21] Steinsaltz, D. (2019) *Statistical Lifetime Models*, University of Oxford

[22] Zhang, D. (2005) *Analysis of Survival Data*, NCSU, https://www4.stat.ncsu.edu/ dzhang2/st745/

[23] https://sites.google.com/site/econometricsacademy/econometrics-models/survival-analysis

[24] Demeritt, J. *An Introduction to Duration Models* https://jdemeritt.weebly.com/uploads/2/2/7/7/22771764/introduction.pdf

[25] Morabia, Alfredo *Observations Made Upon the Bills of Mortality* BMJ 2013;346:e8640

[26] Hlavac, Marek (2018). *stargazer: Well-Formatted Regression and Summary Statistics Tables.* R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

[27] Terry M Therneau *survival-package, A Package for Survival Analysis in S* https://cran.r-project.org/web/packages/survival/index.html

[28] Alboukadel Kassambara *survminer* https://cran.r-project.org/web/packages/survminer/index.html