Influence diagnostics in regression with censored data

This item was submitted to Loughborough University’s Institutional Repository by the/an author.

Additional Information:

- A Doctoral Thesis. Submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy of Loughborough University.

Metadata Record: https://dspace.lboro.ac.uk/2134/11728

Publisher: © Isa Bin Daud

Please cite the published version.
This item was submitted to Loughborough University as a PhD thesis by the author and is made available in the Institutional Repository (https://dspace.lboro.ac.uk/) under the following Creative Commons Licence conditions.

For the full text of this licence, please go to:
http://creativecommons.org/licenses/by-nc-nd/2.5/
<table>
<thead>
<tr>
<th>AUTHOR/FILING TITLE</th>
<th>BIN DAUD, I</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACCESSION/COPY NO.</td>
<td>014427/02</td>
</tr>
<tr>
<td>VOL. NO.</td>
<td></td>
</tr>
<tr>
<td>CLASS MARK</td>
<td></td>
</tr>
<tr>
<td>LOAN COPY</td>
<td>Loan Copy</td>
</tr>
<tr>
<td>30 Jun 1989</td>
<td></td>
</tr>
<tr>
<td>30 Jun 1989</td>
<td></td>
</tr>
<tr>
<td>6 Jul 1992</td>
<td></td>
</tr>
<tr>
<td>2 Jul 1992</td>
<td></td>
</tr>
</tbody>
</table>
INFLUENCE DIAGNOSTICS IN REGRESSION

WITH CENSORED DATA

by

Isa Bin Daud

B.Sc. Hons. (Malaya), M.Sc. (Aston)

A Doctoral Thesis
Submitted in partial fulfilment of the
requirements for the award of Doctor of Philosophy
of the Loughborough University of Technology.


Supervisor: Dr. A.N. Pettitt,
Department of Mathematics.

To my wife Rosnah and our children,
Izah, Remi and Irma.
Acknowledgements

I would like to thank and express my gratitude to Dr. A.N. Pettitt for his constant guidance, invaluable help and suggestions during the course of this work. I am grateful to The University of Agriculture, Malaysia, and the Malaysian Government for granting me the Study leave and financial support.

My thanks also go to Mrs. B. Wright for her immaculate typing of this thesis.
Abstract

The work in this thesis is concerned with the development and extension of techniques for the assessment of influence diagnostics in data that include censored observations. Various regression models with censored data are presented and we concentrate on two models which are the accelerated failure time model, where the errors are generated by mixtures of normal distributions, and the Cox proportional hazards model. For the former, both finite discrete and continuous mixtures are considered, and an EM algorithm is used to determine measures of influence for each case.

For the Cox proportional hazards model, various approaches to approximating influence curves are investigated. One-step or few-step approximations are developed using an EM algorithm and compared with a Newton-Raphson approach. Cook's measures of local influence are also investigated for the detection of influential cases in the data.

The validity of the proportional hazards assumptions is also investigated. The residuals of Schoenfeld are examined for the possibility of being used to detect time dependence of the covariates in the proportional hazards model. Estimates to describe the nature of the time dependency computed from these residuals are presented.
Contents

CHAPTER 1: INTRODUCTION.

1.1 Overview and Definition of some Keywords.

1.1.1 Outliers.
1.1.2 Mixture models.
1.1.3 Censored data.
1.1.4 EM algorithm.
1.1.5 Influence function.
1.1.6 Proportional hazards model.

1.2 Scope of the Thesis.

CHAPTER 2: FINITE DISCRETE MIXTURE MODELS WITH CENSORED DATA.

2.1 Introduction.

2.2 Conditional Distribution of the Mixing Variable Q.

2.2.1 Conditional distribution for the failures.
2.2.2 Conditional distribution for the censored.
2.2.3 Conditional distribution of Q when Q is discrete.

2.3 Conditional Moments.

2.4 Estimation of the Parameters.

2.4.1 Expected value of the complete data likelihood.
2.4.2 Estimation of $\mu$.
2.4.3 Estimation of $\sigma^2$.
2.4.4 Estimation of the mixing proportion.
2.4.5 Estimation of the scale contamination parameter, k.

2.5 Linear Regression Models.

2.6 Marginal Log-Likelihood.

2.7 Summary of the Estimation Procedure.
2.8 Examples and Results.

2.8.1 Darwin data.
2.8.2 Forbes data.
2.8.3 Carcinogenesis data.
2.8.4 Motorette data.
2.8.5 Heart transplant data.
2.8.6 Conclusion.

2.9 Note on Influence and Robustness.

CHAPTER 3 : CONTINUOUS SCALE MIXTURE MODELS WITH CENSORED DATA - STUDENT-t ERROR DISTRIBUTIONS.

3.1 Introduction.

3.2 Marginal Distribution of Q for Student-t Error.

3.3 Posterior Density of Q and the Moments.

3.3.1 Summary of results of gamma distribution.

3.4 Estimation of the Parameters by EM.

3.4.1 Estimation of $\mu$ in single-sample model.
3.4.2 Estimation of the scale parameter $\sigma^2$.

3.5 Estimation of Regression Parameters for the Linear Model.

3.5.1 Observed information.
3.5.2 Finding outliers.

3.6 Newton-Raphson Method.

3.7 Assessment of Influence.

3.7.1 1-Step EM algorithm.
3.7.2 1-Step Newton-Raphson Method.
3.7.3 Comparison between 1-step EM and 1-step Newton-Raphson.
3.7.4 Computation of quadratic norm.
3.7.5 Computation of likelihood distance.
3.7.6 Mean shift outlier model.
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.8</td>
<td>Examples and Results.</td>
<td>99</td>
</tr>
<tr>
<td>3.8.1</td>
<td>Motorette data.</td>
<td>100</td>
</tr>
<tr>
<td>3.8.2</td>
<td>Stanford heart transplant data.</td>
<td>106</td>
</tr>
<tr>
<td>3.9</td>
<td>Conclusion.</td>
<td>112</td>
</tr>
<tr>
<td><strong>CHAPTER 4</strong>: PROPORTIONAL HAZARDS MODEL.</td>
<td></td>
<td>114</td>
</tr>
<tr>
<td>4.1</td>
<td>Introduction.</td>
<td>114</td>
</tr>
<tr>
<td>4.2</td>
<td>Likelihood Function.</td>
<td>116</td>
</tr>
<tr>
<td>4.2.1</td>
<td>Partial likelihood.</td>
<td>116</td>
</tr>
<tr>
<td>4.2.2</td>
<td>Rank marginal likelihood.</td>
<td>117</td>
</tr>
<tr>
<td>4.2.3</td>
<td>Adjustment for tied observations.</td>
<td>121</td>
</tr>
<tr>
<td>4.3</td>
<td>Rank Likelihood as Expectation from Exponential Distribution.</td>
<td>122</td>
</tr>
<tr>
<td>4.4</td>
<td>Estimation and Influence.</td>
<td>126</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Newton-Raphson iteration.</td>
<td>127</td>
</tr>
<tr>
<td>4.4.2</td>
<td>Proportional hazards as a linear model.</td>
<td>129</td>
</tr>
<tr>
<td>4.5</td>
<td>EM Algorithm Approach to Estimation.</td>
<td>130</td>
</tr>
<tr>
<td>4.5.1</td>
<td>Introduction.</td>
<td>130</td>
</tr>
<tr>
<td>4.5.2</td>
<td>Variance-covariance matrix from rank order likelihood.</td>
<td>133</td>
</tr>
<tr>
<td>4.5.3</td>
<td>Rank order expectation for exponential variate.</td>
<td>134</td>
</tr>
<tr>
<td>4.5.4</td>
<td>Poisson random error.</td>
<td>139</td>
</tr>
<tr>
<td>4.5.5</td>
<td>Gamma random error.</td>
<td>144</td>
</tr>
<tr>
<td>4.5.6</td>
<td>Comparison of poisson and gamma random errors.</td>
<td>150</td>
</tr>
<tr>
<td>4.6</td>
<td>Examples and Results.</td>
<td>155</td>
</tr>
<tr>
<td>4.6.1</td>
<td>Carcinogenesis data.</td>
<td>156</td>
</tr>
<tr>
<td>4.6.2</td>
<td>Veteran's administration lung cancer data.</td>
<td>157</td>
</tr>
<tr>
<td>4.6.3</td>
<td>Heart transplant data.</td>
<td>159</td>
</tr>
</tbody>
</table>
CHAPTER 5: DIAGNOSTIC AND INFLUENCE MEASURES.

5.1 Introduction. 163

5.2 Residual Analysis. 164

5.2.1 Lagakos's residuals. 164
5.2.2 Schoenfeld's residuals. 168

5.3 Case Deletion Techniques. 174

5.3.1 Introduction. 174
5.3.2 The "correct" 1-step approximation. 177
5.3.3 Illustration of the "correct" 1-step Newton-Raphson. 181
5.3.4 EM 1-step approximation. 185

5.4 Augmented Linear Predictor Approximation. 188

5.4.1 Newton-Raphson 1-step for the augmented model. 193
5.4.2 EM 1-step from augmented linear model. 208

5.5 Case-Weight Perturbation. 211

5.5.1 Properties of diagnostic $\Delta_k$. 216
5.5.2 Cook's likelihood displacement. 217
5.5.3 Local influence. 220
5.5.4 Case-weights for explanatory variables. 223

5.6 Review of Numerical Results. 227

5.6.1 Brake disc data. 232
5.6.2 Stanford heart transplant data. 239

CHAPTER 6: INVESTIGATING TIME DEPENDENT COVARIATES - SMOOTHING THE SCHOENFELD'S RESIDUALS. 248

6.1 Introduction. 248

6.2 Variance-Covariance of the Schoenfeld Residuals. 250

6.3 Omnibus Statistics. 251

6.4 Smoothing the Schoenfeld Residuals. 253

6.4.1 Constant $g(t)$ in window. 254
Chapter 1

Introduction

The fitting of models to data is an important part of statistical analysis. Model fitting is extremely useful, as it can lead to the understanding of the essential features of a data set. However, these models can only be an approximation of the truth describing the situation within the constraints of the modelling assumptions. Because of this inexactness, the study of stability and variation in the results of an analysis under a slight modification or perturbation of the modelling formulation becomes important. If a minor modification seriously affects the key results of the analysis, then there is cause to re-examine the model or the data set. On the other hand, if a negligible effect is observed, the sample is robust with respect to the induced perturbation, and the assumption of the model may be considered valid to describe the essential features of the data.

Although an assessment of the influence of a model perturbation is generally considered to be useful, few general methods are available for carrying out such an assessment other than in the context of the normal linear or generalized linear regression. For the latter, the assessment and diagnostics are well established (Belsley, Kuh & Welsh, 1980). For review of works by some authors on this subject, see Cook and Weisberg, 1982. However, the models considered by these authors do not generally include censored data. In recent years, authors, like Schoenfeld (1982), Cain and Lange (1984), Kay (1984) and Reid and Creapeau (1985), have developed
methods for the assessment and diagnostics of censored data in the Cox (1972) proportional hazards model.

The work in this thesis is mainly extending the concept of this modest perturbation to models which include censored data. Two models are investigated. They are the scale mixture models and the Cox (1972) proportional hazards model. The outcome of the analysis due to the perturbation is investigated through the influence curves (see Section 1.1.5) and plots of the appropriate residuals. These are the approaches which seem to have found wide acceptance because of their computational simplicity and their providing important diagnostic information.

For most of the work here, the perturbation schemes considered are those where the data is modified. These are case deletion schemes, particularly single case deletion, which has found greatest acceptance in the context of the linear regression model, and the scheme in which the weights attached to individuals or group of cases are modified (Cook, 1977, 1979). These schemes provide the basis for the detection of influential cases in the data set. An observation may be judged influential if important features of the analysis are altered substantially when the model is thus perturbed.

Additionally, in the proportional hazards model, we consider a perturbation scheme in the model formulation. The outcome of the analysis of this "misspecified" model is used to investigate the validity of the proposed model.

Before we outline the scope of this thesis, we shall present in the following subsections the definitions and brief overview
of some of the important keywords used in the thesis.

1.1 OVERVIEW AND DEFINITION OF SOME KEYWORDS.

1.1.1 Outliers.

An outlier in a set of data can be considered as an observation which appears to be inconsistent with the remainder of that set of data as to arouse suspicions that it was generated by a different mechanism. An inspection of a sample containing outliers would show up such characteristics as large gaps between "outlying" and "inlying" observations and the deviation between the outliers and the group of inliers, as measured on some suitably standardized scale. Should such observations be foreign to the main population, they may, by their very nature, cause difficulties for the observer in the attempt to represent the population. They can grossly contaminate estimates of parameters in some model for the population.

In the collection of data there are two basic ways which give rise to outliers. One is through error in collection or misrecording the observation and the other is by virtue of the data coming from some heavy tailed distribution.

The problem of outliers is one of the oldest in statistics, and many authors have proposed numerous statistics to use for the outlier tests. For detailed treatment on the subject, refer to Barnett and Lewis (1978) and Hawkins (1980); but in the work here we shall only be concerned with their detection through their influence on the inference process.
1.1.2 Mixture Models.

Mixture models have received a great deal of attention, and have important applications in engineering and in studies of mortality in demography. They can be envisaged as a device to separate a heterogeneous population into more homogenous sub-populations. The simplest form of a finite mixture model can be viewed as having distribution given by

\[(1-\lambda) F + \lambda G\]

where \(0 \leq \lambda \leq 1\) is the mixing parameter reflecting the chance that an observation arises from a distribution \(G\), rather than from the distribution \(F\). In the context of the outlier problem, the above model defines the alternative hypothesis to the working hypothesis that all the observations are from the distribution \(F\) (Box and Tiao, 1968).

In this thesis for the discrete mixtures, we shall only discuss the distributions of \(F\) and \(G\) as that of normal distributions with equal means but different variances. The model of this nature is sometimes called the symmetric mixtures or scale mixture model. Of course the model above can be generalised to include the infinite mixture situation. This can be presented (Barnett & Lewis, 1978) as

\[\int G(\lambda, x) \, dK(\lambda)\]

where \(K\) is a mixing distribution. For normal scale mixtures the error density is given by (see, for example, Pettitt, 1985)

\[f(\epsilon) = \int_{0}^{\infty} q^{\frac{1}{2}} \phi(q^{1}) p(q) \, dq\]
where the mixing random variable Q has density \( p(q) \) and \( \phi \) is the standard normal density. Here the conditional distribution of \( (c/Q = q) \) is normal with zero mean and variance \( q^{-1} \).

1.1.3 Censored Data.

The main feature and difficulty in the analysis of survival data is the possibility that some individuals may not be observed for the full time until failure. For example in a life-testing experiment in reliability, not all components have failed at the end of the study, and components that have not failed must have failure times exceeding the period of observation. Such an incomplete observation of the failure time is called censoring and data sets containing these observations recorded at the censoring point are called censored data.

A general definition (Cox and Oakes, 1984) of censoring is as follows:

Suppose that, in the absence of censoring, the \( i \)th individual in a sample of \( n \) has failure time \( T_i \), a random variable. We suppose also that there is a period of observation \( c_i \) such that observation on that individual ceases at \( c_i \) if failure has not occurred by then. Then the observations consist of

\[
X_i = \min(T_i, c_i)
\]

together with indicator variable \( \delta_i = 1 \) if \( T_i \leq c_i \) (uncensored), \( \delta_i = 0 \) if \( T_i > c_i \) (censored). \( c_i \) is referred to as realized censoring times for the censored individuals. In contrast, for individuals who are actually observed to fail, it is called the unrealized censoring times.

Various censoring schemes may be implemented to suit the
experimental design; but two main schemes are the so-called Type I and Type II censoring. In the former type all the \( c_i \) are equal, \( c_i = c \), a constant, which is under the control of the investigator; in the latter scheme observation ceases after a predetermined number of failures, so that \( c \) becomes a random variable. (More detailed account of censoring schemes and the applications are found in Lawless (1982), Elandt-Johnson & Johnson (1980) and Cox and Oakes (1984)). For this thesis we shall view the scheme with \( c_i \) to be some predetermined constant but often unknown to the investigator in advance, and that the unrealized \( c_i \) corresponding to the observed failures (uncensored) may never become known. The true life-time of the censored individual is thus only known to exceed \( c_i \) and we shall denote the corresponding observation as a right censored observation.

1.1.4 **EM Algorithm.**

This technique proposed by Dempster et al. (1977) provides the means of computing maximum likelihood estimates from incomplete data. The term incomplete data refers to observed data \( y \) a realization from a sample space \( Y \), such that there exists a sample space \( X \) with a many-one mapping from \( X \) to \( Y \) in which the \( x \) in \( X \) is only observed through the \( y \). The relationships between the sampling densities depending on the parameter \( \phi \) of the incomplete and complete data is given by the Dempster et al. (1977) equation (1.1), that is

\[
g(y/\phi) = \int_R f(x/\phi)dx.
\]

\( R \) is the region of integration for all values of \( x \) consistent
with the mapping of $X \rightarrow Y$.

Assuming the maximum likelihood estimate of $\phi$ satisfies

$$\frac{\partial}{\partial \phi} \log g(y/\phi) = 0$$

we note

$$\frac{\partial}{\partial \phi} \log g(y/\phi) = \int_R \frac{\partial}{\partial \phi} \log f(x/\phi) \cdot f(x/\phi) dx / \int_R f(x/\phi) dx$$

$$= E \left[ \frac{\partial}{\partial \phi} \log f(x/\phi) / y \right]$$

with expectation being defined conditionally with respect to all $x$'s giving a particular value of $y$.

The EM process refers to an algorithm which consists of an expectation or E-step followed by a maximization or M-step. Defining

$$Q(\phi'/\phi) = E \left[ \log f(x/\phi') / y, \phi \right]$$

which is assumed to exist for all pairs ($\phi', \phi$), Dempster et al. (1977) show that the process reduces to the following.

E-step : Compute $Q(\phi/\phi^{(p)})$

M-step : Choose $\phi^{(p+1)}$ to be the value of $\phi$ which maximizes $Q(\phi/\phi^{(p)})$

where the $(p)$ and $(p+1)$ define the current and the next iteration of the algorithm $\phi^{(p)} \rightarrow \phi^{(p+1)}$. The algorithm guarantees that

$$g(y/\phi^{(p+1)}) > g(y/\phi^{(p)})$$

and converges to a solution of

$$\frac{\partial}{\partial \phi} g(y/\phi) = 0$$

that is either a local maximum or a saddle point of $g(y/\phi)$. 
1.1.5 Influence Functions.

At the beginning of the chapter we have discussed briefly the concept of influence through some perturbation schemes. This is readily seen as the study of stability and reliability of the outcome in an analysis and the dependence on the assumed model. Here we shall present the main 'tool' for investigating and quantifying the changes as the result of the induced perturbation. This is the influence function, which was introduced by Hampel (1968, 1974) in order to investigate the infinitesimal behaviour of a real valued statistical functional.

Consider the contamination model of Section 1.1.2 that is

\[(1-\varepsilon) \, F + \varepsilon G\, ,\]

where G is the contaminating distribution giving mass 1 at point x, and F is the basic (working) model. Let \( T_n (x_1, x_2, \ldots, x_n) \) be an estimator based on an independent identically distributed sample \( X_1, X_2, \ldots, X_n \) from the distribution F. If a statistical functional T exists such that as \( n \to \infty \)

\[ T_n (x_1, x_2, \ldots, x_n) \to T(F) \, ,\]

in probability, then, the influence function (sometimes called influence curve) is defined as

\[ \text{IF}(x; T,F) = \lim_{\varepsilon \to 0} \frac{T[(1-\varepsilon)F + \varepsilon G] - T(F)}{\varepsilon} \quad (1.01) \]

It describes the effect of contamination at a point \( x \) on the estimate, standardized by the mass, \( \varepsilon \), of the contamination as \( n \to \infty \), that is it measures the asymptotic bias caused by the
contamination in the observations. The estimator would be called robust if small changes in $F$ did not produce wild fluctuations in the $T(F)$ and hence $IF(\cdot)$.

We will not go into the details of this asymptotic influence function which is mainly a heuristic tool. For more exposition and a rigorous treatment, see Hampel (1971, 1974), Huber (1981), Serfling (1981) and Hampel et al. (1986). Two important properties of the influence function, however, need to be mentioned here. They are

$$
\int IF(x; T,F) \, d\mathbb{P}(x) = 0 \quad , \quad (1.02)
$$

that is the expectation of the $IF(\cdot)$ with respect to the variation of $x$ according to the basic model $F$, is zero and

$$
\int [IF(x; T,F)]^2 \, dF(x) = V(T) \quad (1.03)
$$

where $V(T)$ is the asymptotic variance of $T$. Other summaries which throw light on robustness include:

a) gross-error sensitivity, $\gamma$, defined by

$$
\gamma = \sup_{x} |IF(x; T,F)| \quad (1.04)
$$

the supremum being taken over all $x$ where $IF(x; T,F)$ exists. It measures the worst approximate influence which a fixed amount of contamination can have on the value of the estimator.

b) The local-shift sensitivity, $\lambda$, defined by

$$
\lambda = \sup_{x \neq y} \frac{|IF(x; T,F) - IF(y; T,F)|}{|y-x|} \quad (1.05)
$$
It measures the worst possible effect of shifting an observation slightly from the point \( x \) to some neighbouring point \( y \).

c) The rejection point, \( \rho \), defined by

\[
\rho = \inf \left\{ r > 0 \ ; \ IF(x; T, F) = 0 \text{ when } |x| > r \right\}.
\] (1.06)

Here observations further away than \( \rho \) are rejected completely.

The influence function described above is entirely asymptotic, focusing on the limit of \( T_n(x_1, x_2, \ldots, x_n) \) as \( n \to \infty \); and deriving it with respect to a theoretical sampling population \( F \). In finite sampling situations, the finite sample versions which convey the same information as the infinite version, are used. Several have been suggested of which two, the empirical influence (EIC) and the sample influence curve (SIC) have received greatest attention (Cook and Weisberg, 1982, Chapter 3). Generally the EIC is obtained by substituting the sample cumulative density function \( \hat{F} \) for \( F \) in the influence function (1.01). Likewise the SIC is obtained using \( \hat{F} \), but taking the \( \epsilon \) in (1.01) to be equal to \(-\frac{1}{n-1}\). For linear models with \( e_i = y_i - x_i^T \beta \) (Cook and Weisberg, 1982) we have for the \( i^{th} \) case, empirical vector measures given by

\[
\text{EIC}_i = n(X'X)^{-1} x_i e_i
\] (1.07)

and

\[
\text{SIC}_i = (n-1)(\hat{\beta} - \hat{\beta}_{(i)})
\] (1.08)

\[
= (n-1)(X'X)^{-1} x_i e_i / (1 - v_{ii})
\]

where \( X \) is the design matrix with \( i^{th} \) row \( x_i^T \), \( \hat{\beta}_{(i)} \) is the estimate of \( \beta \) when the \( i^{th} \) case is omitted from the data set,
and $v_{ii}$ is the $i$th diagonal component of $X(X^TX)^{-1}X^T$.

These influence curves, as they are usually called, are readily computed and are proportional to the change in the estimates of the regression parameters when a case is deleted from the sample, thus providing a good measure to detect influential cases.

1.1.6 Proportional Hazards Model.

This model is due to Cox (1972) concerning the analysis of censored failure times using regression model. It is assumed that for each individual there are available values of one or more explanatory variables, $z$. The hazard function, which defines the instantaneous failure rate, is taken to be a function of these explanatory variables and unknown regression coefficients, $\beta$.

The hazard function for an individual, $i$ say, at failure time $T = t$ is given by

$$h_i(t, z) = h_0(t) \exp(z_i^T \beta)$$

where $h_0(t)$ is some arbitrary base line hazard function, usually unknown, corresponding to case $z = 0$.

A more detailed discussion of this model and references of past work is given in Chapter 4.

1.2 SCOPE OF THE THESIS.

This thesis is in two parts. The first part, Chapters 2 and 3, deals with life-time survival models in which the errors are generated by mixtures of normal distributions. Finite discrete and continuous mixture models are considered. The EM algorithm
is used to estimate relevant parameters and this reduces to reweighted least squares which can easily be programmed in GLIM (Baker and Nelder, 1978) or a similar package. The assessment of influence for these models is discussed in Chapter 3 which generalizes the method to include models dealt with in Chapter 2.

The second part is concerned with the proportional hazards model. The likelihood functions are viewed from the semi-parametric approaches of Cox's (1972) partial likelihood and Kalbfleisch and Prentice's (1973) rank marginal likelihood. The EM algorithm is employed to solve the marginal likelihood function for the regression parameter estimates. Here we declare the error in the resulting reweighted least squares to be a gamma distribution.

The techniques of Clayton and Cuzick (1985a), using the EM algorithm, and the Newton-Raphson method for obtaining maximum likelihood estimates are compared by applying them to several data sets.

In Chapter 5 we present a detailed study of influence diagnostic for the proportional hazards model. Reviews and extensions of the existing methods are also described. Various diagnostic aids are examined. In the assessment of influence, the influence curves are of great importance. Necessarily this involves the computation of $\beta - \hat{\beta}(i)$ for each observation, $i = 1, 2, \ldots, n$, where $\beta$ and $\hat{\beta}(i)$ are the estimate of the regression parameter, $\beta$, for the full and when the $i$th observation is omitted from the data set, respectively. This proves to be quite cumbersome especially when $n$ is large, since in the non-linear regression model these estimates are arrived at by an iterative process. In this chapter various approaches to
obtaining the approximation are explored. They include a 1-step EM method, a consequence of the EM algorithm with gamma errors employed for the estimation of the regression parameters in the proportional hazards model. To quantify the influence curve, the likelihood distance measure is investigated for the different approaches. The idea of local influence, as proposed by Cook (1986), is extended to the proportional hazards model, with the relevant quantities derived from the approximation of the influence curve. Whenever appropriate, scatter plots or graphs are used to facilitate detection of influential cases.

In addition to the detection of influential cases in a data set, the assumptions of the proportional hazards and the time constant effects of covariates are investigated. Two types of residual for the proportional hazards model, the Lagakos's (1981) and Schoenfeld's (1982) residuals, are employed to examine the validity of the model. These residuals are discussed in some detail in the beginning of Chapter 5 with explanation of how the plots may be used. The basic formulation from which the Schoenfeld's residuals are arrived at makes them important statistics to investigate the time dependency of the covariates. In Chapter 6 we pursue this investigation and propose a number of estimates to define the nature of the time-dependency. Real and simulated data are analysed accordingly to illustrate the procedures. The nature of the time-dependency is observed from the scatter plots of these estimates.

Finally, Chapter 7 summarises the conclusions arrived at in previous chapters and presents some suggestions for future research.
Chapter 2

Finite Discrete Mixture Models
With Censored Data

2.1 INTRODUCTION.

Mixture models have received a great deal of attention in statistics with the underlying motivation of separating heterogeneous population into more homogeneous subpopulations. Applied to lifetime data, the mixture models approach also provides a smooth estimate of survival function (Jewell, 1982) which may be preferable to the standard non-parametric step function estimates (Kaplan and Meier, 1958). Many authors have dealt with these models. Mandenhall and Hader (1958), Koa (1959) and Rider (1961) have dealt with the finite mixture models, and Doyle et al. (1980) presented properties for infinite mixtures of lifetime data from exponential distribution. Simar (1976) and Hill et al. (1980) consider the maximum likelihood estimation while Laird (1978) obtains some general results about the existence of maximum likelihood estimates of arbitrary mixing measures. Aitkin and Wilson (1980) implement the EM algorithm to determine the mixing probabilities \( p_j, \ j \geq 1 \) and the regression parameters when the data have densities \( g = \sum_{j=1}^{m} p_j f_j \), for normal densities \( f_j \). Their data do not include censored observations. Day (1969) developed an iterative method for the solution of the likelihood equations generated by \( g \), in the more general case of multivariate normal mixtures with common covariance matrix for \( m = 2 \). Jewell (1982) considers the mixture of non-parametric exponential and Weibull distributions for lifetime models, giving
a practical algorithm for computing the maximum likelihood estimates.

In this chapter we shall discuss only the finite scale mixture of normal distributions. We consider $Y = \mu + \sigma \varepsilon$ of which the errors, $\varepsilon$ are deemed to be generated by the scale mixtures of normal distributions with scale variable $Q$ taking finite values, that is, the conditional distribution of $\varepsilon$ given $Q = q$ is normal with mean zero and variance $q^{-1}$. The marginal density of $\varepsilon$ is thus

$$f(\varepsilon) = \sum q^\frac{1}{2} \phi(q^{\frac{1}{2}}) p(q)$$

where $p(q)$ is the marginal probability distribution of $Q$ and the summation is over the number of components, i.e. the number of finite value $Q$ takes. The random variable $Y$ may be subject to censoring, that is for some known value $y$, it is known only that $Y < y$, (left-censoring) or $Y > y$ (right-censoring).

The EM algorithm is implemented for the models with right censoring, (it can be easily extended for grouped or left censored data). It deals with the incomplete data as defined by Dempster et al. (1977). Although censored observations fall within the definition of incomplete data, however in this model the incompleteness is also due to the unobserved scale variable $Q$. From the complete data densities of $(Y, Q)$ the EM algorithm then produces iterative steps that lead to iterative reweighted least squares. The algorithm can then be easily programmed in GLIM (Baker and Nelder, 1973).

2.2 **CONDITIONAL DISTRIBUTION OF THE MIXING VARIABLE, $Q$.**

Consider a random variable $Y$ subjected to scale mixture random
variable $Q$ such that $(Y/Q = q)$ is distributed normally with mean $\mu$ and variance $\sigma^2 q^{-1}$. Note here $Q$ could be discrete or continuous, but for the present argument we shall assume $Q$ to be continuous with density $p(q)$. Further consider for given $Q = q$

$$Y = \mu + \sigma \varepsilon$$

or

$$\varepsilon = \left( \frac{Y - \mu}{\sigma} \right)$$

(2.01)

then the error, $\varepsilon$ is distributed normally with mean zero and variance $q^{-1}$, i.e. $N(0, q^{-1})$. The conditional distribution of $Q$ given $\varepsilon$ (or $Y$), whether $\varepsilon$ is observed, or censored is of great importance in the subsequent analysis of the mixture model. We shall denote this distribution as $p(q/E = r, \mu, \sigma)$ and $p(q/E > r, \mu, \sigma)$ for the failure and censored observations respectively, where $r$ is $(y - \mu)/\sigma$, with $y$ the realization of $Y$.

2.2.1 Conditional Distribution for the Failures.

For the failures we observed that $\varepsilon = r$. Let $f_{\varepsilon, Q}(\varepsilon, q)$ be the joint distribution of $\varepsilon$ and $Q$, then

$$f_{\varepsilon, Q}(\varepsilon, q) = f_1(\varepsilon/Q = q) \cdot p(q)$$

(2.02)

where $f_1(\varepsilon/Q = q)$ is the conditional density of $\varepsilon$ given $Q$ and is $N(0, q^{-1})$, that is

$$f_1(\varepsilon/Q = q) = \frac{1}{\sqrt{2\pi q^{-1}}} \exp \left( -\frac{1}{2} \frac{\varepsilon^2}{q^{-1}} \right)$$

$$= \frac{1}{\sqrt{2\pi}} \exp \left( -\frac{1}{2} q \varepsilon^2 \right)$$

$$= q^{\frac{1}{2}} \phi(q^{-\frac{1}{2}} \varepsilon)$$

(2.03)
where $\phi$ is the standard normal density.

Also we note that the conditional density of $Q$ given $\varepsilon = r$ is given by

$$p(q/\varepsilon = r) = \frac{f_{\varepsilon,Q}(r,q)}{f_{\varepsilon}(r)}$$

where $f_{\varepsilon}(r)$ is the marginal density of $\varepsilon$ evaluated at $r = \frac{Y - \mu}{\sigma}$ and thus a constant. Substituting for $f_{\varepsilon,Q}(\varepsilon,q)$ in (2.02) and using the notation in the section above, we have

$$p(q/\varepsilon = r, \mu, \sigma) \propto q^{\frac{1}{2}} \phi(rq^{\frac{1}{2}}) p(q). \quad (2.04)$$

2.2.2 Conditional Distribution for the Censored.

If $Y$ is right censored then we have observed $Y > y$ and $y$ is known (or $\varepsilon > r$). For given $Q = q$, the probability of this happening is

$$pr(\varepsilon > r/Q = q) = \int_r^\infty f_1(\varepsilon/Q = q) \, d\varepsilon$$

$$= 1 - \int_{-\infty}^r q^{\frac{1}{2}} \phi(q^{\frac{1}{2}} \varepsilon) \, d\varepsilon$$

$$= 1 - \phi(rq^{\frac{1}{2}})$$

where $\phi$ is the standard normal distribution function. Now

$$p(q/\varepsilon > r) = \frac{p(q, \varepsilon > r)}{p(\varepsilon > r)}$$

$$= \frac{pr(\varepsilon > r/Q = q)p(q)}{\int pr(\varepsilon > r/Q = q)p(q) \, dq}$$

where the denominator is the normalizing constant. Hence
2.2.3 Conditional Distribution of $Q$ when $Q$ is Discrete.

Let $m$ be the number of components in the model. For a discrete mixture random variable, let $Q^{-1}$ have probability over $k_j$ and let $p(Q = \frac{1}{k_j}) = \gamma_j$. Denoting the conditional distribution of $Q = \frac{1}{k_j}$, given the $i$th data as $p_{ij}$, then from (2.04) and (2.05) we have

$$p_{ij} = \begin{cases} \gamma_j \frac{1}{k_j} \phi(r_i, k_j^{-\frac{1}{2}}) & \text{if ith case a failure} \\ \gamma_j \left[1 - \phi(r_i, k_j^{-\frac{1}{2}})\right] & \text{if the ith case censored} \end{cases}$$

(2.06)

The proportionality constant may be obtained by noting that

$$\sum_{j=1}^{m} p_{ij} = 1 \quad \forall \ i = 1,2,\ldots,n$$

where $n$ is the number of observations, censored or otherwise.

2.3 CONDITIONAL MOMENTS.

As we shall see later in section 2.4, the EM algorithm approach requires the maximization of the expectation of the complete data set-up likelihood; the expectation being with respect to the joint distribution of $Y$ and $Q$ given $Y$, whether $Y$ is observed failures or censored. Moments for terms like $Q^k Y^k$, $Q^k$ or $\varepsilon^k$ are then useful in the computation of the maximum likelihood. This may be arrived at without difficulty from moments for $Y^k$ (or $\varepsilon^k$) conditional on $Q$ by the relationship

$$p(q|c > r) = \left[1 - \phi(rq^{\frac{1}{2}})\right]p(q).$$

(2.05)
In view of $c = (y - p)/a$ is $N(0, q^{-1})$ for given $Q = q$, we choose to present here the expressions for the moments of $c^k$ rather than $Y^k$ for simplicity.

For failure observations, that is when $Y = y$, the expectation is straightforward the normal variate moments, and

$$E(c^k/Q = q, \varepsilon = r) = r^k$$

(2.08)

For the censored observations however, following the same arguments presented in section (2.2.2) with $r = (y - \mu)/\sigma$

$$E[c^k/Q=q, \varepsilon > r] = \frac{\int_{r}^{\infty} c^k q^{\frac{1}{2}} \phi(c q^{\frac{1}{2}}) dc}{1 - \phi(r q^{\frac{1}{2}})}$$

(2.09)

Note that we have dropped the subscript of $E$ here, but there should be no confusion to its meaning in its presentation above.

Substituting $z = cq^{\frac{1}{2}}$ in equation (2.09), then the integration part becomes

$$q^{-k/2} \int_{rq^{\frac{1}{2}}}^{\infty} z^k \phi(z) dz$$

Putting $I(k,x) = \int_{x}^{\infty} z^k \phi(z) dz$, we have a reduction formula for integer $k$

$$I(k,x) = x^{k-1} \phi(x) + (k-1) I(k-2,x)$$

with $I(0,x) = \int_{x}^{\infty} \phi(z) dz = 1 - \phi(x)$
and \( I(1,x) = \int_x^\infty z\phi(z) \, dz = \phi(x) \). \hspace{1cm} (2.10)

Equation (2.09) can now be written as:

\[
E[\epsilon^k/Q=q, \epsilon > r] = q^{-k/2} I(k,rq^1/2) \left[ 1 - (rq^1) \right]^{-1}. \hspace{1cm} (2.11)
\]

2.4 ESTIMATION OF THE PARAMETERS.

The estimations of the parameters, (including the mixture proportion, \( \gamma_j \), of each components in the mixture) is primarily maximum likelihood. Day (1969) developed an iterative method for the solutions of the likelihood equations in a more general case of multivariate normal mixtures with common covariance matrix for two components, but without the presence of censoring in the data set. This was shown by Dempster et al. (1977) to be a particular case of the EM algorithm. Aitkin and Wilson (1980) apply the EM algorithm to finite normal mixtures in both single case and regression problems for uncensored data. Their scheme for the regression model is that the regression model applies to one component and the rest of the components are single sample problems with different mean. The latter constitute the "outliers". Here we shall use a similar approach, and employ it to find the maximum likelihood estimates of the parameters in finite scale mixture models (with constant mean) applied to censored data.

To recapitulate the underlying principle of the EM algorithm (Chapter 1), relating to our problem here, there are two fundamental steps to proceed.

a) The E-step: compute

\[
Q_e(\tilde{\theta}/\tilde{\theta}) = E \left[ L_o(\tilde{\theta})/\tilde{\theta} \right]
\]
b) The M-step: Determine \( \theta_{p+1} \) as a new value of \( \theta \) which maximizes \( Q_e(\theta / \theta_p) \)

where \( L_o(\theta) \) is the log likelihood for the complete data set-up (see Dempster et al. (1977) for further definition). From here on we shall use \( L_o \) to represent this likelihood. Also note here that the expectation in step (a) is with respect to the joint distribution of \( Q \) and \( Y \) given the data and the parameter \( \theta_p \).

The unobserved \((Y_i, Q_i)\), \( i = 1, 2, \ldots, n \), where \( n \) is the number of observations censored or otherwise, thus will make up for the complete data set-up. For finite mixtures we assume \( Q_i \) is discrete over \( \frac{1}{k_j} \), \( j = 1, 2, \ldots, m \), say, and \( m \) is the number of components in the model. The joint probability density for the complete data when \( Q = \frac{1}{k_j} \) is given by

\[
p(y, \frac{1}{k_j}) = f_j(y) p(\frac{1}{k_j})
\]

where \( f_j(y) \) is the conditional density of \( Y \) given \( Q \), which is, by our assumption, normally distributed with mean \( \mu \) and variance \( \sigma^2 q^{-1} \), that is

\[
f_j(y) = \frac{1}{\sigma \sqrt{2\pi k_j}} \exp\left[-\frac{1}{2} k_j \left(\frac{y - \mu}{\sigma} \right)^2\right]
\]

The marginal distribution of \( y \) is then

\[
p(y) = \sum_{j=1}^{m} f_j(y) \gamma_j
\]

where \( \gamma_j = p(Q = \frac{1}{k_j}) \) as defined in section (2.2.3), is the
mixing proportion. The complete data log likelihood from (2.12) is thus

\[
L_o = \sum_{i=1}^{n} \left\{ \log \left[ f_{j}(y_i) \right] + \log(\gamma_j) \right\}
\]

\[
= -n \log(2\pi \tilde{\sigma}^2) - \frac{n}{2} \log \tilde{k}_j - \frac{1}{2\tilde{\sigma}^2} \sum_{i=1}^{n} (y_i - \tilde{\mu})^2 + n \log(\gamma_j) .
\]

(2.15)

The "-" is to indicate that the arguments capped thus are of the complete log likelihood \(L_o\), as opposed to \(\theta\) which is given to evaluate the conditional expectation. The parameters we wish to estimate apart from \(\mu\) and \(\sigma^2\) are \(k_j\) and \(\gamma_j\) \((j = 1, \ldots, m)\) since we assume in our discrete model \(q_i\) and \(p(q_i)\) are unknown, and need be estimated concurrently in the algorithm.

Let \(L_i\) be the complete log likelihood for observation \(i\), whether censored or otherwise. Then \(L_o = \sum_{i=1}^{n} L_i\). The expectation step of the EM algorithm is

\[
Q_{E} \left( \tilde{\theta} / \tilde{\theta} \right) = E_{Y,Q}(L_o / \tilde{\theta})
\]

\[
= \sum_{i=1}^{n} E_{Y,Q} \left[ L_i / \tilde{\theta} \right] .
\]

(2.16)

Now for the uncensored, we observe \(Y_i = y_i\), and from (2.07)

\[
E_{Y,Q}[L_i] = E_{Q} \left[ E_{Y,Q,Y=y}(L_i) \right]
\]

\[
= E_{Q}[L_i]
\]

since \(E_{Y/Q,Y=y}(L_i) = L_i\). For the uncensored we still need to
evaluate $E_{Y/Q, Y=y} [L_i]$, so that we can express the expectation to that with respect to the conditional distribution of $Q/Y$, which is easily manageable.

The M-step is to maximize $Q_e (\hat{\alpha}/\theta_p)$ with respect to $\hat{\alpha}$. The solutions to the normal equations provide the next iterate $\theta_{p+1}$. These steps are repeated until convergence is achieved. The convergence criterion may be of several alternatives, for example, the maximum likelihood, or the difference between the updated and the previous iterates. We choose the latter in all our analysis since the likelihood criterion is slow for practical purposes. This may be due to the surface being almost flat at its maximum.

2.4.1 Expected Value of the Complete Data Likelihood.

From (2.13) the log likelihood for observation $i$ is

$$L_i = - \log(2\pi \sigma^2) - \frac{1}{2} \log k_j - \frac{(y_i - \mu)^2}{2\sigma^2 k_j} + \log(\gamma_j).$$

As was discussed in the preceding section, for uncensored $i$, the expectation in the E-step, is equivalent to the conditional expectation of $Q$. Defining the density $p(Q_i = \frac{1}{k_j} / Y_i) = p_{ij}$ as in section (2.2.3) then

$$E \left[ L_i / \theta_p \right] = \sum_{j=1}^{m} L_i p_{ij} \text{ (2.17)}$$

where $p_{ij}$ is as in (2.06) evaluated at $\theta_p$. Since $\sum_{j=1}^{m} p_{ij} = 1$ for $i = 1, 2, \ldots, n$, (2.17) may be written as
\[
E\left[\frac{L_i}{\theta^p}\right] = -\log(2\pi \sigma^2) - \frac{1}{2} \sum_{j=1}^{m} \log(\tilde{k}_j) P_{ij}
\]

\[
- \frac{1}{2} \left(\frac{y_i - \mu}{\sigma}\right)^2 \sum_{j=1}^{m} \frac{P_{ij}}{k_j} + \sum_{j=1}^{m} \log(\tilde{y}_j) P_{ij} .
\]

(2.18)

For a censored observation we first need to find the expectation with respect to the distribution of \((Y/Q, Y > y)\). This may be found as follows

\[
E_{Y/Q}\left[\frac{L_i}{\theta^p}\right] = -\log(2\pi \sigma^2) - \frac{1}{2} \log(\tilde{k}_j)
\]

\[
- \frac{1}{2\sigma^2 \tilde{k}_j} E_{Y/Q}\left[(y_i - \tilde{\mu})^2\right] + \log(\tilde{y}_j) .
\]

(2.19)

Consider the term \(E_{Y/Q}\left[(y_i - \tilde{\mu})^2\right]\). This can be written as

\[
E_{Y/Q}\left[\sigma^2 \left(\frac{y_i - \mu}{\sigma}\right)^2 + 2\sigma (\mu - \tilde{\mu}) \left(\frac{y_i - \mu}{\sigma}\right) + (\mu - \tilde{\mu})^2\right]
\]

\[
= \sigma^2 E_{Y/Q}(\epsilon^2) + 2\sigma (\mu - \tilde{\mu}) E(\epsilon) + (\mu - \tilde{\mu})^2
\]

where \(\mu\) is the mean used to calculate the expectation (given at a particular iteration). From (2.11) and (2.10), with \(q = \frac{1}{\tilde{k}_j}\), and dropping the subscript of \(E\),

\[
E\left[\epsilon/\epsilon > r, Q = \frac{1}{\tilde{k}_j}\right] = k_j^\frac{1}{2} \phi(r\tilde{k}_j^{-\frac{1}{2}}) \left[1 - \phi(r\tilde{k}_j^{-\frac{1}{2}})\right]^{-1}
\]

(2.20)

and

\[
E\left[\epsilon^2/\epsilon > r, Q = \frac{1}{\tilde{k}_j}\right] = k_j I(2, r\tilde{k}_j^{-\frac{1}{2}}) \left[1 - \phi(r\tilde{k}_j^{-\frac{1}{2}})\right]^{-1}
\]

\[
= k_j \left\{1 + \frac{r\tilde{k}_j^{-\frac{1}{2}}}{1 - \phi(r\tilde{k}_j^{-\frac{1}{2}})}\right\}
\]

(2.21)

Let \(B_{ij} = k_j^\frac{1}{2} \phi(r\tilde{k}_j^{-\frac{1}{2}}) \left[1 - \phi(r\tilde{k}_j^{-\frac{1}{2}})\right]^{-1} = E\left[\epsilon_{i}/\epsilon_{i} > r_{i}, Q_{i} = \frac{1}{\tilde{k}_j}\right]\)
and

\[ A_{ij} = E\left[ \frac{e_i^2}{e_i} > r_i, \frac{Q_i}{r_i} = \frac{1}{k_j} \right] \]

then

\[ A_{ij} = r_i B_{ij} + k_j . \quad (2.22) \]

(2.19) can now be written as

\[
\hat{E}_{\theta / \theta} \left[ L_i / \theta \right] = - \log(2\pi \sigma^2) - \frac{1}{2} \log(k_j) - \frac{1}{2} \sum_{i=1}^m \left( \sigma^2 A_{ij} + 2\sigma (\mu - \tilde{\mu}) B_{ij} + (\mu - \tilde{\mu})^2 \right) + \log(\hat{\gamma}_j) ,
\]

and the expectation of the above with respect to the conditional
distribution of Q given \( Y > y \) (see section 2.2.3) is

\[
E_{\theta / \theta} \left[ L_i / \theta \right] = E_Q \left\{ E_{\theta / \theta} \left[ L_i / \theta \right] \right\} = \sum_{j=1}^m p_{ij} E_{\theta / \theta} \left[ L_i / \theta \right] = - \log(2\pi \sigma^2) - \frac{1}{2} \sum_{j=1}^m \log(k_j) p_{ij} - \frac{1}{2} \sum_{j=1}^m \left( \sigma^2 A_{ij} + 2\sigma (\mu - \tilde{\mu}) B_{ij} + (\mu - \tilde{\mu})^2 \right) \frac{p_{ij}}{k_j} + \sum_{j=1}^m \log(\hat{\gamma}_j) p_{ij} . \quad (2.23)
\]

Note here, for the censored observation \( p_{ij} = \gamma_j \left[ 1 - \Phi(r_i/k_j) \right] \).

Then from (2.18) and (2.23) the expected value of the complete
data likelihood of (2.16) is

\[
Q_{\theta / \theta} = - \frac{n}{2} \log(2\pi \sigma^2) - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^m \log(k_j) p_{ij} + \sum_{i=1}^n \sum_{j=1}^m \log(\gamma_j) p_{ij} - \frac{1}{2} \sum_{i \in I_u} \sum_{j=1}^m \frac{(y_i - \tilde{\mu})^2}{k_j} p_{ij} + \sum_{i \in EC} \sum_{j=1}^m \psi_{ij} \frac{p_{ij}}{k_j} .
\]

\[ (2.24) \]
where

\[ \psi_{ij} = \sigma^2 A_{ij} + 2\sigma(\mu - \tilde{\mu}) B_{ij} + (\mu - \tilde{\mu})^2, \]

and \( p_{ij} \) given as in (2.06) and rewritten here for easy of reference.

\[
p_{ij} = \begin{cases} 
\gamma_j k_j^{-\frac{1}{2}} \phi(r_i k_j^{-\frac{1}{2}}) & \text{if ith case uncensored} \\
\gamma_j \left[1 - \phi(r_i k_j^{-\frac{1}{2}})\right] & \text{and } \sum_{j} m p_{ij} = 1 .
\end{cases}
\]

U and C in the summation refer to sets of uncensored and censored observations respectively.

2.4.2 Estimation of \( \mu \).

To obtain the next iterate of \( \mu \), say \( \mu_{p+1} \), in line with the M-step of the EM algorithm, (2.24) is differentiated with respect to \( \tilde{\mu} \), and the solution to the normal equation provides the next iterate. The normal equation for the estimation of \( \mu \) when all the observations are uncensored is

\[
\sigma^{-2} \sum_{i \in U} \frac{m}{j=1} (y_{ij} - \tilde{\mu}) \left(\frac{p_{ij}}{k_j}\right) = 0 \tag{2.25}
\]

let \( w_i = \sum_{j=1}^{m} \left(\frac{1}{k_j}\right) p_{ij} \),

then with \( q = \frac{1}{k} \) and \( p_{ij} \) defined in (2.06), \( w_i \) is equivalent to \( E(Q_{ij}) \), where the expectation is with respect to the conditional distribution of \( Q \) given the data. The solution to (2.25) is

\[
\tilde{\mu} = \frac{\Sigma y_i w_i}{\Sigma w_i} . \tag{2.26}
\]
When censoring is present, the normal equation from (2.24) is

\[ \hat{\sigma}^{-2} \left[ \sum_{i \in U} \sum_{j=1}^{m} (y_i - \tilde{\mu}) \frac{p_{ij}}{k_j} + \sum_{i \in C} (\sigma B_{ij} + \mu - \tilde{\mu}) \frac{p_{ij}}{k_j} \right] = 0 \]

Let

\[ B_i^* = \frac{\sum_{j=1}^{m} (B_{ij} \frac{p_{ij}}{k_j})}{\sum_{j=1}^{m} (p_{ij} / k_j)} \]

then with \( w_i \) defined as above, the censored summation in the normal equation becomes

\[ \sum_{i \in C} \left[ (\sigma B_i^* + \mu) - \tilde{\mu} \right] w_i \]

and substituting \( y_i^* \) such that

\[ y_i^* = \begin{cases} y_i & \text{if ith observation uncensored} \\ \sigma B_i^* + \mu & \text{if ith observation censored} \end{cases} \]

the normal equation is reduced to

\[ \hat{\sigma}^{-2} \left[ \sum_{i=1}^{n} (y_i^* - \tilde{\mu}) w_i \right] = 0 \]

which gives the solution

\[ \hat{\mu} = \frac{\sum_{i=1}^{n} y_i^* w_i}{\sum_{i=1}^{n} w_i} \]

(2.27)

This provides the iterative scheme in updating \( \mu \) from the
previous, which can be programmed simply in GLIM (see Baker and Nelder, 1978, for details of GLIM). However the computation of \( w_i \) requires the updated value of \( \tilde{k}_j \) (for the unknown distribution of \( Q \)). As will be seen in section (2.45), the next iterate for \( \tilde{k}_j \) also requires an updated value of \( \mu \). Two options may be employed here. One is to compute (2.27) using the values of \( \tilde{k}_j \) from the previous iteration, or alternatively we may iterate \( \tilde{k}_j \) (in section 2.4.5) for given \( \mu \), first before updating \( \mu \) from (2.25).

2.4.3 Estimation of \( \sigma^2 \).

As in the preceding section differentiating with respect to \( \sigma^2 \) equation (2.24) gives the normal equation for estimating \( \sigma^2 \) as

\[
- \frac{n}{2} \frac{\sigma^{-2}}{\sigma^{-4}} + \frac{1}{2} \frac{\sigma^{-4}}{\sigma^{-4}} \left[ \sum_{i \in \mathcal{U}} \sum_{j=1}^{m} (y_i - \bar{\mu})^2 \frac{p_{ij}}{\tilde{k}_j} + \sum_{i \in \mathcal{C}} \sum_{j=1}^{m} \psi_{ij} \frac{p_{ij}}{\tilde{k}_j} \right] = 0
\]

where

\[
\psi_{ij} = \sigma^2 A_{ij} + 2\sigma(\mu - \bar{\mu}) B_{ij} + (\mu - \bar{\mu})^2 \quad \text{defined in equation (2.24)}.
\]

The updated \( \sigma^2 \) must satisfy

\[
n \tilde{\sigma}^2 = \sum_{i \in \mathcal{U}} (y_i - \bar{\mu})^2 \frac{w_i}{\sum_{i \in \mathcal{C}} \sum_{j=1}^{m} \psi_{ij} \frac{p_{ij}}{\tilde{k}_j}} \quad \text{(2.28)}
\]

The censored terms of (2.28) can be further simplified in terms of \( B_i \) and \( w_i \) as defined in the preceding section, as follows we have

\[
\sum_{j=1}^{m} \left\{ \psi_{ij} \frac{p_{ij}}{\tilde{k}_j} \right\} = \sum_{j=1}^{m} \left[ \sigma^2 A_{ij} + 2\sigma(\mu - \bar{\mu}) B_{ij} + (\mu - \bar{\mu}) \right] \frac{p_{ij}}{\tilde{k}_j}
\]
We note that \( A_{ij} = r_i B_{ij} + k_j \) (see equation 2.22).

Then the first term of the R.H.S. of the above equation becomes

\[
\sum_{j=1}^{m} \sigma^2 (r_i B_{ij} + k_j) \frac{p_{ij}}{k_j}
\]

\[
= \sigma^2 \left[ r_i \sum_{j=1}^{m} \frac{B_{ij} p_{ij}}{k_j} + \sum_{j=1}^{m} p_{ij} \right]
\]

\[
= \sigma^2 \left[ r_i w_i B_i^* + 1 \right]
\] (2.29)

since \( \sum_{j=1}^{m} p_{ij} = 1 \) for every \( i \). \( B_i^* \) is as defined in the above section. The second and the third terms contain \((\mu - \tilde{\mu})\) term.

At convergence \( \mu \rightarrow \tilde{\mu} \). Thus (2.28) can be approximated as

\[
n \sigma^2 \sim \sum_{i \in U} (y_i - \tilde{\mu})^2 w_i + \sum_{i \in C} (1 + r_i w_i B_i^*) \] (2.30)

2.4.4 Estimation of the Mixing Proportion.

Maximizing \( Q_e \) of (2.24) with respect to \( Y_s, \ i = 1, 2, \ldots, m \), will provide us with the scheme to estimate \( Y_s \). This maximization is equivalent to maximizing

\[
\sum_{i=1}^{n} \sum_{j=1}^{m} \log(\gamma_j) p_{ij}
\]

conditional upon \( \sum_{j=1}^{m} \gamma_j = 1 \).

Introducing Lagrange multiplier \( \lambda \), the normal equation for the estimation of \( \gamma_s, \ s = 1, 2, \ldots, m \), for the data set that may
include censored observations is

\[
\frac{\partial}{\partial \gamma_s} \left[ \sum_{i=1}^{n} \sum_{j=1}^{m} (\log \gamma_j) \ p_{ij} + \lambda \sum_{j=1}^{m} (\gamma_j - 1) \right] = 0
\]

i.e.

\[
\sum_{i=1}^{n} \frac{1}{\gamma_s} p_{is} + \lambda = 0
\]

then

\[
\gamma_s = - \frac{1}{\lambda} \sum_{i=1}^{n} p_{is}
\]

(2.31)

But since \( \sum_{j=1}^{m} \gamma_s = 1 \), then equation (2.31) is

\[
1 = - \frac{1}{\lambda} \sum_{s=1}^{m} \sum_{i=1}^{n} p_{is}
\]

which gives the solution of

\[
\lambda = - n
\]

since \( \sum_{s=1}^{m} p_{is} = 1 \), for every \( i = 1, 2, \ldots, n \).

Then the updated estimate of \( \gamma_s \), from (2.31) is

\[
\gamma_s = \frac{1}{n} \sum_{i=1}^{n} p_{is}
\]

(2.32)

2.4.5 Estimation of the Scale Contamination Parameter, \( k \).

Differentiating equation (2.24) with respect to \( \tilde{k}_s \), \( s = 1, 2, \ldots, m \), and setting it to zero gives us the normal equation for the estimation of \( k_s \), assuming \( \mu - \tilde{\mu} \) terms tends to zero,

\[
- \frac{1}{2} \sum_{i=1}^{n} p_{is} \frac{1}{k_s} + \frac{1}{2\sigma^2} \left[ \sum_{i \in C} \frac{(y_i - \tilde{\mu})^2}{k_s^2} p_{is} + \sum_{i \in C} \sigma^2 A_{is} p_{is} \right] = 0
\]
then

\[
\hat{k}_s = \frac{\sum_{i \in U} (y_i - \hat{\mu})^2 p_{is} + \sum_{i \in C} \hat{\sigma}^2 A_{is} p_{is}}{\hat{\sigma}^2 \sum_{i=1}^n p_{is}}, \quad \text{for } s = 1, 2, \ldots, m.
\] (2.33)

Note here that \(A_{is}\) and \(p_{is}\) are evaluated at values of the previous iterate, but \(\hat{\mu}\) and \(\hat{\sigma}^2\) are values at the present iteration. As mentioned in section (2.4.2) we may iterate \(\hat{k}_s\), and compute the corresponding \(y_s\) in (2.32) for the values of \(\mu\) and \(\sigma^2\) at each EM step. We employed the two options mentioned in that section, and found that both converge to the same points. In terms of computational efficiency there is little difference. Though less EM steps may be needed for convergence, more iterations are needed for the \(\hat{k}_s\) in each of the M-step.

2.5 **LINEAR REGRESSION MODELS.**

The extension to regression models is quite straightforward. Suppose that the mean of \(Y\) is \(\mu_i\), and that there exist concomitant variables \(x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})\), such that \(\mu_i = \beta_i x_i\), where \(\beta_i\) is a \(p \times 1\) vector of the regression parameters, which we need to estimate. The results of preceding sections may be used with \(\mu_i\) replacing \(\mu\). From (2.26) and the standard linear regression analysis, the estimate of \(\beta\) then becomes weighted least squares (Pettitt, 1985)

\[
\hat{\beta} = (X^T W X)^{-1} X^T W y^*
\] (2.34)

where \(X\) is the design matrix, \([x_i^T]\),

\[
W = \text{diag} \{w_i\}
\]

\[
y^* = (y_1^*, y_2^*, \ldots, y_n^*)^T
\]
with
\[
Y_i^* = \begin{cases} 
Y_i & \text{for uncensored } i \text{th case} \\
\mu_i + \sigma_i^* & \text{for censored } i \text{th case}
\end{cases}
\]
as given in Section 2.4.2, with the \(\mu\) replaced by \(\mu_i\).

Note that for discrete \(Q\), with \(\Pr(Q = \frac{1}{k_j})\) we have
\[
\omega_i = E(Q_i) = \sum_{j=1}^{m} (p_{ij} / k_j).
\]
where \(m\) is the number of components in the model and \(p_{ij}\) as defined in (2.06). To reduce bias in the estimate of \(\sigma^2\), following Aitkin (1981), we may use \((n-p)\) instead of \(n\) in (2.30), where \(p\) is the dimension of \(S\).

2.6 MARGINAL LOG-LIKELIHOOD.

The marginal log likelihood for the mixture model may be computed without requiring the values of the conditional probability \(p_{ij}\), defined in (2.06). The likelihood for \(n\) independent random variables \(Y_1, Y_2, \ldots, Y_n\) from density \(p(y)\) and survival distribution \(S(y)\), (Cox and Oakes, 1984, p.32) is
\[
\text{lik} = \prod_{i \in U} p(y_i) \prod_{i \in C} S(y_i)
\]
(2.35)
where \(U\) and \(C\) are sets containing the uncensored and censored subjects, respectively. We note that
\[
S(y) = 1 - \int_{0}^{y} p(z) \, dz.
\]
From (2.14) the marginal density for the mixture model with \( m \) components is

\[
p(y_i) = \sum_{j=1}^{m} \gamma_j f_j(y_i)
\]

where

\[
f_j(y_i) = \frac{1}{\sigma \sqrt{2\pi}} \phi \left( \frac{y_i - \mu}{\sigma} \right),
\]

where \( \phi \) is the standard normal density. Let \( r_i = (y_i - \mu)/\sigma \), then with \( S(y) \) defined as above, we have

\[
S(y_i) = \sum_{j=1}^{m} \gamma_j \left[ 1 - \phi(r_i/\sqrt{k_j}) \right]
\]

and the log of the marginal likelihood, \( L_m \), for the mixture model is

\[
L_m = \sum_{i \in U} \log \left[ \frac{1}{\sigma} \sum_{j=1}^{m} (\gamma_j/\sqrt{k_j}) \phi(r_i/\sqrt{k_j}) \right] + \sum_{i \in C} \log \left[ \sum_{j=1}^{m} \gamma_j (1 - \phi(r_i/\sqrt{k_j})) \right]
\]

(2.37)

The estimates obtained in section 2.4 thus maximize \( L_m \).

A point to note here is that maximizing \( Q_e(\hat{\theta}/\theta) \) in (2.24) is equivalent to maximizing the marginal likelihood \( L_m \). In section (3.6) we shall develop a scheme using the Newton-Raphson approximation to maximize \( L_m \) for general conditional distribution resulting in iterative weighted least squares.

2.7 SUMMARY OF THE ESTIMATION PROCEDURE.

The EM algorithm employed in the determination of the parameters of the mixture models, applied to both the single-case and regression problems is found to be equivalent to iterative least squares.
It can be programmed quite easily in GLIM. Appendix C1 gives the macros needed to carry out the computation, and an example run programme to invoke these macros. The step required in the EM procedure can be summarized as follows.

a) Set up initial estimates. These are values of $q_1, \ldots, q_m$ and the probabilities $\gamma_j = P(Q_i = q_j)$, where the index $j$ refers to the corresponding values belonging to the $j$th components. Some plausible starting values are chosen, with $q_1$ always set to unity to correspond to the components that are not contaminated.

b) Obtain initial values of $\mu_1$ and $\sigma$ by fitting standard least squares regression, treating the censored cases as though they failed at their censoring times (Schmee and Hahn's, 1979, iteration 0). $\sigma$ is chosen from the fit such that

$$\sigma = \frac{3}{2} \frac{\text{deviance/degree of freedom}}{\text{deviance}}$$

where deviance is the deviance produced by the GLIM fit.

c) Compute $w_i = E(Q_i)$ and $y_i^*$ for each observation $i = 1, 2, \ldots, n$. These are computed in macro WEST, (see Appendix C1).

d) Define $y_i^*$ as the Y variate, and $w_i$ as the weights for the fitting.

e) Fit model, through macro FITMAC (Appendix C1), which is defined in the run programme.

f) Updates $\sigma^2$, using equation (2.28), with $n$ substituted for $(n-p)$, to reduce biasness (refer section 2.5).
g) Updates $k_j$ and $\gamma_j$ (sections 2.4.5 and 2.4.4). Iterate $k_j$ if required.

h) Check for convergence. The most suitable criterion for convergence is chosen. This may be changes in values of the marginal likelihood $L_m$ or the mixing probability $\gamma_j$ or both together.

i) Repeat step c through to step i if convergence is not met.

This process is in macro ITER (Appendix C1).

2.8 EXAMPLES AND RESULTS.

The algorithm thus far described was implemented on different data sets. We only considered two components models, and determined the component membership by their posterior probabilities.

Observations with high posterior probabilities for the first component are deemed compatible with the $N(\mu, \sigma^2)$ assumption, while components with high probabilities for component 2 may be classified as outliers. Alternatively the final weights in the iterative process, $E(Q_j)$ may be used as a criterion for component membership. As suggested by Pettitt (1985), posterior values of $E(Q_j)$ close to a value of unity may be considered to correspond to the observation from $N(\mu, \sigma^2)$. For single sample model, the algorithm provides the means to separate the "bad" from the "good" observation, and for the linear regression model it further provides the estimation of the regression parameters (see section 2.5). The regression estimates are found using iterative reweighted least squares, using $E(Q)$ as the weights. This is a robust estimate since the outlying points...
are down weighted. In fact for \( E(Q_j) = 0 \), the estimates obtained are equivalent to that obtained when the jth case is deleted from the data set.

This method does not require prior knowledge of a suspect outlier. The algorithm begins from a reasonable choice of \( q_j \) and 
\[
Y_j = \Pr(Q = q_j), \quad j = 1, 2, \ldots, m.
\]

\( q_1 \) is set to 1, to correspond that the errors for this component are from \( N(0,1) \). The initial value of \( \mu_i \) (thus \( \beta \)) and \( \sigma^2 \) are obtained in the initial fit by considering all the weights in the fit to be equal to one. The convergence criterion chosen for our analysis is the loglikelihood.

2.8.1 Darwin Data (Appendix B1).

Recent analyses on this data set were done by Aitkin and Wilson (1980). Their models include three components analysis with different means and variances in the components. Our analysis is on two components model with the assumption that \( \mu \) is the same for both components. There is no censored observation in the data. Convergence is rather slow reflecting the flatness of the loglikelihood surface. If the starting points are too far off, iterations may exceed 100. Two modes of the loglikelihood function were found at values -73.9738 and -73.7627. These correspond to the starting points of \( q_2 \leq 1 \) or \( q_2 \geq 1 \) respectively. Table 2.1a gives the summary of this result.

Table 2.1b gives the result for starting value of \( q_2 = 0.1 \) with probability 0.1 for the component 2. This corresponds to the often quoted mixture distribution, reflecting real data, given by
TABLE 2.1a Estimates at convergence for different starting points of Q in the two components scale mixture model applied to Darwin Data.

<table>
<thead>
<tr>
<th>starting</th>
<th>q2 &lt;= 1</th>
<th>q2 &gt;= 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>log like</td>
<td>-73.9738</td>
<td>-73.7627</td>
</tr>
<tr>
<td>mean</td>
<td>27.06 (5.67)</td>
<td>20.00 (3.69)</td>
</tr>
<tr>
<td>q(2)</td>
<td>0.097</td>
<td>28.54</td>
</tr>
<tr>
<td>P(Q=1)</td>
<td>0.583</td>
<td>0.658</td>
</tr>
<tr>
<td>std(1)</td>
<td>16.67</td>
<td>44.58</td>
</tr>
<tr>
<td>std(2)</td>
<td>53.50</td>
<td>8.34</td>
</tr>
</tbody>
</table>

TABLE 2.1b Result for the two components scale mixture applied to Darwin data with starting values of q_j = (1.0, 0.1) and p_j = (0.9, 0.1)
0.9 N(0, 1) + 0.1 N(0, 9),

which corresponds to Q having the marginal distribution

$$\text{Pr}(Q = 1) = 0.9 \text{ and } \text{Pr}(Q = \frac{1}{9}) = 0.1$$

The mean for this data is found to be 27.6 with standard error of 5.67.

The standard deviations are 16.67 and 53.50 for components 1 and 2 memberships respectively. From the posterior probabilities, $P_{ij}$ or the posterior mean of Q, $E(Q)$ at the final iteration, we observe that cases 1, 2 and 15 are suspect outliers, with their $E(Q)$'s less than 0.2. Observations 13 and 14 are at the threshold of their posterior probabilities, or $E(Q)$'s are around 0.5. The estimate, 27.6, of $\mu$ may be seen also as the approximate value of the mean when observations 1, 2 and 15 are not included in the model, since these observations are down weighted (weights less than 0.2) in the fitting. The $\mu$ obtained by deleting these points is 27.50.

For the same model Aitkin and Wilson estimated $\mu$ to be 26.37 and standard deviation for component 1 membership 16.79. They start out with some suspect outlier points, and their results in the posterior probabilities for the component memberships are the complements of our finding, (that is their component 1 memberships is our component 2 memberships).

For this data set the algorithm seems to be less efficient in terms of the number of iterations required for convergence. However, when the outliers (observations 1, 2, 15) are not included convergence is reached in 9 iterations, and $E(Q_j) = 1$ for all $j = 1, 2, \ldots, 12$.
2.8.2 Forbe's Data (Appendix B2).

This data set represents an excellent example for simple regression analysis and detection of outliers. There are 17 observations all of which are uncensored. As in the simple regression analysis in Weisberg (1980), we fitted a linear model

\[ y = \beta_0 + \beta_1 T \]

where \( y \) is \( 100 \times \log(\text{pressure}) \), and \( T \) is the boiling point in °F. Convergence to four decimal places in the log likelihood is to a global maximum of 7.7895, after 9 iterations when starting at points \( q_2 = 0.1 \) and \( P(Q = q_2) = 0.4 \). The estimates are \( \hat{\beta}_0 = -41.31 \) (1.00), \( \hat{\beta}_1 = 0.891 \) (0.005), and \( \hat{\sigma} = 0.1089 \), with probabilities of component one memberships of over 0.96 with the exception of observation 12. Table 2.2 gives the summary of the result of the analysis with comparisons in the regression parameters estimates produced by the simple linear regression.

The algorithm is able to pick out unambiguously observation 12 (with \( E(Q) = 0.0063 \) as the outlier in the data set. This conforms to the finding in the residual analysis for the simple regression, (see Weisberg, 1980). Omitting this case from the data set, the EM algorithm gives identical results conforming the "robustness" of the estimates produced by this method. In fact in this example, since \( E(Q) \neq 0 \), the regression parameter vector \( \beta = (\beta_0, \beta_1)^T \) is identical to \( \beta_{(12)} \), where the subscript (12) indicates the parameter obtained when observation 12 is deleted from the data set. As seen from Table 2.2, the EM algorithm estimates of \( \beta \) are also identical to the estimates given out by simple linear regression model.
### Table 2.2 Result for the two components scale mixture applied to Forbes' data with starting values $q_j = (1.0, 0.1)$ and $p_j = (0.6, 0.4)$.

**NOTE:** obs. = $100 \times \log(\text{pressure})$

|x| E(Q)| P(Q=q) std dev.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$1.0000$</td>
<td>$1.000$</td>
</tr>
<tr>
<td>2</td>
<td>$0.0063$</td>
<td>$158.730$</td>
</tr>
</tbody>
</table>

#### Comparison of parameter estimates between EM and Linear regression

<table>
<thead>
<tr>
<th>ind</th>
<th>obs</th>
<th>x</th>
<th>$\beta_0$</th>
<th>$\beta_1$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>131.79</td>
<td>194.5</td>
<td>$-41.310$ (0.997)</td>
<td>$0.891$ (0.005)</td>
<td>$0.1089$</td>
</tr>
<tr>
<td>2</td>
<td>131.79</td>
<td>194.3</td>
<td>$-42.131$ (3.339)</td>
<td>$0.895$ (0.016)</td>
<td>$0.379$</td>
</tr>
<tr>
<td>3</td>
<td>135.02</td>
<td>197.9</td>
<td>$-41.302$ (1.000)</td>
<td>$0.891$ (0.005)</td>
<td>$0.113$</td>
</tr>
</tbody>
</table>

**Forbes data:** Comparison of parameter estimates between EM and Linear regression.

Forbes data: Comparison of parameter estimates between EM and Linear regression.
2.8.3 Carcinogenesis Data (Appendix B4).

This data set represents the analysis on a single-sample with censored data. There are 40 observations, 4 of which are censored. The observations are the number of days till mortality, after insult with the carcinogen. The cases are distinguished into two groups by a pretreatment regime. However, in our analysis we found that there is no distinct indication of heterogeneity in the data set. The algorithm converges to $\mu = 230.4$ with standard error of 7.3 and standard deviation 43.88 in 13 iterations. Table 2.3a gives the summary of this result together with that obtained by Schmee and Hahn. The values of $Q$ for the second component and its marginal probability are respectively 0.378 and 0.156. The posterior probabilities for component 1 memberships are all well above 0.7 except for cases 39 and 40 with probabilities 0.68 and 0.32 respectively. Case 40 may be suspect. Looking back at the data, case 40 is the longest surviving individual, and is censored at 344 days.

We then introduced artificial values for cases 1 and 11, replacing 142 by 42 for case 1 and 204 by 2040 for case 11. The algorithm was able to pick out these erroneous cases, from their posterior probabilities, of 0.048 and 0.0. At convergence, $\mu = 229.6$ with standard error (7.016). Note also that the posterior probabilities for case 40 for component 1 membership is also reduced indicating a definite outlier in the model. Table 2.3b gives the summary of this result.
EM  - the EM algorithm applied to the mixture model
S-H  - iterative least square method of Schmee and Hahn

TABLE 2.3a: Comparison of the estimate of mean for the Carcinogenesis data obtained by the two methods indicated above.

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>S-H</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean</td>
<td>230.4 (7.3)</td>
<td>231.6 (7.63)</td>
</tr>
<tr>
<td>Std. deviation</td>
<td>43.48</td>
<td>48.24</td>
</tr>
<tr>
<td>iterations</td>
<td>13</td>
<td>7</td>
</tr>
</tbody>
</table>

2.8.4 Motorette data (Appendix B3).

The data refers to lifetimes of 40 motorettes when subjected to temperature tests. It is heavily censored with less than 50% failures. Schmee and Hahn (1979) fitted a regression line

\[ y = \beta_0 + \beta_1 x \]

where \( y \) is the \( \log_{10} \) (life-times) and \( x = 100/(T + 273.2) \), where \( T \) is the temperature in °C and used an iterative least squares technique, for the normal model and adjusted to accommodate censored observations. Their estimates are \( \hat{\beta}_0 = -5.818 \), \( \hat{\beta}_1 = 4.204 \) and \( \hat{\sigma} = 0.204 \).

Likewise we applied the EM algorithm for finite mixtures on
<table>
<thead>
<tr>
<th>case</th>
<th>days</th>
<th>cen</th>
<th>$P_i(1)$</th>
<th>$E(Q)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>42</td>
<td>0</td>
<td>0.0482</td>
<td>0.0485</td>
</tr>
<tr>
<td>2</td>
<td>143</td>
<td>0</td>
<td>0.9886</td>
<td>0.9886</td>
</tr>
<tr>
<td>3</td>
<td>156</td>
<td>0</td>
<td>0.9934</td>
<td>0.9934</td>
</tr>
<tr>
<td>4</td>
<td>163</td>
<td>0</td>
<td>0.9949</td>
<td>0.9950</td>
</tr>
<tr>
<td>5</td>
<td>164</td>
<td>0</td>
<td>0.9951</td>
<td>0.9951</td>
</tr>
<tr>
<td>6</td>
<td>188</td>
<td>0</td>
<td>0.9976</td>
<td>0.9976</td>
</tr>
<tr>
<td>7</td>
<td>188</td>
<td>0</td>
<td>0.9976</td>
<td>0.9976</td>
</tr>
<tr>
<td>8</td>
<td>190</td>
<td>0</td>
<td>0.9977</td>
<td>0.9977</td>
</tr>
<tr>
<td>9</td>
<td>192</td>
<td>0</td>
<td>0.9978</td>
<td>0.9978</td>
</tr>
<tr>
<td>10</td>
<td>198</td>
<td>0</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
<tr>
<td>11</td>
<td>2040</td>
<td>1</td>
<td>0.0000</td>
<td>0.0003</td>
</tr>
<tr>
<td>12</td>
<td>205</td>
<td>0</td>
<td>0.9982</td>
<td>0.9982</td>
</tr>
<tr>
<td>13</td>
<td>206</td>
<td>0</td>
<td>0.9982</td>
<td>0.9982</td>
</tr>
<tr>
<td>14</td>
<td>209</td>
<td>0</td>
<td>0.9983</td>
<td>0.9983</td>
</tr>
<tr>
<td>15</td>
<td>213</td>
<td>0</td>
<td>0.9983</td>
<td>0.9983</td>
</tr>
<tr>
<td>16</td>
<td>216</td>
<td>0</td>
<td>0.9984</td>
<td>0.9984</td>
</tr>
<tr>
<td>17</td>
<td>216</td>
<td>1</td>
<td>0.9358</td>
<td>0.9358</td>
</tr>
<tr>
<td>18</td>
<td>220</td>
<td>0</td>
<td>0.9984</td>
<td>0.9984</td>
</tr>
<tr>
<td>19</td>
<td>227</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>20</td>
<td>230</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>21</td>
<td>232</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>22</td>
<td>232</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>23</td>
<td>233</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>24</td>
<td>233</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>25</td>
<td>233</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>26</td>
<td>233</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>27</td>
<td>234</td>
<td>0</td>
<td>0.9985</td>
<td>0.9985</td>
</tr>
<tr>
<td>28</td>
<td>239</td>
<td>0</td>
<td>0.9984</td>
<td>0.9984</td>
</tr>
<tr>
<td>29</td>
<td>240</td>
<td>0</td>
<td>0.9984</td>
<td>0.9984</td>
</tr>
<tr>
<td>30</td>
<td>244</td>
<td>1</td>
<td>0.8969</td>
<td>0.8970</td>
</tr>
<tr>
<td>31</td>
<td>246</td>
<td>0</td>
<td>0.9983</td>
<td>0.9983</td>
</tr>
<tr>
<td>32</td>
<td>261</td>
<td>0</td>
<td>0.9980</td>
<td>0.9980</td>
</tr>
<tr>
<td>33</td>
<td>265</td>
<td>0</td>
<td>0.9978</td>
<td>0.9978</td>
</tr>
<tr>
<td>34</td>
<td>280</td>
<td>0</td>
<td>0.9970</td>
<td>0.9970</td>
</tr>
<tr>
<td>35</td>
<td>280</td>
<td>0</td>
<td>0.9970</td>
<td>0.9970</td>
</tr>
<tr>
<td>36</td>
<td>296</td>
<td>0</td>
<td>0.9950</td>
<td>0.9950</td>
</tr>
<tr>
<td>37</td>
<td>296</td>
<td>0</td>
<td>0.9950</td>
<td>0.9950</td>
</tr>
<tr>
<td>38</td>
<td>304</td>
<td>0</td>
<td>0.9932</td>
<td>0.9932</td>
</tr>
<tr>
<td>39</td>
<td>323</td>
<td>0</td>
<td>0.9842</td>
<td>0.9842</td>
</tr>
<tr>
<td>40</td>
<td>344</td>
<td>1</td>
<td>0.0887</td>
<td>0.0890</td>
</tr>
</tbody>
</table>

Table 2.3b Probabilities for component 1 membership and $E(Q)$ in the finite mixture model for the Carcinogenesis data with case 1 and 11 purposely made outliers.
the model. Convergence was attained after 41 iterations to 
\( \hat{q} = (1, 0.047) \) with 
\( P(Q = 0.047) = 0.265 \), and log likelihood 
-8.4547. This gives the mixture distribution as

\[
0.735 N(0, 1) + 0.265 N(0, 21.35)
\]

for \( (y_i - \mu_i)/\sigma \). The regression parameter estimates are as in Table 2.4a, alongside of which are the results obtained by the iterative least squares of Schmee and Hahn (1979), and by maximum likelihood as quoted by Schmee and Hahn (1979), page 421. Referring to Table 2.4c, two distinct component memberships are immediately observed from their posterior probabilities or their posterior means of \( Q \). Cases 11, 21 and 22 falls to component 2 with posterior probabilities close to 1 (or \( E(Q) \) close to zero). These points are identical to those points considered probable outliers in Schmee and Hahn's (1979) scatter plot. When these points are removed from the data set and applying the EM algorithm gives us almost identical results with that of the EM for the full data set, converging to maximum likelihood 1.155. Although they are outliers, these points do not represent influential cases (see section 3.8.1) on the regression parameter estimates, since their omission does not change the parameters significantly.

Table 2.4b gives the comparison of the parameter estimates from the EM algorithm and the iterative least squares method of Schmee and Hahn (1979). Observe that, the results are very similar, except for the standard deviation, \( \sigma \), which is overestimated by the EM algorithm.
Table 2.4a  Comparison of Regression parameter estimates of
the Motorette data obtained by EM algorithm on
mixture model and the Iterative least squares of
Schmee and Hahn (S-H). The estimates given in last column are
those quoted by S-H, and the bounds ( * ) in the parentheses are
the approximate 90% confidence interval.

<table>
<thead>
<tr>
<th></th>
<th>EM full data</th>
<th>S-H full data</th>
<th>max. likelihood</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-5.558 (.251)</td>
<td>-5.818 (.576)</td>
<td>-6.027 (1.484)*</td>
</tr>
<tr>
<td>slope</td>
<td>4.082 (.114)</td>
<td>4.204 (.261)</td>
<td>4.314 (0.684)*</td>
</tr>
<tr>
<td>std. dev</td>
<td>0.097</td>
<td>0.204</td>
<td>0.259</td>
</tr>
</tbody>
</table>

Table 2.4b  Comparison of the regression estimates of
motorette data when observations 11, 21 and
22 are deleted from the data set.

<table>
<thead>
<tr>
<th>case</th>
<th>(obs)</th>
<th>P1</th>
<th>E(Q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.907</td>
<td>0.8041</td>
<td>0.8133</td>
</tr>
<tr>
<td>11</td>
<td>3.246</td>
<td>0.0035</td>
<td>0.0503</td>
</tr>
<tr>
<td>12</td>
<td>3.443</td>
<td>0.5927</td>
<td>0.6118</td>
</tr>
<tr>
<td>16</td>
<td>3.687</td>
<td>0.9235</td>
<td>0.9271</td>
</tr>
<tr>
<td>17</td>
<td>3.716</td>
<td>0.9127</td>
<td>0.9168</td>
</tr>
<tr>
<td>18</td>
<td>3.736</td>
<td>0.5576</td>
<td>0.5784</td>
</tr>
<tr>
<td>21</td>
<td>2.611</td>
<td>0.0000</td>
<td>0.0469</td>
</tr>
<tr>
<td>22</td>
<td>2.611</td>
<td>0.0020</td>
<td>0.0469</td>
</tr>
<tr>
<td>25</td>
<td>3.158</td>
<td>0.8905</td>
<td>0.8957</td>
</tr>
<tr>
<td>26</td>
<td>3.225</td>
<td>0.7650</td>
<td>0.7761</td>
</tr>
<tr>
<td>31</td>
<td>2.611</td>
<td>0.8784</td>
<td>0.8841</td>
</tr>
<tr>
<td>33</td>
<td>2.702</td>
<td>0.9269</td>
<td>0.9303</td>
</tr>
<tr>
<td>40</td>
<td>2.723</td>
<td>0.7294</td>
<td>0.7421</td>
</tr>
</tbody>
</table>

TABLE 2.4c  Posterior probabilities for component 1 membership
and E(Q) for some cases for the two components
scale mixture applied to motorette data with
starting value of q2 < 1
2.8.5 Heart Transplant Data (Appendix B6).

Similar analyses to those above in Section 2.8.4, were done on a larger data set and with more covariates. The Stanford Heart Transplant data for 152 cases who survived for at least 10 days after transplant is used in our analysis with covariates age, T5 mismatch score and age \times age. We applied the EM algorithm for two component mixture to 5 different regression models to include various combinations of covariates with the dependent variable being $\log_{10}$ (failure-times). Convergence was quite rapid; in all cases after less than 15 iterations. No distinct components in all the regression models were apparent. The value of $q_2$ for the second component was close to unity. Hence no outliers were detected.

The regression parameter estimates obtained for the various regression models are given in Table 2.5, where

\[
\begin{align*}
A &= \text{age} \\
T5 &= \text{T5 mismatch score}/100 \\
A2 &= A \times A.
\end{align*}
\]

The estimates obtained by Buckley and James (1979) found by Miller and Helpern (1982) are also given in Table 2.5 for comparison. The results for both methods are quite similar, with the EM method giving slightly larger estimates. We see here that the mixture model can be used successfully to estimate the parameters in the linear regression model with normal errors from a homogenous population with censored observations.

When an outlier is purposely introduced at age 13, replacing
Table 2.5 Stanford Heart transplant data regression estimates for log(life-time) on different linear models with 152 observations. B-J is the Buckley and James method and EM is the algorithm for the two component mixture model.

<table>
<thead>
<tr>
<th></th>
<th>intercept</th>
<th>A</th>
<th>T5</th>
<th>A2</th>
</tr>
</thead>
<tbody>
<tr>
<td>B-J</td>
<td>3.46 (.30)</td>
<td>-0.018 (.007)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2.81 (.14)</td>
<td>-</td>
<td>-0.123 (.111)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3.55 (.31)</td>
<td>-0.018 (.007)</td>
<td>-0.094 (.115)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1.21 (.70)</td>
<td>0.113 (.037)</td>
<td>-</td>
<td>-0.0017 (.0005)</td>
</tr>
<tr>
<td>EM</td>
<td>3.645 (.286)</td>
<td>-0.021 (.007)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>2.935 (.156)</td>
<td>-</td>
<td>-0.141 (.124)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>3.727 (.306)</td>
<td>-0.020 (.007)</td>
<td>-0.098 (.122)</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>1.422 (.679)</td>
<td>0.108 (.037)</td>
<td>-</td>
<td>-0.0017 (.0005)</td>
</tr>
<tr>
<td></td>
<td>1.497 (.700)</td>
<td>0.107 (.037)</td>
<td>-0.052 (.118)</td>
<td>-0.0017 (.0005)</td>
</tr>
</tbody>
</table>

* - found by Miller and Halpern
the survival time of 10 days by 1, the EM algorithm was able to pick out this as an outlier, though convergence is somewhat slow. For regression on age and T5, convergence is reached after 133 iterations resulting in \( q_2 = 0.189 \) and mixing probability, \( \gamma_2 = \Pr(Q = q_2) \), equal to 0.0418. The posterior probability for component 2 membership for the case is 0.8955. The regression parameters estimates are, \( \hat{\beta}_o = 3.856 \ (0.3043) \), \( \hat{\beta}_1 = -0.0233 \ (0.0067) \) and \( \hat{\beta}_2 = -0.0802 \ (0.1198) \) which is identical to the model when the case is deleted from the data set. Again we see that though the case is an outlier, it is not influential to the whole data set up.

2.8.6 Conclusion.

In conclusion we see that the EM algorithm with the finite scale mixture model can be successfully applied to censored data in both single sample and regression problems. It is robust, downweighting the outliers using iterative weighted least squares. We could detect these outliers by the weights, \( E(Q_i) \) having values close to zero or by their posterior probabilities for component memberships. For case \( i \) with weight \( E(Q_i) \approx 0 \), the regression parameter \( \hat{\beta}_2 \), obtained by the algorithm is approximately equal to \( \hat{\beta}(i) \), the estimate when the \( i \)th observation is deleted from the model.

A point to note is that when there is a definite outlier in the data, the algorithm converges to specific value of \( q_1 \) and \( \gamma_j = \Pr(Q = q_j) \), where \( j \) indexes the component. This is in agreement with the formulation of the model. Values of \( q_j \neq 1 \), \( j = 2, 3, \ldots, m \) imply outlier models for which \( \gamma_j \) need be specified
such that \((q_j, \gamma_j)\) defines the maximum likelihood in the \((q, \gamma)\) space. However, if no outlier is present in the data set, \(\gamma_j\) may take value of zero for which \(q_j\) need not be specified, or \(q_j = 1\) for which \(\gamma\) may take any value in the interval \([0, 1]\).

Illustrating this with the two components model, the no outlier values of \(Y_2\) and \(q_2\) are points in the \((q_2, Y_2)\) space on lines OR and NM in Figure 2.1.

![Figure 2.1](image-url)

**Figure 2.1**: Figure to illustrate the values of \(q_2\) and \(Y_2\) in the two component model for which there is no outlier.

The means or the regression coefficients, though, converge to the same points. Thus this algorithm may be used to find the robust estimates in a regression model with normal error for censored data.

As was found in the Darwin data, the likelihood function may be multimodal. The estimates obtained using the EM algorithm may also be at a local mode of the likelihood function or at a stationary point. This can be investigated using different starting values and by evaluation of the second derivative of the loglikelihood function.
Note that for the observation \( Y = \mu + \sigma \varepsilon \), where \( \varepsilon \) has the general density \( f \), assumed symmetric about zero, Pettitt (1985, Section 1), has established the influence of censoring on the estimate \( \mu \). Thus in context of right censoring, and for the commonly occurring symmetric densities, early censoring is qualitatively similar to that of an observation being missing, that is zero influence on the estimate \( \mu \), while large positive observations are effectively similar to uncensored observations.

2.9 NOTE ON INFLUENCE AND ROBUSTNESS.

In this chapter we have developed a scheme using the EM algorithm for estimating population parameters, namely the \( \mu \), (or the regression parameters in the case of linear models), from a population which may be "contaminated" by "rogue" or "outlier" points. Here the "outliers" are thought to come from distributions other than the assumed distribution in the model. Thus for the two component finite mixture models, if \( F \) is the cumulative distribution from which the outliers are deemed to come, the contaminated distribution (see Barnett and Lewis, 1978, page 127) is

\[
(1 - \lambda) F + \lambda G
\]

(2.38)

where \( 0 \leq \lambda \leq 1 \), is the contamination fraction. (2.38) represents mixtures with one contaminating distribution. One could generalize the finite mixture model with several contaminated distributions, giving

\[
\sum_{j=1}^{m} \lambda_j G_j
\]
where \[ G_1 = F \quad \text{and} \quad \lambda_1 = 1 - \sum_{j=2}^{m} \lambda_j. \]

Our analysis thus far only deals with \( F \) and \( G \) being normal distributions with the same mean but with a different variance. Generalizations of this model to infinite scale mixture models will be discussed in Chapter 3.

The EM algorithm for the estimation of the parameters described in our analysis reduces to reweighted iterative least squares; with weights defined by the conditional expectation of the contaminating distribution (see Section 2.4.2). Thus cases far removed in the observation space are down weighted in the analysis. As an example take the Forbes Data in Section 2.8.2. The weight for case 12 is approximately zero; thus the EM estimates of the regression parameters are essentially identical to that obtained by least square estimation when that particular observation is omitted from the data, (see Table 2.2). The EM estimates without case 12 in the data is also identical. From the EM analysis case 12, as an outlier does not contribute influence to the estimation of the parameters, the estimates being the same with or without that particular case. However, there is a slight difference in the estimates obtained by least squares regression with and without the case 12 in the data (Table 2.2 for the comparison of estimates). Even by this method case 12 may not be an influential point in the model; but the effect of the case on standard errors is more marked. Inclusion of this case gives the appearance of less reliable results than would be suggested on the basis of the other 16 cases.
It is possible for the influence of cases that are remote from the bulk of the data to be masked by the robust analysis, since generally the robust procedure downweights extreme points, accommodating them as the rest of the observation under the assumed model. A case might have dominating influence (high leverage), but is not detected due to robustifying the analysis. However, such a point is detected as an outlier. In such a situation we are then confronted with a conflict; if such a leverage point is a proper observation, that is, it belongs to the assumed model, then we lose efficiency in robustifying the analysis. On the other hand if the point represents a gross error and we do not downweight it, the result may be questionable. Influential points may not necessarily be outliers. They may be extreme points in the assumed model.

In Chapter 1, Section 1.1.5, we have defined the influence function as introduced by Hampel (1968). We shall rewrite it here again for convenience of reference.

IF (x; T; F) = \lim_{\varepsilon \to 0^+} \frac{T[(1-\varepsilon) F + \varepsilon G] - T(F)}{\varepsilon}

where T(·) is a statistical functional such that, if \( T_n(x_1, x_2, \ldots, x_n) \) is an estimator based on independent and identically distributed sample \( x_1, x_2, \ldots, x_n \) from cumulative distribution function \( F \), as \( n \to \infty \)

\[ T(F) = T_n(x_1, x_2, \ldots, x_n) \]

and \( G \) is the contaminating distribution giving mass 1 at point \( x \); that is \( P(G = x) = 1 \), and zero elsewhere.
Thus we see that $IF(\cdot)$ is just an ordinary right-hand derivative, evaluated at $\epsilon = 0$, of the statistical functional of the mixture (or contaminated) distribution of (2.38); that is

$$IF(x; T, F) = \frac{\partial}{\partial \epsilon} \left[ T \left( (1-\epsilon) F + \epsilon G \right) \right]_{\epsilon=0}$$ \hspace{1cm} (2.39)

It measures the effect of contamination at point $x$ on the estimate, standardized by the mass of the contamination as $n \rightarrow \infty$. The estimator would be called robust if small changes in $F$ do not produce wild fluctuations in $T(F)$, and hence $IF(\cdot)$. The degree of robustness may also be measured in terms of the quantities derived from the influence curve as given in Section 1.1.5.

Detection of influential cases in a model could tremendously benefit the analyst. The study of influence will yield information on the reliability of the conclusions and their dependence on the assumed model. It also indicates areas in the observation space with inadequate coverage for reliable estimation and prediction. It has become an important diagnostic tool, especially in regression analysis (see Belsley et al. 1980, Cook and Weisberg, 1982).

Essentially the idea of influence analysis is to study the changes in the outcome of the analysis when a small perturbation in the problem formulation is introduced. In this thesis the perturbation scheme considered is that in which the data which include censored observations, are modified by deletion of cases, particularly single case deletion. In section 3.7 scalar measures are introduced to quantify this influence. The Cook distance (Cook, 1977, 1979) is a direct consequence of the sample influence
curve (Section 1.1.5). The influence assessment for the scale mixture model is presented in Section 3.7. Although in that section emphasis is given on the continuous mixture models, the scheme could be adapted to accommodate the finite mixture model described in this section.
Chapter 3

Continuous Scale Mixture Models with Censored Data - Student-t Error Distributions

3.1 INTRODUCTION.

As in Chapter 2, we shall present only the scale mixture model and apply the EM algorithm in both single-case and regression problems with censored data. Again we consider $Y/Q = q$ distributed normally with mean $\mu$ and variance $\sigma^2 q^{-1}$, and $Y = \mu + \sigma \varepsilon$. However here we shall generalize $Q$ to include continuous distributions, with probability density $p(q)$. From (2.02) and (2.03) the marginal distribution of $\varepsilon = (Y - \mu)/\sigma$ is then

$$f(\varepsilon) = \int_{0}^{\infty} q^{-\frac{1}{2}} \phi(q^{-\frac{1}{2}}) p(q) \, dq \quad (3.01)$$

which represents distribution of the scale mixture of a standard normal distribution. Andrew and Mallows (1974) show how to find $p(q)$ by considering that $\varepsilon$ may be generated by the ratio of a standard normal variate to $Q$, and using Laplace transform. For example if the marginal distribution of $\varepsilon$ is Student-t with $k$ degrees of freedom then $Q$ is distributed gamma, $G(\alpha, \beta)$, with $\alpha = \beta = \frac{1}{2} k$; that is the density of $Q$ is:

$$p(q) = \frac{\beta^{\alpha} q^{\alpha-1} e^{-\beta q}}{\Gamma(\alpha)}$$

with $\alpha = \beta = \frac{1}{2} k$. In this chapter we shall take advantage of this relationship, and consider the errors in the models to be those of Student-t distributions with even degrees of freedom,
2α say, so that the marginal distribution \( p(q) \) is \( G(α, α) \). Pettitt (1985) has given explicit results for these models, but we shall present here some of the results for completeness.

As in the previous chapter the EM algorithm provides a means of solving for the parameters in the likelihood (or log likelihood) equation (2.37) for the discrete, or the equivalent for continuous \( Q \). The parameters obtained by this approach are then compared to those obtained by maximizing the likelihood using Newton-Raphson method.

Subsequent subsections will be devoted to identifying outliers and influential points. For influence assessment, the case deletion approach is employed and the variations in the parameters are studied. Scalar measures are also given to provide indication of influence.

3.2 MARGINAL DISTRIBUTION OF \( Q \) FOR STUDENT-t ERROR.

The error components for the scale mixture of normal distributions model has marginal density (3.01), that is

\[
 f(ε) = \int_0^\infty q^{1/2} \phi(eq^{1/2}) p(q) \, dq .
\]

Substituting \( V = Q^{1/4} \) and letting \( G_v(v) \) be the probability distribution of \( V \), then \( f(ε) \) can be written as:

\[
 f(ε) = \int_0^\infty v \frac{v}{\sqrt{2π}} \exp\left(-\frac{1}{2} v^2\right) \, dG_v(v) \quad (3.02)
\]

which is Andrews and Mallow's (1974) equation (2.1) for the random variable generated by independent variables \( Q, Z \) with \( Z \) standard normal such that \( Z/Q = ε \). Andrews and Mallow (1974) also have
proved the theorem which establishes the necessary and sufficient conditions for this representation to exist. However, here we only give their Lemma on the fundamental relationships among the variables, for which $G_V(v)$ may be obtained from given marginal distribution of $\epsilon$.

Lemma 3.2.1.

If $f_X$ has the representation (3.02) and $h(y) = f_X(\sqrt{y})$, then $h$ is the Laplace transform, $\mathcal{L}$, of $H$.

$$H(t) = \int_0^t \left( \frac{u}{t} \right)^{\frac{1}{2}} dG_V\left\{ (2u)^{\frac{1}{4}} \right\} .$$

(3.03)

and where $G_V(v)$ is differentiable we have

$$H'(t) = (2\pi)^{-\frac{1}{4}} G'_V\left\{ (2t)^{\frac{1}{4}} \right\} ,$$

(3.04)

$$G'_V(v) = (2\pi)^{\frac{1}{4}} H'(\frac{1}{2} v^2) .$$

(3.05)

Proof:

Let

$$h(y) = \int_0^\infty \frac{v}{\sqrt{2\pi}} \exp\left\{ -\frac{v^2}{2} - y \right\} dG_V(v) .$$

$$u = \frac{v^2}{2} ,$$

then

$$h(y) = \int_0^\infty e^{-uy} \left( \frac{u}{\pi} \right)^{\frac{1}{4}} dG_V\left\{ (2u)^{\frac{1}{4}} \right\}$$

$$= \int_0^\infty e^{-u(y)} H(u) du$$

$$= \mathcal{L}\{H(t)\}$$
where $H$ is such that

$$H(u) \, du = \left\{ \frac{u}{\pi} \right\} \, dG\left\{ (2u)^{\frac{1}{2}} \right\}$$

hence

$$H(t) = \int_{0}^{t} \left\{ \frac{u}{\pi} \right\} \, dG\left\{ (2u)^{\frac{1}{2}} \right\} .$$

Equations (3.04) and (3.05) follow by differentiating (3.03) with respect to $t$.

In their example 4.1, Andrews and Mallows (1974) show for the Pearson Type VII density,

$$f_{X}(x) = \left\{ \frac{\beta^{\frac{1}{2}}}{\Gamma\left( m \frac{1}{2} \right)} \right\}^{-1} \left\{ 1 + \frac{x^2}{a^2} \right\}^{-m}$$

(3.06)

of which the Student-t is a special case, the density of $V$ is

$$g_{V}(v) = \left\{ \Gamma\left( m \frac{1}{2} \right) \right\}^{-1} \left\{ \frac{1}{2} a^2 v^2 \right\}^{m-\frac{3}{2}} a^2 v \exp\left\{ -\frac{1}{2} a^2 v^2 \right\} .$$

(3.07)

Substituting $Q = V^2$, we have the density of $Q$ as

$$p(q) = 2q^{-\frac{1}{2}} \left\{ \Gamma\left( m \frac{1}{2} \right) \right\}^{-1} \left\{ \frac{1}{2} a^2 q \right\}^{m-\frac{3}{2}} a^2 q^{\frac{1}{2}} \exp\left\{ -\frac{1}{2} a^2 q \right\}$$

$$= \left( \frac{1}{2} a \right)^{m-\frac{1}{2}} q^{\frac{m-\frac{3}{2}}{2}} e^{-\frac{a^2 q}{2}}$$

$$\frac{\Gamma\left( m-\frac{1}{2} \right)}{\Gamma\left( m \frac{1}{2} \right)}$$

let $\alpha = m - \frac{1}{2}$ and $\gamma = \frac{a^2}{2}$, then
\[ p(q) = \frac{\gamma^a q^{a-1} e^{-\gamma q}}{\Gamma(a)} \]  

which is the density for the gamma distribution with parameters \((\alpha, \gamma)\).

The Student-t distribution is a special case of (3.06) with \(\alpha^2 = k\) and \(m = (k+1)/2\), where \(k\) is the degrees of freedom. Then \(\alpha = \frac{1}{2} k\) and \(\gamma = \frac{1}{2} k\), that is for the marginal distribution of \(\varepsilon\) Student-t with \(k\) degrees of freedom, then \(Q\) is distributed gamma with parameters \((\frac{1}{2} k, \frac{1}{2} k)\). Our analysis for the continuous mixtures, in this chapter shall assume that the error is from Student-t with even degrees of freedom, so the \(Q\) is gamma with integer parameters.

### 3.3 POSTERIOR DENSITY OF \(Q\) AND THE MOMENTS.

In general, the conditional density of \(Q\), given \(\varepsilon = r\) if the case is a failure or \(\varepsilon > r\) if the case is censored where for an observation \(y\), \(r = (y-\mu)/\sigma\), for the scale mixture of normal distribution follows as in section (2.2), whether \(Q\) is discrete or continuous, that is

\[
p(q/\varepsilon) = \begin{cases} 
q^{\frac{1}{2}} \Phi(rq^{\frac{1}{2}}) p(q), & \varepsilon = r \text{ for failure} \\
[1 - \Phi(rq^{\frac{1}{2}})] p(q), & \varepsilon > r, \text{ censored} 
\end{cases}
\]

Similarly, the conditional moments of \(\varepsilon\) given \(Q = q\) and the data are
where the expectation is with respect to the conditional distribution of \( \varepsilon \), given \( Q \) and the data, and

\[
I(k,x) = \int_{x}^{\infty} u^k \phi(u) \, du
\]

with the reduction formula as in (2.10).

To obtain the conditional moments of \( Q \) we require results for the following

\[
M_1(r,k) = \int_{0}^{\infty} q^k \phi(q) \, p(q) \, dq \quad \text{(3.11)}
\]

and

\[
M_0(r,k) = \int_{0}^{\infty} q^{1-\varepsilon} \phi(q) \, p(q) \, dq \quad \text{(3.12)}
\]

Note that \( M_1(r,1) \) and \( M_0(r,0) \) are the normalizing constant in (3.09) when \( \varepsilon = r \) and \( \varepsilon > r \) respectively. For \( Q \) distributed as a gamma random variable with \( G(\alpha,\gamma) \), we have

\[
p(q) = \frac{\gamma^\alpha q^{\alpha-1} e^{-\gamma q}}{\Gamma(\alpha)} , \quad q > 0 \quad \text{(3.13)}
\]

then

\[
M_1(r,k) = \int_{0}^{\infty} q^k \frac{e^{-\gamma q} q^{\alpha-1} e^{-\gamma q}}{\sqrt{2\pi} \Gamma(\alpha)} \, dq
\]

\[
= \frac{(2\pi)^{-\frac{1}{2}}}{\Gamma(\alpha)} \gamma^\alpha \int_{0}^{\infty} q^{k+\alpha-1} e^{-(\gamma + \frac{1}{2} k^2)q} \, dq .
\]
Substituting \((y + \frac{1}{2} r^2)q = z\), the integral then becomes

\[
\frac{2^{k+a}}{(2\gamma + r^2)^{k+a}} \int_0^\infty z^{(k+a)-1} e^{-z} \, dz
\]

\[
= \frac{2^{k+a}}{(2\gamma + r^2)^{k+a}} \cdot \Gamma(k+a).
\]

Hence

\[
M_1(r,k) = \frac{2^{\alpha+k-\frac{1}{2}}}{\sqrt{\pi}} \frac{\Gamma(k+a)}{\Gamma(\alpha)} \frac{\gamma^\alpha}{(2\gamma+r^2)^{k+a}}.
\]  

(3.14)

To obtain \(M_0(r,k)\), we need the following relationship for the gamma distribution function

\[
P(q) = \int_0^\infty \frac{q^\alpha u^{\alpha-1} e^{-\gamma u}}{\Gamma(\alpha)} du
\]

\[
= 1 - e^{-\gamma q} \sum_{h=0}^{\alpha-1} \frac{(\gamma q)^h}{h!}.
\]  

(3.15)

Integrating (3.12) by parts we obtain

\[
M_0(r,k) = \left[ 1 - \Phi(q) \right] U(q) + \frac{r}{2} \int_0^\infty U(q) q^{-\frac{1}{2}} \Phi(q) \, dq
\]  

(3.16)

where

\[
U(q) = \int_0^\infty q^k p(q) \, dq
\]

\[
= \int_0^\infty q^{\alpha+k-\frac{1}{2}} \frac{\gamma^\alpha u^{\alpha-1} e^{-\gamma u}}{\Gamma(\alpha)} \, dq
\]

\[
= \frac{\Gamma(\alpha+k)}{\gamma^k \Gamma(\alpha)} \int_0^\infty q^{\alpha+k-1} e^{-\gamma q} \, dq
\]
and from (3.15)

\[ U(q) = G(\alpha, \gamma; k) \left[ 1 - e^{-\gamma q} \sum_{h=0}^{\alpha+k-1} \frac{(\gamma q)^h}{h!} \right] \]

where

\[ G(\alpha, \gamma; k) = \frac{\Gamma(\alpha+k)}{\gamma^k \Gamma(\alpha)} . \]

The first term of (3.16) in the square bracket can then readily be seen equal to zero. Hence we have

\[ M_o(r, k) = \frac{r}{2} \frac{G(\alpha, \gamma; k)}{\Gamma(\alpha)} \int_0^\infty \left[ 1 - e^{-\gamma q} \sum_{h=0}^{\alpha+k-1} \frac{(\gamma q)^h}{h!} \right] q^{-\frac{1}{2}} \phi(rq^\frac{1}{2}) \, dq . \]

Now the first term in the integral is given by

\[ \int_0^\infty q^{-\frac{1}{2}} \phi(rq^\frac{1}{2}) \, dq = \frac{1}{\sqrt{2\pi}} \int_0^\infty q^{-\frac{1}{2}} e^{-\frac{1}{2}r^2 q} \, dq \]

\[ = \frac{1}{rv^2} \Gamma(\frac{1}{2}) \]

\[ = \frac{1}{r} . \]

The second term in the integral can be written

\[ - \frac{1}{\sqrt{2\pi}} \int_0^\infty \left[ e^{-\gamma q} \sum_{h=0}^{\alpha+k-1} \frac{(\gamma q)^h}{h!} \right] q^{-\frac{1}{2}} e^{-r^2 q/2} \, dq \]

\[ = - \frac{1}{\sqrt{2\pi}} \sum_{h=0}^{\alpha+k-1} \frac{\gamma^h}{h!} \int_0^\infty e^{-bq} q^{-\frac{1}{2}} \, dq \]

where \( b = (\gamma + \frac{1}{2}r^2) \). The integral in the above expression is
\[
\frac{1}{b^{h+\frac{1}{2}}} \Gamma(h+\frac{1}{2})
\]

and since \( h! = \Gamma(h+1) \)

(3.16) can now be written as

\[
M_0(r,k) = \frac{G(\alpha, \gamma; k)}{2} \left\{ 1 - \frac{r}{\sqrt{2\pi} \gamma} \sum_{h=0}^{\alpha+k-1} \left[ \frac{2\gamma}{2\gamma+r^2} \right]^{h+\frac{1}{2}} \frac{\Gamma(h+\frac{1}{2})}{\Gamma(h+1)} \right\} . (3.17)
\]

The posterior density for uncensored \( Y \) is thus

\[
p(q/e=r) = \frac{q^{\frac{1}{2}} \phi(rq^{\frac{1}{2}}) p(q)}{M_1(r, \frac{1}{2})}
\]

\[
= \left\{ (2\pi)^{-\frac{1}{2}} \frac{\gamma^\alpha}{\Gamma(\alpha)} q^{(a+\frac{1}{2})-1} e^{-bq} \right\} / M_1(r, \frac{1}{2})
\]

where \( b \) is defined as above. Now

\[
M_1(r, \frac{1}{2}) = (2\pi)^{-\frac{1}{2}} \frac{\Gamma(\alpha+\frac{1}{2})}{\Gamma(\alpha)} \frac{\gamma^\alpha}{b^{a+\frac{1}{2}}}
\]

Then

\[
p(q/e=r) = \frac{b^{a+\frac{1}{2}} q^{(a+\frac{1}{2})-1} e^{-bq}}{\Gamma(\alpha+\frac{1}{2})} \quad (3.18)
\]

which is a gamma density with parameters \((a,b)\), where \( a = a + \frac{1}{2} \)

and \( b = \gamma + \frac{1}{2} r^2 \). Thus \( E(Q) \), where the expectation is with respect to the conditional distribution of (3.18), and higher conditional moments of \( Q \) can be obtained straightforwardly for the uncensored data case, that is
\[ E(Q^k) = \frac{\Gamma(a+k)}{b^k \Gamma(a)} \]

\[ = \frac{\Gamma(a+k+1)}{(\gamma+r^2)^k \Gamma(a+1)} \quad (3.19) \]

Of course (3.19) may be obtained from the ratio of

\[ \frac{M_{11}(r, k+1)}{M_{11}(r, \frac{1}{2})}. \]

In general the moments of \( Q \) conditional on \( \varepsilon \) are

\[ E(Q^k) = \begin{cases} 
\frac{\Gamma(a+k+1)}{(\gamma+r^2)^k \Gamma(a+1)} & \text{for } \varepsilon = r \\
\frac{M_0(r,k)}{M_0(r,0)} & \text{for } \varepsilon > r
\end{cases} \]

In the EM algorithm, as in Chapter 2, we also require the moments of the product \( \varepsilon^k Q^k \) for the complete data specification. For the failure case, \( \varepsilon = r \), the moments are easily obtained since (from 2.07) and (3.10)

\[ E_{\varepsilon Q}\left[ \varepsilon^k Q^k \right] = E_Q\left[ Q^k E_{\varepsilon/Q}(\varepsilon^r) \right] \]

\[ = r^\frac{k}{2} E_Q[Q^k] \]

where \( E_Q \) is the conditional expectation with respect to the distribution density \( p(q/\varepsilon) \).

For the censored case, \( \varepsilon > r \), the inner expectation \( E_{\varepsilon/Q} \) is no longer a constant, and is obtained from equation (3.10)
i.e. the expression for $E\left\{ e^\theta / Q = q, \varepsilon > r \right\}$. Then

$$E_{eQ}[e^k Q^{-1/\ell}] = E_Q\left[ Q^k E_{e/Q}(e^\theta) \right]$$

$$= E_Q\left\{ Q^k Q^{-1/\ell} I(\varepsilon, rQ^1/\ell) \left[ 1 - \phi(rQ^1) \right]^{-1} \right\}$$

(3.20)

where

$$I(\varepsilon, x) = \int_x^\infty u^x \phi(u) \, du$$

with reduction formula defined in (2.10). When $\ell = 1$, we have

$$I(1, rQ^1/\ell) = \phi(rQ^1),$$

and thus

$$E_{eQ}[e^k] = E_Q\left\{ Q^{-1/\ell} \phi(rQ^1) \left[ 1 - \phi(rQ^1) \right]^{-1} \right\}$$

$$= E_Q\left\{ Q^{-1/\ell} \phi(rQ^1) \right\}$$

$$= \frac{1}{d} \int q^{k-1/\ell} \phi(rq^1) \frac{p(q)}{1 - \phi(rq^1)} \cdot p(q) \left[ 1 - \phi(rq^1) \right] \, dq$$

(3.21)

where $d = M_o(r, 0)$, the normalization constant of the posterior distribution in (3.09) for the censored case. Then

$$E_{eQ}[e^k] = \frac{1}{d} \int q^{k-1/\ell} \phi(rq^1) \, p(q) \, dq$$

$$= \frac{M_1(r, k-1/\ell)}{M_o(r, 0)}.$$

Similarly

$$E_{eQ}[e^{2k}] = E_Q\left\{ Q^{k-1} [rQ^1 \phi(rQ^1) + I(0, z)] \right\}$$

$$\frac{1 - \phi(rQ^1)}{1 - \phi(rQ^1)}$$
since \( I(0,z) = 1 - \phi(z) \). Hence from (3.20)

\[
E_{\epsilon, Q}[\epsilon^2 Q] = r E_{\epsilon, Q}[\epsilon Q] + E_{\epsilon, Q}[Q^{-1}]
\]

since \( E_{\epsilon, Q}[Q^{-1}] = E_Q[Q^{-1}] \).

Extending the argument and letting, \( e_k = E_{\epsilon, Q}[\epsilon Q] \), it can be shown in general that \( E[\epsilon^2 Q] \) in (3.20) can be expanded as

\[
x^l e_k + (l-1) x^{l-3} e_{k-1} + \ldots + \left\{(l-1)(l-3)\ldots 2k-1\right\}, \quad \text{for } l \text{ odd } (3.22)
\]

and

\[
x^l e_k + (l-1) x^{l-3} e_{k-1} + \ldots + \left\{(l-1)(l-3)\ldots 3r e(k-\frac{1}{2} l+1)\right\} + \left\{(l-1)(l-3)\ldots 1.E[Q^{\frac{1}{2} l}]\right\}, \quad \text{for } l \text{ even } (3.23)
\]

for \( l \geq 1 \).

3.3.1 Summary of Results for Gamma Distribution.

Let

\[
M_1(r,k) = \int_0^\infty q^k \phi(q^\frac{1}{2}) p(q) dq
\]

and

\[
M_0(r,k) = \int_0^\infty q^k [1 - \phi(q^\frac{1}{2})] p(q) dq ,
\]

if \( p(q) \) is the density of a gamma distribution with parameter
(α, γ) then:

\[ M_1(r, k) = \frac{\Gamma(k+\alpha)}{\sqrt{\pi}} \cdot \frac{\gamma^\alpha}{(2\gamma + r^2)^{k+\alpha}} \] (3.24)

and

\[ M_0(r, k) = \frac{1}{2} \frac{\Gamma(a+k)}{\gamma^\alpha \Gamma(a)} \cdot \left\{ 1 - \frac{r}{\sqrt{2\pi\gamma}} \sum_{h=0}^{\infty} \left( \frac{2\gamma}{2\gamma + r^2} \right)^{h+1} \frac{\Gamma(h+1)}{\Gamma(h-1)} \right\} \] (3.25)

The posterior distribution of Q is

\[
p(q/c) = \frac{b^a q^{a-1} e^{-bq}}{\Gamma(a)} , \quad c = r
\]

and

\[
p(q/c) = \frac{[1 - \Phi(q)] \cdot p(q)}{M_o(r, 0)} , \quad c > r
\] (3.25)

where \( a = \alpha + \frac{1}{2} \) and \( b = \gamma + \frac{1}{2} r^2 \).

The conditional moments of Q are:

\[
E(Q^k) = \begin{cases} 
\frac{\Gamma(a+k+1)}{(\gamma+\frac{1}{2} r^2)^k \Gamma(a+\frac{1}{2})} , & c = r \\
\frac{M_o(r, k)}{M_o(r, 0)} , & c > r
\end{cases} \] (3.26)

The complete data moments of the product \( \varepsilon^k Q^k \), for \( k \geq 1 \) are

a) for the failures, \( c = r \),

\[ E(\varepsilon^k Q^k) = r^k E(Q^k) , \quad k \geq 0 \] (3.27)

b) for the censored \( c > r \)
\[ E[\varepsilon^k Q_k] = \begin{cases} 
\varepsilon_k^{\ell-1} e_k + (\ell-1)\varepsilon_k^{\ell-3} e_{k-1} + \ldots + \left\{ (\ell-1)(\ell-3)\ldots 2 \varepsilon_{k-\frac{1}{2}(\ell-1)} \right\} 
& \text{for } \ell \text{ odd} \\
\varepsilon_k^{\ell-1} e_k + (\ell-1)\varepsilon_k^{\ell-3} e_{k-1} + \ldots + \left\{ (\ell-1)(\ell-3)\ldots 3 \varepsilon_{k-\frac{1}{2}(\ell+1)} \right\} 
& \text{for } \ell \text{ even} \\
& \left\{ (\ell-1)(\ell-3)\ldots 1 E[Q_k^{\ell-2/2}] \right\} 
\end{cases} \]

for \( \ell \geq 1 \), \hspace{1cm} (3.28)

where \( e_k = E(\varepsilon^k Q_k) = M_1(r,k-\frac{1}{2})/M_0(r,0) \) and all the expectations are conditional on \( \varepsilon \).

3.4 ESTIMATION OF THE PARAMETERS BY EM

As was discussed in Chapter 2, the EM algorithm begins by obtaining the expectation of the complete data specification log likelihood, (the E-step) and then computing the values of the parameters that maximize this likelihood (M-step). For the scale mixtures of normal distributions described earlier in the chapter the likelihood, \( \ell_i \), for observation \( i \) is:

\[ \ell_i = f_{Y,Q}(y_i,q_i) = (2\pi\sigma^2)^{-\frac{1}{2}} q_i^\frac{1}{2} \exp \left[ -\frac{1}{2} \left( \frac{y_i - \mu_i}{\sigma} \right)^2 q_i \right] p(q_i) \]

\hspace{1cm} (3.29)

where \( f_{Y,Q}(y_i,q_i) \) is the joint distribution of \( Y \) and \( Q \). The log likelihood, \( L_i \), is thus

\[ L_i = -\frac{1}{2} \log(2\pi\sigma^2) + \frac{1}{2} \log(q_i) - \frac{1}{2} \left( \frac{y_i - \mu_i}{\sigma} \right)^2 q_i + \log(p(q_i)) \]

and for all the \( n \) observations:
\[ L_0 = \sum_{i=1}^{n} L_i = -\frac{n}{2} \log (2\pi \sigma^2) + \frac{1}{2} \sum_{i=1}^{n} \log(q_i) \]

\[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 q_i + \sum_{i=1}^{n} \log(p(q_i)) . \]  

Since \( p(q) \) is assumed known a priori, the parameters in \( L_0 \) that are to be estimated are \( \mu, \sigma^2 \). Putting \( \theta = (\mu, \sigma^2)^T \), the E-step (see Chapter 2, Section 2.4) is then:

\[ Q_e (\hat{\theta} / \theta \cdot p) = E \left[ L_0 / \theta \right] \]

\[ = -\frac{n}{2} \log(2\pi \sigma^2) + \frac{1}{2} \sum_{i=1}^{n} E \left[ \log(q_i) / \theta \right] \]

\[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} E \left[ (y_i - \mu)^2 q_i / \theta \right] \]

\[ + \sum_{i=1}^{n} E \left[ \log(p(q_i)) / \theta \right] \]  

(3.31)

where \( \hat{\theta} \) refers to a given set of parameters, \( (\mu, \sigma^2) \), and "\( \cdot \)" is to indicate the arguments capped thus are of the complete log likelihood \( L_0 \). The M-step is to maximize \( Q_e \) with respect to these parameters, the solution of which provides the next iterate \( \theta_{p+1} \).

Note here that the expectation \( E \) is with respect to the joint distribution of \( Y \) and \( Q \) conditional on the given data.
3.4.1 Estimation of $\mu$ in a Single-sample Model.

To obtain the estimate of $\mu$, (3.31) is differentiated with respect to $\bar{y}$. The normal equation obtained is thus:

$$\hat{\sigma}^{-2} \sum_{i=1}^{n} E \left[ (Y_i - \bar{y}) \frac{Q_i}{\theta_p} \right] = 0$$

(3.32)

which gives the next iterate as

$$\mu_{p+1} = \frac{\sum_{i=1}^{n} E[Y_i Q_i]}{\sum_{i=1}^{n} E[Q_i]}$$

(3.33)

noting that the values in the right-hand-side of (3.33) are all computed at values of the previous iterate $\mu_p$ and $\sigma_p^2$. The expectation for the Student-t error may be obtained directly from the previous section. In general we have:

$$E[Y_i Q_i] = E[(\mu + \sigma \varepsilon) Q_i]$$

$$= \mu E[Q_i] + \sigma E[\varepsilon Q_i]$$

(3.34)

For the failures (uncensored), $\varepsilon = r$ and from (3.27)

$$E[\varepsilon_i Q_i] = r_i E[Q_i].$$

Then

$$E[Y_i Q_i] = (\mu + \sigma r_i) E[Q_i]$$

$$= y_i E[Q_i], \quad Y_i = y_i$$

since $y = \mu + \sigma r$. 
For the censored, from (2.07)

\[ E[\varepsilon Q] = E_Q[Q \cdot E_{\varepsilon/Q}(\varepsilon)] , \]

where the expectations in the right hand side, \( E_Q \) and \( E_{\varepsilon/Q} \), are with respect to the posterior distribution of \( Q \) and distribution of \( \varepsilon \) conditional on \( Q=q \) and \( \varepsilon > r \). From (2.10) and (2.11)

\[ E_{\varepsilon/Q}[\varepsilon / \varepsilon > r] = q^{-\frac{1}{2}} S(r,q) \]

where \( S(r,q) = \phi(rq^{\frac{1}{2}})/(1 - \phi(rq^{\frac{1}{2}})) \).

Then \( E[\varepsilon_i Q_i] = E[Q_i^\frac{1}{2} S(r_i Q_i)] \) (3.35)

and equation (3.34) for the censored observation becomes

\[ E[Y_i Q_i] = \mu E[Q_i] + \sigma E[Q_i^\frac{1}{2} S(r_i Q_i)] \]

\[ = \left\{ \mu + \frac{E[Q_i^\frac{1}{2} S(r_i Q_i)]}{E[Q_i]} \right\} E[Q_i] . \]

Let \( w_i = E[Q_i] \) and

\[ y_i^* = \begin{cases} 
    y_i & \text{, uncensored } i \\
    \mu + \frac{\sigma E[Q_i^\frac{1}{2} S(r_i Q_i)]}{E[Q_i]} & \text{, censored } i 
\end{cases} \]

(3.36)

Then equation (3.33) can now be written as

\[ \mu_{p+1} = \frac{\sum_{i=1}^{n} y_i^* w_i}{\sum_{i=1}^{n} w_i} \]

(3.37)
which shows that \( \mu_{p+1} \) is obtained by weighted least squares with weights \( w_i \) and y-variate \( y^* \), computed from the previous iterate \( \mu_p \) and \( \sigma_p^2 \). Equation (3.37) thus provides us with an iterative scheme for the estimation of \( \mu \) by the EM algorithm. In fact it represents a general scheme for the estimation of \( \mu \) by the EM algorithm for scale mixture models whether continuous or discrete (compare equation (2.27)). The term involving the expectations in \( y^*_i \) in (3.36) is equivalent to \( E_i^* \) in the discrete mixture model. In the continuous model with known distribution of \( Q \), the weights \( w_i \) and the expectation terms can be computed directly, but in the discrete model with unknown \( Q \) as was in the analysis in Chapter 2, those values have to be estimated through the iteration.

For the Student-t error distribution, it can be shown from direct computation or from (3.27) and (3.28) that

\[
y^*_i = \begin{cases} 
  y_i, & \text{uncensored } i \\
  \mu + \frac{\sigma}{M_0(r_i,0)} E[Q_i], & \text{censored } i
\end{cases}
\]

This can be implemented quite easily in GLIM since the weights \( w_i = E(Q_i) > 0 \) (see section 3.2).

3.4.2 Estimation of the Scale Parameter \( \sigma^2 \).

The estimate of the scale parameter \( \sigma^2 \) may be obtained from the differentiation of equation (3.31) with respect to \( \sigma^2 \). This gives the normal equation:

\[
-\frac{n}{2\sigma^2} + \frac{1}{\sigma^4} \sum_{i=1}^{n} E[(y_i - \mu)^2 Q_i] = 0
\]
For the censored data $\tilde{\sigma}^2$ must satisfy

$$n \tilde{\sigma}^2 = \sum_U E[(Y_i - \tilde{\mu})^2 Q_i] + \sum_C [E(Y_i - \tilde{\mu})^2 Q_i]$$  \hspace{1cm} (3.39)

where the summations over $U$ and $C$ refer to the summations over uncensored and censored observations respectively. The parameters capped with "\_" refer to the present iterates. Again the moments are obtained from the relationships in (2.10) and (2.11). We have

$$E[(Y_i - \mu)^2 Q_i] = \sigma^2 E\left[\left(\frac{Y_i - \mu}{\sigma}\right)^2 Q_i\right]$$

$$= \sigma^2 E(\varepsilon_i^2 Q_i) .$$

For uncensored $i$, we have

$$E[(Y_i - \tilde{\mu})^2 Q_i] = \sigma^2 r_i^2 E[Q_i]$$

$$= (y_i - \tilde{\mu})^2 E[Q_i] .$$

For the censored observation

$$E[\varepsilon^2 Q] = E[Q E_{\varepsilon/Q}(\varepsilon^2)]$$

where the inner expectation in the right hand side of the equation is conditional on $\varepsilon > r$ for given $Q$. From (2.11)

$$E[\varepsilon^2] = q^{-1} \left[\frac{(rq^4) \phi(rq^4)}{1 - \phi(rq^4)} + 1\right]$$

$$= rq^{-\frac{1}{4}} S(r, q) + 1$$

and from (2.07)
E[ε^2 Q] = r E_Q[Q S(r, Q)] + 1.

Hence in general for censored data

\[ n \hat{\sigma}^2 = \sum_i (y_i - \tilde{\mu})^2 E[Q_i] + \sigma^2 \sum_i \left\{ 1 + r_i E[Q_i S(r_i, Q_i)] \right\} \]

(3.40)

where the expectations are conditional on \( \varepsilon = r \), for the failures and \( \varepsilon > r \) for the censored. \( \tilde{\mu} \) is the value of the mean at the present iterate while the values of the expectations and \( S(r_i, Q_i) \) and \( \sigma^2 \) are computed from the previous iterates.

For the Student-t error distribution (3.40) holds with

\[ E[Q S(r, Q)] = \frac{M_1(r_i)}{M_0(r, 0)}. \]

3.5 ESTIMATION OF REGRESSION PARAMETERS FOR THE LINEAR MODEL.

In the preceding sections we have established the EM algorithm for obtaining \( \mu \) and \( \sigma^2 \) for a given weighting distribution for \( Q \), for the single-sample model. As in Section 2.5 these results can be easily extended to the estimation of the regression parameters of linear models

\[ \mu_i = E[Y_i] = x_i^T \beta \]

where \( x_i \) is a vector of known regressors with elements \( (x_{i1}, x_{i2}, \ldots, x_{ip}) \) and \( \beta \) the unknown vector regression parameter of length \( p \) say.

Replacing \( \tilde{\mu} \) by \( u_i \) in (3.31) and differentiating with respect to \( \beta_k \), say, and equating the derivative to zero, we have
\[
\sigma^{-2} \sum_{i=1}^{n} E[Y_i - \sum_{k=1}^{p} \beta_k x_{ik}] x_{ik} Q_{ik} = 0, \quad k = 1, \ldots, p
\]

\[
\sum_{i=1}^{n} x_{ik} x_{ik}^T E[Q_{ik}] = \sum_{i=1}^{n} x_{ik} E[Y_i Q_{ik}], \quad k = 1, \ldots, p.
\]

Using the results of the previous sections, that is

\[ w_i = E[Q_{ik}], \quad \text{and,} \quad E[Y_i Q_{ik}] = y_i^* w_i \]

then these equations become

\[
\sum_{i=1}^{n} x_{ik} x_{ik}^T w_i = \sum_{i=1}^{n} x_{ik} w_i y_i^*, \quad k = 1, \ldots, p.
\]

In terms of the matrix notation, this may be written as

\[(X^T W X)^{-1} X^T W y^* \]

\[
\text{equation (3.41)}
\]

assuming \(X^T W X\) is non-singular,

\[ W = \text{diag} \{w_i\}, \quad y^* = (y_1^*, y_2^*, \ldots, y_n^*)^T \]

and \(X\) the design matrix with elements \(x_{ij}, \quad i = 1, 2, \ldots, n;\)
\(j = 1, 2, \ldots, p.\) The estimates of \(\sigma^2\) is given by (3.40) but the
\(\hat{\mu}\) is replaced by \(u_j.\)

Equation (3.41) represents iterative reweighted least squares.

For positive semi-definite \(W,\) it can be programmed quite easily
have listed the GLIM macros and an example run programme to carry
out the computations necessary for Student-\(t\) error distribution (equivalent to the weighting variable from that of gamma distribution) with even degrees of freedom (see also appendix C2). Examples and results from the algorithm applied to some real data are presented in section (3.8).

3.5.1 **Observed Information.**

Following Louis (1982), the observed information matrix when using the EM algorithm may be computed on the final iteration as

\[
I(\hat{\theta}) = E \begin{bmatrix} \frac{\partial^2 L_0}{\partial \hat{\theta} \hat{\theta}^T} & \left( \frac{\partial L_0}{\partial \theta} \right) \left( \frac{\partial L_0}{\partial \hat{\theta}} \right)^T \end{bmatrix} \tag{3.42}
\]

where \(L_0\) is the complete data log likelihood, (equation 3.30).

Pettitt (1985), found expressions for the scale mixtures in the linear model. These are

\[
- \frac{\partial^2 \ell_{ob}}{\partial \hat{\theta} \hat{\theta}^T} = \sigma^{-2} X^T V X \tag{3.43}
\]

\[
- \frac{\partial^2 \ell_{ob}}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{4\sigma^4} \sum_{i=1}^{n} \text{Var}(\varepsilon_{i1}) \tag{3.44}
\]

where \(\ell_{ob}\) is the log likelihood of the observed data, \(X\) is the design matrix in the linear model, \(\hat{\theta}\) the regression parameter and

\[
V = \text{diag} \left\{ E[Q_i] - \text{Var}(\varepsilon_{i1}) \right\} \tag{3.45}
\]

For uncensored observation \(\varepsilon_{i1} = r_{i1} = (y_{i1} - \mu_{i1})/\sigma\) is constant whereas for the censored we have to use the results of Section 3.3.
to obtain $E[\varepsilon_{i}^2 Q_{i}^2]$ and hence $\text{Var}[\varepsilon Q]$. For the Student-t error, from (3.28) we have for the censored observations

$$V = \text{diag} \left\{ [E(\varepsilon_{i} Q_{i})^2 - r_{i} E(\varepsilon_{i} Q_{i})^2] \right\}$$

(3.46)

where $E(\varepsilon_{i} Q_{i}^2) = M_{1}(r_{i}, \ell-1)/M_{0}(r_{i},0)$, for $\ell = 1,2$. For the failures

$$V = \text{diag} \left\{ E[Q_{i}] - r_{i}^2 \text{Var}[Q_{i}] \right\}
$$

the values of $E[Q_{i}]$ and $\text{Var}[Q_{i}]$ can be obtained directly from the gamma distribution with parameters $(a+\frac{1}{2}, \gamma+\frac{3}{2}r_{i}^2)$. Similarly (3.44) may be evaluated.

3.5.2 Finding Outliers.

In our analysis here we use the Student-t error distribution. The mode of the likelihood function obtained by the EM algorithm may not be global (see Pettitt, 1985, Section 5.2). As such we use different starting points to investigate this concern. Having established the maximum likelihood estimates, the next line of inquiry is the detection of outliers. Finding those cases that might be considered outliers in the finite mixture models is quite straightforward, by observing the posterior probabilities of component membership for each case. Student-t analysis in itself is a robust analysis (Fraser 1979, Section 2; O'Hagan, 1979 and West, 1984) which has the property of accommodating outliers. In our continuous mixtures model, like the finite mixtures, cases that do not conform to the hypothesised modelling process are
down-weighted, with weights $E(Q_i)$ where $Q_i$ is the scale mixtures variable from the gamma distribution and the expectation is with respect to the posterior distribution of $Q_i$ conditional on $\varepsilon_i = r_i$ or $\varepsilon_i > r_i$. It is thus reasonable to assume that a value of $E(Q_i)$ close to zero corresponds to that case which may be classified as an outlier, while values close to one refer to the observations being compatible with the normal $N(\mu_i, \sigma^2)$ model.

Pettitt (1985) on the suggestion of West (1984) considers the value of $E[Q_i] - \text{Var}(\varepsilon_i Q_i)$, which is required for the observed information in the previous section. If the value is negative then observation $i$ could be described as an outlier, (since the variance term, $\text{V}(\varepsilon_i Q_i) > 0$ and for an outlier $E(Q_i)$ is approximately zero).

3.6 NEWTON-RAPHSON METHOD.

Thus far in this thesis we have only discussed the use of the EM algorithm in obtaining the maximum likelihood estimates of the parameters in the single-sample or linear model for the scale mixtures model. When the marginal distribution of the error, $\varepsilon$, is known a more obvious and efficient approach is probably to use the Newton-Raphson technique for maximum likelihood estimation. We shall present, in this section, the computational aspect of the method and in later sections extend it to apply to influence assessment. In Section 3.7.3 we shall compare this scheme to that of the EM algorithm.

For an observation $Y_i$, we have the relationship $Y_i = \mu_i + \sigma \varepsilon_i$, and $\mu_i = E[Y_i] = x_i^T \beta$. Suppose that $\varepsilon_i$ has known density $f$ and
distribution function $F$. For continuous $f$, the likelihood
function for censored data (see Cox and Oakes, 1984, page 32; or
Lawless 1982) is:

$$\text{lik} = \prod_{i \in U} \frac{1}{\sigma} f\left(\frac{y_i - x_i^T \beta}{\sigma}\right) \prod_{i \in C} S\left(\frac{y_i - x_i^T \beta}{\sigma}\right)$$  \hspace{1cm} (3.47)

where $S = 1 - F$, the survival function, and $U$ and $C$ are sets of
uncensored and censored observations respectively. Letting
$r_i = (y_i - x_i^T \beta)/\sigma$ and $n_u$ be the total number of the uncensored
observations in the data, the log likelihood is thus:

$$L = -n_u \log \sigma + \sum_{i \in U} \log f(r_i) + \sum_{i \in C} \log S(r_i)$$  \hspace{1cm} (3.48)

and the maximum likelihood estimates of $\beta$ and $\sigma$ are solutions to:

$$\frac{\partial L}{\partial \beta} = -\frac{1}{\sigma} \sum_{i \in U} \frac{f'(r_i)}{f(r_i)} x_i + \frac{1}{\sigma} \sum_{i \in C} \frac{f(r_i)}{S(r_i)} x_i = 0$$

$$\frac{\partial L}{\partial \sigma} = -\frac{n_u}{\sigma} - \frac{1}{\sigma} \sum_{i \in U} \frac{f'(r_i)}{f(r_i)} r_i + \frac{1}{\sigma} \sum_{i \in C} \frac{f(r_i)}{S(r_i)} r_i = 0$$  \hspace{1cm} (3.49)

respectively, where $f'$ is the derivative of $f$. Let

$$\psi_i = \begin{cases} -\frac{f'}{f}, & \text{for uncensored } y_i, \\ \frac{f}{S}, & \text{for censored } y_i, \end{cases}$$  \hspace{1cm} (3.50)

then $\hat{\beta}$, the maximum likelihood estimate of $\beta$, satisfies

$$\sum_{i=1}^{n} x_i \psi_i(r_i) = 0$$  \hspace{1cm} (3.51)
Let \( U(\beta) = \frac{\partial L}{\partial \beta} \), be the score vector of length \( p \),
and \( G(\beta) = -\frac{\partial^2 L}{\partial \beta \partial \beta^T} \), be the \( p \times p \) observed information matrix.

The first order expansion of the Taylor's series for a given trial value \( \beta_t \), on the score statistic at \( \beta \) is

\[
U(\hat{\beta}) = U(\beta_t) - G(\beta_t)(\hat{\beta} - \beta_t) .
\]

At maximum likelihood estimate \( \hat{\beta} \), \( U(\hat{\beta}) = 0 \), and solving the above gives:

\[
\beta_{t+1} = \beta_t + \left[ G(\beta_t)^{-1} \right] U(\beta_t) \quad (3.52)
\]

where \( \beta_{t+1} \) is the next trial value substituting the \( \hat{\beta} \) in the expansion equation above. This gives the iterative process until \( \beta \) estimates agree to a specified extent, and of course, \( U(\hat{\beta}) = 0 \) at convergence. With the definitions in (3.49) and (3.50) we have

\[
U(\beta) = \frac{1}{\sigma} \sum_{i=1}^{n} x_i \psi_i(r_i) ,
\]

with components

\[
U_j(\beta) = \frac{1}{\sigma} \sum_{i=1}^{n} x_{ij} \psi_i(r_i) , \quad j = 1,2,\ldots,p .
\]

Then the \((j,k)\) element of \( G \) is

\[
G_{jk} = -\frac{\partial U_j(\beta)}{\partial \beta_k} = \frac{1}{\sigma^2} \sum_{i=1}^{n} x_{ij} \psi'(r_i) x_{ik}
\]

where \( \psi'(r_i) \) is the derivative of \( \psi(r_i) \) with respect to \( r_i \).

In matrix notation we have
where \( X \) is the design matrix in the linear model,

\[
\mathbf{W}^* = \text{diag} \{\psi'(r_1)\} \text{ and } s \text{ has components } (\psi_1, \psi_2, \ldots, \psi_n)^T.
\]

Equation (3.52) can now be written

\[
\beta_{t+1} = \beta_t + \sigma \left[ X^T W^* X \right]^{-1} X^T s
\]  

(3.55)

with \( \sigma, W^* \) and \( s \) evaluated at the current estimate of \( \beta_t \). The scale parameter \( \sigma \) can be updated from the separate normal equation in (3.49) which from the transformation in (3.50) gives

\[
-\frac{n_u}{\sigma} + \frac{1}{\sigma} \sum_{i=1}^{n} r_i^2 \psi_i'(r_i) = 0
\]

which again can be solved iteratively, by the Newton-Raphson technique described above. The updating formula is given by

\[
\sigma_{t+1} = \sigma_t - \left[ V(\sigma_t) \right]^{-1} M(\sigma_t)
\]  

(3.56)

where

\[
M(\sigma_t) = -\frac{n_u}{\sigma_t} + \frac{1}{\sigma_t} \sum_{i=1}^{n} r_i^2 \psi_i(r_i)
\]

and

\[
V(\sigma_t) = \sigma_t^{-2} \left[ n_u - \sum_{i=1}^{n} \left[ r_i^2 \psi_i'(r_i) + 2r_i \psi_i(r_i) \right] \right]
\]

where \( \psi'(r_i) \) is the derivative of \( \psi(r_i) \) with respect to \( r_i \), and \( r_i, \psi(r_i) \) and \( \psi'(r_i) \) are evaluated at the current estimates of \( \beta_t \) and \( \sigma_t \).
The updating formulae (3.52) and (3.56) could be incorporated into one formula by defining the $(p+1) \times 1$ score vector

$$U_{\theta}(\theta) = \begin{bmatrix} U(\theta) \\ M(\theta) \end{bmatrix}$$

where $\hat{\theta} = (\hat{\beta}, c)^T$, and applying the technique to give the updating formula

$$\hat{\theta}_{t+1} = \hat{\theta}_t + G(\hat{\theta}_t)^{-1} U_{\theta}(\hat{\theta}_t)$$

where $G_{\theta}(\theta) = -\frac{\partial^2 L}{\partial \theta \partial \theta^T}$, the $(p+1) \times (p+1)$ observed information matrix. Here we purposely present in two separate formulae in view of our intention to compare in later sections results found by this method and by the EM algorithm.

3.7 ASSESSMENT OF INFLUENCE.

Regression diagnostic techniques for detecting outlying and influential observations are well known for linear and generalized linear models (Belsley, Kuh and Welsh, 1980). In sections 2.8 and 3.5.1 we have presented some ad-hoc rules for the detection of outliers, which in general undermine the best fit for the bulk of the data. Here we shall discuss a diagnostic method in the detection of influential cases in the data set, for the continuous scale mixtures model which could also be extended to include the finite mixtures model of Chapter 2.

In general the study of influence is a study of stability
and robustness in the output elements of our analysis when the problem formulation is modified. The idea is to introduce small perturbations in the formulation and study how they change the outcome of the analysis. Various perturbation schemes (see Cook and Weisberg, 1982, Chapters 3 and 4) may be introduced, but in the following sections we shall only consider one perturbation scheme, namely that the data are modified by deleting cases either one at a time or in groups. Case deletion diagnostics have found great acceptance and have been applied in many statistical models. Cases which give great change in the outcome of the analysis are considered to exert high influence in the model. In ordinary least squares with design matrix $X$, and regression parameter $\beta$, an exact solution to the change, $\Delta \beta$, is known, and the projection matrix, called the hat matrix

$$H = X(X^T X)^{-1} X^T$$

plays an important role (Cook, 1977, 1979). For single deletion with $k^{th}$ case omitted from the data, this change is given by:

$$\Delta_{k} \hat{\beta} = \hat{\beta} - \hat{\beta}(k) = (X^T X)^{-1} x_k r_k \overline{m_{kk}}$$

(3.57)

where $\hat{\beta}$ and $\hat{\beta}(k)$ are least squares estimates of $\beta$ when there is presence or absence of the $k^{th}$ case in the data, respectively, and $x_k$ corresponds to the vector in the $k^{th}$ row of $X$ and $m_{kk}$ is the $k^{th}$ diagonal element of $I-H$. In non-linear regression, a geometric approach was suggested by Moalgovkar et al. (1984) which presents the construction of a matrix analogous to the hat matrix of the linear least square regression. Pregibon (1981) considered
logistic regression and derived an approximation to $\Delta \hat{\beta}$, where $\hat{\beta}$ is the maximum likelihood estimate of $\beta$. By analogy with ordinary least squares his technique yields a matrix that plays the role of $H$.

For single case deletion, if $\hat{\beta}^{(k)}$ denotes the estimate of $\beta$ based on the sample size $(n-1)$ with the $k^{th}$ case deleted then the sample influence curve (SIC) (see, for example, Cook and Weisberg, 1982) is given by

$$\text{SIC}_k = (n-1)(\hat{\beta} - \hat{\beta}^{(k)})^T.$$  \hspace{1cm} (3.58)

As mentioned above $\text{SIC}_k$ for the standard least square regression can be computed exactly (3.57), but for the mixture models under discussion here, the $\hat{\beta}$'s are found by iterative numerical techniques. This presents a cumbersome task since apart from the computation needed to obtain $\hat{\beta}$, we still require to calculate each $\hat{\beta}^{(k)}$, $k = 1, 2, \ldots, n$. To reduce computation, approximations to $\hat{\beta}^{(k)}$ are of great value. An obvious choice in such a context is the 1-step technique, that is, to compute from the maximum likelihood estimate $\hat{\beta}$, the first step of an iterative process to find $\hat{\beta}^{(k)}$. The 1-step approximations for both EM algorithm and Newton-Raphson method will be discussed in a later section and their relationship will be presented.

To quantify this influence we use the Cook distance (Cook, 1977, 1979; Cook and Weisberg, 1982, section 3.5)

$$D^{(k)}(M, c) = \frac{(\hat{\beta} - \hat{\beta}^{(k)})^T M (\hat{\beta} - \hat{\beta}^{(k)})}{c}.$$  \hspace{1cm} (3.59)
where $M$ is a symmetric (semi-) positive definite matrix and $c$ some positive scale factor. Various $M$ and $c$ have been suggested with their own special interpretation. In robust regression, based on the reweighted least squares approach to computation of $\hat{B}$, the inverse of the variance-covariance matrix of the full model at $\hat{\theta}$ may be used as the $M$ matrix while $c$ can be taken as $p \times \hat{o}$, (Cook and Weisberg, 1982), where $p$ is the number of components of $\hat{\theta}$.

A more general approach is to obtain a measure derived from the contours of the log likelihood function. Let $L(\theta)$ be the log likelihood on parameter $\theta$ based on the full data. The likelihood distance (Cook and Weisberg, 1982) is defined as

$$LD_k = 2\left[L(\hat{\theta}) - L(\hat{\theta}_{(k)})\right]$$  \hspace{1cm} (3.60)

and when using a 1-step approximation $\hat{\theta}_{(k)}$,

$$LD_{k}^{1} = 2\left[L(\hat{\theta}) - L(\hat{\theta}_{(k)})\right]$$  \hspace{1cm} (3.61)

These measures $LD_k$ and $LD_{k}^{1}$ may also be interpreted in terms of asymptotic confidence regions (Cox and Hinkley, 1974, Chapter 9)

$$\left\{ \theta : 2[L(\hat{\theta}) - L(\theta)] < \chi_{a,q}^{2} \right\}$$  \hspace{1cm} (2.62)

where $\chi_{a,q}^{2}$ is the upper $a$ point of chi-squared distribution with $q$ degrees of freedom, and $q$ is the dimension of $\theta$. $LD_k$ can therefore be calibrated by comparison to the $\chi_{q}^{2}$ distribution. If the log likelihood contours are approximately elliptical, $LD_k$ may be approximated by a Taylor expansion of $L(\hat{\theta}_{(k)})$ around $\hat{\theta}$.
thus:

\[ L(\hat{\theta}_{(k)}) \approx L(\hat{\theta}) + (\hat{\theta}_{(k)} - \theta)^T \tilde{L}(\hat{\theta}) + \frac{1}{2} (\hat{\theta}_{(k)} - \hat{\theta})^T \tilde{L}(\hat{\theta})(\hat{\theta}_{(k)} - \hat{\theta}) \]

where \( \tilde{L}(\cdot) = \frac{\partial L}{\partial \theta} \) and \( \tilde{L}(\cdot) = \frac{\partial^2 L}{\partial \theta \partial \theta^T} \),

are the score vector and curvature matrix, and since \( \tilde{L}(\hat{\theta}) = 0 \), at maximum likelihood, we have

\[ LD_k \approx (\hat{\theta}_{(k)} - \hat{\theta})^T A(\hat{\theta}) (\hat{\theta}_{(k)} - \hat{\theta}) \]

where \( A(\hat{\theta}) = -\tilde{L}(\hat{\theta}) \), the observed information matrix.

The likelihood distance can be easily modified to accommodate the situation in which a subset \( \theta_1 \) of \( \theta \) is of special interest. Details of this are found in Cook and Weisberg, (1982, Section 5.2).

3.7.1 1-Step EM Algorithm.

The general arguments that follows are applicable in both the scale mixtures models, finite, discrete or continuous, but examples in Section 3.8 will be given only for the continuous model with Student-t error distribution. The 1-step approximation, \( \hat{\theta}_{(k)}^1 \), to the estimate \( \hat{\theta}_{(k)} \), defined to be the maximum likelihood estimate of \( \theta \) with the \( k \)th case deleted, is formed by taking one step of the iterative process for finding \( \hat{\theta}_{(k)} \) by starting at \( \hat{\theta} \).

Following (3.4.1) the normal equation for the iterative procedure for the scale mixture model with \( k \)th case omitted from the data is
\[ X^T (k) W(k) X(k) \hat{\beta}(k) = X^T (k) W(k) \hat{\gamma}^*(k) \]

where \( X(k) \) and \( W(k) \) are the design, \( X \), and weight matrices, \( W \), (defined in Section 3.4) respectively with the \( k \)th row omitted; and \( \hat{\gamma}^* \) is the response vector with elements \( \hat{y}^*_i \) of (3.36) with \( k \)th component omitted; \( W(k) \) and \( \hat{\gamma}^*(k) \) are evaluated at the current iterate.

Now consider \( W \) and \( \hat{\gamma}^* \) evaluated at the fully iterated estimates of \( \hat{\beta} \) and \( \hat{\sigma} \) for the full data with design matrix \( X \), then isolating the \( k \)th case, we can partition the appropriate matrices as follows:

\[
X = \begin{bmatrix} X(k) \\ X_k^T \end{bmatrix} \quad (n-1) \times p
\]

\[
W = \begin{bmatrix} W(k) & 0 \\ 0 & w_k \end{bmatrix}
\]

and \( \hat{\gamma}^* = \begin{bmatrix} \hat{\gamma}_k^T \\ \hat{\gamma}_k^T \end{bmatrix} \)

where \( X(k) \), \( W(k) \) and \( \hat{\gamma}^*(k) \) are defined above and \( X_k \) is a vector corresponding to the \( k \)th row of \( X \) and \( \hat{\gamma}_k^* \) is the \( k \)th element of \( \hat{\gamma}^* \). Then the \( X^T W X \) term in equation (3.41) can be written as

\[
X^T W X = \begin{bmatrix} X^T (k) & X_k \end{bmatrix} \begin{bmatrix} W(k) & 0 \\ 0 & w_k \end{bmatrix} \begin{bmatrix} X(k) \\ X_k^T \end{bmatrix}
\]

\[
= X^T (k) W(k) X(k) + X_k W_k X_k^T
\]
that is
\[ X^{T} W(k) X(k) = X^{T} W X - x_{k} w_{k} x_{k}^{T} \]
and from Appendix A1, we obtain
\[
\begin{bmatrix} X^{T}(k) & W(k) & X(k) \end{bmatrix}^{-1} = (X^{T} W X)^{-1} + \frac{(X^{T} W X)^{-1} x_{k} w_{k} x_{k}^{T} (X^{T} W X)^{-1}}{1 - a_{k}}
\]
where \( a_{k} = x_{k}^{T} (X^{T} W X)^{-1} x_{k} w_{k} \).

Similarly the right hand side of equation (3.41) can be written as
\[
X^{T}(k) W(k) y^{*}(k) = X^{T} W y^{*} - x_{k} w_{k} y^{*}_{k}.
\]
Writing \( A = (X^{T} W X) \), then from (3.64), the next iterate gives us the 1-step estimate of \( \hat{\beta}(k) \) as
\[
\hat{\beta}(k) = A^{-1} x_{k} w_{k} y^{*}_{k}.
\]
Now \( x_{k}^{T} \hat{\beta} = \hat{\mu}_{k} \) and \( w_{k} = E(Q_{k}) \), so
\[
w_{k} (x_{k}^{T} \hat{\beta} - y^{*}_{k}) = E(Q_{k}) [\hat{\mu}_{k} - y^{*}_{k}].
\]
For \( k \) censored, from (3.35) and (3.36), the term in (3.64) is equal to \(- \sigma E(c_{k} Q_{k})\).

For uncensored, we have from (2.07) and (3.10)
\[ E[\varepsilon_k Q_k] = r_k E(Q_k) \]
\[ = \left( \frac{y_k - \hat{\mu}_k}{\hat{\sigma}} \right) E(Q_k) \]
since, for the uncensored case, \( y_k^* = y_k \). Thus the term in (3.64) also equals \(- \sigma E(\varepsilon_k Q_k)\). Hence the 1-step EM formula for the estimation of \( \beta(k) \) for \( k \) both censored or uncensored, is given by:

\[
\beta_{(k)}^{1} = \hat{\beta} - \frac{\sigma E[\varepsilon_k Q_k]}{1 - a_k} \left[ (X^T W X)^{-1} X_k \right] \tag{3.65}
\]

where \( a_k = x_k^T (X^T W X)^{-1} x_k w_k \). Where appropriate the terms in the right hand side of (3.65) are all evaluated at the fully iterated estimates, \( \hat{\beta} \) and \( \hat{\sigma} \), for the full data. In summary the 1-step estimate can be computed directly at the final iteration for the full data, hence 1-step influence curves or measures may be obtained. GLIM listing by Pettitt and bin Daud, (1986), may easily be adapted to carry out the computation of (3.65).

However for ease of manipulation of matrices, the computation in our analysis is programmed in fortran; (see Appendix C2). Results from some real data are given in Section 3.8.

### 3.7.2 1-Step Newton-Raphson Method.

The iterative scheme for the Newton-Raphson is given in equation (3.55). Using the same arguments as in the preceding section, the 1-step estimate of \( \beta_{(k)} \), starting from the fully iterated estimate \( \hat{\beta} \) for the full data is

\[
\beta_{(k)}^{1*} = \hat{\beta} + \hat{\sigma} (X_{(k)}^T W_{(k)}^* X_{(k)})^{-1} X_{(k)} \beta_{(k)} \tag{3.66}
\]
where the subscripts carry the same meaning used in the above section, and the notation $\hat{\beta}_{(k)}^{1*}$ is used here to distinguish this estimate from that of the 1-step EM algorithm. Representing the subscripted matrices in the right hand side of (3.66) in terms of their full data counterparts as previously, in the preceding section, equation (3.66) then becomes:

$$
\hat{\beta}_{(k)}^{1*} = \hat{\beta} - \frac{\hat{\sigma}(X^T W^* X)^{-1} x_k^* s_k}{1 - a_k^*}
$$

(3.67)

where

$$
a_k^* = x_k^T (X^T W^* X)^{-1} x_k^* w_k^*,
$$

$$
W^* = \text{diag}[\psi'(r_k)], \quad w_k = \psi'(r_k)
$$

where $\psi$ is the function defined in (3.50).

Similarly, by defining for the $k$th observation, an indicator $\delta_k = 1$ if it is a failure and $\delta_k = 0$ if it is censored, $\hat{\sigma}_{(k)}$ may be approximated by $\hat{\sigma}_{(k)}^{1*}$ where

$$
\hat{\sigma}_{(k)}^{1*} = \hat{\sigma} - \left[V_N(\hat{\sigma})\right]^{-1} M_N(\hat{\sigma})
$$

and

$$
M_N = M - \left[x_k^* \hat{s}_k - \delta_k\right] / \hat{\sigma},
$$

$$
V_N = V + \hat{\sigma}^{-2} \left[2x_k^2 w_k + 2r_k s_k - \delta_k\right].
$$

Equation (3.67) is equivalent to that of the robust approach of Cook and Weisberg, (1982), their equation (5.5.5). A precise characterization of the accuracy of this 1-step approximation is, as yet, unavailable. There is no guarantee that $\hat{\beta}_{(k)}^{1*}$ is closer
to $\hat{\beta}(k)$ than $\hat{\beta}$, but if the likelihood function is approximately quadratic in the neighbourhood of $\hat{\beta}$ than $\hat{\beta}^{**}$ should be a sufficiently accurate 1-step approximation.

3.7.3 Comparison between 1-Step EM and 1-Step Newton-Raphson.

The Newton-Raphson method is only applicable when the error distribution is known specifically and the likelihood function is the direct manifestation of this distribution. The method is general for estimating the regression parameters in the linear model for censored data. The EM algorithm discussed here, however, only applies specifically to the scale mixtures of normal distribution models, that is, for a given weighting variable $Q$, the distribution of the error, $\epsilon$, conditional on $Q$ is normal with mean zero and variance $q^{-1}$. In the case of finite discrete mixture the distribution of $Q$ may be unknown, but results established in this chapter for continuous $Q$, require the distribution of $Q$ to be known. It is possible to incorporate into the analysis where $Q$ has a gamma distribution $G(\alpha, \alpha)$, a priori, the situation where $\alpha$ is unknown. In terms of values $q_j$ of $Q_j$ we have the log likelihood

$$
\text{loglik}(\alpha) = (\alpha-1) \sum \log q_j + n \alpha \log \alpha
- \alpha \sum q_j - n \log \Gamma(\alpha)
$$

and thus to implement the EM algorithm we need the "posterior" means $E(\log Q_j | \text{data, parameters})$, $j = 1, \ldots, n$ as well as $E(Q_j | \text{data, parameters})$, $j = 1, \ldots, n$ which are already required. Thus to iterate an $\alpha$ as well as the regression parameters would require solution to
\[ \frac{d}{d\alpha} \log \text{lik}(\alpha) = 0 \]

which gives an added complexity. Additionally the explicit results for \( E(Q_j \mid \text{data, parameters}) \) are only available for integer \( \frac{a}{2} \), so that the terms \( E(Q_j \mid \text{data, parameters}) \) would only be found for non-integer \( \frac{a}{2} \) using numerical quadrature or approximated using interpolation using the values for integer \( \frac{a}{2} \). It is simpler to estimate \( \alpha \), if unknown, by using a profile likelihood in \( \alpha \).

There is a relationship between the weights used in the EM and Newton-Raphson methods.

The observed information matrix from the likelihood of the Newton-Raphson approach (Section 3.6, equation (3.53)) is

\[ G(\beta) = \sigma^{-2} \mathbf{W}^* \mathbf{X} \]

where \( \mathbf{W}^* = \text{diag}\{\psi'(r_i)\} \) with \( \psi \) defined in (3.50). Comparing this to that obtained by Pettitt (1985) for the scale mixtures when \( \varepsilon \) is distributed with density

\[ f(\varepsilon) = \int_{0}^{\infty} q^{1/2} \phi(\varepsilon q^{1/2}) p(q) \, dq \]

then

\[ \mathbf{W}^* = \text{diag} \left[ E(Q_k) - \text{Var}(Q_k) \right] \]

which implies \( w_k^* = w_k - v_k \) \( (3.68) \)

where \( w_k = E(Q_k) \) the weight component used in the EM algorithm and \( v_k = \text{Var}(Q_k) > 0 \). Equation (3.68) gives the relationship for weights of the two methods.

For an outlier, \( w_k \) is approximately zero, (Section 2.8).
From (2.68) and from the definition of $a_k^*$ in (3.67), assuming the matrix $(X^T W X)^{-1}$ is positive definite, $w_k^*$, the weight component for the suspect case in the Newton-Raphson scheme is negative. Thus for the Newton-Raphson 1-step $a_k^*$ may be used as an ad hoc criterion to detect outliers in the data set, depending on whether its value is positive or negative.

With $w_k \approx 0$ an estimate of $\hat{\theta}$ obtained by the EM algorithm is approximately equivalent to $\hat{\theta}^{(k)}$. If the 1-step EM estimate is not too far from the $\hat{\theta}^{(k)}$, then from (3.65), with $a_k = 0$

$$\hat{\theta} - \hat{\theta}^{(k)} = G_k (X^T W X)^{-1} X_k$$

$$\approx 0$$

where $G_k = \hat{\sigma} \cdot E[\varepsilon_k Q_k]$. Thus we see that such extreme points in the data space are not influential in the analysis. Other cases with $E(Q)'s$ greater than zero can have influence and further investigations through influence curves or their measures need to be carried out, (see, for examples and results in Section 3.8).

Two scalar measures to quantify the influence are employed in our analysis, the Cook distance, $D_k(M,c)$, equation (3.59), which we shall call the quadratic norm, and the likelihood distance of equation (3.60) or for the 1-step estimate, equation (3.61).

The 1-step Newton-Raphson estimate is effectively the same as that obtained by maximizing the quadratic approximation of $L_{(k)}(\hat{\theta})$, the log likelihood obtained after deleting the $k^{th}$ case, about the point $\hat{\theta}$, the maximum likelihood estimate of the parameter $\theta$, for the full model, that is
\[ L_{(k)}(\theta) = L_{(k)}(\hat{\theta}) + (\theta - \hat{\theta})^T \dot{L}_{(k)}(\hat{\theta}) \frac{1}{2} (\theta - \hat{\theta})^T \ddot{L}_{(k)}(\hat{\theta})(\theta - \hat{\theta}) \]  

(3.69)

where \( L_{(k)}(\theta) \) is the score vector with the jth element \( \frac{\partial L_{(k)}(\theta)}{\partial \theta_j} \) evaluated at \( \theta = \hat{\theta} \), and the \( \ddot{L}_{(k)}(\hat{\theta}) \) has (i,j)th element \( \frac{\partial^2 L_{(k)}(\hat{\theta})}{\partial \theta_i \partial \theta_j} \) evaluated at \( \hat{\theta} \). Equation (3.69) is maximized at \( \hat{\theta}^{\dagger}_{(k)} \), say where

\[ \hat{\theta}^{\dagger}_{(k)} = \hat{\theta} - \left[ \dot{L}_{(k)}(\hat{\theta}) \right]^{-1} \ddot{L}_{(k)}(\hat{\theta}) \]  

(3.70)

assuming \( - \ddot{L}_{(k)}(\hat{\theta}) \) is positive definite. The estimate \( \hat{\theta}^{\dagger}_{(k)} \) for the Newton-Raphson method in (3.66) is equivalent to that in equation (3.70); since from (3.53) and (3.54) we have:

\[ - \dot{L}_{(k)}(\hat{\theta}) = \frac{1}{\hat{\sigma}^2} X_{(k)} W^*_{(k)} X_{(k)} \]

and

\[ \ddot{L}_{(k)}(\hat{\theta}) = \frac{1}{\hat{\sigma}} X_{(k)}^T S_{(k)} \]

The function \( L(\hat{\theta}) \) or \( L_{(k)}(\hat{\theta}) \) may be multimodal, since it may not be concave. Hence there is a danger using the approximation (3.70) that the Newton-Raphson method leads to a value \( \hat{\theta}^{\dagger}_{(k)} \) such that

\[ L_{(k)}(\hat{\theta}^{\dagger}_{(k)}) < L_{(k)}(\hat{\theta}) \]

i.e. \( \hat{\theta}^{\dagger}_{(k)} \) is a worse approximation to \( \hat{\theta}_{(k)} \) than \( \hat{\theta} \). However if \( \hat{\theta}_{(k)} \) is not too different from \( \hat{\theta} \), and \( L_{(k)}(\hat{\theta}) \) is locally quadratic, the 1-step estimator \( \hat{\theta}_{(k)} \) should be close to the fully iterated value, \( \hat{\theta}_{(k)} \). But for cases that are influential, where \( \hat{\theta} - \hat{\theta}_{(k)} \) is large, the accuracy of the 1-step estimator is likely to be lower. On the other hand an accurate approximation of \( \hat{\theta}_{(k)} \) will not be needed as long as \( \hat{\theta} - \hat{\theta}^{\dagger}_{(k)} \) is sufficiently large to draw our attention for further consideration.
The EM algorithm however, ensures that the next step in the algorithm always increase the likelihood (Theorem 1, Dempster et al. 1977, page 7). Thus we have

\[ L_k(\hat{\theta}) \leq L_k(\hat{\theta}_1) \leq L_k(\hat{\theta}) \]

for a 1-step EM estimate of \( \hat{\theta}_1 \), \( \hat{\theta}_2 \) starting from \( \hat{\theta} \).

### 3.7.4 Computation of Quadratic Norm.

The Cook distance or the quadratic norm, provides us with a scalar measure to quantify the influence. It is defined as in equation (3.59). Here we shall use

\[ D_k(M,c) = \frac{(\hat{\theta} - \hat{\theta}^*)^T M (\hat{\theta} - \hat{\theta}^*)}{c} \]

where \( \hat{\theta}^* \) is an parameter estimate of \( \hat{\theta}_1 \) or its approximation. If \( \hat{\theta}^* \) is a 1-step estimate we shall denote the measure as

\[ D_1^1(M,c) \].

In our analysis, \( M = X^T W X \), where \( W = \text{diag} \{ E(Q_i) \} \) for the EM algorithm, \( W = \text{diag} \{ \psi'(r_i) \} \) with \( r_i = (y_i - \mu_i)/\sigma \) for the NR method. \( c \) is taken to be equal to \( p\sigma^2 \), where \( p \) is the dimension of \( \theta \) and \( \sigma^2 \) is the maximum likelihood estimate of \( \sigma^2 \).

### 3.7.5 Computation of the Likelihood Distance.

The likelihood distance is a more general measure to quantify the influence of a case or subset of cases. The computation is straightforward. For single-case deletion equations (3.60) and (3.61), (or (3.63) for the approximation) may be used. However, equation (3.60) may prove to be cumbersome and expensive, since
we have to calculate each \( \hat{\beta}(j) \), \( j = 1, 2, \ldots, n \) on top of the iterations needed for the computation. We find that the 1-step measure, \( L_{D_k} \) in equation (3.61) is generally quite satisfactory, as shown in the example plot in Figure 3.1, for the motorette data.

If the log likelihood contours are approximately elliptical, then the approximation in (3.63) is useful. For parameter \( \beta \):

\[
L_{D_k} \approx (\hat{\beta}(k) - \hat{\beta})^T A(\hat{\beta}) (\hat{\beta}(k) - \hat{\beta})
\]

where \( A(\hat{\beta}) \) is the observed information matrix given in Section 3.5.1. However, a good approximation of \( \hat{\beta}(k) \) is needed for the \( L_{D_k} \) to be satisfactory. If the contours of the log likelihood function is markedly non-elliptical, Cook and Weisberg (1982) find that the measure can be seriously misleading.

The diagnostic measures above can now be compared to \( \chi_p^2 \), where \( p \) is the dimension of the parameter \( \beta \).

3.7.6 Mean Shift Outlier Model.

We have established in Section 3.7.3 that outliers are not influential in our analysis for the scale mixtures, due to down-weighting of these points. Thus a value of \( E(Q_k) \) close to zero implies that the \( k^{th} \) case is suspect. Another approach used to study outliers is the mean shift outlier model (see Cook and Weisberg, 1982) sometimes, called the augmented model. For the suspect case \( k \), say,

\[
E(Y) = X\hat{\beta} + d_k \phi
\]

(3.71)
where \( \mathbf{d}_k \) is an \( n \times 1 \) vector of zeros except for the \( k \)th element which has unit value, and \( \phi \) is the augmenting parameter. Non-zero values of \( \phi \) imply that the \( k \)th case is an outlier. For the linear regression model, the usual statistic to test \( \phi = 0 \) is that as given by Cook and Weisberg (1982)'s equation (2.2.15), page 21, which is compared to the Student-\( t \) distribution, with \((n-p-1)\) degrees of freedom, where \( n \) and \( p \) are the number of observations and the number of regression parameters respectively.

An alternative statistic based on the likelihood ratio and its asymptotic distribution may be used to test the hypothesis that \( \phi = 0 \). In terms of our outlier model (3.71), to test the hypothesis \( \phi = 0 \), this statistic is equivalent to

\[
LD_{ok} = -2\left(L(\mathbf{\hat{\beta}}) - \max L_{ok}(\mathbf{\beta}, \phi)\right) \tag{3.72}
\]

where \( L(\cdot) \) is the log likelihood for model \( \mathbb{E}(Y) = \mathbf{X}\mathbf{\beta} \), and \( L_{ok}(\cdot) \) is the log likelihood for the outlier model (3.71) with suspect case \( k \).

Introducing the censoring indicator \( \delta_k \), where \( \delta_k = 0 \) if \( k \)th observation is censored, and \( \delta_k = 1 \) if the \( k \)th observation is uncensored, the likelihood equation from (3.47) for the outlier model for \( n \) independent observations is

\[
= \prod_{i=1}^{n} \left[ \frac{1}{\sigma} f(r_i) \right]^{\delta_i} \left[ S(r_i) \right]^{1-\delta_i} \tag{3.73}
\]

where

\[
r_i = \begin{cases} 
\frac{1}{\sigma} (y_i - \mathbf{x}_i^T \mathbf{\hat{\beta}}) & i \neq k \\
\frac{1}{\sigma} (y_i - \mathbf{x}_i^T \mathbf{\beta} - \phi) & i = k.
\end{cases}
\]
Let \( n_u = \sum_{i=1}^{\delta_i} \), the number of the uncensored observations and \( L(k)(\cdot) \) be the log likelihood for model \( E(Y) = X\beta \) with the \( k \)th case deleted from the data set.

For a censored case:

\[
L_{ok}(\beta, \phi, \sigma) = - n_u \log \sigma + \sum_{i=1}^{n} \delta_i \log f(r_i) + \sum_{i \neq k}^{n} (1-\delta_i) \log S(r_i) + \log S(r_k)
\]

\[
= L(k)(\beta) + \log S(r_k)
\]

with \( r_k = (y_i - x_i^T \beta - \phi) \). For any fixed \( \beta \) and \( \sigma \), \( S(r_k) \) is maximized with respect to \( \phi \) by letting \( \phi \to \infty \) so that \( S(r_k) + 1 \).

Hence from the above equation

\[
\max L_{ok}(\beta, \phi, \sigma) = L(k)(\beta) \ .
\]  \hspace{1cm} (3.74)

For uncensored \( k \),

\[
L_{ok}(\beta, \phi, \sigma) = -(n_u - 1) \log \sigma + \sum_{i \neq k}^{n} \delta_i \log f(r_i) + \sum_{i=1}^{n} (1-\delta_i) \log S(r_i) - \log \sigma + \log f(r_k)
\]

\[
= L(k)(\beta) - \log \sigma + \log f(r_k) \ .
\]

Assuming the mode of \( f \) is at 0, as it is for the Student-t distribution then

\[
\max_{\phi} L_{ok}(\beta, \phi, \sigma) = L(k)(\beta) - \log \sigma + \log f(0) \ .
\]  \hspace{1cm} (3.75)
Thus, if $L_{ok}(\beta, \phi, \sigma)$ is maximized with respect to $\beta$ and $\phi$ with $\sigma$ fixed, the maximization is equivalent to maximizing $L_{(k)}(\beta)$ with fixed $\sigma$. Note that this is not true when $\sigma$ is allowed to vary and the likelihood is maximized with respect to $\sigma$ as well as $\beta$.

For fixed $\sigma$, say at the maximum likelihood estimate $\hat{\sigma}$ for the full data set, the max $L_{ok}(\beta, \phi)$ in (3.73) is now given by (3.74) and (3.75) for censored and uncensored $k$, respectively with $\beta$ evaluated at $\hat{\beta}_{(k)}$, the maximum likelihood estimate of $\beta$ when the $k^{th}$ case is omitted from the data. Then $\hat{\beta}_{(k)}$ may be approximated by its 1-step estimate discussed in Sections 3.7.1 and 3.7.2 to give the corresponding 1-step $L_{ok}^{1}$.

3.8 EXAMPLES AND RESULTS.

The continuous scale mixture analysis discussed in this chapter was applied to a number of real data sets, but two are reported here since the analysis produces similar traits for different data sets. In general the estimates pertaining to a given model obtained by the EM algorithm conform well with that obtained by the Newton-Raphson method. However, the latter converges far more quickly. It takes approximately half the number of iterations required for the EM algorithm for the same convergence criterion.

The extreme points are characterized by their $E(Q)$'s in the EM algorithm close to zero or to use Pettitt's (1985) suggestion, $E(Q) - \text{Var}(cQ)$. This is equivalent to the $a^*_{k}$ in (3.67) for the 1-step Newton-Raphson, having negative values. Assessment of
influence is through the norms described in Sections 3.7.4 and 3.7.5. The quadratic norm (3.59), with matrix $M$ equal to the observed information matrix (Section 3.5.1) and $c = \hat{\sigma}^2$, however, gives measures that are too small to be of significance, although for cases that are suspect, the values are very much higher than the bulk of the data. It is rather difficult to interpret statistically, but it does alert us of possible suspect observations which may warrant further consideration and investigation.

Standard errors were computed using the observed information matrix results of Section 3.5.1, but $\sigma$ was assumed known in the analysis.

3.8.1 Motorette Data (Appendix B3).

These data were analysed in Section 2.8.4 using the finite discrete mixture model. Likewise we implemented the technique discussed in this chapter for this data set, fitting the regression $y = \beta_0 + \beta_1x$ (see Section 2.8.4 for explanation of variable). Table 3.1 shows the regressions estimates obtained by the EM algorithm for different degrees of freedom for the Student-t error distribution. The last row gives the estimates obtained by the technique discussed in Chapter 2 for comparison. The Newton-Raphson method gives identical results as that of the EM algorithm, but the number of iterations needed for convergence is considerably less, for all the models.

The Student-t distribution with infinite degrees of freedom is equivalent to Normal distribution. In our analysis the estimates obtained for Student-t with 60 degrees of freedom are essentially
<table>
<thead>
<tr>
<th>Estimates</th>
<th>Std. Dev</th>
<th>Loglike</th>
<th>Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>t-4</td>
<td>-5.6643 (0.613)</td>
<td>0.1534</td>
<td>-10.6656</td>
</tr>
<tr>
<td></td>
<td>4.1397 (0.283)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-6</td>
<td>-5.7125 (0.676)</td>
<td>0.1171</td>
<td>-11.2782</td>
</tr>
<tr>
<td></td>
<td>4.1627 (0.312)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-10</td>
<td>-5.7938 (0.758)</td>
<td>0.2047</td>
<td>-11.9136</td>
</tr>
<tr>
<td></td>
<td>4.2047 (0.349)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t-60</td>
<td>-5.9728 (0.900)</td>
<td>0.2494</td>
<td>-12.7916</td>
</tr>
<tr>
<td></td>
<td>4.2892 (0.412)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FM</td>
<td>-5.558 (0.251)</td>
<td>0.097</td>
<td>-8.4547</td>
</tr>
<tr>
<td></td>
<td>4.082 (0.114)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.1 Regression estimates for the motorette data for different degrees of freedom in the error structure (Student-t) in the continuous scale mixture model. Row FM gives estimates obtained by EM algorithm for the finite discrete scale mixture for comparison.

<table>
<thead>
<tr>
<th>Cases</th>
<th>E(Q)</th>
<th>E(Q)-V(eQ)</th>
<th>LDf</th>
<th>LDout</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>t4</td>
<td>0.4275</td>
<td>-0.1351</td>
<td>0.3341</td>
</tr>
<tr>
<td></td>
<td>t10</td>
<td>0.7491</td>
<td>0.2713</td>
<td>0.3567</td>
</tr>
<tr>
<td>21,22</td>
<td>t4</td>
<td>0.2231</td>
<td>-0.1435</td>
<td>0.5292</td>
</tr>
<tr>
<td></td>
<td>t10</td>
<td>0.5315</td>
<td>-0.0179</td>
<td>1.2178</td>
</tr>
</tbody>
</table>

Table 3.2 Motorette data. Summary of the quantities for the suspect cases under two models - Student-t with 4 and 10 degrees of freedom. LDf and LDout are the log likelihood distance for the fully iterated EM and Chi-square statistic for the outlier model estimated by EM, respectively.
equivalent to those obtained by maximum likelihood estimation with normal error (see last column of Table 2.4a). The model with lower degrees of freedom are more outlier prone, downweighting extreme cases to a larger extent, than models with higher degrees of freedom. Note that the parameters are underestimated in the finite discrete mixture model as compared to the continuous mixtures. In the former model (Section 2.8.4) the $E(Q)'s$ for the probable outliers, (cases 11, 21 and 22) are approximately zero. Hence the estimates given out in the analysis effectively are estimates with these cases omitted from the data. The $E(Q)'s$ for the continuous mixtures are about 0.4275 and 0.2231 for cases 11, 21 (22) respectively. Omitting these cases and employing a model with 10 degrees of freedom on the remaining 37 cases gives the estimates of $\hat{\beta}_0 = -5.598$ and $\hat{\beta}_1 = 4.107$ with standard errors 0.4561 and 0.2117 respectively and $\hat{c} = 0.1403$. These are similar results, except for the standard errors, as produced by the finite discrete mixture model. This is attributed to the modelling scheme.

In the finite model, the analysis is effectively on the "almost homogeneous" data, having "thrown" (downweighted to zero) the outlier, while in the continuous these suspect still carry substantial weights to have them included in the analysis.

The Student-t models find the above mentioned cases as suspect by their values of $E(Q)$, $E(Q) - \text{Var}(cQ)$ or $a_k^\ast$. Summary of these together with log likelihood distance for full step EM, and that of the outlier model (Section 3.7.6), is given in Table 3.2. Cases 21 and 22 under model with 10 degrees of freedom, by the ad hoc rule mentioned earlier are termed outliers (having negative
E(Q) - Var (eQ)) and case 11 is not. However, employing a model with 4 degrees of freedom which is more prone to outlier than a model with 10 degrees of freedom, we find that case 11 is picked out as an outlier. The influence of these cases on the estimates as indicated by the log likelihood distance decreases with the degrees of freedom of the model, as we expect (discussion on Section 3.7.3). The chi-squared outlier test for \( \gamma = 0 \) for the outlier model (Section 3.7.6) gives values of the statistic in the last column of Table 3.2 labelled "LDout". These are computed from the 1-step EM estimates. These are significant compared to \( \chi^2_1 \), rejecting the hypothesis that \( \gamma = 0 \) for these data points.

Figure 3.1 and Figure 3.2 show comparisons of estimates and log likelihood distances between the 1-step and the fully iterated estimations when the model is perturbed by single case deletion scheme for EM algorithm and the Newton-Raphson method. The plots are for the model with 10 degrees of freedom. The estimate \( \hat{\beta}_{(k)} \), the maximum likelihood estimate of \( \beta \) when the \( k \)th case is deleted, is well approximated by both the EM and Newton-Raphson methods, but at points where \( \hat{\beta}_{(k)} \) are far from \( \hat{\beta} \), that is at influential points, the Newton-Raphson methods seems to be less accurate, (see discussion in Section 3.7.3). This is evident in the log likelihood distance plot in Figure 3.2, where the \( LD^1 \), the log likelihood distance for 1-step estimation are larger than the \( LD^f \), the log likelihood distance for the fully iterated estimation. In contrast the \( LD^1 \) plot for the EM algorithm (Figure 3.1), is within the envelope of \( LD^f \). This confirms that in the EM algorithm the 1-step estimate of \( \beta_{(k)} \), starting from \( \hat{\beta} \) always increases the
Fig. 3.1  Plots to compare the 1-step with the fully iterated estimates by EM algorithm for the motorette data.
Fig. 3.2 Plots to compare the 1-step with the fully iterated estimates by N-R method for the motorette data

- Plots for coefficients var. 1
- Plots for coefficients var. 2
- Plots for likelihood distances
value of \( \mathcal{L}_k(\hat{\beta}) \), the likelihood function when the \( k^{th} \) case is deleted: that is

\[
\mathcal{L}_k(\hat{\beta}) \leq \mathcal{L}_k(\hat{\beta}_{(k)}) \leq \mathcal{L}_k(\hat{\beta}_{(k)})
\]

The quadratic norms (defined in Section 3.7.4) also indicates cases 11, 21 (and 22), with values 0.9266 and 0.8346, as having the "highest" influence on the estimates of \( \beta \). Other data points have values of less than 0.4. Again these values are too small to be significant.

Thus in this analysis, we have found outliers, but they appear to have little influence on the regression estimates.

### 3.8.2 Stanford Heart Transplant Data.

This data set (Appendix B6) has already been analysed in Section 2.8.5, under the finite discrete mixture model. It has been analysed quite extensively (see Section 4.6.2) in the context of censored data analysis. We applied the techniques discussed in this chapter to this data set for which the survival times are at least 10 days. There are 152 observations in all with about 33% censoring. The same transformations of the raw data employed in Section 2.8.5 are used, i.e. the log transformation. Table 3.3 gives the regression estimates for the five linear models, under Student-t with 6 degrees of freedom error distribution for the EM algorithm. The estimates found by Newton-Raphson with the same stopping criterion are identical, (to 3 decimal places) with the tabulated result.

The variable \( T5 \), the mismatch score, has little effect in
<table>
<thead>
<tr>
<th>case</th>
<th>age</th>
<th>obs.</th>
<th>LD1</th>
<th>Q*</th>
<th>LDout</th>
<th>E(Q)</th>
<th>E-V</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>29</td>
<td>12</td>
<td>0.041</td>
<td>0.036</td>
<td>5.295</td>
<td>0.554</td>
<td>-0.028</td>
</tr>
<tr>
<td>81</td>
<td>27</td>
<td>22</td>
<td>0.038</td>
<td>0.041</td>
<td>4.284</td>
<td>0.639</td>
<td>0.061</td>
</tr>
<tr>
<td>126</td>
<td>12</td>
<td>86</td>
<td>0.097</td>
<td>0.162</td>
<td>0.843</td>
<td>1.061</td>
<td>0.868</td>
</tr>
<tr>
<td>143</td>
<td>13</td>
<td>10</td>
<td>0.215</td>
<td>0.275</td>
<td>3.984</td>
<td>0.670</td>
<td>0.140</td>
</tr>
</tbody>
</table>

Q*: quadratic norm as computed from the 1-step N-R estimates
E-V: E(Q) - Var(eQ)

Table 3.4 Summary of suspect cases for the Stanford Heart Transplant data - model 4. These are cases with comparatively the highest values of LD1 or LDout in the data.
the regression in all models, but A2, the variable \((age)^2\) is significant. Our analysis henceforth in this section will be with respect to the model fitting the age and \((age)^2\), that is model 4 in Table 3.3.

Summary of cases of interest for the above model is as in Table 3.4. These are all uncensored observations. Cases 42 and 81 are suspect outliers with 2 lowest E(Q) with values of 0.554 and 0.639, but have little influence on the estimates. These cases correspond to patients ages 29 and 27 at the transplant but who died after 12 and 22 days respectively. One would expect at those ages, that they should survive longer. Case 143 has the highest LD\(^1\), for the EM algorithm. The log likelihood distance at fully iterated estimate without case 143 gives us the value of LD = 0.5685 which is still not significant (comparing to \(x^2\) with 3 degrees of freedom). This case corresponds to a very young individual, age 13, who lived only for 10 days after the transplant.

Figure 3.3 gives the scatter plot of the LD\(^1\) on the age at transplant. Although all the LD\(^1\) are not statistically significant four points below the age of 20 appear to be most influential. These points are summarized in Table 3.5.

<table>
<thead>
<tr>
<th>Case No.</th>
<th>Censoring</th>
<th>Life times in days</th>
<th>Age</th>
<th>LD(^1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>1</td>
<td>2006</td>
<td>15</td>
<td>0.1013</td>
</tr>
<tr>
<td>110</td>
<td>1</td>
<td>1116</td>
<td>14</td>
<td>0.0987</td>
</tr>
<tr>
<td>126</td>
<td>0</td>
<td>86</td>
<td>12</td>
<td>0.0973</td>
</tr>
<tr>
<td>143</td>
<td>0</td>
<td>10</td>
<td>13</td>
<td>0.2147</td>
</tr>
</tbody>
</table>

**TABLE 3.5**: Case with highest 1-step EM, LD\(^1\), for the Heart Transplant Data - error \(t_6\).
Fig. 3.3 Plot of likelihood distance obtained from the 1-step EM algorithm on age for the Heart transplant data - model 4

![Plot of likelihood distance obtained from the 1-step EM algorithm on age for the Heart transplant data - model 4](image)

Fig. 3.4 Plot of 1-step, LD^1, likelihood distances for EM and N-R methods applied to Heart transplant data - error t_e

![Plot of 1-step, LD^1, likelihood distances for EM and N-R methods applied to Heart transplant data - error t_e](image)
Points corresponding to young patients who died early after transplant or who lived long (over 1116 days), but censored seem to exert undue influence than the rest of the data.

Figure 3.4 gives comparison of the 1-step likelihood distances for the EM and the Newton-Raphson methods. Generally these measures agree well, apart from the unduly "influential" cases for which the Newton-Raphson 1-step LD\( ^1 \) appear to overestimate. Nevertheless, they are able to draw our attention to suspect cases for further consideration.

To demonstrate further that these techniques are successful in the detection of influential and outlying cases, we employ the same model to a subset of the Stanford Heart Transplant data, namely the data set corresponding to all patients with ages over 50 years. In addition we deliberately include case 143, corresponding to patient age 13 who died after 10 days of transplant, as a "rogue" observation in the data set. There are 38 data points in all with six censored observations with the lifetimes ranging from 10 days to 2474 days. Summary of the estimates obtained by the EM algorithm for this data set with and without the inclusion of the observation at age 13 is tabled in Table 3.6.

<table>
<thead>
<tr>
<th></th>
<th>With observation at age 13</th>
<th>Without observation at age 13</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \beta_0 )</td>
<td>-0.7454 (1.2683)</td>
<td>43.2088 (27.9194)</td>
</tr>
<tr>
<td>( \beta_1 )</td>
<td>0.1636 (0.0698)</td>
<td>-1.4030 (0.9961)</td>
</tr>
<tr>
<td>( \beta_2 )</td>
<td>-0.0020 (0.0009)</td>
<td>0.0118 (0.0088)</td>
</tr>
</tbody>
</table>

| Std.           | 0.7113                    | 0.6987                       |

**TABLE 3.6**: Estimates as obtained by the EM algorithm for the continuous mixture model applied to Heart Transplant data for patients over 50 years old, and a "rogue" point.
Table 3.7 Summary of the quantities output by 1-step EM algorithm for the continuous mixtures model with error t-6 applied on a subset of the Heart Transplant data corresponding to patients 50 and over years of age at transplant - regression model 4. Observation no. 10 is a 'rogue' case, patient at age 13, deliberately included in the data set.
The analyses show that $\beta_0$ is not significant in the model though for the data set without the "rogue" point shows slight evidence of significance of $\beta_0$. However, there are drastic changes in the estimates in terms of magnitude and sign. The full estimate likelihood distance is 35.913 which is highly significant (compared to $x_3^2$).

The summary of the quantities for the 1-step EM is given in Table 3.7. The suspect case is observed to be very influential from the LD but is not considered as an outlying case by its E(Q) or E(Q) - Var (eQ). This is to be expected as the curve in the interval 13 and 50 is solely fitted by the point at age 13.

3.9 CONCLUSION.

The examples and results in the previous section illustrates the usefulness of the techniques thus far described, in determining model parameters and detecting outliers and assessing influential observation. The EM algorithm, although less efficient than the Newton-Raphson method, generally gives better estimates of the 1-step influence curve, than the latter especially at points of high influence for which the Newton-Raphson appear to overestimate. (Compare last plots of Figure 3.1 and Figure 3.2).

The quadratic norm as defined in Section 3.7.4 is difficult to interpret, however, it is consistent with the finding concluded from the likelihood distance. Suspect cases have values of the quadratic norms and likelihood distances exceedingly high from the bulk of the data. This draws our attention for further consideration of these cases. For non-influential points the
values of the 1-step estimate of the likelihood distance, $LD^1$, are very small; suggesting that it is not distributed as $\chi^2$, with $p$ degrees of freedom. On the other hand the "likelihood distance", $LD_{out}$ distributed $\chi^2_1$, for the outlier model is a better statistic for the detection of the outlier. $LD^1$, thus needs to be "quantified" in some other way.
Chapter 4

Proportional Hazards Model

4.1 INTRODUCTION.

Let $T_1, T_2, \ldots, T_n$ be independent failure times with $T_i$ having the distribution function $F_i(t)$ and density function $f_i(t)$. The survivor function, $S_i(t)$, and the hazard rate, $h_i(t)$, are defined respectively as

$$S_i(t) = 1 - F_i(t), \quad (4.01)$$

$$h_i(t) = \frac{f_i(t)}{S_i(t)}. \quad (4.02)$$

For continuous distributions, by the above definitions,

$$h_i(t) = -\frac{d[\log(S_i(t))]}{dt}$$

so that

$$S_i(t) = \exp \left[ - \int_0^t h_i(u) \, du \right] \quad (4.03)$$

since $S_i(0) = 1$. Let $H_i(t) = \int_0^t h_i(u) \, du$ the so-called integrated hazard function, then from (4.02)

$$f_i(t) = h_i(t) \exp \left[ - H_i(t) \right]. \quad (4.04)$$

These relationships are also true for discrete distributions. (Cox and Oakes, 1984, page 14). The hazard rate specifies the instantaneous rate of failure at $T_i = t$ conditional upon survival to time $t$.

A broad family of models that is often used in the analysis of survival data is the Lehmann alternative family or the proportional hazards family. The latter defines the hazard rate as
\[ h_i(t) = h_0(t) \psi \]  
(4.05)

for some constant \( \psi \) and arbitrary unknown \( h_0(t) \). In terms of the survivor function, (4.05) is equivalent to

\[ S_i(t) = \left[ S(t) \right]^\psi \]  
(4.06)

where \( S(t) = \exp \left( - \int_0^t h_0(u) du \right) \)

which was considered by Lehmann (1953) and Savage (1956).

Cox (1972) used this model in a situation where associated with each \( T_i \) is a \( p \times 1 \) covariate vector \( z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^T \). The components of \( z \) may represent various features thought to affect the failure time \( T \). This could be the intrinsic properties of the individual like age, sex, etc., or treatment, or could be synthesized from existing variables to examine interaction effects. In the Cox model \( z \) can be considered to be independent of time, although in general \( z \) can be time-dependent. The hazard for the Cox model for an individual at failure time \( T_i = t \) is

\[ h_i(t, z) = h_0(t) \exp (z_i^T \beta) \]  
(4.07)

where \( h_0(t) \) is some arbitrary baseline hazard function, usually unknown, corresponding to case \( z = 0 \), and \( \beta = (\beta_1, \beta_2, \ldots, \beta_p)^T \) is a \( p \times 1 \) vector of regression coefficients. As we shall see in Section 4.2, \( h_0(t) \) may not be specified in the estimation of \( \beta \) which can be achieved using maximum likelihood.

In Sections 4.4 and 4.5 we will discuss the different approaches of estimating the parameter \( \beta \) of equation (4.07), and in Section 4.6 compare the results obtained from these approaches as applied to different data sets. Influences and validity of proportional hazard assumptions are left to later chapters.
4.2 **Likelihood Function.**

The regression parameter $\beta$, in the model defined by equation (4.07) can be estimated by maximum likelihood. It is thus of prime importance to establish the likelihood function for the model. There are a number of approaches to the problem. Cox (1972) introduced the partial likelihood approach, while Kalbfleisch and Prentice (1973) used the rank marginal likelihood, and Breslow (1974) used a maximum likelihood technique whereby the hazard function is approximated by a step function with discontinuities at each observed failure time. They all arrived at an identical likelihood function. Here we shall discuss in brief the first two approaches.

4.2.1 **Partial Likelihood.**

Cox's partial likelihood approach is based on the conditional probability that an individual fails, given the other individuals still alive or at risk at time $t_i$. This probability is:

$$P(l; \delta_1, \ldots, \delta_k) = \frac{h_l(t_i)}{\sum_{h \in R_i} h_h(t_i)}$$

where $R_i$ is the "risk set" for failure time $t_i$, at which the $i^{th}$ individual fails, that is $R_i = \{j ; t_j > t_i\}$, and $h_h(t_i)$ is the hazard of the $h^{th}$ observation still at risk after failure time $t_i$. The partial likelihood is formed by taking the product over all failure times $t_i$. With the observed explanatory variable $z$, that the failure times may depend upon, from (4.07) the partial likelihood for $k$ distinct failure times is thus
\[
\ell(\beta) = \prod_{i=1}^{k} \frac{\exp(z_i^T \beta)}{\sum_{j \in R_i} \exp(z_j^T \beta)}.
\] (4.08)

Here we see that prior knowledge of \( h(t_i) \) is not required.

Introducing a censoring indicator \( \delta_i \), where \( \delta_i = 1 \) if \( t_i \) is a true failure and \( \delta_i = 0 \) if \( t_i \) is censored, (4.08) can be written, after taking into account the approximation of Peto (1972) and Breslow (1974) for tied observations, (see Section 4.2.3) as

\[
\ell(\beta) = \prod_{i=1}^{n} \frac{\exp(z_i^T \beta \delta_i)}{\sum_{j \in R_i} \exp(z_j^T \beta \delta_j)}.
\] (4.09)

where \( n \) is the total number of observations, whether censored or otherwise.

### 4.2.2 Rank Marginal Likelihood

This approach is due to Kalbfleisch and Prentice (1973). We assume firstly that the observations are distinct and that no censoring is present. Let the failure time \( T_i \), with realization \( t_i \), \( i = 1, 2, \ldots, k \), be such that \( T(1) < T(2) < \ldots < T(k) \). Note that the subscript (i) refers to the label attached to the i\(^{th}\) order statistics. Let \( z(1), z(2), \ldots, z(k) \) be the corresponding p\times1 vectors of explanatory variables. Following Kalbfleisch and Prentice (1980), the marginal likelihood based on the ranks is proportional to:

\[
\Pr\left[T(1) < T(2) < \ldots < T(k)\right]
\]

which is sometimes called the rank likelihood. Thus
\[ \ell(\beta) = \int_R \left[ \prod_{i=1}^{k} f(t(i), z(i)) \right] dt \]  \tag{4.10} 

where \( dt = dt(1) dt(2) \ldots dt(k) \), and the integration is the \( k \)-tuple integration over the rank order information, \( R \), concerning the failure times, that is over \( t(1) < t(2) < \ldots < t(k) \).

The proportional hazards model of (4.07) is

\[ h(t, z) = h_o(t) \exp(z^T \beta). \]

In general, the hazard above could be generated (Kalbfleisch and Prentice, 1973) as the hazard of

\[ u = g(t) = \int_0^t h_o(s) \, ds \]  \tag{4.11} 

where \( t \) is from the exponential distribution with mean \( \exp(-x^T \beta) \) and \( g \) is a monotonic increasing transformation of \( t \). From (4.04) and (4.07) and putting \( \lambda_i = \exp(z_i^T \beta) \), equation (4.10) can now be written as:

\[ \ell(\beta) = \int_R \prod_{i=1}^{k} h_o(t(i)) \lambda(i) \exp\left[ -\lambda(i) \int_0^{t(i)} h_o(s) \, ds \right] dt \]

with the definition of \( dt \) and \( R \) as above. The transformation of (4.11) gives us

\[ dt(i) = \left( \frac{d[g^{-1}(u(i))]}{du(i)} \right) du(i) \]

and

\[ h_o(t(i)) = \frac{du(i)}{dt(i)}. \]

Defining \( du = du(1) \cdot du(2) \ldots du(k) \), the rank likelihood is
\[ \ell(\beta) = \prod_{i=1}^{k} \lambda_{(i)} \exp\left(-\lambda_{(i)} u_{(i)}\right) \frac{d[g^{-1}(u_{(i)})]}{du_{(i)}} \]

\[ = \prod_{i=1}^{k} \lambda_{(i)} \exp(-\lambda_{(i)} u_{(i)}) \, du_{(i)} \quad (4.12) \]

since \( \frac{d[g^{-1}(u_{(i)})]}{dg} = \frac{dt_{(i)}}{du_{(i)}} \).

R here is implicitly defined as \( 0 < u_{(1)} < \ldots < u_{(k)} < \infty \).

We see that (4.12) is identical to the probability of the given rank order when the variables are independent and from an exponential distribution with mean \( 1/\lambda \). By repeated integration or other means (4.12) is

\[ \ell(\beta) = \prod_{i=1}^{k} \lambda_{(i)} \frac{1}{k!} \sum_{j=1}^{k} \lambda_{(j)} \quad (4.13) \]

where (i) indicates the case that fails at \( t_{(i)} \) when the observations are ranked. By defining the set \( R_i = \{ j : t_j > t_i \} \) as before, (4.13) is identical to (4.08).

When there is censoring, the rank statistic is no longer unique since the actual unrealized value of the censored observation can be of any value greater than the censoring time. The marginal likelihood is then computed as the probability of a sample being consistent with the rank ordering. As before, suppose \( t(1) < t(2) \ldots \ldots < t(k) \) are the observed failure times with corresponding covariates \( z(1), z(2), \ldots, z(k) \). Let the interval \( (t(i), t(i)+1) \) contain \( m_i \) censored cases at times \( t_{i1}, t_{i2}, \ldots, t_{im_i} \) with corresponding covariates \( z_{i1}, z_{i2}, \ldots, z_{im_i} \). Kalbfleisch and Prentice (1980)
characterized a set of possible rank vectors to compute the likelihood by the event
\[ t(1) < t(2) < \cdots < t(k) \]
\[ t(i) < t_{i1}, t_{i2}, \cdots, t_{im_i} \quad (i = 1, 2, \ldots, k). \]

(4.14)

Given \( T(i) = t(i) \), the conditional probability of the event \( T(i) < T_{i1}, T_{i2}, \ldots, T_{im_i} \) is
\[ p(t(i)) = \prod_{j=1}^{m_i} \left[ 1 - F_j(t(i)) \right] \]

(4.15)

where \( F_j(t_i) \) is the probability that \( T(i) < T_{ij} \) given \( T(i) = t(i) \).

Following (4.01), (4.03) and (4.07) \( F_j(t_i) \) can be written as:
\[ F_j(t(i)) = 1 - \exp \left[ - \int_0^{t(i)} h_o(u) e^{-\frac{T_{ij}}{\beta}} du \right]. \]

Hence
\[ p(t(i)) = \exp \left\{ - \sum_{j=1}^{m_i} \exp \left( \frac{T_{ij}}{\beta} \right) \int_0^{t(i)} h_o(u) du \right\}. \]

The marginal likelihood is then proportional to the probability of the event (4.14), which is
\[ \mathcal{L}(\beta) = \int_{t(1)=0}^{\infty} \int_{t(2)=t(3)}^{\infty} \cdots \int_{t(k)=t(k-1)}^{\infty} \prod_{i=1}^{k} f(t(i), z(i)) p(t(i)) dt \]

where \( dt = dt(1), dt(2), \ldots, dt(k) \).

With the assumption that \( T_i \) is from the unit exponential distribution and putting \( \lambda_{ij} = \exp(z_{i,j}^T \beta) \), which refers to the \( \lambda \)
of the censored case \( j \) in the interval \([t(i), t(i)+1)\), the likelihood is

\[
L_R = \int R \frac{\prod_{i=1}^k \lambda_i \exp \left\{ - \left[ \sum_{j=1}^{m_i} \frac{\lambda_{ij} + \lambda_i}{\lambda_i} \right] t_i \right\}}{\prod_{j=1}^{m_i} \lambda_i} dt \tag{4.16}
\]

where \( R \) is defined as in (4.10). By repeated integration or by the expectation method (Section 4.3), the likelihood obtained is identical to that of equation (4.08) where the risk set \( R_i \) will include the censored observations whose failure times are greater or equal to \( t_i \).

4.2.3 Adjustment For Tied Observations.

When ties are present the contribution to the likelihood \( L_i \) for case failing at \( t_i \) (see Storer and Crowley, 1985) with multiplicity \( m_i \) is

\[
L_i = \prod_{j \in S_i} \frac{\lambda_j}{c \in R_i} \frac{\sum_{c \in R_i} \lambda_{ij}}{c \in c \cdot \lambda_i} \tag{4.17}
\]

where \( R_i \) is the risk set with \( n_i \) individual at risk and \( S_i \) is the set of labels with \( m_i \) individual failing at \( t_i \). The summation over \( c \) is over all \( \binom{n_i}{m_i} \) combinations of \( m_i \) labels \( (c_1, c_2, \ldots, c_{m_i}) \) that can be chosen from \( n_i \) labels in \( R_i \). For detailed treatment on the subject refer to Kalbfleisch and Prentice (1980), Cox and Oakes (1984), and Cox (1972, 1975).

The likelihood is then the product over all observed failure points \( t_i, i = 1, 2, \ldots, n \), to give

\[
\text{lik} = \prod_{i=1}^n L_i.
\]
Unfortunately, this likelihood is difficult to compute especially if \( n_i \) and \( m_i \) are large. One approximation which is widely used and quite satisfactory is due to Peto (1972) and Breslow (1974). The approximation, with \( \lambda_i = z_i^T \beta \), is

\[
L_i = \prod_{j \in S_i} \exp(z_j^T \beta) / \left( \sum_{j \in R_i} (\exp(z_j^T \beta))^{m_i} \right).
\]  

(4.18)

Introducing the censoring indicator, \( \delta_i = 1 \) for observed failures and \( \delta_i = 0 \) for the censored observations, and extending the observations \( t_1, t_2, \ldots, t_n \) to include censored observations and ties, (4.18) can now be written as

\[
L_i = \frac{\exp(z_i^T \beta \delta_i)}{\left( \sum_{j \in R_i} \exp(z_j^T \beta) \delta_i \right)^{m_i}}, \quad \forall \ i = 1, 2, \ldots, n
\]

which gives an overall likelihood function as in (4.09). We shall use this approximation in our analysis throughout this thesis, and we shall refer (4.09) as partial likelihood in view of the approach of Cox (1972).

4.3 Rank Likelihood as Expectation from Exponential Distribution.

As a consequence of the exponential order statistics in the rank likelihood approach (Section 4.2.2) equation (4.09) may be derived from an expectation argument. We shall present an outline of the argument here and later (Section 4.5) extend it to the EM implementation of the proportional hazards model. Before we endeavour to do so we shall state a lemma from Cox and Hinkley (1974), page 467.
Lemma 4.3.1.

If $X_1 < X_2 < \ldots < X_n$ are $n$ order statistics from a random sample of size $n$ from an exponential distribution with unit mean, then

$$Z_k = (n - k + 1)(X_k - X_{k-1}), \quad k = 1, 2, \ldots, n$$

are independently identically distributed as exponential random variables with unit mean.

Since $X_i$ are ordered, the joint distribution of $X_1 < X_2 < \ldots < X_n$ is

$$f_{X_1 < X_2 < \ldots < X_n}(x_1, x_2, \ldots, x_n) = n! \prod_{k=1}^{n} \exp(-x_k)$$

$$= n! \exp\left[-\sum_{k=1}^{n} x_k\right]$$

$$= n! \exp\left[-\sum_{k=1}^{n} (n-k+1)(x_k - x_{k-1})\right]$$

where $x_0 = 0$.

Define $z_k = (n-k+1)(x_k - x_{k-1})$. Note that the determinant of the Jacobian of the transformation from $x$ to $z$ is 1, and thus we see that $z_k$ is exponential with mean 1. It follows:

$$x_k = \frac{z_k}{n-k+1} + x_{k-1}$$

$$= \frac{z_k}{n-k+1} + \frac{z_{k-1}}{n-k} + \ldots + \frac{z_1}{n}$$

$$= \sum_{j=1}^{n} \frac{c_j z_j}{n!}$$
The expectation approach begins with the rank likelihood of equation (4.12). Rewriting the equation, and substituting the dummy variable $u_i$ by $t_i$ we have

$$
\mathcal{L}(\theta) = \int_{\mathbb{R}} \left[ \prod_{i=1}^{k} \lambda_i \Pi e^{-\lambda(i) t(i)} \right] \left[ \prod_{i=1}^{k} e^{-t_i} \right] dt
$$

$$
= \int_{\mathbb{R}} \left[ \prod_{i=1}^{k} \lambda_i \Pi e^{-\lambda(i) (\lambda(i)^{-1})} \right] \left[ \prod_{i=1}^{k} e^{-t_i} \right] dt
$$

$$
= \frac{1}{k!} \mathbb{E} \left[ \prod_{i=1}^{k} \lambda_i \exp \left\{ -T(i) (\lambda(i)^{-1}) \right\} \right]. \quad (4.19)
$$

The expectation is with respect to the joint distribution of $0 < T(1) < T(2) < \ldots < T(k)$, the order statistics from a random sample of size $k$, from an exponential distribution with mean 1, that is the joint distribution with density function

$$
f(t_1, t_2, \ldots, t_k) = k! \prod_{i=1}^{k} e^{-t(i)} \quad 0 < t(1) < t(2) < \ldots < t(k).$$

From Lemma (4.3.1), let

$$T_i = \sum_{j=1}^{i} c_j Z_j \quad \text{and} \quad c_j = \frac{1}{k-j+1}, \quad Z_j \sim \text{Ex}(1),$$

then equation (4.19) can now be written:

$$
\mathcal{L}(\theta) = \frac{1}{k!} \mathbb{E} \left[ \prod_{i=1}^{k} \lambda_i \exp \left\{ -\sum_{j=1}^{i} c_j Z_j (\lambda(i)^{-1}) \right\} \right]
$$

$$
= \frac{1}{k!} \mathbb{E} \left[ \prod_{i=1}^{k} \lambda_i \prod_{i=1}^{k} \exp \left\{ -\sum_{j=1}^{i} c_j (\lambda(i)^{-1}) Z_j \right\} \right].
$$
\[
\begin{align*}
\prod_{i=1}^{k} \lambda_i & \quad \frac{1}{k!} \sum_{i=1}^{k} \sum_{j=1}^{k} c_j (\lambda_i^{-1}) Z_{ij} \\
& = \frac{1}{k!} \sum_{i=1}^{k} \sum_{j=1}^{k} c_j (\lambda_i^{-1}) Z_{ij}
\end{align*}
\]

where

\[
b_j = c_j \sum_{i=j}^{k} (\lambda_i^{-1}).
\]

Now the \( Z_j \) are independent, hence

\[
E \left[ \prod_{j=1}^{k} e^{-b_j Z_j} \right] = \prod_{j=1}^{k} E \left[ e^{-b_j Z_j} \right].
\]

Further since \( Z_j \sim \text{Ex}(1) \), then

\[
E \left[ e^{-b_j Z_j} \right] = \frac{1}{1 + b_j}, \quad 1 + b_j > 0
\]

and equation (4.21) becomes

\[
E \left[ \prod_{j=1}^{k} e^{-b_j Z_j} \right] = \frac{1}{k!} \sum_{i=1}^{k} \left[ 1 + c_j \sum_{i=j}^{k} (\lambda_i^{-1}) \right] \prod_{j=1}^{k} (1 + b_j) \]

\[
= \frac{1}{\left\{ \prod_{j=1}^{k} \left[ 1 + c_j \sum_{i=j}^{k} (\lambda_i^{-1}) \right] \right\}} \frac{1}{k!}, \quad c_j = 1/(k-j+1)
\]

\[
\left\{ \prod_{j=1}^{k} \frac{\lambda_i}{k-j+1} \right\}
\]
Hence equation (4.20) becomes:

$$
\prod_{i=1}^{k} \frac{1}{k!} \prod_{j=1}^{k} \lambda(i)
$$

$$
= \frac{1}{k!} \prod_{i=1}^{k} \sum_{j=i}^{k} \lambda(i)
$$

(4.23)

Again we note that the index (j) indicates the case that fails at \( t(j) \) when the observations are ranked \( 0 < t(1) < \ldots < t(k) \). Thus (4.23) can be written as

$$
\prod_{i=1}^{k} \frac{1}{\prod_{j=1}^{i} \lambda(j)}
$$

(4.24)

where \( R_i = \{ j : t_j \geq t_i \} \) is defined as before. We have arrived at the likelihood without considering the censored observations. To include them in the likelihood we may use the same line of argument as in Section (4.2.2) and use equation (4.16) instead of (4.12). The likelihood obtained is identical to (4.24) but \( R_i \) now will include the censored observations whose failure times are greater or equal to \( t_i \). Equation (4.09) can be arrived at after introducing the indicator variable \( \delta_i \) and considering the approximation of Peto (1972) and Breslow (1974) for tied observations.

4.4 ESTIMATION AND INFERENCE.

As mentioned earlier, the regression parameters \( \beta \), are estimated by the maximum likelihood method. There are several approaches, of which the standard approach is the Newton-Raphson iterations on the
score function. Other approaches include the linear model approach for survival times and then using GLIM (Baker and Nelder, 1978) to estimate the $\beta$. In Section (4.5) we shall discuss in detail the EM algorithm (Dempster et al., 1977) approach to the estimation.

4.4.1 Newton-Raphson Iteration.

The standard method of finding an estimate of $\beta$ is to maximize the partial likelihood (4.09). Let $\hat{\beta}$ be the solution of $U(\beta) = 0$, where $U(\beta)$ is a p×1 score vector of derivatives of the logarithm of the partial likelihood, $\ell(\beta)$. From (4.09)

$$L(\beta) = \log \left[ \ell(\beta) \right] = \sum_{i=1}^{n} \delta_{i} \left\{ z_{i}^{T} \beta - \log \left[ \sum_{j \in R_{i}} \exp(z_{j}^{T} \beta) \right] \right\}$$

The score vector is then

$$U(\beta) = \frac{\partial L(\beta)}{\partial \beta}$$

$$= \sum_{i=1}^{n} \delta_{i} \left\{ z_{i} - \left[ \sum_{j \in R_{i}} z_{j} \exp(z_{j}^{T} \beta) / \sum_{j \in R_{i}} \exp(z_{j}^{T} \beta) \right] \right\}.$$  \hfill (4.25)

Defining $\lambda_{i} = \exp(z_{i}^{T} \beta)$ as the relative risk of failure of the $i$th individual, and a p×1 vector $a_{i}$, such that

$$a_{i} = \frac{\sum_{j \in R_{i}} z_{j} \lambda_{j}}{\sum_{j \in R_{i}} \lambda_{j}}$$

then $U(\beta)$ is simply

$$\sum_{i=1}^{n} \delta_{i} \left( z_{i} - a_{i} \right).$$  \hfill (4.26)
Note that $a_i$ is the "average" value of $z_j$ when $j$ has a discrete
distribution over $R_i$ with probability $\lambda_j / \sum_{k \in R_i} \lambda_k$, $j \in R_i$.

The $p \times p$ observed information matrix $A(\beta)$ is thus:

$$A(\beta) = - \frac{\partial^2 L(\beta)}{\partial \beta \partial \beta^T}$$

$$= \sum_{i=1}^{n} \delta_i \left\{ \left( \sum_{j \in R_i} \frac{z_j z_j^T \lambda_j}{\sum_{k \in R_i} \lambda_k} \right) - \frac{a_i a_i^T}{\lambda_i} \right\} \quad (4.27)$$

$\hat{\beta}$ can typically be obtained by Newton-Raphson iteration utilizing
(4.25) and (4.27). The scheme is given by

$$\frac{\beta_{s+1}}{\beta_s} = \frac{\beta_s}{\beta_s} + A^{-1}(\beta_s) U(\beta_s) \quad (4.28)$$

where $U(\beta_s)$ and $A(\beta_s)$ are the score vector and the information
matrix, respectively evaluated at $\beta_s$. Appendix C3 gives the
computer listing of this algorithm.

Under certain regularity conditions to permit differentiation
under the integral sign (Cox and Hinkley, 1974) the score vector
$U(\beta)$ has zero expectation and covariance matrix $E[A(\beta)]$. However,
computation of the expectation would require a fuller specification
of the censoring mechanism than is normally available, and in view
of asymptotic theory, the expectation is replaced by the information
matrix instead. The inference for $\hat{\beta}$ is based on the asymptotic
normality of the distribution of $\hat{\beta}$, which has mean of $\beta$ and estimated
covariance matrix $A^{-1}(\hat{\beta})$. These results hold in the absence of ties.
If ties are present then one would expect some asymptotic bias in
both the estimation of $\hat{\beta}$ and its covariance matrix due to the
approximation used in (4.18).
4.4.2 Proportional Hazards as a Linear Model.

A linear model for a survival time $T$ is

$$\log\left\{ g(T) \right\} = -z^T \beta + \epsilon$$

(4.29)

where $g(\cdot)$ is an increasing function and a transformation of the failure times, $\epsilon$ is a random error term distributed as a standard extreme value variate, that is, $\exp(\epsilon)$ is distributed exponentially with unit mean. This implies that $g(t)$ is distributed exponentially with mean $1/\lambda$, where $\lambda = \exp(z^T \beta)$. To obtain the proportional hazard model, the integrated function (Section 4.1) is related to $g(\cdot)$, (see Clayton and Cuzick, 1985a) by

$$g(t) = H_0(t)$$

(4.30)

where

$$H_0(t) = \int_0^t h_0(u) \, du.$$  

$H_0(t)$ may be parameterised as $H_0(t, \alpha)$. Aitkin and Clayton (1980) use $H_0(t, \alpha) = t, t^\alpha, \exp(at)$ for the exponential, Weibull and extreme value fits, respectively. GLIM can be used to estimate $\beta$ by using a Poisson model with the $y$ variate, the censoring indicator $\delta_i$ and offset $\log[H_0(t, \alpha)]$. This is motivated by the general likelihood function for censored data

$$\text{lik} = \prod_{i=1}^{n} \left[ f(t_i) \right]^{\delta_i} \left[ S(t_i) \right]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} \left[ h(t_i) \right]^{\delta_i} S(t_i)$$

and for the proportional hazard
\[
\text{lik} = \prod_{i=1}^{n} \left[ \frac{h_0(t_i)}{o(t_i)} \lambda_i \right]^{\delta_i} \exp \left[ - \lambda_i \frac{h_0(t_i)}{H_0(t_i)} \right]
\]

\[
= \prod_{i=1}^{n} \left[ \frac{\delta_i - \mu_i}{\mu_i e} \right] \left[ \frac{h_0(t_i)}{H_0(t_i)} \right]^{\delta_i}
\]  
(4.31)

where \( \mu_i = H_0(t_i) \lambda_i \).

The first term in the likelihood function is the kernel of the likelihood function of \( n \) independent Poisson variates, \( \delta_i \) with mean \( \mu_i \). The second term does not involve the parameter \( \beta \). Also we see that

\[
\log \mu_i = \log H_0(t_i) + z_i^T \beta
\]

provides the generalised linear model with offset \( \log[H_0(t_i)] \).

A very general approach for fitting parametric models for \( h_0(t_i) \) is described by Clayton (1983). Here again he used GLIM with Poisson random components. The proposed model has

\[
g[h(t)] = \mu + \beta_1 t + \beta_2 t^2 + \beta_3 t^3 + \ldots
\]

with well known special cases, e.g. \( h(t) = \mu \), then the failure time is distributed exponentially.

4.5 EM ALGORITHM APPROACH TO ESTIMATION.

4.5.1 Introduction.

In this section we shall view the EM algorithm for estimation of the regression parameters and use GLIM to perform the iterative least square (see for example, MacCullagh and Nelder, 1983), in the maximisation part of the algorithm. Two GLIM models are considered, one having poisson random component (due to Clayton
and Cuzick, 1985a), and the other gamma random component. In subsection 4.5.6 we shall discuss the differences between the two.

Survival data can be viewed as incomplete data, where the failure time $T_i$ of the censored is not observed, and is only known to exceed the censoring time. In the proportional hazard model, we shall use the rank informations $R_i$ to represent these observations or their transformations. This then is the incomplete data set-up as defined by Dempster et al. (1977) for their EM algorithm.

Let $l_o(\beta)$ and $l_m(\beta)$ be the likelihoods for the complete data specification and the proportional hazard model, respectively. From (4.10) or Dempster et al. (1977) equation (1.1), the likelihood for the model is

$$l_m(\beta) = \int_R l_o(\beta) \, du$$

where $du = du_1, du_2, ..., du_n$, and $u \in R$ represent all values of $u$ consistent with the rank information concerning failure times. Then

$$\frac{\partial}{\partial \beta} \left[ l_m(\beta) \right] = \int_R \frac{\partial}{\partial \beta} \left[ l_o(\beta) \right] \, du$$

$$= \int_R \left\{ \frac{\partial}{\partial \beta} \left[ \log l_o(\beta) \right] \right\} l_o(\beta) \, du$$

Hence

$$\frac{\partial}{\partial \beta} \left\{ \log l_m(\beta) \right\} = \frac{\int_R \left\{ \frac{\partial}{\partial \beta} \left[ \log l_o(\beta) \right] \right\} l_o(\beta) \, du}{\int_R l_o(\beta) \, du}$$

(4.33)
since \[ \frac{\partial}{\partial \beta} \log \left[ \mathcal{L}_m (\beta) \right] = \frac{\partial}{\partial \beta} \left[ \mathcal{L}_m (\beta) \right] / \mathcal{L}_m (\beta). \]

The expectation with respect to rank order information is

\[ E(\cdot) = \frac{\int_{R} (\cdot) \mathcal{L}_o (\beta) \, du}{\int_{R} \mathcal{L}_o (\beta) \, du}, \tag{4.34} \]

which is proper expectation. Let \( \mathcal{L}_o (\beta) \) and \( \mathcal{L}_m (\beta) \) be the logarithms of the complete data specification likelihood and proportional hazard likelihood respectively, then (4.33) becomes:

\[ \frac{\partial}{\partial \beta} \left[ \mathcal{L}_m (\beta) \right] = E \left\{ \frac{\partial}{\partial \beta} \left[ \mathcal{L}_o (\beta) \right] / R, \beta \right\} \tag{4.35} \]

Recapitulating the EM algorithm (Chapter 1) there are two fundamental steps to proceed:

1) the E-step : compute

\[ Q(\beta'/\beta_p) = E \left[ \mathcal{L}_o (\beta'; t) / R, \beta_p \right], \]

2) the M-step : Determine \( \beta_{p+1} \) as a new value of \( \beta' \) which maximizes \( Q(\beta'/\beta_p) \),

where \( \mathcal{L}_o (\beta'; t) \) is the log likelihood if all the \( t_i \) were observed, that is, the loglikelihood for the complete data set-up. Step (2) is equivalent to finding the solution to

\[ E \left[ \frac{\partial}{\partial \beta'} \left[ \mathcal{L}_o (\beta'; t) \right] / R, \beta \right] = 0 \tag{4.36} \]
and from (4.35), the maximum likelihood estimate of $\beta$ which is
the solution to $\frac{\partial}{\partial \beta} \left[ L_m(\beta) \right] = 0$, can be solved iteratively by
solving equation (4.36). This is the EM algorithm.

4.5.2 Variance-Covariance Matrix from Rank Order Likelihood.

As was discussed in section 4.2.2, the inverse of the information
matrix is the asymptotic variance-covariance matrix of the estimates,
$\hat{\beta}$. This may be obtained from the second derivative of the logarithm
of the marginal likelihood. From (4.33), for brevity let
$L_m = \log \left[ \ell_m(\beta) \right]$, $\ell_o = \ell_o(\beta)$ and $\ell_m = \int R \ell_o(\beta) \, du$, then

$$\frac{\partial L_m}{\partial \beta \beta^T} = \frac{\partial}{\partial \beta} \left\{ \int R \frac{\partial}{\partial \beta^T} \log(\ell_o) \right\} \ell_o \, du$$

$$= \frac{1}{\ell_m} \left\{ \int R \frac{\partial^2}{\partial \beta \beta^T} \log(\ell_o) \right\} \ell_o \, du + \int R \left[ \frac{\partial}{\partial \beta} \log(\ell_o) \right] \frac{\partial \ell_o}{\partial \beta} \, du$$

$$+ \frac{1}{\ell_m^2} \left\{ \int R \frac{\partial}{\partial \beta^T} \log(\ell_o) \right\} \frac{\partial \ell_o}{\partial \beta^T} \, du \right\}$$

(4.37)

The second term of the RHS of (4.37) is

$$\frac{1}{\ell_m} \left\{ \int R \left[ \frac{\partial}{\partial \beta} \log(\ell_o) \right] \left[ \frac{\partial}{\partial \beta^T} \log(\ell_o) \right] \right\} \ell_o \, du$$

$$= E \left[ \frac{\partial}{\partial \beta} \log(\ell_o) \cdot \frac{\partial}{\partial \beta^T} \log(\ell_o) \right]$$
and likewise the third term of (4.37) is equal to

$$E \left[ \frac{\partial}{\partial \beta} \log(t_o) \right] E \left[ \frac{\partial}{\partial \beta} \log(t_o) \right]^T$$

where the expectation is as defined in (4.34). But at maximum likelihood estimate

$$E \left[ \frac{\partial}{\partial \beta} \log(t_o) \right] = 0 .$$

Then

$$\frac{\partial L_m}{\partial \beta \beta^T} = E \left[ \frac{\partial^2 \log(t_o)}{\partial \beta \beta^T} \right] + E \left[ \frac{\partial \log(t_o)}{\partial \beta} \right] \left[ \frac{\partial \log(t_o)}{\partial \beta} \right]^T / R, \beta$$

$$= E \left[ \frac{\partial^2 \log(t_o)}{\partial \beta \beta^T} \right] + \text{Cov} \left[ \frac{\partial \log(t_o)}{\partial \beta}, \frac{\partial \log(t_o)}{\partial \beta} \right] / R, \beta .$$

(4.38)

The variance-covariance matrix of the estimate, $\hat{\beta}$, is then the inverse of $A(\hat{\beta})$ where

$$A(\hat{\beta}) = - \frac{\partial L_m}{\partial \beta \beta^T} \bigg|_{\beta = \hat{\beta}} .$$

4.5.3 Rank Order Expectation for Exponential Variate.

It was shown in section (4.2.2) that the marginal likelihood of the proportional hazards model may be generated from exponential distribution. Accordingly the rank order expectation for such distribution has a vital role in the development of the EM algorithm.
The expected value $E(T(i)/R, \beta)$ can be found by standard method for order statistics (see David, 1981, Chapter 3). However in this section we shall arrive at it by using similar argument presented in section (4.3).

Let $T(1) < T(2) < \ldots \ldots < T(k)$, be $k$ distinct ordered variables from the exponential distribution with mean $\frac{1}{\lambda_j}$; $j = 1, 2, \ldots, k$.

From (4.34), the rank order of $T(a)$ is:

$$E_R\left[\frac{T(a)}{R}\right] = \frac{\int_R \left[ \prod_{j=1}^{k} \lambda_j e^{-\lambda_j t(j)} \right] dt}{\int_R \left[ \prod_{j=1}^{k} \lambda_j e^{-\lambda_j t(j)} \right] dt}$$

$$= \frac{1}{k!} \int_R \left[ \prod_{j=1}^{k} \lambda_j e^{-\lambda_j t(j)} \right] dt$$

where $dt = dt(1) \cdot dt(2) \cdot \ldots \cdot dt(k)$ and $\lambda$ the marginal likelihood of the rank order statistics. By the same argument as in section (4.3) and substituting $t(j)$ by $X_j$ we have

$$E_R\left[\frac{T(a)}{R}\right] = \frac{1}{k!} E \left\{ \prod_{j=1}^{k} e^{-X_j(\lambda(j)^{-1})} \right\}$$

$$= \left( \frac{k \prod_{j=1}^{k} \lambda_j}{k! \lambda_m} \right) E \left\{ \prod_{j=1}^{k} e^{-X_j(\lambda(j)^{-1})} \right\}$$

(4.39)

where the expectation in the RHS of the equation is with respect
to the joint distribution of $0 < X_1 < X_2 \ldots < X_k$ the order statistics from a random sample of size $k$, from an exponential distribution mean 1. Substituting $X_{\alpha} = \sum_{\xi=1}^{\alpha} c_{\xi} Z_{\xi}$ where $c_{\xi} = \frac{1}{k-\xi+1}$, the expectation in the RHS of equation (4.37) becomes:

$$E\left\{ \left( \sum_{\xi=1}^{\alpha} c_{\xi} Z_{\xi} \right) \exp \left\{ - \sum_{i=1}^{k} \sum_{j=i}^{k} c_{i} Z_{i} (\lambda_{(j)} - 1) \right\} \right\}. \quad (4.40)$$

By noting that for arbitrary $A_{j}$ and $B_{i}$

$$\sum_{j=1}^{k} \sum_{i=1}^{k} A_{j} B_{i} = \sum_{i=1}^{k} \sum_{j=i}^{k} A_{j} B_{i} = \sum_{i<j} A_{j} B_{i}$$

(4.39) becomes:

$$E\left\{ \left( \sum_{\xi=1}^{\alpha} c_{\xi} Z_{\xi} \right) \exp \left\{ - \sum_{i=1}^{k} \sum_{j=i}^{k} c_{i} Z_{i} (\lambda_{(j)} - 1) \right\} \right\}$$

putting $b_{i} = c_{i} \sum_{j=i}^{k} (\lambda_{(j)} - 1)$ we have

$$E\left\{ \left( \sum_{\xi=1}^{\alpha} c_{\xi} Z_{\xi} \right) \exp \left\{ - \sum_{i=1}^{k} b_{i} Z_{i} \right\} \right\}$$

$$= E\left\{ \left( \sum_{\xi=1}^{\alpha} c_{\xi} Z_{\xi} \right) \prod_{i=1}^{k} e^{-b_{i} Z_{i}} \right\}$$

$$= \sum_{\xi=1}^{\alpha} c_{\xi} E \left[ Z_{\xi} \prod_{i=1}^{k} e^{-b_{i} Z_{i}} \right]. \quad (4.41)$$

It has been proved in section (4.3) that
Differentiating the LHS and RHS respectively with respect to $b_k$

\[
\frac{\partial E}{\partial b_k} \left[ k \prod_{i=1}^{k} \frac{-b_i Z_i}{(1 + b_i)} \right] = E \left\{ \frac{k}{\prod_{i=1}^{k} (1 + b_i)} \right\} \tag{4.42}
\]

and

\[
\frac{\partial}{\partial b_k} \left[ k \prod_{i=1}^{k} \frac{1}{(1 + b_i)} \right] = \left[ - \frac{1}{(1 + b_k)} \right] \frac{1}{\prod_{i=1}^{k} (1 + b_i)} \tag{4.43}.
\]

Hence from (4.42) and (4.43)

\[
E \left[ k \prod_{i=1}^{k} \frac{-b_i Z_i}{(1 + b_i)} \right] = \frac{1}{k} \frac{k}{\prod_{i=1}^{k} (1 + b_i)} = \frac{k!}{k \prod_{i=1}^{k} \lambda_i (r)} \cdot \frac{1}{1 + b_k}.
\]

since

\[
\prod_{i=1}^{k} (1 + b_i) = \frac{1}{k!} \prod_{i=1}^{k} \lambda_i (r) \tag{section 4.3}
\]

Hence (4.41) can now be written as:

\[
\frac{k!}{k \prod_{i=1}^{k} \lambda_i (r)} \sum_{r=i}^{\alpha} \left[ \frac{c_k}{1 + b_k} \right].
\]
\[
\begin{align*}
\frac{k! \sum_{k=1}^{a} c_{\ell} \left( \frac{1}{1 + \sum_{r=\ell}^{k} (\lambda_{r} - 1)} \right)}{\prod_{r=1}^{k} \lambda(r)} &= \frac{k!}{\prod_{i=1}^{k} \sum_{r=i}^{k} \lambda(r)} \\
&= \left[ \frac{k!}{\prod_{i=1}^{k} \sum_{r=i}^{k} \lambda(r)} \right]^{\alpha \sum_{\ell=1}^{a} \frac{1}{\sum_{r=\ell}^{k} \lambda(r)}} \\
&= \left( \frac{k!}{\prod_{i=1}^{k} \sum_{r=i}^{k} \lambda(r)} \right) \left( \frac{1}{\prod_{r=\ell}^{k} \lambda(r)} \right) \\
&\text{since } c_{\ell} = \frac{1}{k-\ell+1} .
\end{align*}
\]

Substituting (4.44) into (4.39), and since from (4.23)

\[
\hat{\ell}_{m} = \frac{k!}{\prod_{i=1}^{k} \sum_{r=i}^{k} \lambda(r)}
\]

we have

\[
E_{R}\left[ T(\alpha) / R \right] = \sum_{\ell=1}^{a} \frac{1}{\prod_{r=\ell}^{k} \lambda(r)} \\
= \sum_{\ell \leq t_{\alpha}} \left[ \sum_{r \in R_{\ell}} \frac{1}{\lambda_r} \right] \\
= \sum_{\ell \leq t_{\alpha}} \left[ \sum_{r \in \mathbb{R}_{\ell}} \frac{1}{\lambda_r} \right] \\
\text{where } R_{\ell} = \left\{ i \mid T_i \geq T_{\ell} \right\} .
\]
4.5.4 Poisson Random Error.

The likelihood for complete data-set up for general censored data with life times \( t_1, t_2, \ldots, t_n \) and censoring indicator \( \delta_i \) (\( \delta_i = 1 \) failure; \( \delta_i = 0 \), censored) is

\[
\mathcal{L}_o = \prod_{i=1}^{n} h(t_i)^{\delta_i} S(t_i)^{1-\delta_i}.
\]

For the proportional hazard model, [as in section (4.4.2)]

\[
\mathcal{L}_o = \prod_{i=1}^{n} \left[ \frac{h_0(t_i)}{H_0(t_i)} \right]^{\delta_i} \exp \left[-\lambda_i H_0(t_i)\right]
\]

\[
= \prod_{i=1}^{n} \left[ \frac{\delta_i}{\mu_i} \exp \left(-\mu_i\right) \right] \left[ \frac{h_0(t_i)}{H_0(t_i)} \right]^{\delta_i}
\]

where \( \lambda_i = \exp(z_i^T \beta) \) and \( \mu_i = \lambda_i H_0(t_i) \).

As was discussed in (4.4.2), since the second term does not depend on \( \lambda_i \), its estimation, hence the estimation of \( \beta \), may be arrived at by maximising the first term which is the likelihood of \( n \) independent Poisson variables \( \delta_i \) with mean \( \mu_i \).

Now

\[
\log \mu_i = \log H_0(t_i) + z_i^T \beta
\]

\[
= \log H_0(t_i) + \eta_i
\]

(4.47)

where \( \eta_i = z_i^T \beta \), the linear predictor of the model.

The general scheme of the iterative weighted least square is to regress the adjusted dependent variable \( v_i \) on the covariates.
\[ v_i = \eta_i + (y_i - \mu_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right) \]

\[ \omega_i = \left( \frac{\partial \mu_i}{\partial \eta_i} \right)^2 \frac{1}{\nu(\mu_i)} \]

where \( y_i \) is the random observations and \( \nu(\mu) \) the variance function.

For the poisson model discussed here we have from (4.47)

\[ \frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i} \]

\[ \nu(\mu) = \mu \]

Hence

\[ v_i = \eta_i + (\delta_i - \mu_i) \frac{1}{\mu_i} \]

\[ \omega_i = \mu_i \]

with

\[ \log(\mu_i) = \log(H_o(t_i)) + \eta_i \]

The offset (see Baker and Nelder, 1978) is thus \( \log[H_o(t_i)] \).

However \( H_o(t_i) \) is unknown, only that it is assumed to be constant between failures, that is if \( t(1) < t(2) < \ldots < t(j) < t(j+1) < \ldots < t(n) \)

are ordered failures then

\[ H_o(t) = H_o(t_j) \quad \text{for} \quad t(j) < t \leq t(j+1) \]

Thus if \( t_i \) is censored

\[ H_o(t_i) = H_o(t(j)) \quad \text{where} \quad t(j) \leq t_i < t(j+1) \]
Now the marginal likelihood for the model is

$$
\mathcal{L}_m = \int_R \prod_{i=1}^n h(t_i)^{\delta_i} S(t_i) \, dt \quad \text{(4.48)}
$$

Substituting $u_i = H_D(t_i)$, with transformation as in section 4.2.2 equation (4.11), and setting $u_i$ equal to the largest $u_i$ corresponding to a failure time smaller than $t_i$, the marginal likelihood (4.48) can now be written as

$$
\mathcal{L}_m = \int_R \prod_{i=1}^n \left[ \lambda_i \exp(-\lambda_i u_i) \right] \, du \quad \text{(4.49)}
$$

where $R$ now represents the rank information for the uncensored $u_i$. The multiple integral is evaluated only over these $u_i$.

Now the inner part of (4.49) is a parametric likelihood for the exponential model with mean $\frac{1}{\lambda_i}$. Then the expectation with respect to the rank information, $R$ as defined above is thus (from section 4.5.2, equation 4.45)

$$
\bar{u} = E_R \left[ \frac{u_i}{R} \right] = \sum_{t_{\beta} < t_{\alpha}} \frac{\delta_{t_{\beta} \in R_{\bar{x}}}}{\sum_{\alpha} \lambda_{\alpha}} \quad \text{(4.50)}
$$

where

$$
R_{\bar{x}} = \left\{ i : u_i \geq T_{\bar{x}} \right\}
$$

Furthermore the EM algorithm requires $E \left[ \frac{\partial}{\partial \beta} L_0(\beta) / R \right]$.

From (4.32) and (4.48) we have

$$
L_0(\beta) = \prod_{i=1}^n \left[ \lambda_i^{\delta_i} \exp(-\lambda_i u_i) \right]
$$

then
since \( \lambda_i = \exp(z_i^T g) \), and the expectation with respect to the rank information is

\[
E_R \left[ \frac{\partial L_o(g)}{\partial \beta} / R \right] = \frac{n}{n} \sum_{i=1}^{n} \delta_i z_i - z_i \lambda_i \ E_R (u_i / R)
\]

\[
= \frac{n}{n} \sum_{i=1}^{n} \delta_i z_i - z_i \lambda_i \ \sum_{j \in R} \left( \sum_{j \in R} \lambda_j \right)
\]

\[
= \frac{n}{n} \sum_{i=1}^{n} \delta_i \left( z_i - a_i \right)
\]

where \( a_i = \sum_{j \in R_i} z_j \lambda_j / \sum_{j \in R_i} \lambda_j \), (see (4.26), and by noting that for arbitrary \( A_j, B_i \) we can change the order of summations such that:

\[
\sum_{i=1}^{n} B_i \sum_{j \in R_i} A_j = \sum_{j=1}^{n} A_j \sum_{i \in R_j} B_i
\]

Thus we see here that the expectation step of the EM algorithm results in the expression of the score vector of equation (4.26).

The success of the GLIM poisson model to arriving at the maximum likelihood estimates of \( g \) may be seen from the following:
Now if we approximate the marginal likelihood (4.49) as that of the complete data likelihood of (4.46) from the "poisson" distribution with \( u_i = H_0(t_i) \) being replaced by its expected value, \( \overline{u}_i \) in equation (4.48), then the maximum likelihood of \( \beta \) is the solution to \( \frac{\partial}{\partial \beta} \log(L_o) = 0 \).

Now from (4.46), with \( \mu_i = \lambda_i \overline{u}_i \)

\[
\log(L_o) = \sum_{i=1}^{n} \left\{ \delta_i \left[ \log(\lambda_i) + \log \overline{u}_i \right] - \lambda_i \overline{u}_i \right\} + \phi(t_i)
\]

\[
= \sum_{i=1}^{n} \left\{ \delta_i \left[ z_i^T \beta + \log \overline{u}_i \right] - \overline{u}_i \exp(z_i^T \beta) \right\} + \phi(t_i)
\]

where \( \phi(t_i) \) some function of the life times independent of the parameter \( \beta \). Then

\[
\frac{\partial}{\partial \beta} \log(L_o) = \sum_{i=1}^{n} \left\{ \delta_i z_i - \overline{u}_i z_i \exp(z_i^T \beta) \right\}
\]

which is identical to the expectation step in (4.51). Then the offset in the log link function (4.47), \( \log H_0(t_i) \), can now be replaced by the logarithm of the expected value of \( H_0(t_i) \), as defined in (4.50).

The iterative weighted least squares for the poisson approach can now be summarized as follows:

1) Compute \( \hat{H}_o(t) = E_r \left[ H_o(t)/R \right] \) from some suitable starting value of \( \beta \).

2) Set \( \mu_i = \hat{H}_o(t_i) \exp(z_i^T \beta) \)

3) Regress \( \eta_i + (\delta_i - \mu_i)/\mu_i \) on the covariates \( z_i \) with weights \( \mu_i \) and with \( \log \mu_i = \log \hat{H}_o(t_i) + \eta_i \).
Now this can be implemented in GLIM. The procedures are as follows:

1) Set initial estimate of $\lambda_i$, usually taken to be 1 for each $i$ (or equivalently set $\beta = 0$).

2) Order all observations, censored and uncensored in ascending order, $T(1), T(2), \ldots, T(n)$. Observations that tie are randomly ordered. Censored observations that tie with an uncensored are given higher rank. The corresponding covariates are mean centered, that is, they sum to zero.

3) Compute $\bar{u}_i = E_R \left[ \frac{H_i(t_i)}{R} \right]$ as in (4.50).

4) Fit a poisson model:

   \begin{align*}
   \text{yvariate} & \quad \delta_i \\
   \text{link} & \quad \log \\
   \text{offset} & \quad \log(\bar{u}_i) 
   \end{align*}

5) Get estimate $\hat{\beta}$ and construct new $\lambda_i = \exp(z^T \hat{\beta})$.

6) Go to step 3 with this $\lambda_i$, and repeat succeeding steps until convergence is achieved.

GLIM macros to carry out the above calculations are described in Clayton & Cuzick (1985a).

4.5.5 Gamma Random Error.

Consider $n$ observations, censored and uncensored. The likelihood for the complete data set-up is

$$
\mathcal{L}_o = \prod_{i=1}^{n} f(t_i)
$$
Here \( t_i \) are the life times which may be imagined time of failure for the censored observations. Since

\[ f(t_i) = h(t_i) S(t_i) \]

then for the proportional hazards model this likelihood is

\[
\ell_o = \prod_{i=1}^{n} h_o(t_i) \lambda_i \exp\left[ - \lambda_i H_o(t_i) \right],
\]

where \( h_o(t_i) \), and hence \( H_o(t_i) \) are unknown. As in preceding section \( H_o(t_i) \) are assumed to be constant between failure times.

Also we have shown in section (4.2.2), that \( H_o(t_i) \) is exponentially distributed with mean \( 1/\lambda_i \).

Substituting \( u_i = H_o(t_i) \), then

\[
\ell_o = \prod_{i=1}^{n} \lambda_i \exp(-\lambda u_i). \tag{4.52}
\]

The marginal likelihood is thus

\[
\ell_m = \int_{\mathbb{R}} \prod_{i=1}^{n} \frac{\lambda_i}{\exp(-\lambda u_i)} \, du_i,
\]

where \( \mathbb{R} \) now represents the \( n \)-tuple integration in region where \( u_i \) is set to equal to largest \( u_i \) corresponding to a failure time smaller than \( t_i \). The expectation of (4.34) is defined accordingly in this region. Now for uncensored, \( u_\alpha \) the expectation as in (4.45) still applies, that is

\[
E_{R}(u_\alpha / R) = \sum_{j \leq t} \frac{1}{\sum_{r \in R_j} \lambda_r}.
\]

However, since the integration also includes the censored cases,
we will need to determine its expectation. Here again the
expectation may be obtained directly by repeated integration,
but we shall derive it by using the Markov property for exponential
order statistics. Let $T_i$ be random variable from exponential
distribution with mean $1/\lambda_i$. Now consider an observation censored
at, say $T_j$ and let $T_{(s_j)} < T_j < T_{(s_j+1)}$, where $T_{(s_j)}$ and $T_{(s_j+1)}$ are
consecutive failure times. Then by the Markov property

$$E_{T_j/T_{s_j}, R} \left[ T_j / T_j > T_{(s_j)}, T_{(s_j)} \right] = T_{(s_j)} + \frac{1}{\lambda_j} \quad (4.53)$$

where the expectation $E_{T_j/T_{s_j}, R}$ is with respect to the conditional
distribution of $T_j$ given $T_j > T_{(s_j)}, T_{(s_j)}$ and the ranks. Then

$$E_{T_j/R}(T_j) = E_{T_j, T_{s_j}} / R \left[ T_j / R, \lambda \right]$$

$$= E_{T_{s_j}} / R \left[ E_{T_j/T_{s_j}, R} \left[ T_j / T_j > T_{(s_j)}, T_{(s_j)} \right] \right]$$

where $E_{T_j, T_{s_j}} / R$ is the expectation with respect to the joint
distribution of $T_j$ and $T_{(s_j)}$ given the ranks; and $E_{T_{s_j}} / R$ is the
expectation with respect to the distribution of the $T_{(s_j)}$ given
the rank. For the sake of brevity we shall disregard the subscripts
of the expectations. There should be no confusion with regards to
their meanings.

Substituting the expectation of (4.53), we have
\[
E\left[T_j/R, \lambda\right] = E\left\{T_{(s_j)} + \frac{1}{\lambda_j}\right\}
\]

\[
= E\left[T_{(s_j)}/R, \lambda\right] + \frac{1}{\lambda_j} .
\]  \hspace{1cm} (4.54)

The expectation on the RHS is the expectation for failure time \(T_{(s_j)}\) as given in (4.45). Introducing the censoring indicator \(\delta_i\) and letting \(u_i = T_i\) (since \(u_i\) is exponential with mean \(1/\lambda_i\)) we have from (4.54) and (4.45)

\[
\bar{u}_i = E\left[u_i/R, \lambda\right] = \frac{1-\delta_i}{\lambda_i} + \sum_{j \in R_i} \frac{\delta_j}{\sum_{r \in R_j} \lambda_r}
\]  \hspace{1cm} (4.55)

where \(u_i\) is the value corresponding to \(t_i\) whether observed or censored.

The \(E\left\{\frac{\partial \log(L)}{\partial \beta} / R\right\}\) which is required in the EM algorithm can now be computed, and from (4.52) with \(\lambda_i = \exp(z_i^T \beta)\) we obtain

\[
E\left\{\frac{\partial \log(L)}{\partial \beta} / R\right\} = \sum_{i=1}^n \left[ z_i - z_i \exp(z_i^T \beta) \{E(u_i/R)\} \right]
\]

\[
= \sum_{i=1}^n \left[ z_i - z_i \lambda_i \left\{ \frac{1-\delta_i}{\lambda_i} + \sum_{j \in R_i} \frac{\delta_j}{\sum_{r \in R_j} \lambda_r} \right\} \right]
\]

\[
= \sum_{i=1}^n \left[ \delta_i z_i - z_i \lambda_i \sum_{j \in R_i} \frac{\delta_j}{\sum_{r \in R_j} \lambda_r} \right]).
\]

By changing the order of summation, and letting
\[
\frac{\bar{a}_i}{a_i} = \sum_{j \in R_i} z_j \lambda_j / \sum_{j \in R_i} \lambda_j
\]

we have an identical equation as that of the poisson scheme, equation (4.51), which is the score vector in equation (4.26).

The Gamma error scheme begins with the approximation of the marginal likelihood by the exponential likelihood (4.52) with the \( u_i \) replaced by its expected value, \( \bar{u}_i \), defined by equation (4.55)

\[
\ell_m = \prod_{i=1}^{n} \lambda_i \exp(-\lambda_i \bar{u}_i)
\]

then

\[
\frac{\partial \log(\ell_m)}{\partial \beta} = \sum_{i=1}^{n} \left[ z_i - z_i \lambda_i \bar{u}_i \right]
\]

\[
= \mathbb{E} \left[ \frac{\partial \log(\ell_o)}{\partial \beta} \right] / R
\]

\[
= \sum_{i=1}^{n} \delta_i (z_i - a_i)
\]

Thus we see here the connection of the EM algorithm with the proposed model under discussion, and how the scheme will lead us to maximum marginal likelihood estimates. In the notation of the generalised linear model for exponential error:

\[
\eta_i = z_i^T \beta
\]

\[
\mu_i = \frac{1}{\lambda_i} \exp(-z_i^T \beta)
\]

then

\[
\log(\mu_i) = -\eta_i
\]

\[
\nu(\mu) = \mu^2
\]
and so
\[
\omega_i = \left( \frac{\partial u}{\partial \eta_i} \right)^2 \frac{1}{\hat{v}(\mu)} = 1
\]
and
\[
v_i = \eta_i + (y_i - \mu_i) \left( \frac{\partial \eta_i}{\partial \mu_i} \right)
\]
\[
= \eta_i + \left( 1 - \frac{y_i}{\mu_i} \right)
\]

The regression of \( v_i \) on the covarates \( z_i \) with weight equal to 1, will result in the iterative weighted least square algorithm for exponential error:
\[
\hat{\beta}_{s+1} = \hat{\beta}_s + (Z^T Z)^{-1} Z^T d
\]
(4.56)
where \( d \) has component \( d_i = \left( 1 - \frac{y_i}{\mu_i} \right) \), and \( Z \) is the nxp design matrix of the covariates. Substituting \( y_i \) for \( \bar{u}_i \), the current estimate of \( d_i \) when \( \lambda_i \) is evaluated at \( \beta \), is:
\[
d_i = 1 - \lambda_i \left\{ \frac{1 - \delta_i}{\lambda_i} + \sum_{t_j \in t_i} \left[ \sum_{r \in \mathbb{R}_j} \lambda_r \right] \left[ \sum_{j} \lambda_j \right] \right\}
\]
\[
= \delta_i - \lambda_i \sum_{t_j \in t_i} \sum_{r \in \mathbb{R}_j} \lambda_r
\]
(4.57)

The GLIM implementation follows similar procedure enumerated in the preceding section for the poisson error model. The \( \bar{u}_i = E[H_0(t_i)/R] \) in step 3 has value defined by (4.55). The model defined in step 4 will be gamma with
4.5.6 Comparison of Poisson and Gamma Random Errors.

The EM approaches thus far discussed were applied to a number of data sets (see section 4.6) and the results are compared to the estimates obtained by the Newton-Raphson iteration discussed in (4.4.1). As seen from tables 4.1 - 4.4, the estimates are, in general, in agreement for the different approaches. However the EM algorithm with poisson error has a slight edge over the EM algorithm with gamma error, in that the convergence is faster and it has closer estimates to that obtained by the Newton-Raphson iterations. This is due to the fact that the method takes into account the censoring in its complete-data specification (see equation 4.4.6). Thus we would expect the algorithm to be more efficient than the gamma error approach, even when the number of censoring is large. Furthermore the y variate in GLIM, being the censoring indicator, is constant throughout the algorithm.

In the case of the gamma error approach, the complete-data specification in the EM requires failure times to be estimated for the censored times. With more censoring, the efficiency is less. If censoring is not present, then there is no difference in the estimates between the two.
The standard errors given by GLIM for both the EM approaches are underestimated. This may be seen from the following:

Suppose that the log likelihood be generated by random variables $U_i$; and let the log likelihood for an observation $u_i$ be $\ell(u_i)$. Then in general the observed information matrix computed by the iterative weighted least square scheme is:

$$A(\hat{\beta}) = - \sum_{i=1}^{n} \frac{\partial^2 \ell(u_i)}{\partial \beta \partial \beta^T} = Z^T W Z$$  \hspace{1cm} (4.58)$$

where $W$ is the weight matrix in the regression and $Z$ the $nxp$ design matrix of covariates.

For the proportional hazard from (4.38) we have

$$A(\hat{\beta}) = - E \left[ \sum_{i=1}^{n} \frac{\partial^2 \ell(u_i)}{\partial \beta / R, \beta} \right] - \text{Cov} \left[ \sum_{i=1}^{n} \frac{\partial \ell(u_i)}{\partial \beta}, \sum_{i=1}^{n} \frac{\partial \ell(u_i)}{\partial \beta^T} \right]$$

$$\text{Expanding } \ell(u_i) \text{ about } \bar{u}_i \text{ where } \bar{u}_i = E(u_i / R, \beta) \text{ as defined in (4.34), we have}$$

$$\ell(u_i) \approx \ell(\bar{u}_i) + (u_i - \bar{u}_i) \frac{\partial \ell(u_i)}{\partial u_i} \bigg|_{u_i = \bar{u}_i}$$

$$\frac{\partial^2 \ell(u_i)}{\partial \beta \partial \beta^T} \approx \frac{\partial^2 \ell(\bar{u}_i)}{\partial \beta \partial \beta^T} + (u_i - \bar{u}_i) \frac{\partial^2}{\partial \beta \partial \beta^T} \left[ \frac{\partial \ell(u_i)}{\partial u_i} \right]_{u_i = \bar{u}_i}$$

Then the first term of the RHS of (4.59) is

$$\sum_{i=1}^{n} \frac{\partial^2 \ell(\bar{u}_i)}{\partial \beta \partial \beta^T} + \sum_{i=1}^{n} E(u_i - \bar{u}_i) \frac{\partial^2}{\partial \beta \partial \beta^T} \left[ \frac{\partial \ell(u_i)}{\partial u_i} \right]_{u_i = \bar{u}_i}$$
\[ = \sum_{i=1}^{n} \frac{\partial^2 \ell(u_i)}{\partial \theta \partial \theta^T} \]

since the second term of the above expression is zero. Similarly the second term of (4.59) is

\[ E \sum_{i=1}^{n} \left\{ \frac{\partial \ell(u_i)}{\partial \theta^T} + (u_i - \bar{u}_i) \frac{\partial}{\partial \theta} \left( \frac{\partial \ell(u_i)}{\partial u_i} \right) \right\} \times \]

\[ \sum_{i=1}^{n} \left\{ \frac{\partial \ell(u_i)}{\partial \theta^T} + (u_i - \bar{u}_i) \frac{\partial}{\partial \theta} \left( \frac{\partial \ell(u_i)}{\partial u_i} \right) \right\} \]

\[ = \sum_{i=1}^{n} E \left\{ (u_i - \bar{u}_i)^2 \frac{\partial^2 \ell(u_i)}{\partial \theta^2 u_i} \right\} \]

since all other expectation terms in the expansion are equal to zero.

In the matrix form this can be written as

\[ B^T M B \]

where \( B \) is the \( n \times p \) matrix with elements

\[ B_{ij} = \frac{\partial^2 \ell(u_i)}{\partial u_i \partial \theta_j} \bigg|_{u_i = \bar{u}_i} \]

and \( M \) is the \( n \times n \) covariance matrix of the order statistics with elements

\[ m_{ij} = E \left[ (u_i - \bar{u}_i)(u_j - \bar{u}_j) / R_i \theta \right] \]

\[ = \text{Cov}(u_i, u_j / R_i \theta) \]
Then equation (4.60) can be approximated by

$$A(\beta) = \sum_{i=1}^{n} \frac{\partial^2 \lambda(u_i)}{\partial \beta \partial \beta^T} - B^T B$$  \hspace{1cm} (4.60)$$

The EM schemes described above make use of $\lambda(u_i)$ in forming the likelihood for the models. Hence GLIM calculates only the first term of equation (4.60). However the loss is quite small in some circumstances (Kalbfleisch and Prentice, 1980, section 4.7) and the estimates are perfectly adequate in most cases. The one drawback is that the convergence is slow compared to Newton-Raphson iteration method. It may not be ideal for routine use, but the approach can be developed later (section 5.4) to provide a scheme for an influence measure.

The observed information matrices computed by GLIM for the two EM schemes are presented below:

a) **Poisson Error**.

The log likelihood for this scheme is

$$\ell(\bar{u}_i) = \delta_i \log \lambda_i - \lambda_i \bar{u}_i, \text{ where } \bar{u}_i = E\left[H_o(t_i)/R, \beta \right].$$

Then

$$\sum_{i=1}^{n} \frac{\partial^2 \lambda(u_i)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^{n} \bar{u}_i z_i z_i^T \lambda_i$$

$$= - \sum_{i=1}^{n} E\left[H_o(t_i)/R, \beta \right] z_i z_i^T \lambda_i$$

$$= - E\left[\sum_{i=1}^{n} H_o(t_i) z_i z_i^T \lambda_i \right],$$
\[ \mathbf{W} = \text{diag}(\mu_i) \text{ and } \mathbf{Z} \text{ the } (n \times p) \text{ design matrix of covariates.} \]

The observed information matrix is thus
\[ A(\beta) = \mathbf{Z}^T \mathbf{W} \mathbf{Z} \]

which is in agreement with (4.58). The variance covariance matrix for the maximum likelihood estimate is the inverse of \( A(\beta) \) evaluated at \( \hat{\beta} \).

b) **Gamma Error.**

The log likelihood for the gamma scheme is
\[ \ell(\mathbf{u}_i) = \sum_{i=1}^{n} \left( \log \lambda_i - \lambda_i \bar{u}_i \right) \]

Then
\[ \frac{\partial^2 \ell(\mathbf{u}_i)}{\partial \beta \partial \beta^T} = - \sum_{i=1}^{n} \left\{ E \left[ H_0(t_i)/R_i, \beta \right] \mathbf{z}_i \mathbf{z}_i^T \lambda_i \right\} \]

which is identical to that produced by the poisson error approach; but expectation defined here is not in the same region (see in equation 4.49 and section 4.5.5), and therefore not equal to the \( \mathbf{Z}^T \mathbf{W} \mathbf{Z} \) from the poisson error. Note though that for known \( H_0(t) \), the \( E_{T_i} \left[ H_0(t_i) \lambda_i \right] = 1 \), where the expectation is with respect to the distribution of \( T_i \). However \( \bar{u} = E(H_0(t_i)/R_i, \beta) \) is imagined to come from exponential distribution with mean \( 1/\lambda \). Then the Fisher information matrix is
\[
E_u \left[ - \frac{\partial^2 g(u_i)}{\partial \beta \partial \beta^T} \right] = \sum_{i=1}^{n} E_u(\tilde{u}) z_i z_i^T \lambda_i
\]

\[= \sum_{i=1}^{n} z_i z_i^T \]

\[= z^T z\]

where the expectation in the above equation is with respect to the distribution of \(\tilde{u}\). Approximating the observed information with the above as in (4.58), we see that the \(A(\hat{\beta})\) for the proportional hazard can be approximated by \(z^T z\), and the inverse of it would give the variance-covariance matrix of \(\hat{\beta}\). A point to note here, for the scheme to work efficiently, the columns of the covariate matrix \(z\) should be mean centered, that is they sum to zero.

4.6 EXAMPLES AND RESULTS.

The following data sets were run to compare the estimates put out by the EM approaches described in section 4.5 to that obtained by the Newton-Raphson iterations. The GLIM fits are done without the grand mean. In all the computations, the convergence criterion employed is the change in the parameter estimates through the iterations. One might use the attainment of the maximum likelihood as the stopping criterion, but we found that convergence is slow. This reflects that the likelihood function has almost flat, plateau-like shape near its maximum.
4.6.1 Carcinogenesis data (Appendix B4).

The estimates for the different methods are shown in Table 4.1. There is a single categorical factor to differentiate the group, which is taken as the covariate in the model. The data set is not heavily censored. Only four observations are censored. All the estimates are identical at -0.5959. However, the standard error computed by GLIM for the EM algorithm with gamma error, EM-E, is smaller than that of the Newton-Raphson method. This is true, in general as was discussed in section (4.5.6) that the standard errors computed by GLIM are underestimated. The loss is small though, less than 9%. The EM algorithm with poisson error, EM-P, as expected gives out better estimate of the standard error compared to the GLIM algorithm for the EM-E. In terms of computational speed, there is little difference in efficiency between the two EM approaches. This may be attributed to the low proportion of censoring in the data set.

<table>
<thead>
<tr>
<th></th>
<th>EM-E</th>
<th>N-R</th>
<th>EM-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>est.</td>
<td>-0.5959 (0.3185)</td>
<td>-0.5959 (0.3484)</td>
<td>-0.5959 (0.3338)</td>
</tr>
<tr>
<td>iter</td>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>sig.</td>
<td>0.061</td>
<td>0.087</td>
<td>0.075</td>
</tr>
</tbody>
</table>

EM-E  EM method with exponential error
N-R    Newton-Raphson iteration method
EM-P   EM method with poisson error.

Cox Proportional hazard estimates for CARCINOGENESIS data.

TABLE 4.1.
The partial log likelihood at $\hat{\beta}$ is -100.7190. The last row in the table gives the level of significance for each estimate using the different methods. Under the null hypothesis of $\beta = 0$. The Newton-Raphson and EM-P show some indication of treatment effect, though the evidence is not strong.

To test whether $\beta = 0$, there are a number of standard statistics available, for example, the likelihood ratio test or tests based on the score statistics which refer to $\chi^2$ distribution. The deviance statistic to test $\beta = 0$ is the difference in the partial deviance for null model (when $\beta = 0$) and partial deviance computed at $\beta = \hat{\beta}$, and is found to be 2.878, which is significant at 10%, but not significant at 5%, referring to $\chi^2(1)$.

4.6.2 Veteran's Administration Lung Cancer Data (Appendix B5).

The data set has 9 censored observations out of 137 life times. There are 8 regressor variables. The parameter estimates are as shown in table 4.2 (alongside are the Kalbfleisch and Prentice, 1980, page 60, estimates using the Weibull distribution, for comparison). As discussed in the preceding sections, with low proportion of censored observations, the estimates are almost identical among the various methods. Again we see that the standard error computed by GLIM is fractionally lower than that computed from the information matrix (4.27). There is also little difference in the number of iterations between the EM methods, being 13 and 11 for the EM with gamma error and EM with poisson error respectively. This is of course due to the small number of censored observations in the data. The partial likelihood deviance for the null model
Parameter Estimates using Proportional Hazard model regression
[* Weibull regression model estimates]
Values in parentheses are the standard errors

| Variable       | EM-E          | N-R          | EM-P          | Wei *
|----------------|---------------|--------------|---------------|------
| z1             | -0.0315 (0.005) | -0.0315 (0.006) | -0.0315 (0.005) | 0.0301 |
| z2             | -0.0065 (0.008) | -0.0065 (0.009) | -0.0065 (0.009) | 0.0061 |
| z3             | 0.0004 (0.008)  | 0.0004 (0.008)  | 0.0004 (0.008)  | -0.0005 |
| z4             | -0.3512 (0.267) | -0.3520 (0.287) | -0.3520 (0.276) | 0.3977 |
| z5             | 0.4970 (0.258)  | 0.4981 (0.270)  | 0.4981 (0.266)  | -0.4285 |
| z6             | 0.8392 (0.288)  | 0.8381 (0.306)  | 0.8381 (0.297)  | -0.7350 |
| z7             | -0.2626 (0.183) | -0.2643 (0.210) | -0.2643 (0.202) | 0.2061 |
| z8             | 0.0042 (0.021)  | 0.0041 (0.023)  | 0.0041 (0.022)  | -0.0041 |

| iter  | 13 | 4 | 11 |

Partial loglikelihood: -471.2998

TABLE 4.2 Cox Proportional Hazard Estimates for the Veteran Administration lung cancer data with 8 regressors in the model

<table>
<thead>
<tr>
<th>Variable</th>
<th>EM-E</th>
<th>N-R</th>
<th>EM-P</th>
</tr>
</thead>
<tbody>
<tr>
<td>z1</td>
<td>-0.0302 (0.004)</td>
<td>-0.0302 (0.005)</td>
<td>-0.0302 (0.005)</td>
</tr>
<tr>
<td>z4</td>
<td>-0.2824 (0.266)</td>
<td>-0.2827 (0.280)</td>
<td>-0.2827 (0.269)</td>
</tr>
<tr>
<td>z5</td>
<td>0.4416 (0.249)</td>
<td>0.4424 (0.264)</td>
<td>0.4424 (0.260)</td>
</tr>
<tr>
<td>z6</td>
<td>0.8803 (0.277)</td>
<td>0.8807 (0.296)</td>
<td>0.8807 (0.284)</td>
</tr>
</tbody>
</table>

| iter  | 11 | 4 | 10 |

Partial loglikelihood: -472.2794

TABLE 4.3 Cox Proportional Hazard Estimates for the Veteran Administration lung cancer data with 4 regressors [performance & cell types] in the model.
is 1002.331, and partial likelihood deviance at $\hat{\beta}$ is 942.596, which gives us the deviance statistic of 59.735 which is highly significant on $x^2(8)$. Further investigations find the performance status (variable $z_1$) and cell types (variables $z_4$, $z_5$, and $z_6$) are highly significant. Estimates for the model with only these variables are as in Table 4.3. Similar characteristics with regards to the methods prevail. The deviance is 57.77 which is highly significant, referring the value to $x^2(4)$.

A full analysis on this data set is found in Kalbfleisch and Prentice, 1980, section 4.5.

4.6.3 Heart Transplant Data (Appendix B6).

These data sets from the Stanford heart transplant programme has been analysed extensively by various authors. The earlier data sets were analysed by Miller (1976), Crowley and Hu (1979), and Buckley and James (1979). Miller and Helpern (1982) and Pettitt (1983) analysed the recent data set. Pettitt's (1983) table 1 draws the comparison of the regression parameter estimates among the different methods. Here we consider regression models with covariates age, T5 score and age $\times$ age, and use the 152 cases for which survival time is at least 10 days. There were 97 deads, and 55 alive, censored, at the time the data was collected. With transformations

$$A = (\text{age} - \text{mean of age})/10,$$

$$A^2 = A \times A,$$

$$T5 = (T5 \text{ score})/100,$$

we fitted six different models and the estimates obtained in each
model and by the 3 methods are given in table 4.4. They are in agreement up to 2 decimal places. The EM method with poisson error gives estimates identical to that of Newton-Raphson iterations; but there is a slight deviation for the estimates produced by the EM algorithm with the gamma error. This is as expected due to reasons described in section 4.5.6. With high proportion of censoring, in this case about 36%, the deviation may be more pronounced. The standard errors computed by GLIM (given in parenthesis) are, as expected, slightly lower. They are too small to render any degree of underestimation for the estimates.

It is also observed that the number of iterations needed to arrive at the maximum likelihood estimates for the EM method with gamma error is about twice of the other two methods. The efficiency of the EM with poisson errors over the EM with gamma errors is more pronounced here than in the other data sets described in preceding sections. This is due to the high proportion of censoring present in the data set. The poisson errors method takes into account of censoring whereas the gamma errors method does not in the complete data specification.

The $\chi^2$ value given in the table are those required to test the null hypothesis, $H: \beta = 0$. They are $D_{\beta=0} - D_{\beta=\hat{\beta}}$, where the deviance, $D_{\beta} = -2*L(\beta)$ is the deviance of the partial loglikelihood, $L(\beta)$, computed at $\hat{\beta}$. These have asymptotic distribution of $\chi^2$ with degrees of freedom $p$, the dimension of $\beta$. For two models, one of which is nested within the other, the difference in the $\chi^2$ can be used to test $H: \beta_1 = 0$ in the presence of other parameters.
<table>
<thead>
<tr>
<th>Model</th>
<th>A</th>
<th>T5/100</th>
<th>A**2</th>
<th>Iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>EM-E: 0.3529 (0.077)</td>
<td>-</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>NR : 0.3939 (0.120)</td>
<td>-</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>EM-P: 0.3939 (0.120)</td>
<td>-</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>EM-E: 0.3493 (0.078)</td>
<td>0.2210 (0.143)</td>
<td>-</td>
<td>14</td>
</tr>
<tr>
<td>II</td>
<td>NR : 0.3508 (0.120)</td>
<td>0.2214 (0.188)</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>EM-P: 0.3508 (0.116)</td>
<td>0.2221 (0.186)</td>
<td>-</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>EM-E: 0.5071 (0.095)</td>
<td>0.2305 (0.144)</td>
<td>0.2382 (0.060)</td>
<td>11</td>
</tr>
<tr>
<td>III</td>
<td>NR : 0.5075 (0.114)</td>
<td>0.2314 (0.187)</td>
<td>0.2385 (0.073)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>EM-P: 0.5075 (0.109)</td>
<td>0.2314 (0.182)</td>
<td>0.2385 (0.071)</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>EM-E: -</td>
<td>0.2349 (0.142)</td>
<td>-</td>
<td>9</td>
</tr>
<tr>
<td>IV</td>
<td>NR : -</td>
<td>0.2360 (0.188)</td>
<td>-</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>EM-P: -</td>
<td>0.2361 (0.186)</td>
<td>-</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>EM-E: -</td>
<td>0.2393 (0.143)</td>
<td>0.0448 (0.049)</td>
<td>10</td>
</tr>
<tr>
<td>V</td>
<td>NR : -</td>
<td>0.2393 (0.187)</td>
<td>0.0449 (0.060)</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>EM-P: -</td>
<td>0.2393 (0.186)</td>
<td>0.0449 (0.059)</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>EM-E: 0.5069 (0.094)</td>
<td>-</td>
<td>0.2343 (0.059)</td>
<td>11</td>
</tr>
<tr>
<td>VI</td>
<td>NR : 0.5069 (0.113)</td>
<td>-</td>
<td>0.2344 (0.072)</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>EM-P: 0.5069 (0.108)</td>
<td>-</td>
<td>0.2344 (0.067)</td>
<td>6</td>
</tr>
</tbody>
</table>

The partial loglikelihoods:
MODEL:  I   II   III  IV  V   VI
chi2:  9.75 11.12 20.13  1.55  2.09 18.65

NOTE: EM-E  EM algorithm with gamma error
      EM-P  EM algorithm with poisson error
      NP   Newton-Raphson method

TABLE 4.4  Cox Proportional Hazard regression estimates on the Stanford Heart Transplant (152) data.
\( \beta_2 \) in the bigger model. Thus to test the hypothesis \( H : \beta_1 = 0 \), 
the difference in the \( \chi^2 \) of models containing parameter \((\beta_1, \beta_2)\) 
and \( \beta_2 \) is computed, and reference is then made to \( \chi^2 \) distribution 
with \( p' = p - p_2 \) degrees of freedom, where \( p \) is the dimension of 
\((\beta_1, \beta_2)\) and \( p_2 \) is the dimension of \( \beta_2 \).

The overall conclusion drawn here is that the T5 mismatch 
score has little effect on survival of the patients after the 
transplant. The age of the patient at transplant appears to be 
the dominant factor. Comparing model VI and I to test \( H : \beta_2 = 0 \), 
where \( \beta_2 \) is the regression coefficient of A2; the \( \chi^2 \) value is 
\((18.65 - 9.75) = 8.9\), which is significant referring to \( \chi^2_1 \).
A quadratic model in age appears to be significantly better than 
a linear model.
Chapter 5

Diagnostic and Influence Measures

5.1 INTRODUCTION.

We have seen in the previous chapter that the estimates of the regression parameters for the proportional hazards model may be obtained through the ranks of the life times. It is thus robust to any increasing transformation of the observations. However the estimates can be sensitive to the presence or otherwise of particular cases. In addition there may be time covariates, such that the hazard is of the type:

\[ h(t, z(t)) = h_0(t) \exp(z(t)^T \beta) \]  \hspace{1cm} (5.01)

where \( z(t) \) is a vector function of the failure time itself. (Treatment in greater detail on time-dependent covariates is found in Kalbfleisch and Prentice, 1980, Chapter 5 and Cox and Oakes, 1984, Chapter 8). There are several approaches possible for establishing the validity of the model as defined in the previous chapter. Graphical methods, based on the properties of the cumulative hazard (Kay, 1977; Lagakos, 1980) is one. More analytical approaches are methods testing the null hypothesis of the model described in the previous chapter against some alternatives. Various alternatives are formulated, undefined in Schoenfeld's (1980) approach or general alternative in the case of Anderson's (1982) approach. Cox (1972) tests the model against alternatives where the estimates vary as linear function of times. Moreau et al. (1985) suggested an alternative where \( \beta \) varies with time but remaining
constant within predefined intervals. We shall look into some of these approaches in our subsequent discussion and then adapt them to the EM approach.

5.2 RESIDUAL ANALYSIS.

Various definition for residuals have been suggested in the literature. Most involve an estimate of a quantity which should behave something like a unit exponential random variable following the general idea of Cox and Snell (1968). The idea is that these residuals could be used in plots to determine the validity of the proposed model, whether important explanatory variables have been omitted or the proportional hazard model assumption is not adequate. Two sets of residuals play an important part in later sections of this thesis. These are those suggested by Lagakos (1981) and by Schoenfeld (1982), and both can be considered as score statistics calculated at \( \hat{\beta} = \hat{\beta} \). The first residuals may be used to examine the omission of important variables from the model and the second to examine the proportional hazard assumption of time-dependence of the covariates.

5.2.1 Lagakos's Residuals.

Suppose a variable \( x \) has been omitted from the model (4.07) and suppose the actual model has the hazard function

\[
h(t, z, x) = h_0(t) \exp(z^T \beta + xy)
\]

(5.02)

where \( \gamma \) is an unknown scalar regression parameter. The partial likelihood is now a function of the parameter \( (\beta^T, \gamma)^T \), rather
than \( \beta \). Then the \((p+1) \times 1\) score vector is:

\[
\mathbf{u}(\beta, \gamma) = \begin{bmatrix}
\sum_{i=1}^{n} \delta_i \left[ x_i - \frac{\sum_{k \in R_i} x_k \exp(z_k^T \beta + \gamma \lambda_k)}{\sum_{k \in R_i} \exp(z_k^T \beta + \gamma \lambda_k)} \right]
\end{bmatrix}
\]

Evaluating the score at \( \beta = \hat{\beta} \) and \( \gamma = 0 \), we have the non-zero score component corresponding to \( \gamma \) given by

\[
\sum_{i=1}^{n} \delta_i \left[ x_i - \frac{\sum_{k \in R_i} x_k \exp(z_k^T \hat{\beta})}{\sum_{k \in R_i} \exp(z_k^T \hat{\beta})} \right]
\]

(5.03)

where \( \hat{\lambda}_k = \exp(z_k^T \hat{\beta}) \). Again by noting that

\[
\sum_{j=1}^{n} A_j \sum_{k \in R_j} B_k = \sum_{k=1}^{n} B_k \sum_{t_j < t_k} A_j
\]

then (5.03) is equal to

\[
\sum_{i=1}^{n} \delta_i x_i - \sum_{i=1}^{n} x_i \hat{\lambda}_i \sum_{t_j < t_k} \left( \delta_j / \sum_{j \in R_i} \hat{\lambda}_j \right)
\]

\[
= \sum_{i=1}^{n} \delta_i x_i - \sum_{i=1}^{n} x_i \hat{\lambda}_i \hat{H}_o(t_i)
\]
where

\[ \hat{H}_o(t_i) = \sum_{t_j \leq t_i} \left[ \frac{\delta_i}{\sum_{r \in R_j} \lambda_r} \right] . \]  

(5.04)

Writing \( e_i = \lambda_i \hat{H}_o(t_i) \), the score statistic (5.03) becomes

\[ \sum_{i=1}^{n} (\delta_i - e_i) x_i . \]  

(5.05)

Let \( e_i^* = \delta_i - e_i \), then the score statistic to test that \( \gamma = 0 \)

is

\[ \sum_{i=1}^{n} e_i x_i . \]

We define \( e_i^* \) to be Lagakos's residual, extending the ideas of

Lagakos (1981) to include censored observations. We also note

that the score (4.25) evaluated at \( \hat{\beta} \) is

\[ U(\hat{\beta}) = \sum_{i=1}^{n} e_i^* z_i = 0 . \]  

(5.06)

An alternative argument of obtaining \( e_i^* \) is by considering

the distribution of the integrated hazard function

\[ H(t/z) = \int_{0}^{t} h(u/z) \, du = H_o(t) \ exp(z^T \hat{\beta}) . \]

The density function of the variate given the integrated

hazard is (from 4.04)

\[ f(t) = h(t) \ exp[- H(t)] \]

and putting \( Y = H(t) \), the density of \( Y \) is then

\[ f_Y(y) = \frac{1}{H'(t)} \ h(t) \ exp(-y) \]

\[ = \ exp(-y) \]
since \( H'(t) = h(t) \). Thus in general, if \( T \) has the integrated hazard function \( H(t) \), then \( H(T) \) has a unit exponential distribution, so that applying this to \( T \) with integrated hazard \( H(t/z) \) implies that \( H_0(t) \exp(zT \beta) \) has the unit exponential distribution. The same result might be expected to hold approximately when the \( T \)'s are transformed by the estimated functions, \( e_i = \hat{H}(t_i) = \lambda_i H_0(t_i), \) where \( \lambda_i = \exp(z_i \beta) \). Cox and Oakes, (1984) called this \( e_i \) as the generalized residual. The base-line integrated function, \( H_0(t_i) \), may be estimated by Breslow's (1972) estimate \( \hat{H}(t_i) \) which is identical to (5.04), for an observed \( t_i \). Now \( \hat{H}(t_i) \) only depends on the rank order of the observed times and not their actual values, so that the residual \( e_i^* \) is similarly dependent. Thus the residual \( e_i \) may be derived in this way. This might suggest the use of a probability plot using the unit exponential distribution in assessing the goodness-of-fit of the model. Cox and Oakes (1984, page 108) and Clayton and Cuzick (1985) propose ordering the \( e_i \) by size, and plotting against their expected values for the unit exponential distribution. This could be very misleading as Lagakos (1981) has shown that the \( e_i \) do not actually have unit exponential distribution even when \( \beta \) replaces \( \beta \) in the definition (5.04) and of \( e_i \). With this definition of \( e_i \), \( e_i \) is the expected values of \( H_0(t_i) \exp(z_i T \beta) \) conditional on the ranks of the \( T \)'s being equal to those of the observed \( t \)'s. But the "generalized residuals" here are estimated from the transformation of \( T \)'s, where \( T \)'s are random variables with integrated hazard function \( H(t, z_i), i = 1, 2, \ldots, n \). Clayton and Cuzick (1985a) exploit this relationship in the EM approach (see section 4.5.4).
Also Crowley and Storer (1983) stated that if the wrong model with $\beta = 0$ is incorrectly fitted to data and there are important covariates then, when there is no censoring, the value of the $j$th smallest $e_i$ is exactly equal to the mean of the $j$th order statistic from the unit exponential distribution. A perfect probability plot results for the wrong model! In a simulation experiment, Day (1985) found that when the correct model involved an indicator variable for group membership and a continuous covariate but either of the two simpler univariate regressions was fitted, better probability plots resulted for the incorrect models than for the correct model!

5.2.2 Schoenfeld's Residual.

This residual, due to Schoenfeld (1982), may be used to assess the time dependence of a covariate in the proportional hazards model. The general model for time dependency of covariates may be written as in (5.01). However, here we consider a particular covariate, say $x$, such that the hazard function may be written as

$$h(t/z) = h_0(t) \exp[z^T \beta + \gamma g(t)]$$

where $\gamma$ is the regressor coefficient for the $xg(t)$, and $x$ is a time constant explanatory variable, the $k$th component of $z$, say, and $g(t)$ a specified function of time. Then as in Section 4.4.1 we can compute the score statistic to test $\gamma = 0$. This statistic, equivalent to that of equation (5.03) is

$$\sum_{i=1}^{n} g(t_i) \delta_i [x_i - a_i(x)]$$

(5.08)
with
\[ a_i(x) = \sum_{j \in R_i} x_j \hat{\lambda}_j / \sum_{j \in R_i} \hat{\lambda}_j \]

where \( \hat{\lambda}_j \) is evaluated as before at \( \gamma = 0 \) and \( \beta = \hat{\beta} \), and \( x \) has
values \( x_j, j = 1, 2, \ldots, n \). If \( x \) is one of the components of \( z \)
then \( x_j - a_i(x) \) is given by the corresponding component of
\( z_i - \hat{a}_i \), with \( \hat{a}_i \) defined in Section 4.4.1. We also know that
\( U(\beta) \) is simply \( \sum \delta_i (z_i - a_i) \). We define the \( p \times 1 \) vector
\( \delta_i (z_i - a_i) \)
i = 1, 2, \ldots, n to be Schoenfeld's residual. Schoenfeld (1982)
considers, the covariate \( z_i = (z_{i1}, z_{i2}, \ldots, z_{ip})^T \) as random
variables, and defines the \( k \), \( k = 1, 2, \ldots, p \), component of \( a_i \)
as the conditional expectation of \( z_{ik} \) given the risk \( R_i \), and
probabilities proportional to \( \exp(z_j S) \), \( j \in R_i \). That is
\[ E(z_{ik} / R_i) = \sum_{j \in R_i} \frac{z_{jk} \exp(z_j S)}{\sum_{j \in R_i} \exp(z_j S)} \] (5.09)

and the residual at failure time \( t_i \) is
\[ \hat{r}_{ik} = z_{ik} - E(z_{ik} / R_i) \] (5.10)

where \( E(z_{ik} / R_i) \) is the value of (5.09) with \( \beta \) replaced by \( \hat{\beta} \).
If the proportional hazard holds \( E(r_{ik}) = 0 \), so plot \( \hat{r}_{ik} \) versus
failure time \( t_i \), will be centred about zero with random scatter.
However, if the hazard function is of type (5.07) Schoenfeld has
shown that
\[ E(r_{ik}) \approx g(t_i) \left\{ E(z_{ik}^2 / R_i) - \left[ E(z_{ik} / R_i) \right]^2 \right\} \]
where $A_i(g)$ is the $k^{th}$ diagonal element of the $p \times p$ matrix

$$ A_i(g) = \frac{\sum_{j \in R_i} z_j z_i^T \lambda_j}{\sum_{j \in R_i} \lambda_j} - \left( \sum_{j \in R_i} \frac{z_j \lambda_i}{\lambda_j} \right) \left( \sum_{j \in R_i} \frac{z_j \lambda_j}{\lambda_i} \right)^T $$

where $\lambda_j = \exp(z_j^T \beta)$. And in accordance with the definition of $E(z_{ik}/R_i)$ in (5.09), the $k^{th}$ diagonal component of the matrix is identical to

$$ \text{Var}(z_{ik}/R_i) = E(z_{ik}^2/R_i) - \left[ E(z_{ik}/R_i) \right]^2. $$

Since the variance term is always positive, the plot of $r_{ik}$ against $t_i$ may suggest the form of $g(t)$. Alternatively the plot of $\hat{r}_{ik}/A_i(g)$ may seem preferable, Kay (1984), as it may naturally reflect the functional form of $g(t)$. Note that if the $k^{th}$ element of $z$ has little effect as $t \to \infty$, then $g(t) \to -\beta_k$.

Schoenfeld (1982, Figure 1) gives a plot of the residuals $\delta_i(z_i - a_i)$ against time for a two-sample problem. Here $p = 1$ and $z_i$ is an indicator variable, with values of 0 and 1 to categorise the group an individual belongs to. The plots show two reasonably tight bands of residuals, one for each of the two groups. The bands are parallel about one unit apart with a positive slope. We illustrate this with the turbocharger data from British Railway, Derby (for description see Appendix B7a). There are 7 covariates included in the fit of the lifetime (in miles) of the turbocharged engine. These are denoted by variables
1, 2, ..., 7, all of which, except the 7th variable are indicator variables taking values of 0 and 1. There are 776 observations with almost 41% censored. The Newton-Raphson estimates for the Cox regression model when all the 7 covariates are fitted, are

\[
\begin{align*}
0.0364 & \quad (0.119) \\
-0.1152 & \quad (0.248) \\
0.0105 & \quad (0.112) \\
0.1580 & \quad (0.274) \\
-0.0500 & \quad (0.221) \\
-0.2629 & \quad (0.440) \\
-0.0391 & \quad (0.026)
\end{align*}
\]

with maximum loglikelihood -2675.852.

Figure 5.1 shows Schoenfeld residuals plot for this data set on the rank of the failure time for variable 1. Similar traits are observed, which is also shown on plots for variable 3, Figure 5.3. However, such a plot for variable 2, Figure 5.2, shows these bands are of different intensity with one band being more intense. Looking into the data pattern it is observed that the grouping (values of 1 or 0) is rather lopsided with too many zeros than one's. With quantitative covariate as variable 7, Figure 5.4, the plots show random scatter centred around zero, which suggest that there is little departure from the proportional hazards model. But the plots for variables 1, 2 and 3 show some time dependence for large failure time. Here the bands curve upwards and then downwards. This trait is also shown in a plot by Cain and Lange (1984) (their figure 1), for the Veterans Administration lung
Fig. 5.1  Plot of Schoenfeld's Residuals on ranks for Variable 1
BR data

Fig. 5.2  Plot of Schoenfeld's Residuals on ranks for Variable 2
BR data
Fig. 5.3  Plot of Schoenfeld's Residuals on ranks for Variable 3  
BR data

Fig. 5.4  Plot of Schoenfeld's Residuals on ranks for Variable 7  
BR data
cancer data. In all these cases the plots are difficult to
interpret since the relative intensity of points must be taken
into account. If g(t) is assumed to change slowly then a smoothed
plot of $\delta_i(z_i - \alpha_i)$ would be useful. This will be discussed in a
later chapter. Moreau, O'Quigley and Mesbah (1985) extend the
residual idea of Schoenfeld to a model where all the components
of $z$ are simultaneously considered and each corresponding g(t) is
taken to be a step-function with unknown steps over known time
intervals. The result is a test based on the corresponding score
statistic, involving the residuals $\delta_i(z_i - \alpha_i)$, having an asymptotic
chi-square distribution.

More explanation on Schoenfeld's residuals will be given in
Section 5.6, and Chapter 6.

5.3 CASE DELETION TECHNIQUES.

5.3.1 Introduction.

The plots described in the preceding sections are useful
tools which may throw some light on the general inadequacy of the
model, especially if there exists time trend in the covariates.
But having decided that $\beta$ is the correct linear predictor for
the proportional hazard model (4.07), we may proceed to investigate
whether there is any deviating data points, or that certain cases
may have great influence in the inference for $\beta$. These cases are
obviously important. In Chapter 3 we have presented some influence
assessment techniques. Here we shall extend them to apply to our
model in question. Basically the idea behind these influence
assessments is to introduce some small perturbations in the problem and then study how these perturbations change the outcome of the analysis. For further reference and detailed treatment on the subject, see Huber (1981), Cook and Weisberg (1982), Hampel (1986). The sample Influence curve of Cook and Weisberg (1982), has the natural appeal as the basis for diagnostic techniques that locate cases that are most influential. It is computed from the observed data, and for may problems the curve or the approximation thereof can be easily computed.

We recall again the definition of the influence curve for single case deletion, (equation 1.08).

\[ SIC_k = (n-1)(\hat{\beta} - \hat{\beta}_{(k)}) \]

where \( \hat{\beta}_{(k)} \) denotes the estimates of \( \beta \) based on sample size \( (n-1) \) with the \( k \)th case deleted. For standard least square regression, \( SIC_k \) can be computed exactly, Section (3.7), but for non-linear regression problems, such as the proportional hazards model, \( \hat{\beta}_{(k)} \) is found by iterative techniques. Thus for non-linear problems various approximations have been suggested, and in particular, for the proportional hazards model, the following: the case deletion location outlier model of Storer and Crowley (1985), the infinitesimal jackknife approach of Cain and Lange (1984) and the influence function approach of Reid and Crépeau (1985).

The fundamental purpose of these techniques is to find explicit formulae to approximate \( \hat{\beta} - \hat{\beta}_{(k)} \). Components of these approximations are then inspected to find the cases with large absolute values. As was discussed in Chapter 3, the problem with
this approach is that the p components have to be combined to obtain a scalar measure of influence of the kth case. One way is to consider the likelihood distance \( LD_k \), of (3.7.5), that is

\[
2 \left\{ L(\hat{\beta}) - L(\hat{\beta} - \Delta \hat{\beta}_k) \right\} 
\]

(5.12)

where \( \Delta \hat{\beta}_k \) is the p×1 vector approximation to \( \hat{\beta} - \hat{\beta}(k) \), and \( L(\cdot) \) is the log partial likelihood. These difference can be calibrated using the chi-squared distribution since a likelihood based confidence for \( \beta \) of size (1-\( \alpha \)) is given by

\[
\beta : 2 \left\{ L(\hat{\beta}) - L(\hat{\beta}) \right\} < \chi^2_{\alpha, p}
\]

where \( \chi^2_{\alpha, p} \) is the upper 100\( \alpha \) percent point of the chi-squared distribution with p degrees of freedom. As presented in Section 3.7, \( LD_k \) can be approximated (provided the log likelihood contours are approximately elliptical) by

\[
LD_k = \Delta \hat{\beta}^T A(\hat{\beta}) \Delta \hat{\beta}
\]

(5.13)

where \( A(\hat{\beta}) \) is the observed information matrix, and \( \Delta \hat{\beta}_k \) is as defined in equation (5.12).

A more general method of assessing influence, which was reviewed and extended by Cook (1986), introduces weights or a perturbation scheme \( w \) to give, in general, the likelihood function \( \ell(\hat{\beta}/w) \) for given \( w \). The log likelihood difference (5.12) can be generalized to \( LD(w) \), which Cook (1986) calls the likelihood displacement

\[
LD(w) = 2 \left[ L(\hat{\beta}) - L[\hat{\beta}(w)] \right]
\]

(5.14)
Cain and Lange (1984) considered the case-weight perturbation scheme

$$
\ell(\beta/w) = \prod_{i=1}^{n} \frac{\exp(z_i^T \hat{\beta} \delta_i \omega_i)}{\left\{ \sum_{j \in \mathcal{R}_i} w_j \exp(z_j^T \hat{\beta}) \right\}}.
$$

(5.15)

Here \( w \) is a vector of ones for the complete data set-up, and if \( w_k = 0 \) and the other \( w_j \)'s are 1, then \( \ell(\beta/w) \) corresponds to the partial likelihood with the \( k^{th} \) case omitted. We shall consider these perturbation schemes and their diagnostics in Section 5.5.

In the next section we shall discuss the case deletions schemes, and their approximations. Comparison is also made between the EM algorithm and the Newton-Raphson method.

5.3.2 The "Correct" 1-step Approximations.

The approach for case deletion used in Chapter 3 is now applied to the proportional hazards model, which is basically a non-linear regression problem. The simplest procedure is to obtain \( \hat{\beta}(k) \) that is, to delete the \( k^{th} \) case that is suspect from the data set and to recompute and analyse the model. This may give more accurate results, but the amount of computation involved is large, since we have to calculate the estimate for each \( k, k = 1,2,\ldots,n \). This may result in an enormous amount of computation, especially when iterative techniques are called for, as is this case. Our main concern is to be able to detect some disparity in the result of the analysis as compared to the complete data set, through the
changes in the estimates due to the omission of the case, and being able to draw the same conclusion from the approximation.

Using the perturbation scheme as in equation (5.15) we have $\ell(\beta|w_k = 0)$, equivalent to the partial likelihood with the kth case omitted from the data set. Our approximations to $\hat{\beta} - \hat{\beta}(k)$ involve 1-step of an iterative technique for maximizing $\ell(\beta)$ applied to $\ell(\beta|w_k = 0)$ starting at $\hat{\beta}$. As in the estimation of $\hat{\beta}$ in Chapter 4, we will consider two iterative techniques; The Newton-Raphson and the EM algorithm.

The Newton-Raphson iterative scheme is as in equation (4.28). For the ease of reference we shall record it here.

$$\beta_{s+1} = \beta_s + \left[A(\beta_s)\right]^{-1} U(\beta_s) \quad (5.16)$$

where $U(\beta_s)$ and $A(\beta_s)$ are in general, a score vector and the observed information matrix respectively, evaluated at $\beta_s$. Unfortunately for the likelihood function $\ell(\beta|w_k = 0)$, the observed information cannot be derived from $\ell(\beta)$ easily. In Section 5.4.1 we shall discuss an alternative approach suggested by Storer and Crowley (1985), so that the Newton-Raphson scheme can be applied. However, for the present, the approximation to $\hat{\beta}(k)$ is computed from (5.16) as the 1-step evaluation of $\beta_{s+1}$ from $\hat{\beta}$, the full data estimate, with the observed information $A(\hat{\beta})$ and the score $U(\hat{\beta})$ recomputed from (4.26) and (4.27) using $R_i^{(k)}$, the risk set for the ith observation when the kth case is omitted from the data set, instead of $R_i$. We shall call this the "correct" Newton-Raphson 1-step approximations. The computation is still lengthy as we still need to recompute for each observation
Table 5.1 Typical output for the Newton-Raphson method for estimating the regression coefficients when a case is omitted from the data set. The output gives the fully iterated estimates and the 1-step [correct] estimates applied on the carcinogenesis data:

<table>
<thead>
<tr>
<th>Number of observations</th>
<th>40</th>
<th>Number of failure times</th>
<th>36</th>
<th>Number of obs. censored</th>
<th>4</th>
</tr>
</thead>
</table>

**ESTIMATES of the beta are -0.5959**

The estimates with their S.E

<table>
<thead>
<tr>
<th>beta</th>
<th>S.E</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.5959</td>
<td>(0.3484)</td>
</tr>
</tbody>
</table>

The log likelihood for the NULL model : = -102.1583
The log likelihood at max. likelihood est: = -100.7130
The partial likelihood deviance : = 2.9786
The information matrix at this estimate is

<table>
<thead>
<tr>
<th>info matrix</th>
<th>0.823822E+01</th>
</tr>
</thead>
</table>

Invert is:

<table>
<thead>
<tr>
<th>Invert</th>
<th>0.121385E+00</th>
</tr>
</thead>
</table>

fully iterated Newton-Raphson - 1 case deletion

<table>
<thead>
<tr>
<th>off</th>
<th>obs</th>
<th>cen</th>
<th>(beta)k</th>
<th>diff</th>
<th>SICk</th>
<th>LDi</th>
<th>Chi</th>
<th>iter</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>142.</td>
<td>0</td>
<td>-0.67350</td>
<td>0.07761</td>
<td>3.02662</td>
<td>0.04986</td>
<td>0.04962</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>143.</td>
<td>0</td>
<td>-0.55162</td>
<td>-0.04428</td>
<td>1.72673</td>
<td>0.01638</td>
<td>0.01615</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>156.</td>
<td>0</td>
<td>-0.67051</td>
<td>0.07461</td>
<td>2.90969</td>
<td>0.04611</td>
<td>0.04586</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>163.</td>
<td>0</td>
<td>-0.67051</td>
<td>0.07461</td>
<td>2.90969</td>
<td>0.04611</td>
<td>0.04586</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>164.</td>
<td>0</td>
<td>-0.55811</td>
<td>-0.03779</td>
<td>1.47362</td>
<td>0.01199</td>
<td>0.01176</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>188.</td>
<td>0</td>
<td>-0.55990</td>
<td>-0.03599</td>
<td>1.40380</td>
<td>0.01089</td>
<td>0.01067</td>
<td>4</td>
</tr>
<tr>
<td>7</td>
<td>188.</td>
<td>0</td>
<td>-0.55990</td>
<td>-0.03599</td>
<td>1.40380</td>
<td>0.01089</td>
<td>0.01067</td>
<td>4</td>
</tr>
<tr>
<td>8</td>
<td>190.</td>
<td>0</td>
<td>-0.55797</td>
<td>-0.03793</td>
<td>1.47914</td>
<td>0.01208</td>
<td>0.01185</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>192.</td>
<td>0</td>
<td>-0.55797</td>
<td>-0.03793</td>
<td>1.47914</td>
<td>0.01208</td>
<td>0.01185</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td>198.</td>
<td>0</td>
<td>-0.65272</td>
<td>0.05683</td>
<td>2.21633</td>
<td>0.02684</td>
<td>0.02661</td>
<td>4</td>
</tr>
<tr>
<td>11</td>
<td>204.</td>
<td>1</td>
<td>-0.58024</td>
<td>-0.01566</td>
<td>-0.61070</td>
<td>0.00225</td>
<td>0.00202</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>205.</td>
<td>0</td>
<td>-0.65455</td>
<td>0.05865</td>
<td>2.28740</td>
<td>0.02858</td>
<td>0.02834</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>206.</td>
<td>0</td>
<td>-0.56840</td>
<td>-0.02749</td>
<td>-1.07215</td>
<td>0.00645</td>
<td>0.00623</td>
<td>4</td>
</tr>
<tr>
<td>14</td>
<td>209.</td>
<td>0</td>
<td>-0.56840</td>
<td>-0.02749</td>
<td>-1.07215</td>
<td>0.00645</td>
<td>0.00623</td>
<td>4</td>
</tr>
<tr>
<td>15</td>
<td>213.</td>
<td>0</td>
<td>-0.56840</td>
<td>-0.02749</td>
<td>-1.07215</td>
<td>0.00645</td>
<td>0.00623</td>
<td>4</td>
</tr>
<tr>
<td>16</td>
<td>218.</td>
<td>0</td>
<td>-0.56840</td>
<td>-0.02749</td>
<td>-1.07215</td>
<td>0.00645</td>
<td>0.00623</td>
<td>4</td>
</tr>
<tr>
<td>17</td>
<td>216.</td>
<td>1</td>
<td>-0.62702</td>
<td>0.03112</td>
<td>1.21378</td>
<td>0.00821</td>
<td>0.00798</td>
<td>4</td>
</tr>
<tr>
<td>18</td>
<td>220.</td>
<td>0</td>
<td>-0.56474</td>
<td>-0.03116</td>
<td>-1.21520</td>
<td>0.00823</td>
<td>0.00800</td>
<td>4</td>
</tr>
<tr>
<td>19</td>
<td>227.</td>
<td>0</td>
<td>-0.56474</td>
<td>-0.03116</td>
<td>-1.21520</td>
<td>0.00823</td>
<td>0.00800</td>
<td>4</td>
</tr>
<tr>
<td>20</td>
<td>230.</td>
<td>0</td>
<td>-0.56474</td>
<td>-0.03116</td>
<td>-1.21520</td>
<td>0.00823</td>
<td>0.00800</td>
<td>4</td>
</tr>
<tr>
<td>21</td>
<td>232.</td>
<td>0</td>
<td>-0.60828</td>
<td>0.01239</td>
<td>0.48315</td>
<td>0.00150</td>
<td>0.00126</td>
<td>3</td>
</tr>
<tr>
<td>22</td>
<td>232.</td>
<td>0</td>
<td>-0.60828</td>
<td>0.01239</td>
<td>0.48315</td>
<td>0.00150</td>
<td>0.00126</td>
<td>3</td>
</tr>
<tr>
<td>23</td>
<td>233.</td>
<td>0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28362</td>
<td>0.00067</td>
<td>0.00044</td>
<td>3</td>
</tr>
<tr>
<td>24</td>
<td>233.</td>
<td>0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28362</td>
<td>0.00067</td>
<td>0.00044</td>
<td>3</td>
</tr>
<tr>
<td>25</td>
<td>233.</td>
<td>0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28362</td>
<td>0.00067</td>
<td>0.00044</td>
<td>3</td>
</tr>
<tr>
<td>26</td>
<td>233.</td>
<td>0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28362</td>
<td>0.00067</td>
<td>0.00044</td>
<td>3</td>
</tr>
<tr>
<td>27</td>
<td>234.</td>
<td>0</td>
<td>-0.61950</td>
<td>0.02860</td>
<td>0.92040</td>
<td>0.00482</td>
<td>0.00459</td>
<td>4</td>
</tr>
<tr>
<td>28</td>
<td>239.</td>
<td>0</td>
<td>-0.60147</td>
<td>0.05558</td>
<td>0.21756</td>
<td>0.00048</td>
<td>0.00026</td>
<td>3</td>
</tr>
<tr>
<td>1-step [correct] Newton-Raphson - 1 case deletion</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------------------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>off</td>
<td>obs</td>
<td>cen (beta)k</td>
<td>diff</td>
<td>SICK</td>
<td>LDi</td>
<td>Chi</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------</td>
<td>------</td>
<td>--------------</td>
<td>------</td>
<td>-------</td>
<td>------</td>
<td>-------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>142.0</td>
<td>0.087355</td>
<td>0.07766</td>
<td>0.02856</td>
<td>0.0492</td>
<td>0.04943</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>143.0</td>
<td>-0.55165</td>
<td>-0.04424</td>
<td>-0.72539</td>
<td>0.01635</td>
<td>0.01604</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>156.0</td>
<td>0.07055</td>
<td>0.07466</td>
<td>2.91161</td>
<td>0.04616</td>
<td>0.04569</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>163.0</td>
<td>0.07055</td>
<td>0.07466</td>
<td>2.91161</td>
<td>0.04616</td>
<td>0.04569</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>164.0</td>
<td>-0.55814</td>
<td>-0.03776</td>
<td>-1.47260</td>
<td>0.01198</td>
<td>0.01169</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>188.0</td>
<td>0.05992</td>
<td>-0.03597</td>
<td>-1.40266</td>
<td>0.01089</td>
<td>0.01061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>188.0</td>
<td>0.05992</td>
<td>-0.03597</td>
<td>-1.40266</td>
<td>0.01089</td>
<td>0.01061</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>190.0</td>
<td>0.05600</td>
<td>0.03790</td>
<td>1.47811</td>
<td>0.01206</td>
<td>0.01177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>192.0</td>
<td>0.05600</td>
<td>0.03790</td>
<td>1.47811</td>
<td>0.01206</td>
<td>0.01177</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>198.0</td>
<td>0.065274</td>
<td>0.06585</td>
<td>2.21702</td>
<td>0.02866</td>
<td>0.02849</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>204.0</td>
<td>0.05824</td>
<td>0.01566</td>
<td>0.61062</td>
<td>0.00225</td>
<td>0.00211</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>205.0</td>
<td>0.05457</td>
<td>0.05867</td>
<td>2.28817</td>
<td>0.02860</td>
<td>0.02822</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>206.0</td>
<td>0.05642</td>
<td>-0.02748</td>
<td>-1.07157</td>
<td>0.00645</td>
<td>0.00619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>209.0</td>
<td>0.05642</td>
<td>-0.02748</td>
<td>-1.07157</td>
<td>0.00645</td>
<td>0.00619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>213.0</td>
<td>0.05642</td>
<td>-0.02748</td>
<td>-1.07157</td>
<td>0.00645</td>
<td>0.00619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>216.0</td>
<td>0.05642</td>
<td>-0.02748</td>
<td>-1.07157</td>
<td>0.00645</td>
<td>0.00619</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>216.0</td>
<td>0.062703</td>
<td>0.03114</td>
<td>1.21432</td>
<td>0.00822</td>
<td>0.00795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>220.0</td>
<td>0.056476</td>
<td>-0.03114</td>
<td>-1.21447</td>
<td>0.00822</td>
<td>0.00795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>227.0</td>
<td>0.056476</td>
<td>-0.03114</td>
<td>-1.21447</td>
<td>0.00822</td>
<td>0.00795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>230.0</td>
<td>0.056476</td>
<td>-0.03114</td>
<td>-1.21447</td>
<td>0.00822</td>
<td>0.00795</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>232.0</td>
<td>-0.60828</td>
<td>0.01239</td>
<td>0.48314</td>
<td>0.00150</td>
<td>0.00126</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>232.0</td>
<td>0.06082</td>
<td>0.01239</td>
<td>0.48314</td>
<td>0.00150</td>
<td>0.00126</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>233.0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28361</td>
<td>0.00067</td>
<td>0.00043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>233.0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28361</td>
<td>0.00067</td>
<td>0.00043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>233.0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28361</td>
<td>0.00067</td>
<td>0.00043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>233.0</td>
<td>-0.60317</td>
<td>0.00727</td>
<td>0.28361</td>
<td>0.00067</td>
<td>0.00043</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>234.0</td>
<td>-0.61951</td>
<td>0.02362</td>
<td>0.92099</td>
<td>0.00483</td>
<td>0.00457</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>239.0</td>
<td>-0.60147</td>
<td>0.00558</td>
<td>0.21756</td>
<td>0.00049</td>
<td>0.00026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>240.0</td>
<td>-0.60147</td>
<td>0.00558</td>
<td>0.21756</td>
<td>0.00049</td>
<td>0.00026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>244.0</td>
<td>-0.70361</td>
<td>0.10792</td>
<td>4.20872</td>
<td>0.09620</td>
<td>0.09546</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>246.0</td>
<td>-0.63615</td>
<td>0.04025</td>
<td>1.56994</td>
<td>0.01359</td>
<td>0.01328</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>261.0</td>
<td>-0.57571</td>
<td>-0.02019</td>
<td>-0.78732</td>
<td>0.00359</td>
<td>0.00334</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>265.0</td>
<td>-0.65830</td>
<td>0.06241</td>
<td>2.43394</td>
<td>0.03232</td>
<td>0.03193</td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>280.0</td>
<td>-0.54838</td>
<td>-0.04752</td>
<td>-1.85324</td>
<td>0.01882</td>
<td>0.01851</td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>280.0</td>
<td>-0.54838</td>
<td>-0.04752</td>
<td>-1.85324</td>
<td>0.01882</td>
<td>0.01851</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>296.0</td>
<td>-0.54603</td>
<td>-0.04987</td>
<td>-1.94488</td>
<td>0.02071</td>
<td>0.02039</td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>296.0</td>
<td>-0.54603</td>
<td>-0.04987</td>
<td>-1.94488</td>
<td>0.02071</td>
<td>0.02039</td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>304.0</td>
<td>-0.85644</td>
<td>0.26055</td>
<td>10.16141</td>
<td>0.55912</td>
<td>0.55647</td>
<td></td>
<td></td>
</tr>
<tr>
<td>39</td>
<td>323.0</td>
<td>-0.48149</td>
<td>-0.11441</td>
<td>-4.46180</td>
<td>0.10793</td>
<td>0.10729</td>
<td></td>
<td></td>
</tr>
<tr>
<td>40</td>
<td>344.0</td>
<td>-0.48149</td>
<td>-0.11441</td>
<td>-4.46180</td>
<td>0.10793</td>
<td>0.10729</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
omitted, those matrices and score vectors for each i in the \((n-1)\) observations. Table 5.1 gives a general output for both 1-step and full-step Newton Raphson iteration applied to Carcinogenesis data (Appendix B4). There is little difference in all the statistics given out here. Figure 5.5 illustrates this further, for the heart transplant data model III, (see Table 4.4). Plotting the likelihood distances \((5.12)\) for the 1-step approximations against the likelihood distance computed from the fully iterated estimates of \(\hat{\beta}_{(k)}\), produce an almost 45° plot, showing that the approximations can represent the correct statistics quite efficiently.

Similarly we can adapt the EM algorithm described in Chapter 4 to compute the "correct" 1-step EM. This may be achieved by recomputing all the details evaluated at \(\hat{\beta} = \hat{\beta}_{(k)}\), with the \(k\)th case omitted and carry out a 1-step computation. Here again, generally the computations is still lengthy since they are computed "fresh" and not derived from \(\ell(\beta/\omega)\), (see equation 5.15), of the full data set. A good approximation to estimating \(\hat{\beta}_{(k)}\), the 1-step estimate, with economical computation is of great value in our assessment of influential cases. In the succeeding sections we shall discuss various methods of estimating \(\hat{\beta}_{(k)}\) or \(\hat{\beta}_{(k)}^{1}\) to be used in the influence diagnostics of the Cox proportional hazards model.

5.3.3 Illustration of the "Correct" 1-step Newton-Raphson.

In this section we shall illustrate by example the influence assessment as computed from the estimates obtained from the 1-step "correct" Newton-Raphson method, as described in the above section.
Fig. 5.5  Plot of the approximate 1-step Likelihood Distances, $LD^1$, computed from the N-R methods on $LD$ for full iterations for the heart transplant data model III.

Note: $LD'^1$ is computed from augmented model (Section 5.4)
The algorithm is implemented on the Stanford heart transplant data (Appendix B6) under model III (see Table 4.4). Fortran program in Appendix C3 can easily be adapted to carry out the computation. Figure 5.6 gives the plots of the influences. These plots are all against the rank of the observations that are deleted from the data set. Plot 1 gives the influence of the likelihood distance as computed from (5.13). It shows cases ranked 2 and 139 have greatest influence as indicated by their high peaks. Not much information is obtained from plot 2 which gives measures that are proportional to sample influence curve $SIC_k$ for variable 1, the age at transplant. In plot 3, for the T5 mismatch score, cases ranked 1 and 82 appear to be influential (in opposite direction); but regression coefficient for this variable is uncorrelated with the other $\hat{\beta}$'s, and on further analysis (see Section 5.4.1) their removal individually do not influence the T5 regression parameter by any significant amount. Plot 4, however, indicates cases ranked 2 and 139 as suspects. This variable which is the squared values of variable 1 (the age) is significant in the model (see 4.6.3), unlike the T5. It is also correlated with variable 1 with correlation coefficient about 0.5. Even though plot 2 does not give any impression of any influential case, but by virtue of its correlation with variable 3, the influential cases are mirrored in the plots for variable 3. This conforms to the plot of the log likelihood distance in plot 1. In Section 5.4.1 the same cases are picked out by measures computed from other approximations of $\hat{\beta}_{(k)}$, namely by the 1-step augmented model (Section 5.4.1).
Fig. 5.6 Plots of influences on rank of deleted cases for the heart transplant data model III

Plot 1

Plot 2

Plot 3

Plot 4
5.3.4 **EM 1-step Approximation.**

As opposed to the "correct" 1-step EM discussed earlier, this scheme makes use of the statistics computed at \( \hat{\beta} \), and extend the computation to yet another step, taking account to adapt to cases that are deleted from the data set. The EM approach best suited for this approximation is the EM with gamma error (Section 4.5.5) where the resulting iterative weighted least squares scheme for the complete data likelihood (equation (4.56)) is:

\[
\hat{\beta}_{s+1} = \hat{\beta}_s + (Z^T Z)^{-1} Z^T d
\]

where \( d \) is defined as in equation (4.57) and has components identical to the Lagakos's residuals \( e^\ast \) discussed in Section 5.1.1.

The covariate matrix \( Z \) are column centred, that is \( \sum_{i=1}^{n} z_{ij} = 0 \) for every \( j = 1,2,\ldots,p \). The 1-step EM, (strictly speaking it is the 1-step of the M process with fixed \( e^\ast \)), estimate for the \( k \)th case deleted, \( \hat{\beta}^{1}_{(k)} \), starting from \( \hat{\beta} = \hat{\beta} \) is

\[
\hat{\beta}^{1}_{(k)} = \hat{\beta} + \left[ Z^T (Z - e^\ast) \right]^{-1} Z^T (Z - e^\ast) \]

\[\text{(5.17)}\]

where \( Z_{(k)} \) is the covariate matrix obtained from \( Z \) with the \( k \)th row omitted, and centred; and \( e^\ast_{(k)} \) is the \((n-1)\) vector of \( e^\ast \) with the \( k \)th case omitted. We use the symbol \( e^\ast \) here instead of \( d \), to be consistent with the concept of residuals, as discussed in (5.1.1). From (5.17) the 1-step approximation to the sample influence curve is found to be

\[
\text{SIC}_k \approx (n-1)(\hat{\beta} - \hat{\beta}^{1}_{(k)})
\]

\[\text{(5.18)}\]
where \( v_{kk} = z_k^T (Z^T Z)^{-1} z_k \), and \( z_k \) is the \( p \) vector corresponding to the \( k^{th} \) row of \( Z \) and \( e_k \) is the \( k^{th} \) element of \( e \). Equation (5.18) is obtained from using standard results for least squares regression, (Cook and Weisberg, 1982, Section 3.4).

Suppose \( Z \) is a \((n \times p)\) matrix whose columns are centred, and suppose \( Z(k) \) is \((n-1) \times p\) matrix from \( Z \) with row omitted, and \( Z(k) \) is also column centred, then

\[
Z = \begin{bmatrix}
Z(k) \\
\ddots \\
\vdots \\
Z_{k}
\end{bmatrix} - \frac{1}{n-1} \begin{bmatrix} 1 \\
\vdots \\
0
\end{bmatrix} z_k^T \tag{5.19}
\]

where \( 1_{n-1} \) is a \((n-1)\) vector of one, and \( z_k^T = (z_{k1}, z_{k2}, \ldots, z_{kp}) \).

Then

\[
Z^T Z = Z^{T(k)} Z^{(k)} + z_k^T z_k - \frac{1}{n-1} \begin{bmatrix} z_k^T \\
\vdots \\
0
\end{bmatrix} \begin{bmatrix} 1 \\
\vdots \\
0
\end{bmatrix} z_k^T
\]

\[- \frac{1}{n-1} z_k \begin{bmatrix} 1 \\\n\vdots \\
z_k
\end{bmatrix} \begin{bmatrix} 1 \\
\vdots \\
0
\end{bmatrix} \begin{bmatrix} Z(k) \\
\vdots \\
Z_k
\end{bmatrix}
\]

\[+ \frac{1}{(n-1)^2} z_k \begin{bmatrix} 1 \\
\vdots \\
z_k
\end{bmatrix} \begin{bmatrix} 1 \\
\vdots \\
0
\end{bmatrix} \begin{bmatrix} 1 \\
\vdots \\
0
\end{bmatrix} z_k^T,
\]

the third and fourth terms of the R.H.S. are zeros, since \( Z(k) \) is column centred, thus the sum of the columns are zero. Hence

\[
Z^T Z = Z^{T(k)} Z^{(k)} + z_k^T z_k - \frac{1}{n-1} z_k z_k^T
\]

\[= Z^{T(k)} Z^{(k)} + \frac{1}{n-1} z_k z_k^T .
\]
Applying the lemmas for matrix simplification (see for example, Mardia, Kent & Bibby, 1979, Appendix A.2.4), stated in Appendix A1

\[
\begin{bmatrix}
Z(k) \\
(k)
\end{bmatrix}^{-1} = 
\begin{bmatrix}
Z_T - \left( \frac{n}{n-1} \right) z_k z_k^T
\end{bmatrix}^{-1}
\]

\[
= (Z_T)^{-1} + \left( \begin{array}{c}
(z_T)^{-1} z_k z_k^T (Z_T)^{-1} \left( \frac{n}{n-1} \right)
\end{array} \right) \left( \begin{array}{c}
1 - \left( \frac{n}{n-1} \right) z_k (Z_T)^{-1} z_k
\end{array} \right) .
\] (5.20)

Similarly partitioning the n-vector \( e^* \) such that \( e^T = (e^T_{(k)}, e^*_k) \) we have

\[
Z e^*_k = \begin{bmatrix}
Z(k) \\
\end{bmatrix} \begin{bmatrix}
e^*_{(k)} \\
e^*_k
\end{bmatrix} - \frac{1}{n-1} z_k \begin{bmatrix} 1 \\
\cdots \\
0
\end{bmatrix} \begin{bmatrix} e^*_{(k)} \\
\cdots \\
e^*_k
\end{bmatrix}
\]

\[
= z^T(k) e^*_{(k)} + z_k e^* - \frac{\alpha_k}{n-1} z_k
\]

\[
= z^T(k) e^*_{(k)} + z_k \left( e^* - \frac{\alpha_k}{n-1} \right)
\]

where \( \alpha_k = \sum_{i \neq k} e^*_i \). But with \( e^* \) defined as in (4.57)

\[
\sum_{i=1}^{n} e^*_i = 0. \text{ Then}
\]

\[
\alpha_k = - e^*_k .
\]

Hence

\[
Z e^* = Z(k) e^*_{(k)} + \left( \frac{n}{n-1} \right) z_k e^*_k
\]

Therefore

\[
Z(k) e^*_{(k)} = Z e^* - \left( \frac{n}{n-1} \right) z_k e^*_k .
\] (5.21)
Now at $\hat{\beta} = \bar{\beta}$, $Z^T e^* = 0$.

Then (5.17) can now be written as

$$\hat{b}_k = \bar{\beta} - \left[ (Z^T Z)^{-1} + \frac{n}{(n-1)} \frac{z_k z_k^T (Z^T Z)^{-1}}{1 - \frac{n}{(n-1)} z_k^T (Z^T Z)^{-1} z_k} \right] \left[ - \frac{n^*}{n-1} z_k e_k^* \right]$$

$$= \bar{\beta} - \left[ \frac{n}{n-1} \right] (Z^T Z)^{-1} z_k e_k^* \left[ 1 - n(n-1)^{-1} v_{kk} \right]^{-1}$$

where $v_{kk} = z_k^T (Z^T Z)^{-1} z_k$. Equation (5.18) then follows.

This result is similar to those found for case deletion in least squares regression problem. Note here that the $e^*_{(k)}$ is the $(n-1)$ subvector of $e^*$ with the $k^{th}$ component omitted, and that there is no readjustment in the individual components $i$ under the new risk set $R_{(k)}$. Despite this, we found that (5.18) gives good approximation to the $SIC_k$, especially with smaller data sets, sample sizes about 40. However, for larger data set, say, the 152 of the Stanford Heart transplant data, this approximation under-estimates the $SIC_k$ by about 40 to 50 per cent, but nevertheless gives a good qualitative impression of influential cases as demonstrated in Figure 5.7. These plots are the equivalent of that of Figure 5.6, but the quantities are computed using the technique in discussion.

5.4 AUGMENTED LINEAR PREDICTOR APPROXIMATION.

To enable the 1-step Newton-Raphson approximation to $\bar{\beta} - \hat{\beta}_{(k)}$ to be easily computed, Storer and Crowley (1985) augmented the
Fig. 5.7 Plots of influences on rank of deleted cases for the heart transplant data model III — by EM

Plot 1

Plot 2

Plot 3

Plot 4
linear predictor $z^T \beta$ for the case in question only. Defining the $n \times 1$ vector $\eta$ as the vector of linear predictors for $n$ cases, this augmented model (or sometimes called outlier model) can be written as

$$\eta = Z \beta_k^* + d_k \gamma$$

(5.23)

where $\beta_k^*$ replaces $\beta$ and $d_k$ is an $n \times 1$ vector of zeros except the $k$th element which has unit value. Denoting the partial likelihood for the proportional hazard model as $\ell(\beta_k^*, \gamma)$, for $k$th case censored we have

$$\ell(\beta_k^*, \gamma) = \frac{n}{\prod_{i \neq k} \prod_{i = 1}^{n} \exp(z_i^T \beta_k^* \delta_i)}$$

$$\prod_{i \in G} \left[ \sum_{j \in R_i} \exp(z_i^T \beta_k^* + d_{kj} \gamma) \right]^{\delta_i}$$

(5.24)

where $d_{kj}$ is the $j$th component of $d_k$.

The denominator can be split up into the product over two groups, $G_1$ and $G_2$, where $G_1$ contains all the risk sets that have $k$th observation, and $G_2$ is without the $k$th observation.

Let

$$b_{ik} = \exp(z_i^T \beta_k^*)$$

then equation (5.24) can be written as

$$\ell(\beta_k^*, \gamma) = \frac{\prod_{i \neq k}^{n} (b_{ik})^{\delta_i}}{\prod_{i \in G_1} \left[ b_{ik} \exp(\gamma) + \sum_{b_i \in G_1 \setminus \{k\}} b_{ik} \right]^{\delta_i}}$$

$$\prod_{i \in G_2, \{k\}} \left[ \sum_{b_i \in G_2 \setminus \{k\}} b_{ik} \right]^{\delta_i}$$

(5.25)
For fixed $\beta_k^*$, $\ell(\beta_k^*, \gamma)$ is maximized with respect to $\gamma$ when the term $b_{ik} \exp(\gamma) + 0$; this is equivalent to letting $\gamma + - \infty$.

Then

$$\max_{\gamma} \ell(\beta_k^*, \gamma) = \frac{\prod_{i \neq k} (b_{ik})^\delta_i}{\prod_{i \in G_1} \left( \sum_{i \in R_i \setminus \{k\}} b_{ik} \right)^{\delta_i} \prod_{j \in G_2} \left( \sum_{i \in R_j} b_{jk} \right)^{\delta_j}}$$

$$= \frac{\exp(z_i^T \beta_k^*)^{\delta_i}}{\prod_{i \neq k} \left( \sum_{i \in R_i} \exp(z_i^T \beta_k^*) \right)^{\delta_i}}$$

which is identical to $\ell(\beta_k^*/\gamma_k = 0)$ (see equation (5.15)), the likelihood of the model with the $k$th case omitted. Hence the maximum likelihood estimate of $\beta_k^*$, $\hat{\beta}_k$, is given by $\hat{\beta}_{(k)}$.

For uncensored cases, in general the maximum likelihood estimate, $\hat{\gamma}$, of $\gamma$ satisfies $- \infty < \hat{\gamma} < \infty$ and $\hat{\gamma}$ depends on $\hat{\beta}_k^*$ when maximization of $\ell(\beta_k^*, \gamma)$ as a function of $\gamma$ only is considered. Here $\hat{\beta}_k^* \neq \hat{\beta}_{(k)}$. The partial likelihood, $\ell(\beta_k^*, \gamma)$ when the $k$th case is a failure (from 5.24) is then

$$\ell(\beta_k^*, \gamma) = \frac{n}{\prod_{i = 1} (b_{ik})^{\delta_i}} \prod_{i \neq k} \left( \sum_{j \in R_i} b_{jk} \exp(d_{kj} \gamma) \right)^{\delta_i}$$
where \( Q = \sum_{j \in R_k \setminus \{k\}} \exp(z_j^T \beta_k^*) \), and \( \ell_k(\beta_k^*, \gamma) \)

is the partial likelihood of (5.24), without the \( k \)th case

likelihood. The first term of the R.H.S. of (5.26), involving

only the \( k \)th case in the numerator and risk set \( R_k \) in the
denominator, may be modified such that

\[
\ell_k(\beta_k^*, \gamma) = \frac{\exp(z_k^T \beta_k^* + \gamma)}{[\exp(z_k^T \beta_k^* + \gamma) + Q]}
\]

is maximized by letting \( \gamma \rightarrow \infty \) (as before

when the augmented case is censored). Then \( \ell_k(\beta_k^*, \gamma, \alpha) \)

is maximized for every \( \beta_k^* \) and \( \gamma \) by letting \( \alpha \rightarrow +\infty \). It is clearly seen from

(5.27) that \( g_k(\beta_k^*, \gamma, \alpha) \) is a monotonic increasing function with

maximum value equal to 1 which occurs when \( \alpha \rightarrow +\infty \). We then find
from (5.27) that

\[
\lim_{\gamma \to -\infty, \alpha \to +\infty} \ell(\beta^*_k, \gamma, \alpha) = \ell(\beta^*_k, \omega_k = 0)
\]

The approach here is to approximate \( \hat{\beta}^{(k)} \) by using a 1-step approximations of \( \hat{\beta}^*_k \) from \( \ell(\beta^*_k, \gamma) \) or \( \ell(\beta^*_k, \gamma, \alpha) \) starting at \( \hat{\beta}^*_k = \beta \), \( \gamma = 0 \) and \( \alpha = 0 \). In the following sections we consider two approximations, one based on the Newton-Raphson, (5.16) and the other on the EM scheme (5.17).

5.4.1 Newton-Raphson 1-step for the Augmented Model.

The score vector of (4.25) can be written in terms of the Lagakos residuals, \( e^*_k \) (Section 5.2.1), such that:

\[
U(\hat{\beta}) = Z^T e^*_k
\]

where \( Z \) is the nxp design matrix of covariates. We shall write \( e^*_k \) here again for convenience of reference.

\[
e^*_i = \delta_i - \lambda_i \left[ \sum_{t_j \leq t_i} \hat{\lambda}_i + \sum_{r \in R} \hat{\lambda}_r \right]
\]

If the suspect case is censored, then we may use the augmented model (5.23). The design matrix is expanded thus:
Then the score is

\[ U(\hat{\beta}_k^*, \gamma) = \left[ \begin{array}{c} U(\hat{\beta}_k^*) \\ U(\gamma) \end{array} \right] = Z_+^T e^* \]

where \( e_k^* \) is the \( k \)th component of \( e^* \). Note here that the \( \lambda \)'s in the expression of \( e^* \) in the augmented model contains \( d_{ik} \gamma \) term, that is:

\[ \lambda_i = \exp(z_i^T \beta_k^* + d_{ik} \gamma) \quad (5.31) \]

where \( d_{ik} \) is the \( i \)th component of the \( n \) vector \( d_k \).

The 1-step Newton-Raphson approximation, \( \hat{\beta}_{(k)}^1 \), to \( \hat{\beta}_k^* \) and thus \( \hat{\beta}_{(k)} \), is the next iterate of (5.16), starting at \( \hat{\beta}_k^* = \hat{\beta} \), the maximum likelihood estimate for the full data set, and \( \gamma = 0 \). Thus:

\[ (\hat{\beta}_k^1, \Delta \gamma)^T = (\hat{\beta}, 0)^T + \left[ A(\hat{\beta}_k^* \gamma)^T \right]^{-1} U(\hat{\beta}_k^*, \gamma) \quad (5.32) \]

where \( A(\hat{\beta}_k^* \gamma) \) is the \( (p+1) \times (p+1) \) observed information matrix.
for the augmented model. \( A(\beta_k^*, \gamma) \) may be partitioned to correspond to the partitioning of \( U(\beta_k^*, \gamma) \), in the following way.

The log of the partial likelihood for the augmented model is

\[
L(\beta_k^*, \gamma) = \sum_{i=1}^{n} \delta_i \left[ (z_i^T \beta_k^* + \gamma d_{ik}) - \log \sum_{j \in R_i} \lambda_j \right]
\]

then

\[
\begin{bmatrix}
\frac{\partial L}{\partial \beta_k^*} \\
\frac{\partial L}{\partial \gamma}
\end{bmatrix} = \begin{bmatrix}
U(\beta_k^*) \\
U(\gamma)
\end{bmatrix} = \begin{bmatrix}
\sum_{i=1}^{n} \delta_i \left\{ \sum_{j \in R_i} \lambda_j \right\} \\
\sum_{i=1}^{n} \delta_i \left\{ d_{ik} \sum_{j \in R_i} \lambda_j \right\}
\end{bmatrix}.
\]

The component

\[
U(\gamma) = \delta_k - \lambda_k \sum_{t_i \leq t_k} \sum_{j \in R_i} \delta_i
\]

since

\[
d_{ik} = \begin{cases} 0 & i \neq k \\ 1 & i = k \end{cases}
\]

and for \( t_i > t_k, \ d_{jk} = 0, \ \forall j \in R_i \).

Now

\[
\frac{\partial^2 L}{\partial \beta_k^* \beta_k^* T} = - \sum_{i=1}^{n} \delta_i \begin{bmatrix}
\sum_{j \in R_i} \lambda_j \\
\sum_{j \in R_i} z_j T z_j \lambda_j \\
\sum_{j \in R_i} \lambda_j
\end{bmatrix}
\]

\[
= - A(\beta_k^*)
\]
as in (4.27), the negative of the $p \times p$ information matrix for the $\beta_k^*$:

$$
\frac{\partial^2 L}{\partial \gamma^2} = -\sum_{i=1}^{n} \delta_i \left[ \sum_{j \in R_i} z_i d_{jk} \lambda_j \right] - \frac{1}{\lambda} \left[ \sum_{j \in R_i} z_j \lambda_j \right] \left[ \sum_{j \in R_i} \lambda_j \right]$$

$$
\left[ \sum_{j \in R_i} \lambda_j \right] \left[ \sum_{j \in R_i} \lambda_j \right] \left[ \sum_{j \in R_i} \lambda_j \right]$$

$$
= - \lambda_k z_k \left[ \sum_{i \leq t_k} \sum_{j \in R_i} \lambda_i \right] + \lambda_k \left[ \sum_{i \leq t_k} \sum_{j \in R_i} \lambda_j \right]
$$

Define residuals (which we shall use in later sections)

$$
\Lambda_{k, 2} = \lambda_k \left[ \sum_{i \leq t_k} \sum_{j \in R_i} \lambda_i \right]
$$

and

$$
\Lambda_{k, 3} = - \lambda_k z_k \left[ \sum_{i \leq t_j} \sum_{j \in R_i} \lambda_j \right]
$$

where $\lambda_j$ are $\lambda_j$ of (5.31) evaluated at $\beta_k^* = \hat{\beta}$ and $\gamma = 0$. Then we have a $p \times 1$ vector

$$
\left( \frac{\partial^2 L}{\partial \gamma^2} \right)_{\beta_k^* = \hat{\beta}, \gamma = 0} = (\Lambda_{k, 2} + \Lambda_{k, 3}).
$$

From (5.33)

$$
\frac{\partial^2 L}{\partial \gamma^2} = - \left[ \lambda_k \left[ \sum_{i \leq t_k} \left( \sum_{j \in R_i} \lambda_j \right) \right] - \lambda_k \left[ \sum_{i \leq t_k} \delta_i \left( \sum_{j \in R_i} \lambda_j \right)^2 \right] \right]
$$
Thus

\[
\begin{bmatrix}
\frac{\partial^2 L}{\partial \gamma^2}
\end{bmatrix}_{\hat{\beta}, \gamma = 0} = -e_k + \hat{\lambda}_k^2 \sum_{t_i \leq t_k} \frac{\delta_i}{\left( \sum_{j \in R_i} \lambda_{ij} \right)^2}
\]

where \(e_k\) is defined in Section 5.2.1.

It can be shown that

\[
\begin{bmatrix}
\frac{\partial^2 L}{\partial \beta_k \partial \gamma}
\end{bmatrix} = (\Delta_{k,2} + \Delta_{k,3})^T
\]
evaluated at \(\hat{\beta}_k^* = \hat{\beta}, \gamma = 0\).

Then the observed information matrix evaluated at \(\hat{\beta}_k^* = \hat{\beta}\) and \(\gamma = 0\) is

\[
A(\hat{\beta}_k^*, \gamma) = \begin{bmatrix}
A(\hat{\beta})_{(p \times p)} & -\left(\Delta_{k,2} + \Delta_{k,3}\right)_{(p \times 1)} \\
-(\Delta_{k,2} + \Delta_{k,3})^T_{(1 \times p)} & c_k
\end{bmatrix}
\]

(5.37)

where the scalar \(c_k\) is

\[
c_k = e_k - \hat{\lambda}_k^2 \sum_{t_i \leq t_k} \frac{\delta_i}{\left( \sum_{j \in R_i} \lambda_{ij} \right)^2}
\]

(5.38)

The score (5.30) at \(\hat{\beta}_k^* = \hat{\beta}, \gamma = 0\) is

\[
U(\hat{\beta}_k^*, \gamma) = \begin{bmatrix}
\mathbf{0}_p \\
e_k
\end{bmatrix}
\]

since \(Z^T \mathbf{e}_k^* = 0\) at the maximum likelihood estimate and \(\mathbf{0}_p\) is a \(p\) vector of zeros. Let the inverse of \(A(\hat{\beta}_k^*, \gamma)\) be
The 1-step Newton-Raphson, from (5.32) is

\[
\begin{pmatrix}
\hat{\beta}_k \\
\Delta \gamma
\end{pmatrix} = \begin{pmatrix}
\begin{array}{c}
\hat{\beta} \\
0
\end{array}
\end{pmatrix} + \begin{pmatrix}
\hat{b} \\
\hat{c}
\end{pmatrix} e_k^*.
\]  

(5.40)

The 1-step Newton-Raphson for estimating \( \beta_k^* \) from \( \beta_k = \hat{\beta} \) is thus

\[
\beta_k^1 = \hat{\beta} + B_{12} e_k^*
\]  

(5.41)

while \( b e_k^* \) gives the change in the augmented parameter \( \gamma \) from zero. To evaluate (5.40) we only require \( p \times 1 \) matrix \( B_{12} \).

Applying the usual lemmas for matrix simplification, (see Appendix A1) to matrix (5.37), and letting \( q = (A_{k,2} + A_{k,3}) \), we have

\[
B_{11} = \left[ A(\hat{\beta}) - \frac{q q^T}{c_k} \right]^{-1}
\]

\[
= A(\hat{\beta})^{-1} + \frac{A(\hat{\beta})^{-1} q q^T A(\hat{\beta})^{-1}}{c_k - q^T A(\hat{\beta})^{-1} q}.
\]

Now

\[
B_{12} = - B_{11} \begin{pmatrix}
-q \\
\frac{c_k}{c_k}
\end{pmatrix}
\]

\[
= A(\hat{\beta})^{-1} \left[ q + \frac{q q^T A(\hat{\beta})^{-1} q}{c_k - q^T A(\hat{\beta})^{-1} q} \right] \frac{1}{c_k}
\]

\[
= A(\hat{\beta})^{-1} q/(c_k - q^T A(\hat{\beta})^{-1} q).
\]
Equation (5.41) then becomes

\[
\hat{\beta}_k^1 = \hat{\beta} + \frac{A(\hat{\beta})^{-1}(\Delta_{k,2} + \Delta_{k,3}) e_k^*}{[c_k - (\Delta_{k,2} + \Delta_{k,3})^T A(\hat{\beta})^{-1}(\Delta_{k,2} + \Delta_{k,3})]}
\]

and the 1-step approximation to \( \hat{\beta} - \hat{\beta}(k) \), when \( k \) is the censored case is

\[
\hat{\beta} - \hat{\beta}(k) = A(\hat{\beta})^{-1} (\Delta_{k,2} + \Delta_{k,3}) g_k
\]

(5.42)

where

\[
g_k = e_k \left\{ e_k - \lambda_k \sum_{t_i \in t_k} \left[ \sum_{j \in R_i} \lambda_j \right] \right\}^2 - (\Delta_{k,2} + \Delta_{k,3})^T A(\hat{\beta})^{-1}(\Delta_{k,2} + \Delta_{k,3})
\]

(5.43)

since \( e_k^* = - e_k \) for censored cases.

Computationally we have found that \( g_k \) is close to 1. Table 5.2 shows the values of \( g_k \) for the heart transplant data model III, (see Table 4.4) to lie in the interval (1.0, 1.4]. A similar result is obtained for the Veteran's Administration lung cancer data (Appendix B5) with all the 8 covariates fitted in the model. For this model \( 1.0 < g_k < 1.2 \). However, the large value of 1.20 for the Veteran's Administration lung cancer data is for observation ranked 118, which has the highest LD for the censored observation, computed from the accurate 1-step Newton-Raphson. Thus, for the censored observation we found that if \( \hat{\beta}(k) \) is far away from \( \hat{\beta} \) the \( g_k \) tends to get large.

In Section 5.5 we shall show that similar approximations are obtained for the censored case, by the case-weights perturbation
Table 5.2 Regression estimates obtained by the 1-step Newton Raphson augmented model when the case omitted from the data is censored. The table lists estimates, likelihood distance, LD1 and gk for some cases from the Stanford Heart transplant data [152] model III.
approach of Cain and Lange (1984). But for the uncensored (failures) cases, no analytical similarity is observed due to the somewhat complicated nature of the resulting approximation, as we shall see in the following.

For failure (uncensored) suspect case, the log of the likelihood \( L(\hat{\beta}_k^*, \gamma, \alpha) \) in (5.28) is

\[
L(\hat{\beta}_k^*, \gamma, \alpha) = \left( z_k^T \hat{\beta}_k^* + \gamma + \alpha \right) - \log \left[ \lambda_k'' + \sum_{j \in R_k - \{k\}} \lambda_j' \right]
\]

\[+ \sum_{i=1}^{n} \delta_i (z_i^T \hat{\beta}_k^* - \log \sum_{j \in R_i} \lambda_j') \]

(5.44)

where \( \lambda_k'' = \exp(z_k^T \hat{\beta}_k^* + \gamma + \alpha) \) and \( \lambda_j' = \exp(z_j^T \hat{\beta}_k^* + d_{jk} \gamma) \).

The score vector \( U(\hat{\beta}_k^*, \gamma, \alpha) \) is then a \((p+2)\) vector of

\[
\left[ U(\hat{\beta}_k^*, \gamma), U(\alpha) \right]^T
\]

with \( U(\hat{\beta}_k^*, \gamma) \) given in (5.30) and \( U(\alpha) \), which is a scalar given by

\[
U(\alpha) = \frac{\partial L(\hat{\beta}_k^*, \gamma, \alpha)}{\partial \alpha} = 1 - \frac{\lambda_k''}{\lambda_k'' + \sum_{j \in R_k - \{k\}} \lambda_j'}
\]

(5.45)

At point \( \hat{\beta}_k^* = \hat{\beta}, \gamma = \alpha = 0 \), this score then becomes

\[
U(\alpha) = 1 - \frac{\lambda_k}{\sum_{j \in R_k} \lambda_j}
\]

where \( \lambda_j = \exp(x_j^T \hat{\beta}) \).

Then the score vector can be written as
\[
U(\theta_k^*, \gamma, \alpha) = \begin{bmatrix}
U(\theta_k^*) \\
U(\gamma) \\
U(\alpha)
\end{bmatrix} = \begin{bmatrix}
0 \\
e_k^* \\
1 - \lambda_k^-
\end{bmatrix}
\]

where \(0\) is a \(p \times 1\) vector of zeros, and

\[
\overline{\lambda}_k = \lambda_k^* / \sum_{j \in R_k} \lambda_j
\]

(5.46)

As in previous analysis, the \((p+2) \times (p+2)\) observed information matrix \(A(\theta_k^*, \gamma, \alpha)\) may be partitioned in the following way,

\[
\begin{bmatrix}
A(\theta_k^*, \gamma) & A_{p1} \\
A_{p1}^T & c_{\alpha \gamma}
\end{bmatrix}
\]

where \(A(\theta_k^*, \gamma)\) is given by (5.37) and \(A_{p1}\) is a \((p \times 1)\) vector corresponding to

\[
- \left( \frac{\partial^2 L}{\partial \theta_k^* \partial \gamma} \right) = \begin{bmatrix}
z_k \lambda''_k \\
\lambda''_k + \sum_{j \in R_k - \{k\}} \lambda_j
\end{bmatrix}
\]

\[
\frac{\lambda''_k \left( z_k \lambda''_k + \sum_{j \in R_k - \{k\}} z_j \lambda_j \right)}{\left( \lambda''_k + \sum_{j \in R_k - \{k\}} \lambda_j \right)^2}
\]

and

\[
c_{\alpha \gamma} = - \left( \frac{\partial^2 L}{\partial \gamma \partial \alpha} \right) = \begin{bmatrix}
\lambda''_k \\
\lambda''_k + \sum_{j \in R_k - \{k\}} \lambda_j
\end{bmatrix}
\]

\[
\frac{\lambda''_k}{\left( \lambda''_k + \sum_{j \in R_k - \{k\}} \lambda_j \right)^2}
\]

(5.47)
Similarly differentiating (5.45) with respect to $\alpha$ we obtain

$$c_{\alpha\alpha} = - \left( \frac{\partial^2 L}{\partial \alpha^2} \right) = c_{\alpha\gamma}$$

At points of $\beta_k^* = \hat{\beta}$ and $\gamma = \alpha = 0$

$$A_{p1} = \frac{z_k \lambda_k}{\sum_{j \in R_k} \hat{\lambda}_j} - \lambda_k \left( \sum_{j \in R_k} \hat{\lambda}_j \right)^2$$

$$= \frac{\lambda_k}{\sum_{j \in R_k} \hat{\lambda}_j} (z_k - a_k)$$

with $a_k$ as defined in Section 4.4.1 evaluated at $\hat{\beta}$.

Let $\Delta_{k,1} = \delta_k(z_k - a_k)$. For $k^{th}$ case being a failure $\delta_k = 1$, hence from (5.46) $A_{p1}$ can be written as

$$A_{p1} = \lambda_k \Delta_{k,1}$$

Similarly from (5.47) evaluated at $\beta_k^* = \hat{\beta}$, $\gamma = \alpha = 0$, we have

$$c_{\alpha\gamma} = c_{\alpha\alpha} = \lambda_k (1 - \lambda_k)$$

The partitioned $(p+2) \times (p+2)$ observed information matrix $A(\beta_k^*, \gamma, \alpha)$ now takes the form of

$$A(\beta_k^*, \gamma, \alpha) = \begin{bmatrix}
A(\beta_k^*, \gamma) & \lambda_k \Delta_{k,1} \\
\lambda_k (1 - \lambda_k) \\
\lambda_k \Delta_{k,1}^T & \lambda_k (1 - \lambda_k) \\
\lambda_k (1 - \lambda_k)
\end{bmatrix}$$

(5.48)
Following through the previous analysis, the 1-step approximation, $\hat{\beta}_k^* - \hat{\beta}_k^1$ is then given by the first $p$ components of

$$- A(\hat{\beta}_k^*, Y, \alpha)^{-1} U(\hat{\beta}_k^*, Y, \alpha) \text{ at } \hat{\beta}_k^* = \hat{\beta}, \ Y = 0 \text{ and } \alpha = 0.$$ 

Simplification in terms $A(\hat{\beta})^{-1}$ similar to that in equation (5.42) is not really possible. However, we may rewrite (4.48) the following way

$$A(\hat{\beta}_k^*, Y, \alpha) = \begin{bmatrix} A(\hat{\beta}_k^*) & (-q, \bar{\lambda}_k \Delta_k, 1) \\ \hline \hline \hline \hline c_k & c_{aa} \\ \hline (-q, \bar{\lambda}_k \Delta_k, 1)^T & c_{aa} \\ \hline c_{aa} & c_{aa} \end{bmatrix}$$

where $q$ and $c_k$ are defined as before, and $c_{aa} = \bar{\lambda}_k (1 - \bar{\lambda}_k)$.

The inverse of $A(\hat{\beta}_k^*, Y, \alpha)$ partitioned correspondingly is thus

$$[A(\hat{\beta}_k^*, Y, \alpha)]^{-1} = \begin{bmatrix} M_0 & M_1 \\ \hline \hline \hline \hline M_1^T & M_2 \end{bmatrix}$$

where $M_0$, $M_1$ and $M_2$ are matrices with dimensions $(p \times p)$, $(p \times 2)$ and $(2 \times 2)$ respectively. The elements of these matrices may be computed from the elements of $A(\hat{\beta}_k^*, Y, \alpha)$ by the matrix simplification given in Appendix A1. Thus at $\hat{\beta}_k^* = \hat{\beta}$, $Y = \alpha = 0$,

$$M_0 = \begin{bmatrix} A(\hat{\beta}) - (-q, \bar{\lambda}_k \Delta_k, 1) & \begin{bmatrix} c_k & c_{aa} \\ \hline \hline \hline \hline c_{aa} & c_{aa} \end{bmatrix}^{-1} \\ \hline \hline \hline \hline (-q, \bar{\lambda}_k \Delta_k, 1)^T \end{bmatrix}^{-1}$$
\[
M_1 = - M_0 (q, \bar{\lambda}_k, \Delta_k) \begin{bmatrix} c_k & c_{aa} \\ c_{aa} & c_{aa} \end{bmatrix}^{-1}
\]

and

\[
M_2 = \begin{bmatrix} c_k & c_{aa} \\ c_{aa} & c_{aa} \end{bmatrix}^{-1} \begin{bmatrix} I - (q, \bar{\lambda}_k, \Delta_k, 1)^T M_1 \end{bmatrix}
\]

(5.49)

The 1-step Newton-Raphson corresponding to (5.40) is

\[
\begin{bmatrix}
\beta_k^0 \\
\gamma \\
a
\end{bmatrix} = \begin{bmatrix}
\beta \\
0 \\
0
\end{bmatrix} + \begin{bmatrix} M_0 & M_1 \\ M_1^T & M_2 \end{bmatrix} \begin{bmatrix} 0 \\ \cdots \\ e_k^* \end{bmatrix}
\]

and the first \( p \) components are

\[
\beta_k^1 = \beta + M_1 \begin{bmatrix} e_k^* \\ 1 - \lambda_k \end{bmatrix}
\]

that is

\[
\hat{\beta} - \beta_k^1 = - M_1 \begin{bmatrix} e_k^* \\ 1 - \lambda_k \end{bmatrix}
\]

\[
= M_0 (q, \bar{\lambda}_k, \Delta_k, 1) \begin{bmatrix} c_k & c_{aa} \\ c_{aa} & c_{aa} \end{bmatrix}^{-1} \begin{bmatrix} e_k^* \\ 1 - \lambda_k \end{bmatrix}
\]

\[
= \frac{M_0}{[c_k - c_{aa}]} (q, \bar{\lambda}_k, \Delta_k, 1) \begin{bmatrix} 1 & -1 \\ -1 & c_k \\ c_{aa} \end{bmatrix} \begin{bmatrix} e_k^* \\ 1 - \lambda_k \end{bmatrix}
\]
Substituting for \( q \) and \( c_{\alpha} \) we have

\[
\hat{\beta} - \hat{\beta}_k = M \frac{\left[ (1-\lambda_k)(\Delta_k, 2 + \Delta_k, 3) - e_k^* (\lambda_k, \Delta_k, 1 + \Delta_k, 2 + \Delta_k, 3) + c_k \Delta_k, 1 \right]}{c_k - \lambda_k (1-\lambda_k)}
\]

(5.50)

Equations (5.42) and (5.50) are the direct consequence of the approach followed by Storer and Crowley (1985). The expressions are somewhat complicated, but they may be computed straightforwardly from results obtained in the computation of the maximum likelihood estimates \( \hat{\beta} \). Applying these to different data set, and comparing with the difference obtained using an exact 1-step Newton-Raphson discussed in Section 5.3.2, we found that there is little difference. As an example, Figure 5.8 shows the closeness in the measures, \( LD_k \) of (5.13) computed from \( \hat{\beta} - \hat{\beta}_k \) of the "correct" 1-step for case deleted model, and \( \hat{\beta} - \hat{\beta}_k \), of the augmented model, applied to Stanford Heart transplant data, (model III, Table 4.4). This suggests that in terms of 1-step diagnostic approximations, the augmented model provides good approximations to those for the case deleted model with likelihood \( \ell(\hat{\varphi}/w_k = 0) \).
Fig. 5.8 Comparison of the likelihood distances $LD_k$ between the augmented and the deletion models for the Stanford Heart Transplant data - model III
5.4.2 EM 1-step from Augmented Linear Model.

We can now apply the EM scheme (4.56) to the augmented model (5.23). The \( n \times (p+1) \) model matrix is now given by \( Z_+ = [Z, d_k] \), where \( d_k \) is now column centred, that is, if \( d_{jk} \), \( j = 1, 2, \ldots, n \) and \( 1 \leq k \leq n \), are the elements of \( d_k \), then

\[
\begin{cases}
  - \frac{1}{n-1} & j \neq k \\
  1 & j = k
\end{cases}
\]

Note also that the columns of \( Z \) are centred. The 1-step scheme starting at \( \beta = \hat{\beta} \) and \( \gamma = 0 \) is thus:

\[
\begin{pmatrix}
  \hat{\beta}^1 \\
  \gamma
\end{pmatrix} = \begin{pmatrix}
  \hat{\beta} \\
  0
\end{pmatrix} + (Z_+^T Z_+)^{-1} Z_+^T \hat{e}^*
\]

(5.52)

where \( \hat{e}^* \) is the Lagakos residual. Now

\[
Z_+^T \hat{e}^* = \begin{bmatrix}
Z^T \hat{e}^* \\
Y^T e^*
\end{bmatrix}
\]

at \( \beta = \hat{\beta} \) and \( \gamma = 0 \), \( Z_+^T \hat{e}^* = 0 \), and

\[
d_+^T e^* = - \frac{1}{n-1} \sum_{i \neq k} e_i^* + e_k^*
\]

\[
e_k^* + \frac{1}{n-1} e_k^*
\]

\[
= \left\{ \frac{n}{n-1} \right\} e_k^*
\]

(5.53)

since

\[
\sum_{i=1}^{n} e_i^* = 0.
\]
Hence at $\hat{\beta} = \hat{\beta}$ and $\gamma = 0$

\[
Z_+ e^* = \begin{bmatrix} 0 \\ \left( \frac{n}{n-1} \right) e_k^* \end{bmatrix}
\]

Now

\[
Z_+^T Z_+ = \begin{bmatrix} Z^T Z & Z^T d_k \\ d_k Z & d_k d_k \end{bmatrix}
\]

Let the inverse of $Z_+^T Z_+$ have corresponding components

\[
(Z_+^T Z_+)^{-1} = \begin{bmatrix} A_{11} & A_{12} \\ A_{12}^T & A_{22} \end{bmatrix}
\]

From (5.52) the first $p$ components of the $(p + 1)$ vector are

\[
\hat{\beta}_k = \hat{\beta} + A_{12} \left( \frac{n}{n-1} \right) e_k^*
\]

As previously, using Appendix A1, for matrix simplification we have

\[
\hat{\beta} - \hat{\beta}_k = A_{11} Z^T d_k \left( \frac{n}{n-1} \right) e_k^*
\]

\[
= A_{11} Z^T d_k e_k^*
\]

since $d_k^T d_k = n(n-1)^{-1}$. Now from Appendix A1

\[
A_{11} = \left[ Z^T Z - \frac{Z^T d_k d_k^T Z}{n(n-1)^{-1}} \right]^{-1}
\]
\[
\begin{align*}
&= (Z^T Z)^{-1} \left\{ I + \frac{Z^T d_k}{n(n-1)^{-1}} \left[ I - \frac{(d_k^T Z)(Z Z^T)^{-1} Z^T d_k}{n(n-1)^{-1}} \right]^{-1} d_k^T \right\} Z^T Z^{-1} \\
&= (Z^T Z)^{-1} + \frac{(Z^T Z)^{-1} Z^T d_k d_k^T Z (Z^T Z)^{-1}}{n(n-1)^{-1} (1-w_k)}
\end{align*}
\]

where \( w_k = \frac{d_k^T Z (Z^T Z)^{-1} Z^T d_k}{n(n-1)^{-1}} \), since the term in the square brackets is scalar. Then

\[
\hat{\beta}_k - \beta_k = (Z^T Z)^{-1} Z^T d_k e_k^* + \frac{(Z^T Z)^{-1} Z^T d_k d_k^T Z (Z^T Z)^{-1} Z^T d_k e_k^*}{n(n-1)^{-1} (1-w_k)}
\]

\[
= (Z^T Z)^{-1} Z^T d_k e_k^* \left[ 1 + \frac{n(n-1)^{-1} w_k}{n(n-1)^{-1} (1-w_k)} \right]
\]

\[
= \frac{(Z^T Z)^{-1} Z^T d_k e_k^*}{1-w_k}.
\]

Now \( d_k \) can be written as

\[
d_k = \begin{pmatrix} n \\ n(n-1) \end{pmatrix} 0_{-k} - \begin{pmatrix} 1 \\ n(n-1) \end{pmatrix} 1
\]

where \( 0_{-k} \) is an \( n \times 1 \) vector of zeros, except for the \( k \)th component which is unity, and \( 1 \) is an \( n \times 1 \) vector of ones.

Then

\[
Z^T d_k = \begin{pmatrix} n \\ n(n-1) \end{pmatrix} z_k
\]

since \( Z^T 1 = 0 \) because of \( Z \) being column centred. We have thus

\[
w_k = n(n-1)^{-1} z_k^T (Z^T Z)^{-1} z_k
\]

\[
= n(n-1)^{-1} v_{kk}
\]
where \( v_{kk} = z_k^T (Z^T Z)^{-1} z_k \) as defined in (5.22).

Then
\[
\hat{\beta} - \beta_k = \left( \frac{n}{n-1} \right) (Z^T Z)^{-1} z_k e_k^* \left[ 1 - n(n-1)^{-1} v_{kk} \right]^{-1}
\]
which is identical to (5.22) for case deletion model. Here we have found analytical evidence to support Storer and Crawleys' (1985) finding mentioned in the previous Section 5.3.4, that the augmented model (5.23) and case deletion model gives similar results. Here the iterative method is based on the EM algorithm rather than the Newton-Raphson technique.

5.5 CASE-WEIGHT PERTURBATION.

The likelihood, \( \hat{l}(\beta/\omega) \) for the perturbation scheme is given in (5.15). Recalling the discussion in Section 5.3.1, we note here that if all components of \( \omega \) is equal to one, then \( \hat{l}(\beta/\omega = 1) \) correspond to the likelihood in (4.09), for the full data set; and if
\[
\omega_i = \begin{cases} 
1 & i \neq k \\
0 & i = k 
\end{cases}
\]
then \( \hat{l}(\beta/\omega_k = 0) \) corresponds to the partial likelihood from the data set with the \( k \)th case omitted.

The method of Cain and Lange (1984) is to consider varying only one component of \( \omega \), corresponding to the suspect case, while the other components of \( \omega \) are set to 1. Let \( \hat{l}(\beta/\omega_k) \) denote the likelihood with the suspect \( k \)th case, and let \( U(\beta/\omega_k) \) be the score vector resulting from differentiating the logarithm of
\( \ell(\beta/w_k) \) with respect to \( \beta \). If \( \hat{\beta}(w_k) \) maximizes \( \ell(\beta/w_k) \) then \( \hat{\beta}(w_k) \) satisfies \( U(\hat{\beta}(w_k)/w_k) = 0 \). We now consider \( U(\hat{\beta}(w_k)/w_k) \) as a function of \( w_k \) and differentiate it with respect to \( w_k \), to obtain

\[
\left[ \frac{\partial}{\partial w_k} U(\beta/w_k) \right] + \left[ \frac{\partial}{\partial \beta} U(\beta/w_k) \right] \frac{\partial \hat{\beta}(w_k)}{\partial w_k} = 0.
\]

The derivatives are evaluated at \( \beta = \hat{\beta}(w_k) \). Since all the resulting functions are continuous in \( w_k \), we can let \( w_k \to 1 \). Then the second term in the square brackets in the L.H.S. of the above equation tends to \(-A(\hat{\beta})\) (see equation 4.27), by virtue that in the limit \( w_i \to 1 \) for all \( i \). Thus at \( \beta = \hat{\beta} \) and \( w_k = 1 \) we have

\[
\frac{\partial \hat{\beta}(w_k)}{\partial w_k} = A(\hat{\beta})^{-1} \left[ \frac{\partial}{\partial w_k} U(\beta/w_k) \right].
\]

We have from Taylor series, the first order approximation

\[
\hat{\beta}(w_k) \approx \hat{\beta}(w_0) + \frac{\partial \hat{\beta}(w_k)}{\partial w_k} (w_k - w_0) \tag{5.54}
\]

at \( w_0 = 1 \), \( \hat{\beta}(w_0) = \hat{\beta} \), then we have in general

\[
\hat{\beta} - \hat{\beta}(w_k) \approx (1-w_k) \left[ \frac{\partial \hat{\beta}(w_k)}{\partial w_k} \right] \tag{5.55}
\]

By putting \( w_k = 0 \), \( \hat{\beta}(w_k) \) is equivalent to \( \hat{\beta}(k) \), the estimate obtained by deleting the \( k \)th case completely from the data set.

Equation (5.55) thus provides an approximation to \( \hat{\beta} - \hat{\beta}(k) \), namely

\[
\hat{\beta} - \hat{\beta}(k) \approx A(\hat{\beta})^{-1} \left[ \frac{\partial}{\partial w_k} U(\beta/w_k) \right] \tag{5.56}
\]
where the derivative is evaluated at $w_k = 1$ and $\beta = \hat{\beta}$. We need now to determine $\frac{\partial}{\partial w_k} U(\beta / w_k)$, to complete the approximation term in (5.56). From $\ell(\beta / w)$ in (5.15), the score vector $U(\beta / w)$ is given by

$$U(\beta / w) = \sum_{i=1}^{n} \delta_i \left\{ w_i z_i - w_i a_i(w) \right\} \tag{5.57}$$

where the $p \times 1$ vector $a_i(w)$ is

$$a_i(w) = \frac{\sum_{j \in R_i} w_j z_i \lambda_j}{\sum_{j \in R_i} w_j \lambda_j}$$

We also require the differential of $a_i(w)$ with respect to $w_k$

$$\frac{\partial a_i(w)}{\partial w_k} = \frac{\sum_{j \in R_i} z_j \lambda_j \frac{\partial w_j}{\partial w_k}}{\sum_{j \in R_i} w_j \lambda_j} - \left[ \sum_{j \in R_i} w_j \lambda_j \right] \lambda_k \frac{\partial a_i}{\partial w_k}$$

since all the $w_j$'s are set to 1, and only $w_k$ is to vary, then

$$\frac{\partial a_i}{\partial w_k} = \begin{cases} 0 & j \neq k \\ 1 & j = k \end{cases}$$

and evaluating at $w = 1$

$$\frac{\partial a_i(w)}{\partial w_k} = z_k \lambda_k - \sum_{j \in R_i} \lambda_j \frac{a_i}{j \in R_i}$$
where

$$a_i = \left\{ \sum_{j \in R_i} z_j \lambda_j \right\} / \left\{ \sum_{j \in R_i} \lambda_j \right\}$$

, see equation (4.26).

Differentiating (5.57) with respect to $\omega_k$, and noting that

$$\frac{\partial a_i(w)}{\partial \omega_k} = 0 \text{ for } t_i > t_k \text{ since } t_k, \text{ and hence } \omega_k, \text{ is not included in}

the risk set } R_i \text{, we have for } t_i \leq t_k

$$

$$\frac{\partial U(\hat{\theta}/w)}{\partial \omega_k} = \delta_k z_k - \sum_{i=1}^{n} \frac{\partial}{\partial \omega_k} \left[ \delta_i \omega_{i} a_i(w) \right]

= \delta_k z_k - \delta_k a_k - \sum_{t_i \leq t_k} \delta_i \omega_i \frac{\partial a_i(w)}{\partial \omega_k}

= \delta_k (z_k - a_k) + \lambda_k \sum_{t_i \leq t_k} \left( \sum_{j \in R_i} \lambda_j \right) - z_k \lambda_k \sum_{t_i \leq t_k} \sum_{j \in R_i} \lambda_j

$$

with evaluation of the derivative at $w = 1$. Now let

$\Delta_{k,1} = \delta_k (z_k - a_k)$. At $\hat{\theta} = \hat{\theta}$, the second and third terms of the

R.H.S. of the above equation are $\Delta_{k,2}$ and $\Delta_{k,3}$ respectively (see

equations 5.34 and 5.35). Then

$$\left[ \frac{\partial U(\hat{\theta}/w)}{\partial \omega_k} \right]_{w=1, \hat{\theta}} = \Delta_{k,1} + \Delta_{k,2} + \Delta_{k,3}

$$

and hence (5.55) becomes

$$\hat{\theta} = \hat{\theta}(k) \approx A(\hat{\theta})^{-1} \Delta_k \quad (5.58)

$$

where

$$\Delta_k = \Delta_{k,1} + \Delta_{k,2} + \Delta_{k,3} \quad .$$
(5.58) can be considered as a local approximation to \( \hat{\beta} - \hat{\beta}_{(k)} \) since it is derived from the first order approximation (5.54). It is the diagnostic given by Cain and Lange (1984, equation 2). Now \( \Delta_{k,1} \) is included only if the \( k \)th case is a failure, and is then the Schoenfeld's (1982) residual discussed in Section 4.4.2. For a censored \( k \)th case, \( \Delta_{k,1} \) vanishes and we have similar approximations, equation (5.42), as that obtained for the augmented model.

The sum of \( \Delta_{k,2} \) and \( \Delta_{k,3} \) measures the effect of the \( k \)th case on all risk sets which include it. The value is increasing with large \( t_k \) since the summation in these terms are over all failure times no larger than \( t_k \). This agrees with intuition that early censored cases should have little effect, and cases with large \( t_k \) should have influences which are similar irrespective of being a failure or censored. The effect of an early censoring is like a missing observation.

\( \hat{\beta} - \hat{\beta}_{(k)} \) is well approximated using (5.58). However for the most influential cases we found that in general, the approximation underestimates. Nevertheless it is able to indicate cases that are influential.

Reid and Crépeau (1985) arrived at the identical formula
\[ A(\hat{\beta})^{-1} \Delta_k \]
for their empirical influence function \( \hat{I}_k \). Their derivation of the influence functions is outlined in their Appendix, page 8. The authors also show that \( \sum \hat{I}_k = 0 \), and \( \sum \hat{I}_k^T \hat{I}_k \) is asymptotically equivalent to \( A(\hat{\beta})^{-1} \).
5.5.1 Properties of the Diagnostic $\Lambda_k$.

From the preceding section we have

$$\Lambda_k = \Delta_{k,1} + \Delta_{k,2} + \Delta_{k,3}.$$  

The Breslow's (1972) estimate of the integrated baseline hazard is as (5.04), that is

$$\hat{H}_o(t_k) = \sum_{j \in \tau_k} \sum_{i \not\in \tau_k} \delta_j \frac{\lambda_i}{\sum_{j \in r_j} \lambda_j}.$$  

With the definitions of $\Delta_{k,2}$ and $\Delta_{k,3}$ as in (5.34) and (5.35) respectively we note that

$$\Delta_{k,2} = -\lambda_k \frac{\partial}{\partial \beta} \hat{H}_o(t_k)$$  

$$\Delta_{k,3} = -\lambda_k z_k \hat{H}_o(t_k)$$

where $\hat{H}_o(t_k)$ is considered as a function of $\lambda_j = \exp(z_j^T \beta)$.

Similarly with $e_k = \lambda_k \hat{H}_o(t_k)$; so when differentiating with respect to $\beta$ we have

$$\frac{\partial e_k}{\partial \beta} = \lambda_k \frac{\partial \hat{H}_o(t_k)}{\partial \beta} + \lambda_k z_k \hat{H}_o(t_k)$$

$$= - (\Delta_{k,2} + \Delta_{k,3})$$

so that

$$\Delta_k = \Delta_{k,1} - \frac{\partial e_k}{\partial \beta}.$$  

Thus we see here that the rate of change of $e_k$ with respect
to $\beta$ is the most significant for the case-weight influence. It is more dominant than $\delta_k(z_k - a_k)$ for the more influential cases. This is illustrated in our examples for the heart transplant date (Appendix B6) model III (see Table 4.4) and BR disc data Appendix B7a in Section 5.6.

Note also the $e^* = \delta - e_k$, then $\Delta_k$ can be written as

$$\Delta_k = \delta_k(z_k - a_k) + \frac{\partial e^*}{\partial \beta}.$$ 

Now $\sum_{k=1}^{n} e_k^* = 0$. This implies that $\sum_{k=1}^{n} \frac{\partial e^*}{\partial \beta} = 0$.

Also $\sum_{k=1}^{n} \delta_k(z_k - a_k) = U(\bar{\beta}) = 0$, then

$$\sum_{k=1}^{n} \Delta_k = 0$$

which conforms to the result of Reid and Crépeau (1985) that

$$\sum_{k=1}^{n} \hat{I}_k = 0, \text{ since } \sum_{k=1}^{n} A(\beta)^{-1} \Delta_k = 0 \text{ is true}$$

when $\sum_{k=1}^{n} \Delta_k = 0$.

Since $\Delta_k$ is approximately proportional to $\hat{\beta} - \beta(\hat{\beta})$ plots of $\Delta_k$ on rank of observation (or observations) may give some insight in the assessment of influence.

5.5.2 Cook's Likelihood Displacement.

The likelihood displacement as reviewed and extended by Cook (1986) is defined in (5.14), that is
\[ LD(w) = 2\left[ L(\hat{\theta}) - L(\hat{\theta}(w)) \right] \]

where \( \hat{\theta} \) and \( \hat{\theta}(w) \) are the maximum likelihood estimates for the unperturbed and perturbed model respectively. The idea is to use the weights \( w \) for each case to investigate the direction in which the greatest changes occur. From a Taylor series expansion we have

\[ L(\hat{\theta}(w)) \approx L(\hat{\theta}) + (\hat{\theta}(w) - \hat{\theta})^T \left( \frac{\partial L(\theta)}{\partial \theta} \right) + \frac{1}{2} (\hat{\theta}(w) - \hat{\theta})^T \left( \frac{\partial^2 L(\theta)}{\partial \theta^2} \right) (\hat{\theta}(w) - \hat{\theta}) \]

where the derivatives are evaluated at \( \hat{\theta} = \hat{\theta} \) and \( w = 1 \). Hence the term

\[ \frac{\partial L(\theta)}{\partial \theta} = 0 \]

and the second order derivative is equal to the negative of the observed information matrix \( A(\hat{\theta}) \). Then the above equation becomes

\[ 2\left[ L(\hat{\theta}) - L(\hat{\theta}(w)) \right] = (\hat{\theta}(w) - \hat{\theta})^T A(\hat{\theta})(\hat{\theta}(w) - \hat{\theta}) \quad (5.59) \]

Similarly \( \hat{\theta}(w) \) can be approximated by the following

\[ \hat{\theta}(w) \approx \hat{\theta}(w_o) + \left( \frac{\partial \hat{\theta}(w)}{\partial \theta} \right)_{\theta = \hat{\theta}} (w - w_o) \]

for some \( w_o \). If the \((n \times 1)\) vector \( w_o = 1 \), then \( \hat{\theta}(w_o) = \hat{\theta} \), (see Section 5.3.1).

Then we have

\[ \hat{\theta}(w) - \hat{\theta} = \left( \frac{\partial \hat{\theta}(w)}{\partial \theta} \right)_{\theta = \hat{\theta}, w_o} (w - w_o) \]
It was shown in Section 5.5 that for a component of \( w \), say \( w_k \)

\[
\frac{\partial \hat{\theta}(w_k)}{\partial w_k} = A(\hat{\theta})^{-1} \left[ \frac{\partial U(\theta/w_k)}{\partial w_k} \right]_{\theta = \hat{\theta}, w_k = 1}
\]

\[= A(\hat{\theta})^{-1} \Delta_k \]

Extending this to apply to the \( n \times 1 \) vector \( w \), we have

\[
\frac{\partial \hat{\theta}(w)}{\partial w^T} = A(\hat{\theta})^{-1} \Delta
\]

where \( \Delta \) is \( p \times n \) matrix whose elements are \( \Delta_{ij} = \frac{\partial^2 L(\theta/w)}{\partial \theta_i \partial \theta_j} \) where \( L(\theta/w) \) is the logarithm of the perturbed likelihood \( L(\theta/w) \), (equation (5.15)). Then

\[
\hat{\theta}(w) - \hat{\theta} = A(\hat{\theta})^{-1} \Delta(w - w_o)
\]

Substituting this in (5.59), we have

\[
LD(w) = 2 \left[ L(\hat{\theta}) - L(\hat{\theta}(w)) \right]
\]

\[= (w - w_o)^T \Delta^T A(\hat{\theta})^{-1} A(\hat{\theta}) A(\hat{\theta})^{-1} \Delta(w - w_o)
\]

\[= (w - w_o)^T \Delta^T A(\hat{\theta})^{-1} \Delta(w - w_o) \quad (5.60)
\]

which provides the local approximation to \( LD(w) \) around \( w_o \).

Putting the elements of \( w \) equal to 1 except the suspect \( k \)th case which is zero, we have in general, the likelihood distance
when the $k^{th}$ case is omitted from the model, that is

$$LD_k = \Delta_k^T A^{-1}(\hat{\beta}) \Delta_k.$$ 

From the Cain and Lange diagnostic in (5.58), we have

$$\Delta_k \approx A(\hat{\beta})(\hat{\beta} - \hat{\beta}_{(k)}).$$

Hence

$$LD_k \approx (\hat{\beta} - \hat{\beta}_{(k)})^T A(\hat{\beta})(\hat{\beta} - \hat{\beta}_{(k)}),$$

which is identical to the approximation to $LD_k$ in (3.63) in Chapter 3.

The main drawback against single case deletion statistics in determining influential cases is that some may be masked in the process. The case deletion approach is equivalent to constraining the $w$ to values of 0 and 1. But Cook (1986) suggests choosing simultaneously $w$ such that the approximation (5.60) to $LD(w)$ is maximised; thus finding a local measure of influence for characterizing the behaviour of an influence graph around $w_o$.

5.5.3 Local Influence.

The equation given by (5.60) is a local approximation to $LD(w)$ around $w_o$. In a general development of local influence, Cook (1986) suggests choosing the weight $w$ so that this local approximation is maximized. This amounts to finding a direction for a $n \times 1$ vector, $\xi$; of unit length which maximizes $\xi^T B \xi$ where $B = \Delta^T A(\hat{\beta})^{-1} \Delta$. In this scheme $A(\hat{\beta})^{-1}$ is a $p \times p$ positive definite matrix and $\Delta$ is a $p \times n$ matrix, so that $B$ is a $n \times n$ symmetric semi-
positive definite with rank no more than p. Maximizing $\ell^T B \ell$ is equivalent to finding a unit length eigenvector, $\ell_{\text{max}}$, which corresponds to the largest eigenvalue, $\lambda_{\text{max}}$ of B. Thus

$$
\ell_{\text{max}}^T B \ell_{\text{max}} = \ell_{\text{max}}^T \lambda_{\text{max}} \ell_{\text{max}} = \lambda_{\text{max}} \ell_{\text{max}}^T \ell_{\text{max}} = \lambda_{\text{max}}
$$

that is the maximum value of $\ell^T B \ell$ correspond to the largest eigenvalue, $\lambda_{\text{max}}$ of B.

Cook (1986) also considers a statistic $C_{\text{max}}$, the maximum curvature, which is equal to $2\lambda_{\text{max}}$ here. He argues that values of $C_{\text{max}}$ greater than 2 or $\lambda_{\text{max}}$ greater than 1 indicates notable local sensitivity. Since $\ell_{\text{max}}$ is invariant to multiplication of its elements by -1, the cases which have large local influences are those which have elements of $\ell_{\text{max}}$ which are large in absolute value. Substituting $(w-w^0)$ in (5.60) for $\pm \ell_{\text{max}}$ causes the approximation to $\text{LD}(w)$ to be maximized for changes, $w-w^0$, of unit length. Thus the plots of the elements of $\ell_{\text{max}}$ against case number and covariates can throw some light on the local influence, and can be important diagnostic tools. Trivially, we note, if $p = 1$, then $\ell_{\text{max}}$ is proportional to $\Delta^T$ and $\lambda_{\text{max}} = (\hat{\Theta})^{-1} \| \Delta \|$.

To find the local influence measures $\ell_{\text{max}}$ and $\lambda_{\text{max}}$ we need to compute the eigenvectors and eigenvalues of the $n \times n$ matrix B. For some analyses $n$ may be large, and finding these values direct
using computer routines available would require large amounts of memory space. Furthermore the matrix $B$ is of rank $p$, which implies that at most there are only $p$ non-zero eigenvalues. We employ the following means to obtain $\lambda_{\text{max}}$ and $\lambda_{\text{max}}$ without forming the matrix $B$.

If $C$ is a $p \times n$ matrix, and $U = C^TC$ and $V = CC^T$ then as a consequence of the singular value decomposition theorem (see, for example, Mardia et al., 1979, page 473), the non-zero eigenvalues of $U$ and $V$ are equal. Further if $v$ is an eigenvector of $V$ with corresponding eigenvalue $\lambda$, then $u = C^Tv$ is an eigenvector of $U$ with corresponding eigenvalue. Note also that $U$ and $V$ are symmetric and they are positive semi-definite matrices. The positive eigenvalues of $V$ correspond to the positive eigenvalues of $U$ and vice-versa. Thus if $v_{\text{max}}$ corresponds to the eigenvector of $V$ having largest eigenvalue, then $u_{\text{max}}$, similarly defined for $U$, is given by $C^Tv_{\text{max}}$.

To apply this to matrix $B$, we note that $A(\hat{\beta})^{-1}$ is positive definite which allows a Cholesky decomposition (Stewart, 1973) thus we can represent the $p \times p$ matrix $A(\hat{\beta})^{-1}$ as

$$A(\hat{\beta})^{-1} = M^TM$$

where $M$ is an upper triangular $p \times p$ matrix. Then

$$B = \Delta^T A(\hat{\beta})^{-1} \Delta$$

$$= (M\Delta)^T M \Delta$$

putting the $p \times n$ matrix $C = M\Delta$, then the non-zero eigenvalues of $B$ are equal to the eigenvalues of the $p \times p$ matrix $V = CC^T$, and
\( \lambda_{\text{max}} \) is proportional to \( \Delta T \mathbf{M}^T \mathbf{v}_{\text{max}} \). These can be found numerically by using the NAG routines F01BKF and F02ABF for the Cholesky decomposition and the eigenvalue (and eigenvectors) computations respectively.

5.5.4 Case-weights for Explanatory Variables.

Cook (1986) extended the idea of case-weights perturbation and local influence of the previous chapter, to investigate the effects of minor perturbation of explanatory variables in linear regression. Here we adopt his methodology for the proportional hazards model. Initially we shall assume the first of the explanatory variable that is being perturbed, such that if \( x \) is the value of the first variable, then under the perturbed scheme we have for the first explanatory variable

\[
x + s w
\]

where \( s \) is some scalar chosen to take into account of the scale of \( x \). Thus the perturbed explanatory variable for the \( i \)th observation is

\[
z_i^T(w_i) = (x_i + s w_i, v_i^T)
\]

where \( v_i \) is a \((p-1)\) vector of \( z_i \) with the first component omitted.
Correspondingly we put
\[
\delta^T = (\gamma, \theta^T)
\]

where \( \gamma \) is a scalar and \( \theta \) is \((p-1)\) vector.

Letting \( y_i = x_i + s w_i \), and following (4.07) and (5.02) the model then has a hazard function for the \( i \)th observation

\[
h(t; y_i, v_i) = h_0(t) \exp(y_i \gamma + v_i^T \theta)
\]
Let $U(\gamma/w)$ and $U(\theta/w)$ be the score vectors for given $w$ with respect to $\gamma$ and $\theta$ respectively. Then

$$U(\theta/w) = \left\{ \begin{array}{l}
U(\gamma/w) \\
U(\theta/w)
\end{array} \right.$$

$$= \left[ \begin{array}{c}
\sum_{i=1}^{n} \delta_i \left( y_i - \frac{\sum_{j \in R_i} y_j}{\sum_{j \in R_i} \lambda_j} \right) \\
\sum_{i=1}^{n} \delta_i \left( v_i - \frac{\sum_{j \in R_i} v_j}{\sum_{j \in R_i} \lambda_j} \right)
\end{array} \right]$$

where $\overline{\lambda}_j = \exp(y_j \gamma + \frac{T}{v_j \theta})$.

Then

$$\frac{\partial U(\gamma/w)}{\partial w} = s \left\{ \delta_k - \lambda_k (1 + y_k \gamma) \sum_{t_i \in t_k} \left[ \frac{\delta_i}{\sum_{j \in R_i} \lambda_j} \right] \right.$$  

$$+ \gamma \lambda_k \sum_{t_i \in t_k} \delta_i \left[ \left( \sum_{j \in R_i} y_j \overline{\lambda}_j \right) / \left( \sum_{j \in R_i} \overline{\lambda}_j \right)^2 \right] \right\}$$

$$+ \gamma \lambda_k \sum_{t_i \in t_k} \frac{\delta_i}{\lambda_j} \left[ \left( \sum_{j \in R_i} v_j \overline{\lambda}_j \right) / \left( \sum_{j \in R_i} \overline{\lambda}_j \right)^2 \right]$$

At the maximum likelihood estimates $\hat{\gamma}$ and $\hat{\theta}$, and $w = 0$, we
then have the values of the score vectors under the weight pesrurbation for the \( k \)th case as in equations (5.64) and (5.65) with \( \gamma, \lambda_j \) replaced by \( \hat{\gamma} \) and \( \hat{\lambda}_j \) respectively, where

\[
\hat{\lambda}_j = \exp(x_j \hat{\gamma} + v_j \hat{\theta})
\]

and \( y_j \) in equation (5.64) is replaced by \( x_j \). Then using expressions for \( \epsilon^*_k, \epsilon_k \) (Section 5.2.1) and \( \Delta_{k,2} \) (see (5.34)) we find that these equations can be written as

\[
\frac{\partial}{\partial \omega_k} U(\gamma/\omega) = s\left\{ \epsilon^*_k + \hat{\gamma}(\Delta_{k,2} - \epsilon_k \epsilon_k^*) \right\}
\]

and

\[
\frac{\partial}{\partial \omega_k} U(\theta/\omega) = s\hat{\gamma}(\Delta_{k,2} - \epsilon_k \epsilon_k^*) .
\]

If \( C_1 \) is the \( p \times n \) matrix with column equal to

\[
s^{-1}\left[ \frac{\partial}{\partial \omega_k} U(\gamma/\omega), \frac{\partial}{\partial \omega_k} U(\theta/\omega) \right]^T, \quad k = 1, 2, \ldots, n
\]

then

\[
C_1 = d_1(e^*_\omega)^T - \hat{\beta}_1 Z^T D(e) + \hat{\beta}_1 \Delta(2)
\]

where \( d_1 \) is a \( p \times 1 \) vector with 1 as the first element and zero elsewhere,

\[
e^*_\omega = (e^*_1, \ldots, e^*_n)^T \text{ the vector of residual defined in Section 5.1.1,}
\]

\( \hat{\beta}_1 \) is the first component of \( \hat{\beta} \); i.e. \( \hat{\gamma} \),

\( D(e) = \text{diag}(e_j) \)

and \( \Delta(2) \) is the \( p \times n \) matrix \((\Delta_{1,2}, \Delta_{2,2}, \ldots, \Delta_{n,2})\).
Following through the analysis in Section 5.4, if $\hat{\beta}(w)$ represents the estimate from the perturbed model, then from (5.54) applying to vectors $w$ and $w_o$, we have

$$\hat{\beta}(w) = \hat{\beta} - s A(\hat{\beta})^{-1} C_1 (w-w_o) \quad (5.69)$$

and

$$LD(w) = s^2 (w-w_o)^T C_1^T A(\hat{\beta})^{-1} C_1 (w-w_o) \quad (5.70)$$

where $w_o$ corresponds to the unperturbed model, that is $w_o = 0$.

As in the previous section we can consider the $n \times n$ matrix $B_1 = C_1^T A(\hat{\beta})^{-1} C_1$ and the eigenvector $\ell_{\text{max}}$ corresponding to the largest eigenvalue of $B_1$. Components of $\ell_{\text{max}}$ which are large in absolute value will indicate influential cases.

We can extend this scheme further to perturbing the $\ell$th component of $Z$ rather than the first. Then we obtain $C_\ell$, instead of $C_1$ where

$$C_\ell = d_\ell (e^*_\ell)^T - \hat{\beta}_\ell Z^T D(e) + \hat{\beta}_\ell \Delta^{(2)} \quad (5.71)$$

where $d_\ell$ is now a $p \times 1$ vector of zeros except the $\ell$th element which is one, and $\hat{\beta}_\ell$ is the $\ell$th component of $\hat{\beta}$.

Following Cook (1986) we can consider simultaneous perturbation of each component of $Z$. This leads to consideration of the matrix $B = C^T A(\hat{\beta})^{-1} C$, where $C$ is the $p \times np$ matrix $(C_1, C_2, \ldots, C_p)$. However, finding the eigenvector $\ell_{\text{max}}$ of the $np \times np$ matrix $B$ directly, is in general a cumbersome numerical task. The technique used in the previous section, using the Cholesky decomposition may be used, but we might be faced with computer memory storage problem, particularly when $n$ is large.
5.6 REVIEW OF NUMERICAL RESULTS.

In this section we shall review some of the results obtained by the techniques and methodology discussed in the previous sections as applied to some real data. Some of these examples have been given in the sections where appropriate.

Generally in the assessment of influence and in the diagnostics of the departure from proportional hazards model, plots are valuable tools to facilitate comparison in the key results of the analysis derived from the original and the "perturbed" models. In Section 5.2.2, the plots of Schoenfeld's residuals (Figure 5.1 - Figure 5.4) demonstrate the validity of the proportional hazards model with regards to the time dependency of the covariates. For the British Railway data (Appendix B7a) there appear to be some time dependency for large failure time. Plots of Lagakos's (1981) residuals, may be used to detect the effect of the omission of important variables from the model (see Section 5.2.1).

In the assessment of influence, our perturbation scheme is effectively single case deletion, and as a measure to quantify the changes in the result, we use the Cook likelihood distance. The approximations of this measure as found by the different methods discussed in this chapter are found to be in agreement. Figure 5.9 shows the likelihood distances as computed from results obtained by the different schemes applied on the Carcinogenesis data (Section 4.6.1). There is only one 1-step EM plot in the figure, since the results obtained by the augmented scheme are identical to that given by the deletion scheme for the EM algorithm (see Section 5.4.2). For this data set the suspect is case 38 with
Fig. 5.9 Plot of likelihood distances as computed by the various methods against ranks of the observations for the carcinogenesis data

![Plot of likelihood distances](image1)

Fig. 5.10 Plots of sample influence curves on ranks of observations for the carcinogenesis data

![Plots of sample influence curves](image2)
values of $SIC_k$ and likelihood distance of 10.154 and 0.558 respectively (see Table 5.1). Note that the values quoted are those from the fully iterated estimation. This case corresponds to a rat in group 1 which died at a much later stage after the insult of the carcinogen DMBA, than the rest in that group. Similar conclusion can be drawn from the plots of the sample influence curve, $SIC_k$, (since there is only one covariate), against the rank of the observation. Figure 5.10 shows the plots of $SIC_k$ obtained by the 1-step (correct) Newton-Raphson and the Newton-Raphson perturbed model.

For models with more than one covariate, the plots of the likelihood distance, $LD_k$ against the rank of the cases or some specific covariate is preferable to plots of the $SIC_k$ for individual covariates, since the $LD_k$ gives overall scalar measure of the influence. The plots for the Stanford Heart Transplant data on the 152 observations (see Section 4.6.3), model III (Table 4.4) on the rank of the observations, point to two particularly suspect cases (Figure 5.8). These are cases with ranks 2 and 139 corresponding to observations of 10 days (age 13, and T5 score 149) and 2006 days (age 15 and T5 score 126); the latter case being censored. This is in agreement with our findings in Section 3.8.2 for the same data set, that young patients who died early, or censored at a much later stage after the treatment, exert greater influence in the estimation of $\beta$.

Other data sets, for example the Veteran's Administration lung cancer data (Appendix B5) and British Railway data (Appendix B7) were also analysed. Generally the outcome of the analysis
with regards to the assessment of influence from the approximations are in agreement with all the methods described in the previous sections. Each of the methods has its own merit. The Newton-Raphson method, though more efficient computationally, the 1-step estimate of $\hat{\theta}_{(k)}$ cannot be derived easily from the quantities obtained from the maximum likelihood estimate, $\hat{\theta}$ for the full data. A technique as described in Section 5.4.1 is required, and even then gives terms which are difficult to simplify. On the other hand the 1-step EM algorithm gives a compact formulation similar to that found for case deletion in the least squares regression problem (see Section 5.3.4). It is also with this technique that we found analytical evidence that augmented model (Section 5.4) and case deletion give identical results.

In Section 5.5.1, we have discussed some properties of $\Delta_k$ and its three parts, $\Delta_{k,1}$, $\Delta_{k,2}$ and $\Delta_{k,3}$. Plots of these quantities are useful in the assessment of influential cases. Note that Schoenfeld's residual (Section 5.2.2) is $\Delta_{k,1} = \delta_k (z - \bar{z}_k)$, so some of the characteristics of this statistic is reflected in $\Delta_k$.

To explain the trend and systematic pattern in the scatter plots, we first consider a univariate regression model, such that the estimated regression is not large and that the $\hat{\lambda}_j = \exp(\hat{\beta} z_j)$, do not vary much. As an approximation and without loss of generality we take $\hat{\lambda}_j = 1$. Suppose also that the risk sets $R_j$ consist of cases which have values approximately uniformly distributed over the set of all sample values, which is consistent with $\hat{\lambda} = 1$. We then have the following approximation. If $\bar{z}$ is the average of $z$ over the whole data set, then with the above
assumption, the expected value \( \bar{a}_k = \sum_{j \in R_k} \lambda_j z_j / \sum_{j \in R_k} \lambda_j \) over the risk set \( R_k \) is approximately equal to \( \bar{z} \). For failed case, the Schoenfeld's residual, \( \Delta_{k,1} = z_k - \bar{z} \). Additionally if \( H_k \) is \( \hat{H}_o(t_k) \), the integrated base hazard (see (5.04)), then
\[
\Delta_{k,2} = \hat{H}_k \bar{z} \quad \text{and} \quad \Delta_{k,3} = -\hat{H}_k z_k,
\]
so that
\[
\Delta_{k,2} + \Delta_{k,3} = -H_k (z_k - \bar{z}) \quad (5.72)
\]
and
\[
\Delta_k = (\delta_k - H_k)(z_k - \bar{z}) \quad (5.73)
\]
Since the \( \lambda_j \sim 1 \), \( H_k \) does not show any dependence on \( z_k \) and \( H_k \) in (5.73) can be replaced by its average value \( \bar{H} \). Thus, if \( \Delta_k \) when considered as a function of \( z_k \) is plotted against \( z_k \), the result will be two bands of points, one, for failures, scattered about the line \( \Delta = (1-H)(z-\bar{z}) \), the other for censored cases, about the line \( \Delta = -\bar{H}(z-\bar{z}) \).

On the other hand if we assume there is a strong positive association between survival time and the covariate (assumed continuous) with values of \( z \) almost in the same rank order as the survival times, for \( \hat{\beta} \) large and negative we have
\[
\hat{\lambda}_1 >> \hat{\lambda}_2 >> \hat{\lambda}_3 \ldots \ldots \quad \text{etc.}
\]
for failure times \( t_i \) indexed to be in rank order, commencing at \( i = 1 \). Under these conditions, a crude analysis gives \( a_k \sim z_k \) and
\[
\sum_{j \in R_i} \lambda_j \sim \lambda_i
\]
\[
\sum_{t_i \leq t_k} \frac{\delta_i z_i}{\lambda_i} \propto \frac{\delta_k z_k}{\lambda_k}.
\]

Then
\[
\Delta_{k,2} + \Delta_{k,3} = \lambda_k \sum_{t_i \leq t_k} \frac{\delta_i a_i}{\lambda_j} - z_k \lambda_k \sum_{j \in R_i} \sum_{t_i \leq t_k} \frac{\delta_i}{\lambda_j}
\]

\[
= -\delta_k (z_k - a_k)
\]

\[
= -\Delta_{k,1}
\]

and so \(\Delta_k \propto 0\). Thus a plot of \(\Delta_k\) against \(z_k\) should give a scatter of points about \(\Delta_k = 0\).

An alternative measure to assess the influence of a case is the \(\ell_{\text{max}}\) criterion discussed in Section 5.5.3. This is a consequence of the local approximation to LD(w), and thus \(\ell_{\text{max}}\) is a measure of local influence around \(w_0\).

We analysed several data sets to illustrate these, but two will be reported in the following.

5.6.1 Brake Disc Data (Appendix 7b).

This data set of industrial reliability is from British Rail, Derby. It concerns the reliability of disc brakes fitted to railway carriages of high speed trains in the Western region of England. The "life-times" considered here are the distances travelled by the carriage until the disc brake unit failed in some minor way. Two explanatory variables are considered, both indicator variables, having values of zeros and ones. The first
$z_1$, gives the type of material of the disc and the second, $z_2$, gives the fixing position. This data set represents heavily censored data. There are 357 cases of which only 98 are uncensored, while the others are withdrawals from the study for some reason. The first failure occurs at the observation ranked 34 by increasing size. Note that for a proportional hazards model the first 33 observations are thus redundant in the data set.

The proportional hazards model was fitted with covariate $z_1$ and $z_2$. The summary of the estimates obtained by Newton-Raphson in 6 iterations are as in Table 5.3.

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Std. error</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_1$</td>
<td>0.923</td>
<td>0.730</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>0.332</td>
<td>0.211</td>
</tr>
</tbody>
</table>

estimated correlation $\hat{\sigma}_{12} = -0.13$

log likelihood $= - 480.7693$

**TABLE 5.3**: The regression coefficients for the proportional hazards model fitted to the Disc data.

Figure 5.11, shows plots of $\Delta_{k,1}$, $\Delta_{k,2} + \Delta_{k,3}$ and $\Delta_k$ against the rank of cases for the $z_1$ variable, the material. For the $\Delta_{k,1}$ plot, the two outstanding points correspond to the two failure cases with value of $z_1 = 1$ having ranks 34 and 128. This is to be expected as they are the only 2 uncensored cases having
Fig. 5.11 Plots of $\Delta$'s for 'material' for BR disc data against rank of distance travelled.

Plot 1

Plot 2

Plot 3

Legend:
- x: uncen
- o: cen
values of $z_l = 1$. (There are only 11 non-redundant cases with values of $z_l = 1$). Here $\bar{z} \approx 0$, and as explained in the above, a plot of $\Delta_{k,1} = z_k - \bar{z}$ give scatter points about $\Delta_{k,1} = 0$ except for the two uncensored cases with values $z_l = 1$. Similarly plot 2, for $\Delta_{k,2} + \Delta_{k,3}$ is explained. Here from (5.72), with $\bar{z} \approx 0$, $\Delta_{k,2} + \Delta_{k,3} = -H_k z_k$, giving most of the points on the axis and the non-zero points correspond to those cases with $z_l = 1$. Plot 2 indicates case 197 as probable influential point. Plot 3 of $\Delta_k$ against the rank gives all three cases which could be influential, namely cases ranked 134, 128 and 197.

Figure 5.12 gives the equivalent plots for the variable corresponding to the fixing position variable, $z_2$. Here we have $\bar{z} \approx \frac{1}{2}$. This variable is well balanced over the two values of $z_2$. The result in the plot of $\Delta_{k,1}$, is that we see two bands of points, almost 1 unit apart. The points with positive residuals correspond to cases with values of $z_2 = 1$, and conversely the points with negative residuals refer to those cases with values of $z_2 = 0$.

The plot indicates some influential cases, perhaps with ranks 321, 322 and 323. Cases ranked greater than 334 have $\Delta_{k,1} = 0$. This is explained by noting that the largest in observation with $z_2 = 1$ corresponds to the case ranked 334 which implies cases having ranks greater than 334 have $\Delta_{k,1} = 0$.

The plot $\Delta_{k,2} + \Delta_{k,3}$ (plot 2) for this variable is mostly consistent with $\Delta_{k,2} + \Delta_{k,3} = -H_k z_k$ suggested in the discussion above. We note that $H_k$ is an increasing function of the rank of the observation. Points with $\Delta_{k,2} + \Delta_{k,3} > 0$ correspond to cases with $z_2 = 0$, and the converse corresponds to cases with
Fig. 5.12 Plots of $\Delta'$s for 'position' for BR disc data against rank of distance travelled

**Plot 1**

- $\Delta_{k,1}$
- $\Delta_{k,2} + \Delta_{k,3}$

**Plot 2**

- $\Delta_{k,2} + \Delta_{k,3}$

**Plot 3**

- $\Delta_{k}$

Rank of distance travelled:

- Plot 1: 0, 50, 100, 150, 200, 250, 300, 350, 400
- Plot 2: 0, 50, 100, 150, 200, 250, 300, 350, 400
- Plot 3: 0, 50, 100, 150, 200, 250, 300, 350, 400

Symbols:

- X: uncen
- O: cnet
$z_2 = 1$. The plot of $\Delta_k$ (plot 3) gives no discernable influential cases.

We analysed the data further using the technique of the local Influence of Section 5.5.3. For these data the non-zero eigenvalues are 0.982 and 1.055. A plot of the absolute value of the elements of vector $\xi_{\text{max}}$ corresponding to the maximum eigenvalue $\xi_{\text{max}} = 1.055$, against rank of the observation is given in Figure 5.13. In Table 5.4 we give details of cases with the five largest absolute values of $\xi_{\text{max}}$. The plot shows clearly the five influential cases given in the Table, and these five cases account for most of the variability of the elements of $\xi_{\text{max}}$, since their sum of squares is 0.91 out of a total of 1.00. (Note: $\|\xi_{\text{max}}\| = 1$). For the 3 largest (corresponding to cases 34, 179 and 128) this sum of squares is 0.86. Table 5.4 also gives the $\hat{\beta}_{(k)}$ for variable 1 as approximated by (5.58). The effect of removal of these cases changes the $\hat{\beta}_1$ form 0.92 to that as tabulated, but in each case the value of $\hat{\beta}_2$ is hardly affected. Cases ranked 34, 128 and 197 are highly influential especially the last which increases $\hat{\beta}_1$ to an almost significant size.

<table>
<thead>
<tr>
<th>rank</th>
<th>$z_1$</th>
<th>$z_2$</th>
<th>censoring</th>
<th>$\xi_{\text{max}}$</th>
<th>$\hat{\beta}_1$</th>
<th>$\hat{\beta}_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>34</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.62</td>
<td>0.454</td>
<td>0.332</td>
</tr>
<tr>
<td>197</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-0.51</td>
<td>1.311</td>
<td>0.332</td>
</tr>
<tr>
<td>128</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.46</td>
<td>0.575</td>
<td>0.332</td>
</tr>
<tr>
<td>116</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-0.18</td>
<td>1.062</td>
<td>0.332</td>
</tr>
<tr>
<td>95</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-0.11</td>
<td>1.032</td>
<td>0.332</td>
</tr>
</tbody>
</table>

**TABLE 5.4**: Details of cases with the five largest absolute value of components of $\xi_{\text{max}}$ for the Brake Disc Data.

Note: Values of censoring indicator 0 = censored; 1 = uncensored.
Fig. 5.13 Plot of elements of $|l_{max}|$ against rank of distance travelled for BR disc data
5.6.2 Stanford Heart Transplant Data (Appendix B6).

To illustrate further for a model with covariates which are not indicator variables we analysed the Stanford Heart Transplant data set once again under the same model III (Table 4.4), that is regressing on covariates A, T and A2 where

\[
\begin{align*}
A &= (\text{age} - 41.9)/10 \\
T &= T5 \text{ mismatch score}/100 \\
A2 &= A \times A
\end{align*}
\]

with the linear predictor \( A_1 + T.\hat{\beta}_2 + A2.\hat{\beta}_3 \), the estimates of \( \hat{\beta}'s \) for the different methods used are given in Table 4.4 in Chapter 4. The Newton-Raphson method gives \( \hat{\beta}_1 = 0.508, \hat{\beta}_2 = 0.231 \) and \( \hat{\beta}_3 = 0.239 \) with standard errors 0.114, 0.188 and 0.073 respectively. \( \hat{\beta}_1 \) and \( \hat{\beta}_3 \) are significantly different from zero, while \( \hat{\beta}_2 \) is not, but we still keep the T5 in the regression equation to investigate its probably noisy effect on the influence measures. The estimates of \( \hat{\beta}_1 \) and \( \hat{\beta}_3 \) are approximately uncorrelated with \( \hat{\beta}_2 \) while their correlation is 0.5.

Table 5.5 gives the output of the estimation by the Newton-Raphson method and some details of some suspect cases.

In Figure 5.14 we give plots of \( \Delta_{\text{k,1}}, \Delta_{\text{k,2}} + \Delta_{\text{k,3}} \) and \( \Delta_{\text{k}} \) against the rank order of the observation for the A or age component. These show that the cases ranked 2 and 139 have the largest, but have opposite, influence. This is consistent with our finding in Section 3.8.2. These cases correspond to an individual aged 13 years and is dead 10 days after transplant (rank 2) and an individual aged 15 years but censored after
The N-R solution to COX REGRESSION in 5 iterations

ESTIMATES of the beta are

0.5075 ( 0.114 ), 0.2314 ( 0.188 ), 0.2385 ( 0.073 )

The log likelihood at max. likelihood est is -415.9050

The information matrix at this estimate is

\[
\begin{pmatrix}
0.102806E+03 & 0.119085E+01 & -0.801626E+02 \\
0.119085E+01 & 0.284578E+02 & -0.461432E+01 \\
-0.801626E+02 & -0.461432E+01 & 0.252583E+03
\end{pmatrix}
\]

Inverse is:

\[
\begin{pmatrix}
0.129262E-01 & 0.124645E-03 & 0.410469E-02 \\
0.124645E-03 & 0.352454E-01 & 0.683439E-03 \\
0.410469E-02 & 0.683439E-03 & 0.527428E-02
\end{pmatrix}
\]

<table>
<thead>
<tr>
<th>case</th>
<th>obs</th>
<th>cen</th>
<th>age</th>
<th>T5</th>
<th>estimates</th>
<th>diff.</th>
<th>( \Delta_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>10</td>
<td>1</td>
<td>56</td>
<td>276</td>
<td>0.4929</td>
<td>0.0146</td>
<td>0.9442</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.1797</td>
<td>0.0517</td>
<td>1.4536</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2307</td>
<td>0.0078</td>
<td>0.5502</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>1</td>
<td>13</td>
<td>149</td>
<td>0.5216</td>
<td>-0.0141</td>
<td>-3.2613</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2186</td>
<td>0.0128</td>
<td>0.2432</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2157</td>
<td>0.0228</td>
<td>6.8224</td>
</tr>
<tr>
<td>16</td>
<td>42</td>
<td>1</td>
<td>19</td>
<td>63</td>
<td>0.5250</td>
<td>-0.0175</td>
<td>-2.4928</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2480</td>
<td>-0.0166</td>
<td>-0.5312</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2301</td>
<td>0.0084</td>
<td>3.5961</td>
</tr>
<tr>
<td>34</td>
<td>65</td>
<td>1</td>
<td>45</td>
<td>168</td>
<td>0.5084</td>
<td>-0.0009</td>
<td>0.1465</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2178</td>
<td>0.0136</td>
<td>0.3991</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2413</td>
<td>-0.0028</td>
<td>-0.6956</td>
</tr>
<tr>
<td>40</td>
<td>86</td>
<td>1</td>
<td>12</td>
<td>128</td>
<td>0.5111</td>
<td>-0.0036</td>
<td>-1.7707</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2262</td>
<td>0.0052</td>
<td>0.0623</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2209</td>
<td>0.0175</td>
<td>4.7083</td>
</tr>
<tr>
<td>82</td>
<td>576</td>
<td>0</td>
<td>53</td>
<td>225</td>
<td>0.5243</td>
<td>-0.0168</td>
<td>-1.2518</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2851</td>
<td>-0.0537</td>
<td>-1.5167</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2452</td>
<td>-0.0067</td>
<td>-0.1034</td>
</tr>
<tr>
<td>106</td>
<td>1116</td>
<td>0</td>
<td>14</td>
<td>54</td>
<td>0.4983</td>
<td>0.0092</td>
<td>2.4193</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2171</td>
<td>0.0143</td>
<td>0.5016</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2587</td>
<td>-0.0182</td>
<td>-5.3965</td>
</tr>
<tr>
<td>139</td>
<td>2006</td>
<td>0</td>
<td>15</td>
<td>126</td>
<td>0.4918</td>
<td>0.0157</td>
<td>4.0122</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2432</td>
<td>-0.0118</td>
<td>-0.1777</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.2685</td>
<td>-0.0300</td>
<td>-8.7942</td>
</tr>
</tbody>
</table>

Note: values of censoring indicator 0 = censored; 1 = failures

Table 5.5 Details of some of the cases from the Stanford Heart Transplant data which have high components of . The values are computed using the Newton-Raphson 1-step approximation, on regression model III (see Table 4.4).
Fig. 5.14  Plots of $\Delta$'s for 'age' for Stanford Heart Transplant data against rank of survival time - model III

**plot 1**

$\Delta_{k,1}$

**plot 2**

$\Delta_{k,2} + \Delta_{k,3}$

**plot 3**

$\Delta_k$

[Legend: x uncen, o cen]
2006 days. Plots of the $\Delta$'s for the A2 (Figure 5.15) also pick up cases ranked 2 and 139 as the most influential. This is explained by the fact that in the Cox regression, $\hat{\beta}_1$ and $\hat{\beta}_3$ are correlated. The $\Delta$'s for suspect cases for this covariate are about twice the values of those for covariate A but of opposite sign so cases that are influential for plot on A are also reflected here. Plots of the $\Delta$'s for T5 score (Figure 5.16) are visually very similar to those for A but cases ranked 1 and 82 give the impression of high influence. Their $\Delta_k$ values are about 1.5 and -1.5 respectively. Since $\hat{\beta}_2$ is uncorrelated with the other $\hat{\beta}$'s, removal of cases 1 and 82 individually from the data set would change the value of $\hat{\beta}_2$, (using the approximation of (5.18)) by less than 0.06, about 0.35 of a standard error of $\hat{\beta}_2$ (0.188). These two cases therefore, do not influence the T5 regression parameter by any significant amount.

Removal of the case ranked 2 changes $\hat{\beta}_1$ by 0.014 while $\hat{\beta}_3$ is decreased by 0.023. These changes are not significant in terms of the standard errors of the parameters. Similarly, removal of case ranked 139 decreases $\hat{\beta}_1$ by 0.016 and increases $\hat{\beta}_3$ by 0.030.

The local influence analysis of Section 5.5.3 gives us plots of the absolute value of the components of eigenvector $\mathbf{e}_{\text{max}}$ against rank of the observations and age at transplant as in Figures 5.17 and 5.18 respectively. The value of the corresponding eigenvalue $\lambda_{\text{max}}$ of $B = \mathbf{e}^T A(\hat{\beta})^{-1} B$ is found to be 1.169, suggesting the existence of notable local sensitivity. Similar points found in the $\Delta$'s plots are also reflected in these plots, most notably cases ranked 2 and 139. Ordering the cases by their
absolute values of $\xi_{\text{max}}$ give the five largest values corresponding to cases all having ages less than 20 years (see Table 5.6). The sum of squares of their values of $\xi_{\text{max}}$ is 0.50. Since the total sum of squares of the $\xi_{\text{max}}$ components is 1, these five cases account for considerable local influence. Also in Table 5.6 we give the values of $\xi_{\text{max}}$ for the other cases having age less than 20 years. These values are all greater than or about the average value of a component of $|\xi_{\text{max}}|$, that is $(152)^{-\frac{1}{2}}$ or 0.081. The total sum of squares for the 10 cases with ages less than 20 years is 0.56. We note that of these ten components of $\xi_{\text{max}}$, five are negative and five are positive, showing that there is a lack of stability of the fitted model for age less than 20 years. Figure 5.18, the plot of elements of $|\xi_{\text{max}}|$ against age reveals quite plainly the large values of $|\xi_{\text{max}}|$ for age less than 20 years and the instability of the fitted model for these cases, that is the breakdown of the quadratic model in this area.
Fig. 5.15  Plots of $\Delta$'s for 'age*age' for Stanford Heart Transplant data against rank of survival time – model III

plot 1

plot 2

plot 3

uncen

 cen
Fig. 5.16 Plots of $\Delta$'s for 'T5' for Stanford Heart Transplant data against rank of survival time - model III

plot 1

plot 2

plot 3

$\Delta_k, \Delta_{k,2}, \Delta_{k,s}$

$\times$ uncen

$\circ$ cen

rank of survival time
Fig. 5.17  Plot of elements of $|\lambda_{max}|$ against 'rank' for Stanford Heart Transplant data – model III

Fig. 5.18  Plot of elements of $|\lambda_{max}|$ against 'age' for Stanford Heart Transplant data – model III
<table>
<thead>
<tr>
<th>Rank of case</th>
<th>Value of $l_{\text{max}}$</th>
<th>age</th>
<th>T5</th>
<th>censoring</th>
</tr>
</thead>
<tbody>
<tr>
<td>139</td>
<td>0.46</td>
<td>15</td>
<td>126</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>-0.37</td>
<td>13</td>
<td>149</td>
<td>1</td>
</tr>
<tr>
<td>40</td>
<td>-0.24</td>
<td>12</td>
<td>126</td>
<td>1</td>
</tr>
<tr>
<td>106</td>
<td>0.23</td>
<td>14</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td>96</td>
<td>0.20</td>
<td>19</td>
<td>198</td>
<td>0</td>
</tr>
<tr>
<td>16</td>
<td>-0.14</td>
<td>19</td>
<td>53</td>
<td>1</td>
</tr>
<tr>
<td>56</td>
<td>-0.13</td>
<td>19</td>
<td>102</td>
<td>1</td>
</tr>
<tr>
<td>95</td>
<td>0.10</td>
<td>20</td>
<td>112</td>
<td>0</td>
</tr>
<tr>
<td>105</td>
<td>0.08</td>
<td>18</td>
<td>25</td>
<td>0</td>
</tr>
<tr>
<td>109</td>
<td>-0.08</td>
<td>18</td>
<td>70</td>
<td>1</td>
</tr>
</tbody>
</table>

Note: Cases ranked 139, 2, 40, 106 and 96 are the cases with the largest component of $|l_{\text{max}}|$. Cases ranked 16, 56, 95, 105 and 109 are the other cases with age less than 20 years.

**TABLE 5.6**: Components of the eigenvector $l_{\text{max}}$ corresponding to $\lambda_{\text{max}} = 1.169$ for some cases for the Stanford Heart Transplant Data - model III.
6.1 INTRODUCTION.

In the previous chapters we have shown the usefulness of plots as diagnostic tools to assess the influence of cases in the proportional hazards model. Also in Sections 5.2.1 and 5.2.2 we have discussed the two basic residuals in the proportional hazards model, the Lagakos's and the Schoenfeld's residuals, and seen how the plots may present discernable patterns to suggest deviations from the formulated model. In this chapter we shall investigate further, the departure from the proportional hazards model when time constant explanatory variables are not reasonable. For this the Schoenfeld's residuals in Section 5.2.2 are of great importance in the analysis. The general proportional hazards model for time dependent covariates has a hazard function given in 5.01. However, Schoenfeld (1982) considers a misspecified model with hazard function given in (5.07), that is

$$h(z/t) = h_o(t) \exp(\beta^T \beta + \gamma x g(t))$$  \hspace{1cm} (6.01)

where $\gamma$ is the regression coefficient for the $xg(t)$, $x$ is a time constant explanatory variable and $g(t)$ is some function of the failure times. Thus the model with hazard as in (6.01) allows interaction between the covariate $x$ and the time. The score statistic to test $\gamma = 0$ is given in (5.08) and the Schoenfeld's residuals are given by (5.10) [or the equivalent representation]
given in Section 5.5), that is
\[ r_i = z_i - a_i \]
where \( a_i \) is as defined in Section 4.4.1, that is the conditional expectation of \( z_i \) over the risk set \( R_i \). We shall rewrite it here for ease of reference
\[ a_i = E[z_i/R_i] = \sum_{j \in R_i} z_j \lambda_j / \sum_{j \in R_i} \lambda_j \quad (6.02) \]
where \( \lambda_j = \exp(z_j^T \beta) \).

If \( x \) is the \( k \)th component of \( z \), the residuals at failure time \( t_i \) evaluated at estimate \( \hat{\beta} \) of \( \beta \) is
\[ \hat{r}_{ik} = z_{ik} - E[z_{ik}/R_i]. \]
Note that \( \sum_{i \in D} \hat{r}_{ik} = 0 \), \( k = 1,2,\ldots,p \)
where \( D \) is the set of failure times.

Schoenfeld (1982) has shown that if \( g(t) \) varies about zero slowly, then, the expectation with respect to the discrete distribution over the risk set at time \( t_i \), \( R_i \) (see Section 5.2.2),
\[ E(\hat{r}_{ik}) \approx g(t_i) A_i(\hat{\beta})_{kk} \quad (6.03) \]
where \( A_i(\hat{\beta})_{kk} \) is the variance term, that is the \( k \)th diagonal element of \( A_i(\hat{\beta}) \) defined in Section 5.2.2 such that
\[ \sum_{i=1}^{n} A_i(\hat{\beta}) = A(\hat{\beta}), \]
the observed information matrix (4.27) estimated at \( \hat{\beta} \). This suggests plotting \( \hat{r}_{ik} A_i(\hat{\beta})_{kk}^{-1} \) against \( t_i \) (Kay 1984), or some
predictable function of \( t_i \) in order to assess the form of \( g(t_i) \).
We will discuss these plots in our examples in Section 6.5.

6.2 VARIANCE-COVARIANCE OF THE SCHOENFELD RESIDUALS.

We note that from the definition of \( r_i \) and \( a_i \), we have

\[
E(r_i) = E[z_i / R_i] - a_i = 0
\]

\[
\text{Var}(r_i) = E[r_i r_i^T] = A_i(\beta)
\]

and

\[
\frac{\partial r_i}{\partial \beta} = - A_i(\beta),
\]

with expectation over the discrete distribution determined by
the risk set \( R_i \); and the \( r_i \) are uncorrelated (Cox, 1975).

Schoenfeld (1982) has established the asymptotic distribution
of the \( r_i \). Here we shall note his results. Expressing \( r_i \), the
residuals computed at the maximum likelihood estimate \( \hat{\beta} \) of \( \beta \),
in terms of \( \hat{r}_i \), using Taylor series, we have

\[
\hat{r}_i \sim r_i - A_i(\beta) (\hat{\beta} - \beta)
\]

Similarly we have the score statistic

\[
U(\hat{\beta}) \sim U(\beta) - A(\beta) (\hat{\beta} - \beta)
\]

where \( A(\beta) \) is the observed information matrix given in (4.27).

At maximum likelihood estimate \( \hat{\beta} \) of \( \beta \), we have

\[
\hat{\beta} - \beta = A(\beta)^{-1} U(\beta)
\]

\[
= A(\beta)^{-1} \sum_{k \in D} r_k
\]
since \[ U(\beta) = \sum_{k \in D} r_k \] from (4.26) and \( D \) is a set of all failures. Hence equation (6.04) can be written as

\[
\hat{r}_i \sim r_i - A_i(\beta) A(\beta)^{-1} \sum_{k \in D} r_k .
\] (6.05)

The variance covariance matrix can be expressed as

\[
\text{cov}(\hat{r}_i, \hat{r}_j) = E[\hat{r}_i \hat{r}_j^T] = E[r_i r_j^T] - A_i(\beta) A(\beta)^{-1} \times \sum_{k \in D, \ell \in D} E[r_k r_\ell^T] \times A(\beta)^{-1} A_j(\beta)
\]

Since the \( r_i \) are uncorrelated we have

\[
\sum_{k \in D} E(r_k r_k^T) = \sum_{k \in D} A_k(\beta) = A(\beta) .
\]

Hence

\[
\text{cov}(\hat{r}_i, \hat{r}_j) = \delta_{ij} A_i(\beta) - A_i(\beta) A(\beta)^{-1} A_j(\beta) \] (6.06)

where

\[
\delta_{ij} = \begin{cases} 0 & \text{if } i \neq j \\ 1 & \text{if } i = j \end{cases}
\]

6.3 OMNIBUS STATISTICS.

Omnibus goodness-of-fit statistics have been considered by several authors. With a lack of any a priori functional form for \( g(t) \), it can be approximated by a piece-wise linear function constant over a priori specified time intervals. This is the approach of Schoenfeld (1980), Anderson (1982) and Moreau et al.
Suppose that the time dependence is modelled by

$$h(t/z) = h_0(t) \exp\left\{ z^T \beta + \sum_{k=1}^{p} \theta_k(t) z_k \right\}$$  \hspace{1cm} (6.07)

where $z^T = (z_1, z_2, \ldots, z_p)$ and each $\theta_k(t)$ is linear over the partition $W_1, W_2, \ldots, W_q$ of the time domain into disjoint intervals, that is

$$\theta_k(t) = \gamma_{ik}, \quad t \in W_i \quad ; \quad k = 1, 2, \ldots, p.$$  

Moreau et al. (1985), Section 2, have shown that the non-zero part of the efficient score evaluated at $\hat{\theta}$ and $\gamma_{ik} = 0$, for testing all the pq $\gamma$'s are zero is equal to the pq×1 vector

$$\left( \sum_{i \in W_1} \hat{r}_i, \ldots, \sum_{i \in W_q} \hat{r}_i \right)^T$$  \hspace{1cm} (6.08)

where $\hat{r}_i$ is the Schoenfeld residuals in Section 6.1.

Let $\hat{r}_{i,s} = \sum_{i \in W_s} \hat{r}_i$, a pq×1 vector, the $s$th component of $(6.08)$, and

$$C_s = \sum_{i \in W_s} A_i \quad s = 1, 2, \ldots, q.$$  

Then the quadratic form score test statistic of Moreau et al. (1985), to test that the $\gamma$'s are zero, is given by

$$\sum_{s=1}^{q} \left( \hat{r}_{i,s} \right)^T C^{-1} \hat{r}_{i,s}$$  \hspace{1cm} (6.09)

which has a large sample chi-squared distribution on $p(q-1)$ degrees of freedom on the null hypothesis of no time dependence; $p$ degrees of freedom are lost from (6.09) since
\[
\sum_{i \in D} \hat{z}_i = 0.
\]

The score statistic above uses the covariance matrix of \( r_i \) rather than \( \hat{r}_i \), and is given by

\[
\text{cov}(r_i, r_j) = \delta_{ij} A_i.
\]

From (6.06), the expression for \( \text{cov}(\hat{r}_i, \hat{r}_j) \), it can be seen that \( \hat{r}_i \) are asymptotically uncorrelated since the inverse of the observed information matrix, \( A(\beta)^{-1} \), tends to the zero matrix as \( n \to \infty \), so that a quadratic form of the \( \hat{r}_i \) using the more accurate approximation of (6.06) for the \( \text{cov}(\hat{r}_i, \hat{r}_j) \) would be equivalent to (6.09) asymptotically.

An obvious modification of (6.09) is to choose the intervals \( W_1, W_2, \ldots, W_q \) of the time domain so that there are approximately equal number of failures in each set \( W_s \), \( s = 1, 2, \ldots, q \). This choice of partition is dependent upon the sample but independent of the \( \hat{r}_i \) since this choice of partition depends on failures but not their labels. Given that in most applications, the variance term \( A_i(\beta) \) remains relatively constant over varying \( t_i \), except for large \( t_i \), then the contributions to each term of the quadratic form (6.09) should be similar assuming \( \gamma = 0 \).

### 6.4 SMOOTHING THE SCHOENFELD RESIDUALS.

For the time dependent model of (6.01), denote the Schoenfeld residual for the variable \( x \) as \( r_i(x) \) and the variance \( A_i(x) \) which is the variance element in the \( pxp \) matrix \( A_i(\beta) \) corresponding to the \( x \) covariate. From (6.03)
This suggests taking \( \frac{r_i(x)}{A_i(x)} \) as an estimate of \( Yg(t_i) \) or plotting this quantity against \( t_i \) or some function of \( t_i \). This plot, however, in general, will appear very rough; for zero \( g(t) \), the \( \frac{r_i(x)}{A_i(x)} \) are approximately uncorrelated with variance \( 1/A_i(x) \). In the following we suggest several methods of smoothing these plots which may reveal the nature of the function \( g(t) \). These techniques assume a priori the nature of \( g(t) \) within a specified time interval, or window.

6.4.1 Constant \( g(t) \) in Window.

If we assume \( g(t) \) is slowly changing function smoothing the following way may bring about the nature of \( g(t) \). Assume \( g(t) \) is constant over a specified window. Then \( g(t) \) in that particular window, say \( W_j \), can be estimated from

\[
\hat{r}_j(x) = g(t_i) A_j(x) + \varepsilon_j ; \quad j \in W_i
\]

where the error \( \varepsilon_j \) has approximate variance \( A_j(x) \). Consider estimating \( g(t_i) \) in (6.10) by weighted least squares. Denoting \( S^k_i(x) \) as this estimate corresponding to covariate \( x \), then

\[
S^k_i(x) = \frac{\sum_{j \in W_i} \hat{r}_j(x)}{\sum_{j \in W_i} A_j(x)}
\]

where the window \( W_i \) contains a set of indices of failed cases having ranks whose values differ by no more than \( k \) from that of \( t_i \).
Plotting $S_i^k(x)$ against the rank of $t_i$ should indicate the nature of the non-constant $g(t)$. Again using the approximation from (6.06)

$$\text{cov}(\hat{r}_i, \hat{r}_k) = \delta_{ik} A_i(x)$$

we obtain the expression

$$\text{var}\{S_i^k(x)\} = \left\{\sum_{j \in W_i} A_j(x)\right\}^{-1}$$

(6.12)

when $\gamma = 0$. This can be used to give a tolerance interval for $S_i^k(x)$ for each value of $i$ when $\gamma = 0$.

6.4.2 Locally Linear $g(t)$.

An alternative smoothing procedure is to assume that $g(t_j)$ is locally linear in $\log(t_j)$ for $j$ in the window $W_i$, that is

$$g(t_j) = a_i + \beta_i \log t_j, \quad j \in W_i$$

(6.13)

[Note that functions of $t$ other than $\log$ could be considered in (6.13)]. Then from the model established in (6.10), and denoting $\hat{r}_j$ and $A_j$ for $\hat{r}_j(x)$ and $A_j(x)$ respectively, we have

$$\frac{\hat{r}_j}{A_j} = a_i + \beta_i \log t_j + e_j$$

(6.14)

where the error $e_j$ has approximate variance $\frac{1}{A_j}$.

The $a_i$ and $\beta_i$ for each $W_i$ can be estimated by the weighted least squares giving

$$\hat{\beta}_i = \frac{\left[\sum_{j \in W_i} y_j^* t_j^* - \left(\sum_{j \in W_i} y_j \sum_{j \in W_i} t_j^*\right) / V\right]}{\sum_{j \in W_i} (t_j^*)^2 - \left(\sum_{j \in W_i} t_j^*\right)^2 / V}$$
and 
\[ \hat{a}_i = y - \hat{\beta}_i t \]
where 
\[ v = \sum \lambda_j, \quad \hat{y} = \frac{\sum y_j}{v} \quad \text{and} \quad \hat{t} = \frac{\sum t_j}{v} \]

with 
\[ y_j^* = \frac{r_j}{\lambda_j} \quad \text{and} \quad t_j^* = \frac{1}{\lambda_j} \log(t_j) \]

The smoothed value of \( \frac{r_i(x)}{\lambda_i(x)} \) is given by

\[ L_{i}^{k}(x) = \hat{a}_i + \hat{\beta}_i \log(t_i) \]  \hspace{1cm} (6.15)

For hazard functions that converge for large \( t \), it might be more appropriate to consider

\[ g(t_i) = \alpha_i + \beta_i / (c + t_j) \], \hspace{0.5cm} j \in W_i \]  \hspace{1cm} (6.16)

where \( c \) is a positive number introduced to overcome the problems of small values of \( t_j \). The smoothed version is also given by (6.15) but \( (c + t_i)^{-1} \) replaces \( \log(t_i) \). The smoother \( L_{i}^{k}(x) \) is identical to the weighted scatter plot smoother of Hastie and Tibshirani (1986).

6.4.3 Accelerated Failure Time.

All explanatory variables can show time dependence of a similar form if survival time \( T \) is represented by the model

\[ \psi(T) = -z^T \beta + \varepsilon \]

where \( \psi(\cdot) \) is a monotone increasing function and \( \varepsilon \) is an error distribution with hazard function \( \lambda_{\varepsilon} \). Pettitt (1984) has shown the hazard function \( T \) can be approximated by
\[ h(t/z) = h_o(t) \exp \left\{ \frac{T}{\varepsilon} \theta \nu(t) \right\} \]  \hspace{1cm} (6.17)

where
\[ \nu(t) = \lambda^t \left\{ \phi(t)/\lambda^t \left( \psi(t) \right) \right\} \]
\[ h_o(t) = \psi'(t) \lambda^t \left\{ \psi(t) \right\} \]

If we take (6.17) as an exact relationship for \( h(t/z) \) then the expected value of \( z_i \) over the risk set \( R_i \) is
\[
E[z_{i/R_i}] = \frac{\sum_{j \in R_i} z_j \exp (z^T \theta \nu(t_j))}{\sum_{j \in R_i} \exp (z^T \theta \nu(t_j))} .
\]

Now letting \( \nu(t_i) \) vary about \( \overline{\nu} = 1 \) and expanding this mean about \( \nu(t_i) = 1 \) we obtain from
\[
r_i = z_i - \overline{a_i}
\]

The expected value over \( R_i \) is
\[
E(r_i/\nu) = E(z_i - a_i/\nu) = E(z_i/\nu) - a_i \tag{6.18}
\]
since \( a_i \) is fixed over the risk set \( R_i \) and is equivalent to \( E(z_i/\nu=1) \). Expanding \( E(z_i/\nu) \) about \( \nu = 1 \) we have the following approximation
\[
E(z_i/\nu) \approx E(z_i/\nu=1) + (\nu-1) \left[ \frac{\partial}{\partial \nu} E(z_i/\nu) \right]_{\nu=1}
\]
\[
= a_i + (\nu-1) A_i \theta
\]
Hence

$$E(\hat{r}_i) = (v(t)-1) A_i \hat{\beta}$$  \hspace{1cm} (6.19)

with $A_i$ evaluated at $\hat{\beta} = \hat{\beta}$.

If the model (6.17) appears plausible then $v(t)$ can be estimated by considering similar smoothers to $S$ (Section 6.4.1) and $L$ (Section 6.4.2). Assuming the residual $\hat{r}_j$ has approximate mean given by (6.19) and are approximately uncorrelated with variances equal $A_j$, then we can write for the window $W_i$

$$\hat{r}_j = (v(t_i)-1) A_j \hat{\beta} + \epsilon_j ; \quad j \in W_i$$

where $\epsilon_j$ is the error with variance equal to $A_j$.

Consider

$$G(v) = \sum_{j \in W_i} \varepsilon_j^T A_j^{-1} \varepsilon_j$$

$$= \sum_{j \in W_i} (r_j - (v-1) A_j \hat{\beta})^T A_j^{-1} (r_j - (v-1) A_j \hat{\beta})$$

$$\frac{\partial}{\partial v} G(v) = 2 \sum_{j \in W_i} \left[ - \hat{\beta}^T A_j A_j^{-1} (r_j - (v-1) A_j \hat{\beta}) \right].$$

Setting $\frac{\partial G(v)}{\partial v} = 0$, then the estimate of $(v-1)$, $C^k_i$, say, is given by

$$C^k_i = \frac{\hat{\beta}^T \sum_{j \in W_i} \hat{r}_j}{\hat{\beta}^T \left[ \sum_{j \in W_i} A_j \right] \hat{\beta}}$$  \hspace{1cm} (6.20)

The estimate $C^k_i$ is also found to be identical to the 1-step estimate of $(v(t)-1)$ over disjoint intervals defined on the time
domain for a model with the hazard function given by (6.17).

Assuming \( \theta \) fixed, and supposing

\[
v(t_j) = v_i \quad \text{for} \quad t_j \in W_i
\]

then for \( W_i \) containing failed cases only the log likelihood, (from Section 4.4.1) is

\[
L(v_i) = \sum_{j \in W_i} \left[ z_j^T \theta v_i - \log \sum_{k \in E_i} \exp(z_k^T \theta v_i) \right]
\]

and the score

\[
U(v_i) = \sum_{j \in W_i} \left[ z_j^T \theta - \frac{\sum_{k \in E_i} z_k^T \theta}{\sum_{k \in E_i} \eta_k} \right]
\]

where

\[
\eta_k = \exp(z_k^T \theta v_i), \quad \text{and}
\]

\[
\frac{\partial U(v_i)}{\partial v_i} = \sum_{j \in W_i} \left[ -\frac{\sum_{k \in E_i} (z_k^T \theta)^2 \eta_k}{\sum_{k \in E_i} \eta_k} + \frac{\sum_{k \in E_i} z_k^T \theta \eta_k}{\sum_{k \in E_i} \eta_k} \right]
\]

Solving for \( U(v_i) = 0 \) the 1-step estimate of \( v_i \) is thus

\[
U(v_i) = U(v_i=1) + (v_i-1) \left. \frac{\partial U(v_i)}{\partial v_i} \right|_{v_i=1}
\]

since \( v = 1 \) corresponds to the proportional hazard model. Then

\[
(v_i-1) = -U(v_i=1) \left/ \left. \frac{\partial U(v_i)}{\partial v_i} \right|_{v_i=1} \right.
\]
The numerator from (6.22) with \( v_i = 1 \) gives

\[
\sum_{j \in \mathcal{W}_i} \left[ z_j^T \hat{\beta} - \frac{\sum_{k \in \mathcal{R}_j} z_k^T \hat{\beta} \lambda_k}{\sum_{k \in \mathcal{R}_j} \lambda_k} \right]
\]

where \( \lambda_k = \exp(z_k^T \hat{\beta}) \). Since \( z_j^T \hat{\beta} = \hat{\beta}^T z_j \), the above equation can now be written as

\[
\hat{\beta}^T \sum_{j \in \mathcal{W}_i} \hat{r}_j
\]

Similarly the numerator from (6.23) with \( v_i = 1 \) gives

\[
\hat{\beta}^T \left( \sum_{j \in \mathcal{W}_i} A_j \right) \hat{\beta}
\]

where \( A_j \) is defined in Section 5.2.2, and evaluated at \( \hat{\beta} = \hat{\beta} \).

Thus the 1-step estimate of \( v(t) - 1 \) over a window \( \mathcal{W}_i \) is estimated equivalently by (6.20).

6.5 **EXAMPLES.**

6.5.1 **Simulated Data.**

We generated a sample of size 50 from the unit exponential distribution and 50 observations from a Weibull distribution with hazard function \( h(t) = a t^{a-1} \) for some value of \( a \). Labelling the samples 1, 2 respectively, the ratio of the hazard functions is given by

\[
\frac{h_2(t)}{h_1(t)} = a t^{a-1}
\]
The relationship between the variables from the distributions is

\[ T_2^a = T_1 \]

so that

\[ \log T_2 = \log T_1 \]

The mean of \( \log T \) over the two samples is

\[
E(\log T) = \frac{1}{2} E(\log T_2 + \log T_1)
= \frac{1}{2} \left( \frac{1}{a} + 1 \right) E(\log T_1) .
\]

Since \( T_1 \) is negative exponential with mean 1 then

\[
E(\log T_1) = -C
\]

where \( C \) is the Euler's constant, 0.5772.... Denoting

\[
E(\log T) \text{ by } \bar{x} = -(a+1) C/2a,
\]

we can write the right hand side of (6.24) as

\[
\exp \left[ \log a + (a-1) \bar{x} + \left( \log t - \bar{x} \right)(a-1) \right]
\]

with

\[
(a-1) \bar{x} = -\frac{(a-1)(a+1)C}{2a}
\]

\[
= -0.2885 \frac{(a^2-1)}{a}.
\]

Equation (6.24) becomes

\[
\frac{h_2(t)}{h_1(t)} = \exp \left[ \log a - \frac{0.2885(a^2-1)}{a} + (a-1)\left( \log t - \bar{x} \right) \right]
\]

(6.25)

Taking \( z \) in (6.01) to be the scalar indicator variable for sample 2 (from Weibull distribution) and \( x = z \), we have

\[
h_2(t) = h_0(t) \exp(\beta + \gamma g(t))
\]
and comparing with (6.25) we have

\[ \beta = \log a - 0.2885 \frac{(a^2 - 1)}{a} \]

\[ \gamma = (a - 1) \]

and \( g(t) = \log t - \bar{x} \).

We generated a number of samples and give here a typical result for \( a = 2 \). We fitted the proportional hazards model for the two-sample problem assuming no time dependence and obtained \( \hat{\beta} = 0.009 \) with estimated standard error 0.21. From the above, the "theoretical" value of \( \beta \) is 0.261. We computed the score test statistic (6.09) using 10 disjoint partitions of the time domain each containing 10 failures. The value of this statistic is 30.5 which is highly significant when referred to the chi-squared distribution with 9 degrees of freedom.

In Figure 6.1 we plot the Schoenfeld residuals, \( \hat{r}_i \), against the rank order of the failure times for this two-sample problem. For most times \( t_i \), except the largest with ranks greater than about 80, the residuals have values about \( \frac{1}{2} \) for \( z = 1 \) and \( -\frac{1}{2} \) for \( z = 0 \). It is difficult to assess much about the nature of \( g(t_i) \) from such a plot for these smaller \( t_i \) since it is obviously the average of these values over time intervals which gives this information. For the \( t_i \) values having ranks greater than 80 there is an excess of cases with \( z = 0 \) than with \( z = 1 \), giving \( a_i \) values which decreases from \( \frac{1}{2} \) to 0 as \( t_i \) increases. Finally for the largest \( t_i \)'s there are no cases left with \( z = 1 \) giving \( a_i = 0 \), hence \( \hat{r}_i = 0 \). Thus it is only for the larger \( t_i \) cases that some
Fig. 6.1  Plot of Schoenfeld residuals for the simulated data with $a=2$ against the rank of the observations.
Fig. 6.2  Plot of estimate, $S^k_1$, for the simulated data [ $a = 2$ ] against the log of the failure times.

Fig. 6.3  Plot of estimate, $L^k_1$, for the simulated data [ $a = 2$ ] against the log of the failure times.
deviation from the proportional hazards model is easily discernable from the plot of the residuals.

In Figure 6.2 we give the plot of $S^k_i(x)$ of (6.11) for $k = 5$ against $\log t_i$. There is some indication of the theoretical functional form of $\gamma g(t)$ namely $\log(t) + 0.433$ especially the unit slope for $\log(t)$. Figure 6.3 shows the plot of $L^k_i(x)$, (6.15) for $k = 10$ and also the best weighted least squares straight line fit to all the 100 residuals, which is very close to the theoretical line $\log(t) + 0.433$, having slope 0.969 and intercept equal to 0.4231. The shape of $L^k_i(x)$ in Figure 6.3 begins to capture something of the linear nature of $g(t)$ for smaller $t_i$ somewhat better than $S^k_i(x)$ in Figure 6.2.

6.5.2 Stanford Heart Transplant Data.

We consider again this data set which has been analysed quite extensively throughout this thesis. For this we use model III (Table 4.4), with three explanatory variables $z_1$, $z_2$ and $z_3$ in the regression model where the variables are as defined in Section 4.6.3. The estimates for the proportional hazards model applied on this data set are given in Table 4.4, Section 4.6.3.

In Figure 6.4 we plot the values of $S^k_i(z_j)$, $j = 1,2,3$ and $k = 2$ against the rank of $t_i$ in the sample of failures and censored observations, that is $n - N_i + 1$, where $N_i$ is the number of cases at risk at time $t_i$ and $n$ is the number of observations. With $k = 2$ these plots are not particularly smooth and together show little but random behaviour. The 95% pointwise tolerance interval for $S^2_i(z_j)$ is also given (dotted curve) for $j = 1,2,3$ and it can be
Fig 6.4 Plot of $S^k_1$ and 2*STD [asymptotic] for the Heart Transplant Data - model III

Variable 1 – AGE

Variable 2 – T5 score

Variable 3 – AGE**2
seen that most of the values are within the interval.

The score test statistic of (6.09) of time dependency based on 10 disjoint partitions of the time domain chosen to include about 10 failures in each partition is found to be 42.0 which is not significant referring to chi-squared with $3 \times (10-1) = 27$ degrees of freedom.

The time constancy of the covariates on the hazard function is also discernable from the plots of $L_i^k(z_j)$, (6.15), $j = 1, 2, 3$ in Figure 6.5, for $k = 5$. These plots have slope approximately equal to zero.

Figure 6.6 shows the plot of the smoother $C_i^k$ of (6.20) for $k = 5$, used to estimate $\nu(t) - 1$ in the model (6.17), where all the covariates show a similar form of time dependency. It gives a sinusoidal like curve with troughs at observations with ranks about 16, 60 and 110 estimating $\nu(t)$ to be zero, and peaks at ranks 36, 80 and 130, estimating $\nu(t)$ to be about 3.2 and 1.5 respectively. This appears interesting but is inexplicable.

6.5.3 British Rail Turbocharger Data. (Appendix 7a).

This example involves some reliability data obtained from British Railways concerning the distances travelled by power carriages until the failure of a turbo charger unit. So here the observations $t_1, t_2, \ldots, t_n$ are distances rather than times. A proportional hazards model with seven variables $z_1, z_2, \ldots, z_7$ was fitted. Here variables $z_1, z_2, \ldots, z_6$ are indicator variables referring to various design modifications of the turbo charger and $z_7$ is the number of days since 1st January 1982 until the
Fig 6.5 Plot of \( L_t^k \) \( [ k = 5 ] \) against log of failure time for the Heart Transplant Data – model III

**Variable 1 – AGE**

\[ L_t^k \]

**Variable 2 – T5 score**

\[ L_t^k \]

**Variable 3 – AGE**

\[ L_t^k \]
Fig. 6.6  Plot of the estimate, $c^k_i$, for $k = 5$ against the rank of the failure time for the Stanford Heart transplant data - model III
The estimated values of the parameters are as given in Section 5.2.2. The only variable showing any significant effect is \( z_7 \), which could be taken to represent technological change since 1/1/82, and the value \( \hat{\beta}_7 = -0.039 \) indicates improvement.

The plots of the Schoenfeld residuals for variables \( z_1, z_2, z_3 \) and \( z_7 \) against the rank order of the observations are shown in Figures 5.1, 5.2, 5.3 and 5.4 respectively in Section 5.2.2. For the indicator variables \( z_1 \) and \( z_3 \) the residuals are approximately \( \frac{1}{2} \) and \( -\frac{1}{2} \) for variable values 1 and 0 respectively, and the plots give two definite bands in these values for cases with ranks less than 700. An indication of time dependency is seen for the few cases whose ranks are greater than 700. The Schoenfeld plot for variable \( z_2 \) in Figure 5.2 has most of the residuals approximately zero, indicating the unbalanced values of \( z_2 \), and in this case more of \( z_2 \) with values zero. The plot also reflects some time dependence of the covariate for cases with ranks over 700. Plots for the other indicator variables are similar to this, however the plot for explanatory variable \( z_7 \) (Figure 5.4, Section 5.2.2) which is a quantitative variable shows no discernable time trend for the covariate, with the residuals scattered in the interval \([-5, 6)\).

We further investigate the distance dependency of the variables using the score test statistic of (6.09), with partition of the time domain chosen to include 21 failure distances in each interval, giving 22 intervals in total. The value of this statistic is found to be 142.0 which can be referred to the chi-squared distribution.
on $7 \times (22-1) = 142$ degrees of freedom. The value of the statistic indicates no distance dependency.

The distance constancy of $z_7$ or the "technological improvement" on the hazards function can be investigated by $S_i^k(z_7)$ or $L_i^k(z_7)$. In Figure 6.7, we plot $S_i^k(z_7)$ for $k = S^0$ against the rank of $t_i$ amongst failed cases and withdrawals. The main point to be drawn from this plot is that the estimate of $\gamma g(t)$ for distances $t_i$ having ranks equal to 330 to 455 is about $-0.1$, that is about 2½ times the value, $-0.039$, of $\hat{\beta}_7$. There are other fluctuations in the plot of $S_i^k(z_7)$ but these do not appear quite significant, except for values of $S_i^k(z_7)$ close to 0.1 for distances $t_i$ having ranks equal to about 90 to 130. Thus the plot suggests that the "technological improvement" is increased from the basic distance constant proportional hazards model for the middle distances travelled, corresponding to the ranks of observations between 330 and 455. It appears to be decreased somewhat for smaller distances travelled for observations having ranks between 90 and 130.

It is possible that the effect of distance dependence of $z_7$ can manifest as spurious effects in other variables. This is reflected in plots of $S_i^k(z_j)$ for $z_j$ other than $z_7$, which are similar to that for $z_7$. As an example see Figure 6.8, the plot of $S_i^k(z_3)$. Variable $z_3$ has little effect in the proportional hazards model, $\hat{\beta}_3 = 0.011$ with standard error 0.11, but Figure 6.8 shows that $S_i^k(z_3)$ is about five times $S_i^k(z_7)$. This non-zero distance dependent effect for $z_3$ could be accounted for by an interaction (distance dependent) effect on the hazard function. It appears more plausible to explain the plot for $z_3$ in terms of the distance
Fig. 6.7 Plot of estimate, $S^k$, for the BR turbocharge data 
[$k = 50$] against rank – for variable $z_7$

Fig. 6.8 Plot of estimate, $S^k$, for the BR turbocharge data 
[$k = 50$] against rank – for variable $z_9$
dependence for the variable $z_7$.

Since there is some similarity of shape amongst the plots of $S^k_{ij}(z_j)$ for different $j$, this suggests that model (6.17) is plausible; that is a common distance dependent effect $v(t)$ acting simultaneously on all in the covariates in the model. In Figure 6.9 we plot the values of $C^k_i$, (6.20), for $k = 50$ against the rank $t_i$. The plot is an estimate of $v(t) - 1$ (see Section 6.4.3). Here the plot shows that the estimate of $v(t)$ increases steadily from a near zero value for $t_i$ having ranks less than 100 to value of about 2.5 for $t_i$ having ranks equal to about 400. The estimate then decreases to about $\frac{1}{2}$ for the largest values of $t$. This suggests that the covariates have little effect on the hazard function of early failures, have their greatest effect for failures with ranks about 400 and then their effect decreases. Since we are considering the failure of a mechanical item, the early failures could be due to items which are supplied with faults which manifest themselves quickly, independently of the covariate values, much in the same way that the early part of the well known "bathtub" hazard function is formed. Again the effect of the covariates on the hazard function appears diminished for the longest distances travelled.

For a single explanatory variable in the model, models (6.01) and (6.17) are equivalent and are related thus

$$\left\{v(t) - 1\right\} \beta = \gamma g(t).$$

In our example here if the only non-zero $\beta_j$ is $\beta_7$, and noting the Figure 6.7 gives estimate of $\gamma g(t)$, Figure 6.9 an estimate of $v(t-1)$ and $\hat{\beta}_7 = -0.039$, then we can see the similarity between the estimates.
Fig. 6.9 Plot of the estimate, $C_i^k$, for $k = 50$ against the rank of the distance travelled for the BR turbocharged data
Chapter 7

Conclusions and Extensions

7.1 CONCLUSIONS.

The purpose of this chapter is to summarize some of the conclusions drawn from the work done in the preceding sections and to suggest areas for further research. The work in this thesis is primarily the assessment of influence in regression involving censored data, but we have concentrated the application to two models namely the scale mixtures (Chapters 2 and 3) and the Cox proportion hazards model (Chapters 4, 5 and 6). It is by no means that some of the generalities found cannot be extended to apply to other models. In the assessment of influence the emphasis is on the detection of influential cases rather than how to deal with them once they are found since the final judgement must necessarily depend on their context in the set of data.

Subsequent treatment has to depend on the prior assessment and reliability as required by the data provider.

This thesis also gives another exposure of the application of the EM algorithm (Dempster et al., 1977) in solving otherwise difficult likelihood equations for the parameter estimation. For the two models where we employ the EM algorithm, we found that it is less efficient in terms of computational speed. Even though, it has its merits. It reduces models to simple reweighted least squares which can be easily programmed in GLIM or such like packages. For the scale mixture models the EM algorithm gives a better estimate of the 1-step influence curves than that of the
Newton-Raphson. Applying the EM algorithm with gamma errors in
the reweighted least squares algorithm for the Cox proportional
hazards model, we find that the observed information matrix can
be approximated by $Z^T Z$, where $Z$ is the design matrix in reweighted
least squares with $Z$ column centred, that is the column of $Z$ sums
to zero. Also it is in the EM 1-step approximation of the influence
curve in the proportional hazard models that we found analytical
evidence to support Crowley and Storer's findings that the augmented
model (5.23) and the case deleted model (5.17) give similar results.

In Chapter 2 we analysed the discrete scale mixture using
entirely the EM algorithm. The results obtained, as shown by the
examples given in Section 2.8, are comparable to those found by
other authors for the same data sets. In Chapter 3, for the analysis
of continuous scale mixtures, the regression estimates found by the
EM algorithm are identical to those found by the Newton-Raphson
method; but the 1-step estimate, (3.65), of the influence curves
by the EM algorithm, for cases that are influential gives better
results. The Newton-Raphson 1-step approximation (3.67) is
equivalent to the quadratic approximation of the likelihood function
(see Kennedy and Gentle, 1980, Chapter 10). If the influence curves
are not too different from zero, and the likelihood function is
locally quadratic, then the 1-step estimator should be close to the
fully iterated value; but for cases that are influential, the
accuracy of the Newton-Raphson 1-step estimator is likely to be
lower. This is well illustrated in the likelihood distance plots
in Figure 3.2, where the likelihood distance from the 1-step over-
estimates that of the fully iterated by more than twice the value.
In summary, the EM 1-step estimate of influence curves for the scale mixture models gives better results than that of the Newton-Raphson.

The estimates of the regression parameters in the Cox proportional hazards model found by the EM algorithm using gamma errors in the resulting reweighted least squares are in agreement with those found by the widely used Newton-Raphson method and by the EM (Poisson errors) technique of Clayton and Cuzick, (1985a). However, in terms of computational efficiency, the EM (gamma) algorithm is not particularly encouraging especially when the proportion of censored cases is high. With few censored observations there appears to be little difference between the EM (gamma) and the EM (Poisson) techniques. This difference in the efficiency between the two techniques is due to the modelling scheme for the complete data specification. The EM (Poisson) takes into account the censorings in its specification while the EM (gamma) accommodates them as the incomplete data. Further EM (poisson) declares the censoring indicator as the y-variate in GLIM. Unlike the y-variate declared in EM (gamma), which is the $E[H(t_i)/R_i]$, the y-variate remains the same throughout the iterations. Full details of the comparison between these two techniques are given in Section 4.5.6.

In Chapter 5 we focussed our attention on the assessment of influence and diagnostic for the Cox proportional hazards model. Primarily it is the assessment of the influence curves, through some perturbation schemes. Three schemes to effect this perturbation were discussed. These are the case deletion perturbation, the
augmented model, Section 5.4 and case weights perturbation of Section 5.5. Approximations of the influence curves are obtained through these schemes, and in general, the influence curves and the subsequent scalar measures to quantify them are in agreement within some specific tolerance for the different methods. Here also are found analytical evidence through the 1-step EM algorithm with gamma errors applied to the augmented model of Section 5.4 that the augmented model give similar results as the case deletion model.

The 1-step EM approximation to compute the influence curves appear to be the most efficient and easy to program in GLIM. But due to the limited capability of GLIM to deal with matrices which are needed in further analysis, the influence curves are mainly computed from the A matrix of Section 5.5. Plots of A's or the approximation of the influence curves against rank order of the observations are very useful in the detection of influential cases. However the A's and the influence curves are vector valued and there is no natural ordering of multidimensional vectors on the basis of the influence. Scalar values that characterize these vectors are then needed. In this thesis these are in the forms of the Cook likelihood distance, LD (5.12) and for the local influence measures of Section 5.5.3, the components of the eigenvector \( \ell_{\max} \) corresponding to the maximum eigenvalue of the matrix obtained from (Section 5.5). Plots of LD and components of \( \ell_{\max} \) can also be revealing. Additionally by plotting the absolute values of the components of \( \ell_{\max} \) against the covariate values, not only can it show individual influential cases, general deviations
from the fitted model may also be detected. This is illustrated by the plot for the Stanford Heart transplant data, Figure 5.18.

The two basic residuals for the proportional hazards model are the Lagakos's residual, $e_i^*$, and the Schoenfeld's, $(p \times 1)$ vector valued residuals $\delta_i(z_i - a_i)$. These residuals can be used to assess the validity of the proportional hazard assumptions.

Lagakos's residuals have been discussed in brief in Section 5.2.1, (no further investigation is carried out in this work) and their plots can be used to investigate whether important explanatory variables have been omitted from the model. The Schoenfeld residuals on the other hand can be used to examine the time dependent effect. As discussed in Section 5.2.2, properties of the residuals can be investigated for the misspecified model with hazard function given by

$$h(z/t) = h_0(t) \exp(z^T \beta + \gamma g(t)).$$

Basic plots of these residuals are generally too rough as to reveal the nature of $g(t)$ clearly. In Chapter 6 we investigated and derived estimates of $g(t)$ (or the equivalent) in the misspecified model from the Schoenfeld residuals. The plots of these estimates are effectively smoothing the Schoenfeld residuals. The smoothed plots would generally give a more discernable pattern to suggest the nature of $g(t)$.

We summarize here by suggesting the following possible strategy for the analysis of the proportional hazards model.

a) Lagakos (1981) residuals $e_i^*$ should be plotted against explanatory variables included in the model to discover
whether there are any effects not accounted for by the fitted model. A smoothed plot of the residuals, smoothed by running medians or such like (see Velleman and Hooglin, 1981) would be better than a scatter plot.

b) Smoothed version of Schoenfeld (1982), vector residuals \( \mathbf{r}_i \) (see Chapter 6) should be plotted against the rank order of failure times for each component. Plots varying in a systematic way indicate some time dependence of the hazard function relationship with the covariable corresponding to that particular component of \( \mathbf{r}_i \).

c) The detection of influential cases can be done in two stages. First compute the matrix \( \Delta \) (see Section 5.5.1). Obtain the maximum eigenvalue, \( \lambda_{\text{max}} \), for the matrix \( B = \Delta^T \hat{\Sigma}^{-1} \hat{\Delta} \) (see Section 5.5.3), and the corresponding \( n \times 1 \) eigenvector \( \mathbf{e}_{\text{max}} \). If \( \lambda_{\text{max}} \) is greater than one, then there exist influential cases which are given by those cases having large absolute values of \( \mathbf{e}_{\text{max}} \). So either plot the absolute values of the components of \( \mathbf{e}_{\text{max}} \) against case number or order the cases by their absolute values of \( \mathbf{e}_{\text{max}} \).

d) The second stage of the detection of influential cases is given by considering the values of the sample influence curve, \( \text{SIC}_k \) or the approximation given by \( n \hat{\Delta}_k (\hat{\Delta}_k^{-1} \Delta_k \) (see Section 5.5.1). Each component of the \( p \times 1 \) vector \( \text{SIC}_k \) should be plotted against case number or cases ordered by their \( \text{SIC}_k \) values for each component. The influence of cases can be
associated with particular explanatory variables. For the overall detection of influential cases, plots of the likelihood distance $LD$, (5.12), against case number could prove useful.

7.2 EXTENSIONS

In this thesis we have analysed and explored the influence of several data sets which are assumed to contain no nested structures which could lead to dependent observations. The observations are considered to be independent here. A line of approach for future research is to consider a model with possible nested structures. The random effects and repeated measures models (see, for example, Pettitt, 1986) are possible candidates. A simple random effects is defined by

$$Y_{ij} = \beta_j + \varepsilon_{ij} \quad (7.01)$$

while a mixed effects model is given by

$$Y_{ij} = \mathbf{x}_{ij}^T \gamma + \beta_j + \varepsilon_{ij} \quad (7.02)$$

where the $\beta_j$ are independent random variables representing the variability between "blocks" and independent $\varepsilon_{ij}$ represent the variability among the observations within blocks, $\gamma$ is a fixed $p \times 1$ regression parameter corresponding to the covariate vector $\mathbf{x}_{ij}$, and the $Y_{ij}$ is the log of the failure times $T_{ij}$. The $\beta_j$ and $\varepsilon_{ij}$ are assumed to be independent of one another and distributed as normal random variables with means zero and variances $\sigma_{\beta}^2$ and $\sigma_{\varepsilon}^2$ respectively. Here "block" could represent
a group of siblings when medical data are considered; or when industrial data are considered, "block" could represent repeated failures on the same item (although here there might be some serial correlation) or failures on items from the same batch.

To robustify such models the technique of Chapters 2 and 3 may be explored defining mixture distributions for the $\beta_{ij}$ and $\epsilon_{ij}$ distributions. Thus $\beta_{ij}$ could be taken to have a Student-$t$ distribution with $\nu$ degrees of freedom and $\epsilon_{ij}$ Student-$t$ with $\eta$ degrees of freedom. These outlier prone distributions allow for outlying blocks and outlying observations within a block.

The proportional hazards equivalent of (7.01) and (7.02) have been considered by Clayton and Cuzick (1985b), however, the canonical model represented in (7.02), $y_{ij}$ is the monotone transformation of the failure time $t_{ij}$, and $\beta_{ij}$ and $\epsilon_{ij}$ are such that $\exp(-\beta_{ij})$ is distributed as a gamma random variable with parameters $(k, k)$, say, and $\exp(\epsilon_{ij})$ is distributed exponentially with mean 1. Matched pair and sibling problems have been investigated by these authors.

Another possible area of research for the extension of the technique used in Chapters 2 and 3 is when there are multiple failures, that is individuals may fail due to several distinct types or causes. Problems arising in the analysis of such data are commonly referred to as competing risk problems. The mixture model for this multivariate failure time may be formulated such

$$(Y_1, Y_2, \ldots, Y_k) / Q = q \sim MN(\mu, \Sigma/q)$$

where $Y_j = \log T_j$ and $T_1, \ldots, T_k$ are the $k$ failure times due
to distinct causes and \( \text{MN}(\mu, \Sigma) \) is the multivariate normal distribution with mean \( \mu = (\mu_1, \mu_2, \ldots, \mu_k)^T \) and \((k \times k)\) covariance matrix \( \Sigma \).

For the proportional hazards model, areas in the influence diagnostic are still open for further investigation. The influence of covariates in the model may be examined further. We have used as an ad hoc rule that the maximum eigenvalue, in the local influence analysis, greater than one signifies the existence of influential cases. The relevance of the value one needs to be interpreted, and the eigenvectors corresponding to the rest of the eigenvalues can be investigated, to see whether they are useful or merely contain noise. Research on other scalar measures to quantify the influence curves that reflect global influence is useful.

In concept, the time dependence covariate, \( z(t) \), for the proportional hazards model is the extension of the time constant, \( z \) of the basic proportional hazards model of (4.07), replacing the \( z \) by \( z(t) \). All the statistics resulting from the analysis of model (4.07) can be extended to the general time dependent model (5.01). We have utilized Schoenfeld residuals to investigate the nature of this time dependency for models defined in (6.01) and (6.17), proposing a number of estimates relevant to these models. These estimates can be investigated to provide further evidence for the validity of the proportional hazards assumption.
References


Annals of Statist., 10, 2, 497-484.


PETTITT, A.N. and BIN DAUD, I. (1986), Fitting Student-t distributions to censored data. GLIM Newsletter, 13, 22-27.


Appendix A


Let \( A \) be a \( p \times p \) and rank \( p \) symmetric matrix, and suppose that \( a \) and \( b \) are \( q \times p \) rank \( q \) matrices. Then, provided that the inverses exist,

\[
(A + a^T b)^{-1} = A^{-1} - A^{-1} a^T (I + bA^{-1} a^T)^{-1} bA^{-1}.
\]
Appendix B

The following data sets are used for examples and discussion at various places in the thesis.

B1: Darwin Data.

The data represent the differences in heights of 15 pairs of self-fertilized and crossed-fertilized plants grown on the same plot. It was originally analysed by Box and Tiao (1968). The observations are as follows:

\[-65 \quad -48 \quad 66 \quad 8 \quad 14 \quad 16 \quad 23 \quad 24 \quad 28 \quad 29\]

\[41 \quad 49 \quad 56 \quad 60 \quad 75\]

The following table gives the boiling points of water in °F and $100 \times \log$ of pressure (pressure in inches of mercury) in 17 locations in the Alps and Scotland, collected by J.D. Forbes in the 1850's.

<table>
<thead>
<tr>
<th>Case</th>
<th>deg. F</th>
<th>100*log(P)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>194.5</td>
<td>131.79</td>
</tr>
<tr>
<td>2</td>
<td>194.3</td>
<td>131.79</td>
</tr>
<tr>
<td>3</td>
<td>197.9</td>
<td>135.02</td>
</tr>
<tr>
<td>4</td>
<td>198.4</td>
<td>135.55</td>
</tr>
<tr>
<td>5</td>
<td>199.4</td>
<td>136.46</td>
</tr>
<tr>
<td>6</td>
<td>199.9</td>
<td>136.83</td>
</tr>
<tr>
<td>7</td>
<td>200.9</td>
<td>137.82</td>
</tr>
<tr>
<td>8</td>
<td>201.1</td>
<td>138.00</td>
</tr>
<tr>
<td>9</td>
<td>201.4</td>
<td>138.06</td>
</tr>
<tr>
<td>10</td>
<td>201.3</td>
<td>138.05</td>
</tr>
<tr>
<td>11</td>
<td>203.6</td>
<td>140.04</td>
</tr>
<tr>
<td>12</td>
<td>204.6</td>
<td>142.44</td>
</tr>
<tr>
<td>13</td>
<td>209.5</td>
<td>145.47</td>
</tr>
<tr>
<td>14</td>
<td>208.6</td>
<td>144.34</td>
</tr>
<tr>
<td>15</td>
<td>210.7</td>
<td>146.30</td>
</tr>
<tr>
<td>16</td>
<td>211.9</td>
<td>147.54</td>
</tr>
<tr>
<td>17</td>
<td>212.2</td>
<td>147.80</td>
</tr>
</tbody>
</table>
B3: Motorette Data.

The table gives the results of temperature accelerated life tests on electrical insulation in 40 motorettes, originally reported by Crawford (1970). Ten motorettes were tested at each of four temperatures. Testing was terminated at different times for each temperature, which are the censored times for motorette still on test without failure. The failure times (in hours) are as tabulated.

<table>
<thead>
<tr>
<th>150</th>
<th>170 cen</th>
<th>190 cen</th>
<th>220 cen</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 units censored at 8064</td>
<td>1764 1</td>
<td>408 1</td>
<td>408 1</td>
</tr>
<tr>
<td></td>
<td>2772 1</td>
<td>408 1</td>
<td>408 1</td>
</tr>
<tr>
<td></td>
<td>3444 1</td>
<td>1344 1</td>
<td>504 1</td>
</tr>
<tr>
<td></td>
<td>3542 1</td>
<td>1344 1</td>
<td>504 1</td>
</tr>
<tr>
<td></td>
<td>3780 1</td>
<td>1440 1</td>
<td>504 1</td>
</tr>
<tr>
<td></td>
<td>4860 1</td>
<td>1680 0</td>
<td>528 0</td>
</tr>
<tr>
<td></td>
<td>5196 1</td>
<td>1680 0</td>
<td>528 0</td>
</tr>
<tr>
<td></td>
<td>5488 0</td>
<td>1680 0</td>
<td>528 0</td>
</tr>
<tr>
<td></td>
<td>5488 0</td>
<td>1680 0</td>
<td>528 0</td>
</tr>
</tbody>
</table>

Note: values of censoring indicator 1 : failures
0 : censored.
**B4: Carcinogenesis Data.**

The following table from Kalbfleisch and Prentice (1980), originally from Pike (1966), gives the times (in days) from insult with carcinogen DMBA to mortality from vaginal cancer in rats. Two groups were distinguished by a pretreatment regime.

<table>
<thead>
<tr>
<th>Group 1</th>
<th>143</th>
<th>164</th>
<th>188</th>
<th>188</th>
<th>190</th>
<th>192</th>
<th>206</th>
<th>209</th>
<th>213</th>
<th>216</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>220</td>
<td>227</td>
<td>234</td>
<td>246</td>
<td>265</td>
<td>304</td>
<td>216*</td>
<td>244*</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Group 2</td>
<td>142</td>
<td>156</td>
<td>163</td>
<td>198</td>
<td>205</td>
<td>232</td>
<td>232</td>
<td>233</td>
<td>233</td>
<td>233</td>
</tr>
<tr>
<td></td>
<td>233</td>
<td>239</td>
<td>240</td>
<td>261</td>
<td>280</td>
<td>280</td>
<td>296</td>
<td>296</td>
<td>323</td>
<td>204*</td>
</tr>
<tr>
<td></td>
<td>344*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* indicates censored observation.
B5 : Veteran's Administration Lung Cancer.

(Source Kalbfleisch and Prentice, 1980). In this clinical trial, males with advanced inoperable lung cancer were randomized to either a standard or test chemotherapy. The primary end point for therapy comparison was time (in days) to death. Only 9 out of 137 survival times were censored. The data is recorded in the attached floppy disc in 'vlung.dat'. It includes the following information on the covariables:

\[ z_1 \] : Performance status (Karnofsky rating).

\[ z_2 \] : Age in years of patient.

\[ z_3 \] : Months from diagnosis to entry.

\[ z_4 \] 
\[
\text{squamous} = 1 \quad , \quad \text{otherwise} \ 0
\]

\[ z_5 \] 
\[
\text{tumor cell type} \quad \text{small} = 1 \quad , \quad \text{otherwise} \ 0
\]

\[ z_6 \] 
\[
\text{adenocarcinoma} = 1 \quad , \quad \text{otherwise} \ 0
\]

\[ z_7 \] : Treatment \ [\text{standard} = 1 \ ; \ \text{test} = 0] 

\[ z_8 \] : Prior therapy \ [\text{yes} = 10 \ , \ \text{no} = 0 ]
B6 : Stanford Heart Transplant data.

(Source Miller and Helpern, 1982).

The Stanford heart transplantation program was begun in October 1967. By February 1980, 184 patients had received heart transplants. A few of these had multiple transplants. The survival times, in days, uncensored and censored for the 184 patients are displayed in Table 1 in Miller and Helpern, 1982. The Table also includes the ages of patients at transplant and the T5 mismatch scores which measure the degree of tissue incompatibility between the initial donor and recipient hearts with respect to HLA antigens. The regression analyses reported at various parts of the thesis only consider 152 patients with complete records and who survived for at least 10 days after transplant. This data set is in the floppy disk in segment "hx152.dat".
B7 : British Railways Data.

The data is from British Rail, Derby. Two sets of data are considered in the examples. They are:

B7a : Turbocharger data.

The data is from the reliability study concerning the "life-times" of turbocharger units in power carriages. The "life-times" is the distance travelled by the carriage until the failure of a turbocharger unit. Seven covariates are included, six of which are indicator variables referring to various design modifications of the units and the non-indicator variable (referred as variable 7) is the number of days since the 1st January 1982 until the turbocharger was fitted. Segment "bturbo.dat" in the floppy disc list this data set.

B7b : Disc data.

This data set concerns the study of the reliability of the brake discs on British Railways High Speed Trains, in the Western Region of England. The detailed design of the experiment is not discussed here. The data gives the distance travelled by the railway carriages until the disc brake unit failed in some way. Two indicator covariates are included. They refer to the type of material used (variable 1) and fixing position of the unit (variable 2). The data contains 357 cases of which 98 are failures while the others are withdrawals from the study for some reason. The segment "bdisc.dat" in the floppy disc contains this data set.
Appendix C

This appendix describes briefly the computer listing referred to in the various parts of this thesis. The programmes are listed in the attached floppy disc which can be read by any IBM/IBM compatible PC, or PC/AT microcomputer running under PC-DOS operating system.

C1 : FINMIX.GLM.

GLIM macros used for the finite mixture models and an example of run programme (the heart transplant data)

C2 : a) CONMIX.GLM.

GLIM listing to analyse the continuous mixture models with Student-t error distributions. An example run programme (the motorette data is also included).

 b) FORMIX.FOR

Fortran listing, the equivalent of C2(a) for the analysis of the continuous mixture models with Student-t error using Newton-Raphson method (and adapted for 1-step influence).

C3 : PROHZ.FOR.

Fortran listing to compute the estimates of the regression coefficients in the Cox proportional hazards model by Newton-Raphson Method.
C4 : PROGM.GLM.

GLIM programme to compute the estimates of the regression coefficients in the Cox proportional hazards model using EM algorithm with gamma error.
CONTENTS OF THE FLOPPY DISC

The attached floppy disc contains the following segments:

vlung.dat [see appendix B5]
hx152.dat [see appendix B6]
b turbo.dat [see appendix B7a]
b disc.dat [see appendix B7b]
f inmix glm [see appendix C1]
conmix glm [see appendix C2a]
formix for [see appendix C2b]
prohz for [see appendix C3]
progm glm [see appendix C4]