OUTCOME MEASUREMENT ERROR IN SURVIVAL ANALYSIS

Ph.D. Thesis

by

WILLIAM MARK HIRST

September 1998

OUTCOME MEASUREMENT ERROR IN SURVIVAL ANALYSIS

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy

by

WILLIAM MARK HIRST

September 1998

To

Noreen, my family and my friends

Outcome Measurement Error in Survival Analysis

William Mark Hirst

Abstract:

Introduction: Measurement error is a well known statistical problem. Covariate error in survival data has received much attention but outcome error has not. Cancer registration data may suffer from a "window of uncertainty" in the date of diagnosis of a given individual which may lead to error in survival time (the "outcome" variable) and hence bias in modelling survival. True and observed survival are related by a measurement model.

Aims: The aims of this thesis are: (i) to examine the effect of survival error on the parameter estimates of a Cox regression model, (ii) to develop methodology to correct for survival error for the Cox model, and (iii) to demonstrate how developed methodology can be of value to cancer registries.

Methods: An experiment is undertaken to assess the effect of a measurement model on the parameters of a Cox regression. A method for weighting risk sets in the likelihood of a Cox model in order to achieve a superior partial likelihood is derived. This is based on approximations for tied data and the incorporation of a measurement model for outcome error. The effectiveness of this likelihood is examined using the data from the initial experiment.

Data: The data are all primary lung malignancies in the Merseyside and Cheshire region for the years 1974 - 1993 inclusive. The total number of patients was 42,222. A measurement error analysis is considered for 500 of the patients diagnosed in 1993.

Results: The effectiveness of the Cox likelihood in the presence of error is dependent on the size of the error, the size of the relative risk and characteristics of the true survival times. For large error bias is towards the null hypothesis. The correction procedure is effective in reducing the bias when it is of considerable size, but parameter estimates have larger variance. The variance of bias is reduced if its mean is not large. For cancer registry data, survival error due to diagnostic uncertainty is unlikely to be a problem when estimating the effect of covariates, but predicted survival is reduced when a correction for bias is employed.

Conclusion: This thesis introduces a correction procedure for outcome error in the Cox model which shows considerable promise in resolving the problem of bias. A prototype analysis for cancer registration data demonstrates that the procedure is applicable to real data problems.

Acknowledgements

I would like to express my thanks to all the people who have helped me during this project, particularly my supervisors Professor Deborah Ashby and Dr. Lyn Williams for their guidance, encouragement and support. I am grateful to Professor Raj Bhansali, Dr. Paula Williamson, Professor Phil Brown and Dr. Dave Downham for their help. I also express my thanks to Paul Monaghan and Simon Fear for extensive discussion and input, often into the early hours. I would like to thank all the staff of the Merseyside and Cheshire Cancer Registry.

Personal thanks go to Noreen, for putting up with and encouraging me during my studies, and my family for their support. I must also thank Rob for being an understanding friend and flatmate during the more stressful times.

I am also grateful to the North West Cancer Research Fund for the sponsorship of the project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Analysis of survival data has seen many advances in recent years, and the Cox proportional hazards model is commonly used to examine the effect of explanatory variables on the survival time of an individual.

Much attention has been paid to the fact that explanatory variables, particularly in epidemiology, may be measured with error. There is a wealth of literature documenting the fact that such error, depending on its structure, can lead to severe bias in the estimation of the effect of covariates. The Cox model has not escaped attention, and authors have been able to ascertain the extent of the bias, and suggest correction procedures to deal with it.

Little attention has been paid to error in the outcome variable, namely survival. There is in fact less scope for such errors in many analyses, such as clinical trials. Some authors have discussed particular types of outcome error for survey data.

We examine a previously unexplored statistical problem, namely uncertainty in diagnosis recording in cancer registration data. Due to the fact that different patients receive different levels of diagnosis that can occur weeks apart, it is important that some patients do not have artificially extended survival in comparison with others. If this is the case, it is desirable to be able to incorporate such uncertainty in a Cox analysis.

Estimation in the Cox model is achieved via a procedure dependent on the order of failure times, which will be clouded by noise in the recording of times. One such problem is tied data, and approximations have been suggested when failures are grouped due to rounding. The recording of diagnosis for cancer patients is not however a grouped rounding problem. We examine how Cox model estimates are affected by more general outcome measurement error.

The problem of bias leads us to desire a correction procedure. A simple and generally applicable approximate correction for bias is derived in this thesis. This is then applied to diagnosis error in cancer registration.

1.2 Structure of the Thesis

Chapter 2 introduces the reader to the system of cancer registration in the U.K. and discusses briefly where measurement error might be coming from. The epidemiology of lung cancer is discussed, with emphasis placed on recent incidence reports from the Merseyside and Cheshire cancer registry. Finally the lung cancer data used for analysis in this thesis is introduced.

Chapter 3 introduces survival analysis, and reviews existing non-parametric, semi-parametric and fully parametric methodology. Of particular importance is the Cox proportional hazards model - the most widely used model in survival analysis. Methodology is illustrated using the lung cancer data. Initial analyses are carried out under the assumption of no measurement error.

Chapter 4 is a review of methodology available for covariate and outcome error. The Cox proportional hazards model is covered in detail.

A new method for the incorporation of measurement error in the partial likelihood of the Cox model is introduced in chapter 5. This is an extension of approximate likelihoods used for tied data, and is closely related to the Efron likelihood for ties. In addition experiments are undertaken to establish the effect of measurement error on the parameter estimates of the Cox model. Data from these experiments are used to verify the correction procedure, and show that bias due to measurement error is virtually eliminated.

The new methodology was motivated by concern over recording of date of diagnosis by cancer registries. Chapter 6 examines more deeply how the measurement error arises. Once the nature of the measurement error is established a correction is undertaken using the new methodology introduced in chapter 5.

The conclusion of this work is given in chapter 7 and directions and ideas for further work outlined.

1.3 Notation and Abbreviations

Throughout the thesis the following notation is used:

T: True survival time

S: Observed survival time

X: True covariable

Z: Observed covariable

U: Measurement error relating the true survival time/covariate to the observed survival time/covariate

Notation used for survival analysis is introduced as needed in chapter 3.

1.3.1 List of Medical / Health Service Abbreviations

NHS: National Health Service

MRCR: Mersey Regional Cancer Registry

MCCR: Merseyside and Cheshire Cancer Registry

SRR: Standardised Registration Ratio

IACR: International Association of Cancer Registries

UKACR: United Kingdom Association of Cancer Registries

ONS: Office of National Statistics

WHO: World Health Organisation

ICD: International Classification of Disease

Chapter 2

Introduction to Cancer Registration and Lung Cancer

2.1 Introduction to Cancer and Cancer Registration

In England and Wales, cancer affects approximately 1 in 3 people during their life and is responsible for about 1 in 4 deaths. Rates of incidence are usually higher in men and the majority of cancer sufferers are over 65. As a rule teenagers are least likely to be afflicted by cancer. In the region of 1 in 200 cases are aged 15 or under. The age distribution of specific cancers may however vary (UKACR, 1994) .

In general there are a large number of established risks and causes of cancer. Commonly known examples are smoking and lung cancer or exposure to the sun and skin cancer. Occupational and demographic factors may increase risk and diet has been associated with certain cancers (for instance, a western diet may

explain some of the difference in breast cancer incidence between Europeans and Asians). Increased risk may have genetic origins. Other suspected associations are viral, for instance infection from the Epstein-Barr virus is associated with certain lymphomas (Vessey and Gray, 1985). However, there is scope for much more research investigating the causes of cancer with the aim of prevention. Until prevention is possible, it is important to examine actual cases of cancer in order to establish prognostic factors associated with better survival.

In terms of the area covered by the Mereyside and Cheshire Cancer Registry (MCCR) recent incidence reports demonstrate a number of facts. Urban areas such as Liverpool compare unfavourably with more affluent areas of Cheshire such as Macclesfield (Hussey and Ashby, 1990; Youngson *et al.*, 1991; Youngson *et al.*, 1992). Most cancers are related to poor lifestyle indicators, and the Standardised Registration Ratio's (SRR's) for these cancers tend to be high across the region. An SRR provides a comparison of relative registration with the wider England and Wales population adjusted for differences in age and sex. The situation is more favourable for breast cancer, for which the Mersey region is in line with government targets (Williams *et al.*, 1994). Overall it is evident that the Mersey region compares unfavourably with the rest of England and Wales (see table 2.1). Although now nearly a decade later, big changes from the 1980's in registration patterns are unlikely. The incidence report for the years 1990-1995 is due for publication in late 1998.

Table 2.1: Standardised Registration Ratios 1986-1990 for Mersey RHA (All Cancers)

District	Males (95 % C.I.)	Females (95 % C.I.)
Chester	125 (120,131)	129 (124,135)
Crewe	91 (88,95)	91 (88,95)
Halton	127 (121,134)	131 (124,137)
Liverpool	147 (143,150)	139 (135,142)
Macclesfield	109 (104,114)	115 (110,120)
South Sefton	125 (120,131)	122 (117,127)
Southport and Formby	124 (119,130)	131 (125,137)
St. Helens & Knowsley	128 (124,132)	122 (118,126)
Warrington	119 (114,124)	124 (119,130)
Wirral	131 (128,135)	126 (123,130)
Mersey Region	127 (125,128)	125 (124,127)

England and Wales = 100
Source: Youngson et al 1992

2.2 History of Cancer Registration

Cancer registration can be traced back to the early years of the twentieth century, when in approximately 1900 Germany unsuccessfully attempted a cancer "census" based on questionnaires to doctors. The first registry was set up in Hamburg in 1929 in order to gain a superior follow up to patient care (Wagner, 1991) . The Mersey Regional Cancer Registry (MRCR) was founded in 1944, and emerged from the Liverpool Cancer Control Organisation set up in 1939 as a result of the Cancer Act (MRCR, 1990) . In 1994 a merger between the Mersey and North West Health Regions meant the creation of the North West Regional Health Authority, covering a population of about 6.6 million people. The MRCR has since been called the Merseyside and Cheshire Cancer Registry (MCCR), covering seven District Health Authorities and a population of 2.4 million people.

In recent history a number of organisations on a national, European and international basis have been created to increase the potential of cancer registration. The MRCR was a founder member of the International Association of Cancer Registries (IACR) set up in Tokyo in 1966. This now boasts over 200 members from 80 countries and has its headquarters in Lyons, France. Also, the European Network of Cancer Registries was formed in 1990. The United Kingdom Association of Cancer Registries (UKACR) was created as an umbrella organisation in 1992. Other groups exist in order to share experiences and discuss methodology, practices and training. These meet several times a year and include the Cancer Registries Consultative Group, the Cancer Registries I.T. group and the Cancer Surveillance Group, who together with subgroups address issues of education,training and data quality (UKACR, 1994). Thus in its relatively short history cancer registration has expanded rapidly, and with growing advances in technology is a valuable worldwide resource for research.

2.3 Uses of Cancer Registration

The UKACR report in their handbook (UKACR, 1994), *"The cancer registration system is the most powerful tool available for the epidemiological study of cancer"* . The reasons for this are clear. As a population based register that includes both fatal and surviving cases from a long period of time, with a wide range of data sources and items, the cancer registry can provide data of minimal selection bias (Jensen and Storm, 1991) for the research worker in a number of areas. Registries are able to provide information on long term incidence and survival,

identifying any improvement or lack of it over time (when the data are analysed with care). In particular the registry offers an unrivalled opportunity for the comparison of regional differences in cancer. As Jensen and Storm argue, a well presented incidence report *"serves an important function as part of the health information of a country or region"*.

Other uses suggested in the Cancer Registry handbook are summarised below. Through the study of incidence and survival a greater insight into the areas that require resources can be gained, and thus the registry can help in the allocation of funding. For the evaluation of screening programmes the registries can provide data for retrospective studies and assess the effectiveness of a screening programme (this may not be a trivial exercise due to the nature of survival measurement). In terms of aetiological studies (i.e. those of a risk factor) the cancer registry can be used to study post treatment effects, and usually together with other sources (such a medical notes, biological samples or questionnaires) provide good information for carrying out studies. Study of survival times in relation to treatment is an area to be approached with caution. An effective treatment may not increase observed survival times since more advanced cases may be treated in this manner. Green and Byar (1984) argue that registry based studies are not an alternative to randomised trials in this area, and the focus of the registry should be for epidemiological studies. However the registry can identify patients for a cohort study and aid in the planning of case-control studies and clinical trials.

2.4 Data Items and Collection

Data recorded on the registration form are keyed into the computer system by the MCCR personnel. The current computer system of the MCCR has been operational since 1989, and represents a major advance in ease of recording,use and validation of data compared with previous systems.

The UKACR and ONS have outlined a Minimum Data Set (MDS) in which each item of information is regarded as essential. MCCR record other optional data items as routine and this is in common with most other registries. The three fundamental measurements for this analysis are date of diagnosis, date of death and date of birth as these allow the calculation of survival and comparisons between age groupings. The recording of sex and place of residence are also vital for comparative purposes. Survival and incidence reports will tend to focus on these particular data items.

There are a large number of sources of information for the MCCR, ranging from hospital case notes to post-mortem reports. After notification of a tumour, registry staff will search for additional information such as case notes and treatment records. It is difficult to follow up patients, especially those with a long survival period, and this has led to the registry flagging patients as "dead" or "not known to be dead". Ideally an active follow up of each patient would be implemented by the registry, but usually passive follow up is achieved through the NHS Central Register and the ONS (for instance, the notification of death certificates). There is a small proportion of cases however that remain 'immortal', i.e. after notification of the cancer no follow up has been achieved.

Powell(1991) points out the need to evaluate sources of data, observing cost and potential use. He asserts the need for effective record linkage in order to avoid multiple notifications.

The classification of cancers is that of the ICD-9 codes laid out by the World Health Organisation (WHO) in 1977. An ICD code provides information on topography (location) and morphology (behaviour). Together with an identification of subtype on the ICD code, there are two main classifications used to assess the stage of a cancer. The TNM method records the tumour size (T), indication of spread to lymph nodes(N) and spread to distant metastases(M). The second classification ranges from 0 - IV each describing a further level of invasion (0 being in-situ and IV distant spread). Stage at diagnosis is not a routinely collected data item by MCCR, partly because of the inconsistent manner in which it is available in clinical notes.

2.5 Measurement Problems in Cancer Registration

The primary purpose of this thesis is to examine survival measurement for cancer registry data and to develop methodology to cope with any particular problems associated with it. The simplest definition of survival is (date of death) - (date of diagnosis). Briefly here, and in more detail in chapter 6, we examine the recording of date of diagnosis by cancer registries. It is immediately apparent that comparison of survival times may be clouded if the observed duration of survival is lengthened by artificially shifting the 'zero' time. If this occurs for

all patients or only a subset of patients then comparisons of survival and fitting of models to prognostic factors may contain considerable bias.

The start date of a cancer might logically be date of disease onset, but this is impossible to ascertain. It may be possible via screening to detect a cancer in the pre-clinical phase. Following symptoms, an individual will present at a GP and is then usually referred to a hospital. A clinical diagnosis may be possible by the GP, otherwise a variety of macroscopic and microscopic verifications are undertaken following date of first attendance at a hospital. Following diagnosis a patient may receive treatment designed to treat the cancer, or may receive palliative care in order to alleviate symptoms alone.

Hence there is considerable scope for defining date of diagnosis for patients with a full history available. If no history is available other than date of death, then the patient has a theoretical survival of zero. There is controversy over the best process of defining date of diagnosis (see chapter 6), and no research has been undertaken to gauge the effect of this "window of diagnosis" on conclusions of survival.

2.6 Introduction to Lung Cancer

Lung cancer has received extensive research over a long period of the twentieth century, and hence its aetiology is reasonably well understood. Despite this, it is the third most common cause of death in the U.K., and represents a quarter of all cancer mortality (NHS, 1998) . On a worldwide basis lung cancer is commoner

in western countries (Williams, 1992).

2.6.1 Epidemiology of Lung Cancer

It is widely established that the greatest cause of primary lung cancer is cigarette smoking. Epidemiologists have readily shown that increased smoking is responsible for the transformation of lung cancer from a rare disease at the turn of the century to a major cause of death today. Indeed by late middle age the lung cancer rate for regular smokers is more than ten times that of never smokers (Doll and Peto, 1981)

Increased risk for 'current smokers' is in the region of fifteen times higher compared with 'never smokers', and is highest for heavy smokers who have smoked for a long period of time. For ex-smokers, death levels reduce and are are similar to 'never smokers' about 25 years after stopping (Williams, 1992). It is estimated about 2% of all cases are attributed to passive smoking (Williams *et al.*, 1993).

Pollution and industrial exposures such as asbestos are associated with increased risk. Radon, a radioactive naturally occurring gas has received increasing attention by authors (Chaffey and Bowie, 1994). It is generally believed that chemical and industrial exposures interact synergistically with smoking to increase risk (NHS, 1998).

There is some speculation that improved diet may provide protection against lung cancer, but the protective mechanism is not understood. Alcohol may have

an adverse effect (Carpenter *et al.*, 1998), while tea drinking (Ohno *et al.*, 1995; Mendilaharsu *et al.*, 1998) and physical activity (Thune and Lund, 1997) may be beneficial . There is also a possible link between depression and lung cancer (Knekt *et al.*, 1996). More research is required however to establish such associations.

Elevated risk is also linked with living in a urban area and belonging to a lower social class. This may be explained by smoking patterns.

The profile of highest risk includes (Williams, 1992) :

- living in a westernised society;
- being a man;
- being a smoker;
- being aged 60 or more;
- living in an urban environment.

Strategies for controlling lung cancer focus on prevention, as prognosis once the disease is diagnosed is poor. Obviously the single most important measure of prevention is the trend in smoking, which until recently had been dropping. Since 1994, smoking amongst young people and women is on the increase (NHS, 1998).

2.6.2 Diagnosis and Treatment of Lung Cancer

Screening has not proved effective in improving the outcome of lung cancer cases. For the Merseyside and Cheshire Region about 55 % of cases are diagnosed

with the aid of microscopic evidence. Histologic confirmation typically involves a bronchoscopy or biopsy. Cytologic confirmation is achieved via microscopic examination of an individual's sputum. Macroscopic examination involves X-rays, but typically 36 % of cases are clinical diagnosed.

About 60 % of all lung cancer cases in the Mersey region do not receive specific treatment other than to alleviate suffering. This varies with age, and older patients are more likely to receive palliative treatment. For patients who do receive treatment, this is typically a combination of one, two or all three treatments available, namely surgery, radiotherapy and chemotherapy (Williams *et al.*, 1993).

2.6.3 Incidence of Lung Cancer for the Merseyside and Cheshire Region

As part of one of the four specific cancers targeted by the government for 'The Health of the Nation', the MCCR published a bulletin (Williams *et al.*, 1993) documenting incidence, treatment and a strategy for future prevention discussing government targets and particular problems associated with the Mersey region. The bulletin covered the 16 year period between 1975 and 1990, with a more detailed analysis of incidence for the years 1983 - 1990. Table 2.2 shows the SRR's by region for the period 1986 - 1990 and underlines the fact that incidence in the poorer, urban areas is greater than that for wealthier, rural regions.

The bulletin gave a valuable insight into the current position of the Mersey

Table 2.2: Standardised Registration Ratios 1986-1990 for Mersey RHA (Lung Cancer)

District	Males (95 % C.I.)	Females (95 % C.I.)
Chester	106 (97,117)	115 (99,132)
Crewe	96 (89,105)	91 (79,104)
Halton	136 (123,150)	150 (129,174)
Liverpool	168 (161,176)	225 (212,239)
Macclesfield	94 (85,103)	103 (89,119)
South Sefton	143 (132,155)	177 (158,198)
Southport and Formby	105 (94,117)	132 (114,152)
St. Helens & Knowsley	141 (132,150)	152 (139,167)
Warrington	115 (104,126)	138 (121,158)
Wirral	129 (121,136)	149 (137,162)
Mersey	130 (127,133)	154 (150,160)

England and Wales = 100
Source: Youngson et al 1992

region for lung cancer. In particular the Mersey region is *"top nationally for both men and women for lung cancer incidence"*.

2.7 Introduction to the Lung Cancer Dataset

The data used in this thesis are cases of primary lung cancer (ICD 162) in the Mersey region on the registry database for the years 1974 - 1993 inclusive. Incidence for each sex by year is given in figure 2.1. These suggest that trends identified in the lung cancer bulletin have continued - the overall number of cases is not in decline and the proportion of female cases is increasing. The annual average number of male and female cases over the full 20 year period is 1496 and 615, compared with 1386 and 464 for the first five years of the study and 1380 and 760 for the final five years of study. Therefore any reductions in incidence for males is offset by an increase in the number of female cases. For the years

1975 - 1990 there are 33,914 cases compared with 34,118 cases used for the lung cancer bulletin, however the shortfall of 204 cases may be due to re-classification of cases as extra-regional or non-primary lung cancer.

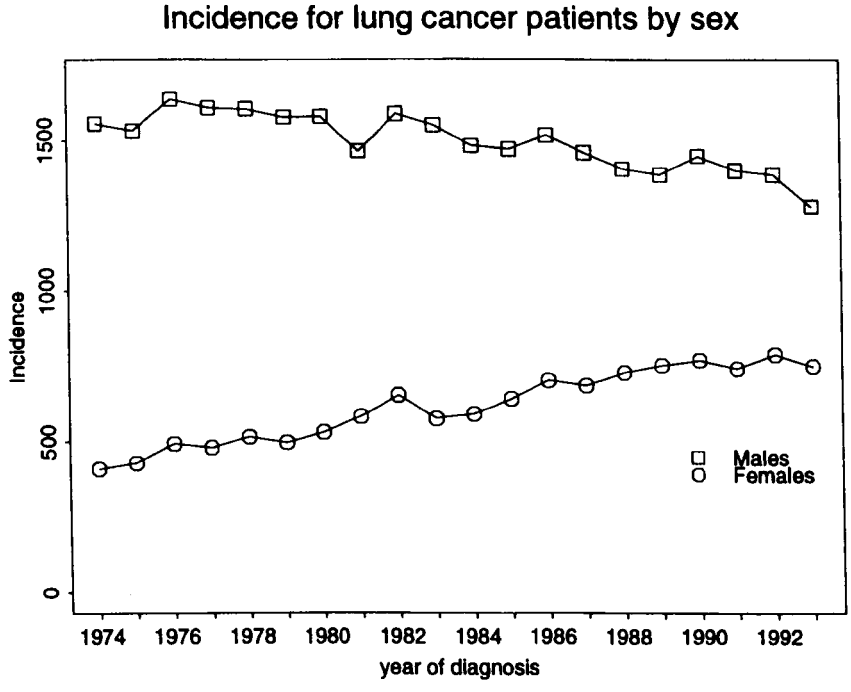


Figure 2.1: Incidence of lung cancer patients

There are 2092 cases with zero or negative survival times and these are removed as it is not theoretically possible to admit zero or negative survival times in a survival analysis (see chapter 3). A summary is given in table 2.3. Such cases may arise from typing errors by registry staff or patients who genuinely died on the same day as diagnosis, but cases with zero survival are typically patients whose survival history comes from a Death Certificate Only notification. For these cases the death certificate is the initial and only point of contact with the registry. For the Mersey region as a whole an initial notification by death certificate accounts for about $\frac{1}{3}$ of all cases, but history is successfully traced

for about $\frac{2}{3}$ of these. The proportion of cases notified by death certificate and the proportion of these whose records can be traced varies dramatically between registries. Typically survival for cases initiated by death certificate and then traced is poorer, and hence estimates of survival with Death Certificate Only cases excluded may be biased in favour of better survival. One strategy is to estimate the survival time for Death Certificate Only cases from characteristics of survival in the traced population. However the absolute difference in survival estimation for cancers with short survival (such as lung cancer) by removing Death Certificate Only cases is likely to be less severe than for cancers with superior survival (such as breast cancer) (Berrino *et al.*, 1995).

Table 2.3: Cases with zero or negative survival - survival times

survival time	0	-1 - -7	-8 - -14	-15 - -99	$\leq$ -100	total
no. of cases	1805	169	100	7	11	2092

A summary of cases used in the full survival analysis of lung cancer is in table 2.4.

Table 2.4: Cases included in survival analysis

year	males	females	total	male:female ratio
1974	1541	404	1945	3.80:1
1975	1510	420	1930	3.58:1
1976	1623	487	2110	3.32:1
1977	1576	463	2039	3.36:1
1978	1551	497	2048	3.12:1
1979	1476	461	1937	3.19:1
1980	1538	520	2058	2.99:1
1981	1455	576	2031	2.52:1
1982	1581	644	2225	2.44:1
1983	1542	570	2112	2.70:1
1984	1479	585	2064	2.52:1
1985	1452	632	2084	2.30:1
1986	1495	695	2190	2.16:1
1987	1433	677	2110	2.13:1
1988	1269	648	1917	1.93:1
1989	1235	658	1893	1.84:1
1990	1302	654	1956	1.88:1
1991	1249	666	1915	1.89:1
1992	1230	681	1911	1.76:1
1993	1070	585	1655	1.71:1
total number			40130	

2.8 Summary

As cancer is responsible for 1 in 4 deaths in England and Wales, research into its causes and control is vital. In this chapter the reader has been familiarised with the system of cancer registration, an unrivalled resource for the study of incidence and survival of cancer. However, as there are many sources of data and quality may vary, a brief discussion has been made regarding the problems of survival measurement and how the method of diagnosis for an individual could artificially extend their survival.

The epidemiology of lung cancer has been discussed, in particular how smoking has made lung cancer a major cause of death in the western world. Those at highest risk are males living in less affluent urban areas who have a poor lifestyle.

An introduction to the data used throughout this thesis was given. The full datset consisted of 42,222 cases between 1974 and 1993 but due to the removal of cases with 'impossible' survival the dataset to be used for analysis consists of 40,130 patients. While overall incidence over the period has remained steady, the proportion of female cases has increased from about 20 % in 1974 to about 35 % in 1993.

Analysis of survival for the lung cancer dataset will be used to illustrate survival methodology in chapter 3. A re-analysis of 500 cases diagnosed in 1993 is undertaken in chapter 6 incorporating a model for survival time measurement error using methodology introduced in chapter 5.

Chapter 3

Introduction to Survival Analysis

In this chapter we introduce the reader to survival analysis. After each technique is reviewed we apply it to the full lung cancer dataset introduced in the previous chapter. All analyses assume survival and prognostic factors are accurately recorded, although a brief discussion of uncertainty around date of diagnosis was given in chapter 2.

Survival analysis has become an extensively researched statistical topic, and many texts are available to the researcher for reference. Two styles have become common. A traditional approach is adopted by authors such as Collett (1994), Kalbfleisch and Prentice (1980) and Cox and Oakes (1984). We have adopted this style for this thesis, as it is more accessible to epidemiologists and workers in the cancer field. A more mathematically rigorous approach has become increasingly common based on the theory of counting processes, and is the preferred style of authors such as Fleming and Harrington (1991).

3.1 Introduction

Survival data or failure time data are typified by a measurement to a particular event such as time to death from time of diagnosis, time to recurrence from time of remission for cancer studies or time to failure of machinery in industrial / engineering applications. Every study should have a defined origin and end point and every survival episode should have a start time and an end time. Figure 3.1 shows the history of 5 patients in a study of 1000 days, where recruitment takes place in the first 500 days. Patients are either lost to follow up during the study, die or are still alive at the end of the study. In order to define survival data for such patients, one must take each individual's starting point and end point, so survival times all have the same origin as in figure 3.2. The definition of the start time for an individual is rarely contentious, for instance in a clinical trial situation all patients may have start time as date of randomisation. However this is not the case for cancer registry data, where there is ambiguity surrounding the definition of date of diagnosis. This will be discussed further in chapter 6.

If a patient is lost to follow up, or has died from a cause not associated with the definition of failure for the study, their time is recorded as being at least that last recorded. Likewise is the case for patients still alive at the end-point of the study. Such cases are called *censored* and are a distinguishing feature of survival data. The presence,type of censoring and censoring mechanism have a great effect on a particular analysis.

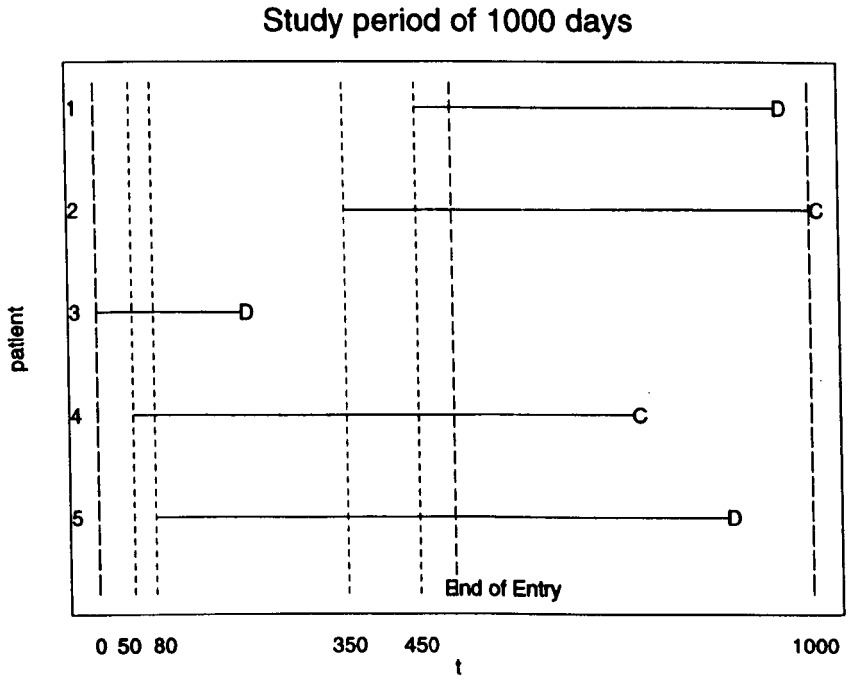


Figure 3.1: History of 5 patients in a hypothetical study of 1000 days

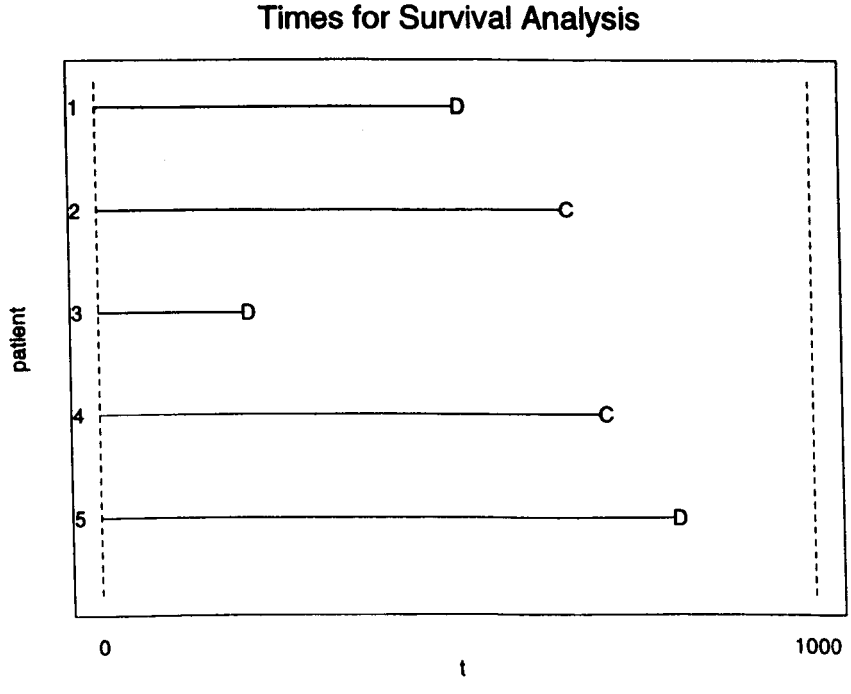


Figure 3.2: Generation of times for survival analysis of the 5 patients

3.2 Censoring

3.2.1 Types of Censoring

Left Censoring

In this situation the survival time is known to be less than the observed time, such as when detection of the event takes place after the event has occured, for instance at the three monthly check-up.

Right Censoring

Here the survival time is known to be after the observed survival time. This could be due to loss to follow up or more usually because the subject survives beyond the end point of the study. This is the most common form of censoring and is typical of cancer registry patients who have not died at the point of data being taken for analysis.

Interval Censoring

If the failure occurs during an interval but only the start and end points of the interval are observed the time is interval censored. Typically this type of censoring occurs when patients are seen on a regular basis, and detection of the event occurs when the patient is seen.

3.2.2 Classes of Censoring

Independent Censoring

Kalbfleisch and Prentice describe this as "conditional independence" between the censoring and the failure mechanism. In other words censored times are

representative of all at risk at a given time and whether or not censoring occurs is not dependent on the risk at a given time.

Random Censoring

Censoring is regarded as random if censored times are stochastically independent of each other and of the failure times.

Informative Censoring

If the censoring mechanism depends on the parameters of interest then the censoring can be considered informative. Collett gives an example where a patient might be withdrawn from the study due to a side-effect from a treatment and hence the censoring is informed by treatment received. One suggestion is to plot survival and censoring times against such an explanatory variable and check if a pattern emerges.

Type I censoring

Here a study has a fixed end-point and hence the censoring time is fixed in advance.

Type II censoring

If the end point of a study is dependent on a given number of failures occuring this is called type II censoring.

3.2.3 Censoring in the Lung Cancer Data

We will assume that the censoring is random and type I for the lung cancer data. The censoring date is the 27th of May 1997, the date the data were drawn from the registry computer records for analysis, hence patients have survived at least three years five months even if their diagnosis date is the final entry date of December 31st 1993. All censored patients are right censored, as we know they have survived up to the date the data were drawn from the registry. There may be a small number of patients who died before this date but this information was not entered on to the registry database at the time, however this is likely to be a tiny proportion of the total number of cases. A summary of censoring is given in table 3.1. Note how small the frequency of censoring is, as long term survival from lung cancer is rare.

	frequency	percent	cumulative frequency
Censored	1520	3.6	1520
Dead	40702	96.4	42222

Table 3.1: Censoring for the lung cancer patients

3.3 Representations of Survival Time

Survival times have a probability density function valid only for non-negative times. Writing the density function as $f_T(t)$ then the distribution function is defined as:

$$F(t) = P(T < t) = \int_0^t f_U(u)du \tag{3.1}$$

The survivor function for a survival time distribution is defined as the probability of surviving beyond a given time and hence is defined as:

$$S(t) = P(T > t) = 1 - F(t) \text{ where } S(0) = 1 \text{ and } \lim_{t\to\infty} S(t) = 0 \tag{3.2}$$

note

$$f_T(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

The hazard function is the instantaneous death rate for an individual surviving to time t . Given that an individual has survived up to t the hazard is the probability of failing at time t. The hazard function is defined as:

$$\begin{aligned} \lambda(t) &= \lim \delta t \to 0\{\frac{P(t \le T < t + \delta t | T \ge t)}{\delta t}\} \\ &= \frac{\lim \delta t \to 0\{\frac{F(t+\delta t)-F(t)}{\delta t}\}}{P(T \ge t)} = \frac{\frac{dF(t)}{dt}}{S(t)} = \frac{f_T(t)}{S(t)} \end{aligned} \tag{3.3}$$

Hence there are important relationships between the probability density function, the survivor function and the hazard function of a particular variable t. Note also

$$\lambda(t) = \frac{f_T(t)}{S(t)} = \frac{\frac{-dS(t)}{dt}}{S(t)} = -\frac{d}{dt}\log S(t) \tag{3.4}$$

The cumulative hazard is the sum of the instantaneous hazard rates over time

$$\Lambda(t) = \int_0^t \lambda(u)du = -\log S(t) \text{ and } S(t) = \exp(-\Lambda(t)) \tag{3.5}$$

3.3.1 Non-Parametric Estimates of the Survivor, Hazard and Cumulative Hazard functions

If we have a set of survival data and wish to estimate the survivor, or equivalently the cumulative hazard function, without parameterising the distribution, there are two main types of estimate based on the the proportion of individuals at risk for some form of partitioning of the time axis. The product-limit method considers risk at the observed death times, whereas the actuarial method considers risk in intervals independent of observed failure.

Product Limit / Kaplan-Meier Estimate

Given $t_i, i = 1, \ldots, n$ survival times of which r are failure times and $n - r$ are censored times, with d_i failures at t_i, and n_i individuals at risk just prior to t_i then the "non-parametric maximum likelihood estimate" of the survivor function (Kalbfleisch and Prentice, 1980) is a step function, given as:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right) \qquad \text{for } t_{(k)} \leq t < t_{(k+1)}, k = 1, \ldots r, t_{(r+1)} = \infty \tag{3.6}$$

Throughout the thesis the notation $t_{(.)}$ will denote an ordered survival time. Note $\hat{S}(0) = 1$, and $\hat{S}(t_{(r)}) = 0$ if the rth failure time is the largest observed time, $\hat{S}(t_{(r)}) > 0$ if the rth failure time is less than an observed censored time. The most popular estimate of the variance is Greenwood's formula, derived by

regarding the proportion surviving in each interval as binomial,and defined as:

$$\mathrm{Var}(\hat{S}(t)) \approx (\hat{S}(t))^2 \{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\} \tag{3.7}$$

Greenwood's formula is only an approximation and therefore confidence intervals may be outside the range $[0, 1]$. One can either simply round to 0 or 1 or take a transformation such as the complementary log-log transformation (which has the range $(-\infty, \infty)$).

$$\mathrm{Var}(\log(-\log \hat{S}(t))) \approx \frac{1}{(\log \hat{S}(t))^2} \{\sum_{j=1}^{k} \frac{d_j}{n_j(n_j - d_j)}\} \tag{3.8}$$

Another tack is to estimate the cumulative hazard function at each death time:

$$\hat{\Lambda}(t) = \sum_{j=1}^{k} \frac{d_j}{n_j} \qquad \text{for } t_{(k)} \leq t < t_{(k+1)}, k = 1, \ldots r, t_{(r+1)} = \infty \tag{3.9}$$

This is known as Nelson's estimate. An approximation for tied death times is also given, as the above estimate is conservative for tied data due to rounding. If n_j are at risk just prior to t_j, and d_j deaths occur at t_j, then the modified contribution is:

$$\sum_{m=0}^{d_j - 1} \frac{1}{n_j - m} \tag{3.10}$$

The relationship between the cumulative and the survivor function can then be employed:

$$\hat{S}(t) = \exp(-\hat{\Lambda}(t)) \tag{3.11}$$

The variance of $\hat{\Lambda}(t)$ is approximated by:

$$\text{var}(\hat{\Lambda}(t)) \approx \frac{d_j}{(n_j)^2} \tag{3.12}$$

3.3.2 Estimates of the Survivor Function for Lung Cancer Data

For the 40,130 positive survival times in the lung cancer dataset we now estimate the survivor functions and related quantities of interest. A Kaplan - Meier estimate of the survivor function for all the cases is given in figure 3.3 and the median and other percentiles are given in table 3.2. Half of all patients die within 3 months of diagnosis and only 5 % of the patients survive more than five years.

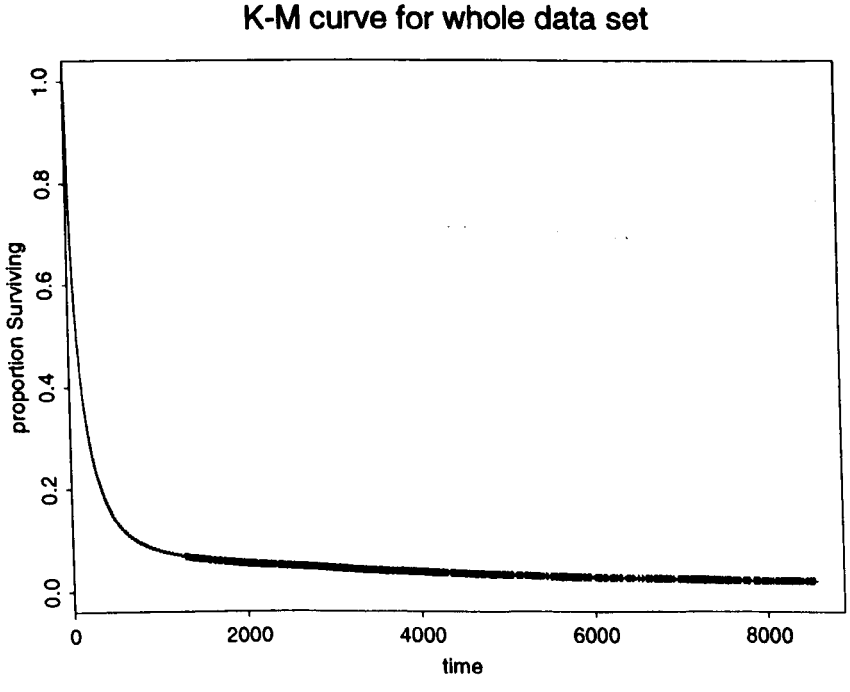


Figure 3.3: Survivor curve for lung data

percentile	5	25	50	75	95
survival time(days)	7	30	90	261	1824

Table 3.2: Percentiles of survival for lung data

Important covariates are age (a continuous covariate) and sex (a binary covariate). Other useful indicators such as stage are not routinely recorded (see chapter 2). The continuous variable age was categorised according to the deciles of the age distribution, in order to detect if there is a noticeable trend in survival across the distribution of age.

Age(yrs)	0.2-	54.8-	59.7-	62.9-	65.7-	68.2-	70.8-	73.4-	76.3-	80.2-124.9
frequency	4013	4022	4011	4015	4020	4001	4009	4026	4007	4006

Table 3.3: Creation of 10 age groups for continuous covariable age according to the deciles of the age distribution

Kaplan-Meier curves for the ten levels of age and sex are shown in figures 3.4, 3.5, 3.6 and 3.7. Table 3.4 shows the estimated median survival with confidence intervals.

3.3.3 Comparison of Survivor Curves

When we have g different groups of data, we wish to test whether the estimated survivor functions differ across the groups. The most commonly used test statistic is the log-rank test . This test is based on the concept of difference between observed and expected survival, and it is assumed the number of deaths at each of the r observed times follows a hypergeometric distribution. For each of the first $g-1$ groups the (observed - expected) contribution is:

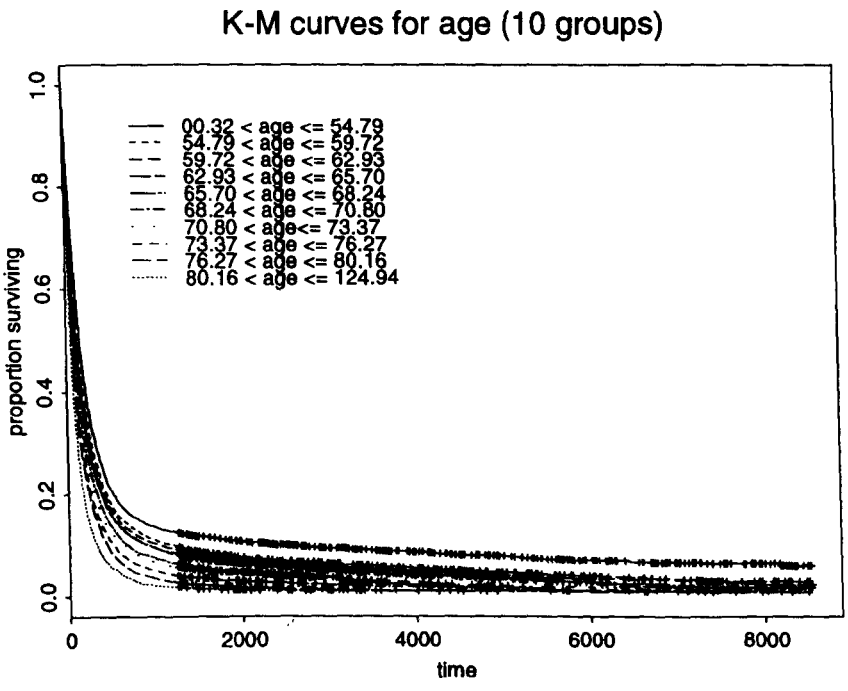


Figure 3.4: Survivor curves for age

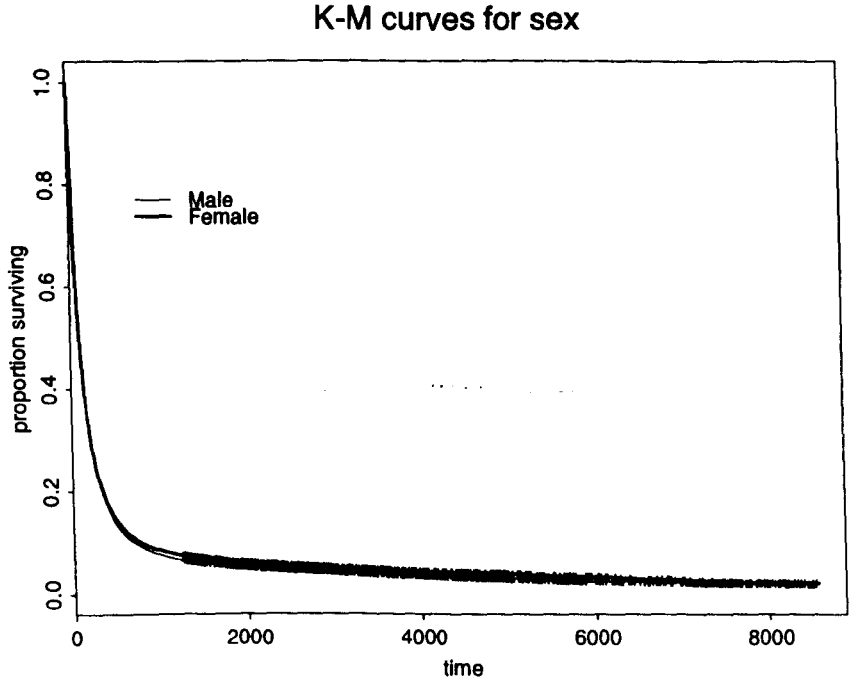


Figure 3.5: Survivor curves for sex

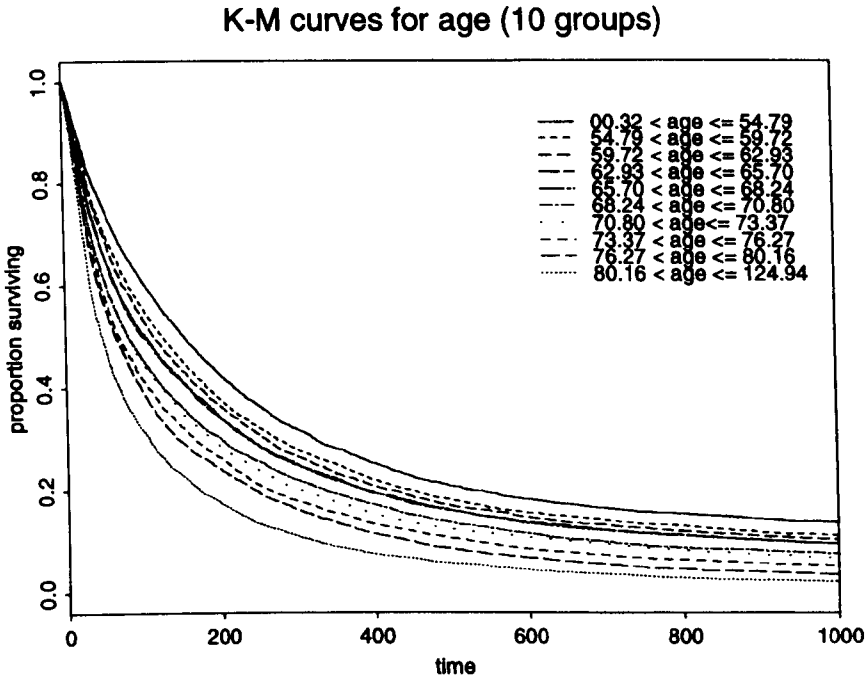


Figure 3.6: Survivor curves for age - survival up to 1000 days

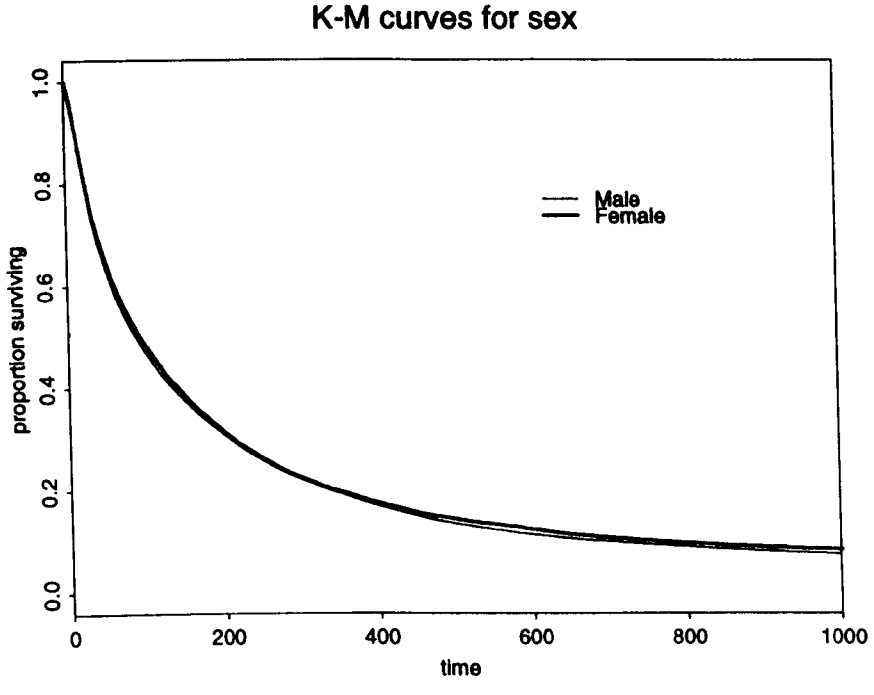


Figure 3.7: Survivor curves for sex - survival up to 1000 days

covariable		median survival (days)	95 % C.I.
age group	0.2 -	154	(146,162)
	54.8 -	127	(119,135)
	59.7 -	116	(109,124)
	62.9 -	102	(94,111)
	65.7 -	105	(97,112)
	68.2 -	82	(78,88)
	70.8 -	83	(77,88)
	73.4 -	71	(66,76)
	76.3 -	67	(63,71)
	80.2 - 124.9	48	(45,51)
sex	Male	89	(86,91)
	Female	93	(91,97)

Table 3.4: Median survival estimates for age and sex

$$u_i = \sum_{j=1}^{r} u_{(ij)} = \sum_{j=1}^{r} (o_{(ij)} - e_{(ij)}) = \sum_{j=1}^{r} (d_{(ij)} - \frac{n_{(ij)} d_j}{n_j}) \qquad i = 1, \ldots, g-1 \tag{3.13}$$

This has variance:

$$V_{(ii)} = \sum_{j=1}^{r} = V_{j(ii)} = \sum_{j=1}^{r} \frac{n_{(ij)}(n_j - n_{(ij)}) d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \tag{3.14}$$

Also the covariance of between groups i and k is:

$$V_{(ik)} = \sum_{j=1}^{r} (V_{j(ik)}) = \sum_{j=1}^{r} (\frac{-n_{(ij)} n_{(kj)} d_j (n_j - d_j)}{n_j^2 (n_j - 1)}) \tag{3.15}$$

Given the $g-1$ vector $\underline{u}$ and the variance-covariance matrix V then $\underline{u}^T V \underline{u}$ is $\chi^2_{(g-1)}$ and hence the hypothesis of no difference in survival between groups can be tested. Weighting the contribution to u_i gives a class of significance tests. In particular weighting by the numbers at risk at each time gives the Wilcoxon test.

3.3.4 Log-Rank tests for Lung Cancer Data

Having estimated the survivor functions for sex and age group, we now wish to test if there is a significant difference between the estimated levels of survival. We employ the log-rank test in order to do this. Sex has 2 levels, and is hence tested against a χ^2_1, while age has 10 levels and is tested against a χ^2_9 distribution, the results being in table 3.5.

covariable	$\Sigma\frac{(O-E)^2}{E}$	p
age group	1613	0
sex	10.5	0.000116

Table 3.5: Log-Rank tests for binary covariables

There is clear evidence that age has a strong effect on survival. Although the result is significant for sex, a dataset of this size will detect statistical significance when there is no real clinical significance in the result. The difference for sex may be due to a difference in age distribution across the sexes. The age distribution for males and females is shown in figure 3.8. These are extremely similar, though the median for males is more than that for females. After the median the age distribution of female cases shows they are marginally older than men.

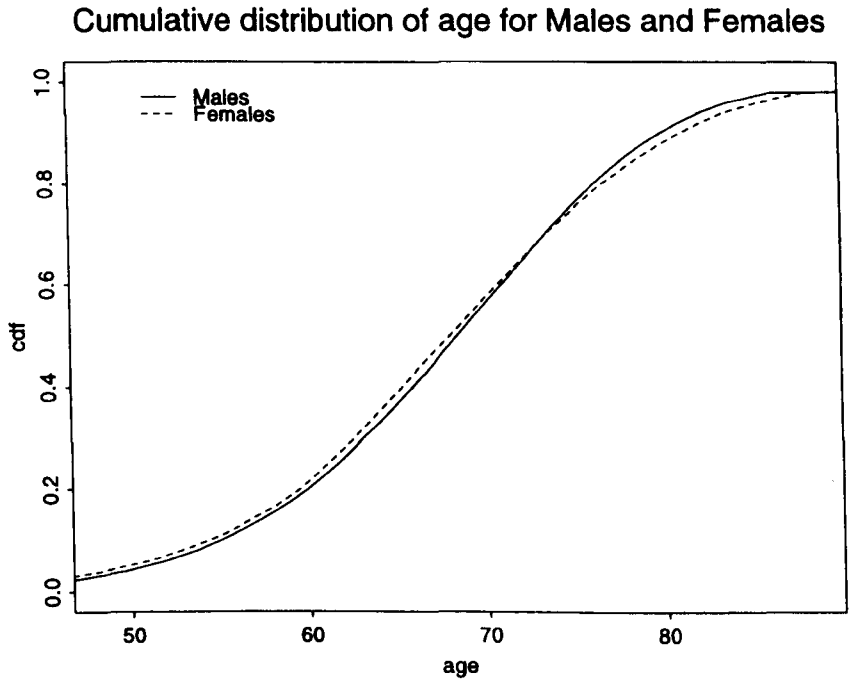


Figure 3.8: Age across the sexes

3.3.5 Parametric Distributions for Survival Data

There are many distributions applicable to survival data and an extensive review of these are in Kalbfleisch and Prentice. We are only considering the Weibull and exponential distributions. These may follow a proportional hazards representation (see section 3.4) when fitted to explanatory variables. In this thesis we are only considering proportional hazards. Other distributions include the gamma which has a difficult to represent hazard, and the log-logistic.

The exponential distribution

For the exponential distribution we have:

$$f_T(t) = \tau \exp(-\tau t) \text{ where } t > 0, \tau > 0 \quad (3.16)$$

$$S(t) = \exp(-\tau t) \text{ and } \lambda(t) = \tau \quad (3.17)$$

The exponential distribution has a constant hazard over time. The mean is $\frac{1}{\tau}$, the variance $\frac{1}{\tau}^2$ and the median $\frac{1}{\tau}\log 2$. It is also memoryless in that given survival up to a certain time the distribution of survival is still exponential with parameter τ.

The Weibull distribution

For the Weibull distribution we have:

$$f_T(t) = \tau\rho t^{\rho-1} \exp(-\tau t^{\rho}) \text{ where } t > 0, \tau > 0, \rho > 0 \quad (3.18)$$

$$S(t) = \exp(-\tau t^{\rho}) \text{ and } \lambda(t) = \tau\rho t^{\rho-1} \quad (3.19)$$

Note the exponential distribution is a special case of the Weibull distribution with $\rho = 1$. In fact ρ is called a shape parameter as it determines the shape of the hazard over time. If $\rho < 1$ the hazard decreases over time and if $\rho > 1$ it increases over time. This is illustrated in figure 3.9. The Weibull distribution has mean $\frac{\Gamma(1+\frac{1}{\rho})}{\tau^{\frac{1}{\rho}}}$ and median $\{\frac{\log 2}{\tau}\}^{\frac{1}{\rho}}$. Kalbfleisch and Prentice (1980) use $\tau = \tau^{\rho}$ in their representation.

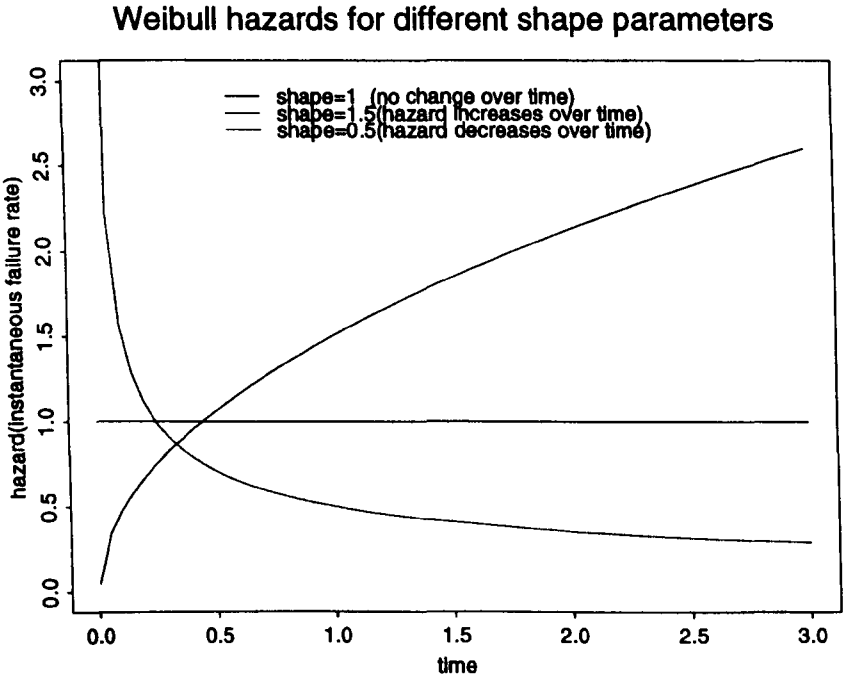


Figure 3.9: Weibull hazard function for different shape parameters

3.4 Proportional Hazards

If we have explanatory variables associated with the survival outcome, then the assumption of *proportional hazards* implies that the hazard for a set of covariates $\underline{x}$ acts multiplicatively on the *baseline hazard* when $\underline{x} = 0$. This is mathematically denoted as:

$$\lambda(t|x) = \lambda_0(t)g(\underline{\beta}^T\underline{x}) \tag{3.20}$$

The function $g(.)$ is the *relative risk* for covariates. Care must be taken to ensure that $g(.)$ is positive and thus a linear form is rarely employed. We will now only consider the log-linear model:

$$\lambda(t|x) = \lambda_0(t)\exp(\underline{\beta}^T\underline{x}) \tag{3.21}$$

Notice that if there is more than one covariate, the model assumes that their joint effect is multiplicative, and that an increase of one in the value of a covariable x_1 with relative risk $\exp(\beta_1)$ represents an increase in the hazard of $\exp(\beta_1)$. For example:

$$\begin{aligned}
\lambda(t|x_1, x_2) &= \lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2) \\
\lambda(t|x_1 = 1, x_2 = 1) &= \lambda_0(t)\exp(\beta_1 + \beta_2) = \lambda_0(t)\exp(\beta_1)\exp(\beta_2) \\
\lambda(t|x_1 = 2, x_2 = 1) &= \lambda_0(t)\exp(2\beta_1 + \beta_2) = \lambda_0(t)\exp(\beta_1)\exp(\beta_1)\exp(\beta_2)
\end{aligned}$$

We can also easily show the effect of the covariables on the survivor and cumulative hazard functions.

$$\Lambda(t|x) = \Lambda_0(t)\exp(\underline{\beta}^T\underline{x}) \tag{3.22}$$

$$S(t|x) = \{S_0(t)\}^{(\exp(\underline{\beta}^T\underline{x}))} \tag{3.23}$$

The baseline hazard may be parameterised as a Weibull or exponential baseline, but the most common approach is to regard it as arbitrary and concentrate on estimation of the parameters of relative risk.

3.5 The Cox Proportional Hazards Model

The proportional hazards model with an unspecified baseline hazard was first proposed by Cox (1972). The Cox model has been described as *semi-parametric*

- the baseline is not parameterised but the effect of covariates on the baseline is. When primary interest is in the effect of covariates on an arbitrary baseline it is assumed that the intervals between survival times offer "no information" on $\underline{\beta}$ and hence only the order of the survival times contributes to the estimation of $\underline{\beta}$. The formulation of the likelihood in Cox's (1972) paper was justified by a conditioning argument.

p(failure for individual j|all individuals surviving up to t_j) =

$$\frac{\lambda_{(\text{individual j})}(t|\underline{\beta})}{\sum_{(l\in\text{total at risk at } t_j)}\lambda_l(t|\underline{\beta})} = \frac{\lambda_0(t)\exp(\underline{\beta}^T\underline{x}_j)}{\sum_l \lambda_0(t)\exp(\underline{\beta}^T\underline{x}_l)} = \frac{\exp(\underline{\beta}^T\underline{x}_j)}{\sum_l \exp(\underline{\beta}^T\underline{x}_l)} \quad (3.24)$$

The suggested likelihood is the product of above terms for the r death times of the n individuals in the study. The conditioning argument was "unduly cryptic" and the likelihood required additional clarification. This came in the forms of marginal likelihood (Kalbfleisch and Prentice, 1973) and partial likelihood (Cox, 1975).

3.5.1 Parameter Estimation and Inference

The technique of marginal likelihood arises from the fact that a transformation of the order statistic $\{t_{(1)}\ldots t_{(r)}\}$ does not lead to a change in the rank statistic. Thus the rank statistic is "sufficient for $\underline{\beta}$ in the absence of knowledge of $\lambda_0(t)$" (Kalbfleisch and Prentice, 1973). The marginal likelihood of $\underline{\beta}$ in the absence of tied times and censoring ($r = n$) is defined as:

$$\begin{aligned} L(\underline{\beta}) &\propto \int_0^\infty \int_{t_{(1)}}^\infty \cdots \int_{t_{(n-1)}}^\infty \Pi_1^n f(t|x) dt_{(n)} \ldots dt_{(1)} \\ &= \int_0^\infty \int_{t_{(1)}}^\infty \cdots \int_{t_{(n-1)}}^\infty \Pi_1^n \lambda(t|x) S(t|x) dt_{(n)} \ldots dt_{(1)} \\ &= \int_0^\infty \int_{t_{(1)}}^\infty \cdots \int_{t_{(n-1)}}^\infty \Pi_1^n \lambda(t|x) \exp(-\Lambda(t|x)) dt_{(n)} \ldots dt_{(1)} \\ &= \frac{\exp(\sum_{j=1}^n \underline{\beta}^T \underline{x_j})}{\prod_{j=1}^n \sum_{l \in R(t_{(j)})} \exp(\underline{\beta}^T \underline{x_l})} \end{aligned} \tag{3.25}$$

where the notation $R(t_{(j)})$ denotes the *risk set* for the jth ordered failure time. Kalbfleisch and Prentice also derive the likelihood for right censoring and show the likelihood is exactly the same as that of Cox.

•

Cox (1975) introduced the concept of *partial likelihood* and derives the partial likelihood (PL) for his proportional hazards model with r failures from n individuals:

$$\begin{aligned} PL(\underline{\beta}) &= \prod_{j=1}^r f(\text{failure case } t_{(j)} | \text{censoring in} [t_{(j-1)}, t_{(j)}], \text{failure case } t_{(j-1)}; \underline{\beta}) \\ &= \frac{\exp(\sum_{j=1}^r \underline{\beta}^T \underline{x_j})}{\prod_{j=1}^r \sum_{l \in R(t_{(j)})} \exp(\underline{\beta}^T \underline{x_l})} \end{aligned} \tag{3.26}$$

Inference on $\underline{\beta}$

The large sample properties of the partial likelihood allow the *maximum partial likelihood* estimate to be treated as a *maximum likelihood* estimate. In particular the estimates are asymptotically consistent and multivariate normal:

$$\begin{aligned} \hat{\underline{\beta}} &\sim N(\underline{\beta}, \frac{1}{I(\underline{\beta})}) \\ I(\underline{\beta}) &= [\gamma_{ij}] = E(\frac{-\partial^2(\log L(\underline{\beta}))}{\partial\beta_i\beta_j}) \end{aligned} \tag{3.27}$$

We can therefore calculate confidence intervals for $\underline{\beta}$. The *score test* of the hypothesis $\underline{\beta} = \underline{0}$ is defined as

$$\frac{(\frac{\partial \log L(\underline{0})}{\partial \underline{\beta}})^2}{I(\underline{0})} \tag{3.28}$$

and is tested against a χ^2 distribution. For the case of a single binary covariate this is exactly the same as the log-rank test. A further test is the *Wald* test, where

$$\frac{\hat{\underline{\beta}}}{s.e.(\hat{\underline{\beta}})} \tag{3.29}$$

is tested against the standard normal distribution (or equivalently the square of this statistic is tested against a χ^2 distribution).

Model selection can be carried out via comparison of $-2\log L$. If an additional variable is included, the difference between $-2\log(\hat{L})$ from the model with and without that variable can be tested against a χ^2 distribution. If the result is not significant, the simpler model is preferred.

3.5.2 Treatment of Tied Data

The proportional hazards model is a continuous time model and hence does not permit tied data, however most real datasets include tied observations and hence a new likelihood must be specified. This is an important idea to this thesis - rounding creating ties is a form of outcome variable error. We now outline "exact" and approximate likelihoods for tied data. The idea of approximating the likelihood when survival data display errors is extended more generally later in the thesis (chapter 5), hence it is important here to examine established approximations for ties.

Let us consider three tied times ($t_{(1)} = t_{(2)} = t_{(3)}$) and a further two times at risk i.e. ($t_{(4)}, t_{(5)} > t_{(3)}$) with relative risks $\psi(j) = \exp(\underline{\beta}^T \underline{x_j}), j = 1, \ldots, 5$. The true partial likelihood at $t_{(1)} = t_{(2)} = t_{(3)}$ is now one of six possibilities:

$$\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(2)}{\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(3)}{\psi(3)+\psi(4)+\psi(5)}$$

$$\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(3)}{\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(2)}{\psi(2)+\psi(4)+\psi(5)}$$

$$\frac{\psi(2)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(1)}{\psi(1)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(3)}{\psi(3)+\psi(4)+\psi(5)}$$

$$\frac{\psi(2)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(3)}{\psi(1)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(1)}{\psi(1)+\psi(4)+\psi(5)}$$

$$\frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(1)}{\psi(1)+\psi(2)+\psi(4)+\psi(5)} \frac{\psi(2)}{\psi(2)+\psi(4)+\psi(5)}$$

$$\frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} \frac{\psi(2)}{\psi(1)+\psi(2)+\psi(4)+\psi(5)} \frac{\psi(1)}{\psi(1)+\psi(4)+\psi(5)}$$

Kalbfleisch and Prentice (1973) derive a marginal likelihood and the contribution at a death time with ties is the sum of the possible likelihoods i.e. in our example

the sum of the six possibilities. Defining $D(t_{(j)})$ as the set of failures at $t_{(j)}$ and m_j as the number of failures at $t_{(j)}$:

$$L_{marginal} = \prod_{j=1}^{r} \Pi_{k \in D(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_k) \sum_{\text{all perms}} \prod_{i=1}^{m_j} \frac{1}{\sum_{l \in R(t_{(j)}, \text{one perm})} \exp(\underline{\beta}^T \underline{x}_l)} \tag{3.30}$$

Cox (1972) considered a conditional logistic model as a discrete time analogy to the continuous case. This is defined as:

$$\frac{\lambda(t|z)dt}{1 - \lambda(t|z)dt} = \frac{\lambda_0(t)dt}{1 - \lambda_0(t)dt} \exp(\underline{\beta}^T \underline{x}) \tag{3.31}$$

The contribution to the likelihood denominator is then the sum over all combinations of choosing m_j individuals from the risk set at $t_{(j)}$. For our example we have:

$$\begin{aligned}(\psi(1)\psi(2)\psi(3)) * (1/(\quad & \psi(1)\psi(2)\psi(3) + \psi(1)\psi(2)\psi(4) + \\ & \psi(1)\psi(2)\psi(5) + \psi(1)\psi(3)\psi(4) + \\ & \psi(1)\psi(3)\psi(5) + \psi(1)\psi(4)\psi(5) + \\ & \psi(2)\psi(3)\psi(4) + \psi(2)\psi(3)\psi(5) + \\ & \psi(2)\psi(4)\psi(5) + \psi(3)\psi(4)\psi(5))\end{aligned} \tag{3.32}$$

This is more formally expressed as:

$$L_{discrete} = \prod_{j=1}^{r} \frac{\Pi_{k \in D(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_k)}{\sum_{(l \in \text{all combs size } m_j \text{ from risk set})} \exp(\underline{\beta}^T \underline{x}_l)} \tag{3.33}$$

Both likelihoods assume censored observations at death time occur after all the failure times.

Approximations to the Partial Likelihood

The calculations required in order to fit the "exact" likelihoods make their use unpopular. Other approximations have become the norm for practical fitting of the Cox model. One such approximation is due to Breslow (1974), and assumes each tied failure time occurs before each other tied time. For our example we have the likelihood:

$$\frac{3! * \psi(1)\psi(2)\psi(3)}{(\psi(1) + \psi(2) + \psi(3) + \psi(4) + \psi(5))^3} \tag{3.34}$$

More generally, if we have m_j failures at $t_{(j)}$:

$$L_{Breslow} = \prod_{j=1}^{r} \frac{\Pi_{k \in D(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_k)}{(\sum_{(l \in R(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_l))^{m_j}} \tag{3.35}$$

This approximation is fine if the proportion of tied times to the total risk set is small. Oakes (1981) examined the expectation of the second derivative of the likelihood and found Breslow's approximation overestimates the variance of the first derivative. He then suggested weighting each tied failure time by $\frac{m_j+1}{2m_j}$ in each of the m_j risk sets. Despite the improvement on the Breslow approximation the Oakes approximation is not, as far as this author is aware, a

feature of commercial packages for survival analysis.

Another approximation which weights each of the tied failure times is that of Efron (1977):

$$(\psi(1)\psi(2)\psi(3)) * (1/\ (\ (\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)) * (\frac{2}{3}\psi(1)+\frac{2}{3}\psi(2)+\frac{2}{3}\psi(3)+\psi(4)+\psi(5)) * (\frac{1}{3}\psi(1)+\frac{1}{3}\psi(2)+\frac{1}{3}\psi(3)+\psi(4)+\psi(5)))) \tag{3.36}$$

Note if the tied data all have the same relative risk, the Efron likelihood is identical to the exact marginal likelihood for ties. We then have:

$$L_{Efron} = \prod_{j=1}^{r} \frac{\Pi_{k\in D(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_k)}{\prod_{k=1}^{m_j}\{\sum_{(l\in R(t_{(j)}))} \exp(\underline{\beta}^T \underline{x}_l) - \frac{(k-1)}{m_j}\sum_{(l\in D(t_{(j)}))} \exp(\underline{\beta}^T \underline{x}_l)\}} \tag{3.37}$$

3.5.3 Cox Model Fits to Lung Cancer Data

Computer packages to fit the Cox model usually use the Breslow approximation as the default option. The S-plus survival library (Therneau, 1994) has the Efron approximation as the default option. We fit models for age (continuous covariate and binary covariate created around the median 68.24- see table 3.6) and binary covariate sex. The results for continuous covariate age are in table 3.7 and for binary covariate age are in table 3.8. Neither fit shows any significance for an interaction between age and sex, and for the case of the continuous covariate age the effect of sex is approximately the effect of one year of age. The hazard

is multiplied by $\exp(0.0206) = 1.02$ for each year of age, thus the effect of 10 years of age is to increase the hazard by $\exp(10 * 0.0206) = 1.229$.

Age	frequency
≤ 68.24	20081
> 68.24	20049

Table 3.6: Creation of binary covariate for age

model eqn	coeff age (s.e.) p	coeff sex (s.e.) p	coeff (age*sex) p
age + sex + (age*sex)	0.019 (0.002) 0	-0.120 (0.077) 0.16	0.001 (0.001) 0.30
age + sex	0.0206 (0.0005) 0	-0.03 (0.011) 0.007	-
age	0.0206 (0.0005) 0	-	-

Table 3.7: Fits to lung data - covariate age,factor sex

Table 3.8: Fits to lung data - binary covariate age,factor sex

model eqn	coeff age (s.e.) p	coeff sex (s.e.) p	coeff (age*sex) p
age + sex + (agefac*sex)	0.293 (0.031) 0	-0.044 (0.016) 0.006	0.029 (0.023) 0.20
age + sex	0.331 (0.010) 0	-0.030 (0.011) 0.009	-
age	0.331 (0.010) 0	-	-

Note: Age here is a binary covariate created around the median of 68.24

It is also of interest to see if there has been an improvement in survival over time. To check this a fit to the Cox model of age in the presence of year of diagnosis was carried out (see table 3.9). The coefficient of age compared to the age alone model is similar. Figure 3.10 suggests the distribution of age has increased marginally over the twenty year period of study, and hence the large improvement in survival is unlikely to be explained by this alone.

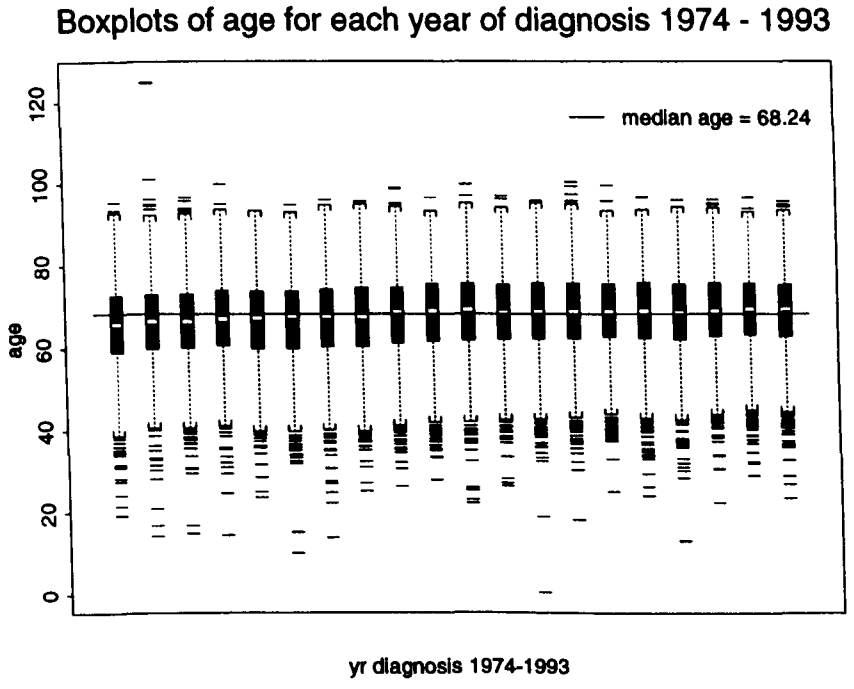


Figure 3.10: Age distribution for each year of diagnosis

model eqn	coeff age (s.e.) p	coeff yr diag (s.e.) p
age + yrdiag	0.0224 (0.0005) 0	-0.0236 (0.0009) 0

Table 3.9: Fit to lung data - age + year of diagnosis

3.5.4 Estimation of the Baseline Hazard

Once an estimate of $\underline{\beta}$ has been obtained, attention then turns to estimation of the baseline hazard, or equivalently the baseline survivor function/cumulative hazard using our estimate $\hat{\underline{\beta}}$. The original estimate from Cox (1972) was based on the conditional logistic model and is rarely applied in practice. Kalbfleisch and Prentice consider a discrete time model defined for the hazard in the jth interval $(t_{j-1}, t_j]$ as:

$$\lambda(t_j|z)dt = 1 - \{1 - \lambda_0(t_j)dt\} \exp(\underline{\beta}^T \underline{x}) \tag{3.38}$$

It was shown that the maximum likelihood estimate of the baseline hazard $\lambda_0(t_{(j)}) = 1 - \hat{\xi}_j$ is given by the solution to:

$$\sum_{(l \in D(t_{(j)}))} \frac{\exp(\hat{\underline{\beta}}^T \underline{x}_l)}{1 - (\hat{\xi}_{(j)})^{\exp(\hat{\underline{\beta}}^T \underline{x}_l)}} = \sum_{(l \in R(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x}_l) \tag{3.39}$$

With a single failure at $t_{(j)}$ the above can be solved analytically to give:

$$\hat{\xi}_{(j)} = 1 - \frac{\exp(\hat{\underline{\beta}}^T \underline{x}_{(j)})}{\sum_{(l \in R(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x}_l)} \tag{3.40}$$

Approximations to the baseline hazard have been suggested, in particular:

$$\begin{aligned} \hat{\xi}_{(j)} &\approx \exp \frac{-m_j}{\sum_{(l \in R(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x}_l)} \\ \hat{\lambda}_0(t_{(j)}) = 1 - \hat{\xi}_{(j)} &\approx \frac{m_j}{\sum_{(l \in R(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x}_l)} \end{aligned} \tag{3.41}$$

The latter approximation can be used in conjunction with Breslow's approximation for ties in the partial likelihood. The resulting estimate of the survivor function is equivalent to the non-parametric Kaplan-Meier curve. If the Efron approximation for ties is employed, a different approximation of the baseline hazard should then be adopted (Therneau, 1994):

$$\hat{\lambda}_0(t_{(j)}) \approx \sum_{k=1}^{m_j} \frac{1}{\{\sum_{(l \in R(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x}_l) - \frac{(k-1)}{m_j} \sum_{(l \in D(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x}_l)\}} \tag{3.42}$$

Recall our example of the previous section for tied data, we have three tied times and a further two times at risk with relative risks $\psi(j) = \exp(\underline{\beta}^T \underline{x}_j), j = 1, \ldots, 5$. The relevant estimates of the baseline hazard are:

$$\hat{\lambda}_0(\text{Breslow})(t_{(j)}) \approx \frac{3}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)}$$

$$\hat{\lambda}_0(\text{Efron})(t_{(j)}) \approx \left(\frac{1}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} + \frac{1}{\frac{2}{3}\psi(1)+\frac{2}{3}\psi(2)+\frac{2}{3}\psi(3)+\psi(4)+\psi(5)} + \frac{1}{\frac{1}{3}\psi(1)+\frac{1}{3}\psi(2)+\frac{1}{3}\psi(3)+\psi(4)+\psi(5)}\right)$$

Note that when $\underline{\beta} = \underline{0}$ i.e. $\psi(j) = 1$, $j = 1, \ldots, 5$ the estimates are equivalent to the contributions at each death time of the uncorrected and corrected Nelson cumulative hazard estimates, namely $\frac{3}{5}$ and $\frac{1}{5} + \frac{1}{4} + \frac{1}{3}$.

The resulting estimates are step functions, and given the cumulative hazard estimate $\hat{\Lambda}_0(t)$ we estimate the survivor function by $\hat{S}_0(t) = \exp(-\hat{\Lambda}_0(t))$. When the survivor function for a given set of covariable values has been estimated, it is then possible to examine quantities of interest such as the median predicted survival.

3.5.5 Estimation of the Baseline Hazard for Cox Model Fits to Lung Cancer Data

As we have employed the Efron correction for ties (the default option in Splus) we use the equivalent estimate of the baseline hazard. For the Cox model fit which had binary covariate age as its only term, figures 3.11 and 3.12 show the estimated survivor curves and cumulative hazard functions. Note that the cumulative hazards are parallel as constrained by the assumption of proportional

hazards.

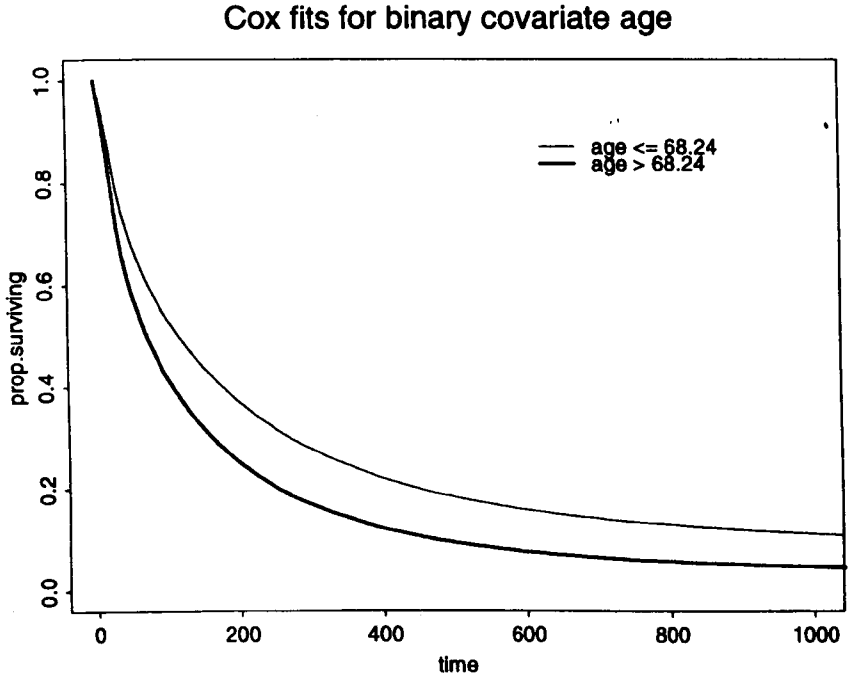


Figure 3.11: Estimated survivor curves for factor age from Cox model fit

It is therefore possible to predict survival for an individual. Table 3.10 gives the predicted median survival for binary covariate age. For the model fitted for continuous covariate age and continuous covariate year of diagnosis table 3.11 shows the predicted median survival for ages 55 and 80 (representing the 10th and 90th percentile of the age distributions) and years 1974 and 1993. For a given age predicted survival has increased about two-fold over the period under study.

covariable		median survival (days)	95 % C.I.
age	≤ 68.24	116	(112,118)
	> 68.24	72	(70,73)

Table 3.10: Predicted median survival estimates for binary covariate age

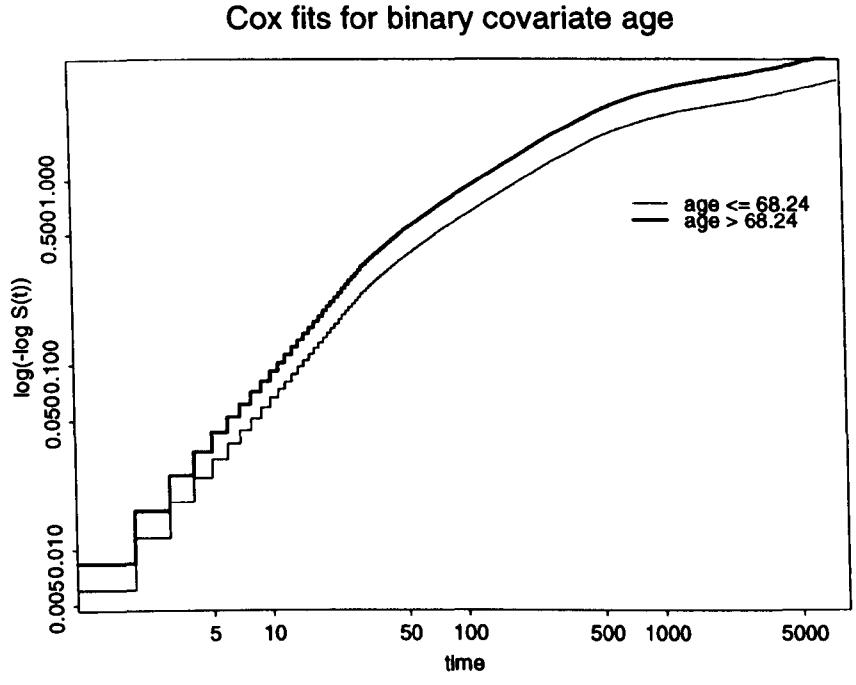


Figure 3.12: Estimated cumulative hazard for factor age from Cox model fit

age	year diagnosis	median survival (days)	95 % C.I.
55	1974	99	(96,103)
80	1974	47	(46,49)
55	1993	195	(187,203)
80	1993	85	(82,88)

Table 3.11: Predicted median survival estimates for continuous covariate age and year of diagnosis

3.5.6 Testing the Assumption of Proportional Hazards

The assumption of proportional hazards implies the hazard ratio for two different values of a covariate is constant over time. For a binary covariate we have $\lambda(t|x = 1)/\lambda(t|x = 0) = \exp(\beta)$. If the coefficient of x varies over time, then the assumption of proportional hazards is inappropriate, however the purpose of this thesis is not to assess violations from proportional hazards in the true survival data. Current work on mis-specification of survival models is being undertaken

by Mr Paul Monaghan as a sister project to this, and the reader is referred to his thesis. Outcome error may imply that the observed survival data are no longer proportional hazards (see chapter 5), but we still assume the underlying true data follow a proportional hazards representation.

A review of graphical methods is presented by Hess (1995). The simplest method is to compare the fitted model with Kaplan-Meier curves, but Therneau (1994) argues against this as violations are difficult to assess. However, examination of the Kaplan-Meier curves alone does give us a handle on the assumption of proportional hazards. The survivor curves cannot cross, and a plot of the log cumulative hazards should show parallel curves.

If one suspects that the coefficient does change over time it is possible to split the time axis and fit the model within each time period. The selection of the intervals is however unclear. Many residuals and significance tests of the proportional hazards assumption have been suggested (Collett chapter 5) but we are not considering these.

3.6 Parametric Proportional Hazards

If we assume a parametric form of the baseline hazard, then different methods for estimation of $\underline{\beta}$ are appropriate. Parametric Weibull models have the same shape parameter ρ and a different scale parameter $\tau = \exp(\underline{\beta}^T \underline{x})$. If one suspects the baseline hazard is from a Weibull distribution, a simple test of this is to plot the log-cumulative hazard against log time, as

$$\log \Lambda(t) = \log \tau + \rho \log t \tag{3.43}$$

is linear for the Weibull model. If the gradient of the plot is ≈ 1 then there is further evidence the data are from the exponential model.

The likelihood for n observations, where $\delta_i = 1$ if the ith observation is a death time and $\delta_i = 0$ if the ith observation is censored is defined as:

$$L(t|\underline{\beta}) = \prod_{i=1}^{n} (f_T(t|\underline{\beta}))^{\delta_i} (S(t|\underline{\beta}))^{(1-\delta_i)} \tag{3.44}$$

Parameter estimates are then obtained by maximising the likelihood with respect to $\underline{\beta}$ and confidence intervals, significance tests and model selection can be performed using the same techniques as outlined for the Cox model.

When the data truly have a parametric baseline and $\underline{\beta} \neq 0$ then the *efficiency* of the parametric maximum likelihood estimate will be increased compared to estimation via the Cox partial likelihood (not specifying the baseline). This means the estimate of $\underline{\beta}$ will have reduced variance. Discussion of this is given in Kalbfleisch and Prentice (Chapter4), in particular the case of the exponential distribution. Intuitively correctly specifying the baseline hazard for all times via parameterisation is likely to be more efficient than assuming no information on the baseline hazard between death times.

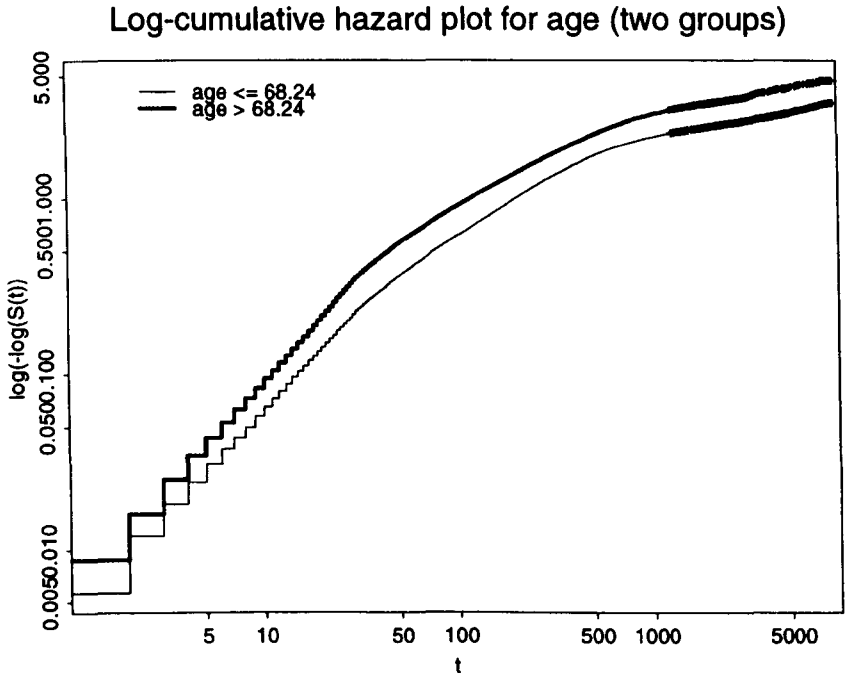


Figure 3.13: Estimated cumulative hazard for factor age

3.6.1 Appropriateness of Weibull Assumption to Lung Cancer Data

Figure 3.13 shows the estimated cumulative hazard function for the binary covariate age. The assumption of proportional hazards does not seem unreasonable (the vertical distance between the lines is approximately constant), but curves are clearly not linear. This would lead us to conclude that the fitting of a parametric Weibull model to the lung cancer data is inappropriate.

3.7 Summary

In this chapter we have reviewed methodology for non-parametric and semi-parametric survival methods. The idea of proportional hazards and the Cox

model were introduced. The Cox model assumes covariates act multiplicatively on an arbitrary baseline hazard. Estimation is achieved via partial likelihood. Where survival times display ties (a form of rounding measurement error) exact likelihoods have been proposed but due to their computational expense are unpopular. Therefore approximations to the partial likelihood are the norm, with S-plus employing the Efron (1977) approximation as a default. This is a vital springboard to the development of an approximation for more general outcome error in chapter 5. Fully parametric analyses were also briefly discussed.

An analysis of the full lung cancer dataset was undertaken. Fitted Kaplan-Meier curves demonstrated how overall survival for patients is extremely poor, with the median survival being 3 months. The effect of age is an important one; the log-rank test for ten age groups created around the deciles of the age distribution displayed strong significance. Cox fits for continuous and binary covariate age reinforced this result within the framework of proportional hazards. Ten years of age increased the hazard by 1.229. Sex is unlikely to affect survival. There is also strong evidence that survival has improved during the period under study. Predicted survival for a 55 year old in 1974 was 99 days (95 % C.I. (96,103)) and predicted survival for a 55 year old in 1993 was 195 days (187,203)).

In the next chapter we familiarise the reader with covariate and outcome measurement error and give a full literature review of measurement error methodology with particular reference to the Cox model.

Chapter 4

Review of Measurement Error

In this chapter we review the topic of measurement error, giving particular attention to the Cox proportional hazards model introduced in the previous chapter. Covariate and outcome error are both reviewed. The particular measurement error problem considered in this thesis is that of a window of measurement for certain patient's diagnosis time (see chapters 2 and 6). This has obvious implications for recording of survival time and may imply that covariates such as age at diagnosis are also subject to measurement error.

4.1 Measurement Error in Statistical Models

Measurement error in the two variable, linear regression model is not a new problem. The earliest reference in the introduction to the Errors-in-Variables workshop (Byar and Gail, 1989) published by Statistics in Medicine is from 1878. The field has grown enormously since the mid 20th century and work in the last twenty years covers many types of statistical models and many different ways

of coping with the problem of measurement error. Two main texts have been published on measurement error: Fuller (1987) extensively covers the case of linear models and Carroll, Ruppert and Stefanski (1995) cover nonlinear models, in particular generalised linear models.

The need to address the problem of measurement error arises from the *bias* in regression parameter estimates in the presence of measurement error. This is often called *attenuation* and estimates ignoring measurement error are frequently called *naive*. The usual direction of such bias, particularly in univariate analyses, is towards the null hypothesis of no covariate effect, however this is not always the case. Figure 4.1 shows a graphical demonstration of this bias for hypothetical data.

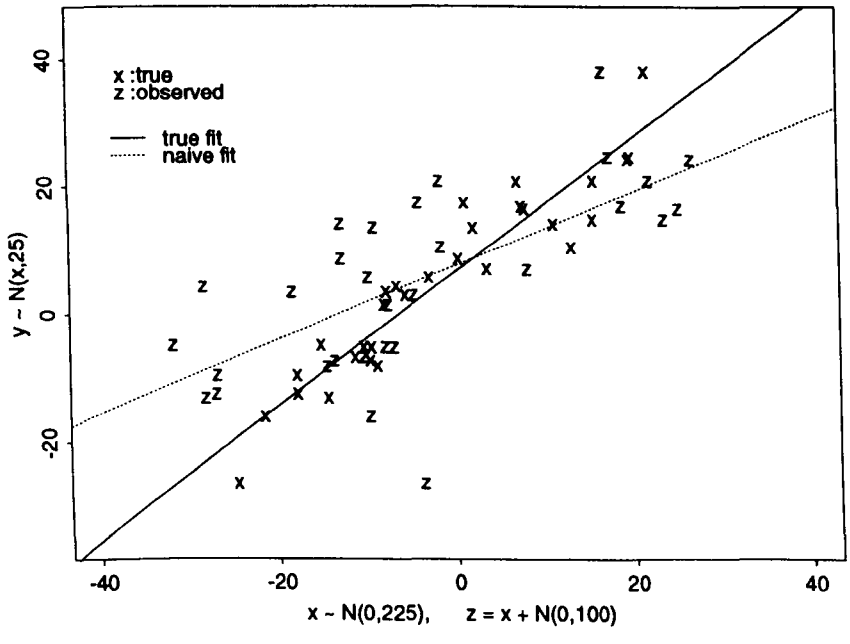


Figure 4.1: Attenuation due to measurement error - hypothetical example

There are many ways to approach a measurement error problem, and the method adopted depends on many considerations. Broadly speaking, the difference between an error free analysis and an error correction analysis is that at least one further model (in addition to the usual model of outcome on covariables) has to be specified. This is often referred to as a *measurement model* (Clayton, 1991) and relates the unobserved true variables to their observed counterparts. A further assumption regarding the underlying distribution of the true variables will also frequently be considered.

4.1.1 Types of Error, Measurement Models and Assumptions

Measurement errors can be broadly split into four categories. Random errors imply that the mean of the observed data will consistently estimate the mean of the true data whereas systematic errors have an overall mean bias compared to the true data. Within person error represents errors across replicated measurements for the same individual, and can be either random or systematic. Between person error represents errors across a sample of individuals and again may or may not be consistent for the mean of the whole group (Willet, 1989).

Let us consider the usual simple linear regression model:

$$Y_i = \alpha + \beta X_i + \epsilon_i \qquad \text{where} \qquad \epsilon_i \sim N(0, \sigma_\epsilon^2) \tag{4.1}$$

Instead of observing X_i we observe Z_i, a distortion of X_i due to error.

Differential and Non-Differential Measurement Error

If we assume that given the true X_i then the Z_i have no further value in the model of the outcome variable Y_i then the measurement error is regarded as *non-differential.* Other terms for this are that Z_i is a *surrogate* for X_i or that Z_i is *conditionally independent* of Y_i given X_i . This can be expressed mathematically as:

$$f(Y|X,Z) = f(Y|X) \tag{4.2}$$

Differential error is more problematic and is not considered further as a correction is usually impossible (Willet, 1989). One example of differential error is relating diet to breast cancer. The true data are diet pre-diagnosis but the observed diet is post-diagnosis. The observed diet may thus have changed as a result of diagnosis, and is hence correlated with the outcome variable (see Carroll, Ruppert and Stefanski(page 16)).

Errors-in-Variables and Berkson models

The traditional *errors-in-variables* assumption is that the relation between true and observed is $Z_i = X_i + u_i$ or a more general function where the observed Z_i is a distortion of the true X_i with a random error u_i. This is often seen as the most natural measurement model. The basic Berkson (1950) model states that $X_i = Z_i + u_i$. Moran (1971) suggests the Berkson model when one aims for a particular X but fails to achieve it such as in a laboratory experiment when an instrument is set to a particular value. In terms of estimation the two models are not interchangeable even in the most simple form given above. Note in the Berkson case the observed data Z_i are uncorrelated with the measurement errors

u_i.

Functional and Structural Models

Another consideration is how to regard the unobserved true X_i's. Two main types of assumption are common. The *functional* assumption regards the true data as fixed constants whereas the *structural* model assumes an underlying distribution for the true values (usually normal). For the case of simple linear regression, a further model, the *ultrastructural* model has been considered where the X_i's are normally distributed with common variance but have an individual mean μ_i (Cheng and Van Ness, 1994).

The functional model is regarded as appropriate when the sample X_i's are not a random sample of a general population, but rather come from a designed experiment (see Fuller and Carroll, Ruppert and Stefanski (page 7)). Carroll, Ruppert and Stefanski generalise the definition of functional models to those where any underlying distributional assumption is not important to the correction procedure.

4.1.2 Examples of Measurement Error

There are a rich literature of examples of measurement errors - epidemiological exposures are particularly prone. When a binary factor or a categorical factor is measured, the measurement error is called *misclassification.* Examples of such errors are stage data in cancer studies. Whittemore (1990) provides an example of Poisson regression of breast mortality from the San Francisco Cancer Registry where under a misclassification analysis the assumption of proportional

hazards is accepted. Other possibilities of misclassification arise from survey sampling such as level of smoking history (e.g. number of packets smoked a day) or surveys of diet and nutrition. A recent thesis by Seyed Hassan Saneii (1997) considered psychiatric survey data where binary and ordinal covariates were missing or subject to misclassification. Diet related mismeasurements, such as fat intake or alcohol consumption, whether continuous or categorical, are a continuous theme in the literature (Clayton, 1991; Rosner *et al.*, 1989; Huakka, 1995).

Another popular example is measurement of air pollution exposure, in particular the level of NO_2 an individual has received (Whittemore and Keller, 1988; Stephens and Dellaportas, 1992; Hasabelnaby *et al.*, 1989; Tosteson *et al.*, 1989). A similar type of exposure problem that has been a popular example in the proportional hazards papers is the dose of radiation exposure in the wake of the atomic bombing of Japan. Prentice considers this as a Berkson formulation and Nakamura as an errors-in-variables problem.

Specific medical measurements that can be measured with error are blood pressure measurements (Hughes *et al.*, 1995) and CD4 counts in relation to AIDS (Satten and Longini, 1996).

4.2 Correction for Measurement Error

A number of methods are available in order to correct for measurement error . One common theme is that in order to perform a measurement error analysis,

the extent of error needs to be estimated. This is usually done via *validation* data or *replication* data. Validation data would typically be available for a small subset of the data, in which both the true X_i and observed Z_i are available and hence the extent of error can be determined. Replication data are available if more than one imperfect Z measurement is available.

4.2.1 Simple Linear Regression

For the simple linear regression case we have the true model :

$$E(Y|X) = \alpha + \beta X \tag{4.3}$$

$$\text{var}(Y|X) = \sigma_\epsilon^2 \tag{4.4}$$

When the usual linear regression model is incorrectly applied to the observed data Z, we have an *induced* regression model:

$$E(Y|Z) = \alpha + \beta E(X|Z) \tag{4.5}$$

The above demonstrates the fundamental difference between the errors-in-variables and the Berkson cases.

$$\text{e-i-v} \quad : \quad Z_i = X_i + u_i \Longrightarrow E(X|Z) \neq Z \tag{4.6}$$

$$\text{Berkson} \quad : \quad X_i = Z_i + u_i \Longrightarrow E(X|Z) = Z \tag{4.7}$$

The reason for the attenuation in the errors-in-variables case is due to the regression of the underlying true on the observed. In the basic Berkson

formulation there is no attenuation of β. In fact for errors-in-variables if the true data are $N(\mu, \sigma_X^2)$ then (Clayton, 1991):

$$E(X|Z) = \{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2}\}Z + (1 - \{\frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2}\})\mu \tag{4.8}$$

A full derivation of the distribution of $X|Z$ is given in appendix A.1. For the errors-in-variables case the degree of attenuation for β thus depends on the "reliability ratio" :

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_u^2} = \frac{\sigma_Z^2 - \sigma_u^2}{\sigma_Z^2} \tag{4.9}$$

where σ_X^2 is the variance of the true exposures and σ_u^2 the variance of the measurement error (Clayton, 1991; Carroll, 1989; Fuller, 1987). Since $0 < \lambda < 1$ then the true β_x is always underestimated. The induced model for the error-in-variables case is:

$$E(Y|Z) = \alpha + \beta(\lambda Z + (1 - \lambda)\mu) = \alpha + \beta\mu(1 - \lambda) + \lambda\beta Z \tag{4.10}$$

If λ can be estimated or certain assumptions about it (such as $\sigma_X^2 = \sigma_u^2$) are made then the "naive" estimate of β can be corrected :

$$\beta_X = \beta_Z(\lambda^{-1}) \tag{4.11}$$

The only effect in the Berkson case is to increase the residual variance, a trait of both types of model.

$$\text{e-in-v} \quad : \quad \text{var}(Y|Z) = \sigma_\epsilon^2 + \frac{\beta^2\sigma_u^2\sigma_X^2}{\sigma_u^2 + \sigma_X^2} \tag{4.12}$$

$$\text{Berkson} \quad : \quad \text{var}(Y|Z) = \sigma_\epsilon^2 + \beta^2 \tag{4.13}$$

Note that at $\beta = 0$ the naive induced model is equivalent to the true model. This implies that the usual score test of whether $\beta = 0$ is valid, but has reduced power.

Variance of the regression parameters

A correction for measurement error will consistently estimate the slope parameter, but by recognising the extra variability within the model will increase the variance of the regression parameters. The review of Cheng and Van Ness (1994) outlines the difficulty in obtaining valid confidence intervals and in some cases their non-existence. The simplest calculation of confidence intervals in this problem is a bootstrap resampling mechanism as described in the *Regression Calibration Algorithm* in Carroll,Ruppert and Stefanski chapter 3.

Denoting the naive estimate as β_{naive} then we have $E(\beta_{naive}) = \lambda\beta$ and $\text{var}(\beta_{naive}) = \sigma^2_{naive}$. The method of moments correction for beta $\beta_{corr} = \lambda^{-1}\beta_{naive}$ has expectation $E(\beta_{corr}) = \beta$ and variance $\text{var}(\beta_{corr}) = \lambda^{-2}\sigma^2_{naive}$. This leads to a "bias versus variance tradeoff". Consideration of the mean squared error of both estimates can then be implemented if required.

Multivariate Models

If, in addition to the incorrectly measured covariable, there are other explanatory variables within the regression then the estimates of these variables can also be biased, sometimes showing the opposite to the true effect. This can only be

avoided if the variables are independent. Examination of this is in various papers such as Carroll (1989) and Marshall and Hastrup (1996) and has implications for design. Nakamura and Akazawa (1994a) examine an unbalanced confounding variable in the Cox Model.

4.2.2 Literature review

Methods for the correction of measurement error come under a broad umbrella. This section provides a literature review for models encountered in epidemiology other than survival models.

The previous section developed regression calibration in the context of simple linear regression. The corrected estimate is consistent in this case, but in other models it is only an approximate correction procedure. The idea of calibration is simple, in that one replaces the observed covariate by the expectation of the true covariate given the observed value. The case of logistic regression is covered by Rosner, Willet and Spiegelman (1989) who also derive corrected confidence intervals. Tosteson et al (1989) consider binary and ordinal probit models and derive a test for the conditional independence assumption. Carroll (1989) states that Poisson and gamma regression with a log-linear link function consistently estimates the slope parameter for randomised studies.

A recent method for the treatment of measurement error is the SIMEX algorithm (simulation extrapolation). This is a simple idea that demonstrates the effect of and corrects for measurement error. The idea is that the naive estimate at $\lambda = 0$ has measurement error σ_u^2. Additional error data $\lambda_M \sigma_u^2$ is generated for

M datasets and added to the original data to give the total measurement error $\sigma_u^2+\lambda_M\sigma_u^2 = (1+\lambda_M)\sigma_u^2$. Hence a corrected estimate is obtained by extrapolating to the case $\lambda = -1$ (Carroll *et al.*, 1995).

Likelihood, approximate likelihood via expansion and quasi-likelihood methods have not received extensive treatment by authors. The structural likelihood for the true data given the observed covariate Z where θ are the parameters of the measurement model is:

$$L(Y|Z,\underline{\beta},\theta) = \int f(Y|X,\underline{\beta})f(X|Z,\theta)dx$$

(Carroll, 1989). Quasi-likelihood methods model the mean and variance of the above, and in many ways are simpler to fit. Note that these methods are structural as deriving $f(X|Z)$ requires the specification of $f(X)$.

Rosner et al (1989) expand the likelihood for logistic regression, while Schafer (1993) treats the probit model. An example of a likelihood analysis for misclassification and Poisson data is Whittemore (1990).

Functional methods have the advantage of not requiring the specification of the distribution of the underlying true covariates. The functional likelihood is defined as :

$$L(Y|\underline{\beta}, X_1, \ldots X_n) = \Pi_{i=1}^n f(Y|, \underline{\beta}, X)f(Z|X,u)$$

This is an over-parameterised problem, as the likelihood has to be maximised with respect to the n true X covariates as well as β. Thus for linear regression

we estimate $n + 3$ parameters $(X_i, \alpha, \beta, \sigma)$. For linear regression this likelihood is consistent (Cheng and Van Ness, 1994). However the functional model is not usually consistent for non-linear models, in particular the logistic model (Nakamura, 1990; Stefanski and Carroll, 1987).

Stefanski and Carrol (1985) discuss the functional logistic model and generalise this to generalised linear models in their later 1987 paper. They propose a *conditional score* . Conditioning the density of Y on the observed data and a sufficient statistic for the true data the new model is also a generalised linear model whose score equations are unbiased for β. An alternative method ,*corrected score* is proposed by Nakamura (1990). He derives a score equation under the observed data with global expectation equal to the score of the corrected data. This exists for nearly all generalised linear models except the logistic case.

Bayesian methods for measurement error have received more attention with the advent of computer intensive methods such as GIBBS sampling. Conditional independence is an inherent idea in the implementation of such methodology. Generalised linear models for both Berkson and errors-in-variables problems were considered by Stephens and Dellaportas (1992, 1995). Richardson and Gilks (1993) provide a review of conditional independence models for epidemiological studies with covariate measurement error. Computer software that allows the specification of a vast range of models has been developed (Spiegelhalter *et al.*, 1995) and examples of measurement error fits using the BUGS software are given.

A bootstrap approach for generalised linear models is presented by Huakka (1995). The method was fitted to a multivariate logistic model and compared with the likelihood approximation estimate of Rosner, Willet and Spigelman. Huakka found his bootstrap approach gave smaller confidence intervals.

A suite of computer programs for S-Plus designed to fit many correction procedures including regression calibration and maximum likelihood for both binary and continuous covariates is presented by Bashir and Duffy (1995).

Measurement error is not restricted to epidemiology, and estimation for other types of problem such as complex survey data (Fuller, 1995) have been addressed. However epidemiological applications are common. Despite this use of the methodology has not been widespread. Lack of validation data is one possibility. Willet (1989) cites this and other reasons. Many procedures are bivariate, lack confidence intervals and focus on unpopular models such as the probit model. He also suggests a bias towards the null hypothesis may (wrongly) be regarded as acceptable by some, while others may be put off by over-technical presentation.

4.3 Covariate Error and the Cox Model

4.3.1 Attenuation and the Cox Proportional Hazards Model

Recall from the previous chapter that the usual Cox model is defined as:

$$\lambda(t, x) = \lambda_0(t) \exp(\underline{\beta}^T \underline{x}) \tag{4.14}$$

For simplicity consider a single covariate X. As for the case of linear regression, if instead of observing the covariate X we observe the stochastically related variable $Z = X+U$ and a Cox model is fitted via the usual partial likelihood with the Z treated as true data then the resulting estimate of β will be attenuated towards the null hypothesis of $\beta = 0$. The extent of this attenuation is however not as clear as in the usual case.

Nakamura and Akazawa (1994a) and Hughes (1993) both examine the extent of attenuation in the Cox model. Hughes considers the naive partial likelihood and derives some important results. Under the case of no censoring, he shows that the asymptotic expectation of the naive score statistic is not dependent on the baseline hazard $\lambda_0(t)$. The expected score can then be solved numerically to determine the degree of bias. Hughes gives two figures that show the degree of bias for different values of β and levels of measurement error when the true covariable has a standard normal distribution.Some of the results are shown in table 4.1, where the approximations are read from figure 1 (Hughes p.1058) and β_Z = (degree of attenuation)β_X.

Recall that for linear regression the degree of dilution of β is determined by $\lambda = \frac{\sigma_X^2}{\sigma_X^2+\sigma_u^2}$. Hughes' results show that for the case of no censoring (or a failure rate of 100 %) the actual degree of bias can be much larger. This increases as β increases. Nakamura and Akazawa consider simulated data with 80 % failure and conclude that if $|\beta|$ is small the degree of bias is λ but for $\beta = 1$ the degree

β_{true}	σ_u^2	degree of attenuation
0.5	0.1	0.983
0.5	0.5	0.625
0.5	1	0.452
0.5	2.5	≈ 0.25
1	0.1	0.855
1	0.5	0.546
1	1	0.379
1	2.5	≈ 0.19
1.5	1	≈ 0.30
1.5	2.5	≈ 0.16
2	1	≈ 0.25
2	2.5	≈ 0.13

Table 4.1: Levels of bias for Cox model with no censoring

of bias is close to $\frac{\sigma_X^2}{\sigma_X^2+(\sigma_u+0.1)^2}$.

Hughes also derives the score under two censorship schemes, the first is when censoring occurs if the patient is still alive at the end of a study and the second is a particular form of random censorship. Under random censorship the baseline hazard does not drop out of the asymptotic naive score equation and hence the degree of attenuation is also dependent on the baseline hazard. The effect of this in the resulting simulation study was marginal however.

The degree of attenuation is reduced when censorship is present and with only 20 % failure the degree of bias is extremely close to λ, the reliability ratio for linear regression. The relationship between attenuation for no censorship and attenuation for high censorship is approximately linear, and one can thus interpolate the degree of bias for any censorship level.

Nakamura and Akazawa also study the effect on a second binary treatment variable Δ when Z is a confounding variable and is unbalanced between levels of treatment. The naive estimate of Δ can lead to a reverse conclusion of effect. This is analogous with results on multivariate linear regression.

4.3.2 Correction for Attenuation and the Cox Model

Correction for covariate error in the Cox model has received some attention by authors. The first papers to appear in the literature are that of Prentice (1982a, 1982b). More recent papers are Pepe, Self and Prentice (1989) and Clayton (1991) who further the previous work of Prentice. A different approach to measurement error is that of Nakamura (1992).

Adjusting β directly from the naive estimate

The discussion of the level of attenuation in the previous section allows an extremely simple correction for β. Hughes paper allows us to calculate the level of dilution for high censorship and no censorship under a considerable array of situations after λ is estimated, and through interpolation for the degree of censoring a good estimate of the level of attenuation can be found. The inverse of this is our "adjustment factor" for the observed β_Z in order to estimate β_X.

As the results of Hughes are asymptotic, he undertakes a simulation study to examine the adjustment factor for small samples. Although the distribution of the corrected estimates were skew they still performed well in overall coverage of the true β.

Regression Calibration and the Cox Model

A regression calibration method for the Cox model is the subject of the early papers on covariate error. The basic approach to regression calibration is described in chapter 3 of Carroll, Ruppert and Stefanski (1995), namely to replace the observed covariate with an estimate for true given observed via the measurement model. This method is only appropriate in the Cox model if further assumptions are made.

Prentice (1982a) induces a hazard model:

$$\lambda(t,z) = \lambda_0(t) E_{\{T \geq t, z\}} \exp(\beta x) \tag{4.15}$$

via the *conditional independence* assumption:

$$\lambda(t,x,z) = \lambda(t,x) \tag{4.16}$$

and the *censorship* assumption:

$$\lambda(t,x,\text{no censorship in [0,t)}) = \lambda(t,x) \tag{4.17}$$

i.e. that the measured covariate does not inform the censorship mechanism.

The key point is that the induced hazard rarely takes the proportional hazards form and the induced relative risk will usually include the baseline hazard. Prentice (1982b) shows for example that if the true relative risk is of the form $1+\beta X$ and $X|Z \sim N(\mu_Z, \sigma_Z^2)$ then the induced model is:

$$\lambda(t|z) = \lambda_0(t)[1 + \{\mu_Z - \int_0^t \lambda_0(u)du\beta\sigma_Z^2\}\beta] \tag{4.18}$$

A more general form of the induced model with exponential relative risk is also given and is shown to be:

$$\lambda(t|z) = \lambda_0(t)\frac{\int \exp(\beta x)\exp\{-\exp(\beta x)\int_0^t \lambda_0(u)du\}f(x|z)dz}{\int \exp(\beta x)\int_0^t \lambda_0(u)du f(x|z)dx} \tag{4.19}$$

The usual partial likelihood for the error free model is:

$$L(\beta) = \Pi_{j=1}^{r}\frac{\exp(\beta x)}{\Sigma_{l\in R(t_{(j)})}\exp(\beta x_l)} \tag{4.20}$$

where the product is over the r death times from the n observed survival times.

Using the principle of regression calibration this can be replaced by the likelihood:

$$L(\beta) = \Pi_{j=1}^{r}\frac{E_{\{T\geq t_j,z_j\}}\exp(\beta x_j)}{\Sigma_{l\in R(t_{(j)})}E_{\{T\geq t_j,z_l\}}\exp(\beta x_l)} \tag{4.21}$$

It is useful to consider when one can apply calibration without fear of the inclusion of the baseline hazard in the induced relative risk. This is reasonable if $pr(T \geq t|X) \approx pr(T \geq t|z)$. For this to hold, a further *rare failure* assumption is introduced, where $pr(T \geq t|X) \approx 1$. In the example used by Prentice, where failure was succumbing to thyroid cancer following the atomic bombing of Japan this was reasonable as less than 2% of those studied actually failed. A further possibility is if $f(X|Z)$ is "concentrated" implying that the measurement error is small. These assumptions allows interesting comparisons with the work of

Hughes. When censorship is large (failure is rare) then the attenuation level is close to that of linear regression as is the case if measurement error is small. Thus under these conditions regression calibration as practiced for linear models is appropriate.

The reason for the inclusion of the baseline hazard is that we require the distribution of $X|Z$ within each risk set. If failure is large then the distribution is likely to change over time and this is dependent on the baseline hazard as well as the relative risk. Certain distributions do hold their structure in these circumstances and these are discussed in Pepe, Self and Prentice.

A nice result given by Prentice (1982a) is that when $\beta = 0$ the induced model is the same as the true model and hence the score test for $\beta = 0$ is still appropriate, as is the usual case for other induced models.

Clayton (1991) proposed a modification to the regression calibration method of Prentice that removes the complication of the $\{T \geq t, z\}$ conditioning. Clayton instead suggests calibrating the model within each risk set. He also suggests a normal discriminant model under the assumption of normality for the true covariable and measurement error leading to the estimate for β:

$$\hat{\beta} = \frac{\Sigma_i \hat{\lambda}_i [z_i - \bar{z}_i]}{\Sigma_i \hat{\lambda}_i \hat{\sigma}_i^2} \tag{4.22}$$

where $\hat{\lambda}_i$ is the estimated reliability ratio in the ith risk set and σ_i^2 is the estimate of the variance of the underlying true covariates. The results for this estimate and the regression calibration estimate were very similar.

The induced model 4.18 for the linear relative risk with normal errors was dependent only of the cumulative baseline hazard. Prentice (1982b) and Pepe, Self and Prentice (1989) suggest using a Kaplan-Meier type estimate of the cumulative baseline and then iterating with the partial likelihood.

Pepe, Self and Prentice also present a product partial likelihood for joint estimation of the parameters of the measurement model and the parameters of the regression model when a validation study is available. The resulting estimates are consistent and asymptotically normal, but the same issues for regression calibration and the induced model still apply, i.e. the method is most appropriate for studies where failure is rare. The product partial likelihood for exponential relative risk with normal errors is derived. In addition a study is undertaken to determine the size of data set in relation to the size of validation study for the most cost effective design.

Although we are not considering time dependent covariates in this thesis, Tsiatis, DeGruttola and Wulfson (1995) consider a regression calibration for time dependent covariables with missing values and measurement error. Covariates follow a Gaussian stochastic process that is estimated and then the conditional expectation is used in the partial likelihood.

Corrected Score and the Cox Model

Nakamura (1992) proposes an approximate corrected score for the Cox model. This follows from his 1990 paper where an exact corrected score is derived for certain generalised linear models (Nakamura, 1990). The approximation using a Taylor series expansion arises from the fact that no exact corrected score is available for the Cox model. Recall that a corrected score for the observed data has a global expectation equal to the score of the true covariable. For the Cox model with one covariate the score for the jth failure is defined as follows:

$$U_j(\beta, x, T) = x_{(j)} - \frac{\Sigma_{l \in R(t_{(j)})} x_l \exp(\beta x_l)}{\Sigma_{l \in R(t_{(j)})} \exp(\beta x_l)} \tag{4.23}$$

For additive normal errors with variance σ_U^2 Nakamura defines the (first order) correction that holds if $\beta \sigma_U^2 \beta$ is small:

$$U_j^*(\beta, z, T) = U_j(\beta, z, T) + \sigma_U^2 \beta \tag{4.24}$$

A second order correction is also given :

$$U_j^*(\beta, z, T) = U_j(\beta, z, T) + \sigma_U^2 \beta \{1 - \frac{\Sigma_{l \in R(t_{(j)}} \exp(2\beta z_l)}{\{\Sigma_{l \in R(t_{(j)}} \exp(2\beta z_l)\}^2}\} \tag{4.25}$$

An approximate corrected observed information is also derived, and hence variance estimates are found in the traditional manner. Nakamura and Akazawa (1994a) and Nakamura (1992) examine the performance of β^* (the corrected estimate) and state this depends on $\beta \mathrm{SD}(X)$, $\frac{\sigma_U}{\mathrm{SD}(X)}$ and $\beta \sigma_U$. If these quantities are small the estimates should perform well. In the simulations where a binary treatment variable was also present the corrected estimate performed well in

estimating the treatment effect.

A computer program for fitting the corrected score equations is given in Nakamura and Akazawa (1994b). One problem with the practical use of this methodology, particularly when $|\beta\sigma_U|$ is relatively large (e.g. $\beta = 1, \sigma_U = 1$) is that negative values of the corrected information matrix can be encountered and hence estimation has to be stopped before completion.

Parametric Survival Models

Parametric survival models with covariate measurement error have received scant attention in the literature, although the exponential and gamma models are covered in the wide literature on generalised linear models (however the presence of censoring will complicate the corrected score or likelihood methodology). The Weibull model was briefly discussed by Prentice, but he concluded the induced model for normal errors with a linear hazard ratio was *"perhaps too complicated to expect much use"*. One important feature of a log-linear exponential model with no censoring is that using regression calibration will consistently model the slope parameter (Carroll, Ruppert and Stefanski).

4.4 Outcome Error in Statistical Models

The topic of outcome measurement error has not commanded the same attention from authors as covariate error, the principal reason being that outcome error in the case of many models (such as simple linear regression) will not bias the estimates of the model parameters, but merely increase the residual variance.

However Carroll, Ruppert and Stefanski do include a chapter on outcome error in their book. The principal method suggested is modelling the extra variance via quasi-likelihood, although the full likelihood of the observed data s_i, where $s_i = t_i + u_i$, is derived on page 236. The indicator $\Delta = 1(0)$ denotes the presence(absence) of validation data.

$$\prod_{i=1}^{n}\{f(s_i|t_i,x_i,\underline{\beta})f(t_i|x_i,\underline{\beta})\}^{\Delta_i}\{f(s_i|x_i,\underline{\beta})\}^{1-\Delta_i} \tag{4.26}$$

Correction for outcome measurement error in continuous time survival models is a problem not previously addressed, although in principle parametric models can be fitted using the above technique. The main difficulty is deriving the distribution of s_i, usually requiring numerical integration. Holt, McDonald and Skinner (1991) examine Weibull outcome error in the context of event history analysis via simulation, in particular different errors for different levels of a covariate and error varying according to size of t_i. They show regression coefficients have bias and this bias decreases with increased censoring. No correction for error is undertaken, but the authors point out that a least-squares fit via the log transformation of data incorporating independent multiplicative error will give an unbiased estimate of $\underline{\beta}$.

Sudman and Bradburn (1973) consider a particular type of outcome error in surveys. This is called telescoping, and is due to participants recall, where dates are brought forward from the true date towards the date the survey is taken. An example is age of first cigarette, where individuals might bring forward their smoking start time towards to the date of reporting.

4.5 Summary

In this chapter we have introduced the reader to, and reviewed, the topic of measurement error in statistical models. Particular attention was paid to the Cox proportional hazards model, where much work has been done on the problem of covariate measurement error. No work has however been done on the problem of survival time error for the Cox model, which is a feature of cancer registration data.

Many of the concepts introduced for covariate error are pertinent to the next chapter on outcome error, in particular the idea of a measurement model relating the true variable to the observed one. In the next chapter we consider both Berkson and errors-in-variables models (we use the term "Berkson" to imply that the distribution of true given observed is not dependent on the underlying true value). We examine the effect of and introduce a new approximate partial likelihood that incorporates weighted risk sets according to a measurement model. The concepts of a corrected score equation and regression calibration are important to the new procedure.

In chapter 6 we apply the new method to the lung cancer data introduced in chapter 2 and analysed in chapter 3 under the assumption of no measurement error.

Chapter 5

Outcome Error in the Cox Proportional Hazards Model

5.1 Introduction

Outcome error was reviewed in the previous chapter. No correction for outcome error has been derived for the Cox proportional hazards model. In this chapter we investigate the effect of outcome error on estimation in the Cox model and introduce a new method for dealing with survival error. The motivating need for new methodology is the problem of recording of diagnosis in cancer registration data.

5.2 Outcome Error in the Cox Proportional Hazards Model

Recall from chapters 3 and 4 that estimation for the Cox model is achieved via the *partial likelihood*:

$$\begin{aligned} PL(\beta) &= \prod_{j=1}^{r} f(\text{failure case } t_{(j)} | \text{censoring in} [t_{(j-1)}, t_{(j)}], \text{failure case } t_{(j-1)}; \underline{\beta}) \\ &= \frac{\exp(\sum_{j=1}^{r} \underline{\beta}^T \underline{x}_j)}{\prod_{j=1}^{r} \sum_{l \in R(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_l)} \end{aligned} \tag{5.1}$$

Due to the nature of the likelihood, the crucial factor in parameter estimation is the order of the failure times $t_{(j)}$ $j = 1, .., r$ and the respective risk sets at each failure time.

5.2.1 Rounding Error and the Partial Likelihood

Consider a survival data set where each failure is recorded up to the day of survival. In other words, if a patient dies during the fourth day of the study their recorded time is three. This can lead to tied data due to "rounding error". Continuous survival data does not permit tied data and various approximations for the true partial likelihood were given in section 3.5.2.

The simplest approximation was suggested by Breslow (1974) where one assumes each survival time occurs immediately before each other survival / censoring time. For a data set of n observations with r failures and death set $D(t_{(j)})$ for

the m_j failures at time $t_{(j)}$ we then have the approximate likelihood:

$$L(\beta) = \Pi_{j=1}^{r}[\frac{\Pi_{k\in D(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_k)}{[\Sigma_{l\in R(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_l)]^{m_j}}] \tag{5.2}$$

A better approximation is due to Efron (1977)

$$L(\beta) = \Pi_{j=1}^{r}[\frac{\Pi_{k\in D(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_k)}{\Pi_{k=1}^{m_j}[\Sigma_{l\in R(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_l) - (k-1)\frac{1}{m_j}\Sigma_{l\in D(t_{(j)})} \exp(\underline{\beta}^T \underline{x}_l)]}] \tag{5.3}$$

The Efron approximation includes each tied time in the denominator according to the probability of it being in each risk set, given that each possible risk set at the time $t_{(j)}$ is equally likely to occur. Two tied times $t_{(1)} = t_{(2)}$ would be included with weights 1 in the first risk set and $\frac{1}{2}$ in the second. Three ties have weights $1, \frac{2}{3}$ and $\frac{1}{3}$ in the first,second and third risk sets respectively.

It is also possible to examine this problem as a Berkson measurement error problem. For each tied time we then have a measurement model for $true|observed$ according to the rounding down of the times.

$$true|obs \sim U(obs, obs+1) \tag{5.4}$$

This model for $true|obs$ allows us to specify the probability each tied time is in each risk set and hence a weighted likelihood in the spirit of Efron. In fact for

this model the weights are identical to Efrons, but this will be explained more formally later in this chapter.

The treatment of censoring presents a complication to this idea since in all the previous approximations each censored time is assumed to occur after each death time. One possible model is:

$$\begin{aligned} true|obs_{fail} &\sim U(obs, obs + 0.99) \\ true|obs_{cens} &= obs + 0.995 \end{aligned}$$

The reduction in the upper level of the model for true given observed does not affect the weighting for each failure time provided the model is applied to each failure time.

5.2.2 Survival Time Measurement Error and the Partial Likelihood

Let us consider now that instead of observing the actual true survival time t_i we instead observe a related time s_i according to some measurement model. This would lead us to suspect bias in our estimates from the partial likelihood. For covariate error models we have a handle on the size and direction of this bias given our knowledge of the measurement and regression models employed.

The previous section discussed the fact that due to rounding error we do not know the true order of survival times, and hence could only approximate to the

underlying true partial likelihood. Measurement error on survival time presents us with a similar argument.

Consider five true survival observations t_i as in figure 5.1 with $\exp(\underline{\beta}^T \underline{x}_i) = \psi(i), i = 1, \ldots, 5$ then the partial likelihood is as follows

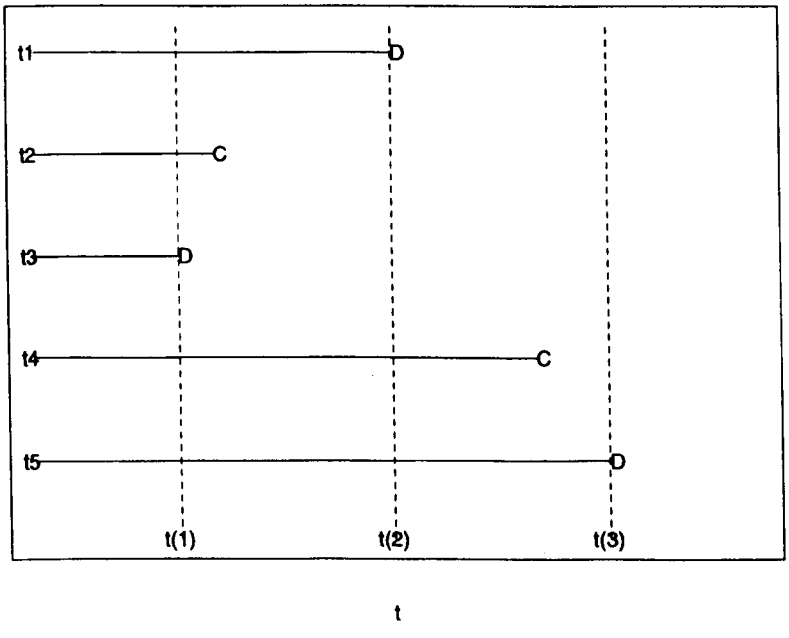


Figure 5.1: True survival times $t_1, \ldots, t_5$

$$\frac{\psi(3)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} * \frac{\psi(1)}{\psi(1)+\psi(4)+\psi(5)} * \frac{\psi(5)}{\psi(5)} \qquad (5.5)$$

If instead of observing t_i we observed $s_i = t_i + u_i \; i = 1, .., n$ but the extra noise did not change the *order* of the times i.e. $s_{(j)} = t_{(j)} \; j = 1, .., n$ then $\underline{\beta}$ would be estimated identically. This suggests that the Cox model may have a certain robustness against outcome measurement error. However this is just one realisation from a whole distribution of possible values. If the measurement error

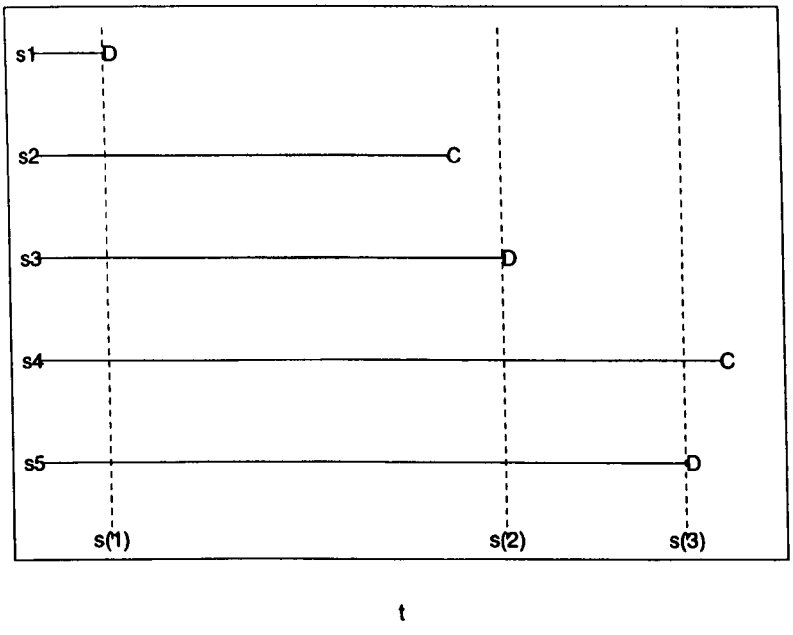


Figure 5.2: Observed survival times $s_1, \ldots, s_5$

led to the observed survival times s_i as in figure 5.2 then the partial likelihood would be as follows

$$\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)+\psi(4)+\psi(5)} * \frac{\psi(3)}{\psi(3)+\psi(4)+\psi(5)} * \frac{\psi(5)}{\psi(4)+\psi(5)} \quad (5.6)$$

5.2.3 An Example

At this point some discussion of how measurement error might bias our estimate of $\underline{\beta}$ may be helpful. The spread of survival times of the true data are determined by the baseline hazard $\lambda_0(t)$ and the relative risk. The shape of the baseline is likely to be of importance, as is the extent of the relative risk. As the rank statistic is marginally sufficient for $\underline{\beta}$, and is unaffected by the addition or subtraction of a constant, it appears intuitive that the mean of the measurement

error is unlikely to affect the distribution of $\underline{\beta}$ for the observed data.

Examination of the way the Cox model estimates $\underline{\beta}$ when the hazards are not truly proportional allows us an insight into the bias measurement error may bring to a Cox fit. If the hazard ratio is not constant over time, then a weighted average of the ratios at each death time provide us with an estimate of $\underline{\beta}$. Hence if $\underline{\beta} = \underline{0}$ the Cox model for the observed data is still valid. Consider a specific example.

Suppose that the true survival data t_i were generated from an exponential distribution $t_i \sim \exp(\tau_i = \exp(\underline{\beta}^T \underline{x}))$ but instead of observing the true data we observe $s_i = t_i + u_i$ where $u_i \sim U(0, b)$. In this case it is simple to calculate the distribution of the observed data for a particular τ using the following probability identities.

$$\begin{aligned} f_{S,T} &= f_T f_{S|T} \\ f_S &= \int_{\text{range of T}} f_{S,T} dT \end{aligned}$$

Using these we have the following:

$f_S = \int_0^s \frac{1}{b}\tau \exp(-\tau t)dt = \frac{1}{b}[1 - \exp(-\tau s)]$ for $0 < s \leq b$

$f_S = \int_{s-b}^s \frac{1}{b}\tau \exp(-\tau t)dt = \frac{1}{b}[\exp(-\tau(s-b)) - \exp(-\tau s)]$ for $b < s < \infty$

Hence we can calculate the survivor functions and hazard functions of the observed data:

$$S(s) = \begin{cases} \frac{1-\exp(-\tau s)-\tau(s-b)}{\tau b} & 0 < s \leq b \\ \frac{\exp(-\tau(s-b))-\exp(-\tau s)}{\tau b} & b < s < \infty \end{cases} \quad \lambda(s) = \begin{cases} \frac{\tau[1-\exp(-\tau s)]}{1-\exp(-\tau s)-\tau(s-b)} & 0 < s \leq b \\ \tau & b < s < \infty \end{cases}$$

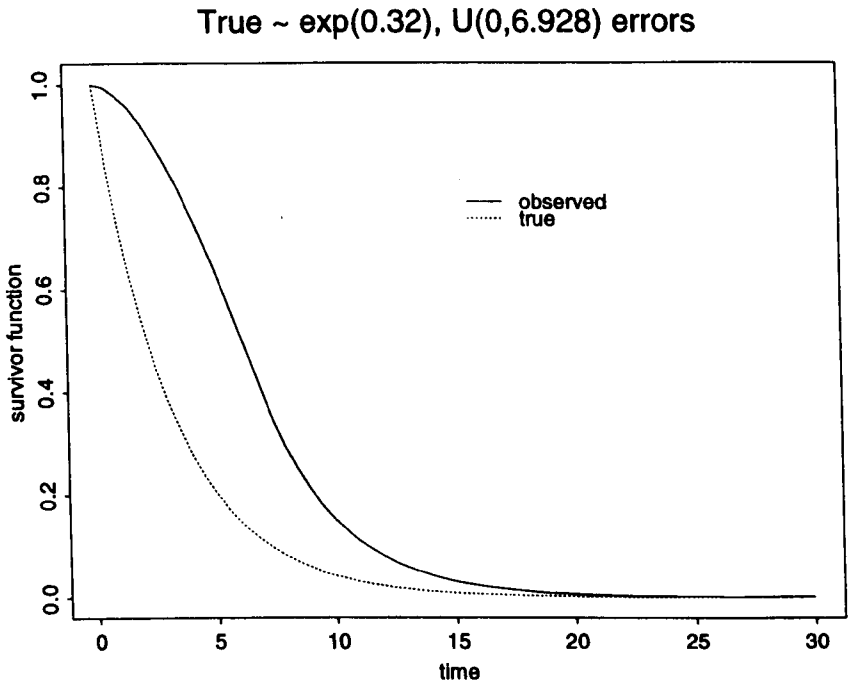


Figure 5.3: Survival for true and observed data : exponential true data with U(0,b) errors

Figures 5.3 and 5.4 show the survivor functions and hazard functions for true exponential(0.32) data with U(0,6.928) errors.

Note for two groups with true parameters τ and 2τ the hazards of the observed data are no longer proportional, and hence the hazard ratio is not constant over time. The hazard ratio for two groups with true parameters τ and 2τ is shown graphically for a particular case in figure 5.5. As the hazard ratio is less than the true hazard ratio for a subset of the data we would expect the direction of bias

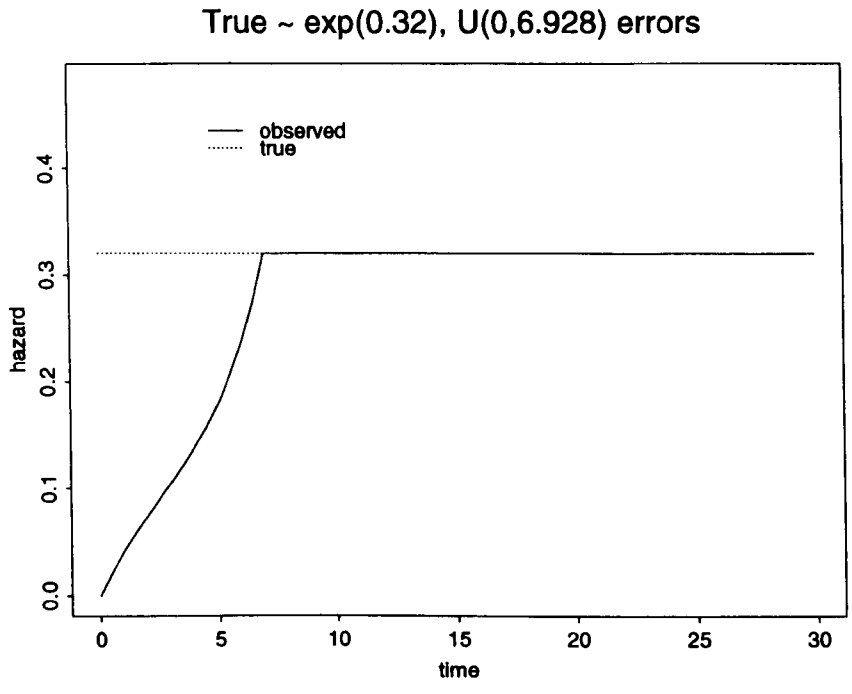


Figure 5.4: Hazard for true and observed data : exponential true data with U(0,b) errors

to be towards the null hypothesis. The extent of bias depends on the proportion of the data observed in the region where the reduction in the hazard ratio is most severe i.e. on the percentiles of the observed (or true) data. For Weibull data with uniform errors the integration is not possible analytically. This is also true for normal errors, but intuitively it appears the hazard ratio would never return to the true level as the distribution of errors is limitless.

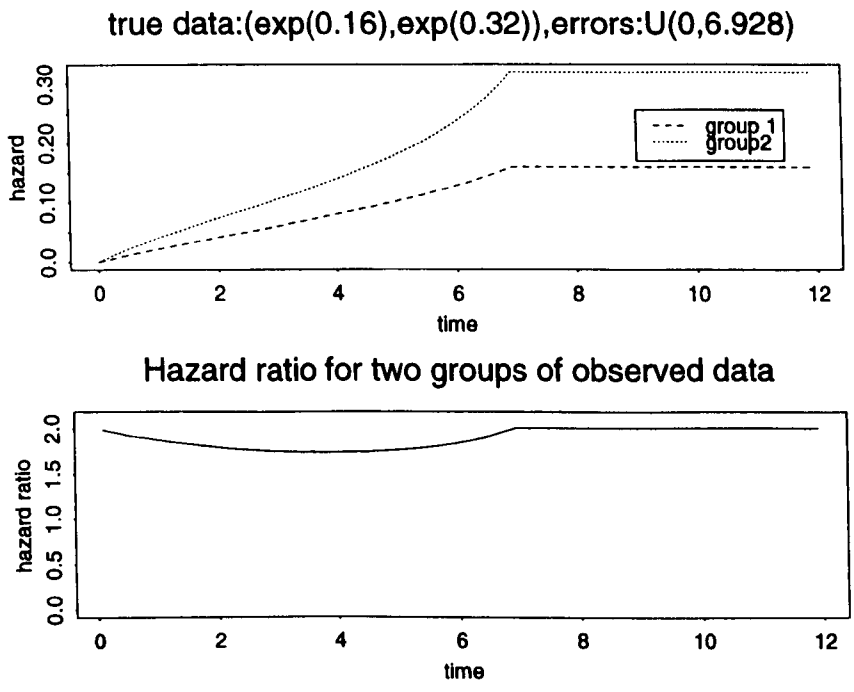


Figure 5.5: Hazards for observed data : exponential true data with U(0,b) errors (2 groups)

5.3 A Simulation Study to Investigate the Effect of Outcome Error

Consider a single binary covariate with relative risk $\exp(\beta)$. We wish to determine the effect of outcome error on the estimate of β, and hence carry out a series of monte-carlo simulations to examine the distribution of $\beta_{observed}$ in a number of situations. Each experiment will be carried out for two measurement models, the usual errors-in-variables model and a biased uniform error of the same standard deviation. For each true dataset an observed dataset is generated. For each set of parameters for the true model we generate pairs of true and observed datasets. For the errors-in-variables model with normal errors the observed survival data may include negative times. These are removed from the Cox analysis. Most experiments assume a flat baseline (i.e. survival

times are from the exponential distribution). We wish to simulate an overall median survival for the two groups in each individual experiment. This will be determined by baseline hazard and relative risk. We therefore derive the median of a mixture of two Weibull distributions with the same shape parameter.

5.3.1 The Median of a Mixture of Two Weibull Distributions

For two groups of equal number from the Weibull distribution, with scale parameter τ for group 1 and $\tau \exp(\beta)$ for group 2 we have the mixture density:

$$f(t|x) = \frac{1}{2}\tau\rho t^{\rho-1}(\exp(-\tau t^{\rho})) + \frac{1}{2}\tau \exp(\beta)\rho t^{\rho-1}(\exp(-\tau \exp(\beta)t^{\rho})) \qquad (5.7)$$

In order to calculate the median we set $F(t|x) = \frac{1}{2}$:

$$\begin{aligned}
\frac{1}{2} &= \int_0^m \frac{1}{2}\tau\rho t^{\rho-1}(\exp(-\tau t^{\rho}))dt + \int_0^m \frac{1}{2}\tau \exp(\beta)\rho t^{\rho-1}(\exp(-\tau \exp(\beta)t^{\rho}))dt \\
&= [-\frac{1}{2}(\exp(-\tau t^{\rho}))]_0^m + [-\frac{1}{2}(\exp(-\tau \exp(\beta)t^{\rho}))]_0^m \\
&= 1 - \frac{1}{2}\exp(-\tau m^{\rho}) - \frac{1}{2}\exp(-\tau \exp(\beta)m^{\rho}) \\
\Longrightarrow 1 &= \exp(-\tau m^{\rho})) + \exp(-\tau \exp(\beta)m^{\rho}) \qquad (5.8)
\end{aligned}$$

Writing $G = m^{\rho}$,$H = \exp(-\tau G)$,$J = \exp(\beta)$, then this is the solution to the equation $1 = H + H^J$. When $J = 2$ $H = \frac{\sqrt{5}-1}{2}$ and when $J = 3$ H can be solved iteratively to give $H = 0.682$. The median is thus defined as:

$$m = \{\frac{1}{\tau}\log(\frac{1}{H})\}^{\frac{1}{\rho}} \tag{5.9}$$

Table 5.1: Scale parameter τ for baseline group for various characteristics from the Weibull mixture with median m

shape par ρ	relative risk $\exp(\beta)$	median m		
		3	12	30
1	2	0.160	0.040	0.016
0.5	2	0.278	0.139	0.088
1.5	2	0.093	0.012	0.0029
1	3	0.128	0.032	0.013

The data are generated using the rweibull(par1,par2) and rexp(τ) functions in S-plus. The function rweibull is parameterised slightly differently with par1 = ρ and $\tau = (\text{par2}^{1/\rho})^{-1}$. In total 8 different combinations of τ,n and $\exp(\beta)$ were chosen. For each of these data medians of 3,12, and 30 (months) were considered (note median 3 corresponds to the median survival of 90 days for all lung cancer patients, 12 to a 1 yr median survival and 30 to median survival of $2\frac{1}{2}$ years). Errors arising from both measurement models with standard deviations $\frac{1}{30}, \frac{1}{2}, 1$ and 2 months (or about 1 day, 2 weeks, 30 and 60 days) were considered. Hence a total number of 192 individual experiments were carried out, with characteristics given in tables 5.2, 5.3 and 5.4. Code to create the simulated data is in appendix E.5.

Table 5.2: Characteristics of individual experiments (exponential baseline - rr = 2)

n	median surv	Error			
50	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
100	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
200	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
500	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
1000	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$

Table 5.3: Characteristics of individual experiments (Weibull baselines - rr = 2)

n	ρ	median surv	Error			
500	0.5	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
			$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
		12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
			$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
		30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
			$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	1.5	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
			$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
		12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
			$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
		30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
			$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$

Table 5.4: Characteristics of individual experiments (exponential Baseline - rr = 3)

n	median surv	Error			
500	3	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	12	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$
	30	$N(0,(\frac{1}{30})^2)$	$N(0,(\frac{1}{2})^2)$	$N(0,1)$	$N(0,2^2)$
		$U(0,0.115)$	$U(0,1.732)$	$U(0,3.464)$	$U(0,6.928)$

5.3.2 Results

In order to continue the flow of the thesis for the reader, detailed tabulations of the results are given in appendix B. Tables B.1, B.2, B.3, B.4, B.5 give the results for the different size data with an exponential baseline and relative risk 2. For rr 3 with exponential baseline the results are in table B.6. Results when the data assume a Weibull baseline are in tables B.7 and B.8.

The simulations demonstrate a number of things. The most crucial point is that the Cox model does indeed demonstrate a robustness to outcome error - in only two experiments was the mean of the observed estimates more than one decimal place away from the mean of the true estimates. Most experiments show a bias towards the null hypothesis of $\beta = 0$. No positive mean bias exceeded 0.007.

It is clear that the median survival has a large effect on the mean and variance of the bias. Experiments with median survival 3 have larger bias with increased variance than those with median survival 12 or 30. The size of the dataset does not have an effect on the mean of the bias but, as one would expect, as n increases the variance of the bias is reduced.

The degree of bias is however dependent on the baseline hazard. The observed β performs best for the case when $\rho = 0.5$ and worst when $\rho = 1.5$. It is unclear without further investigation how the hazard ratio is affected when $\rho \neq 1$. One would anticipate from the results of these experiments that as ρ increases the loss of proportionality of hazards also increases. Theoretical results would be desirable here.

The bias is greater with relative risk 3 than the equivalent experiments with relative risk 2. Figure 5.6 demonstrates this with greater clarity. The hazard ratio is pulled by a larger degree to 1 by the measurement error as the relative risk increases.

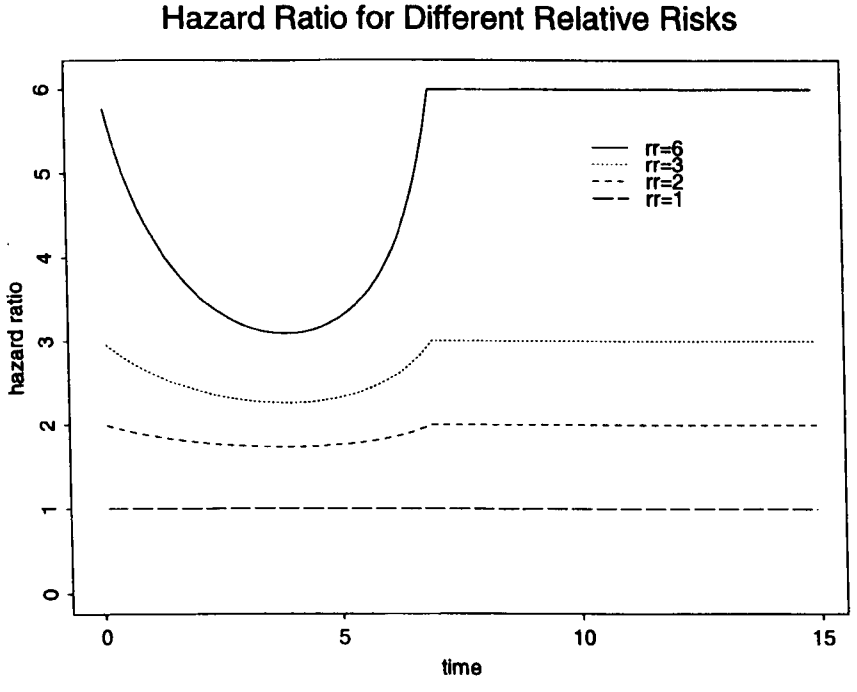


Figure 5.6: Hazards ratios for different relative risks - exp(0.16) baseline, U(0,6.928) errors

As a rule bias is marginally more severe for the uniform measurement model but has reduced variance. It is however unclear as to how the removal of negative survival times from the normal measurement model affects the mean and variance of the bias.

It is also desirable to have a measure of the degree of attenuation for each experiment. This is an equivalent of the measure λ described in chapter 4 for covariate measurement error in simple linear regression. Therefore we define the

measure:

$$\begin{aligned} E(\beta_{obs}) &= (\text{attenuation})E(\beta_{sim}) \\ (\text{attenuation}) &= E(\beta_{obs})/E(\beta_{sim}) \end{aligned} \tag{5.10}$$

In other words the measure of attenuation is the mean of the β estimate for the 200 true datasets over the mean of the β estimates for the 200 observed datasets in each case. The results are in table 5.5.

The observed estimates have mean within 1% of the true mean for all cases except where the median survival is three or $\rho = 1.5$. When $\rho = 1$ and the median survival is 3 the observed mean for error standard deviation 1 is within 3 % of the true mean. For error standard deviation 2 and $\rho = 1$ typical attenuation levels are 93 - 95 % for normal errors and 91 - 93% for uniform errors. When $\rho = 1.5$ attenuation is more severe, and observed means are about 7% less than true means for error standard deviation 1 and 20 % less for error standard deviation 2.

It would seem therefore that despite the robust nature of the Cox model when survival error is present, a correction procedure is necessary when overall survival is poor and measurement error is large. Until it is possible to formulate the true nature of bias in terms of the baseline hazard, the relative risk and the type and degree of measurement error, a simple and flexible correction that gives the analyst a safer level of attenuation is desirable. We therefore formulate an approximate likelihood, similar in style to Efron's approximation to rounding error.

Table 5.5: Attenuation(all experiments)

sim char.	median surv.	Error standard deviation							
		$\frac{1}{30}$		$\frac{1}{2}$		1		2	
		N	U	N	U	N	U	N	U
n=50	3	1.002	1.000	0.995	0.999	0.991	0.990	0.939	0.935
$\rho = 1$	12	1.000	1.000	1.003	0.998	1.000	1.004	0.995	0.999
rr=2	30	0.999	1.000	1.000	0.999	1.003	0.998	0.998	1.002
n=100	3	1.000	1.001	1.008	0.992	0.987	0.972	0.948	0.907
$\rho = 1$	12	1.000	1.000	1.003	0.999	1.002	0.999	1.008	0.992
rr=2	30	1.000	1.000	1.001	1.000	1.002	1.000	1.000	1.000
n=200	3	1.000	1.000	0.996	0.990	0.980	0.973	0.937	0.914
$\rho = 1$	12	1.000	1.000	0.998	0.999	0.999	0.997	0.996	0.990
rr=2	30	1.000	1.000	1.000	1.000	0.998	0.999	1.000	0.998
n=500	3	1.001	1.000	1.007	0.995	1.009	0.990	1.008	0.976
$\rho = 0.5$	12	0.999	1.000	1.003	0.999	1.004	0.998	1.007	0.995
rr=2	30	1.000	1.000	1.002	1.000	1.003	0.999	1.004	0.998
n=500	3	1.000	1.000	0.994	0.993	0.979	0.975	0.927	0.918
$\rho = 1$	12	1.001	1.000	0.999	0.999	0.997	0.998	0.994	0.993
rr=2	30	1.000	1.000	0.999	1.000	1.000	1.000	0.998	0.998
n=500	3	1.000	1.000	0.979	0.981	0.928	0.933	0.781	0.794
$\rho = 1.5$	12	1.000	1.000	0.998	0.999	0.992	0.995	0.978	0.981
rr=2	30	1.000	1.000	1.000	1.000	0.999	0.999	0.995	0.997
n=500	3	1.000	1.000	0.996	0.992	0.981	0.973	0.927	0.915
$\rho = 1$	12	1.001	1.000	1.000	0.999	0.998	0.998	0.996	0.992
rr=3	30	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.998
n=1000	3	1.000	1.000	0.997	0.994	0.985	0.978	0.939	0.923
$\rho = 1$	12	1.000	1.000	1.000	1.000	1.001	0.999	0.997	0.994
rr=2	30	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999

key: sim char. - simulation characteristics

N: Normal errors, U: Uniform Errors

mean($\beta_{observed}$) = attenuation * mean(β_{sim})

5.4 Formulation of an Approximate Partial Likelihood for Measurement Error

In order to specify the partial likelihood of the underlying true survival times their true order is required. The measurement error implies we do not know the true order and hence the true risk set at each survival time. Regardless of the order of the survival times the numerator in the partial likelihood (equation 5.1) remains identical. Section 5.2.1 showed a possible approximation for the true risk sets in the denominator of the partial likelihood for the case of tied data when the true order of times at a particular time is unknown. The next section uses a similar argument when measurement error implies that true times may not have the same order as observed times.

5.4.1 Approximating True Risk Sets Given the Measurement Model

In order to formulate a corrected likelihood we require the probability that each observed individual is in each true risk set. To do this we can specify the joint distribution of all true times given the corresponding observed times. Since each of these times are independent, this is the product of their individual $true|obs$ distributions.

5.4.2 Example 1: Berkson Tied Data

Consider three tied observed survival times $s_1 = s_2 = s_3$ from the Berkson rounding error model

$$t_i|s_i \sim U(obs, obs + 1)$$

The joint distribution of $t_1|s_1, t_2|s_2, t_3|s_3$ is then:

$$f_{T_1|S_1,T_2|S_2,T_3|S_3}(t_1|s_1, t_2|s_2, t_3|s_3) = 1 * 1 * 1$$

We can then specify the probability of each arrangement. For example the probability $t_1 < t_2 < t_3$ is:

$$P(t_1 < t_2 < t_3) = \int_0^1 \int_0^{t_3} \int_0^{t_2} dt_1 dt_2 dt_3 = \int_0^1 \int_0^{t_3} t_2 dt_2 dt_3 = \int_0^1 \frac{t_3^2}{2} dt_3 = \frac{1}{6}$$

In fact we have six equally likely orderings of the true survival times

$$(t_1, t_2, t_3), (t_1, t_3, t_2), (t_2, t_1, t_3), (t_2, t_3, t_1), (t_3, t_1, t_2), (t_3, t_2, t_1)$$

Thus we can specify the probability each true time is in each risk set:

$$P(t_1 \in \text{1st risk set}) = 1, P(t_1 \in \text{2nd risk set}) = \frac{2}{3}, P(t_1 \in \text{3rd risk set}) = \frac{1}{3}$$
$$P(t_2 \in \text{1st risk set}) = 1, P(t_2 \in \text{2nd risk set}) = \frac{2}{3}, P(t_2 \in \text{3rd risk set}) = \frac{1}{3}$$
$$P(t_3 \in \text{1st risk set}) = 1, P(t_3 \in \text{2nd risk set}) = \frac{2}{3}, P(t_3 \in \text{3rd risk set}) = \frac{1}{3}$$

Thus the Berkson rounding error formulation gives us the probabilities that correspond to the Efron approximation for ties. When one employs a rounding error measurement model alone we would thus recommend the use of the Efron

approximate partial likelihood.

5.4.3 Example 2: Berkson Normal Error Model

Now consider three observed survival times s_i that follow the Berkson normal error model.

$$t_i = s_i + u_i \text{ where } u_i \sim N(0, \sigma_u^2) \tag{5.11}$$

Suppose $s_{(1)} = 25, s_{(2)} = 50, s_{(3)} = 75$ and $\sigma_u^2 = 14^2$. The six possible arrangements now have much different probabilities of occurring. For instance the arrangement t_1, t_2, t_3 is much more likely than the arrangement t_3, t_2, t_1. Hence the probability t_1 is in the 3rd risk set is much smaller than that for t_3. Calculating the individual probabilities of each arrangement is an extremely time consuming exercise. For 50 times we would have $50! \approx 3 * 10^{64}$ individual arrangements. Hence a simpler procedure is required to avoid this. We now outline such a procedure.

5.4.4 Matrix P_{ij}

Let P_{ij} be an n*p matrix where n is the ordered number of observations and p is the ordered number of failures. Define P_{ij} as the probability that the ith observation is greater than the jth failure time i.e. the first row is the probability that the first observed survival time is greater than the first,second,third .. pth failure time. If i is the lth failure time set $P_{il} = .$ since we do not require the probability a given time is greater than itself.

For the previous example, $t_i = s_i + u_i$, and $u_i \sim N(0, \sigma_u^2)$ .Consider the first two observed survival times, s_1 and s_2 and their corresponding true times t_1 and t_2. Note $E(t_1|s_1) = s_1$ and $E(t_2|s_2) = s_2$. Then $t_1|s_1 \sim N(s_1, \sigma_u^2)$ and $t_2|s_2 \sim N(s_2, \sigma_u^2)$.

Then $P(t_2 > t_1) = P(t_2 - t_1 > 0)$ and $t_2 - t_1 \sim N(s_2 - s_1, 2\sigma_u^2)$. Then $P(t_2|s_2 - t_1|s_1 > 0) = 1 - \Phi(\frac{s_1 - s_2}{\sqrt{2}\sigma_u})$. Hence $P_{ij} = 1 - \Phi(\frac{s_j - s_i}{\sqrt{2}\sigma_u})$

For our example $s_1 = 25, s_2 = 50, s_3 = 75$ and we have:

$$P_{ij} = \begin{pmatrix} . & 1 - \Phi(\frac{(50-25)}{14\sqrt{2}}) & 1 - \Phi(\frac{(75-25)}{14\sqrt{2}}) \\ 1 - \Phi(\frac{(25-50)}{14\sqrt{2}}) & . & 1 - \Phi(\frac{(75-50)}{14\sqrt{2}}) \\ 1 - \Phi(\frac{(25-75)}{14\sqrt{2}}) & 1 - \Phi(\frac{(50-75)}{14\sqrt{2}}) & . \end{pmatrix} = \begin{pmatrix} . & 0.103 & 0.006 \\ 0.897 & . & 0.103 \\ 0.994 & 0.897 & . \end{pmatrix} \tag{5.12}$$

5.4.5 Treatment of Censored Times

Censored observations present a complication to the calculation of P_{ij}. For tied data it is usual to assume that censored times occur after all times they are tied with. A similar assumption here would be to assume that if a censored time can be greater than a failure time according to the measurement model then it is definitely greater than that failure time. This is equivalent to setting all $0 < P_{ij} < 1, j = 1, \ldots, r$ for a censored time to 1.

5.4.6 Matrix C_{ij}

Let C_{ij} be an n*p matrix where n is the ordered number of observations and p is the ordered number of failures. Define C_{ij} as a weight representing the probability that the true time corresponding to the *ith* observation is in the *jth* risk set.

Hence $C_{i1} = 1$ for all i failure times and $0 \leq C_{ij} \leq 1 (j \geq 2)$.

In order to specify C_{ij} we require an approximation to the probability that the *ith* observation is the *jth* time. For a non-rounding measurement model the *ith* row of the P matrix is likely to tell us that the *ith* time might be bigger than a proportion of the other times. These probabilities typically will have a reasonable spread (i.e. we can say that the *ith* time is extremely likely to be greater than some times but unlikely to be greater than most times). For the subset of times that the *ith* time communicates with we then wish our early weights to be 1 or close to 1, our middle weights to be close to $\frac{1}{2}$ and our later weights to be 0 or close to 0. We thus propose the following method to calculate the C matrix.

5.4.7 Algorithm to Calculate C_{ij} When One Knows P_{ij}

1) Calculate the number of times that the *ith* time is definitely greater than the *jth* failure time j=1, ...,r (i.e. where $P_{ij} = 1$). Call this *ngreater*.

2) Calculate the number of times that the *ith* time might be greater than the *jth* failure time j=1, ...,r (i.e. where $0 < P_{ij} < 1$). Define this set of times

as the *commset* (communication set) for the ith time with $ncomm$ elements. If $ncomm = 0$ proceed to step 5.

3) Calculate the approximate probability that the ith time is the $(ngreater+1)th$ ... $(ngreater+ncomm+1)th$ time. To do this average the elements of *commset*, giving us p and then allocate according to the binomial probabilities :

$$\approx P(i\text{th time is } (ngreater + j + 1)\text{th time}) = ({}^{ncomm}C_j)p^j(1-p)^{ncomm-j}$$
$$\text{for } j = 0, \ldots, ncomm$$

4) Calculate C_{ij} - the weight representing the probability that the ith time is in the $(ngreater + 1)th$ ... $(ngreater + ncomm + 1)th$ risksets, defined as follows: $P(ith \text{ time} \in jth \text{ riskset }, (ngreater + 1) \leq j \leq (ngreater + ncomm + 1)) =$

$$\sum_{[k=ngreater+j+1]}^{ngreater+ncomm+1} \approx P(ith \text{ time} = kth \text{ time}) \text{ for } j = 0, \ldots, ncomm$$

5) If $ncomm = 0$ and the ith time is censored set C_{ij} to 1 if $j \leq ngreater$ and $C_{ij} = 0$ if $j = ngreater + 1$

If $ncomm = 0$ and the ith time is not censored set C_{ij} to 1 if $j \leq ngreater + 1$

If $ncomm > 0$ set C_{ij} to 1 if $j \leq ngreater$

6) Set C_{ij} to 0 if $j > (ngreater + ncomm + 1)th$ time.

5.4.8 Assumptions Used in Formulating C_{ij}

The calculations in the algorithm are a great simplification from the actual calculations required. In fact two assumptions are employed (in step 3) in order

to use the algorithm:

Assumption 1: The probability the *ith* time is greater than all times in its *commset* is equal to the average of all probabilities in the *commset*. This also implies that all combinations of times in the commset are theoretically possible

Assumption 2: The probability that the *ith* time is greater than the *jth* time is conditionally independent of the probability the *ith* time is greater than the $kth (k \neq j)$ time. This is expressed mathematically as:

$$P(ith \text{ time } > jth \text{ time } | ith \text{ time } > kth \text{ time }) = P(ith \text{ time } > jth \text{ time }) \tag{5.13}$$

Although these assumptions do not in practise hold, if all combinations have a small or zero probability then the approximation should prove adequate.

5.4.9 Example of the C Calculation

Berkson normal error model

Recall we have three observed survival times $s_{(1)} = 25, s_{(2)} = 50, s_{(3)} = 75$ with corresponding $\psi(i) = \exp(\underline{\beta}^T \underline{x}_i)$ that follow the Berkson normal error model. Ignoring the error model and fitting using the observed data gives the partial likelihood

$$\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)} * \frac{\psi(2)}{\psi(2)+\psi(3)} * \frac{\psi(3)}{\psi(3)} \tag{5.14}$$

Given the P matrix:

$$P_{ij} = \begin{pmatrix} . & 0.103 & 0.006 \\ 0.897 & . & 0.103 \\ 0.994 & 0.897 & . \end{pmatrix} \tag{5.15}$$

we have to work out the approximate probability that each time is the true 1st,2nd and third time. For the first observed survival time $s_{(1)}$ $ngreater = 0$ and $ncomm = 2$. We have elements $P_{(12)} = 0.103$ and $P_{(13)} = 0.006$. Hence the calculation follows a Bin(2,0.0545) distribution.

$$\begin{aligned} &\approx P(s_1 \text{is 1st time}) \approx {}^2C_0 * (1 - 0.0545)^2 = 0.894 \\ &\approx P(s_1 \text{is 2nd time}) \approx {}^2C_1 * (0.0545) * (1 - 0.0545) = 0.103 \\ &\approx P(s_1 \text{is 3rd time}) \approx {}^2C_2 * (0.0545)^2 = 0.003 \end{aligned}$$

Hence we can work out the weights for the probability that the corresponding true time t_1 is in the true 1st,2nd and 3rd risk sets.

$$\begin{aligned} P(t_1 \in \text{1st risk set}) &= C_{11} = 0.894 + 0.103 + 0.003 = 1 \\ P(t_1 \in \text{2nd risk set}) &= C_{12} = 0.103 + 0.003 = 0.106 \\ P(t_1 \in \text{3rd risk set}) &= C_{13} = 0.003 \end{aligned}$$

Following the same procedure for the second and third observed survival times we have the C matrix

$$C_{ij} = \begin{pmatrix} 1 & 0.106 & 0.003 \\ 1 & 0.75 & 0.25 \\ 1 & 0.997 & 0.894 \end{pmatrix} \tag{5.16}$$

5.4.10 Approximating the Partial Likelihood for Measurement Error

Armed with the matrix C_{ij} we can then carry out a correction for measurement error in the Cox model. We simply weight each $\psi(i)$ in the jth risk set by the estimate of C_{ij}.

Recall from our example we had the uncorrected likelihood.

$$\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)} * \frac{\psi(2)}{\psi(2)+\psi(3)} * \frac{\psi(3)}{\psi(3)}$$

Now using the C_{ij} matrix we then have the approximate partial likelihood.

$$\frac{\psi(1)}{\psi(1)+\psi(2)+\psi(3)} * \frac{\psi(2)}{0.106\psi(1)+0.75\psi(2)+0.997\psi(3)} * \frac{\psi(3)}{0.003\psi(1)+0.25\psi(2)+0.894\psi(3)} \tag{5.17}$$

This is more formally expressed (where D is the set of deaths and $\underline{\beta}$ being a p vector of covariates):

$$L = \prod_{j \in D} \frac{\exp \underline{\beta}^T \underline{x}_j}{\sum_{i=1}^{n} C_{ij} \exp(\underline{\beta}^T \underline{x}_i)} \tag{5.18}$$

The corresponding log-likelihood is as follows:

$$\log L = \sum_{j \in D} \left[\underline{\beta}^T \underline{x}_j - \log \sum_{i=1}^{n} C_{ij} \exp(\underline{\beta}^T \underline{x}_i) \right] \tag{5.19}$$

and the score equations are:

$$U_k = \sum_{j \in D} \left[x_{jk} - \frac{\sum_{i=1}^{n} C_{ij} x_{ik} \exp(\underline{\beta}^T \underline{x}_i)}{\sum_{i=1}^{n} C_{ij} \exp(\underline{\beta}^T \underline{x}_i)} \right] \text{ for } k = 1 \ldots p \tag{5.20}$$

with the $p * p$ information matrix having components on the diagonal:

$$IN_{kk} = \sum_{j \in D} \left[\frac{\sum_{i=1}^{n} C_{ij} \exp(\underline{\beta}^T \underline{x}_i) \sum_{i=1}^{n} C_{ij} x_{ik}^2 \exp(\underline{\beta}^T \underline{x}_i) - (\sum_{i=1}^{n} C_{ij} x_{ik} \exp(\underline{\beta}^T \underline{x}_i))^2}{(\sum_{i=1}^{n} C_{ij} \exp(\underline{\beta}^T x_i))^2} \right] \tag{5.21}$$

for $k = 1 \ldots p$. The components of IN_{km} for $k, m = 1 \ldots p, k \neq m$ are

$$\sum_{j \in D} \left[\left\{ \sum_{i=1}^{n} C_{ij} \exp(\underline{\beta}^T \underline{x}_i) \sum_{i=1}^{n} C_{ij} x_{ik} x_{im} \exp(\underline{\beta}^T \underline{x}_i) - \sum_{i=1}^{n} C_{ij} x_{ik} \exp(\underline{\beta}^T \underline{x}_i) \sum_{i=1}^{n} C_{ij} x_{im} \exp(\underline{\beta}^T \underline{x}_i) \right\} / (\sum_{i=1}^{n} C_{ij} \exp(\underline{\beta}^T \underline{x}_i))^2 \right] \tag{5.22}$$

Hence we can proceed to estimate $\underline{\beta}$ via a Newton-Raphson iterative process. Appendix C.1 gives an alternative notation for the usual Cox likelihood due to Nakamura and shows the new likelihood in this notation. Appendix E.7 gives a set of programs based on this notation for S-plus that can fit this likelihood in the at present rather restrictive circumstances of one covariate. Additional programming in order to extend the programs for several covariates is not intended in the course of this thesis. Reasons for this are that the computation of the P matrix for large datasets is extremely expensive and hence only small datasets (and in turn a small number of covariates) are considered .

5.4.11 The Likelihood When No Error is Present

If no error is present then the approximate likelihood is identical to the Cox likelihood. This is due to the fact that all elements of P_{ij} are either 1 or 0 - ie each time is either greater or less than all other times. Hence the *commset* for each time is always empty and C_{ij} reduces to a matrix showing the 'at risk' status of each observation.

If data display ties only and no additional error then there is no need to approximate the elements of C_{ij} and P_{ij} is redundant. In this situation the use of the exact C matrix results in the Efron approximation for ties (see sections 5.2.1 and 5.4.2).

5.4.12 Errors-in-Variables and the Likelihood Approximation

The errors-in-variables model does not allow us to specify $t|s$ without specifying the distribution of true, and hence no longer allows a semi-parametric Cox model. By conditioning on the observed data for this case the distribution of $t|s$ is likely to be dependent on the parameters of the distribution of t and hence on $\underline{\beta}$. We thus need simple approximations to $t|s$ in order to proceed without violating the semi-parametric nature of the model and introducing dependency on our unknown parameters of $\underline{\beta}$.

Normal errors

If the underlying true data were normal (which in survival datasets they clearly are not), and the error distribution were normal the distribution of $t|s$ is also normal (see section 4.2.1):

$$t|s \sim N(\lambda s + (1-\lambda)\mu_{(t)}, \lambda\sigma^2_{(error)}) \text{ where } \lambda = \frac{\sigma^2_{(t)}}{\sigma^2_{(error)} + \sigma^2_{(t)}} \tag{5.23}$$

Note for survival data even with large error $\sigma^2_{(t)} \gg \sigma^2_{(error)}$ and $\lambda \approx 1$ and hence a 95% confidence interval for $t|s$ is approximately:

$$(s - 1.96 * \sigma_{(error)}, s + 1.96 * \sigma_{(error)}) \tag{5.24}$$

Thus we could approximate the distribution for the calculation of the matrices P_{ij} and C_{ij} by merely using a validation subsample to estimate the standard deviation $\sigma_{(error)}$ of the measurement error:

$$t_i|s_i \sim U(\max[0, (s_i - 1.96 * \sigma_{(\hat{error})})], (s_i + 1.96 * (\sigma_{(\hat{error})})) \tag{5.25}$$

Uniform errors

For the U(0,b) errors case with an exponential baseline the distribution of $t|s$ can easily be derived using the identity $f_{t|s} = \frac{f_{t,s}}{f_s}$ where $f_{t,s}$ and f_s are given in section 5.2.3. This is defined as:

$$f_{t|s} = [\tau \exp(-\tau)]/[1 - \exp(-\tau s)] \text{ for } 0 < s < b, 0 < t < s$$

$f_{t|s} = [\tau \exp(-\tau t)]/[\exp(-\tau(s-b)) - \exp(-\tau s)]$ for $s > b, s - b < t < s$

This is roughly flat across the interval $[\max(0, s-b), s]$. Hence if we can estimate b we will have a good approximation for the distribution of $t|s$. If we estimate the mean $\mu_{(error)} = \frac{b}{2}$ and the variance of the error from validation data we have an estimate of $\sigma^2_{(error)} = b^2/12$ and this gives:

$$t_i|s_i \sim U(\max(0, s_i - \hat{\mu_{(error)}} - \hat{\sigma_{(error)}}\frac{\sqrt{(12)}}{2}), s_i - \hat{\mu_{(error)}} + \hat{\sigma_{(error)}}\frac{\sqrt{(12)}}{2}) \quad (5.26)$$

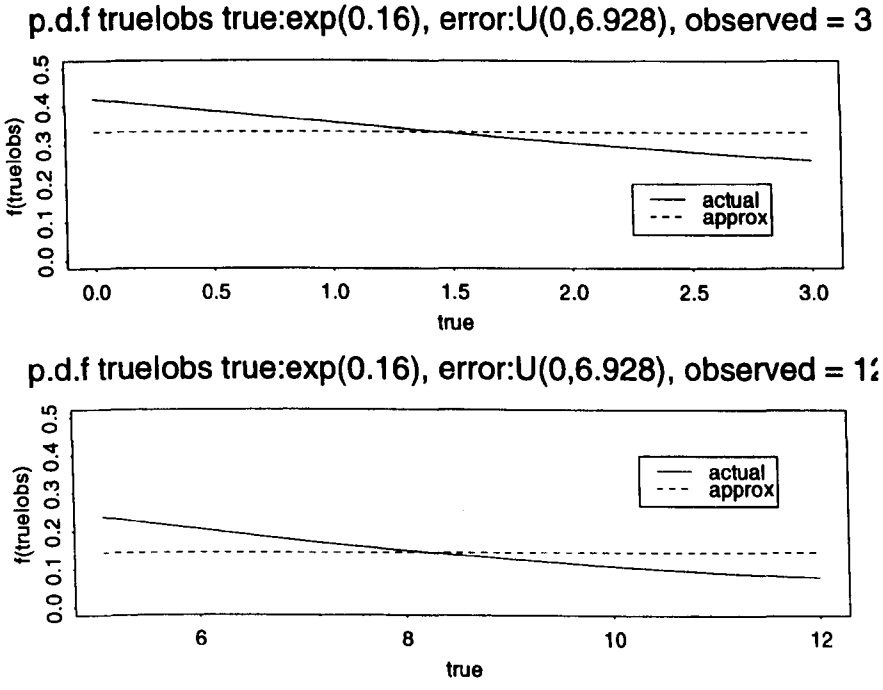


Figure 5.7: Actual and approximate distribution of true| observed survival for exponential data with uniform errors

An example is given in figure 5.7. In order to calculate the P matrix we thus need the density function of the subtraction of two uniform variables.

5.4.13 Probability Density Function of U(a,b) - U(c,d)

We have $X \sim U(a,b)$ and $Y \sim U(c,d)$. Let $W = X - Y$. This is an extension of the common undergraduate problem of determining the sum of two $U(0,1)$ random variables, see for examples Rice (1988) p.100. Via the convolution integral we obtain:

$$f_W(w) = \begin{cases} \frac{w-(a-d)}{(b-a)(d-c)} & (a-d) \leq w \leq \min(b-d, a-c) \\ \min(\frac{1}{b-a}, \frac{1}{d-c}) & \min(b-d, a-c) < w < \max(b-d, a-c) \\ \frac{(b-c)-w}{(b-a)(d-c)} & \max(b-d, a-c) \leq w \leq b-c \end{cases} \quad (5.27)$$

Example : $W = U(8,14) - U(2,5)$

$$f_W(w) = \begin{cases} \frac{w-3}{18} & 3 \leq w \leq 6 \\ \frac{1}{6} & 6 < w < 9 \\ \frac{12-w}{18} & 9 \leq w \leq 12 \end{cases} \quad (5.28)$$

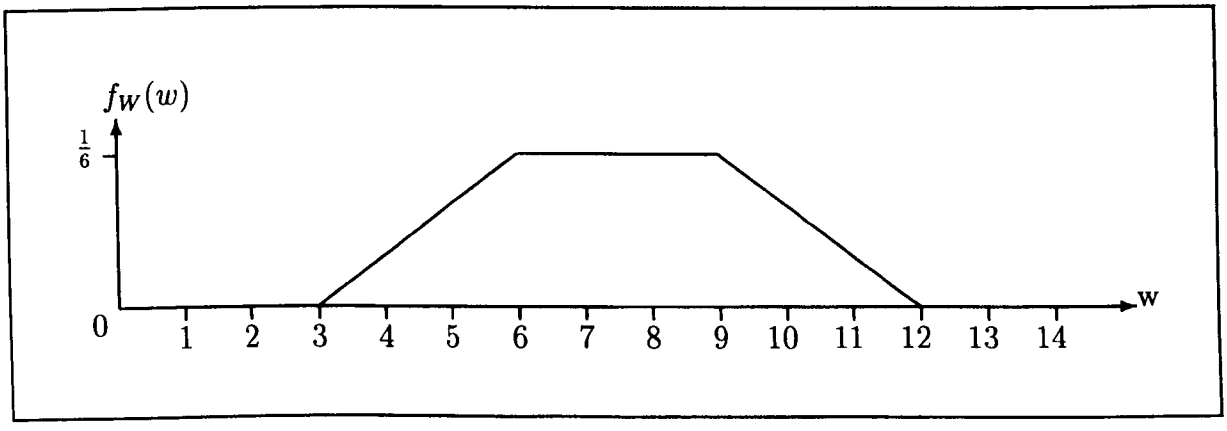


Figure 5.8: $W = U(8,14) - U(2,5)$

A plot of the p.d.f. is given in figure 5.8

Probability $U(a,b) > U(c,d)$

We wish to calculate the P matrix for the data we are approximating by the uniform distribution. Hence we wish to find the probability that a $U(a,b)$ variable is greater than a $U(c,d)$ variable. This is equivalent to calculating the probability that $W = U(a,b) - U(c,d) > 0$. We can do this via the distribution function for the variable W defined in the previous subsection.

We wish to calculate the distribution function $F_W(w) = P(W \leq w)$. For simplicity we now assume $\frac{1}{b-a} < \frac{1}{d-c}$.

$$F_W(w) = \begin{cases} \int_{(a-d)}^{w} \frac{s-(a-d)}{(b-a)(d-c)} ds & (a-d) \leq w \leq (a-c) \\ \int_{(a-d)}^{(a-c)} \frac{s-(a-d)}{(b-a)(d-c)} ds + \int_{(a-c)}^{w} \frac{1}{(b-a)} ds & (a-c) < w < (b-d) \\ \int_{(a-d)}^{(a-c)} \frac{s-(a-d)}{(b-a)(d-c)} ds + \int_{(a-c)}^{(b-d)} \frac{1}{(b-a)} ds + \int_{(b-d)}^{w} \frac{(b-c)-s}{((b-a)(d-c)} ds & (b-d) \leq w \leq (b-c) \end{cases} \tag{5.29}$$

$$F_W(w) = \begin{cases} \frac{1}{(b-a)(d-c)} [\frac{w^2}{2} - w(a-d) + \frac{(a-d)^2}{2}] & (a-d) \leq w \leq (a-c) \\ \frac{2w-2(a-c)+(d-c)}{2(b-a)} & (a-c) < w < (b-d) \\ \frac{2(b-a)+(c-d)}{2(b-a)} + \frac{[(b-c)w - \frac{w^2}{2} - (b-c)(b-d) + \frac{(b-d)^2}{2}]}{(b-a)(d-c)} & (b-d) \leq w \leq (b-c) \end{cases} \tag{5.30}$$

Hence the probability $W > 0$ is given by $1 - F_W(0)$:

$$F_W(0) = \begin{cases} \frac{(a-d)^2}{2(b-a)(d-c)} & (a-d) \leq 0 \leq (a-c) \\ \frac{(d-c)-2(a-c)}{2(b-a)} & (a-c) < 0 < (b-d) \\ \frac{2(b-a)+(c-d)}{2(b-a)} + \frac{2ac+2bd-2ad-(b^2+c^2)}{2(b-a)(d-c)} & (b-d) \leq 0 \leq (b-c) \end{cases} \tag{5.31}$$

Note if $\frac{1}{d-c} < \frac{1}{b-a}$ then the probability we require is $F_W(0)$ as defined above.

5.5 Estimation of the Baseline Hazard

The likelihood we have derived is of the same form as that of Efron, in that the contribution to the risk set at each death time is weighted between 0 and 1. Unfortunately we do not know the true death times but in the spirit of regression calibration we can estimate these by taking the mean of our (approximate) distribution of $t_i|s_i$, which is of the form $U(a_i, b_i)$. We thus have:

$$E(t_i|s_i) = \frac{a_i + b_i}{2} \tag{5.32}$$

Recall the estimate for the baseline hazard used in conjunction with Efron's likelihood is:

$$\hat{\lambda_0}(t_{(j)}) \approx \sum_{k=1}^{m_j} \frac{1}{\{\sum_{(l \in R(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x_l}) - \frac{(k-1)}{m_j} \sum_{(l \in D(t_{(j)}))} \exp(\hat{\underline{\beta}}^T \underline{x_l})\}} \tag{5.33}$$

When $\underline{\beta} = 0$ the sum of these estimates is identical to the Nelson estimate of the cumulative hazard modified for ties. Using the expected death time $E(t_{(j)}|s_{(j)})$ we have an estimate of the baseline hazard:

$$\hat{\lambda_0}(E(t_{(j)}|s_{(j)})) \approx \sum_{\substack{\text{k=no. of previous failures} + 1 \\ \text{k=no. of previous failures} + 1}}^{+m_j} \frac{1}{\{\Sigma_{i=1}^{n} C_{ik} \exp(\underline{\beta}^T \underline{x_i})\}} \tag{5.34}$$

5.5.1 The Case of No Covariates

We would also wish to have an estimate of the survivor and hazard functions for the case of no covariates, i.e. an equivalent of the non-parametric estimate. If we substitute $\underline{\beta} = \underline{0}$ in the above estimate we then have

$$\hat{\lambda}(E(t_{(j)}|s_{(j)})) \approx \sum_{\substack{\text{k=no. of previous failures + 1} \\ \text{k=no. of previous failures + 1}}}^{\text{k=no. of previous failures + 1+}m_j} \frac{1}{\sum_{i=1}^{n} C_{ik}} \tag{5.35}$$

The sum of each contribution provides us with an estimate of the cumulative hazard and hence an estimate of the survivor function.

5.6 Relationship with Interval Censoring

The employment of the uniform approximation for the distribution of $t|s$ is in effect creating interval censored data i.e. an interval in which exact failure is unknown but assumed equally likely to occur. Methods exist for interval censoring (see Collett ch.8) but these assume *grouping* of data. In other words throughout the period of study all individuals are followed up at given times e.g. 3 months, 6 months, 12 months and if failure/censoring has occurred between the previous visit and the current visit the event is interval censored. In fact grouping of data is in effect creating Berkson rounding error.

The approximate likelihood suggested in this thesis provides a method for dealing with interval censoring when the intervals can crossover e.g. patient 1 (0,10) and patient 2 (5,15). In other words interval censoring can be included

via a Berkson rounding measurement model for which each rounding interval is potentially unique.

5.7 Verification of the Correction using the Simulated Data

In order to verify the correction procedure the correction was tested against the observed data for each experiment in section 5.3 where $n = 50, 100, 200$. Although in principle it is possible to test the correction for datasets of size 500 and 1,000 computer memory consideration and speed of calculation make this impossible at present. As the simulated datasets do not show a large bias in the observed data measurement model the true data with median 3 were also tested with error standard deviations of 4,6 and 8. The computer code required to calculate the P matrix is given in appendix E.1. Code to calculate the uniform probabilities is in appendix E.3. A routine for implementation of the C matrix algorithm is in appendix E.4. A parent function to implement the correction via one command is in E.6.

Fully tabulated results are given in appendix D, tables D.1, D.2 and D.3. Notice there is a column headed "na" in each table. This represents the number of cases where at least one column of the C matrix has consisted entirely of 0's. In other words at a given death time the calculation has returned zero risk. This is due to the approximate nature of the calculation and rounding employed in the computer code. Although these cases were removed from the re-analysis in reality it would be simple to ensure no column of the C matrix had only zero

elements. One suggestion is to include each failure time with weight one in the risk set corresponding to it.

Where the initial bias was negligible (median survival 3 with error standard deviations $\frac{1}{30}$ and $\frac{1}{2}$, median survival 12 or 30 with all errors) the approximate corrected likelihood did not incur extra bias in most cases. Where it did the extra mean bias is not of concern. The variance of bias is universally improved. This guards against the tails of the distribution of bias. There are no missing values for any of these cases.

Where the mean bias in the naive fits was considerable (median survival 3 with error standard deviations 1,2,4,8) the new fits showed a great improvement to the extent that the new mean β was only once over 1 decimal place away from the mean of the true fits (n=50, normal errors with error standard deviation 8). The levels of attenuation for naive and corrected fits are given in table 5.6.

There is little evidence that the missing values represent the cases with most severe bias in the original fits - see table 5.7. Indeed the correction still holds well when there are few missing values and the original attenuation is considerable.

Where correction for attenuation is considerable the variance of the new β is increased. This increase is starker for the case of normal errors than uniform errors.The extra variance is a feature of all measurement error corrections (see chapter 4). The extra variance in the new β is not fully reflected in the mean of the variance estimates obtained from the corrected likelihood - see table 5.8. Hence significance tests and confidence intervals are too conservative if the new

Table 5.6: Naive and corrected attenuation

sim. char.	median surv.	error type	Error standard deviation $\frac{1}{30}$ Nai	$\frac{1}{30}$ Cor	$\frac{1}{2}$ Nai	$\frac{1}{2}$ Cor	1 Nai	1 Cor	2 Nai	2 Cor
n=50	3	N	1.002	1.000	0.995	0.985	0.991	0.985	0.939	1.001
	12	N	1.000	1.000	1.003	0.996	1.000	0.991	0.995	0.985
	30	N	0.999	1.000	1.000	0.999	1.003	0.997	0.998	0.993
n=100	3	N	1.000	1.000	1.008	0.994	0.987	0.996	0.948	1.002
	12	N	1.000	1.000	1.003	0.999	1.002	0.997	1.008	0.994
	30	N	1.000	1.000	1.001	1.000	1.002	0.999	1.000	0.998
n=200	3	N	1.000	0.999	0.996	0.999	0.980	1.003	0.937	1.020
	12	N	1.000	1.000	0.998	0.999	0.999	0.998	0.996	0.999
	30	N	1.000	1.000	1.000	0.999	0.998	0.999	1.000	0.998
n=50	3	U	1.000	0.999	0.999	0.995	0.990	0.993	0.935	0.989
	12	U	1.000	1.000	1.004	0.997	0.999	0.995	1.000	1.000
	30	U	1.000	1.000	0.999	0.998	0.998	0.997	1.002	0.997
n=100	3	U	1.001	1.000	0.992	0.994	0.972	0.990	0.907	0.985
	12	U	1.000	1.000	0.999	0.999	0.999	0.997	0.992	0.994
	30	U	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.998
n=200	3	U	1.000	1.000	0.990	0.996	0.973	0.995	0.914	0.991
	12	U	1.000	1.000	0.999	0.999	0.997	0.997	0.990	0.996
	30	U	1.000	1.000	1.000	1.000	0.999	0.999	0.998	0.998

sim. char.	median surv.	error type	Error standard deviation 4 Nai	4 Cor	6 Nai	6 Cor	8 Nai	8 Cor
n=50	3	N	0.824	0.992	0.619	0.953	0.522	0.846
n=100	3	N	0.798	0.984	0.654	0.948	0.541	0.885
n=200	3	N	0.799	1.021	0.652	0.993	0.526	0.924
n=50	3	U	0.802	0.984	0.681	0.952	0.581	0.939
n=100	3	U	0.787	0.996	0.659	0.978	0.556	0.945
n=200	3	U	0.775	0.969	0.649	0.953	0.550	0.943

key: sim char. - simulation characteristics

N: Normal errors, U: Uniform Errors

Nai: Naive fit, Cor: Corrected fit

$\text{mean}(\beta_{naive}) = \text{attenuation}_{naive} * \text{mean}(\beta_{sim})$

$\text{mean}(\beta_{corr} = \text{attenuation}_{corr} * \text{mean}(\beta_{sim})$

Note: All simulations: $\rho = 1$ (exponential), rr=2 ($\beta_{true} = 0.693$)

Table 5.7: True and observed β where the corrected likelihood encountered zero risk

n	e.t.	s.d.	n.m.	mean β_{sim}	var β_{sim}	mean β_{obs}	var β_{obs}	m.b.$_{obs}$	v.b.$_{obs}$
50	N	6	33	0.5266	0.1033	0.2549	0.1365	-0.2717	0.1839
50	U	6	24	0.6802	0.0784	0.4517	0.0844	-0.2285	0.0817
100	N	6	36	0.6838	0.0435	0.4189	0.0650	-0.2650	0.0487
100	U	6	26	0.6260	0.0484	0.3918	0.0409	-0.2342	0.0240
200	N	6	31	0.6565	0.0228	0.3829	0.0354	-0.2736	0.0207
200	U	6	18	0.7229	0.0207	0.4133	0.0189	-0.3096	0.0182
50	N	8	73	0.6788	0.1128	0.2900	0.1734	-0.3889	0.1764
50	U	8	57	0.6857	0.0784	0.3724	0.0818	-0.3134	0.0903
100	N	8	78	0.6917	0.0435	0.3717	0.0687	-0.3201	0.0635
100	U	8	70	0.6532	0.0484	0.3633	0.0384	-0.2900	0.0419
200	N	8	77	0.6718	0.0228	0.3331	0.0349	-0.3387	0.0276
200	U	8	65	0.6877	0.0207	0.3618	0.0186	-0.3258	0.0211

key: e.t. = error type (N normal, U uniform)
s.d. = error standard deviation, n.m. number missing (new fits)
m.b. = mean bias, v.b. = variance of bias

variance is employed in these cases. This will be discussed in chapter 7.

A nice graphical representation of the error correction can be seen in figures 5.9, 5.10, 5.11, 5.12, 5.13 and 5.14. These are plots of the quantiles of the true, naive and corrected fits against the quantiles of the standard normal distribution for the case of median survival 3 with error standard deviations 2,4,6 and 8. It is clear from these that the naive fits show considerable bias whereas the new fits have the approximately the same mean as the true fits. The increased gradient in the new fits show how the new β has increased variance.

Table 5.8: Variance of new β and mean of the variance estimates

Err	ms	var β_{new}	mean(f.v.)	var β_{new}	mean(f.v.)	var β_{new}	mean(f.v.)
		n=50		n=100		n=200	
N1	3	0.1050	0.0958	0.0435	0.0461	0.0241	0.0266
U1	3	0.0782	0.0942	0.0584	0.0462	0.0223	0.0228
N2	3	0.1020	0.0947	0.0437	0.0459	0.0243	0.0226
U2	3	0.0813	0.0937	0.0593	0.0461	0.0222	0.0227
N3	3	0.1045	0.0946	0.0457	0.0459	0.0253	0.0227
U3	3	0.0868	0.0937	0.0621	0.0461	0.0226	0.0228
N4	3	0.1165	0.0964	0.0523	0.0465	0.0297	0.0230
U4	3	0.1021	0.0950	0.0685	0.0467	0.0237	0.0230
N5	3	0.1401	0.1017	0.0687	0.0483	0.0397	0.0239
U5	3	0.1178	0.0989	0.0659	0.0483	0.0257	0.0233
N6	3	0.1411	0.1022	0.0846	0.0500	0.0491	0.0246
U6	3	0.1215	0.1012	0.0719	0.0499	0.0289	0.0241
N7	3	0.1698	0.1028	0.0995	0.0506	0.0447	0.0245
U7	3	0.1153	0.1026	0.0740	0.0506	0.0354	0.0248
N1	12	0.1055	0.0959	0.0435	0.0461	0.0241	0.0226
U1	12	0.0784	0.0943	0.0585	0.0462	0.0223	0.0228
N2	12	0.1034	0.0955	0.0435	0.0461	0.0240	0.0226
U2	12	0.0784	0.0940	0.0584	0.0461	0.0222	0.0227
N3	12	0.1028	0.0952	0.0434	0.0460	0.0240	0.0226
U3	12	0.0796	0.0939	0.0585	0.0461	0.0221	0.0227
N4	12	0.1020	0.0947	0.0437	0.0459	0.0243	0.0226
U4	12	0.0813	0.0937	0.0593	0.0461	0.0222	0.0227
N1	30	0.1057	0.0959	0.0435	0.0461	0.0241	0.0226
U1	30	0.0784	0.0943	0.0585	0.0462	0.0223	0.0228
N2	30	0.1047	0.0957	0.0435	0.0461	0.0240	0.0226
U2	30	0.0781	0.0942	0.0584	0.0462	0.0223	0.0228
N3	30	0.1037	0.0956	0.0436	0.0461	0.0240	0.0226
U3	30	0.0783	0.0941	0.0584	0.0462	0.0222	0.0227
N4	30	0.1030	0.0953	0.0434	0.0460	0.0240	0.0226
U4	30	0.0792	0.0939	0.0584	0.0461	0.0221	0.0227

key: ms = median survival (true data)

mean(f.v.) = mean(fitted variance estimates)

$N1 : N(0, \frac{1}{30}^2)$,$N2 : N(0, \frac{1}{2}^2)$, $N3 : N(0, 1^2)$,

$N4 : N(0, 2^2)$,$N5 : N(0, 4^2)$,$N6 : N(0, 6^2)$,$N7 : N(0, 8^2)$

$U1 : U(0, 0.115)$,$U2 : U(0, 1.732)$,$U3 : U(0, 3.464)$,$U4 : U(0, 6.928)$,

$U5 : U(0, 13.856)$,$U6 : U(0, 20.785)$,$U7 : U(0, 27.713)$

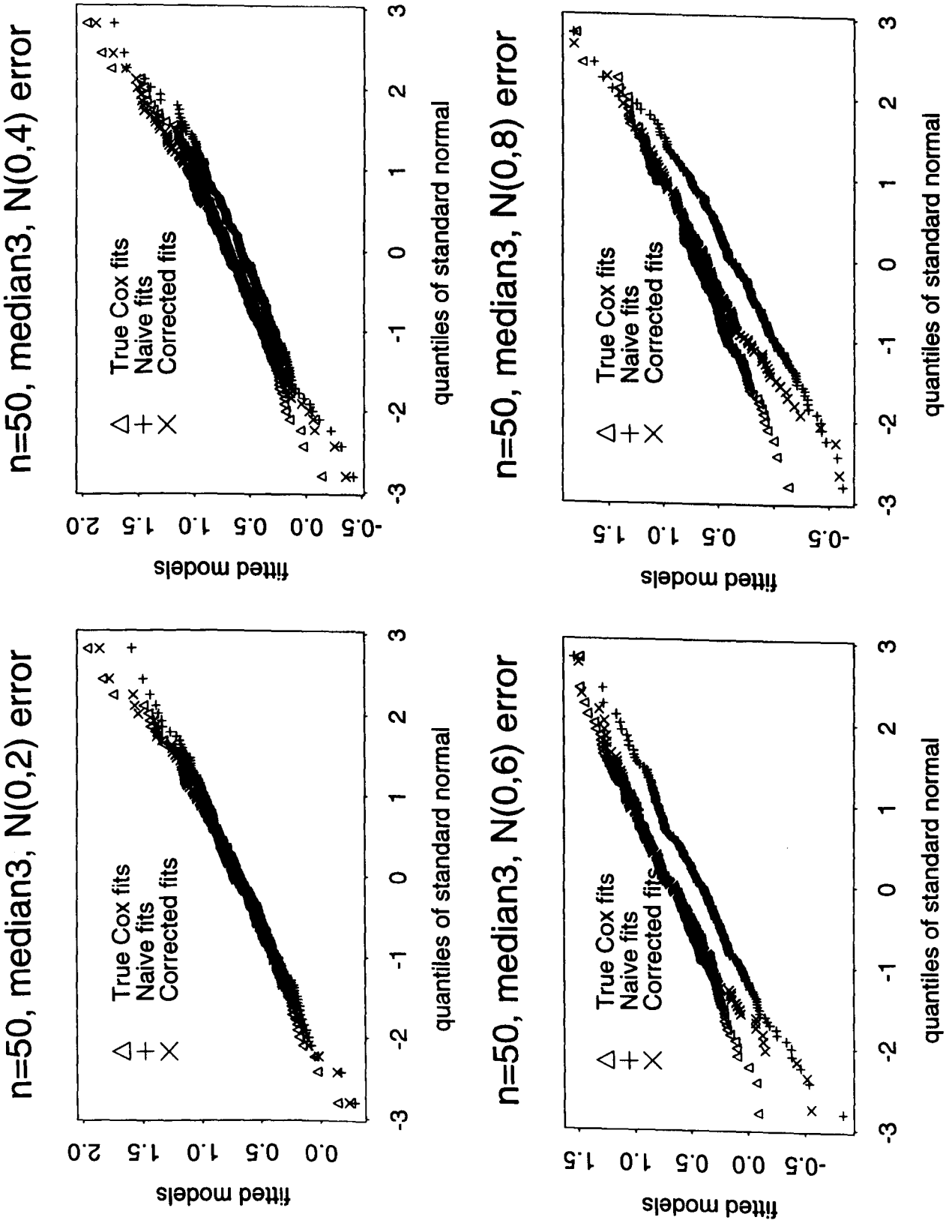


Figure 5.9: Q-Q plots against standard Normal - n=50, normal error

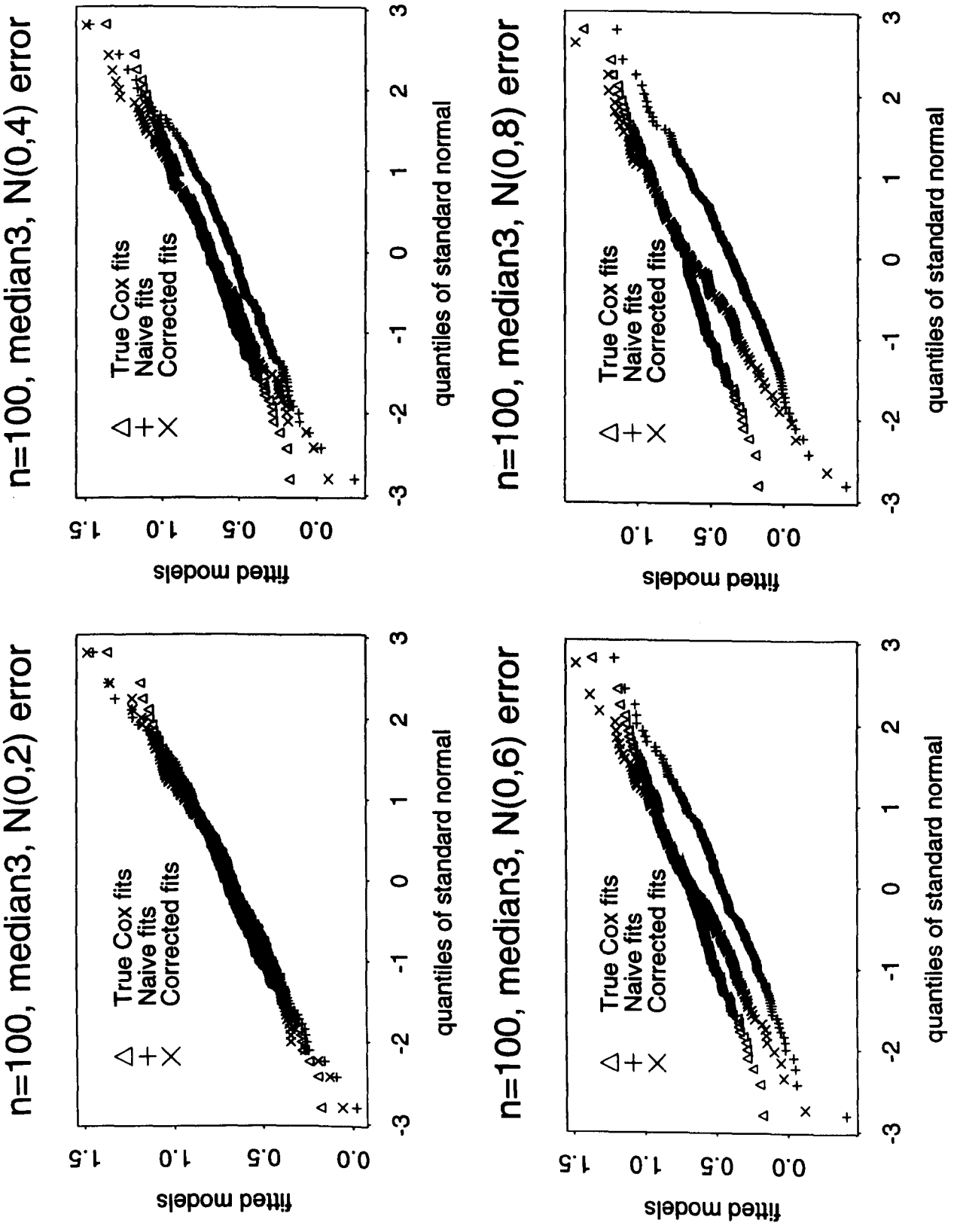


Figure 5.10: Q-Q plots against standard Normal - n=100, normal error

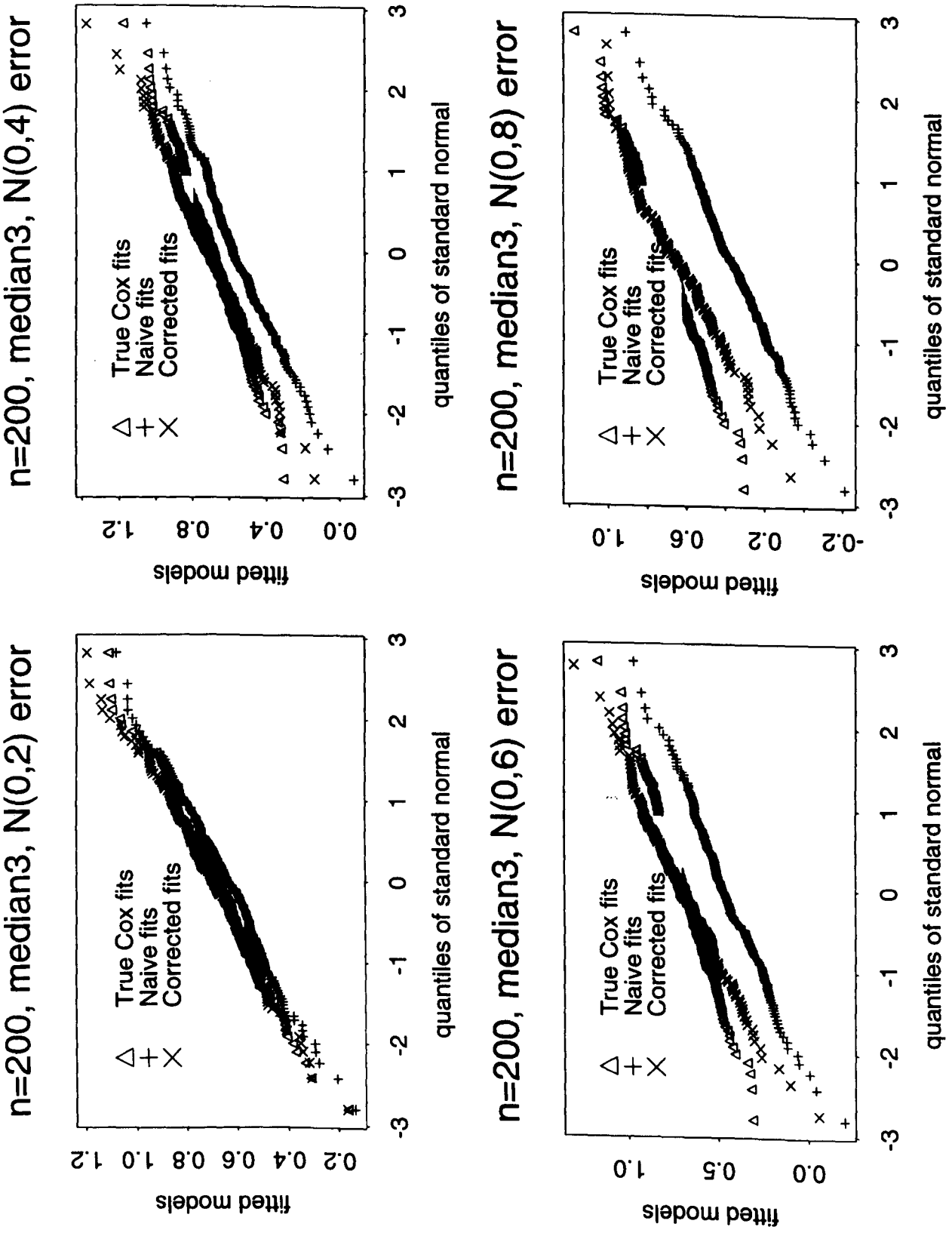


Figure 5.11: Q-Q plots against standard Normal - n=200,normal error

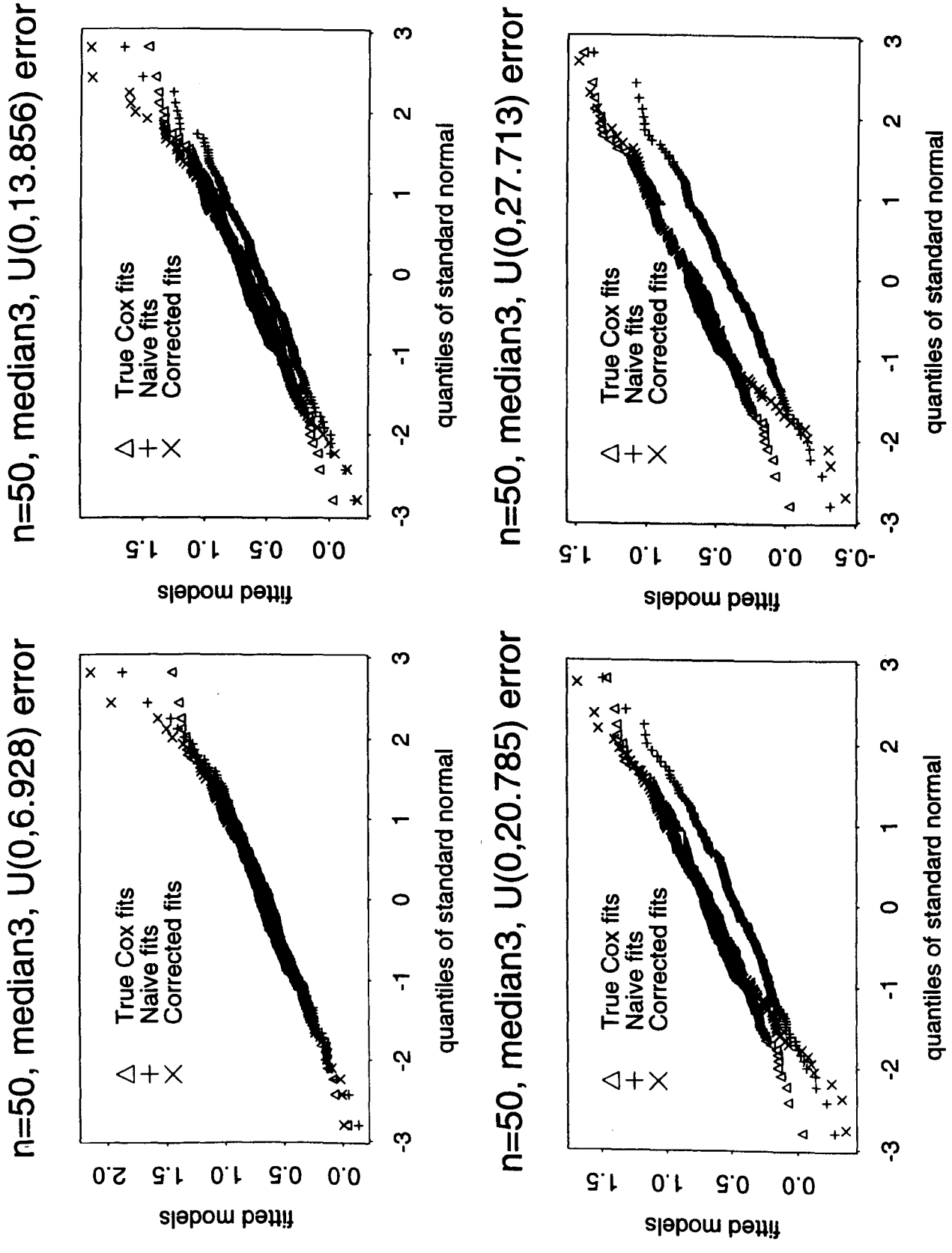


Figure 5.12: Q-Q plots against standard Normal - n=50, uniform error

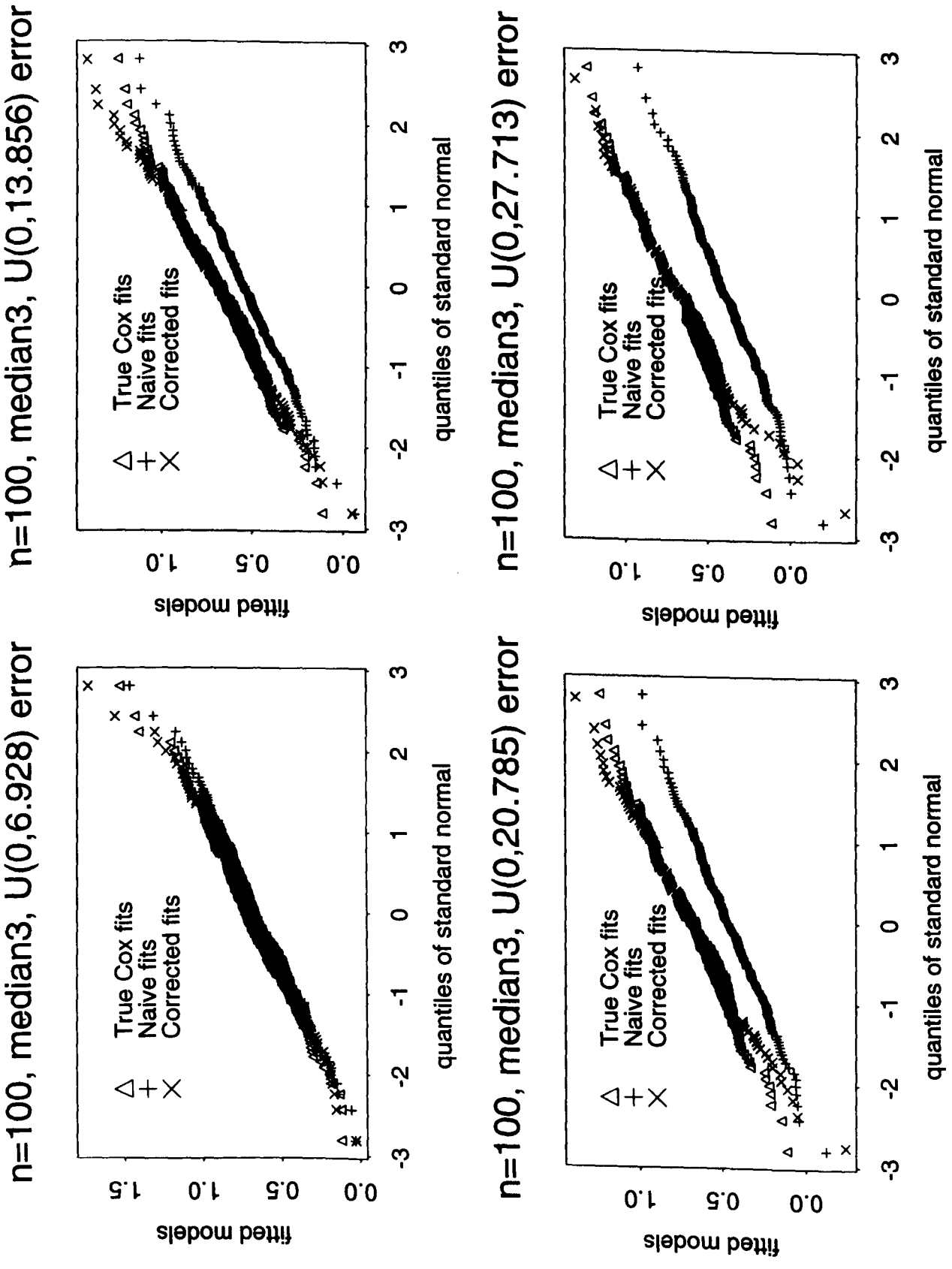


Figure 5.13: Q-Q plots against standard Normal - n=100, uniform error

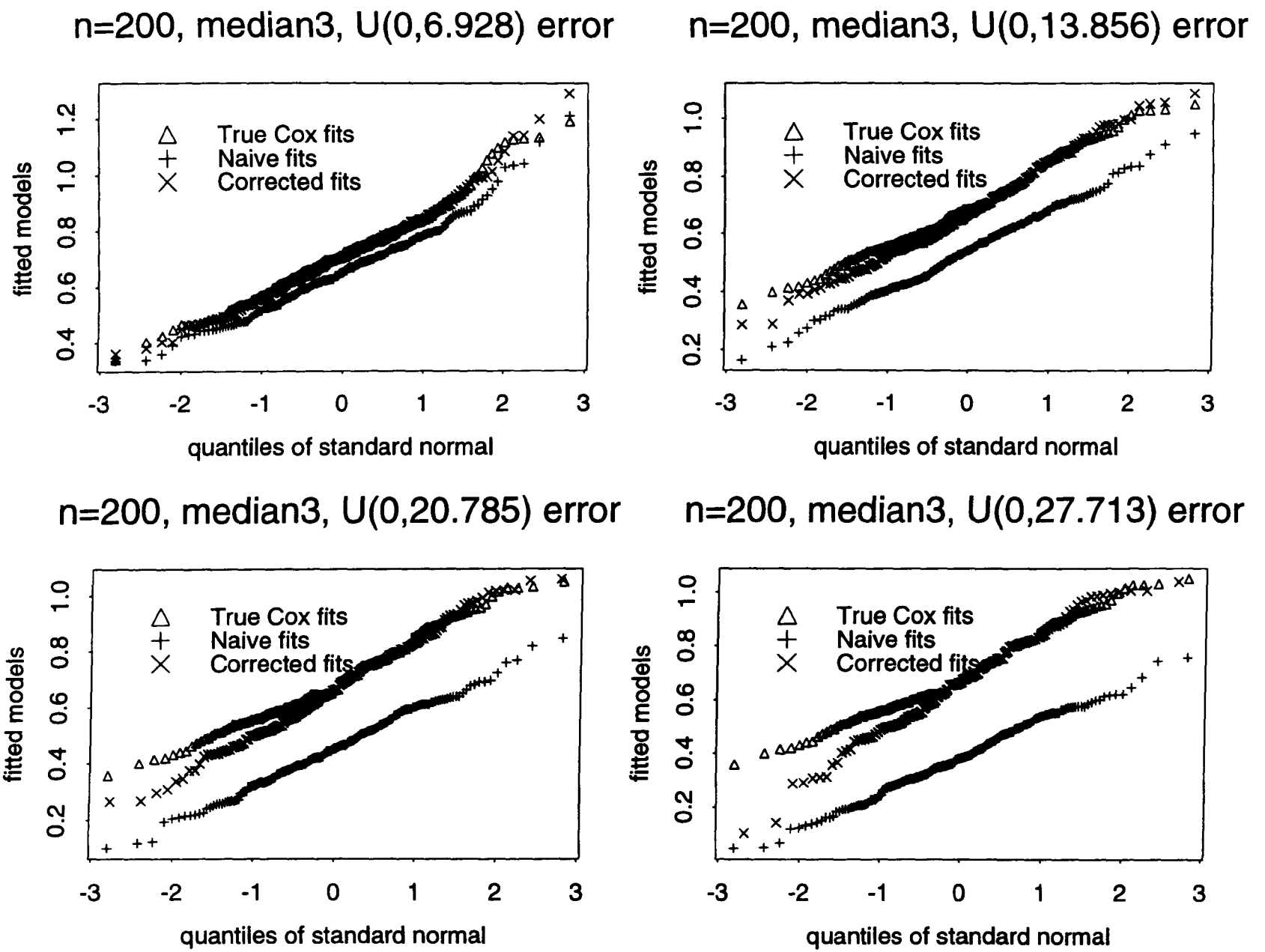


Figure 5.14: Q-Q plots against standard Normal - n=200, uniform error

5.8 Summary

In this chapter we have examined a previously unexplored statistical problem, namely general outcome error in the Cox model. The only work prior to this thesis was on rounding error, where tied data presented a complication to the partial likelihood.

It was unclear how survival error would affect estimation via the partial likelihood, a procedure dependent on the order of the survival times and the risk set at each ordered time. Intuitive arguments suggested that the direction of bias would be towards the null hypothesis of no covariate effect, namely that the relative risk over time is reduced because of error. An experiment was undertaken to give us greater understanding of the bias. It was established that the size of bias is dependent on the shape of the baseline hazard , the relative risk and the percentiles of survival. If initial survival is poor a greater proportion of individuals fail where the effect of error is large, and hence bias is increased. Bias is also increased as the relative risk is increased, indeed the proportional hazards model is valid for the observed data if there is no covariate effect. The impact of the baseline hazard on bias questions the merits of deriving theoretical results on bias analogous to the work of Hughes (1993) for covariate error. As the baseline hazard clearly affects the nature of bias one would have to specify it in order to specify the level of attenuation and hence the semi-parametric nature of the Cox model is violated. The variance of bias depends on the size of the dataset.

We presented a new method to incorporate outcome error into the partial

likelihood. This utilises simplifying assumptions in order to specify a weight representing the probability that a given survival time is in each risk set. When no error is present the likelihood is identical to the Cox likelihood, and for the case of rounding error exact weights are available and the likelihood is that of Efron. Theoretical results on the performance of the approximate likelihood were not attempted, but a justification was sought via simulation. The results were extremely encouraging. Where the observed Cox fits showed considerable bias the corrected fits were greatly improved. An advantage of the new procedure is its general nature - it is potentially applicable to any outcome error problem and does not require specification of the baseline hazard. The approximate likelihood thus offers the cancer registry a powerful tool. A prototype registry analysis is given in the next chapter.

In short in this chapter we have investigated and made progress in a new field for statisticians. Where one fears survival data might contain outcome errors the new procedure offers an estimation method that incorporates measurement error. As this work is still in development, further work is required to correct the variance of the covariate effect and establish the effectiveness of the correction in a wider variety of situations including censoring. This will be discussed in chapter 7.

Chapter 6

Outcome Error Analysis of Lung Cancer Data

6.1 Accuracy of Cancer Registry Data

If the registry had a complete and accurate data set, it would be possible to perform full survival analyses and fit mathematical models to each cancer in terms of risk factors such as social class, lifestyle, occupation and district. However, as the registry has a wide range of sources of data and collects a large number of data items errors are inevitable, and if the cancer registry has no confidence in the data it holds it would be of little value. This chapter examines how errors occur, the scope of those errors and the possible implications of them. For the problem of measuring date of diagnosis briefly discussed in chapter 2 we examine more closely where error might be coming from and how survival might be affected, and then employ the new method introduced in the previous chapter to correct for measurement error.

6.1.1 Sources of Error and Quality Control

Genuine errors in the data held by the registry can have a number of origins. Perhaps the most straightforward type of error is that in simple recording, "keying-in" errors by the registry staff, and missing data. Some of these may become apparent through computer checks. For instance, if a person's date of death is prior to their date of birth, an error has clearly occurred. Other more subtle errors can occur if a secondary tumour or recurrence of tumour are recorded as a separate case on the registry database.

Skeet (1991) and Brewster (1995) have both studied the question of assessing data quality and provide guidelines in ensuring an acceptable level of quality.

Brewster suggests the two important criteria in assessing data quality are completeness of case ascertainment and accuracy of detail and points to the following as useful indicators :

- Proportion of histologically diagnosed patients.
- Proportion of Death Certificate Only (DCO) cases where the death certificate is the initial and only point of contact with the registry
- Ratio of mortality/incidence. This can be critically assessed from knowledge of the cancer concerned.

Skeet makes a number of useful observations in ensuring the level of data quality remains acceptable. A focus should be made on ensuring consistency in essential data items such as sex whilst others such as marital status can be regarded as less important. It is argued computer systems should be designed to validate these in particular. Skeet argues that "blind re-abstractions" should periodically

be carried out to test the quality of the data. A re-abstraction of data is where the original sources of data are sought again in order to check these against what the registry has recorded. A number of such studies are now examined.

6.2 Studies in the Accuracy of Cancer Registration Data

A brief overview of five studies where re - abstraction of cancer registry data was carried out is given below. This gives us a handle on the proportion and extent of errors one can expect when presented with a cancer registry dataset. Three of the studies are from Scotland. Lapham and Waugh (1992) examine the accuracy of tumour pathology registration by comparing with pathology and hospital notes for $\approx$ 10% of all 1988 cases at the Tayside cancer registry. Of the 200 cases, 197 were traced. Brewster, Crichton and Muir (1994) reabstracted 2,200 cases from the whole of Scotland (6.9 % of all 1990 cases) and traced records for 92 % (2021 cases). A further paper by the same authors (Brewster *et al.*, 1995) is a more detailed examination of the 340 lung cancer cases of the 1990 study data, of which 309 were traced. West (1976) examines a 1 in 5 sample of cases from the South Wales Cancer Registry. Of the 1,800 cases records were found for 1,460 (81 %). Finally we examine a site specific study of bladder cancer in the Thames region (Gulliford *et al.*, 1993). The study population was all men aged less than 75 and 466 of the 609 cases were traced. Table 6.1 gives a brief overview of the results of these studies. The overall conclusions of the studies were that the quality of cancer registration data is good, but that there is room for improvement.

Table 6.1: Results of re-abstraction studies of cancer registry data

		no. in error (% of total traced)					
study	c.t.	site	d.o.b.	d.o.d.	Ann. date	yr diag	sex
Lapham & Waugh	197	8 (4)	-	-	-	-	-
Brewster et al (full)	2021	109 (5)	27 (1)	-	243*[1] (12)	131 (5)	8 (1)
Brewster et al (lung)	309	13 (4)	4 (1)	-	34*[1] (10)	18 (6)	3(1)
West	1460	93 (6)	164 (11)		198 (13.6)	112 (8)	
Gulliford et al	466	-	32 (6.2)	16 (7 *[2])	83 (16)	24 (5)	

c.t. = number of cases traced
d.o.b. = date of birth
d.o.d. = date of death
Ann. date = Anniversary date (date first treated)
*[1] classed as matched if dates lie within six weeks *[2] the total dead was 218

6.2.1 Accuracy of Lung Cancer Data

One important message of the study of Brewster et al is that although identification of site (ICD-9 3 digit code) was extremely good and hence incidence figures display a high degree of accuracy, the recording of the fourth digit i.e. specific subsite was poor and for the lung cancer data was inaccurate in 56.5 % of cases (in particular unspecified cases were assigned a specific subsite). This however will not affect our analysis as we are concerned only with ICD 162 cases (i.e. primary lung cancer) and not specific subsites. Morphology also had a high discrepancy rate with 47.2 % of cases given a new morphology on abstraction.

In terms of our analysis, recording of date of birth and sex were sufficiently accurate to allow us the assumption of no measurement error. However the recording of date treatment commenced (anniversary date) was only 90 % accurate up to six weeks, and 6 % of cases were allocated to the wrong year. This is the recorded value of date of diagnosis. Post validation survival estimates were not however significantly different.

In fact there is even greater issue regarding the calculation of survival. As Brewster et al state in their 1994 paper :

"We believe the term 'date treatment commenced' is misleading and should be abandoned in favour of 'date of diagnosis' (which is theoretically applicable to all patients)"

The rules for calculating date treatment commenced were as follows (see Appendix of 1994 paper):

- For patients who have received in patient care - insert date of first admission for investigation or treatment.
- For patients who have received out-patient care only, i.e. with no record of in-patient care for this cancer - insert date of first out-patient consultation
- For patients who have received domiciliary care only (i.e. with no record of hospital care (out-patient or in-patient) for this cancer) - insert date of diagnosis (or estimated date).

One suggestion is to use the rules regarding recording of 'incidence date' as

outlined by Maclennon (1991) which are, in order of priority:

- date of first consultation at, or admission to, a hospital, clinic or institution for the cancer in question;
- date of first diagnosis of the cancer by a physician or the date of the first pathology report - a population-based registry should seek this information only when necessary for recording the incidence date;
- date of death (year only), when the cancer is first ascertained from the death certificate and follow-back attempts have been unsuccessful; or
- date of death preceding an autopsy, when this is the time at which the cancer is first found and was unsuspected clinically (without even a vague statement, such as 'tumour suspected', 'malignancy suspected')

For MCCR the rules governing recording date of diagnosis changed for 1993. Previous to this year the date of diagnosis was recorded as anniversary date, leading to the problems outlined above. The 1993 definition of date of diagnosis is as follows:

- The date of diagnosis is derived from pathology reports (i.e. date of first report)
- For cases without histological confirmation, this becomes the date of first attendance or hospital admission during which the diagnosis was made.
- If no hospital attendance, date of diagnosis by GP etc.
- If no other information is available the date of death is used. (i.e. the death certificate provides the only information on diagnosis)

Although this differs with the Maclennon definition of 'incidence date', it ties in with his definition of 'most valid basis of diagnosis' - i.e. that the *"minimum requirement of a cancer registry is differentiation between neoplasms that are verified microscopically and those that are not"*. If one is to use this, then the date of diagnosis should be defined by the date of microscopic examination for cases where this is available. In other words, date of first pathology report represents a gold standard for date of diagnosis of cancer.

Recording of diagnosis thus presents a particular problem for survival analysis - cases that do not have microscopic confirmation may have an artificially longer survival time, and cases that were diagnosed by their GP alone have even more potential for increased observed survival. This phenomenon can be termed "window of diagnosis" and is illustrated in figure 6.1.

We wish to analyse survival on a "level playing field" for each case of diagnosis and thus employ a measurement model estimated from internal validation data in order to do this.

6.2.2 Using Internal Validation Data to Estimate the Measurement Model

We now perform a measurement error analysis on a subset of the cases for 1993. The reason for choosing the 1993 data was that this is the only year of the full dataset for which the above date of diagnosis applies. We take a random sample of 500 cases from the 1,655 occuring in 1993 in order to estimate the

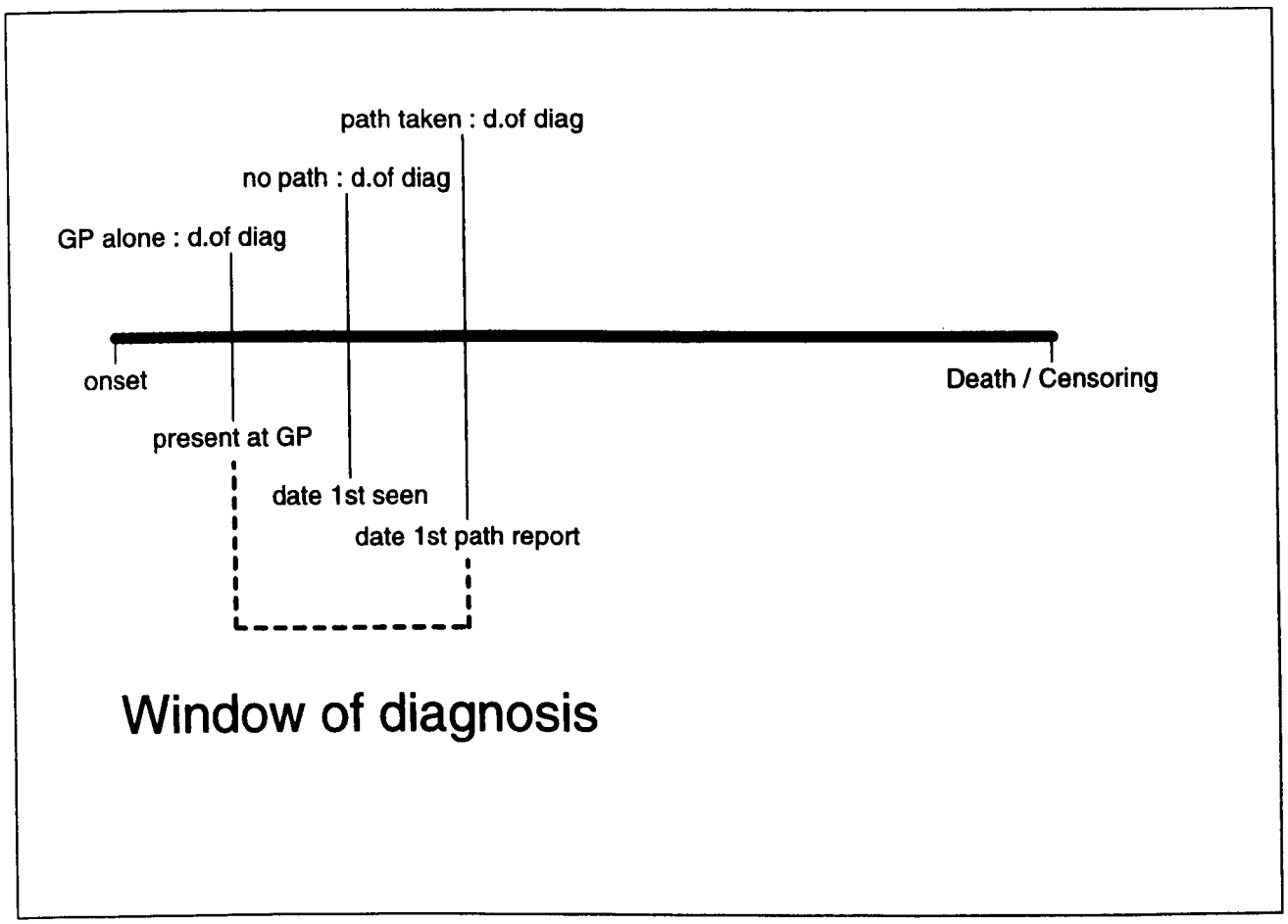


Figure 6.1: Defining date of diagnosis - potential for artificially lengthening survival

measurement error model and perform a corrected analysis. The sample was taken due to memory considerations on S-plus. When calculating large matrices the system can run out of memory and work is lost.

In order to define the true validation data we first need to identify cases who have had a microscopic diagnosis. These are defined by the proof variable. If

proof = 1,2 histology was confirmed and if proof=3 the patient gave a specimen for cytology. A clinical diagnosis is defined as proof=8. Values of 4-7 represent haematological and imaging techniques whilst 9 implies the type of proof is not known. The proof for the sample of 500 cases is given in table 6.2.

proof	1	2	3	4	5	6	7	8	9	total
frequency	241	62	40	0	0	13	15	127	2	500

Table 6.2: Variable proof for the sample of 500 patients

For the 343 cases with a microscopic confirmation a summary of date first seen at hospital and date of GP referral is given in table 6.3. For a valid measurement model it is imperative to assume that the date of diagnosis is indeed the date of first pathology report. We therefore discard cases where this date is the same as that of date first seen. We will assume for these cases that the date of first report was unretrievable and hence date first seen at hospital was recorded. These now enter our non-validation group. For validation cases that have a date of diagnosis but date first seen at hospital is missing we assume that the date of diagnosis is valid but contribution to defining the measurement model is impossible. Therefore there are 104 cases that can be used to define a measurement model for non-validation cases. Of these 28 patients have three individual ordered dates - these can be used to specify the measurement model when the observed date of diagnosis is that of date seen GP alone.

For the 157 cases with a non-microscopic confirmation a summary of date first

Table 6.3: Date information for patients with microscopic verification

	d. d = d. 1	d. d > d. 1	d. 1 missing	d.d > d. 1 > d. GP
n	216	104	23	28

d. d = date of diagnosis
d. 1 = date 1st seen at hospital
d. GP = date of GP referral

seen at hospital and date of GP referral is given in table 6.4. For cases where the date of diagnosis is after date first seen at hospital we will assume this is the recorded date of a subsequent hospital visit when a macroscopic diagnosis was made. For the 16 cases where all three dates are equal we must model the date of diagnosis assuming the patient was never referred to a hospital.

Table 6.4: Date information for patients with non-microscopic verification

	d. d = d. 1	d. d > d. 1	d. 1 missing	d.d = d. 1 = d. GP
n	133	24	0	16

d. d = date of diagnosis
d. 1 = date 1st seen at hospital
d. GP = date of GP referral

For the 104 cases in the validation group, a histogram of the time from date first seen at a hospital to date of diagnosis is given in figure 6.2. For the 28 cases used to validate time from GP referral to diagnosis, the histogram is given in figure 6.3. A table of the percentiles of the distribution of times is given in table 6.5. For simplification we assume a uniform distribution with lower limit set at the 85 percentile for each case -this is equivalent to saying diagnosis is equally likely to occur at any time across the range of the distribution. There is little support for this from the histograms, but calculations for uniform distributions

were derived in the previous chapter and using a flat distribution for true times given observed times proved effective for normal errors despite the bell shape of the normal distribution. Each case, regardless of validation status is rounded to the nearest day, hence a Berkson rounding error model is also placed on each time.

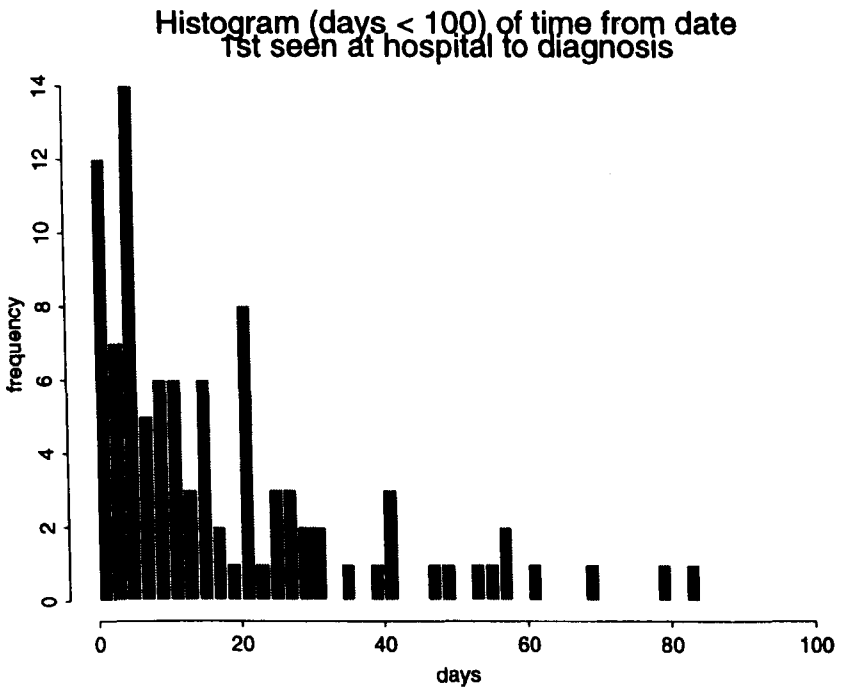


Figure 6.2: Time from hospital visit to diagnosis for validation group (full 104 cases)

percentile	15	35	50	75	85	100
first seen at hospital	3	8	14	29	52	218
gp referral	9	21	24	51	64	166

Table 6.5: Percentiles of date first seen at hospital and date of GP referral to date of diagnosis

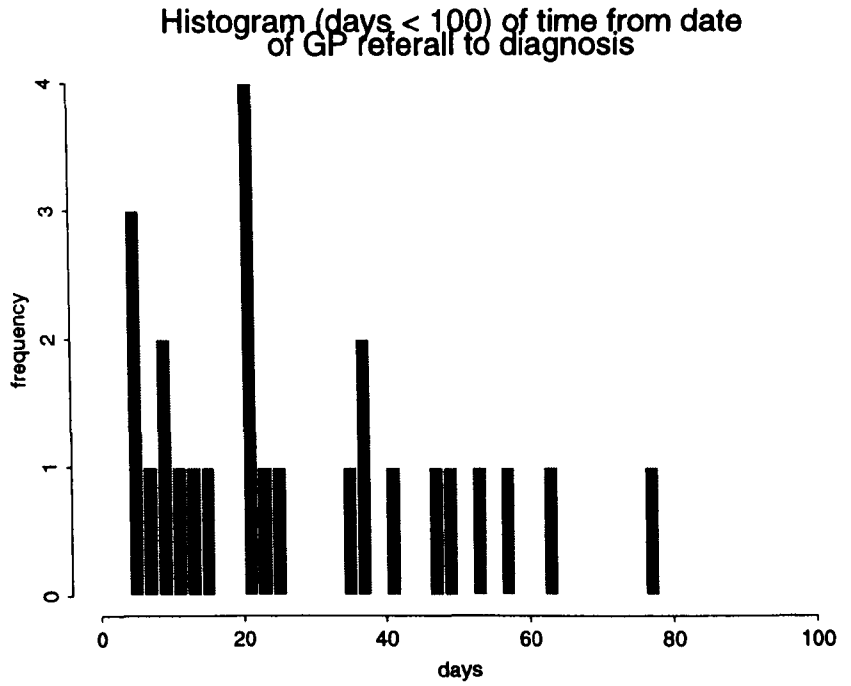


Figure 6.3: Time from GP referral to diagnosis for validation group (28 validation cases with full information)

In order to calculate the P matrix for the 500 cases, a validation status variable must be allocated to the survival times. This has values 0,1,2 depending on the measurement model applicable to each individual. A flow diagram showing how this variable was allocated is given in figure 6.4. The measurement model associated with each level of that status is:

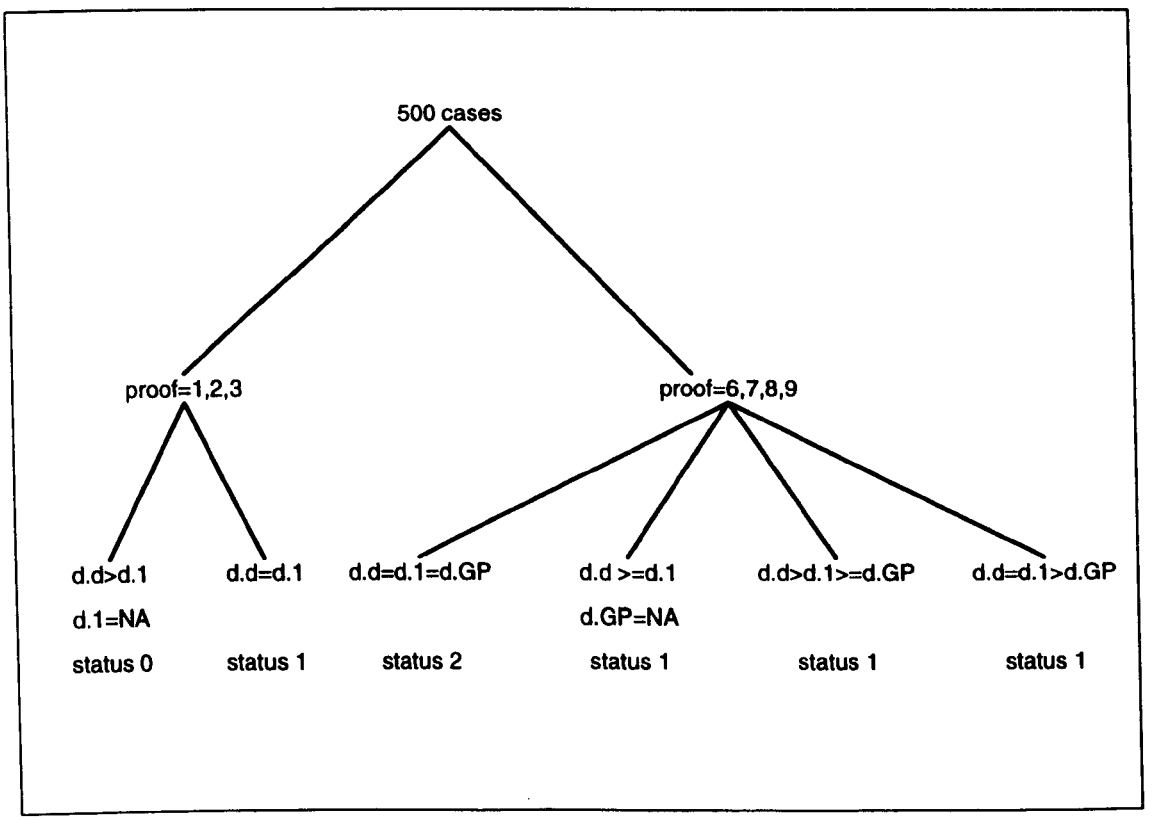
$$\text{status 0}: \quad (\text{true survival } t_i | \text{observed time } s_i) \sim U(s_i, s_i + 1)$$

$$\text{status 1}: \quad (\text{true survival } t_i | \text{observed time } s_i) \sim U(max(0, s_i - 52), s_i + 1)$$

$$\text{status 2}: \quad (\text{true survival } t_i | \text{observed time } s_i) \sim U(max(0, s_i - 64), s_i + 1)$$

In total 127 cases had validation status 0, 357 status 1 and 16 status 2.

Figure 6.4: Defining a validation status variable for each patient



d. d = date of diagnosis
d. 1 = date 1st seen at hospital
d. GP = date of GP referral

6.3 Naive and Corrected Fits for the Sample of 500 1993 Cases

6.3.1 Calculation of the P and C Matrices for the Sample

Of the sample of 500 patients, 51 were censored (see table 6.6). Hence the P matrix and C matrix have dimensions 500 * 449 as there are only 449 risk

sets contributing to the new likelihood. Appendix E.2 gives the code for the P calculation and appendix E.4 the C calculation. Censored times were assumed to be greater than all times in their *commset*. Calculation of the P matrix took about 3hrs but the time required to calculate the C matrix was negligible.

Table 6.6: Censoring for the 1993 sample of 500 patients

	frequency	percent	cumulative frequency
Censored	51	10.2	51
Dead	449	89.8	500

6.3.2 Naive and Corrected Fits for the Sample

Naive and corrected Cox fits were applied to the sample for covariates age, factor age (created around the median of age which was 68.90 yrs) and factor sex. The assumptions for the naive analysis are as follows:

- All cases are genuine 1993 cases.
- All death times are recorded correctly.
- All classification of death is correct.
- All dates of diagnosis are recorded correctly.
- All dates of birth are recorded correctly.
- All recordings of sex and age are correct.

The assumptions for the corrected analysis include the measurement model for error:

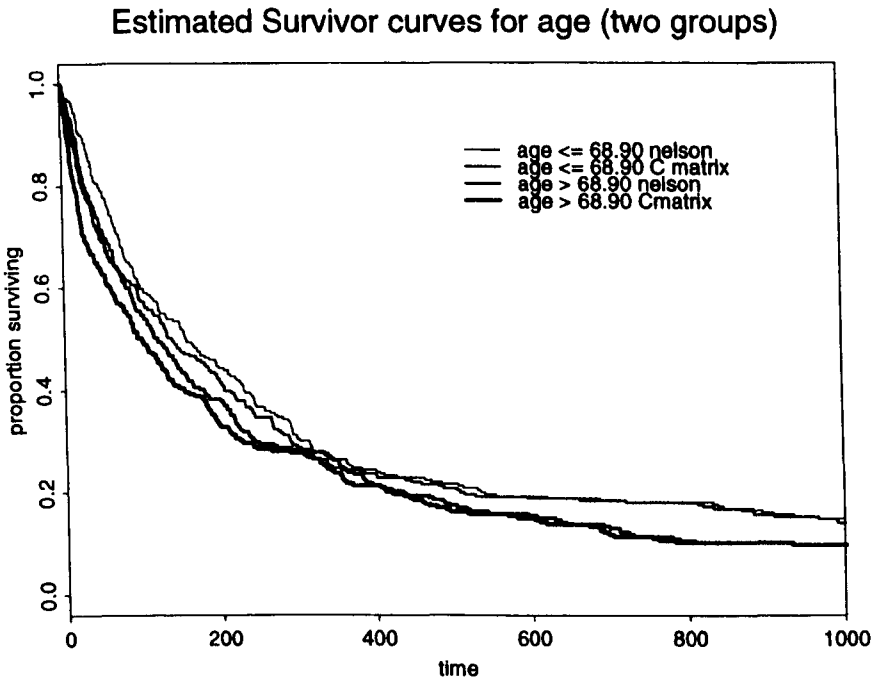


Figure 6.5: Naive survivor function (using modified Nelson cum. hazard estimate) and corrected survivor function estimates for factor age

- All cases are genuine 1993 cases.
- All death times are recorded correctly.
- All classification of death is correct.
- Dates of diagnosis are subject to error according to validation status 0,1,2.
- All dates of birth are recorded correctly.
- All recordings of sex are correct and the assigned measurement model will not unduly affect recording of age or classification of factor age.

The corrected survivor function estimates for factor age using the method outlined in section 5.5 are shown in figure 6.5 along with the naive estimate calculated using the modified Nelson cumulative hazard estimate for tied data closely allied to the correction routine. Code for fitting these is in appendix E.8. The survivor curves calculated using the measurement model and C matrix

slightly worsen the estimated survival at a given time - the estimated median survival is in table 6.7.

covariable		naive median survival (days)	corrected median survival (days)
age	≤ 68.90	156	133
	> 68.90	120	95

Table 6.7: Predicted median survival estimates for binary factor age

For the Cox modelling each covariate was fitted individually for the correction method due to the restrictions of the new Newton-Raphson routine. Table 6.8 gives the results of the naive and corrected fits (the corrected p-values were calculated using the Wald test). The coefficients of age or factor age were virtually the same in the presence of sex, as was sex in the presence of age or factor age implying that the restriction to one covariate in the corrected fits does not invalidate any conclusions. Compared to the analysis of the full dataset the effects of age and factor age are reduced in both the naive and corrected estimates, and the effect of sex is reversed but is now not significant as one would expect with a smaller dataset. The correction is small, which is in line with our knowledge gained from the simulation studies - the measurement model does not suggest a huge level of error as the error standard deviation for validation variable status 2 is about $\frac{1}{7}$ of the median survival for the whole group ($\approx$ 19 compared to a median survival of 139).

Figure 6.6 shows the estimated survivor curve for factor age for the naive and corrected fits and table 6.9 shows the estimated median survival for each factor of age.

Table 6.8: Fits to 1993 sample - covariate age,factor sex

fit type	model eqn	coeff age (s.e.) p	coeff sex (s.e.) p
Cox naive	age + sex	0.0183 (0.00482) 0.00014	0.0249 (0.0993) 0.800
Cox naive	age	0.0184 (0.00482) 0.00014	-
corrected	age	0.0187 (0.00482) 0.00010	
Cox naive	agefac + sex	0.189 (0.0945) 0.046	0.038 (0.0993) 0.700
Cox naive	agefac	0.188 (0.0945) 0.047	-
corrected	agefac	0.196 (0.0945) 0.038	-
Cox naive	sex	-	0.0349 (0.0993) 0.725
corrected	sex	-	0.0341 (0.0993) 0.731

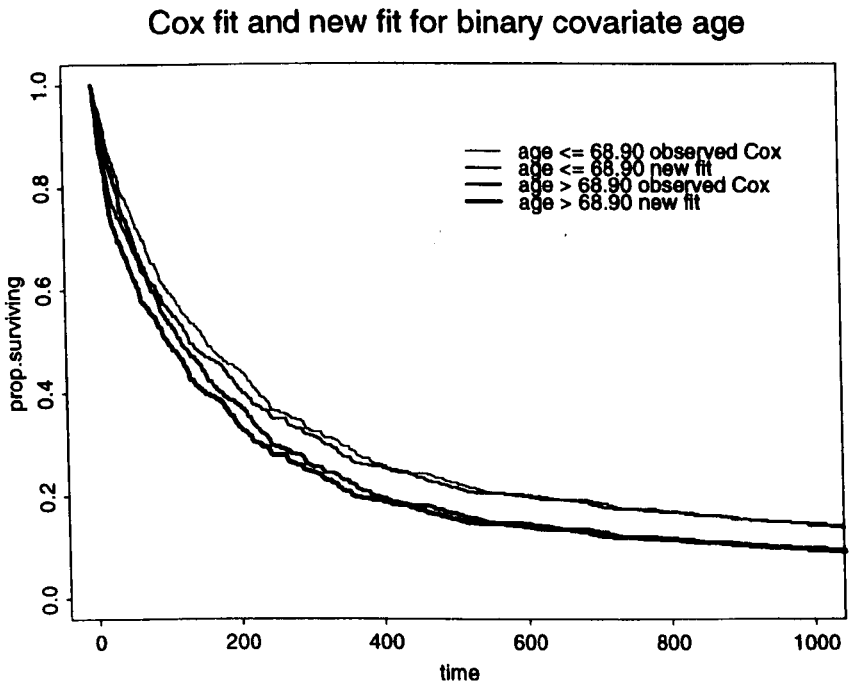


Figure 6.6: Naive Cox and corrected Cox survival estimates for factor age

covariable		naive median survival (days)	corrected median survival (days)
age	$\leq$ 68.90	155	131
	$>$ 68.90	121	101

Table 6.9: Predicted median survival estimates for binary covariate age - naive and corrected cox analysis

6.4 Summary

In this chapter we have re-analysed a sample of the 1993 data employing a measurement model to compensate for the "window of diagnosis". This analysis is not intended to be a full outcome measurement error analysis for the lung cancer data but illustrates how the new method introduced in chapter 5 can be used by the cancer registry. In order to correctly define the measurement model the minimum requirement would be to examine the paper records in the registry. A full re-abstraction of case notes would be ideal. The inclusion of all cases will become possible when calculation of the P matrix in S-plus is less memory intensive.

The analysis does however suggest that measurement error will not unduly bias the estimates of covariate effect. Naive and corrected coefficients for age and sex were extremely similar. The measurement model employed did reduce predicted survival as the observed times are not unbiased estimates of the true times. The predicted median survival estimates for binary covariate age were reduced by about three weeks for each level.

Chapter 7

Further Work and Conclusions

7.1 Summary of Thesis

In this thesis we have motivated a particular outcome error problem in survival data, namely a "window of diagnosis" for recorded episodes of cancer in the system of cancer registration. The Cox model is the routinely employed method for the analysis of survival data with available covariates. Together with non-parametric and fully parametric modelling, the Cox model was introduced in chapter 3. Procedures for estimation were reviewed. An analysis of the 40,130 lung cancer cases with positive survival was also undertaken, with the assumption of no error. Age was identified as an important covariate, and survival improved over the twenty year period under study. Overall survival for lung cancer remains poor however, with the median survival for all cases being three months.

Measurement error was introduced in chapter 4. The emphasis was on covariate error in regression models, a widely researched statistical topic. Many of the

concepts introduced were vital to the new work in chapter 5. Work on the Cox model was extensively reviewed.

Outcome error is a less well researched topic. We conducted an experiment to establish the nature of bias due to measurement error in the Cox model. It was demonstrated that despite an inherent robustness to measurement error, a correction procedure was desirable when potential bias is large. Due to the nature of the partial likelihood rounding error has been considered by authors. The Efron approximate likelihood for ties weights each tied time in each risk set according to the probability it is in the risk set. We extended this idea to more general measurement models, and using simplifying approximations suggested a new weighting of risk sets in the partial likelihood. This proved successful in bias correction for the particular experiments that were considered.

Application of the new method was undertaken in chapter 6. Five hundred 1993 cases were randomly sampled and internal validation was used to establish a measurement model for error. The effect on covariate estimates was marginal but as the measurement model was biased predicted survival was reduced.

In this chapter we examine what further work is required. As outcome error in the Cox model is a previously unconsidered problem the development of the new methodology is still in an embryonic state, and further experiments with additional theoretical justification are required. We also conclude the work that has been done.

7.2 Further Statistical Work

7.2.1 Effectiveness of the Approximate Partial Likelihood

In chapter 5 a limited simulation study was undertaken to verify the correction procedure for a normal unbiased errors-in-variables model and a uniform biased errors-in-variables model of the same standard deviation. For each of these experiments, an exponential baseline was assumed and no censorship was present. Clearly more work is required on verifying the correction for a wider variety of situations.

Little attention was paid to the effect of censoring on the estimate of the Cox model with outcome error. This is because the lung cancer data had few censored cases and these were in the extremes of the survival distribution. Hence censoring would have little effect on parameter estimation, other than to increase the proportion surviving at later death times. This led us to believe that censoring had a protective effect on bias in estimation, as is the case for covariate error. We did allude to the presence of censoring in calculating the P and C matrices, assuming if a censored time might be greater than a survival time then it is assumed to be so. However simulation studies for different types and proportions of censoring are desirable in the future, to understand the implications in both the naive and corrected likelihoods.

We established that different shaped baseline hazards led to different levels of bias. In particular a Weibull shape parameter of $\rho = 1.5$ led to greater bias than $\rho = 1$ (exponential). Again simulations for the corrected likelihood are desirable

here.

Extensions to multivariate problems are also important. The code given in the appendix needs modification in order to fit corrected Cox models with many covariates. Investigation is required into how the new likelihood corrects for bias when continuous covariates and confounding variables are present.

The theoretical properties of the new likelihood have not been established. Q-Q plots given in chapter 5 show the new fits are approximately normally distributed with mean equal to the true Cox fits. Asymptotic expectation and results are however important. Due to the approximations involved it is unclear how such expectations could be established. For instance, the expectation of the P and C matrices as the number of failures tends to infinity.

7.2.2 Variance of the New β

The simulation studies in chapter 5 showed that, particularly for the case of normal errors, that the estimated variance of the corrected β estimate was too small, and hence hypothesis tests are anti-conservative. Carroll, Ruppert and Stefanski (1995) outline a procedure to estimate corrected confidence intervals using the resampling pairs bootstrap. This involves sampling M pairs of outcome and covariate values with replacement from the original data and then estimating the corrected variance by:

$$\hat{var}(\hat{\beta}) = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\beta}^{(m)} - \bar{\beta})^2 \tag{7.1}$$

It would appear this procedure is potentially applicable, though some caution is necessary as the resampled data would exhibit heavy ties. The main drawback of the procedure is the need to undertake $M + 1$ calculations of the C and P matrices (1 for the original fit, and M for the variance bootstrap). Work is required to test whether this would prove effective. At present one should guard against making conclusions on P-values calculated using the variance estimate.

7.2.3 Relationship with Model Mis-specification

Our example showed that fitting the Cox model to the observed data is equivalent to mis-specifying the proportional hazards assumption. Theoretical results could be obtained by examination of the likelihood under the observed data in order to gauge the asymptotic expectation of the bias. This is closely related to the work of Mr Paul Monaghan on the sister project to this. Omission of covariates (i.e. increased residual variance) or mis-specifying proportional hazards leads to bias in estimates via the partial likelihood.

The work of Hughes reviewed in chapter 4 did exactly this for the problem of covariates following an errors-in-variables model. He demonstrated that expected levels of attenuation can approximately be obtained by specifying the error standard deviation and level of censoring alone. Hence a direct correction of the naive estimate is possible. Our work in chapter 5 showed that the degree of bias is clearly dependent on the baseline hazard, implying that theoretical bias results could not be obtained based on the measurement model, level of censoring and size of error. Therefore although examination of the observed likelihood under various baselines may prove interesting, it is unlikely to yield a

simple correction.

This outlines an advantage of the procedure suggested in this thesis, namely that specification of the baseline hazard is not required, up to approximation of the distribution of $true|observed$.

7.2.4 Further Work in Cancer Epidemiology

The re-analysis carried out in chapter 6 proved a useful exercise in demonstrating the potential of the new methodology for including outcome error in a Cox analysis of cancer registration data. The analysis employed was essentially a prototype and by no means optimal.

Observed survival times that were zero or negative were not included, nor were Death Certificate Only cases (see chapters 2 and 6). It is perfectly feasible to include such cases in a corrected analysis if careful thought is given to the measurement model employed.

The assumed "gold-standard" data is also questionable. In order to verify the assumed measurement model examination of within registry paper records or re-abstraction of case notes is required. It would also be desirable to examine the date of diagnosis for those cases with a microscopic verification but whose date of diagnosis was not assumed to be recorded (i.e. the observed diagnosis was equal to the observed date first seen). The measurement model employed in the chapter 6 analysis was the same as for cases with macroscopic verification, but this is unlikely to be the case.

There is also little reason that, when carefully considered, a re-analysis could not be applied to the whole dataset when calculation of the P matrix becomes computationally feasible. Recall that pre-1993 patients were assigned dates of diagnosis by anniversary date, or date first treated. One would have to closely examine records in order to establish measurement models for true date of diagnosis for those cases that received treatment, and those who did not.

It is also necessary when assuming error in date of diagnosis that there is also error in age at diagnosis. For simplicity we made the assumption that error was sufficiently small in age as to ignore it. This is not unreasonable, particularly for the binary covariate age as the only concern are ages close to the boundary of the defined binary covariate.

7.3 Conclusions

7.3.1 Statistical Conclusions

The Cox Proportional Hazards Model

For continuous survival data, or discrete tied data under the assumption of rounding error the Cox proportional hazards model is by far the most commonly employed model to estimate the effects of prognostic factors on a survival outcome. This is estimated using partial likelihood, based on the factorisation of conditional survival probabilities. When the data exhibit ties, a variety of "exact" and approximate likelihoods are available. These were reviewed in

chapter 3. Many authors agree that the Efron approximation for ties is a simple and satisfactory procedure.

Covariate Error and the Cox Proportional Hazards Model

Since the first paper appeared on covariate error in the Cox model (Prentice, 1982) there has been considerable further research. The nature of attenuation is well understood and correction for covariate error is now a well established statistical principle in Cox modelling. Use of the correction is however less common. All the methods examined in chapter 4 of this thesis assume a Gaussian error, however this need not always be the case - our employed model for diagnosis error would not result in Normal errors for recorded age.

Attenuation due to Outcome Error in the Cox Model

We have established by simulation that the Cox model is indeed fairly robust to outcome error. Attenuation is unaffected by the mean of the error distribution. When overall survival is poor and measurement error is large, expected bias in the observed covariate effect is considerable. The direction of bias is towards the null hypothesis of no covariate effect. The variance of the distribution of bias is affected by the size of the dataset, and as one would expect small datasets demonstrate greater variance in the bias. Bias is increased when the size of the covariate effect is increased, and is affected by the shape of the baseline hazard.

Correction for Outcome Error

Although extreme conditions are required for attenuation to be considerable when outcome error is present, it is possible to severely underestimate a covariate effect by fitting a naive Cox fit to the observed data. No methodology was available to apply a corrected fit for general outcome error problems in the Cox model prior to this thesis. By extending the idea of approximating the partial likelihood for rounding error we have presented a simple and easily applicable procedure for approximating to the true partial likelihood when a measurement model for outcome error is assumed. This proved effective in correcting for attenuation in a simulation study. The correction procedure is, if employed with care, potentially applicable to any type of error, and is also closely related to the problem of interval censoring. More work is however required to establish how well the correction works for a wider set of baseline hazards, censoring levels, types of censorship and multiple covariate problems.

Joint Correction for Outcome and Covariate Error

Although this has not been attempted in the thesis, there is no reason why the correction procedure could not be used in conjunction with covariate correction procedures. The usual assumption however for regression calibration is that failure is rare. This implies that changes in risk sets over time are small and hence the distribution of the $true|observed$ covariate does not substantially change over the period under study. In this situation one would anticipate that the effect of outcome error is also small, as a small number of well spread failure times would result in most elements of the C matrix being either close to 1 or close to 0.

7.3.2 Conclusions for Cancer Epidemiology

It is reassuring to the cancer registry that the work in this thesis has shown only extreme measurement error in survival causes the epidemiologist concern when fitting the Cox model. Our main concern has been the "window of diagnosis" - the potential for lengthening of survival due to the point at which diagnosis is made for an individual cancer episode. It would seem from chapter 6 that concern over severe bias due to this is not great - however the selection and accuracy of validation cases is still in question. A fully desirable analysis would at least require examination of paper records in the registry and ideally a re-abstraction of case notes. Generally however there would only be a problem for an extremely lethal cancer with very poor recording of dates.

However this is not the only type of data imperfection one encounters when analysing a cancer registry dataset. Our analysis in chapter 6 was largely to illustrate the potential use of the developed methodology by cancer registry researchers.

There are a number of wider uses and issues that arise from this discussion. The main concern is to put all cases on a "level playing field" in order to properly estimate the effect of prognostic factors. Between registry comparisons of survival are also problematic as recording of dates and quality of data may vary considerably. In general more consistent definitions of dates with more careful application are required. Detection in certain regions may be superior but survival only appears superior due to this early detection. Likewise the introduction of screening may only artificially increase a patient's survival. Hence there are many possible applications of the new methodology to cancer

registration data.

7.4 Overall Conclusion

The examination of outcome error in survival data is a previously unconsidered problem. Hence the field covered in this thesis is , with the exception of rounding error, a novel one. We have proposed a correction for outcome error in estimation of non-parametric survivor curves and the semi-parametric Cox model that proves successful in correcting for bias. This has many potential applications, and is a useful new weapon in the armoury of the cancer registry. There is, however, need for additional work in order to gauge the effectiveness of the method in a wider variety of settings.

Appendix A

A.1 Derivation of the Distribution of true|observed for the Normal Errors-in-Variables Model

Under the assumption that the observed variable z is related to the true variable x via the traditional errors-in-variables model where x is normally distributed we have:

$$z = x+u, \text{ where } x \sim N(\mu, \sigma_x^2),\ u \sim N(0, \sigma_u^2),\ z|x \sim N(x, \sigma_u^2),\ z \sim N(\mu, \sigma_x^2+\sigma_u^2) \tag{A.1}$$

Employing the identity $f_{x,z} = f_x f_{z|x}$ we can state the joint density of x, z:

$$f_{x,z} = \frac{1}{\sigma_x\sqrt{2\pi}}\frac{1}{\sigma_u\sqrt{2\pi}}\exp(-\frac{1}{2\sigma_x^2}(x-\mu)^2)\exp(-\frac{1}{2\sigma_u^2}(z-x)^2) \tag{A.2}$$

Then using $f_{x|z} = \frac{f_{x,z}}{f_z}$

$$\begin{aligned} f_{x|z} &= \frac{\sqrt{(\sigma_x^2+\sigma_u^2)}\sqrt{2\pi}\exp(-\frac{1}{2\sigma_x^2}(x-\mu)^2)\exp(-\frac{1}{2\sigma_u^2}(z-x)^2)}{\sqrt{(\sigma_x^2\sigma_u^2)2\pi}\exp(-\frac{1}{2\sigma_x^2+\sigma_u^2}(z-\mu)^2)} \\ &= \frac{1}{\sqrt{2\pi}\sqrt{(\frac{\sigma_x^2\sigma_u^2}{\sigma_x^2+\sigma_u^2})}}\exp(-\frac{1}{2}\{\frac{(x-\mu)^2}{\sigma_x^2}+\frac{(z-x)^2}{\sigma_u^2}-\frac{(z-\mu)^2}{\sigma_x^2+\sigma_u^2}\}) \end{aligned}$$

Writing $\lambda = \frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}$ and $1-\lambda = \frac{\sigma_u^2}{\sigma_x^2+\sigma_u^2}$ then :

$$\begin{aligned} f_{x|z} &= \frac{1}{\sqrt{2\pi}\sqrt{(\lambda\sigma_u^2)}}\exp(\frac{1}{2\lambda\sigma_u^2}\{(1-\lambda)(x-\mu)^2+\lambda(z-x)^2-\lambda(1-\lambda)(z-\mu)^2\}) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{(\lambda\sigma_u^2)}}\exp(\frac{1}{2\lambda\sigma_u^2}\{(1-\lambda)(x^2-2\mu x+\mu^2)+\lambda(z^2-2xz+x^2) \\ &\quad -(\lambda(1-\lambda))(z^2-2\mu z+\mu^2)\}) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{(\lambda\sigma_u^2)}}\exp(\frac{1}{2\lambda\sigma_u^2}\{x^2-2x(\lambda z+(1-\lambda\mu))+(\lambda z+(1-\lambda)\mu)^2\}) \\ &= \frac{1}{\sqrt{2\pi}\sqrt{(\lambda\sigma_u^2)}}\exp(\frac{1}{2\lambda\sigma_u^2}\{x-(\lambda z+(1-\lambda)\mu)\}^2) \end{aligned}$$

Hence $x|z \sim N(\lambda z+(1-\lambda)\mu, \lambda\sigma_u^2)$.

Appendix B

B.1 Full Results of Simulation Studies - Naive Fits to Ascertain Level of Bias

We desire to determine the effect of outcome error on the naive fit of a Cox regression when outcome error is present. Recall a total of 8 different combinations of τ (Weibull shape parameter) , n (size of total dataset) and $\exp(\beta)$ were chosen. For each combination we then set the baseline hazard so the overall total median survival for both groups (i.e. the full n patients) was 3,12, and 30 (months). Outcome errors arising from both the usual normal and a biased $U(0, b)$ measurement model with standard deviations $\frac{1}{30}, \frac{1}{2}, 1$ and 2 months were considered. Hence a total number of 192 individual experiments were carried out. Each experiment samples 200 values from the distribution of the observed β. A table for the level of attenuation for each case is given in chapter five (table 5.5). Detailed results for each experiment are now given:

Table B.1: Results for naive fits: exponential data: rr=2 : (n=50)

n=50, $\beta_{true} = 0.693 = \log 2$,seed=1					
normal errors: mean $(\beta_{sim}) = 0.7221$, var $(\beta_{sim}) = 0.1063$					
uniform errors: mean $(\beta_{sim}) = 0.7020$, var $(\beta_{sim}) = 0.0784$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0,(\frac{1}{30})^2)$	3	0.7234	0.1053	0.0013	0.0004
$U(0,0.115)$	3	0.7021	0.0785	0.0001	0.0001
$N(0,(\frac{1}{2})^2)$	3	0.7186	0.1096	-0.0035	0.0073
$U(0,1.732)$	3	0.7015	0.0821	-0.0005	0.0037
$N(0,1)$	3	0.7157	0.1083	-0.0065	0.0175
$U(0,3.464)$	3	0.6951	0.0850	-0.0069	0.0091
$N(0,2^2)$	3	0.6781	0.1005	-0.0440	0.0381
$U(0,6.928)$	3	0.6567	0.0864	-0.0453	0.0235
$N(0,(\frac{1}{30})^2)$	12	0.7223	0.1056	0.0002	0.0001
$U(0,0.115)$	12	0.7022	0.0786	0.0002	0.0000
$N(0,(\frac{1}{2})^2)$	12	0.7246	0.1038	0.0024	0.0016
$U(0,1.732)$	12	0.7007	0.0782	-0.0013	0.0005
$N(0,1)$	12	0.7224	0.1056	0.0003	0.0038
$U(0,3.464)$	12	0.7048	0.0806	0.0028	0.0014
$N(0,2^2)$	12	0.7186	0.1096	-0.0035	0.0073
$U(0,6.928)$	12	0.7015	0.0821	-0.0005	0.0037
$N(0,(\frac{1}{30})^2)$	30	0.7216	0.1060	-0.0005	0.0000
$U(0,0.115)$	30	0.7020	0.0783	0.0000	0.0000
$N(0,(\frac{1}{2})^2)$	30	0.7224	0.1051	0.0002	0.0005
$U(0,1.732)$	30	0.7011	0.0784	-0.0009	0.0002
$N(0,1)$	30	0.7245	0.1038	0.0023	0.0013
$U(0,3.464)$	30	0.7008	0.0782	-0.0012	0.0004
$N(0,2^2)$	30	0.7209	0.1038	-0.0012	0.0028
$U(0,6.928)$	30	0.7031	0.0788	0.0011	0.0010

Table B.2: Results for naive fits: exponential data: rr=2 : (n=100)

n=100, $\beta_{true} = 0.693 = \log 2$,seed=2					
normal errors: mean $(\beta_{sim}) = 0.7072$, var $(\beta_{sim}) = 0.0435$					
uniform errors: mean $(\beta_{sim}) = 0.7113$, var $(\beta_{sim}) = 0.0585$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0,(\frac{1}{30})^2)$	3	0.7073	0.0444	0.0001	0.0002
$U(0,0.115)$	3	0.7117	0.0585	0.0005	0.0000
$N(0,(\frac{1}{2})^2)$	3	0.7126	0.0485	0.0054	0.0031
$U(0,1.732)$	3	0.7058	0.0587	-0.0054	0.0008
$N(0,1)$	3	0.6981	0.0494	-0.0091	0.0074
$U(0,3.464)$	3	0.6911	0.0582	-0.0202	0.0028
$N(0,2^2)$	3	0.6705	0.0542	-0.0368	0.0149
$U(0,6.928)$	3	0.6454	0.0565	-0.0658	0.0088
$N(0,(\frac{1}{30})^2)$	12	0.7070	0.0439	-0.0002	0.0001
$U(0,0.115)$	12	0.7114	0.0585	0.0001	0.0000
$N(0,(\frac{1}{2})^2)$	12	0.7095	0.0449	0.0022	0.0007
$U(0,1.732)$	12	0.7108	0.0588	-0.0004	0.0001
$N(0,1)$	12	0.7083	0.0450	0.0011	0.0013
$U(0,3.464)$	12	0.7108	0.0590	-0.0005	0.0003
$N(0,2^2)$	12	0.7126	0.0485	0.0054	0.0031
$U(0,6.928)$	12	0.7058	0.0587	-0.0054	0.0008
$N(0,(\frac{1}{30})^2)$	30	0.7073	0.0438	0.0001	0.0000
$U(0,0.115)$	30	0.7113	0.0588	0.0001	0.0000
$N(0,(\frac{1}{2})^2)$	30	0.7082	0.0449	0.0010	0.0003
$U(0,1.732)$	30	0.7116	0.0583	0.0003	0.0000
$N(0,1)$	30	0.7087	0.0450	0.0015	0.0005
$U(0,3.464)$	30	0.7114	0.0584	0.0001	0.0001
$N(0,2^2)$	30	0.7022	0.0455	-0.0001	0.0012
$U(0,6.928)$	30	0.7110	0.0589	-0.0003	0.0002

Table B.3: Results for naive fits: exponential data: rr=2 : (n=200)

n=200, $\beta_{true} = 0.693 = \log 2$,seed=3					
normal errors: mean $(\beta_{sim}) = 0.6888$, var $(\beta_{sim}) = 0.0240$					
uniform errors: mean $(\beta_{sim}) = 0.7189$, var $(\beta_{sim}) = 0.0222$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0,(\frac{1}{30})^2)$	3	0.6890	0.0245	0.0002	0.0001
$U(0,0.115)$	3	0.7189	0.0222	0.0000	0.0000
$N(0,(\frac{1}{2})^2)$	3	0.6860	0.0260	-0.0028	0.0014
$U(0,1.732)$	3	0.7120	0.0218	-0.0069	0.0004
$N(0,1)$	3	0.6747	0.0259	-0.0141	0.0025
$U(0,3.464)$	3	0.6995	0.0215	-0.0194	0.0012
$N(0,2^2)$	3	0.6452	0.0271	-0.0436	0.0066
$U(0,6.928)$	3	0.6574	0.0202	-0.0615	0.0040
$N(0,(\frac{1}{30})^2)$	12	0.6886	0.0243	-0.0002	0.0000
$U(0,0.115)$	12	0.7189	0.0223	0.0000	0.0000
$N(0,(\frac{1}{2})^2)$	12	0.6873	0.0250	-0.0016	0.0003
$U(0,1.732)$	12	0.7181	0.0221	-0.0009	0.0000
$N(0,1)$	12	0.6884	0.0254	-0.0004	0.0006
$U(0,3.464)$	12	0.7168	0.0220	-0.0022	0.0001
$N(0,2^2)$	12	0.6860	0.0260	-0.0028	0.0014
$U(0,6.928)$	12	0.7120	0.0218	-0.0069	0.0004
$N(0,(\frac{1}{30})^2)$	30	0.6888	0.0242	-0.0001	0.0000
$U(0,0.115)$	30	0.7189	0.0223	0.0000	0.0000
$N(0,(\frac{1}{2})^2)$	30	0.6890	0.0249	0.0001	0.0001
$U(0,1.732)$	30	0.7187	0.0223	-0.0002	0.0000
$N(0,1)$	30	0.6877	0.0250	-0.0012	0.0003
$U(0,3.464)$	30	0.7186	0.0222	-0.0004	0.0000
$N(0,2^2)$	30	0.6887	0.0252	-0.0001	0.0006
$U(0,6.928)$	30	0.7176	0.0220	-0.0014	0.0001

Table B.4: Results for naive fits: exponential data: rr=2 : (n=500)

n=500, $\beta_{true} = 0.693 = \log 2$,seed=4					
normal errors: mean $(\beta_{sim}) = 0.6973$, var $(\beta_{sim}) = 0.0076$					
uniform errors: mean $(\beta_{sim}) = 0.6908$, var $(\beta_{sim}) = 0.0084$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0, (\frac{1}{30})^2)$	3	0.6971	0.0076	-0.0002	0.0000
$U(0, 0.115)$	3	0.6907	0.0084	-0.0001	0.0000
$N(0, (\frac{1}{2})^2)$	3	0.6934	0.0076	-0.0039	0.0005
$U(0, 1.732)$	3	0.6857	0.0086	-0.0051	0.0001
$N(0, 1)$	3	0.6826	0.0084	-0.0147	0.0010
$U(0, 3.464)$	3	0.6735	0.0089	-0.0173	0.0005
$N(0, 2^2)$	3	0.6467	0.0088	-0.0506	0.0026
$U(0, 6.928)$	3	0.6342	0.0092	-0.0566	0.0015
$N(0, (\frac{1}{30})^2)$	12	0.6977	0.0076	0.0004	0.0000
$U(0, 0.115)$	12	0.6909	0.0084	0.0001	0.0000
$N(0, (\frac{1}{2})^2)$	12	0.6967	0.0075	-0.0006	0.0001
$U(0, 1.732)$	12	0.6903	0.0084	-0.0005	0.0000
$N(0, 1)$	12	0.6953	0.0074	-0.0020	0.0002
$U(0, 3.464)$	12	0.6892	0.0085	-0.0016	0.0000
$N(0, 2^2)$	12	0.6934	0.0076	-0.0039	0.0005
$U(0, 6.928)$	12	0.6857	0.0086	-0.0051	0.0001
$N(0, (\frac{1}{30})^2)$	30	0.6975	0.0076	0.0001	0.0000
$U(0, 0.115)$	30	0.6908	0.0084	0.0000	0.0000
$N(0, (\frac{1}{2})^2)$	30	0.6968	0.0076	-0.0005	0.0000
$U(0, 1.732)$	30	0.6906	0.0084	-0.0002	0.0000
$N(0, 1)$	30	0.6970	0.0076	-0.0003	0.0001
$U(0, 3.464)$	30	0.6905	0.0084	-0.0003	0.0000
$N(0, 2^2)$	30	0.6958	0.0074	-0.0015	0.0002
$U(0, 6.928)$	30	0.6897	0.0084	-0.0011	0.0000

Table B.5: Results for naive fits: exponential data: rr=2 : (n=1000)

n=1000, $\beta_{true} = 0.693 = \log 2$,seed=5					
normal errors: mean $(\beta_{sim}) = 0.6946$, var $(\beta_{sim}) = 0.0040$					
uniform errors: mean $(\beta_{sim}) = 0.7000$, var $(\beta_{sim}) = 0.0044$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0, (\frac{1}{30})^2)$	3	0.6945	0.0040	-0.0001	0.0000
$U(0, 0.115)$	3	0.7000	0.0044	0.0000	0.0000
$N(0, (\frac{1}{2})^2)$	3	0.6928	0.0045	-0.0018	0.0003
$U(0, 1.732)$	3	0.6961	0.0044	-0.0039	0.0001
$N(0, 1)$	3	0.6844	0.0046	-0.0102	0.0007
$U(0, 3.464)$	3	0.6843	0.0043	-0.0157	0.0002
$N(0, 2^2)$	3	0.6521	0.0051	-0.0424	0.0015
$U(0, 6.928)$	3	0.6458	0.0043	-0.0512	0.0007
$N(0, (\frac{1}{30})^2)$	12	0.6944	0.0040	-0.0002	0.0000
$U(0, 0.115)$	12	0.6999	0.0044	0.0000	0.0000
$N(0, (\frac{1}{2})^2)$	12	0.6949	0.0041	0.0003	0.0001
$U(0, 1.732)$	12	0.6998	0.0044	-0.0002	0.0000
$N(0, 1)$	12	0.6950	0.0042	0.0004	0.0001
$U(0, 3.464)$	12	0.6991	0.0049	-0.0009	0.0000
$N(0, 2^2)$	12	0.6928	0.0043	-0.0018	0.0003
$U(0, 6.928)$	12	0.6961	0.0044	-0.0039	0.0001
$N(0, (\frac{1}{30})^2)$	30	0.6946	0.0039	0.0000	0.0000
$U(0, 0.115)$	30	0.6999	0.0049	0.0000	0.0000
$N(0, (\frac{1}{2})^2)$	30	0.6943	0.0040	-0.0003	0.0000
$U(0, 1.732)$	30	0.7000	0.0044	0.0000	0.0000
$N(0, 1)$	30	0.6946	0.0041	0.0001	0.0000
$U(0, 3.464)$	30	0.6998	0.0044	-0.0001	0.0000
$N(0, 2^2)$	30	0.6948	0.0041	0.0002	0.0001
$U(0, 6.928)$	30	0.6995	0.0044	-0.0005	0.0000

Table B.6: Results for naive fits: exponential data: rr=3 : (n=500)

n=500, $\beta_{true} = 1.099 = \log 3$,seed=4					
normal errors: mean $(\beta_{sim}) = 1.1047$, var $(\beta_{sim}) = 0.0086$					
uniform errors: mean $(\beta_{sim}) = 1.0975$, var $(\beta_{sim}) = 0.0094$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0, (\frac{1}{30})^2)$	3	1.1046	0.0086	-0.0001	0.0000
$U(0, 0.115)$	3	1.0973	0.0094	-0.0003	0.0000
$N(0, (\frac{1}{2})^2)$	3	1.1001	0.0085	-0.0046	0.0006
$U(0, 1.732)$	3	1.0890	0.0096	-0.0085	0.0002
$N(0, 1)$	3	1.0837	0.0091	-0.0209	0.0013
$U(0, 3.464)$	3	1.0684	0.0098	-0.0291	0.0006
$N(0, 2^2)$	3	1.0241	0.0098	-0.0805	0.0031
$U(0, 6.928)$	3	1.0041	0.0102	-0.0934	0.0020
$N(0, (\frac{1}{30})^2)$	12	1.1052	0.0087	0.0006	0.0000
$U(0, 0.115)$	12	1.0975	0.0094	-0.0001	0.0000
$N(0, (\frac{1}{2})^2)$	12	1.1042	0.0086	-0.0005	0.0001
$U(0, 1.732)$	12	1.0966	0.0094	-0.0009	0.0000
$N(0, 1)$	12	1.1021	0.0084	-0.0026	0.0003
$U(0, 3.464)$	12	1.0949	0.0095	-0.0026	0.0000
$N(0, 2^2)$	12	1.1001	0.0085	-0.0046	0.0006
$U(0, 6.928)$	12	1.0890	0.0096	-0.0085	0.0002
$N(0, (\frac{1}{30})^2)$	30	1.1050	0.0086	0.0004	0.0000
$U(0, 0.115)$	30	1.0974	0.0094	-0.0001	0.0000
$N(0, (\frac{1}{2})^2)$	30	1.1042	0.0087	-0.0004	0.0001
$U(0, 1.732)$	30	1.0972	0.0094	-0.0004	0.0000
$N(0, 1)$	30	1.1042	0.0087	-0.0004	0.0001
$U(0, 3.464)$	30	1.0968	0.0095	-0.0007	0.0000
$N(0, 2^2)$	30	1.1031	0.0084	-0.0016	0.0002
$U(0, 6.928)$	30	1.0957	0.0095	-0.0019	0.0000

Table B.7: Results for naive fits: Weibull data : $\rho = 0.5$: rr=2 : (n=500)

n=500, $\beta_{true} = 0.693 = \log 2$,seed=4					
normal errors: mean $(\beta_{sim}) = 0.6973$, var $(\beta_{sim}) = 0.0076$					
uniform errors: mean $(\beta_{sim}) = 0.6908$, var $(\beta_{sim}) = 0.0084$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0, (\frac{1}{30})^2)$	3	0.6981	0.0075	0.0008	0.0002
$U(0, 0.115)$	3	0.6907	0.0084	-0.0001	0.0000
$N(0, (\frac{1}{2})^2)$	3	0.7021	0.0083	0.0048	0.0009
$U(0, 1.732)$	3	0.6876	0.0085	-0.0032	0.0001
$N(0, 1)$	3	0.7037	0.0090	0.0063	0.0014
$U(0, 3.464)$	3	0.6837	0.0086	-0.0071	0.0002
$N(0, 2^2)$	3	0.7026	0.0094	0.0053	0.0020
$U(0, 6.928)$	3	0.6739	0.0088	-0.0169	0.0004
$N(0, (\frac{1}{30})^2)$	12	0.6967	0.0075	-0.0006	0.0001
$U(0, 0.115)$	12	0.6909	0.0084	0.0001	0.0000
$N(0, (\frac{1}{2})^2)$	12	0.6994	0.0079	0.0021	0.0005
$U(0, 1.732)$	12	0.6903	0.0084	-0.0005	0.0000
$N(0, 1)$	12	0.7002	0.0081	0.0029	0.0007
$U(0, 3.464)$	12	0.6894	0.0084	-0.0014	0.0000
$N(0, 2^2)$	12	0.7021	0.0083	0.0048	0.0009
$U(0, 6.928)$	12	0.6894	0.0084	-0.0014	0.0000
$N(0, (\frac{1}{30})^2)$	30	0.6972	0.0076	-0.0001	0.0001
$U(0, 0.115)$	30	0.6908	0.0084	0.0000	0.0000
$N(0, (\frac{1}{2})^2)$	30	0.6989	0.0076	0.0016	0.0003
$U(0, 1.732)$	30	0.6906	0.0084	0.0000	0.0000
$N(0, 1)$	30	0.6993	0.0079	0.0002	0.0004
$U(0, 3.464)$	30	0.6904	0.0084	-0.0004	0.0000
$N(0, 2^2)$	30	0.7001	0.0082	0.0028	0.0006
$U(0, 6.928)$	30	0.6897	0.0084	-0.0011	0.0000

Table B.8: Results for naive fits: Weibull data : $\rho = 1.5$: rr=2 : (n=500)

n=500, $\beta_{true} = 0.693 = \log 2$,seed=4					
normal errors: mean $(\beta_{sim}) = 0.6973$, var $(\beta_{sim}) = 0.0074$					
uniform errors: mean $(\beta_{sim}) = 0.6908$, var $(\beta_{sim}) = 0.0084$					
Error	median surv	mean (β_{obs})	var (β_{obs})	mean (bias)	var (bias)
$N(0, (\frac{1}{30})^2)$	3	0.6972	0.0074	-0.0002	0.0005
$U(0, 0.115)$	3	0.6906	0.0084	-0.0002	0.0000
$N(0, (\frac{1}{2})^2)$	3	0.6825	0.0074	-0.0149	0.0005
$U(0, 1.732)$	3	0.6778	0.0087	-0.0130	0.0004
$N(0, 1)$	3	0.6469	0.0078	-0.0504	0.0017
$U(0, 3.464)$	3	0.6448	0.0090	-0.0460	0.0013
$N(0, 2^2)$	3	0.5444	0.0087	-0.1529	0.0051
$U(0, 6.928)$	3	0.5485	0.0093	-0.1423	0.0034
$N(0, (\frac{1}{30})^2)$	12	0.6973	0.0076	0.0000	0.0000
$U(0, 0.115)$	12	0.6908	0.0084	0.0000	0.0000
$N(0, (\frac{1}{2})^2)$	12	0.6959	0.0076	-0.0014	0.0001
$U(0, 1.732)$	12	0.6899	0.0085	-0.0009	0.0000
$N(0, 1)$	12	0.6918	0.0076	-0.0055	0.0002
$U(0, 3.464)$	12	0.6871	0.0086	-0.0037	0.0001
$N(0, 2^2)$	12	0.6819	0.0074	-0.0154	0.0005
$U(0, 6.928)$	12	0.6774	0.0087	-0.0134	0.0004
$N(0, (\frac{1}{30})^2)$	30	0.6973	0.0076	0.0000	0.0000
$U(0, 0.115)$	30	0.6908	0.0084	0.0000	0.0000
$N(0, (\frac{1}{2})^2)$	30	0.6971	0.0076	-0.0002	0.0000
$U(0, 1.732)$	30	0.6905	0.0084	-0.0003	0.0000
$N(0, 1)$	30	0.6965	0.0076	-0.0008	0.0000
$U(0, 3.464)$	30	0.6901	0.0084	-0.0007	0.0000
$N(0, 2^2)$	30	0.6941	0.0076	-0.0033	0.0001
$U(0, 6.928)$	30	0.6885	0.0085	-0.0023	0.0001

Appendix C

C.1 Estimation in the Cox Proportional Hazards Model

Using the notation of Nakamura consider estimation for the Cox model with one covariate:

Cox model : $\lambda(t|x) = \lambda_0(t)\exp(\beta x)$

For each $j \in D$, the set of deaths with R being the risk set at each j :

$S_j(\beta, X) = \Sigma_{i \in R}\exp(\beta x_i)$

$S_j'(\beta, X) = \Sigma_{i \in R}(x_i \exp(\beta x_i))$

$S_j''(\beta, X) = \Sigma_{i \in R}(x_i^2 \exp(\beta x_i))$

The contributions at each $j \in D$ are:

likelihood : $L_j(\beta, X) = \frac{\exp \beta x_j}{S_j(\beta, X)}$

Log-likelihood : $l_j(\beta, X) = \beta x_j - \log S_j(\beta, X)$

Score : $U_j(\beta, X) = x(j) - \frac{S_j'(\beta, X)}{S_j(\beta, X)}$

Information : $IN_j(\beta, X) = \frac{[S_j(\beta, X)S_j''(\beta, X) - (S_j'(\beta, X))^2)]}{[S_j(\beta, X)]^2}$

i.e.

likelihood : $L = \Pi_{j \in D} L_j(\beta, X)$

Log-likelihood : $l = \Sigma_{j \in D} l_j(\beta, X)$

Score : $U = \Sigma_{j \in D} U_j(\beta, X)$

Information : $IN = \Sigma_{j \in D} IN_j(\beta, X)$

C.2 Estimation for the Cox Model Using the C_{ij} Correction Matrix

The same idea is applied to estimation for one covariate for the new model:

For each $j \in D$, the set of deaths with R being the risk set at each j :

$S_j(\beta, X) = \Sigma_{i=1}^{n} C_{ij} \exp(\beta x_i)$

$S_j'(\beta, X) = \Sigma_{i=1}^{n} C_{ij} (x_i \exp(\beta x_i))$

$S_j''(\beta, X) = \Sigma_{i=1}^{n} C_{ij} (x_i^2 \exp(\beta x_i))$

The contributions at each $j \in D$ are:

likelihood : $L_j(\beta, X) = \frac{\exp \beta x_j}{S_j(\beta, X)}$

Log-likelihood : $l_j(\beta, X) = \beta x_j - \log S_i(\beta, X)$

Score : $U_j(\beta, X) = x_j - \frac{S_j'(\beta, X)}{S_j(\beta, X)}$

Information : $IN_j(\beta, X) = \frac{[S_j(\beta, X) S_j''(\beta, X) - (S_j'(\beta, X))^2)]}{[S_j(\beta, X)]^2}$

i.e.

likelihood : $L = \Pi_{j \in D} L_j(\beta, X) =$

$$\Pi_{j\in D}\frac{\exp \beta x_j}{\Sigma_{i=1}^{n} C_{ij} \exp(\beta x_i)} \tag{C.1}$$

Log-likelihood : $l = \Sigma_{j\in D} l_j(\beta, X) =$

$$\Sigma_{j\in D}[\beta x_j - \log \Sigma_{i=1}^{n} C_{ij} \exp(\beta x_i)] \tag{C.2}$$

Score : $U = \Sigma_{j\in D} u_j(\beta, X) =$

$$\Sigma_{j\in D}[x_j - \frac{\Sigma_{i=1}^{n} C_{ij} x_i \exp(\beta x_i)}{\Sigma_{j=1}^{n} C_{ij} \exp(\beta x_i)}] \tag{C.3}$$

Information : $IN = \Sigma_{j\in D} IN_j(\beta, X)$

$$\Sigma_{j\in D}[\frac{\Sigma_{i=1}^{n} C_{ij} \exp(\beta x_i) \Sigma_{i=1}^{n} C_{ij} (x_i)^2 \exp(\beta x_i) - (\Sigma_{i=1}^{n} C_{ij} x_i \exp(\beta x_i))^2}{(\Sigma_{i=1}^{n} C_{ij} \exp(\beta x_i))^2}] \tag{C.4}$$

Appendix D

D.1 Full Results of Simulation Studies - Naive and Corrected Fits

The likelihood incorporating the C matrix to weight risk sets via a measurement model was tested against the simulated data when n=50,n=100 and n=200 for both the uniform and normal error models. Larger error standard deviations of 4,6 and 8 (where the naive fits demonstrate considerable attenuation) were also considered for median survival 3. The column "na" represents the number of times the C matrix encountered zero risk (due to rounding and the approximations involved) and thus these cases were excluded.

Table D.1: Naive and corrected fits: exponential data: rr=2 : (n=50)

n=50, $\beta_{true} = 0.693 = \log 2$,seed=1										
normal errors: mean $(\beta_{sim}) = 0.7221$, var $(\beta_{sim}) = 0.1063$										
uniform errors: mean $(\beta_{sim}) = 0.7020$, var $(\beta_{sim}) = 0.0784$										
Err	ms	m.β_{obs}	v.β_{obs}	m.b.$_{obs}$	v.b.$_{obs}$	m.β_{new}	v.β_{new}	m.b.$_{new}$	v.b.$_{new}$	na
$N1$	3	0.7234	0.1053	0.0013	0.0004	0.7222	0.1050	0.0001	0.0004	0
$U1$	3	0.7021	0.0785	0.0001	0.0001	0.7013	0.0782	-0.0007	0.0001	0
$N2$	3	0.7186	0.1096	-0.0035	0.0073	0.7115	0.1020	-0.0107	0.0025	0
$U2$	3	0.7015	0.0821	-0.0005	0.0037	0.6986	0.0813	-0.0034	0.0029	0
$N3$	3	0.7157	0.1083	-0.0065	0.0175	0.7114	0.1045	-0.0107	0.0079	0
$U3$	3	0.6951	0.0850	-0.0069	0.0091	0.6968	0.0868	-0.0052	0.0091	0
$N4$	3	0.6781	0.1005	-0.0440	0.0381	0.7226	0.1165	0.0010	0.0287	1
$U4$	3	0.6567	0.0864	-0.0453	0.0235	0.6943	0.1021	-0.0077	0.0294	0
$N5$	3	0.5953	0.1130	-0.1269	0.0974	0.7160	0.1401	-0.0080	0.0938	2
$U5$	3	0.5633	0.0865	-0.1387	0.0535	0.6905	0.1178	-0.0138	0.0701	2
$N6$	3	0.4295	0.1365	-0.2638	0.1353	0.6608	0.1411	-0.0655	0.1158	33
$U6$	3	0.4778	0.0844	-0.2242	0.0752	0.6683	0.1215	-0.0367	0.1041	24
$N7$	3	0.3581	0.1734	-0.3522	0.1865	0.6239	0.1698	-0.1470	0.1602	73
$U7$	3	0.4079	0.0818	-0.2941	0.0890	0.6555	0.1153	-0.0530	0.1114	57
$N1$	12	0.7223	0.1056	0.0002	0.0001	0.7224	0.1055	0.0003	0.0000	0
$U1$	12	0.7022	0.0786	0.0002	0.0000	0.7018	0.0784	-0.0002	0.0000	0
$N2$	12	0.7246	0.1038	0.0024	0.0016	0.7192	0.1034	-0.0029	0.0004	0
$U2$	12	0.7007	0.0782	-0.0013	0.0005	0.6995	0.0784	-0.0025	0.0004	0
$N3$	12	0.7224	0.1056	0.0003	0.0038	0.7159	0.1028	-0.0062	0.0010	0
$U3$	12	0.7048	0.0806	0.0028	0.0014	0.6998	0.0796	-0.0022	0.0011	0
$N4$	12	0.7186	0.1096	-0.0035	0.0073	0.7115	0.1020	-0.0107	0.0025	0
$U4$	12	0.7015	0.0821	-0.0005	0.0037	0.6986	0.0813	-0.0034	0.0029	0
$N1$	30	0.7216	0.1060	-0.0005	0.0000	0.7223	0.1057	0.0002	0.0000	0
$U1$	30	0.7020	0.0783	0.0000	0.0000	0.7019	0.0784	-0.0001	0.0000	0
$N2$	30	0.7224	0.1051	0.0002	0.0005	0.7217	0.1047	-0.0004	0.0001	0
$U2$	30	0.7011	0.0784	-0.0009	0.0002	0.7007	0.0781	-0.0013	0.0001	0
$N3$	30	0.7245	0.1038	0.0023	0.0013	0.7201	0.1037	-0.0020	0.0003	0
$U3$	30	0.7008	0.0782	-0.0012	0.0004	0.6995	0.0783	-0.0025	0.0003	0
$N4$	30	0.7209	0.1038	-0.0012	0.0028	0.7122	0.1030	-0.0050	0.0007	0
$U4$	30	0.7031	0.0788	0.0011	0.0010	0.6998	0.0792	-0.0022	0.0008	0

key: ms = median survival (true data), m. = mean, v. = variance
m.b. = mean bias, v.b. = variance of bias, na = missing values

$N1 : N(0, \frac{1}{30}^2)$,$N2 : N(0, \frac{1}{2}^2)$, $N3 : N(0, 1^2)$,
$N4 : N(0, 2^2)$,$N5 : N(0, 4^2)$,$N6 : N(0, 6^2)$,$N7 : N(0, 8^2)$
$U1 : U(0, 0.115)$,$U2 : U(0, 1.732)$,$U3 : U(0, 3.464)$,$U4 : U(0, 6.928)$,
$U5 : U(0, 13.856)$,$U6 : U(0, 20.785)$,$U7 : U(0, 27.713)$

Table D.2: Naive and corrected fits: exponential data: rr=2 : (n=100)

n=100, $\beta_{true} = 0.693 = \log 2$,seed=2										
normal errors: mean $(\beta_{sim}) = 0.7072$, var $(\beta_{sim}) = 0.0435$										
uniform errors: mean $(\beta_{sim}) = 0.7113$, var $(\beta_{sim}) = 0.0585$										
Err	ms	m.β_{obs}	v.β_{obs}	m.b.$_{obs}$	v.b.$_{obs}$	m.β_{new}	v.β_{new}	m.b.$_{new}$	v.b.$_{new}$	na
$N1$	3	0.7073	0.0444	0.0001	0.0002	0.7071	0.0435	-0.0002	0.0000	0
$U1$	3	0.7117	0.0585	0.0005	0.0000	0.7112	0.0584	-0.0001	0.0000	0
$N2$	3	0.7126	0.0485	0.0054	0.0031	0.7029	0.0437	-0.0044	0.0009	0
$U2$	3	0.7058	0.0587	-0.0054	0.0008	0.7071	0.0593	-0.0042	0.0008	0
$N3$	3	0.6981	0.0494	-0.0091	0.0074	0.7042	0.0457	-0.0030	0.0027	0
$U3$	3	0.6911	0.0582	-0.0202	0.0028	0.7043	0.0621	-0.0070	0.0027	0
$N4$	3	0.6705	0.0542	-0.0368	0.0149	0.7084	0.0523	0.0011	0.0094	0
$U4$	3	0.6454	0.0565	-0.0658	0.0088	0.7003	0.0685	-0.0109	0.0098	0
$N5$	3	0.5645	0.0611	-0.1428	0.0322	0.6957	0.0687	-0.0149	0.0371	5
$U5$	3	0.5512	0.0442	-0.1491	0.0199	0.6975	0.0659	-0.0073	0.0262	4
$N6$	3	0.4625	0.0650	-0.2447	0.0539	0.6705	0.0846	-0.0419	0.0726	36
$U6$	3	0.4618	0.0409	-0.2385	0.0308	0.6845	0.0719	-0.0268	0.0440	26
$N7$	3	0.3824	0.0687	-0.3248	0.0681	0.6261	0.0995	-0.0910	0.1013	78
$U7$	3	0.3894	0.0384	-0.3109	0.0386	0.6619	0.0740	-0.0637	0.0511	70
$N1$	12	0.7070	0.0439	-0.0002	0.0001	0.7071	0.0435	-0.0001	0.0000	0
$U1$	12	0.7114	0.0585	0.0001	0.0000	0.7114	0.0585	0.0001	0.0000	0
$N2$	12	0.7095	0.0449	0.0022	0.0007	0.7065	0.0435	-0.0008	0.0001	0
$U2$	12	0.7108	0.0588	0.0004	0.0001	0.7102	0.0584	-0.0001	0.0001	0
$N3$	12	0.7083	0.0450	0.0011	0.0013	0.7050	0.0434	-0.0023	0.0003	0
$U3$	12	0.7108	0.0590	-0.0005	0.0003	0.7090	0.0585	-0.0022	0.0003	0
$N4$	12	0.7126	0.0485	0.0054	0.0031	0.7029	0.0437	-0.0044	0.0009	0
$U4$	12	0.7058	0.0587	-0.0054	0.0008	0.7071	0.0593	-0.0042	0.0008	0
$N1$	30	0.7073	0.0438	0.0001	0.0000	0.7072	0.0435	0.0000	0.0000	0
$U1$	30	0.7113	0.0585	0.0001	0.0000	0.7114	0.0585	0.0001	0.0000	0
$N2$	30	0.7082	0.0449	0.0010	0.0003	0.7070	0.0435	-0.0002	0.0000	0
$U2$	30	0.7116	0.0583	0.0003	0.0000	0.7109	0.0584	-0.0003	0.0000	0
$N3$	30	0.7087	0.0450	0.0015	0.0005	0.7066	0.0436	-0.0006	0.0001	0
$U3$	30	0.7114	0.0584	0.0001	0.0001	0.7105	0.0584	-0.0008	0.0001	0
$N4$	30	0.7072	0.0455	-0.0001	0.0120	0.7056	0.0434	-0.0016	0.0002	0
$U4$	30	0.7110	0.0589	-0.0003	0.0002	0.7095	0.0584	-0.0017	0.0002	0

key: ms = median survival (true data), m. = mean, v. = variance
m.b. = mean bias, v.b. = variance of bias, na = missing values

$N1 : N(0, \frac{1}{30}^2)$,$N2 : N(0, \frac{1}{2}^2)$, $N3 : N(0, 1^2)$,
$N4 : N(0, 2^2)$,$N5 : N(0, 4^2)$,$N6 : N(0, 6^2)$,$N7 : N(0, 8^2)$
$U1 : U(0, 0.115)$,$U2 : U(0, 1.732)$,$U3 : U(0, 3.464)$,$U4 : U(0, 6.928)$,
$U5 : U(0, 13.856)$,$U6 : U(0, 20.785)$,$U7 : U(0, 27.713)$

Table D.3: Naive and corrected fits: exponential data: rr=2 : (n=200)

n=200, $\beta_{true} = 0.693 = \log 2$,seed=3										
normal errors: mean $(\beta_{sim}) = 0.6888$, var $(\beta_{sim}) = 0.0240$										
uniform errors: mean $(\beta_{sim}) = 0.7189$, var $(\beta_{sim}) = 0.0222$										
Err	ms	m.β_{obs}	v.β_{obs}	m.b.$_{obs}$	v.b.$_{obs}$	m.β_{new}	v.β_{new}	m.b.$_{new}$	v.b.$_{new}$	na
$N1$	3	0.6890	0.0245	0.0002	0.0001	0.6884	0.0241	-0.0004	0.0000	0
$U1$	3	0.7189	0.0222	0.0000	0.0000	0.7188	0.0223	-0.0001	0.0000	0
$N2$	3	0.6860	0.0260	-0.0028	0.0014	0.6880	0.0243	-0.0009	0.0004	0
$U2$	3	0.7120	0.0218	-0.0069	0.0004	0.7161	0.0222	-0.0028	0.0003	0
$N3$	3	0.6747	0.0259	-0.0141	0.0025	0.6911	0.0253	0.0023	0.0014	0
$U3$	3	0.6995	0.0215	-0.0194	0.0012	0.7154	0.0226	-0.0035	0.0011	0
$N4$	3	0.6452	0.0271	-0.0436	0.0066	0.7027	0.0297	0.0138	0.0054	0
$U4$	3	0.6574	0.0202	-0.0615	0.0040	0.7127	0.0237	-0.0063	0.0038	0
$N5$	3	0.5598	0.0344	-0.1406	0.0180	0.7149	0.0397	0.0116	0.0179	5
$U5$	3	0.5437	0.0193	-0.1580	0.0104	0.6802	0.0257	-0.0226	0.0120	1
$N6$	3	0.4566	0.0354	-0.2438	0.0265	0.6958	0.0491	-0.0126	0.0341	31
$U6$	3	0.4552	0.0189	-0.2464	0.0158	0.6683	0.0289	-0.0312	0.0198	18
$N7$	3	0.3685	0.0349	-0.3318	0.0322	0.6470	0.0447	-0.0713	0.0409	77
$U7$	3	0.3859	0.0186	-0.3157	0.0201	0.6620	0.0354	-0.0464	0.0029	65
$N1$	12	0.6886	0.0243	-0.0002	0.0000	0.6888	0.0241	-0.0001	0.0000	0
$U1$	12	0.7189	0.0223	0.0000	0.0000	0.7189	0.0223	0.0000	0.0000	0
$N2$	12	0.6873	0.0250	-0.0016	0.0003	0.6879	0.0240	-0.0010	0.0000	0
$U2$	12	0.7181	0.0221	-0.0009	0.0000	0.7181	0.0222	-0.0008	0.0000	0
$N3$	12	0.6884	0.0254	-0.0004	0.0006	0.6876	0.0240	-0.0012	0.0001	0
$U3$	12	0.7168	0.0220	-0.0022	0.0001	0.7171	0.0221	-0.0018	0.0001	0
$N4$	12	0.6860	0.0260	-0.0028	0.0014	0.6880	0.0243	-0.0009	0.0004	0
$U4$	12	0.7120	0.0218	-0.0069	0.0004	0.7161	0.0222	-0.0028	0.0003	0
$N1$	30	0.6888	0.0242	-0.0001	0.0000	0.6888	0.0241	0.0000	0.0000	0
$U1$	30	0.7189	0.0223	0.0000	0.0000	0.7189	0.0223	0.0000	0.0000	0
$N2$	30	0.6890	0.0249	0.0001	0.0001	0.6883	0.0240	-0.0006	0.0000	0
$U2$	30	0.7187	0.0223	-0.0002	0.0000	0.7187	0.0223	-0.0002	0.0000	0
$N3$	30	0.6877	0.0250	-0.0012	0.0003	0.6879	0.0240	-0.0009	0.0000	0
$U3$	30	0.7186	0.0222	-0.0004	0.0000	0.7183	0.0222	-0.0006	0.0000	0
$N4$	30	0.6887	0.0252	-0.0001	0.0006	0.6877	0.0240	-0.0012	0.0001	0
$U4$	30	0.7176	0.0220	-0.0014	0.0001	0.7175	0.0221	-0.0015	0.0001	0

key: ms = median survival (true data), m. = mean, v. = variance
m.b. = mean bias, v.b. = variance of bias, na = missing values

$N1 : N(0, \frac{1}{30}^2)$,$N2 : N(0, \frac{1}{2}^2)$, $N3 : N(0, 1^2)$,
$N4 : N(0, 2^2)$,$N5 : N(0, 4^2)$,$N6 : N(0, 6^2)$,$N7 : N(0, 8^2)$
$U1 : U(0, 0.115)$,$U2 : U(0, 1.732)$,$U3 : U(0, 3.464)$,$U4 : U(0, 6.928)$,
$U5 : U(0, 13.856)$,$U6 : U(0, 20.785)$,$U7 : U(0, 27.713)$

Appendix E

E.1 S-Plus Code for P calculation for Simulated Data

```
#calculate p for simulated data
#requires survival time, mean and s.d. of error
simcalcp <- function(dataset, mu, sdev)
{
len <- length(dataset)
p <- array(NA, dim = c(len, len))
for(i in 1:(len - 1)) {
A <- max(0, dataset[i] - mu - (2 * sdev))
B <- max(0.001, dataset[i] - mu + (2 * sdev))
for(j in i:len) {
C <- max(0, dataset[j] - mu - (2 * sdev))
D <- max(0.001, dataset[j] - mu + (2 * sdev))
p[i, j] <- unifprobs(A, B, C, D)
p[j, i] <- 1 - p[i, j]
}
p[i, i] <- -1
}
p[len, len] <- -1
return(p)
}
```

E.2 S-Plus Code for P Calculation for Lung Cancer Sample

```
#calcualate p for 1993 sample data
#
nrows <- 500
ncols <- length(lung93samp$dead[lung93samp$dead == 1])
lungpmx <- array(NA,dim=c(nrows,ncols))
for(i in 1:nrows){
cat(i," ")
counter <- 0
for(j in 1:nrows){
if (lung93samp$dead[j] == 1){
counter <- counter + 1
#two validation times - status 0
if ((lung93samp$val[i] == 0) && (lung93samp$val[j] == 0)){
A <- lung93samp$surv[i]
B <- lung93samp$surv[i] + 1
C <- lung93samp$surv[j]
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
#one validation status 0, 1 non-validation status 1
#status 1 true|obs ~ U(max(0,obs-52),obs+1)
if ((lung93samp$val[i] == 0) && (lung93samp$val[j] == 1)){
A <- lung93samp$surv[i]
B <- lung93samp$surv[i] + 1
C <- max(0,lung93samp$surv[j] - 52)
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
if ((lung93samp$val[i] == 1) && (lung93samp$val[j] == 0)){
A <- max(0,lung93samp$surv[i] - 52)
B <- lung93samp$surv[i] + 1
C <- lung93samp$surv[j]
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
#one validation status 0, 1 non-validation status 2
#status 2 true|obs ~ U(max(0,obs-64),obs+1)
if ((lung93samp$val[i] == 0) && (lung93samp$val[j] == 2)){
A <- lung93samp$surv[i]
```

```
B <- lung93samp$surv[i] + 1
C <- max(0,lung93samp$surv[j] - 64)
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
if ((lung93samp$val[i] == 2) && (lung93samp$val[j] == 0)){
A <- max(0,lung93samp$surv[i] - 64)
B <- lung93samp$surv[i] + 1
C <- lung93samp$surv[j]
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
#one validation status 1, 1 non-validation status 2
#status 1 true|obs ~ U(max(0,obs-52),obs+1)
#status 2 true|obs ~ U(max(0,obs-64),obs+1)
if ((lung93samp$val[i] == 1) && (lung93samp$val[j] == 2)){
A <- max(0,lung93samp$surv[i] - 52)
B <- lung93samp$surv[i] + 1
C <- max(0,lung93samp$surv[j] - 64)
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
if ((lung93samp$val[i] == 2) && (lung93samp$val[j] == 1)){
A <- max(0,lung93samp$surv[i] - 64)
B <- lung93samp$surv[i] + 1
C <- max(0,lung93samp$surv[j] -52)
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
#both non-validation status 1
#status 1 true|obs ~ U(max(0,obs-52),obs+1)
if ((lung93samp$val[i] == 1) && (lung93samp$val[j] == 1)){
A <- max(0,lung93samp$surv[i] - 52)
B <- lung93samp$surv[i] + 1
C <- max(0,lung93samp$surv[j] - 52)
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
#both non-validation status 2
#status 2 true|obs ~ U(max(0,obs-64),obs+1)
if ((lung93samp$val[i] == 2) && (lung93samp$val[j] == 2)){
A <- max(0,lung93samp$surv[i] - 64)
B <- lung93samp$surv[i] + 1
```

```
C <- max(0,lung93samp$surv[j] - 64)
D <- lung93samp$surv[j] + 1
lungpmx[i,counter] <- unifprobs(A, B, C, D)
}
}
}
if (lung93samp$dead[i] == 1) lungpmx[i,sum(lung93samp$dead[1:i])] <- -1
}



#censored times - assume if can be at risk is definetly at risk
counter <- 0
for(i in 1:nrows){
if (lung93samp$dead[i] == 0) {
counter <- counter + 1
cat(counter," ")
tempmx <- lungpmx[i,]
tempmx[tempmx > 0] <- 1
lungpmx[i,] <- tempmx
}
}
```

E.3 S-Plus Code for Calculation of P(U(a,b)>U(c,d))

```
#function to calculate P(U(A,B) > U(C,D))
unifprobs <- function(A, B, C, D)
{
if((B - A) > (D - C)) {
k1 <- c(A, B)
k2 <- c(C, D)
}
else {
k1 <- c(C, D)
k2 <- c(A, B)
}
if((k1[1] - k2[2]) >= 0){temp <- 0}
if((k1[1] - k2[2]) < 0 && (k1[1] - k2[1]) >= 0) {
temp <- ((k1[1] - k2[2])^2)/(2 * (k1[2] - k1[1]) * (k2[2] - k2[1]))
}
if((k1[1] - k2[1]) < 0 && (k1[2] - k2[2]) >= 0) {
temp <- ((k2[2] - k2[1]) - (2 * (k1[1] - k2[1])))/(2 * (k1[2] - k1[1]))
}
if((k1[2] - k2[2]) < 0 && (k1[2] - k2[1]) >= 0) {
temp <- (((2 * k1[1] * k2[1]) + (2 * k1[2] * k2[2]) - (2 * k1[1
] * k2[2]) - (k1[2]^2) - (k2[1]^2))/(2 * (k1[2] - k1[1]
) * (k2[2] - k2[1])))
}
if((k1[2] - k2[1]) < 0){temp <- 1}
if((B - A) > (D - C)){temp <- 1 - temp}
temp <- round(temp, 2)
return(temp)
}
```

E.4 S-Plus Code for Calculation of C Matrix

```
#calculate C - need P matrix and vector of censoring indication
calculatec <- function(probs, dead)
{
#define C array
k <- length(probs[1, ])
riskmx <- array(NA, dim = c(length(probs[, 1]), k))
#calulate ngreater,ncomm,commset
for(i in 1:length(probs[, 1])) {
currentrow <- probs[i, ]
ngreater <- length(currentrow[currentrow == 1])
ncomm <- length(currentrow[currentrow < 1]) - length(currentrow[
currentrow <= 0])
commset <- currentrow[currentrow < 1]
commset <- commset[commset > 0]
#perform steps 3,4,5 & 6 of the algorithm if there is communication
if(ncomm > 0) {
risklik <- rep(0, k)
ave <- mean(commset)
temp <- 1 - pbinom(seq(-1, (ncomm - 1)), ncomm, ave)
if(ngreater > 0) {risklik[1:ngreater] <- 1 }
risklik[(ngreater + 1):(ngreater + ncomm + 1)] <- temp
}
#perform steps 5,6 if there is no communication
if(ncomm == 0) {
risklik <- rep(0, k)
if(dead[i] == 0) {risklik[1:ngreater] <- 1}
if(dead[i] == 1) {risklik[1:(ngreater + 1) <- 1}
}
riskmx[i, ] <- round(risklik, 2)
}
return(riskmx)
}
```

E.5 Code for Creating and Fitting Simulated Data with Normal or Uniform Errors

E.5.1 Normal Errors

```
# code to create simulated data with normal errors
# n - dataset size, p = no. of repeats
# sd error standard deviation
# bline is basline shape parameter value
# rho is weibull scale parameter
# rr is relative risk
# seedy is seed
simnormcreation <- function(n, p, sd, bline, rho=1, rr, seedy)
{
set.seed(seedy)
dataname <- NA
dataname$seedy <- seedy
dataname <- dataname[2]
dataname$sd <- sd
dataname$bline <- bline
dataname$rr <- rr
dataname$cov <- c(rep(0, n/2), rep(1, n/2))
dataname$truecox <- rep(NA, p)
dataname$truecoxvar <- rep(NA, p)
dataname$obsbeta <- rep(NA, p)
dataname$obsvarbeta <- rep(NA, p)
dataname$newbeta <- rep(NA, p)
dataname$newvarbeta <- rep(NA, p)
for(i in 1:p) {
cat("Fit: ", i, "\n")
# if rho not equal 1 Weibull data
if (rho != 1){surv <- c(rweibull(n/2, rho, ((1/bline)^(1/rho))),
                        rweibull(n/2, rho, ((1/(rr * bline))^(1/rho))
                        ))}
# if rho equals 1 Exponential data
if (rho == 1){surv <- c(rexp(n/2, bline), rexp(n/2, (bline * rr)))}
# true cox fits
truecox <- coxph(Surv(surv, rep(1, n)) ~ dataname$cov)
dataname$truecox[i] <- round(truecox$coefficients[1], 3)
dataname$truecoxvar[i] <- round(truecox$var[1, 1], 4)
cat("True beta: ", dataname$truecox[i], " Old var:", dataname$
truecoxvar[i], "\n")
```

```
#observed data creation
tempsurv <- surv + rnorm(n, 0, sd)
#remove negaitve times
posvalsurv <- tempsurv[tempsurv > 0]
posvalcov <- dataname$cov[tempsurv > 0]
#observed Cox fit
posvalcox <- coxph(Surv(posvalsurv, rep(1, length(tempsurv[
tempsurv > 0]))) ~ posvalcov)
dataname$obsbeta[i] <- round(posvalcox$coefficients[1], 3)
dataname$obsvarbeta[i] <- round(posvalcox$var[1, 1], 4)
cat("Old beta: ", dataname$obsbeta[i], " Old var:", dataname$
obsvarbeta[i], "\n")
#new Cox fits - if not desired make into comments using # at start of lines
newfittemp <- onetotalfit(tempsurv[order(tempsurv)], rep(1, n),
dataname$cov[order(tempsurv)], 0, dataname$sd)
dataname$newbeta[i] <- newfittemp$beta
dataname$newvarbeta[i] <- newfittemp$varest
}
return(dataname)
}
```

E.5.2 Uniform Errors

```
# code to create simulated data with uniform errors
# n - dataset size, p = no. of repeats
# upper is upper level (b) of U(0,b) errors
# bline is basline shape parameter value
# rho is weibull scale parameter
# rr is relative risk
# seedy is seed
simunifcreation <- function(n, p, upper, bline, rho=1, rr, seedy)
{
set.seed(seedy)
dataname <- NA
dataname$seedy <- seedy
dataname <- dataname[2]
dataname$upper <- upper
dataname$bline <- bline
dataname$rr <- rr
dataname$cov <- c(rep(0, n/2), rep(1, n/2))
dataname$truecox <- rep(NA, p)
dataname$truecoxvar <- rep(NA, p)
dataname$obsbeta <- rep(NA, p)
dataname$obsvarbeta <- rep(NA, p)
dataname$newbeta <- rep(NA, p)
dataname$newvarbeta <- rep(NA, p)
for(i in 1:p) {
cat("Fit: ", i, "\n")
# if rho not equal 1 Weibull data
if (rho != 1) surv <- c(rweibull(n/2, rho, ((1/bline)^(1/rho))),
                        rweibull(n/2, rho, ((1/(rr * bline))^(1/rho))
                        ))
# if rho equals 1 Exponential data
if (rho == 1) surv <- c(rexp(n/2, bline), rexp(n/2, (bline * rr)))
# true cox fits
truecox <- coxph(Surv(surv, rep(1, n)) ~ dataname$cov)
dataname$truecox[i] <- round(truecox$coefficients[1], 3)
dataname$truecoxvar[i] <- round(truecox$var[1, 1], 4)
cat("True beta: ", dataname$truecox[i], " Old var:", dataname$
truecoxvar[i], "\n")
#observed data creation
 tempsurv <- surv + runif(n, 0, upper)
#observed cox fits
tempcox <- coxph(Surv(tempsurv, rep(1, n)) ~ dataname$cov)
```

```
dataname$obsbeta[i] <- round(tempcox$coefficients[1], 3)
dataname$obsvarbeta[i] <- round(tempcox$var[1, 1], 4)
cat("Old beta: ", dataname$obsbeta[i], " Old var:", dataname$
obsvarbeta[i], "\n")
#new Cox fits - if not desired make into comments using # at start of lines
newfittemp <- onetotalfit(tempsurv[order(tempsurv)], rep(1, n),
dataname$cov[order(tempsurv)], 0, dataname$sd)
dataname$newbeta[i] <- newfittemp$beta
dataname$newvarbeta[i] <- newfittemp$varest
}
return(dataname)
}
```

E.6 Parent Code for Fitting Simulated Data

```
#call to fit new fit for simulated data
#factor is for U(0,b) errors - it deflates errorsd by sqrt(12)/4
#hence 2*(errorsd*factor) = b/2
#the errormean = b/2 hence we have the correct upper and lower levels
onetotalfit<-function(obstimes, dead, cov, errormean, errorsd, factor = 1)
{
if(factor != 1) {errorsd <- (errorsd * factor)}
#calculate p
p <- simcalcp(obstimes, errormean, errorsd)
#calculate C
tempc <- calculatec(p)
#check for encountering zero risk
canido <- cando(tempc)
#If possible do the new fit
if(canido == 1) {onefit <- newtraph(tempc, cov, 0, dead, 5)}
#If not possible return missing values
if(canido == 0) {
onefit <- NA
onefit$beta <- NA
onefit$varbeta <- NA
onefit <- onefit[2:3]
}
list(beta = onefit$beta, varest = onefit$varest)
}


#checks for zero risk in the C matrix
cando <- function(cmx)
{
        temp <- length(cmx[, 1])
#require second half of C elements only
       temp2 <- trunc(temp/2)
        value <- 1
        for(i in temp2:temp) {
                if(sum(cmx[, i]) == 0) {value <- 0}
        }
        return(value)
}
```

E.7 S-Plus Code for Newton - Raphson Procedure for Estimating β

```
#newton raphson iterative process
#requires C matrix, covariate value, starting point, censoring indicator
#and maximum number of iterations
newtraph <- function(cmx, cov, beta, cens, itermax)
{
#treat covariate as continuous
if(is.factor(cov) == T) cov <- as.numeric(cov) - 1
#create initial values
n <- length(cov)
betanew <- rep(NA, itermax + 1)
betanew[1] <- beta
iter <- 0
count <- 0
diff <- 1
#count the number of patients who fail
for(i in 1:n) {
if(cens[i] == 1) {count <- count + 1}
}
#repeat until estimate is found or iterations reach maximum number
while(abs(diff) > 0.0001 && iter < itermax) {
iter <- iter + 1
U <- rep(NA, count)
I <- rep(NA, count)
counter <- 1
for(i in 1:n) {
#calculate score at each failure time
if(cens[i] == 1) {
U[counter] <- calcscore(cmx, cov, betanew[iter], i, counter)
I[counter] <- calcinf(cmx, cov, betanew[iter], counter)
counter <- counter + 1
}
}
#sum individual score contributions
score <- sum(U)
#calculate Jacobian
jacob <- (-1) * sum(I)
invjacob <- solve(jacob)
diff <- (invjacob * score)
```

```
betanew[iter + 1] <- betanew[iter] - diff
cat("iteration", iter, "new estimate", betanew[iter + 1], "\n")
}
#calculate s.e. estimate via Information
inf <- rep(NA, count)
counter <- 1
for(i in 1:n) {
if(cens[i] == 1) {
inf[counter] <- calcinf(cmx, cov, betanew[iter + 1], counter)
counter <- counter + 1
}
}
infor <- sum(inf)
varest <- solve(infor)
beta <- round(betanew[iter + 1], 3)
varest <- round(varest, 4)
cat("Beta estimate :", beta, "  Var estimate :", varest, "\n")
#return values
list(beta = beta, varest = varest)
}
```

The previous routine also calls the following sub-routines :

```
#calculate score equation
calcscore <- function(cmx, data, beta, i, counter)
{
temp1 <- s1(cmx, data, beta, counter)
temp0 <- s0(cmx, data, beta, counter)
temp <- temp1/temp0
U <- data[i]
U <- U - temp
return(U)
}
#calculate information
calcinf <- function(cmx, data, beta, counter)
{
temp0 <- s0(cmx, data, beta, counter)
temp1 <- s1(cmx, data, beta, counter)
temp2 <- s2(cmx, data, beta, counter)
```

```
temp <- (temp0 * temp2)
temp <- temp - (temp1^2)
temp <- temp/(temp0^2)
return(temp)
}

s0 <- function(cmx, data, beta, counter)
{
s0res <- sum(cmx[, counter] * exp(beta * data))
return(s0res)
}

s1 <- function(cmx, data, beta, counter)
{
s1res <- sum(cmx[, counter] * data * exp(beta * data))
return(s1res)
}

s2 <- function(cmx, data, beta, counter)
{
s2res <- sum(cmx[, counter] * data * data * exp(beta * data))
return(s2res)
}
```

E.8 S-Plus Code for Baseline Hazard and Non-Parametric Hazard Estimate for 1 Group

```
#fit the baselinehazard or
#estimates cumulative hazard for 1 group for lung data
#needs survival times,censoring indicator
#if fitting basehazard needs estimate of beta plus covariate values
#set whichtype to 1 non-parametric cumulative hazard estimation
#of single sample
#cmx required also
basehazard<-function(obstimes, dead, betahat, cov, val, whichtype=2, cmx)
{
#estimate expected true survival
#need to change if implementing for other measurement models
tempor <- rep(NA, length(obstimes))
for(i in 1:length(obstimes)) {
if(val[i] == 0){lower <- obstimes[i]}
if(val[i] == 1){lower <- max(0, obstimes[i] - 52)}
if(val[i] == 2){lower <- max(0, obstimes[i] - 64)}
upper <- obstimes[i] + 1
tempor[i] <- (upper + lower)/2
}
#order times and related values
temporder <- order(tempor)
temp <- obstimes[temporder]
tempdead <- dead[temporder]
tempcov <- cov[temporder]
tempor <- tempor[temporder]
#get timepoints of each death
ndead <- sum(tempdead)
whodead <- tempor[tempdead == 1]
pointsoftime <- 1
for(i in 2:ndead) {
#isolate each survival time
if((whodead[i] != whodead[i - 1])){pointsoftime <- pointsoftime + 1}
}
#returns an array with 4 rows
#1st row is death times
#2nd row is no.of deaths at each time
times <- array(0, dim = c(4, pointsoftime))
conts <- array(0, ndead)
```

```
times[1, 1] <- whodead[1]
times[2, 1] <- length(whodead[whodead == whodead[1]])
counter <- 2
for(i in 2:ndead) {
if(whodead[i] != whodead[i - 1]) {
times[1, counter] <- whodead[i]
times[2, counter] <- length(whodead[whodead == whodead[i]])
counter <- counter + 1
}
}
#calculate the hazard contibutions at each deathtime
for(i in 1:ndead) {
if(whichtype == 2){conts[i] <- sum((cmx[, i]) * exp(betahat * cov))}
if(whichtype == 1){conts[i] <- sum(cmx[, i])}
conts[i] <- 1/conts[i]
}
#3rd and 4th row are the individual contribution to the hazard
#and the cumulative hazard at a death time
counter <- 1
for(i in 1:pointsoftime) {
newcounter <- counter + times[2, i]
times[3, i] <- sum(conts[counter:(newcounter - 1)])
times[4, i] <- sum(times[3,  ])
counter <- newcounter
}
return(times)
}
```

References

Bashir, S.A. and Duffy, S.W. (1995). Correction of risk estimates for measurement error in epidemiology. *Methods of Information in Medicine*, **34**, 503–510.

Berkson, J. (1950). Are there two regressions? *JASA*, **45**, 164–180.

Berrino, F., Esteve, J. and Coleman, M.P. (1995). Basic issues in estimating and comparing the survival of cancer patients. IARC EUROCARE study.

Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, **2**, 89–100.

Brewster, D., Crichton, J. and Muir, C. (1994). How accurate are Scottish cancer registration data? *Br. J. Cancer*, **70**, 954–959.

Brewster, D., Muir, C. and Crichton, J. (1995). Registration of lung cancer in Scotland: an assessment of data accuracy based on review of medical records. *Cancer Causes and Control*, **6**, 303–310.

Byar, D.P. and Gail, M.H (1989). Workshop on errors-in-variables, 1987. *Statistics in Medicine*, **8**.

Carpenter, C.L., Morgenstern, H. and London, S.J. (1998). Alcoholic beverage

consumption and lung cancer risk among residents of Los Angeles County. *Journal of Nutrition*, **128,** 694–700.

Carroll, R.J. (1989). Covariance analysis in generalized linear measurement error models. *Statistics in Medicine*, **8,** 1075–1093.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement error in non linear models.* Chapman and Hall.

Chaffey, C.M. and Bowie, C. (1994). Radon and health - an update. *Journal of Public Health Medicine.*

Cheng, Chi-Lun. and Van Ness, J.W. (1994). On estimating linear relationships when both variables are subject to errors. *JRSS* B, **56,** 167–183.

Clayton, D.G. (1991). Models for the analysis of cohort and case control studies with inaccurately measured exposures. In *Statistical Models for Longitudinal Studies of Health* (eds J.H. Dwyer, M. Feinlib, P. Lipsert and H. Hoffmeister), pp. 301–331. OUP.

Collett, D. (1994). *Modelling survival data in medical research*, 1 edn. Chapman and Hall, London.

Cox, D.R. (1972). Regression models and life tables. *JRSS* B, **34,** 187–220.

Cox, D.R. (1975). Partial likelihood. *Biometrika*, **62,** 269–276.

Cox, D.R. and Oakes, D. (1984). *Analysis of survival data.* Chapman and Hall, London.

Doll, R. and Peto, R. (1981). *The causes of cancer.* Oxford Medical Publications.

Efron, B. (1977). The efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*, **72,** 557–565.

Fleming, T.R. and Harrington, D.P. (1991). *Counting processes and survival analysis.* Wiley, New York.

Fuller, W.A. (1987). *Measurement error models.* Wiley.

Fuller, W.A. (1995). Estimation in the presence of measurement errors. *International Statistical Review*, **63,** 121–147.

Green, S.B. and Byar, P.B. (1984). Using observational data from registries to compare treatments : The fallacy of omnimetrics. *Statistics in Medicine*, **3,** 361–370.

Gulliford, M.C., Bell, J., Bourne, H.M. and Petruckevitch, A. (1993). The reliability of cancer registry records. *Br. J. Cancer*, **67,** 819–821.

Hasabelnaby, N.A., Ware, J.H. and Fuller, W.A. (1989). Indoor air pollution and pulmonary performance, investigating errors in exposure measurements. *Statistics in Medicine*, **8,** 1109 - 1126.

Hess, K.R. (1995). Graphical methods for assessing violations of the proportional hazards assumption in Cox regression. *Statistics in Medicine*, **14,** 1707–1723.

Holt, D., McDonald, J.W. and Skinner, C.J. (1991). The effect of measurement error on event history analysis. In *Measurement Errors in Surveys* (eds Biemer, Groves, Lyberg, Mathiowetz and Sudman), pp. 665–685. Wiley.

Huakka, J.K. (1995). Correction for covariate measurement error in generalized linear models - a bootstrap approach. *Biometrics*, **51,** 1127–1132.

Hughes, M.D. (1993). Regression dilution in the proportional hazards model. *Biometrics*, **49,** 1056 - 1066.

Hughes, M.D., Thompson, S.G. and Pocock, S.J. (1995). Optimal sequential screening guidelines for quantitative risk factors measured with error. *JASA*, **90,** 19–26.

Hussey, R.M. and Ashby, D. (1990). Geographical pattern on cancer incidence in Mersey region 1983-1987. *Mersey Regional Cancer Registry.*

Jensen, O.M. and Storm, H.H. (1991). Purposes and uses of cancer registration. In *Cancer registration principles and methods, France* (eds O.M. Jenson, D.M. Parkin, R. Maclennan, C.S. Muir and R.G. Skeet). IARC scientific publications.

Kalbfleisch, J.D. and Prentice, R.L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika*, **60,** 267–278.

Kalbfleisch, J.D. and Prentice, R.L. (1980). *The statistical analysis of failure time data.* Wiley, New York.

Knekt, P., Raitasalo, R., Heliovaara, M., Lehtinen, V., Pukkala, E., Teppo, L., Maatela, J. and Aromaa, A. (1996). Elevated lung cancer risk amongst persons with depressed mood. *Americal Journal of Epidemiology*, **144,** 1096–1103.

Lapham, R. and Waugh, N.R. (1992). An audit of the quality of cancer registration data. *Br. J. Cancer*, **66,** 552–554.

Maclennon, R. (1991). Items of patient information which may be collected by registries. In *Cancer registration principles and methods, France* (eds

O.M. Jenson, D.M. Parkin, R. Maclennan, C.S. Muir and R.G. Skeet). IARC scientific publications.

Marshall, J.R. and Hastrup, J.L. (1996). Mismeasurement and the resonance of strong confounders: Uncorrelated errors. *American Journal of Epidemiology*, **143,** 1069–1078.

Mendilaharsu, M., DeStefani, E., Deneopellegrini, H., Carzoglio, J.C. and Ronco, A. (1998). Consumption of tea and coffee and the risk of lung cancer in cigarette smoking men: a case-control study in Uruguay. *Lung Cancer*, **19,** 101–107.

Moran, P.A.P (1971). Estimating structural and functional relationships. *Journal of Multivariate Analysis*, **1,** 232–255.

MRCR (1990). Obvectives of cancer registration.

Nakamura, T. (1990). Corrected score function for errors-in-variables model: Methodology and application to generalized linear models. *Biometrika*, **77,** 127–37.

Nakamura, T. (1992). Proportional hazards model with covariates subject to measurement error. *Biometrics*, **48,** 829–838.

Nakamura, T. and Akazawa, K. (1994a). Corrected likelihood for the proportional hazards measurement error model and its application. *Environmental Health Perspectives*, **102,** 21–24.

Nakamura, T. and Akazawa, K. (1994b). Computer program for the proportional hazards measurement error model. *Computer Methods and Programs in Biomedicine*, **45,** 202–212.

NHS (1998). Management of lung cancer. *Effective Health Care*, **4**.

Oakes, D. (1981). Survival times: aspects of partial likelihood (with discussion). *International Statistical Review*, **49**, 235–264.

Ohno, Y., Wakai, K., Genka, K., Ohmine, K., Kawamura, T., Tamakoshi, A., Aoki, R., Senda, M., Hayashi, Y., Nagoa, K., Fukuma, S. and Aoki, K. (1995). Tea consumption and lung cancer risk - a case-control study in Okinawa, Japan. *Japanese journal of Cancer Research*, **86**, 1027–1034.

Pepe, M.S., Self, G. and Prentice, R.L. (1989). Further results on covariate measurement error in cohort studies with time to response data. *Statistics in Medicine*, **8**, 1167–1178.

Powell, J. (1991). Purposes and uses of cancer registration. In *Cancer registration principles and methods, France* (eds O.M. Jenson, D.M. Parkin, R. Maclennan, C.S. Muir and R.G. Skeet). IARC scientific publications.

Prentice, R.L. (1982a). Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, **69**, 331–42.

Prentice, R.L. (1982b). Covariate measurement errors in the analysis of case-control and cohort studies. In *Survival Analysis* (eds R. Johnson and J. Crowley), pp. 137–151. IMS lecture notes.

Rice, J.A. (1988). *Mathematical statistics and data analysis.* Wadsworth and Brooks, California.

Richardson, S. and Gilks, W.R. (1993). Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, **12**, 1703–1722.

Rosner, B., Willet, W. and Spigelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within person measurement error. *Statistics in Medicine*, **8**, 1051–1069.

Saneii, S.H. (1997). Measurement error modelling for ordered covariates in epidemiology. Ph.D. Thesis. University of Liverpool.

Satten, G.A. and Longini, I.M. (1996). Markov chains with measurement error: Estimating the "true" course of a marker of the progression of HIV disease. *Applied Statistics*, **45**.

Schafer, D. (1993). Likelihood analysis for probit regression with measurement errors. *Bometrika*, **80**, 899–904.

Skeet, R.G. (1991). Quality and quality control. In *Cancer registration principles and methods, France* (eds O.M. Jenson, D.M. Parkin, R. Maclennan, C.S. Muir and R.G. Skeet). IARC scientific publications.

Spiegelhalter, D.J., Thomas, A., Best, N.G. and Gilks, W.R. (1995). *BUGS:Bayesian inference Using Gibbs Sampling version 0.50.* MRC Biostatistics unit, Cambridge.

Stefanski, L.A. and Carroll, R.J. (1985). Covariate measurement error in logistic regression. *Annals of statistics*, **13**, 1335–1351.

Stefanski, L.A. and Carroll, R.J. (1987). Efficient scores and optimal scores in generalized linear measurement error models. *Biometrika*, **74**, 703–716.

Stephens, D.A. and Dellaportas, P. (1992). Bayesian analysis of generalized linear models with covariate measurement error. In *Bayesian Statistics*, volume 4, pp. 813–820.

Stephens, D.A. and Dellaportas, P. (1995). Bayesian analysis of errors-in-variables regression models. *Biometrics*, **51**, 1085–1095.

Sudman, S. and Bradburn, N.M. (1973). Effects of time and memory factors on response in surveys. *JASA*, **68**, 805–815.

Therneau, T.M. (1994). A package for survival analysis in S. Technical Report. Mayo Clinic.

Thune, I. and Lund, E. (1997). The influence of physical activity on lung cancer risk - A prospective study of 81,156 men and women. *International Journal of Cancer*, **70**, 57–62.

Tosteson, T.D., Stefanski, L.A. and Schafer, D.W. (1989). A measurement error model for binary and ordinal regression. *Statistics in Medicine*, **8**, 1139–1147.

Tsiatis, A.A., DeGruttola, V. and Wulfsohn, M.S. (1995). Modelling the relationship of survival to longitudinal data measured with error. applications to survival and CD4 counts in patients with AIDS. *JASA*, **90**.

UKACR (1994). Cancer registry handbook.

Vessey, M.P. and Gray, M. (1985). *Cancer risks and prevention.* Oxford Medical Publications.

Wagner, G. (1991). History of cancer registration. In *Cancer registration principles and methods, France* (eds O.M. Jenson, D.M. Parkin, R. Maclennan, C.S. Muir and R.G. Skeet). IARC scientific publications.

West, R.R. (1976). Accuracy of cancer registration. *Brit. J. prev. soc. Med.*, **30**, 187–192.

Whittemore, A.S. (1990). Errors-in-variables regression problems in epidemiology. *Contemporary Mathematics*, **112**.

Whittemore, A.S. and Keller, J.B. (1988). Approximations for regression with covariate measurement error. *JASA*, **83**.

Willet, W. (1989). An overview of issues related to non-differential exposure measurement in epidemiologic studies. *Statistics in Medicine*, **8**, 1031–1040.

Williams, C. (1992). *Lung cancer : The facts.* Oxford Medical Publications.

Williams, E.M.I., Somerville, M., Youngson, J.H., Smith, C., White, F.E., Leinster, S.J. and Zabhoor, H.D. (1994). Breast cancer bulletin. *Merseyside and Cheshire Cancer Registry.*

Williams, E.M.I., Youngson, J.H., Ashby, D. and Donnelly, R.J. (1993). Lung cancer bulletin. *Mersey Regional Cancer Registry.*

Youngson, J.H., Ashby, D. and Hussey, R.M. (1991). Incidence of cancer in Mersey region and it's constituent health districts 1983-1987. *Mersey Regional Cancer Registry.*

Youngson, J.H., Ashby, D. and Williams, E.M.I. (1992). Incidence of cancer in Mersey region and it's constituent health districts 1986-1990. *Mersey Regional Cancer Registry.*