Estimating Medical Care Costs:
An Examination under Conditions of Censoring

Maria Raikou

Ph.D. Thesis

Department of Economics
City University, London

London, February 2003

Contents

List of Tables and Figures

Declaration

I hereby grant powers of discretion to the librarian of City University, London to allow this thesis to be copied in whole or in part without further reference to the author. This permission covers only single copies made for the purposes of research subject to the normal conditions of acknowledgement.

Papers

At the time of submission of this thesis a paper consisting of a substantial part of chapter 3 has been published as: Raikou, M., Briggs, A., Gray, A., McGuire, A., 2000. Centre-specific or average unit costs in multi-centre studies? Some theory and simulation. Health Economics 9, 191-198.

A second paper is forthcoming in Pharmacoeconomics based on the Appendix to chapter 2 and is to be published as: Raikou, M. and McGuire, A., forthcoming, The Economics of Type 2 Diabetes, Pharmacoeconomics.

Acknowledgements

Over the years it took to reach the stage of being able to write these words support has been provided from a number of sources. This thesis, initiated at City University, London, was funded in part by a London NHS Executive grant and I am grateful for their financial support.

I also wish to acknowledge the assistance provided by a number of individuals at various points in time. I would like to thank Professor David Cox of Nuffield College, Oxford for directing me to some of the sources upon which this study was built and for comments on certain issues and Dr Heejung Bang of Harvard University for answering specific questions on some of the estimators used in chapter 4. I am also grateful to a number of individuals involved in the UK Prospective Diabetes Study from which the data used in this thesis is drawn, specifically to Professors Alastair Gray and Rury Holman both of the University of Oxford for early support. In particular however I am indebted to Professor Robert Turner, without whom the UKPDS would never have met the success it did, both for his personal encouragement and support early in the thesis, and more importantly for his faith and trust in me during the whole period I worked under his direction.

This has taken a long time and I guess many suffered from it. Given that they were left without choice they did their best to provide me with support over the years and the least I can do is to say thanks. The person who suffered the most is Professor Alistair McGuire who had to develop nerves of steel during the process but remained standing until the very end. His input and patience over the years cannot be mentioned enough. Although I had to teach him a lot during the thesis he is a quick learner and he can be very perceptive on occasions though not very frequently. Despite the fighting I could not have wished for a better "supervisor". I also owe a great deal to my family who still put up with me after all these years (thank you Dimitri for waking me up every morning, Irene for being your difficult self and Katerina for being more grown up than me at times). While these four individuals have suffered the most some others have also provided various forms of support – many in the form of drink. Special thanks go to Edward who always thought I would do it and whose support at times made all the difference, to Yanni for his critical interventions and to my uncle Adoni who says I should also thank Karl Marx and Vladimir Lenin – although the thesis has not left me time to find out why.

If the words "dedicated to" mean anything, then two individuals should succeed them. Robert Turner, my first "boss" whom I have missed many times and Alistair McGuire a true friend more than a supervisor who chummed me all the way.

Ευχαριστίες

Στη διάρκεια των χρόνων που χρειάστηκαν μέχρι να φτάσει η στιγμή να γραφτούν αυτές οι λέξεις, μου προσφέρθηκε υποστήριξη από πολλές κατευθύνσεις. Η παρούσα διατριβή, η οποία ξεκίνησε στο Πανεπιστήμιο City του Λονδίνου, χρηματοδοτήθηκε εν μέρει μέσω υποτροφίας που μου χορήγησε το τμήμα Έρευνας και Ανάπτυξης του Εθνικού Συστήματος Υγείας (Λονδίνο) και τους είμαι ευγνώμων για την οικονομική υποστήριξη.

Παράλληλα θα ήθελα να επισημάνω τη βοήθεια που μου προσέφεραν συγκεκριμένα πρόσωπα σε διάφορες χρονικές περιόδους. Θα ήθελα να ευχαριστήσω τον Καθηγητή David Cox του Nuffield College του Πανεπιστημίου της Οξφόρδης για τις κατευθύνσεις του σε μερικές θεμελιώδεις για τη διατριβή πηγές και τα πολύτιμα σχόλια του, και τη Δόκτορα Heejung Bang του Πανεπιστημίου του Harvard για τις διευκρινήσεις της σχετικά με κάποιους από τους εκτιμητές (estimators) που χρησιμοποιούνται στο κεφάλαιο 4. Ευχαριστώ επίσης τους Καθηγητές Alastair Gray και Rury Holman του Πανεπιστημίου της Οξφόρδης για τη συνεισφορά τους στο αρχικό στάδιο της παρούσας εργασίας μέσω της συμμετοχής τους στην UK Prospective Diabetes Study από όπου αντλήθηκαν τα δεδομένα που χρησιμοποιούνται στη διατριβή αυτή. Ιδιαίτερα υποχρεωμένη όμως νοιώθω απέναντι στον Καθηγητή Robert Turner, χωρίς τον οποίο η UK Prospective Diabetes Study δεν θα είχε γνωρίσει το βαθμό επιτυχίας που γνώρισε τελικά, όχι μόνο για την ενθάρρυνση και υποστήριξη του στο ξεκίνημα της διατριβής αλλά κυρίως για την πίστη και την εμπιστοσύνη που μου έδειξε καθ'όλο το χρονικό διάστημα που δούλεψα υπό την καθοδήγηση του.

Η διεκπεραίωση της παρούσας εργασίας πήρε πολύ καιρό και έχω την αίσθηση ότι πολλοί υπέφεραν κατά τη διάρκεια. Δεδομένου ότι στερήθηκαν επιλογής έκαναν ό,τι καλύτερο μπορούσαν για να με υποστηρίξουν όλα αυτά τα χρόνια και το λιγότερο που μπορώ να κάνω είναι να πω ευχαριστώ. Εκείνος που υπέφερε το περισσότερο είναι ο Καθηγητής Alistair McGuire, ο οποίος αναγκάστηκε να αναπτύξει ατσάλινα νεύρα κατά τη διάρκεια της όλης διαδικασίας αλλά παρέμεινε όρθιος μέχρι το τέλος. Δεν υπάρχουν λέξεις που να αποδίδουν το μέγεθος της εισφοράς και της υπομονής του. Παρόλο που χρειάστηκε να του διδάξω πολλά κατά τη διάρκεια της διατριβής, μαθαίνει γρήγορα και μπορεί να επιδείξει αξιοσημείωτες ικανότητες αντίληψης και κατανόησης κατά περιόδους - αν και όχι ιδιαίτερα συχνά. Παρά τις συγκρούσεις και τους τσακωμούς δεν θα μπορούσα να έχω ελπίσει σε καλύτερο "επιβλέπων" καθηγητή. Πολλά επίσης χρωστάω στην οικογένεια μου που με ανέχεται ακόμα μετά από τόσα χρόνια (σ'ευχαριστώ Δημήτρη που με ξυπνάς κάθε πρωί, Ειρήνη που είσαι ο δύσκολος εαυτός σου και Κατερίνα που κάποιες φορές είσαι πιο ώριμη από μένα). Αν και οι παραπάνω υπέφεραν τα περισσότερα, υπήρξαν και άλλοι που με στήριξαν με διάφορους τρόπους - κυρίως μέσω βραδυών οινοποσίας και κονιάκ. Ιδιαίτερο ευχαριστώ ανήκει στον Edward που πάντα πίστευε ότι θα τα καταφέρω και του οποίου η υποστήριξη κατά καιρούς έκανε τη διαφορά, στο Γιάννη για τις παρεμβολές του σε στιγμές κρίσης και στο θείο μου Αντώνη που ισχυρίζεται ότι θα έπρεπε επίσης να ευχαριστήσω τους Karl Marx και Vladimir Lenin - αν και η εκπόνηση της διατριβής δεν μου άφησε χρόνο να ανακαλύψω το γιατί.

Αν οι λέξεις "αφιερωμένο σε" σημαίνουν κάτι, τότε δύο είναι εκείνοι που πρέπει να τις ακολουθήσουν. Ο Robert Turner, το πρώτο μου "αφεντικό" την απουσία του οποίου έχω αισθανθεί πολλές φορές και ο Alistair McGuire, αληθινός φίλος περισσότερο από επιβλέπων καθηγητής που ήταν δίπλα μου σε κάθε βήμα απ'τήν αρχή μέχρι το τέλος.

Abstract

This thesis is concerned with two specific cost measurement issues commonly raised within the context of economic evaluation of health care interventions. Both these issues arise due to limited availability of cost information from medical studies. The first relates to the process of cost data collection while the second relates to the statistical analysis of cost data. Investigation of this subject matter is undertaken with reference to clinical trials although this setting does not restrict the generality of the findings. The typical pattern of cost collection records volumes of resource use at the patient level but not resource unit cost in the treatment centre where the resource was utilised. The calculation of treatment cost then is normally based on some average unit cost estimate obtained from a variety of sources as opposed to centre specific unit cost information. The question arises as to whether the source of unit cost information has an impact on the calculated total treatment costs. This is addressed using a simulation setting and assuming specific underlying production and cost relations which determine the behaviour of treatment centres in delivering a health outcome. The results show that assuming the treatment centres operate in a manner consistent with economic theory, using average instead of centre specific unit cost information will lead to biased estimates of the total cost of treatment. The issue of primary concern in the thesis relates to the incompleteness of cost information for analysis due to censoring. Censoring occurs whenever patients are not observed for the full time to event and affects both effectiveness and cost data. Any analysis of such data that fails to account for the presence of censoring will result in biased estimates of the statistics of interest. This issue has only recently been addressed in the literature within the context of cost analysis and a well established methodology for dealing with this problem is lacking. There are a limited number of parametric and non-parametric estimators which have been proposed in an attempt to adjust cost estimates for censoring all of which are considered here. A subset of those lack theoretical justification and as such lead to erroneous inferences, while those whose use is justified on theoretical grounds have not been empirically assessed under conditions of heavy censoring using real medical data. This is undertaken in the present analysis using a clinical trial dataset which displays extreme levels of censoring. Although the theoretical investigation shows that under specific assumptions the approaches provide consistent estimators of mean cost while accounting for the loss of information due to censoring, the analysis reveals various performance patterns ranging from generally stable estimators under the conditions considered to estimators which become increasingly unstable with increasing levels of censoring.

Chapter 1

Introduction

The necessity of adopting economic evaluation in the health care sector arises because the market fails to fulfil the conditions required to ensure efficient allocation of resources. Economic evaluation then provides a method for determining the point of efficiency, that is the point at which the allocation of resources leads to maximisation of social welfare. In the process of achieving the optimal resource allocation, alternative states have to be evaluated each one associated with different individual welfare levels. Given that any alternative state of resource allocation will normally result in an improvement in welfare for some individuals and a deterioration for others, interpersonal comparisons of utility have to be made in order to determine whether there is a net gain in social welfare. The choice becomes then either to consider situations in which unambiguous welfare improvements are possible or to consider a wider range of situations by making interpersonal comparisons. In the former case, evaluation of alternative states is undertaken based on the Pareto principle according to which welfare improvement occurs if resource allocation is such that an individual is made better off without making another individual worse off. In the latter case, value judgements must be made to determine whether there are net gains in welfare. Given that evaluations based on the Pareto optimality criterion do not encompass value judgements and interpersonal comparisons of utility levels, the Pareto principle has to be accompanied by the implementation of the Hicksian compensation test which leads to efficiency being defined in terms of potential Pareto improvement.

In this context cost-benefit analysis is implemented specifically as a means of achieving Pareto welfare. The method itself is consistent with the Pareto principle and does not encompass interpersonal comparisons in determining the optimal resource allocation state. It can however be accompanied by information on individual welfare changes resulting from implementation of the alternative under consideration determined based on the Hicksian compensation criterion. As such cost benefit analysis in conjunction with the compensation criterion can in theory provide a mechanism for determining optimal resource allocation patterns consistent with the maximisation of social welfare. In the health care sector where the monetary valuation of outcomes is complex cost-effectiveness commonly replaces cost-benefit analysis as a method of identifying patterns of health care resource allocation. If relative valuations can be attached to health states then cost-effectiveness may encompass cost-utility analysis. All these methods, cost-benefit, cost-effectiveness and cost-utility analysis, have specific problems of implementation which have long been discussed in the welfare economics literature. Most recently three particular themes have come to dominate the literature in health economics.

First there has been increasing consideration of the specific technical conditions under which cost-effectiveness and cost-utility analyses relate to cost-benefit analysis whose objective is to identify Pareto optimal states consistent with the maximisation of social welfare. Under the assumption that consumers and producers are utility maximisers and the health care sector is budget constrained, a range of models have been developed each specifying a utility function based on specific underlying assumptions. These are then used to reveal the specific conditions under which patient preferences, normally represented in the model by some quality adjusted life year concept, can be related to cost-benefit analysis and traditional notions of welfare economics (Garber, 2000). Weinstein and Stason (1976) and Weinstein and Zeckhauser (1973) discuss the relationship of cost-effectiveness to cost-benefit analysis through the use of linear programming techniques.

Secondly there has been growing criticism of the traditional definition of welfare as based on Pareto optimality and utility maximisation (Williams and Cookson, 2000; Tsuchiya and Williams, 2001). It has been suggested that the definition of welfare ought to take account of concepts that are not solely utility based. The justification for this approach derives from Sen's argument that welfare is not only defined by means of utility but is also related to fundamental attributes, which he refers to as basic capabilities (Sen, 1982). On this basis, proponents of the notion of extra-welfarism have suggested that efficiency may be defined with regards to the maximisation of health and not utility *per se*. As such these capabilities may be related to cardinal measurements of health benefit allowing the problems imposed by interpersonal comparisons to be overcome. Within this context the role of economic evaluation is not to determine the optimal allocation of health care resources that will maximise utility-based welfare, but rather to supply the relevant decision makers with information that assists their assessment of the appropriate allocation of health care resources. That is, interpersonal comparisons are considered through the explicit cardinal measurement of health benefit, using for example QALYs, but ultimately it is the decision maker who defines the trade-offs across individuals in specifying the social welfare function. Under this interpretation cost-effectiveness and cost-utility analyses are not necessarily related to cost-benefit analysis as there is no attempt to follow Paretian notions of efficiency. Cost-effectiveness and cost-utility analyses become appropriate allocative tools in their own right.

The third theme that has dominated the literature has dealt with measurement issues given that any economic evaluation involves measurement of the costs incurred by and the benefits derived from a health care intervention. Particular emphasis has been given to the measurement of the benefits derived from a given health outcome partly reflecting the fact that the various types of economic evaluation adopt different definitions of the health outcome. Cost-benefit values health outcomes by a monetary metric, cost-utility by a measure of relative value of health states and cost-effectiveness more generally by some appropriate physical quantity. A considerable literature has addressed the issue of how monetary measurements of health states may be obtained more recently concentrating on the techniques of conjoint analysis (Deiner et al, 1998). At the same time there is a sizeable literature that considers the measurement of the relative valuations of health states using non-

monetary values (Dolan, 2000, 2001). Here the topics analysed include the mathematical properties required to elicit relative valuations using a variety of instruments including the visual analogue score, the time trade-off method and the standard gamble. A directly related literature considers the relative strengths and weaknesses of the various measures. A different component of the literature has concentrated on the issue of data uncertainty with the largest part of this literature addressing the question of how the uncertainty surrounding the incremental cost-effectiveness ratio statistic should be handled (Briggs, 2001). Consideration of the appropriate methodology for calculating confidence intervals for this ratio statistic constitutes another substantial part of this literature followed by investigation of alternative ways in which the ratio may be presented as for instance in the case of the net benefit approach which essentially results in a transformation of the incremental cost-effectiveness ratio. Other measurement related issues that have received some attention in the literature include adjustments for missing data, the measurement of indirect costs and the transferability of findings across different regulatory environments.

This thesis relates to the general aspect of measurement issues. Having focussed on measurement problems raised within the context of analysing health outcome data, the literature has generally given less attention to the issues that arise in the analysis of cost data. With respect to cost the matter most commonly addressed is the definition and measurement of indirect costs (Sculpher, 2001). The measurement of direct costs has received less attention. There is a relatively small literature which considers the appropriate definition of direct costs and their relationship to opportunity costs and charges (Brouwer et al, 2001; Dranove, 1996). There is limited consideration however of the impact that different data collection methods and different methodological approaches employed in analysing cost data have on the estimates of cost statistics.

The limited information available with regards to direct cost measurement in general and the lack of a well-established methodology in dealing with particular statistical issues arising in the analysis of treatment costs are themselves a justification for the subject matter that follows. However the emphasis placed on the analysis of treatment cost is also due to another growing tendency within the cost-effectiveness literature. This is the increasing adoption of economic analysis alongside clinical trials. Undertaking an economic evaluation alongside a clinical trial presents specific analytical problems due to the experimental design itself. A clinical trial is designed to test a hypothesis of a clinically important difference between two alternative treatments. As a result data collection is primarily concentrated on accumulating the necessary clinical information possibly at the expense of economic data requirements. In addition to being of secondary importance, recording cost information on every cost generating event may also be a relatively demanding process. Consequently, it is commonly the case that only the minimum amount of cost data is available from the trial itself. A further consequence of the nature of the experimental design is that the study will end once the difference in the clinical endpoint between the trial arms is attained. This means that the study will terminate prior to every patient reaching the prespecified endpoint, a condition referred to as censoring, resulting in information on some patients not being available for the whole

duration of interest. Any statistical analysis should account for this information loss, an issue well recognised within the context of time to event data analysis. The same concerns arise in the analysis of cost data where the censoring mechanism results in similar loss of information.

As an extension to this literature, two specific cost measurement issues are addressed in this thesis. The first relates to the collection of cost data and the second to the treatment of censoring in the statistical analysis of costs. Both issues are considered within a clinical trial setting although this setting does not restrict the generality of the findings. The issue relating to cost collection is investigated using a simulated dataset whereas investigation of the condition of censoring, which constitutes the major objective of the thesis, is undertaken using a dataset drawn from a prospective randomised clinical trial briefly described below.

The data were drawn from the UK Prospective Diabetes Study (UKPDS 33, 1998). A total of 5102 newly diagnosed type 2 diabetic patients defined as having fasting plasma glucose (fpg) greater than 6mmol/l on two occasions, aged 25-65 years (mean age 53) were recruited to 23 UK study centres. After initial diet treatment a total of 4209 had fasting plasma glucose between 6.1 mmol/l and 15 mmol/l without symptoms of hyperglycaemia and were followed-up. Of these, 342 overweight patients were randomised to metformin leaving 3867 individuals who entered the main randomisation and were allocated either to conventional policy (1138 patients) achieved primarily through diet or to intensive policy (2729 patients) based on either insulin (1156) or sulphonylurea (1573). The aim of the conventional policy was to maintain patients free of diabetic symptoms and with a fasting plasma glucose of less than 15 mmol/l, whereas the aim of the intensive policy was to achieve a fasting plasma glucose concentration of less than 6mmol/l. Figure 1.1 shows the main randomisation process.

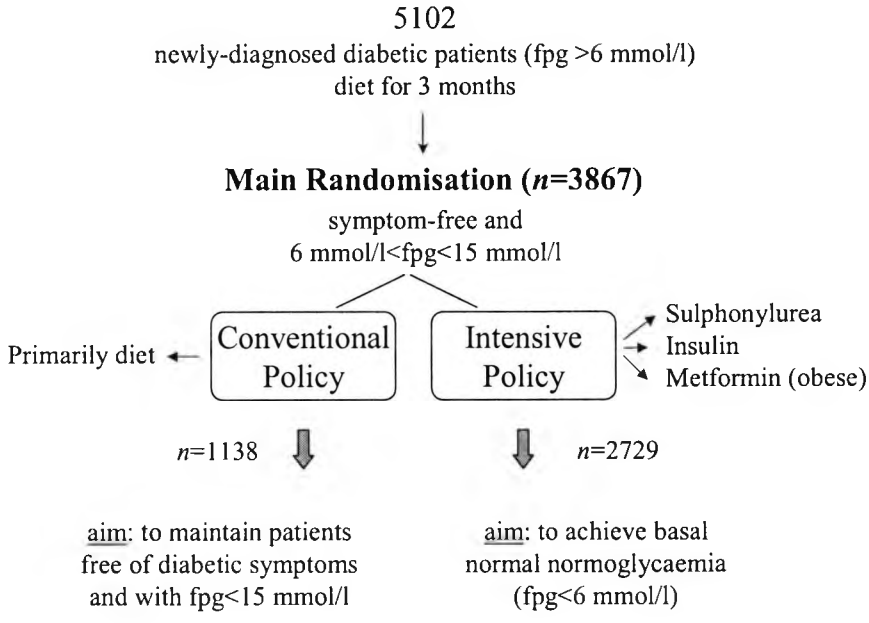


Figure 1.1. The UKPDS main randomisation

The major clinical endpoints analysed were death or the development of diabetic complications including coronary heart disease, cerebrovascular disease, amputations, laser treatment for retinopathy, cataract extraction and renal failure. All analyses and comparisons were performed on an intention-to-treat basis. The trial started in 1978 and ended in 1998 with a median follow-up to death, the last date at which clinical status was known or to the trial end of 10 years.

For each patient in the study data were collected at 3 monthly clinic visits on the doses of all agents used for the treatment of diabetes as well as all other co-medications and the number of home blood glucose tests. Data on the date and duration of each hospital admission were collected at every clinic visit. These were coded using ICD-9 and ICD-10 classifications for prime cause of admission and OPCS-4 codes for all procedures undertaken. In addition a separate record was maintained of all angiograms, angioplasties and bypass grafts for coronary or peripheral vascular disease. Data on non-inpatient health care resource use were collected from all patients in the trial using a questionnaire distributed at routine clinic visits between January 1996 and September 1997 and by post to those who had not attended a clinic during that period. This questionnaire collected information on all home, clinic and telephone contacts with general practitioners, nurses, chiropodists, opticians, dieticians, eye and other specialists over the preceding four months. These cross-sectional data were analysed using multiple regression techniques to estimate for each patient the annual non-hospital resource use adjusted for significant variables including age, gender, body mass index, duration of diabetes and time from a non-fatal diabetes-related endpoint. Unit costs for all resources used by the trial patients were obtained from national statistics and from centres participating in the trial. These unit costs were then combined with the resource volumes to obtain a cost per patient over their time in the trial.

The data on non-inpatient costs were not included in the analysis undertaken in the thesis as the underlying resource use data were not collected during the trial for all patients but were estimated from a regression model. In addition all costs used and reported in subsequent chapters are in 1997 UK £s and are not discounted as the analysis is concerned with assessing the differences among alternative methods in accounting for censoring and not the difference in the costs between the trial arms. For these reasons the cost estimates reported in subsequent chapters are not directly comparable to those reported in the UKPDS economic paper (UKPDS 41, 2000).

The clinical trial reported that the intensive blood glucose control policy significantly reduced the risk of any diabetes related endpoint by 12% (P=0.029), but did not significantly reduce diabetes related deaths or all-cause mortality. The diabetes related endpoints were myocardial infarction, congestive heart failure, stroke, renal replacement therapy, amputation, cataract extraction, vitreous haemorrhage and death from any cause. With respect to the analysis undertaken in the thesis the failure event was death by any cause. Figure 1.2 presents the Kaplan-Meier estimates for the survival probabilities by randomisation group and shows graphically the lack of clear difference in

mortality between the conventional and intensive policy populations as reported by the trial (UKPDS 33, 1998).

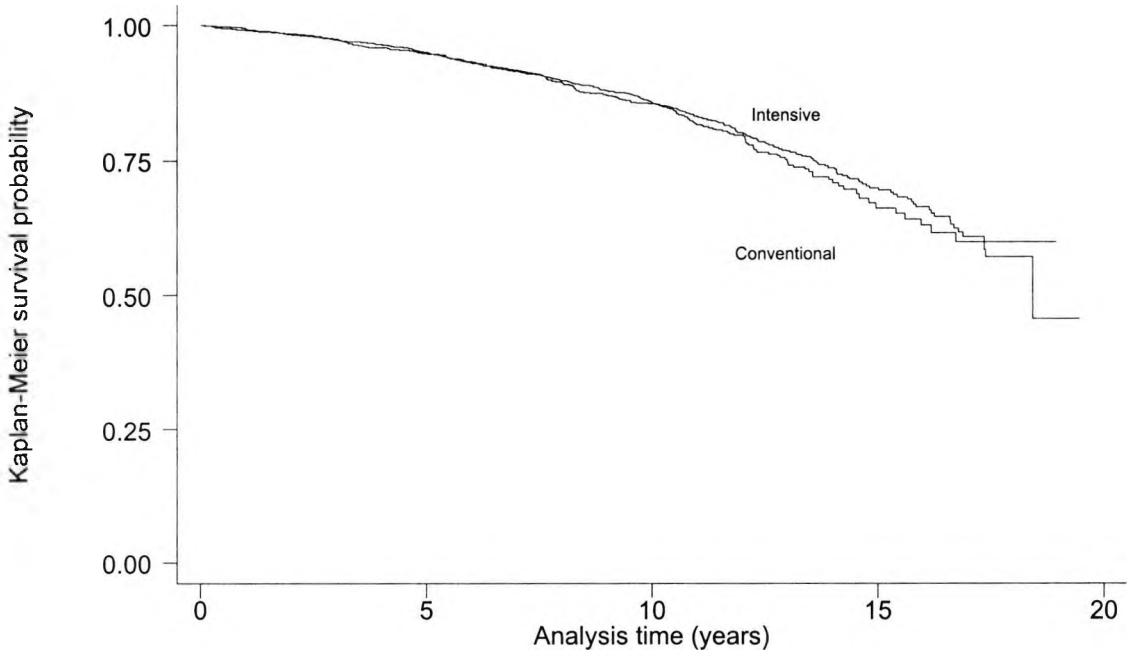


Figure 1.2. Kaplan-Meier estimates for the survival probability by randomisation group

The thesis is structured as follows. Chapter 2 presents an overview of the literature that addresses the methodological issues concerned with the collection and analysis of cost data alongside clinical trials. This review is not systematic, indeed it is not intended to be given the existence of a number of recent reviews in this area, but is meant to provide the necessary background information as a means of introduction to the analysis presented in the subsequent chapters and in particular to the analysis of censored cost data. Emphasis therefore is given to the various statistical issues that arise in the analysis of cost data and the respective methodologies that have been proposed to address these issues. The main findings suggest that in general in medical studies and in particular in clinical trials cost data will typically present analytical problems due to a number of reasons relating both to the specific application under consideration but more importantly to the complexity of the cost distributions in general. Censoring appears to be the most commonly encountered feature of cost data drawn from a clinical trial and given the experimental design of such studies it is likely that it will continue to be so. The importance of successfully handling the presence of censoring in deriving estimators for the statistics of interest has been long established in the context of time to event data but only recently in the context of cost data. Consequently the number of the proposed approaches that attempt to deal with the bias induced by censoring in the cost estimates is limited. Moreover some of these approaches lack theoretical justification and have been erroneously recommended in the literature. Finally the recently developed methodologies whose use is justified

in the context of censored cost data analysis on theoretical grounds have not been empirically assessed under conditions of heavy censoring using real clinical data.

An appendix to chapter 2 (Appendix A.2.1) presents a systematic review of the economic literature relating to type 2 diabetes. The motivation for undertaking this review was that as stated above the dataset used in analysis in the later chapters was drawn from a clinical trial in type 2 diabetes. Although this review was undertaken primarily for completeness, as the methodology studied in the subsequent chapters is not dependent on the disease area, it served to indicate that although type 2 diabetes is a major disease area the number of relevant economic analyses is limited. Most of the studies are either cost-of-illness studies or pure modelling studies. The most important cost-effectiveness analyses undertaken to date appear to be the ones accompanying the UKPDS.[1] Moreover, while these studies provide essential information to the understanding of the costs associated with type 2 diabetes and its complications and to the assessment of the cost-effectiveness of the proposed treatment interventions, it is revealed that no individual study in this disease area has addressed the issue of censoring in the cost data.

Chapter 3 considers an issue raised within the UKPDS economic analysis and liable to arise in any multi-centre study relating to the collection of costs rather than their statistical analysis. The typical pattern of cost data collection within a multi-centre study involves recording volumes of resource use at the patient level but not the recording of unit cost information from the individual participating centres. As such the available cost information reflects centre-specific resource utilisation but not centre-specific unit costs. In these circumstances, the unit cost attached to the resource volumes is some average unit cost obtained either from a sample of the participating centres or from a number of published sources. The question then arises as to whether using such an average unit cost in the calculation of total treatment cost instead of the unit cost in the centre where the resource was utilised has an impact on the cost estimates. This question is addressed using a simulation setting where an underlying production function is assumed which determines the behaviour of treatment centres in delivering a health outcome. Under a formal specification of the production and cost relations the simulation exercise considers a number of specific interactions between unit cost and resource volumes and compares the estimates of treatment costs resulting from the various assumptions employed within the simulation setting. The results show that under the assumption that treatment centres operate in a manner consistent with economic theory as assumed by economic evaluation, ignoring the underlying production-cost relations will lead to biased estimates of treatment costs.

Chapters 4 and 5 concentrate on the impact of censoring on the estimates of cost statistics. Chapter 4 investigates the theoretical properties, underlying assumptions and empirical performance of a number of non-parametric estimators of average cost which attempt to account for the loss of

[1] The author of this thesis was a member of the UKPDS economic group and was involved in a number of the individual economic analyses associated with the UKPDS.

information imparted by censoring. The estimators are studied within the counting process and martingale framework given that their asymptotic properties have been established using this theory as applied to survival analysis. The development of the theory of counting processes, stochastic integration and martingales is therefore presented but only to the extent required for the applications of interest. The proposed estimators are then assessed under extreme censoring conditions using the UKPDS data which exhibits 82% censoring by the trial end. The analysis shows that of the estimators expected from the theory to provide consistent estimates of average cost in the presence of censoring, only two appear to do so under heavy censoring conditions. Furthermore, both these adequately performing estimators require information on individuals' cost history recorded at intermediate points in time over the study period.

Chapter 5 addresses the same issue as chapter 4 but with reference to semiparametric and parametric approaches. Being of a semiparametric or parametric nature, the set of estimators considered in this chapter involve additional assumptions with regard to the distribution of costs relative to the non-parametric estimators of the preceding chapter. On the other hand such estimators, if they perform adequately, will allow extrapolation of within study estimates to different populations or to points in time beyond the end of the study. The estimators' asymptotic properties are again investigated using the counting process and martingale theory when required and when the idea underlying their development originates from the study of time to event data, the approach is first considered within this context and the extension to the analysis of cost to event data follows. The proposed estimators are empirically assessed under the same censoring conditions as the non-parametric alternatives using the UKPDS data and the resultant estimates are compared to the best performing estimators identified in the preceding analysis. This comparison shows that of the parametric methodologies whose use in the analysis of censored cost data is theoretically justified, only one appears to perform adequately under the circumstances considered in this analysis. Moreover this estimator shares the two most important features of the best performing non-parametric estimator, namely both these approaches use intermediate cost information on each individual in the study and in both cases censoring is accounted for using the same type of adjustment in the estimating equations.

A final chapter considers the central themes investigated in the thesis concentrating on the main results of the analysis and concluding on the implications of the findings for the study of medical costs under conditions of censoring.

Chapter 2

Review of methodological issues in the collection and analysis of cost data

2.1. Introduction

The objective of this chapter is to present some of the fundamental methodological issues arising in the collection and analysis of cost data generated from a clinical trial and to highlight specific statistical issues of concern. The purpose of addressing these methodological issues is to set the context for the subsequent chapters which are concerned with various aspects of cost collection and analysis both from a theoretical viewpoint as well as from an empirical one, the latter being undertaken using clinical trial data. This chapter therefore provides the general background to the substantive contents of the thesis and while reviewing the relevant literature is not a systematic review. More specifically, chapter 3 investigates the impact of different levels of cost data collection on estimates of total treatment costs. Accordingly, while the introduction to chapter 3 makes specific reference to the literature that addresses data collection issues, this chapter provides an overview of the arguments various investigators have forwarded on which costs ought to be collected and how specific these should be. Chapters 4 and 5 consider a number of non-parametric and parametric estimators of mean treatment cost under conditions of censoring and assess the estimators' performance using data collected alongside a clinical trial. This condition occurs frequently in medical studies and affects both measures of cost and effectiveness. The importance of ignoring the presence of censoring when deriving estimates of medical costs is increasingly acknowledged. This issue is addressed in great detail in the subsequent chapters and is therefore only briefly discussed in this review chapter. In addition, censoring was the most prominent feature of the trial data used for the analysis in this thesis whereas the remaining of the potentially problematic issues presented below were either almost absent or completely absent from these data.

An appendix to this chapter (Appendix A.2.1.) presents a systematic review of the economic literature relating to type 2 diabetes. The review of this disease specific literature was undertaken for the sake of completeness as the empirical analysis in later chapters uses data from a prospective randomised controlled clinical trial in type 2 diabetes. This review was deemed of secondary importance on the grounds that at no point was it expected that the results of the analysis presented in the subsequent chapters would be disease-specific. On the contrary, the explored analytical methods are completely independent of the disease area considered and consequently their applicability is in no way restricted by the nature of the clinical area the data analysed are drawn from.

The chapter therefore is structured as follows. As a means of introduction a short discussion of the role of economic analysis alongside clinical trials is given highlighting points of concern raised in the literature. This is followed by a discussion of the problems commonly encountered in collecting and analysing cost data together with the methodology employed in an attempt to overcome these difficulties.

2.2. Economic analysis alongside clinical trials

A number of studies have considered cost data issues in the context of economic evaluation alongside clinical trials. Johnston et al (1999) give an overview of the issues. Given the existence of this review the findings are not replicated here. Instead a summary of the aspects which are of relevance to the issues addressed in the thesis is presented. There is a related literature which assesses the relative advantage of economic evaluation undertaken alongside a clinical trial over evaluation based solely on modelling techniques (Buxton et al, 1997; Morris, 1997; Kuntz and Weinstein, 2001). Again only the main points are stated here. In the recent collection edited by Drummond and McGuire (2001) costing issues are also considered in a number of chapters and the main relevant points are included in the exposition below.

Data for the purposes of economic evaluation can be obtained from a number of sources including clinical trials and a variety of models or a combination of the two instruments. There are a number of advantages and disadvantages associated with each data source, though generally clinical trials are preferred to modelling approaches largely because they are believed to give a better indication of cost-effectiveness given their reliance on real albeit experimental data. Under certain circumstances however modelling is unavoidable either because economic data have not been collected alongside the trial or more commonly because there is an interest in generalising the trial results to different population settings or to points in time that exceed the duration of the clinical study. Modelling therefore can be useful in situations where the trial data present limitations with respect to a number of factors such as the length of the study duration, the range of assessed outcomes or the experimental design itself. In these situations modelling allows for example extrapolation from intermediate clinical endpoints to final outcomes, the linking of disease-specific measures of health state preference values to a standard utility value and generalisation of the findings to standard clinical practice settings. Moreover there are situations where use of a statistical model provides the only means for analysis due to specific data problems as is the case for instance when some data are missing or the variable of interest is censored. Clinical trial based economic evaluations and economic evaluations based on modelling approaches are therefore not necessarily mutually exclusive analytical techniques.

In comparisons between the two instruments however the established view is that clinical trials provide the most powerful means of assessing the efficacy of health care interventions. Assessment

of an intervention's efficacy requires that all confounding factors be controlled for to the largest possible extent through the experimental design and patient entry criteria. From the perspective of an economic evaluation however the outcome of interest is not efficacy but effectiveness. Assessing an intervention in terms of effectiveness requires that the health outcome derived from the intervention represent a relative gain compared to an alternative under conditions of a standard clinical practice setting. By contrast, an intervention is deemed efficacious if there is a relative clinical benefit attributed to the intervention under the controlled ideal conditions defined by the experimental study design. It is important therefore that efficacy information derived from clinical trial data be adjusted to reflect standard clinical practice conditions for the purposes of economic evaluation. Otherwise, given the ideal conditions specified by the trial design, efficacy will normally represent the upper bound of effectiveness.

Aside from the efficacy issue, the randomised controlled trial is of relatively limited use to economic evaluation for the following related reasons. Randomisation, while it tends to balance prognostic factors across the study groups is at the same time linked to a number of exclusion criteria on the basis of which the groups are defined. As a result the generalisability of the findings to different population settings is limited. In addition clinical trials are typically undertaken in specialist settings and use the most recent medical equipment. A strict protocol is often followed and treatment as well as disease progression are carefully monitored with every effort being made to ensure that both patients and clinicians comply with the trial requirements. Consequently the treatment pattern observed within a trial setting is likely to be substantially different from the pattern corresponding to a standard practice setting. This discrepancy will lead to difficulties in assessing not only the effectiveness of the intervention but also its cost. In an attempt to reflect real clinical practice more closely, some clinical trials follow a naturalistic protocol in the sense that the study patients are relatively typical of the normal caseload, the intervention being assessed is compared with current standard practice, the setting used and the physicians involved are fairly representative of the population and the trial protocol is flexible. Such pragmatic studies are normally based on a cohort study design and attempt to provide information both on the incidence of disease and the impact of the intervention under consideration. Naturalistic studies however are normally time-consuming, often require large sample sizes, incur significant loss of follow-up, are costly to undertake and consequently relatively uncommon. Moreover, while naturalistic studies attempt to reflect real conditions and are by definition less tightly controlled than randomised controlled trials, they are still restricted by the settings in which they are conducted and the population involved and therefore they are not entirely generalisable.

The concerns raised above though important should not be interpreted as undermining the usefulness of the experimental design for the purposes of economic evaluation. Furthermore modelling approaches can be employed to overcome some of these difficulties. Given that the objective of a clinical trial is different from the objective of the economic analysis conducted alongside it, that is the trial will normally be set up to test a hypothesis of clinical importance as

opposed to one of economic importance, it is important that this discrepancy be accounted for. This is another reason why modelling techniques are increasingly used within such an analytical setting.

2.3. Cost collection

Concentrating on resource use and cost issues it appears that the number of studies that have used patient level resource and cost data alongside a clinical trial is limited although increasing. Briggs and Gray (1998) identified a number of such studies as part of a wider systematic review concerned with the handling of uncertainty in the cost-effectiveness literature. They reported on 368 published studies up to the end of calendar year 1995 which were identified as cost-effectiveness or cost-utility studies. Of these, only 22 studies had used patient specific data collected alongside a randomised controlled trial and a further 19 studies had used such data collected within another form of clinical study. That is, only 11% of the identified studies in their review had analysed patient specific resource/cost data drawn directly from the clinical study. Of this total, only 3.3% reported some conventional measure of variance. In other words the vast majority of economic evaluations appear not to be utilising the information required to perform statistical analysis of cost data drawn from a clinical study, and even where such information was reported only a subset derived measures of variation associated with the estimates of the average cost of treatments. This finding is consistent with the conclusion reached by Barber and Thompson (1998). They identified 45 randomised clinical trials published in the year 1995. They reported that while each of these trials was believed to have recorded cost data at the individual patient level, less than 20% reported standard measures of variability although 56% performed statistical tests to compare resource costs across the treatment groups.

Issues generated by the cost data collection process alongside a clinical trial have been addressed in a number of publications (Drummond, 1994; Drummond and Davies, 1991; Drummond and Stoddart, 1984; Buxton et al, 1997; Glick et al, 2001). Aside from addressing the question of how to integrate cost collection into the trial design, these studies also consider whether power calculations ought to be undertaken on the economic endpoints, whether clinical trials which generally measure efficacy can appropriately define effectiveness, whether the population selected for inclusion into the trial through controlled entry and design is appropriate for consideration by an economic evaluation as the population in a trial is normally more homogeneous than the heterogeneous population faced by decision makers, whether the selection of the control arm forms an appropriate comparator for the purposes of an economic evaluation and what forms the most appropriate methodology for the analysis of the collected data.

A small number of studies have addressed the specific question of which particular cost components are relevant to the estimation of treatment cost within an economic evaluation (Brouwer et al, 2001; Gold et al, 1996). Glick et al (2001) outline a general strategy for designing an economic evaluation

to be undertaken alongside a clinical trial. With regards to the question of what proportion of total costs ought to be collected as part of the trial they offer the rather simplistic answer that as many components of cost as possible should be measured. Their rationale is that minimising the loss of information reduces the likelihood that cost differences among the competing treatments will be due to study artefacts although they also note that there are no *a priori* guidelines. There have been a limited number of studies which have addressed this question empirically. Whynes and Walker (1995) analysed costs collected alongside a trial for colorectal cancer on 360 patients over a 3-year period during which individuals could have up to 14 individual cost generating events. The authors undertook a detailed costing exercise on an individual patient basis incorporating all 14 cost generating events and compared the resultant cost estimates with costs derived by using the same information at a more aggregate level. Their results indicated that estimating costs based on 4 specific cost categories rather than 14 accounted for approximately 92% of total costs on average but this finding held for only 44% of individual patients. On this basis the authors recommend collecting cost data at a patient level as this information reveals differences in cost patterns among individual patients which would not have been identified if an aggregate costing exercise had been undertaken. A similar study in the area of mental health (Knapp and Beecham, 1993) concluded that of the 21 identified cost categories, 5 accounted for approximately 95% of total cost while 10 categories accounted for 98% of total cost. The findings from these studies suggest that while aggregation of cost elements into categories is possible with direct implications for the process of cost data collection, it would be difficult to identify important cost categories at the outset of the study both due to the variation in the distribution of cost across individual patients and due to the variation in the distribution of cost across disease areas.

Graves et al (2002) assess the quality of the methods used to derive patient level cost in economic evaluations conducted alongside clinical trials. They review the same 45 studies reported by Barber and Thompson (1998) referred to above. The quality of the methods employed by the studies was assessed using twelve criteria that covered general costing issues, the methods used to identify and quantify the resource elements and the reporting of data. The results indicated that the vast majority of the reviewed studies failed to attain acceptable levels of quality as specified by the authors' criteria. The authors therefore concluded that although the statistical analysis of cost data collected alongside a clinical trial is fundamental to the evaluation of health care technologies, the process of cost collection and collation is equally important and as the title of their article states "No amount of statistical analysis can compensate for poor quality cost data" (op. cite).

Once the relevant data have been collected the issue becomes to determine the most appropriate method for analysis. In situations where information on resource use is not available from the trial economic analysis is based on some modelling approach and the necessary information is typically obtained from published sources. In these circumstances statistical analysis of the distribution of treatment cost based on the observed trial population sample is not possible. Uncertainty surrounding the cost estimates may then be assessed by univariate or multivariate sensitivity

analysis techniques in which model parameters vary over a specified range of values and the effect on the cost estimates is considered. If however the collection of patient level resource use data is incorporated within the trial design, statistical analysis of the trial data allows estimates of the parameters of the distribution of cost to be derived based on the observed population sample. Issues of concern then relate to whether tests of statistical significance should be performed at all or whether ranges should be presented instead given that although the clinical trial is considered a means of hypothesis testing, an economic evaluation undertaken alongside it will generally not be. That is, the trial is designed and powered to test a null hypothesis of no difference in effect but not an analogous hypothesis of no difference in cost between the alternative treatments even though the study design might incorporate cost data collection. Another concern relates to whether statistical tests, if performed, should be undertaken on the individual resource quantities, unit costs or both as each of these elements may be responsible for different degrees of variation in cost. Finally, a large literature has considered the methodology employed in calculating confidence intervals for the cost-effectiveness ratio statistic (see Briggs, 2001 for an overview of this literature).

The main finding which emerges from the preceding discussion is that although economic evaluation is increasingly undertaken alongside clinical trials, the proportion of economic studies that analyse data collected alongside a clinical trial in a manner that allows statistical inferences to be made regarding the distribution of cost is low. The primary reason why this is the case appears to be the lack of available cost data collected alongside a trial. Nevertheless, there are indications of an increasing tendency towards an incorporation of requirements for the collection of data on economic variables into the trial design. In these circumstances the issue becomes one of data analysis rather than data collection.

2.4. Cost analysis

Within the context of cost analysis a number of statistical problems arise, some of which require treatment by a specific statistical methodology while others stem from more general concerns. Before discussing the specific methodological problems most commonly encountered in the analysis of cost data drawn from a clinical trial, an important issue of a more general nature relates to the representativeness of a typical clinical trial population. Mullahy and Manning (1996) state for instance that even with randomisation there may still be selection bias arising from the individual's decision to participate, that eligibility criteria may lead to inappropriate exclusion of subsets of patients and that there may be selective compliance of treatments. For all these reasons the patient sample studied by the trial may not be representative of the true underlying treatment population. This is undoubtedly an important potential problem in any experimental design. A related issue raised by the same investigators is concerned with the omitted variables problem which imparts bias in the estimates. Viewed within a linear regression setting, the authors argue that the randomisation process must appropriately account for confounding variables. In other words, for the bias due to

omitted variables to be avoided, randomisation must ensure orthogonality between the regressors (i.e. the trial variables) and the unobservables (the confounding variables).

The statistical methodology employed in the analysis of cost and effectiveness trial data assumes in general that the process of randomisation both controls for confounding factors and ensures the representativeness of the trial population. Under this assumption the objective of the statistical analysis is to derive estimates of cost and effectiveness for the alternative interventions which are typically combined to form a ratio statistic defined as

$$R = \frac{(\mu_{C1} - \mu_{C2})}{(\mu_{E1} - \mu_{E2})}$$

where μ_{C1} and μ_{C2} are the average total costs of the two competing interventions (interventions 1 and 2) respectively, μ_{E1} and μ_{E2} denote the respective average total effects, and the ratio R referred to as the incremental cost effectiveness ratio (ICER) represents the incremental cost per unit of additional health outcome and is derived as the difference in the average total cost between the two alternative interventions relative to the difference in the average total effectiveness. As O'Brien et al (1994) state in circumstances where the sample is randomly drawn from the true underlying patient population, the ICER statistic can be estimated by replacing the unknown population parameters by their sample estimators and is given by

$$\hat{R} = \frac{(\hat{\mu}_{C1} - \hat{\mu}_{C2})}{(\hat{\mu}_{E1} - \hat{\mu}_{E2})}$$

The issue then becomes to obtain appropriate estimators for the mean values appearing in the expression above. The most obvious estimator would be the one based on the assumption of normality of the distribution of costs (and similarly of effects). However the assumption of normality for the distribution of costs is rarely valid. As a result such estimators of cost statistics will typically result in biased estimates. As the following sections show cost distributions tend to be particularly complex and their pattern depends to a substantial extent on the specific application. Consequently deriving estimators for the parameters of interest may be more appropriately achieved using an approach that does not impose specific distributional assumptions on cost. The choice of the estimator depends on a number of issues of a statistical nature the most common of which are presented below.

A regression framework is adopted as this provides a useful analytical device that allows elaboration of the problems and their proposed solutions. Thus considering a typical dataset which includes measurements on a set of covariates X the classical regression model relates cost to these covariates as follows

$$C = \beta' X + \varepsilon$$

where C is the random variable denoting cost, β is a $p \times 1$ vector of unknown regression parameters and ε is a zero-mean error term assumed normally distributed with constant variance σ^2. Letting i identify individuals the model for individual i is

$$C_i = \beta' X_i + \varepsilon_i \ , \ i = 1,...,n$$

The estimator for mean cost is then

$$\hat{C} = \hat{\beta}' \widetilde{X}$$

where $\hat{\beta}$ is the vector of the estimated regression parameters resulting from the least squares normal equation and $\widetilde{X}$ denotes the covariates vector evaluated at the mean values of the covariates. Being parametric, a regression approach may be preferred to a non-parametric estimator if for example there is interest in the extrapolation of costs beyond the end of the trial period or if interest is in assessing covariate effects on cost or in deriving cost predictions for different patient populations. On the other hand, a parametric approach typically imposes a particular distributional form on cost, in this case a normal distribution, which may not be justified. Additional concerns which influence the choice of the analytical methodology relate to the pattern of the observed cost data and include positive skewness in the cost observations, a substantial proportion of zero costs, censoring, missing data and lack of independence in the cost observations in situations where each individual in the sample contributes multiple data points over time. Each of these problems is considered in turn below. Heyse et al (2001) and Lipscomb et al (1998) cover similar ground. In what follows interest lies in deriving estimates of mean cost over the duration of the study as this forms the objective of the analysis in subsequent chapters.

2.4.1. Skewness in the cost data

The distribution of costs is typically positively skewed with a small number of patients incurring very high costs which implies that methods that assume a normal distribution and constant variances, such as the classical linear regression model may not be appropriate. In addressing this issue Briggs and Gray (1998) investigated the distributional properties of five datasets in detail. They found that all exhibited non-normality. To account for non-normality the authors considered a number of transformations, namely the log transform, the square root transform and the reciprocal transform. They state that although the data allow such transformations to be performed interpretation problems arise, an issue also raised by Manning (1998). The reason why such difficulties arise is that with transformed dependent variables, the mean response is not equal to the response of an individual with mean covariate values, but it equals the mean of the retransformed

estimate of the dependent variable which depends on the distribution of the covariates and not just their mean. Hence the property of the ordinary least squares regression estimates where the mean response equals the mean covariate values multiplied by the estimated regression parameters thus implying a straightforward interpretation for the coefficients disappears when the dependent variable is transformed. Despite these general problems, transformations are useful in regression problems and under conditions of positively skewed cost data the most commonly employed transformation is the logarithmic. This model is defined by

$$\ln(C_i + 1) = \beta' X_i + \varepsilon_i$$

where ε_i is normally distributed with mean 0 and constant variance σ^2, so that the unexplained part of C_i is now assumed to be lognormally distributed and the value of 1 (or any small positive value) is added to the observed costs to ensure that the logarithmic function is defined when there are zero cost observations in the data. The corresponding mean cost is

$$\hat{C}_i = \exp(\hat{\beta}' X_i) - 1$$

If however the errors ε_i in the above model are not normally distributed the above expression will lead to biased estimates of mean cost (Duan, 1983; Manning, 1998). To correct for this bias, Duan (1983) proposed a nonparametric alternative referred to as the smearing estimator, considered in detail in chapter 5, which provides consistent mean estimates even when the error distribution in the above model is normal. In addition the smearing estimator can also be used to account for this bias in the mean estimates for transformations other than the logarithmic.[1] The mean cost with smearing is then given by

$$\hat{C}_i = [\exp(\hat{\beta}' X_i)]S - 1$$

where $S = \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i}$ is the smearing estimator, $\hat{\varepsilon}_i$ is the ordinary least squares residual for the cost observation C_i and n is the total sample size.

2.4.2. Zero costs

Another issue arising in studies of treatment cost is that it is possible that a substantial number of patients have zero costs recorded as a result of the treatment lowering the probability of a cost generating event occurring. An estimator of average cost not taking account of this issue such as the

[1] The smearing estimator does not account for heteroscedasticity. If the error terms in the model $\ln(C_i + 1) = \beta' X_i + \varepsilon_i$ do not exhibit constant variance σ^2 a heteroscedastic smearing estimator is recommended (Manning, 1998).

simple arithmetic average would result in an underestimate of the mean cost. Similarly, a regression model including the zero cost observations would result in biased estimates. Duan et al (1983) and Lipscomb et al (1998) consider two-part models to address the problem generated by zero cost observations. The first part fits a logistic or probit model to the dichotomous variable defined by whether the patient incurred costs or not and thus predicts the probability that the patient has any cost generating events. The second part fits standard linear models to the cost data (possibly transformed) for those patients whose costs are positive. Expected values for individual patients are then derived by multiplying the two components together. The two-part model is given by

$$C_i = \begin{cases} 0 & \text{with probability } \rho_i \\ \beta' X_i + \varepsilon_i & \text{with probability } (1-\rho_i) \end{cases}$$

where $\rho_i = \dfrac{1}{1+\exp[-(\eta' X_i)]}$

with η being a vector of logistic regression coefficients and X being the same explanatory variables used in the cost regressions. The corresponding estimator for the individual's mean cost is

$$\hat{C}_i = \frac{\hat{\beta}' X_i}{1+\exp(\hat{\eta}' X_i)}$$

where the error terms in the regression model using the positive costs are zero-mean with constant variance σ^2. Applying the logarithmic transformation to the positive cost observations, the two-part model results in the following estimator for mean cost

$$\hat{C}_i = \frac{\exp(\hat{\beta}' X_i)}{1+\exp(\hat{\eta}' X_i)}$$

and with smearing, the two-part estimator of mean cost is

$$\hat{C}_i = \frac{[\exp(\hat{\beta}' X_i)]S}{1+\exp(\hat{\eta}' X_i)}$$

where the smearing factor S is defined above. With or without smearing, the logistic model component of the two-part model remains the same. Two-part models therefore allow consideration of factors which might affect both the probability of an individual having a cost generating event and the level of cost incurred given that there was a cost generating event. The importance of not accounting for a high proportion of zero costs in the data when estimating average cost can be seen by the following example. If a treatment lowers the probability that a patient has a cost generating

event but does not lower the level of incurred cost, ignoring the first issue might result in the treatment appearing to reduce the costs on average.

2.4.3. Censoring

There has been a long history of concern with censored data in analysing the effectiveness of treatment interventions within a clinical trial setting. More recently censoring has also been considered within the context of cost analysis. This section merely gives an introduction to the issue as the problem of censoring and the relevant literature are investigated in detail in chapters 4 and 5. Censoring arises because a number of observations fail to reach some pre-specified clinical endpoint, for instance death, before the end of the study period.[2] Until recently the problem of censored cost data had not received much attention and the analysis was based on the following two "naïve" estimators. The first, referred to as the uncensored cases estimator, only uses the uncensored cases in the estimation of mean cost while the second, referred to as the full-sample estimator, uses all cases but does not differentiate between censored and uncensored observations. Both these estimators will always be biased. The full-sample estimator is always biased downward because the costs incurred after censoring times are not accounted for whereas the uncensored-cases estimator is biased toward the costs of the patients with shorter survival times because larger survival times are more likely to be censored (Lin et al, 1997).

Censoring within the context of time-to-event data analysis has been dealt with using survival analysis techniques some of which make specific assumptions about the distribution of time-to-event while others are completely free of distributional assumptions. Initial attempts to account for censoring in the estimation of censored cost statistics applied parametric and non-parametric survival analysis methodology to cost (Dudley et al, 1993; Fenn et al, 1995, 1996). Underlying all the standard survival analysis approaches within the context of time to event data analysis is the assumption of independence between time to event and time to censoring. This implies that

[2] Although interest in this thesis is restricted to consideration of censoring in the analysis of cost data a more general point arises from the relationship between censoring and the alternative definitions of treatment effect. Within the context of an economic analysis conducted alongside a clinical trial the treatment effect is normally based on efficacy data drawn from the randomised controlled experiment. For the purposes of an economic analysis however the treatment effect should reflect the health gain attributable to the intervention under consideration as based on the general population likely to receive the treatment. In the treatment effects literature there are three basic parameters of interest. The local average treatment effect which is the average effect of treatment in the compliers, the global average treatment effect estimated based on the entire study population –i.e. including compliers and non-compliers- and the intent-to-treat parameter which is the average effect of treatment assignment (e.g. Angrist, Imbens and Rubin, 1996; Robins and Greenland, 1996). These alternative definitions of treatment effect could lead to differing levels of generalisability of the results of the economic analysis as well as differing levels of censoring and/or different censoring mechanisms. For example if the treatment is responsible for non-compliance and non-compliers are withdrawn from the study this will imply a non-random censoring mechanism requiring different estimation strategies for the parameters of interest from the ones employed under random censoring. In this case the mechanism that causes censoring might need to be explicitly modelled. As stated later in the thesis in most medical studies censoring is assumed to arise completely at random and consequently all estimators examined in chapters 4 and 5 have been developed based on the assumption of random censoring.

censored individuals do not constitute a particularly high or low risk subgroup so that removal of certain observations from the sample due to censoring leaves the remaining event times of the individuals who are still under observation having the same joint distribution as if there had been no censoring. A case of dependent censoring will arise for example if high-risk individuals tend to be censored because then the remaining individuals in the sample will represent a low risk population and the estimated hazard will underestimate the true hazard. In the context of cost analysis the independence assumption is translated into independence between the cost at event and the cost at censoring. Such an assumption would only be valid if all individuals accumulate cost at the same rate over time which implies a one-to-one mapping between time t and cost accumulated by time t. Typically however the rate of cost accrual varies amongst individuals with those in poorer health states using more resources and therefore incurring higher costs per unit of time. This means that individuals with a higher rate of cost accumulation will incur higher costs both at failure time and at the censoring time inducing positive correlation between cost at failure and cost at censoring. The assumption of independence between the variable of interest and its censoring variable is therefore violated and on this basis all traditional survival analysis techniques are inappropriate for estimating censored cost statistics (Lin et al, 1997; Etzioni et al, 1999).

Lipscomb et al (1998) applied the stratified variant of the Cox proportional hazards model[3] with time being the stratification variable in deriving estimates of patient cost within each stratum and although censoring was not present in their data, they suggested use of this model under censoring conditions on the basis that stratification by time circumvents the problem of dependent censoring as this specification imposes no constraint as to how cost varies over time within a given time period. Etzioni et al (1999), as presented in detail in chapter 5, criticise this approach on the basis that the accrual of costs at different rates across individuals will result in dependent censoring within the subgroups defined by the covariate levels even when the stratified variant of the proportional hazards model is adopted. An additional criticism relates to the proportionality assumption underlying the validity of the Cox regression model and as Etzioni et al (1999) show this assumption will not generally hold in circumstances when individuals accumulate costs at different rates.

As a result of the inappropriateness of the standard survival analysis techniques for censored cost data analysis, a number of alternative methods have been introduced all of which are considered in detail in chapters 4 and 5. Their main difference lies in that some are completely non-parametric whereas some make distributional assumptions about how cost varies with time or given a set of

[3] The non-stratified Cox proportional hazards model was used to analyse censored cost data by Dudley et al (1993) who also used the Weibull regression approach in the context of assessing the effect of clinical factors on the cost of coronary artery bypass graft surgery using alternative models. However their primary concern was not to address the issue of censoring as this was present in only 2.6% of the total number of observations in their data. Nevertheless they commented that these models could be used to account for higher levels of censoring although they expressed their concern with regards to the potential bias imparted by dependent censoring. Hay (1989) introduced the notion of using the non-stratified Cox proportional hazards model, as well as the Kaplan-Meier estimator, in estimating censored medical costs but did not proceed to estimation.

covariates. Lin et al (1997) proposed two non-parametric approaches in estimating censored cost statistics. The first uses only the individuals' total costs at the last contact dates while the second requires information on individual cost histories.[4] Both methods partition the study period into subintervals and derive an estimate of mean cost over the study period by weighting an estimate of interval-specific mean cost with an interval-specific Kaplan-Meier probability of survival or death depending on the method. Both estimators are shown to be consistent under the assumption that the censoring distribution is discrete. An alternative methodology was proposed by Bang and Tsiatis (2000) who introduced a class of non-parametric estimators that do not depend on the pattern of the censoring distribution. Of the two main estimators presented by these authors, like the Lin et al estimators, the first uses information on total costs whereas the second requires information on individual cost histories. These are supplemented by a further two estimators that attempt to improve efficiency by recapturing information lost due to censoring. Aside from the non-parametric methods mentioned above, a number of parametric regression models for estimating cost statistics under conditions of censoring have been introduced. Carides et al (2000) proposed a two-stage estimator which involves explicit parameterisation of the relationship between cost and failure time. Lin (2000) introduced a regression technique which directly assesses the effect of a set of covariates on cost whilst adjusting for censoring in the cost observations through use of a probability weight in the estimating equations. This approach can be used both when only total costs are available and when individual cost histories are recorded. In the latter case the technique addresses the issue of censoring while at the same time accounting for the correlation amongst repeated observations for the same subject. Furthermore the approach can accommodate covariate dependent censoring.

2.4.4. Missing data

In many applications data could be missing for a number of reasons. With specific reference to medical studies, measurements on some important prognostic factors could be missing for a subset of patients either for reasons unknown to the analyst and unrelated to the other observations in the sample being complete or for a reason relating to the nature of the factor, e.g. when obtaining the measurement requires some invasive and costly medical procedure as in the case of biopsy compared to a blood sample, or for a reason relating to the design of the study, e.g. when measurement on the variable requires visits to clinics at fixed points in time and therefore the availability of the data is determined by the timing of the visits specified in turn by the design of the study. If data are missing for reasons unrelated to the completeness of the remaining observations in the sample, the problem is referred to as the ignorable case because if efficiency is not the primary concern the estimation can be undertaken by ignoring the problem, that is by analysing the complete cases only (Griliches, 1986). If however the missing data are related to the phenomenon being

[4] Etzioni et al (1996), who was also one of the co-authors on the paper by Lin et al (1997), introduced one of these estimators, which was referred to as the Kaplan-Meier sample average (KMSA) estimator, but the authors did not study the estimator's statistical properties.

studied then analysing the information of the complete cases alone will result in estimators both biased and inefficient.

Rubin (1976) considered various mechanisms that may cause missing data and identified the weakest conditions under which it is appropriate to ignore the process that causes missing data. In situations where this process cannot be ignored modelling of the process is required. Assuming that interest lies in making statistical inferences about the parameter θ of the data with φ being the parameter of the missing data process, Rubin (1976) defines the weakest conditions under which it is appropriate to ignore the process that causes the missing data as follows. If statistical inference means sampling distribution inference,[5] then the process causing missing data can be ignored if the missing data are missing at random and the observed data are observed at random. "The missing data are missing at random if for each possible value of the parameter φ, the conditional probability of the observed pattern of missing data given the missing data and the value of the observed data is the same for all possible values of the missing data. The observed data are observed at random if for each possible value of the missing data and the parameter φ, the conditional probability of the observed pattern of missing data, given the missing data and the observed data, is the same for all possible values of the observed data. If statistical inference means direct-likelihood or Bayesian inference,[6] then the process responsible for the missing data can be ignored if the missing data are missing at random and the parameter φ is distinct from θ. The parameter φ is distinct from θ if there are no a priori ties between φ and θ either via parameter space restrictions or prior distributions." (Rubin, 1976, p.582).

Methods for handling missing data are related to the mechanism that causes the missing values. For example, this mechanism could be independent of the data, could depend on the value of the corresponding covariate for the case with the missing value or could depend on the value of more than one covariates. In most cases missing data are analysed based on the assumption that the missing data are missing at random in the sense of Rubin (1976) as stated above or on the stronger assumption that the missing data are missing completely at random, that is the process that causes the missing data does not depend on the values of the variables of interest in the data. A test statistic for testing whether the data are missing completely at random has been proposed by Little (1988). Under either of the above assumptions the mechanism causing the missing data can be ignored. If this is the case, which implies that the missing data mechanism is not modelled, the main approaches to analysing such data are briefly presented below.

---

[5] A sampling distribution inference is inference resulting solely from comparing the observed value of a statistic with the sampling distribution of that statistic under various hypothesised underlying distributions. Within the sampling distribution inference context, the parameters θ and φ have fixed hypothesised values.

[6] A direct-likelihood inference is inference resulting solely from ratios of the likelihood function for various values of the parameter. Within the context of direct-likelihood inference, the parameters θ and φ take values in a joint parameter space. A Bayesian inference is inference resulting solely from posterior distributions corresponding to specified prior distributions. Within the Bayesian inference context, the parameters θ and φ are random variables whose marginal distribution is specified by the product of the prior densities $p(\theta)p(\varphi|\theta)$.

Assuming the classic linear regression model given by

$$E(Y|X) = \beta_0 + \beta' X \text{, where } \text{var}(Y|X) = \sigma^2$$

the problem is to derive estimates of the parameters and their associated variance when some of the data are missing. Although missing data could be present both in the outcome variable Y as well as the regressors X, the methods discussed below concentrate on the case of missing X's but this does not restrict the outcome variable which could have some missing values as well. The main approaches for handling the problem of missing data for some of the regressors now follow.

The complete case analysis is a least squares method that results from minimising the sum of squared residuals with respect both to the parameters and the missing values. This is performed by assigning a zero residual to any incomplete case with missing values which effectively results in the removal of that case from the estimation of the regression parameters. Completely discarding the cases with missing values in this manner leads to loss of information which could be a substantial problem if the proportion of cases with missing values is high. On this basis the approach has been deemed useful only for providing a baseline method for comparisons. Another method is first to impute the missing X values and then to perform the regression of Y on X using the imputed X values either by ordinary least squares or by a weighted least squares regression that attaches lower weights to the incomplete cases. A simple approach for imputing the missing X values is to replace the missing X's by their unconditional sample means. This method will result in poor estimates if there is substantial correlation in the data. Assuming for example a positive correlation between Y and X such that high values of X are associated with high values of Y, if every missing X observation is replaced with its unconditional sample mean, then high values of the imputed X do not imply higher values of Y. An improvement is to use information on the observed X's in a case to impute the missing X's. This can be achieved by imputing for a missing X value for a particular case by linear regression on the observed X's in that case estimated from the complete cases. Another method is to impute the missing X by using both Y and the observed X's for imputation. Another approach is to assume a joint distribution for Y and X, a typical choice being a multivariate normal distribution, and to estimate the parameters of this distribution by maximum likelihood (Beale and Little, 1975). The estimates of the distribution parameters are then substituted into the regression model yielding maximum likelihood estimates of the regression parameters.[7]

The imputation methods stated above could result in low standard error estimates because errors in the imputations are not taken into account. A solution to this problem was proposed by Rubin (1978, 1978a) and is referred to as multiple imputation. According to this approach instead of

[7] For small sample inference a Bayesian approach could be preferable, according to which a prior is added to the likelihood and inference is based on the posterior distribution. Some applications of the Bayesian approach to multivariate problems with incomplete dependent variables are referenced by Little (1992) but application of the approach to regression problems with missing X's is limited.

imputing a single mean for each missing value, $N \geq 2$ values are drawn from the predictive distribution of the missing values given the remaining covariates and Y and then complete data analysis are repeated N times once with each imputation substituted. The final estimate of the parameter is then the sum of the parameter values obtained for each imputation divided by the number of imputations. Multiple imputation has the advantage that once the imputations are constructed analysis proceeds by complete-data methods. Multiple imputations could be predictions based on an explicit model or they could be based on an implicit model for the missing values. An example of the latter case is the hot deck imputations which match incomplete cases to complete cases using information on covariates and then impute values from the complete cases. All the above approaches assume data missing at random in the sense defined above and therefore do not model the missing data mechanism. Little (1992) covers similar ground as discussed above but in greater detail.

Robins, Rotnizky and Zhao (1994) proposed a class of estimators to account for a subset of regressors having missing values either by design or happenstance. Their estimating equations make use of the inverse of the probability of non-missingness and the estimators for the regression parameters of the conditional mean model are consistent under the following conditions. The data are missing at random in the sense of Rubin (1976) as defined above which implies that the probability of an observation being missing for individual i may depend on subject i's observed data including the outcome variable Y_i but not on the missing data, the probabilities of missingness are bounded away from zero and the probabilities of missingness are known or can be estimated parametrically. Comparisons between estimators belonging to their general class and estimators previously proposed to account for missingness, of which the main ones are given above, showed asymptotic equivalence between the estimators being compared each time but their respective estimators were always more efficient. Efficiency was improved by retrieving information lost due to missingness from subjects with incomplete data, an approach also adopted by Dagenais (1973), Gourieroux and Montfort (1981), Beale and Little (1975), Pepe and Fleming (1991), and Carroll and Wand (1991). Compared with the latter set of estimators, the estimators proposed by Robins et al (1994) are again more efficient.

Censoring and missing data are of the same nature as both these issues imply incomplete information. Consequently a large part of the statistical theory developed for handling missing data can also be applied to the problem of censoring. There are obviously circumstances in which the problem of missing data on some of the covariates can arise alongside the problem of censoring of the dependent variable. For example, Lin and Ying (1993) have considered this issue within the context of the Cox proportional hazards regression model where time to failure forming the outcome variable is subject to censoring. Under conditions where some of the regressors have missing values and although the model adjusts for censoring of the dependent variable, the authors state that not appropriately accounting for the missing covariate values will result in biased and inefficient parameter estimates. More specifically, they argue that discarding cases with missing

covariate values could result in considerable reduction in efficiency especially if the discarded cases correspond to uncensored failure times, including in the model only the covariates with complete measurements on every subject could distort the partial likelihood based inference, and imputing the missing values can impart considerable bias into the parameter estimators. They propose an estimator for the regression parameters which is consistent and asymptotically normal under the assumption of the data missing at random which is shown to be more efficient than the complete case analysis estimator especially when failures are infrequent. This illustrates the point that there may be situations in which missingness affects both the outcome variable, in this case through the censoring mechanism, and the covariates. Given that censoring is a missing data process the statistical theory for analysing missing covariate values can be applied to censored outcome variables. For example, the idea of using the inverse of the probability of non-missingness mentioned above has been used in applications where the outcome variable is censored (Koul et al, 1981; Robins & Rotnitzky, 1992; Robins, Rotnitzky & Zhao, 1994). In this context the weight is the inverse of the probability of an observation not being censored which is used in the estimating equations to derive consistent parameter estimates adjusted for censoring. Moreover, as will be shown in detail in chapters 4 and 5, the same idea underlies a number of approaches which attempt to provide unbiased and consistent estimators of cost statistics in the presence of censoring.

2.4.5. Dependency among multiple measurements on the same subject

An issue inherent in longitudinal data, that is data consisting of repeated measurements of the variables of interest on each individual usually obtained at various points in time, is the dependency among the repeated measurements for any subject. In such circumstances ordinary least squares regression is not an appropriate estimation procedure as the assumptions concerning the error terms are no longer valid.[8] The general procedure to analysing such data in the econometric and statistical literature is to adopt an alternative model to ordinary least squares, referred to as the generalised linear regression model, which accommodates more general patterns for the distribution of the disturbances.

In this context a class of generalised estimating equations for the regression parameters has been proposed by Liang and Zeger (1986) which result in consistent estimates of the regression parameters and their variances without requiring specification of the joint distribution of a subject's

[8] Partly in an attempt to resolve the problem of dependency among the multiple observations for each subject, Lipscomb et al (1998) used the stratified variant of the Cox proportional hazards model, as stated above, with time being the stratification variable in deriving estimates of patient cost within each stratum. They argue that a great advantage of this model is that it does not make any assumptions about the distributional form of the error terms, and as such it is likely to be amongst the best alternatives when interest lies in modelling complex distributions such as distributions of cost. In addition, they suggest that this model overcomes the problem of dependency between multiple observations on the same individual, as it assumes different baseline hazards for cost across different strata. Given however the criticism of this approach mentioned above and presented in detail in chapter 5, it seems unlikely that such a methodology would be useful.

observations. This approach has wide application if interest is in modelling the dependence of the outcome variable on the covariates and not in the pattern of change of the outcome variable over time. If this is the case, the approach models the marginal expectation of the outcome variable as a function of the covariates at each point in time whilst accounting for the correlation among the repeated measurements for a given subject by treating the time dependence among repeated measurements for an individual as a nuisance. When the time dependence is of primary importance then, as stated by the authors, models for the conditional distribution of the outcome variable given its past values would be more appropriate. The authors argue that if observations gained from different subjects are independent, the estimates of the regression parameters will be consistent even if the correlation structure is misspecified provided that the model for the marginal means of the outcome variable at each point in time is correctly specified.

More importantly, this approach can also be adopted in the event of some observations being missing, in which case the same results hold provided that data are missing completely at random in the sense of Rubin (1976). Within this framework of generalised linear regression Lin (2000) derived estimators for the regression parameters when repeated measurements on the outcome variable were obtained within the context of censored cost analysis. The approach is considered in detail in chapter 5.

2.5. Concluding remarks

The emphasis in this overview of the literature has been on the problems most commonly encountered in the collection and analysis of cost data generated from a clinical trial. These difficulties are partly due to the trial design and partly due to the general nature of cost data. The specific issues arising in any given analysis will also depend on the cost pattern observed in the particular application considered. In the case of the UKPDS data described in chapter 1 for instance, of the problems discussed above, censoring was the prominent feature of the data. Although this clinical dataset exhibits extreme levels of censoring reaching 82% by the trial end, it is likely that censoring will be a common if not the most common characteristic of cost data drawn from any clinical trial. The loss of information due to censoring leads to biased estimates of the statistics of interest. In the context of time to event data analysis a number of alternative estimators provide consistent estimates of failure time statistics under censoring conditions. In the context of cost to event data development of estimators attempting to adjust the estimates for censoring has been much more limited and very recent. Recently proposed estimators are investigated in detail in chapters 4 and 5 and their performance is empirically assessed under extreme censoring conditions using the UKPDS data. Prior to addressing the impact of censoring, the next chapter considers an issue relating to the collection rather than the statistical analysis of cost data.

Chapter 3

Cost collection: Centre specific versus average unit costs in multi-centre studies

3.1. Introduction

The analysis of cost data has received little attention compared to the discussion of the various applications of different forms of economic evaluation or the analysis of effectiveness data. As emphasised earlier the estimation of treatment costs is a crucial element in the calculation of an incremental cost-effectiveness ratio and yet there has been relatively limited examination of the concepts underlying the estimation procedure, measurement problems or the impact that reliance on different cost data has on the estimated level of cost. As noted in the review chapter a small number of studies have recently started to address these issues. This chapter considers some specific aspects of the collection of cost data and uses a simulation model to assess the impact that different levels of aggregation in the collection of unit cost data have on the estimation of treatment costs.

Dranove (1996) argues that identification of the relevant cost components to be included in a given study and their measurement will be determined by the perspective adopted for the analysis. In the vast majority of clinical trials the cost elements are confined to costs directly related to the treatment but as Dranove states these may also include social care or non-medical patient related costs. For a societal based analysis he points out that all cost elements must be identified. The analysis in this thesis is based on data collected alongside a clinical trial which only recorded cost information on the components directly related to the treatment being assessed. As such discussion is focussed on direct treatment costs from a health care providers perspective. This is not to diminish the other categories of cost that may be included in the incremental cost-effectiveness estimation, for example productivity costs, but rather reflects the general pattern of cost collection alongside a clinical trial according to which treatment cost is the major cost component. Assuming a multi-centre trial setting in which information on resource use is collected on a patient level, the question arises as to what would constitute the optimal unit cost measurement to be subsequently attached to the resource volumes in order to derive an estimate of the cost of treatment. More specifically the question addressed in this chapter is whether unit cost data should be collected on an individual trial centre basis or whether an average unit cost provides an adequate measure in the estimation of treatment cost.

Treatment costs are normally estimated in two stages (Johannesson, 1996). First the volumes of resource use attributable to the intervention under study are quantified and then the unit costs of the resources are attached to the resource volumes to derive estimates of treatment cost. Although in general the data could be obtained from a number of sources, within a trial design it is common that

resource use information is collected prospectively on a per patient basis. It is then necessary to identify the source of the unit cost data for the various resources utilised (Drummond, 1994). In theory unit costs could also be collected prospectively but in practice data restrictions lead to limited capture of economic data alongside a trial. This partly reflects the priorities set by the clinical investigators with respect to the data being collected within a trial design where recording of clinical information is the primary objective. Of the economic data required to enable economic assessment of the treatment under consideration unit costs are usually deemed of secondary importance and as such this data is normally obtained from alternative sources. In support of concentrating on volumes of resource use as opposed to resource unit costs in the prospective collection of data, it has been argued that for the purposes of making statistical inference on economic variables variation in total costs will be generally due to variation in resource volumes across treatment centres reflecting variation in clinical practice while unit costs are not expected to vary substantially across centres (Spiegelhalter et al, 1996). This however appears to represent an extreme view. Hospital charges such as Extra Contractual tariffs or Health Resource Groups (HRG) costs are known to vary across treatment centres and assuming that such charges to an extent reflect costs this information provides contrary evidence to the point stated above.

From a theoretical perspective the distinct consideration of resource volumes and unit costs may be seen as a useful analytical device in relating the concepts of production and cost functions. The production function represents the transformation of volumes of input into outcomes, while the cost function provides the relationship between total costs (input volumes multiplied by unit costs) and outcomes. A clinical trial may be characterised as an evaluation of the transformation of inputs (that is the bundle of resource volumes that make up the individual treatments) into outcomes (that is the health outcome) in an attempt to assess efficacy. In practice however if unit costs are subsequently attached to resource volumes in abstraction from consideration of the production relations this may lead to a miscalculation of total costs. Different types of production units may have different unit costs, for example teaching hospitals generally have higher unit costs than non-teaching hospitals. The scale of production may also give rise to different unit costs. Ignoring the relationship between resource volumes and unit costs may thus introduce bias into the total treatment cost calculations.

Moreover in theory cost ought to represent the true opportunity cost of the resource. Under perfectly operating markets this opportunity cost would be the minimum price required to keep the resource in its current use rather than some alternative usage. Such perfectly constructed prices are not available in the health care sector and a number of proxies are used. Dranove (1996) gives a useful discussion of the problems which arise in using such proxies covering such issues as the use of charges, the allocation of fixed costs and the difficulties imposed by the existence of joint products. Johannesson (1996) covers similar ground but also discusses the distortion introduced due to the existence of monopoly.

Such theoretical considerations complicate the process of collecting cost information. In identifying which costs ought to be included in an economic evaluation Gold et al (1996) argue that ease of measurement is not a justifiable criterion. Glick et al (2001) suggest, as noted in chapter 2, that pragmatism should dictate which costs ought to be collected within a trial. As they state "The best approach is to measure as many services as possible, because minimising the services that go unmeasured reduces the likelihood that differences among them will lead to study artefacts. However there are no *a priori* guidelines about how much data are enough, nor are there data on the incremental value of specific items in the economic case report form" (op. cite., p.121). Johnston et al (1999) on the other hand suggest that it is enough to identify the key cost generating events with the aim of minimising data collection while maximising the ability to measure the difference in cost. This would imply concentrating the costing exercise on the major components of total cost by studying the expected frequency of occurrence of different events and identifying resources associated with a high unit cost. Johnston et al (1999) also suggest that it may be possible to rely on a sub-sample of patients if it can be assumed that this will be truly representative of the full sample population.

Having identified the cost generating events to be included in the analysis the issue then becomes to specify the unit costs of the resource elements entering the cost estimation process particularly if the data under consideration are obtained from a number of individual settings possibly reflecting variations in clinical practice and consequently variations in resource use and unit costs. A major benefit of conducting a multi-centre study is that sample size requirements can be met more quickly as the number of centres increases. On the other hand it is recognised that multi-centre studies give rise to issues of heterogeneity and selection of the participating centres will have an impact on the transferability and generalisability of any accompanying economic analysis results. A small but growing literature addresses such general concerns (see for example Drummond et al, 1992; Coyle, 1996; Jonsson and Weinstein, 1997; O'Brien, 1996; Schulman et al, 1998; Wilke et al, 1998; Drummond and Pang, 2001). On the more specific issue of which are the appropriate data to collect within a multi-centre trial setting consideration has been more limited and opinions are divergent. For example, Johnston et al (1999) note that differences in resource use, unit costs and outcomes may occur and introduce economic bias and refer to Ellwein and Drummond (1996) who discuss the difficulty in rectifying such bias. Johnston et al (1999) also note that "In examining centre differences, a recognition of the potential relationship between resource use and unit cost is required because if unit costs are a function of resource use at individual centres, this implies that centre specific unit costs should be used" (op. cite., p.18) thus highlighting the importance of the centre specific production and cost function relations in determining how costs ought to be collected. Coyle and Drummond (1996) and the Australian guidelines on economic evaluation of pharmaceuticals (Commonwealth of Australia, 1995) both recommend that a single set of unit cost data, applied to centre specific resource volumes, may suffice. Glick et al (2001) give a simple illustrative example in which using an average of unit costs gives rise to different results compared

to using individual centre specific costs.[1] Drummond et al (1998) have argued that the costing process should be context specific and that there is little to discuss in terms of a general methodological approach.

Given the lack of consensus on the appropriate costing methodology to be adopted within a multi-centre trial setting, the analysis in this chapter attempts to address the issue of cost data collection within such a setting by assessing the performance of two alternative approaches in estimating average treatment costs which differ on the basis of the unit cost information entering the estimation process. Aside from the general interest in identifying a methodological approach in these circumstances, in the present study consideration of the appropriate unit costs to be attached to the resource volumes in deriving cost estimates was also motivated by the fact that the empirical dataset used for the analysis in the following chapters was drawn from a prospective multi-centre randomised controlled trial which collected patient level data on resource use but not centre specific unit costs. Consequently estimation of treatment costs was based on the resource volumes as recorded within the trial while unit costs were obtained from national statistics and from a number of centres participating in the trial. To assess the validity of such an approach in estimating treatment cost the approach together with its competing alternative are theoretically considered within the context of economic theory assuming an underlying production function and their performance is empirically gauged using a simulation experiment.

3.2. General setting and method

The pattern of cost data collection most commonly encountered in multi-centre clinical trial settings records information on resource use for all individuals in the trial but not information on the centre specific unit cost of the resource element. The patient cost is then calculated by attaching a standard unit cost to each resource item. The mean cost per patient is subsequently estimated by averaging across all patients in the trial without differentiating among the participating centres. This standard unit cost is normally some estimate of an average unit cost based for example on a sample of the centres participating in the trial or on published national data. As such, the estimated average cost reflects the variation in volumes of resource use across the participating centres but not the potential variation in the unit costs. An alternative approach is to combine centre specific unit cost data with resource volume data for each patient to calculate a treatment cost per patient before averaging across patients.

Although seldom explicitly considered in this context, economic theory suggests that there should be some defined predictable relationship between the mix of resource volumes used in producing treatments and the relative costs of these resources. Theory would suggest that if operating efficiently, each treatment centre would define technical efficiency with regard to a production

[1] This example post-dates the published paper (Raikou et al, 2000) which was based on this chapter.

function which related factor mix to maximum output. Moreover if the treatment centre were operating as a cost minimising firm, it would choose the least cost input combination to produce the desired level of output. In other words it would operate at the point of both technical and productive efficiency, i.e. the point on the isoquant where the slopes of the isoquant and the isocost curves are equal. If there were a change in the relative input prices, economic theory would predict that there would be a substitution away from the relatively more expensive input which would result in a change in the mix of resource volumes. If such conditions were to hold across all treatment centres in a multi-centre trial, the centres would display different mixes of resource volumes in producing the same level of output (such as one successfully treated case) as a response to the differing relative factor input costs that they might face.

If this occurred, the estimate of average cost per treated case across different treatment centres based on some average unit cost applied to centre specific resource volumes would differ from the estimate of average cost per treated case based on centre specific unit costs applied to the corresponding centre specific resource volumes. The first method would lead to biased estimates of the cost per treated case since by using the average unit cost of inputs it fails to take into account the substitution of relatively less expensive inputs for more expensive ones. If however resource use at individual treatment centres is not responsive to unit cost changes, such that there is no relationship between the variation in costs and variation in resource use, no difference in the estimates of average cost per treated case between the methods would be expected.

The present analysis attempts to consider more closely the implications of these two different methods of calculating treatment costs in multi-centre studies by exploring the theoretical reasons which might lead to systematic differences in the estimates derived using these two alternatives and by addressing the question of whether any such differences are affected by specific assumptions concerning the change in the relative input prices. The alternative costing methodologies are considered within the framework of economic theory and are empirically assessed by a simulation experiment designed to identify potential differences in the resultant estimates of treatment cost. The alternative approaches to cost estimation are initially explored under general circumstances where concern is with whether or not individual treatment centres respond to changes in the unit cost of resources in a manner that is consistent with economic theory. Subsequently, consideration is given to the response to changes in unit costs of factors when a specific change, an increase in the input price of one of the factors of production arising for instance from the introduction of a new health technology, is introduced.

With interest lying in the total costs of producing a specific level of output (say, a successfully treated case) across a number (n) of treatment centres the following assumptions are made. Each centre has only two inputs available to produce a successfully treated case, for example outpatient visits (denoted by V_a) and inpatient days (denoted by V_b). Each centre faces local unit costs for the two inputs of C_a and C_b respectively. Hence the total costs of generating a single unit of output

could be formulated in two ways consistent with the description above. First, by calculating an average unit cost to attach to each centre specific resource item resulting in the following expression for the cost of producing a treated case (averaged across centres):

$$\frac{1}{n}\sum_{i=1}^{n}\left\{V_{ai}\times\left(\frac{1}{n}\sum_{i=1}^{n}C_{ai}\right)+V_{bi}\times\left(\frac{1}{n}\sum_{i=1}^{n}C_{bi}\right)\right\}=\mathrm{E}(C_a)\mathrm{E}(V_a)+\mathrm{E}(C_b)\mathrm{E}(V_b) \tag{3.1}$$

Alternatively total costs could be calculated by using centre specific unit costs and volumes resulting to a treatment cost (averaged across centres) given by

$$\frac{1}{n}\sum_{i=1}^{n}\left\{(V_{ai}\times C_{ai})+(V_{bi}\times C_{bi})\right\}=\mathrm{E}(V_aC_a)+\mathrm{E}(V_bC_b) \tag{3.2}$$

This algebraic formulation of the two methods of cost estimation highlights the potential problem. In general the expectation of a function (equation 3.2) does not equal the function of the expectations (equation 3.1). The two expressions are equal only if the individual costs and volumes of each resource component are independent. This implies that if there is no predictable economic relationship between unit factor costs and volumes, no difference would be expected in the estimates gained from (3.1) and (3.2) above and it would be of little concern which of the two methods was employed. If however an underlying predictable relationship between the factor input volumes and their corresponding unit costs as dictated by economic theory and upheld by economic evaluation exists, then the two methods (3.1) and (3.2) would be expected to yield different estimates of the total cost per successfully treated case. Two distinct scenarios are therefore considered: (*a*) Treatment centres are assumed to operate according to the principles of economic theory and therefore respond to changes in the relative input prices through a predictable substitution of one input for the other, and (*b*) treatment centres operate on their production function but they do not respond to changes in relative input prices. These two situations are illustrated in terms of a typical economic model of production in Figure 3.1. The number of outpatient visits are plotted on the horizontal axis and the number of inpatient days on the vertical axis.

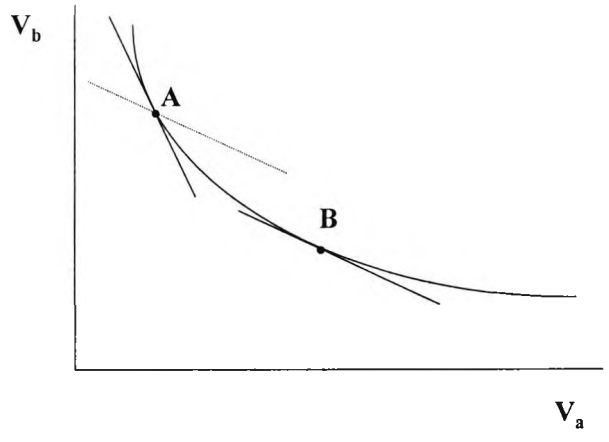


Figure 3.1. Different responses to price changes

Assuming initially that all centres operate at the point of both technical and productive efficiency, the optimal combination of inpatient days and outpatient visits is at point A in the figure, where the isocost line (with slope equal to the ratio of initial factor prices) is tangential to the isoquant. Introducing a change in relative input prices, for example due to the cost of an inpatient day becoming more expensive, with output being held constant so that focus is concentrated purely on the substitution effect gives rise to the following situation. A cost minimising centre in a manner consistent with economic theory would substitute away from inpatient days toward outpatient visits to a new optimal point B where the new ratio of input unit costs is tangential to the isoquant. By contrast, a centre which does not respond to the change in input unit costs would continue to employ the same combination of resources at point A. Hence the total costs of production for a centre which responds to changes in the input unit cost ratio will be lower than the total costs of production for a centre which does not respond to unit cost changes and that the magnitude of this difference in total cost will be determined by the extent of input substitution.

The elasticity of substitution provides a measure of the responsiveness of the factor input ratio (V_b / V_a) to changes in the unit costs of the factors (C_a / C_b) and is defined (in terms of the present example) as

$$\sigma \equiv \frac{\text{relative change in } (V_b / V_a)_{opt}}{\text{relative change in } (C_a / C_b)}$$

where $(V_b / V_a)_{opt}$ denotes the optimal factor volume ratio at the point of both technical and productive efficiency and $\sigma \in [0, \infty)$. The limiting case of $\sigma = 0$ is where the two inputs must be used in a fixed proportion as complements to each other. The other limiting case, with σ infinite, is where the two inputs are perfect substitutes for each other. If $\sigma > 1$ the elasticity of substitution is said to be elastic while if $\sigma < 1$ the elasticity of substitution is inelastic.

The predictions of the above theory are that where centres are assumed to respond to unit cost changes (scenario *a*) there will be a systematic difference between the two methods of calculating costs per treated case and the magnitude of the difference will be related to the elasticity of substitution. By contrast, where centres are assumed to be unresponsive to changes in the input cost ratio (scenario *b*) there will be no systematic difference between the two methods of cost calculation.

A simulation experiment was designed to test these predictions and to address the question of what degree of substitutability would generate a statistically significant difference between the two methods. It was assumed that the unit costs of the two inputs (C_a and C_b) varied randomly across individual treatment centres. In order to determine production responses to unit cost changes in a manner that is consistent with economic theory a specific production function was defined across the treatment centres. The constant elasticity of substitution (CES) production function was chosen

as it allows concentration on the substitution effect of the relative price change by keeping output constant and at the same time enables the role of the elasticity of substitution to be studied (see for example Heathfield, 1971). The CES production function is defined (in terms of the particular example) as

$$Q = A\left[\delta V_b^{-\rho} + (1-\delta)V_a^{-\rho}\right]^{-1/\rho}$$

where A is the efficiency parameter $(A > 0)$ and serves as an indicator of the state of technology, δ is the distribution parameter $(0 < \delta < 1)$ and relates to the relative factor shares in the product, and ρ is the substitution parameter $(-1 < \rho \neq 0)$ and determines the value of the elasticity of substitution such that $\sigma = \frac{1}{1+\rho}$. By definition therefore the elasticity of substitution remains constant along the same isoquant and by choosing appropriate values for ρ the degree of substitutability is varied. It should be noted that the CES production function does not allow the elasticity of substitution to attain a value of 1 given that the function is undefined for $\rho = 0$. Nevertheless it can be demonstrated that as $\rho \to 0$ the CES function approaches the Cobb-Douglas production function which is characterised by a unitary elasticity of substitution (see for example Chiang, 1984) and is defined as

$$\lim_{\rho \to 0} Q = A V_b^{\delta} V_a^{1-\delta}$$

In the simulation experiment the CES production function (and the Cobb-Douglas production function in the special case of $\sigma = 1$) were employed to calculate the initial factor volumes for the centres given their initial ratio of unit costs. This guaranteed that the production process is well behaved with regards to satisfying the least cost input combination condition according to which the marginal product ratio (i.e. the slope of the isoquant) is equal to the factor price ratio.

A given percentage change was subsequently initiated in relative unit costs. In scenario *a* where centres respond to unit cost changes the new optimal position on the isoquant was calculated with regards to the new ratio of unit costs. Total costs were then calculated using the two methods described above and the difference between the two methods was recorded. In order to examine the influence of different distributions of factor input unit costs on the results various distributions were prespecified. The simulation experiment was therefore undertaken assuming that input factor unit costs were drawn from uniform, normal and logistic distributions. The same simulations were undertaken for scenario *b* where the treatment centres are assumed not to respond to the change in unit costs, i.e. centres remained on the initial point on the isoquant although facing the new relative input prices. Again total costs were calculated by both methods and the difference was recorded.

For both scenarios these simulations of the difference between the two total cost estimates were repeated 1000 times. In this way an empirical estimate of the sampling distribution of the difference between the two methods was generated. Where no systematic difference between the two methods is expected, the distribution should be centred around zero with approximately 50% of observations above and below zero. By contrast, where a systematic difference is expected, the strength of that difference can be judged by the proportion of results that lie either side of zero. This is akin to the traditional p-value in hypothesis testing, such that if less than 25 observations out of 1000 (2.5%) lie above zero (or alternatively if less than 25 observations out of 1000 lie below zero), a 95% confidence interval would not include zero and the null hypothesis of no difference between the methods would be rejected at the 5% level. In the case under consideration here, a one-sided hypothesis test is more appropriate such that the null hypothesis of no difference would be rejected if less than 50 observations out of 1000 (5%) lie below zero.

In addition for scenario *a* the simulation experiment was repeated for a number of different values of the elasticity of substitution ranging between $\sigma = 0.1$ and $\sigma = 10$ relating to nearly perfect complements or highly substitutable inputs. This allowed consideration of the effect that the degree of input substitutability has on the significance of the difference between the two methods of estimating the cost per treated case. Finally, it was assumed that the centres varied in their response to unit cost changes in a stochastic rather than deterministic manner. A normal distribution was thus imposed on the centre factor volumes after they had responded to changes in relative unit costs. The mean of this normal distribution was the mix of the factor input volumes which would have emerged given the deterministic response, that is assuming no uncertainty in the response. The variance of the distribution was set at increasingly high levels to mimic greater degrees of uncertainty in the response to changes in relative unit costs.

3.3. Results

Considering first the case of independence between the ratio of the input volumes and the ratio of the factor costs, that is scenario *b*, the results of the simulation experiment showed the differences in the estimates being normally distributed around zero indicating that, as predicted, there is no systematic difference between the two methods of cost calculation. Turning to scenario *a* where factor volumes respond to changes in relative unit costs as dictated by the production function and the elasticity of substitution different findings are reached. Figure 3.2 outlines the results where the three distributions relate to those from which the unit costs are drawn and the individual centres factor input responses to the change in relative unit costs are at this stage deterministic. That is, there is no uncertainty surrounding the new optimal mix of the factor inputs. As can be seen even at relatively low values of the elasticity of substitution and irrespective of the assumed underlying distribution for unit costs there is a significant difference between the average treatment costs calculated by the two methods as indicated by the p-values. These p-values are based on testing the

mean differences in the calculated costs. In other words once it is assumed that treatment centres respond to changes in the unit costs of inputs in the manner dictated by economic theory, as is implicit in economic evaluation, the two methods of calculating treatment costs give rise to statistically different estimates.

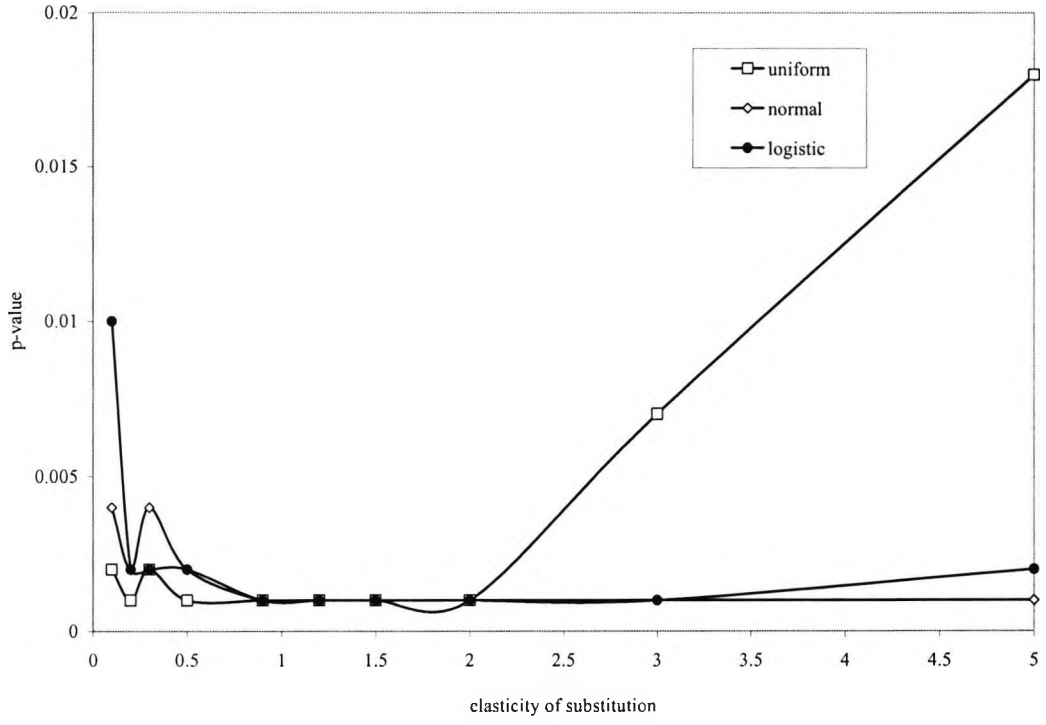


Figure 3.2. Deterministic relationship between unit costs and volumes

Concentrating on the case where the centres respond to the relative change in unit costs in a stochastic manner, the results are consistent with the previous findings. As shown in Figure 3.3, where the increasing degree of uncertainty is represented by an increasing coefficient of variation (*cv*), it can be seen that when the level of uncertainty in the response is relatively low the method of cost calculation appears to matter. Thus when the coefficient of variation is below 0.2 the mean values of the total cost are statistically different at the 10% level for all values of the elasticity of substitution. However as the uncertainty in the response increases, as measured by the coefficient of variation, there is a tendency towards a situation that replicates independence between input volumes and unit costs, as in scenario *b*, with the two methods giving similar estimates. It should be noted that even a coefficient of variation with a relatively small value (e.g. 0.2) expresses a wide dispersion around a mean response.

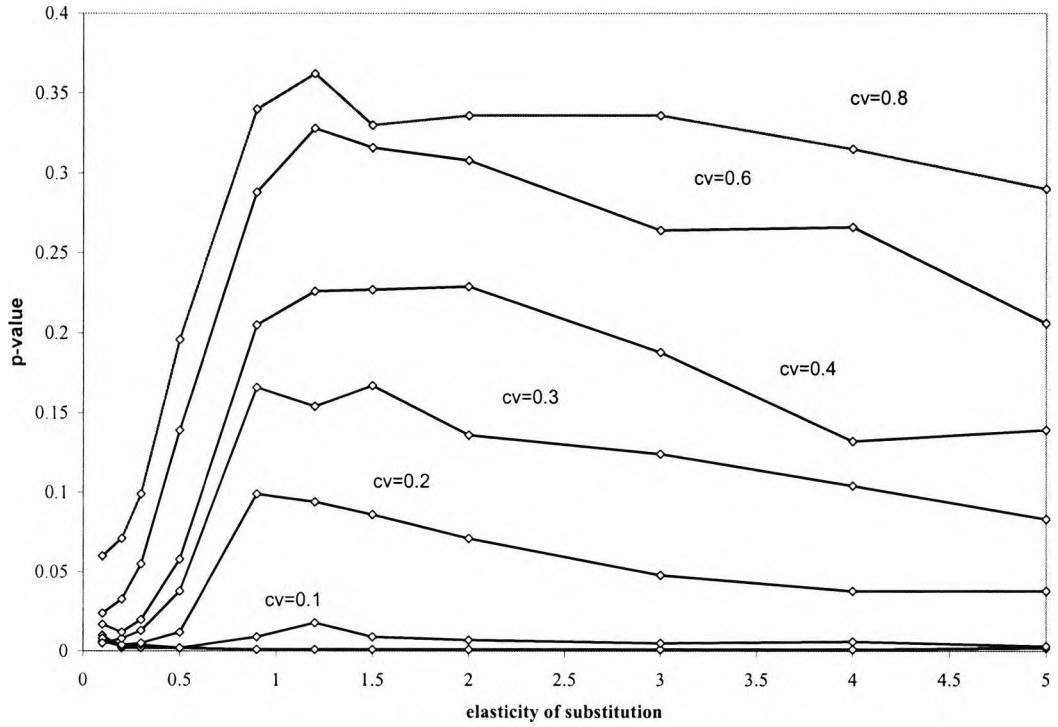


Figure 3.3. Statistical significance for different values of elasticity of substitution and stochastic response to changes in unit costs (input unit costs drawn from a normal distribution)

3.4. Discussion

To date limited attention has been given to the identification and collection of cost data drawn from clinical trials for the purposes of economic evaluation. The analysis in this chapter has attempted to assess the impact of different methods of cost collection on the estimated treatment costs within a multi-centre trial setting by exploring potential differences in the resultant estimates between two alternative methods of cost calculation. The first alternative and the most commonly adopted in practice attaches an average unit cost to the individual resource volumes whereas the second attaches individual centre specific unit costs to resource elements. The conceptual exposition above shows that under the assumption that treatment centres behave in a manner consistent with production theory as is implicitly required by economic evaluation, attaching an average unit cost to the resource volumes would lead to biased cost estimates. This prediction was confirmed by the results of the simulation experiment for different statistical distributions of unit costs and for a range of degrees of input substitutability representing nearly perfect substitutes to nearly perfect complements. Moreover the same results were reached when the response to relative changes of unit costs was assumed to have a stochastic component. Only under conditions of absence of response to changes in relative input prices does the choice of costing methodology appear not to have an impact. The present analysis has not taken the output effect of a relative input price change into account. In reality relative price changes would give rise to both a substitution and an output effect. Consideration here was restricted to the former in order to isolate the impact of input

substitutability on production responses to changes in relative unit costs and subsequently on the two methods of cost calculation. Focussing on the substitution effect however does not affect the generality of the findings.

These findings emphasise a need for more detailed information on the production process in the health care sector. The results reached in this chapter indicate that the method of calculating treatment costs that ignores the substitution effect will result in biased estimates on the assumption that treatment centres operate in a rational economic manner. In reality however little is known about the degree of factor substitution in this sector. What is known is that there is considerable variation in treatment patterns across centres but the sources and mechanisms of this variation are not well understood. A starting point in gaining better understanding of the health care production process could be achieved by placing greater importance on the measurement and reporting of the unit costs of the resources used to produce a given health outcome.

Chapter 4

Cost analysis: Non-parametric estimators of treatment cost under conditions of censoring

4.1. Introduction

Increasingly cost information is collected alongside clinical trials as the basis for cost-effectiveness analysis. As discussed in chapter 2 a number of problems arise in the analysis of such data including issues concerning missing observations, skewed data and censoring. Discussion here concentrates on the last issue. There are various types of censoring, such as right censoring or left censoring. Left censoring, which is not as common in clinical trials, involves a loss of information due to individual observations entering the study at different points of progression to end-point. This chapter does not address this issue. Right censoring occurs whenever some individuals are not observed for the full duration of interest which results in information being incomplete for these patients. Consequently, estimators of statistics of interest are biased if no account is taken of censoring with the bias increasing as the degree of censoring increases. Parametric or non-parametric modelling approaches may be used to adjust the estimators for this loss of information which is observed both in effectiveness and cost data. Both parametric and non-parametric approaches have been applied to the analysis of effectiveness data when effectiveness is assessed in terms of time to event yielding estimators that appropriately adjust the estimates for censoring. It is only recently however that attention has turned to the issue of censored cost data. Given that parametric approaches involve explicit assumptions regarding the distribution of costs which may not be justified by the data, initial attempts to adjust estimators of cost statistics for censoring involved application of non-parametric survival analysis techniques to cost data. The Kaplan-Meier estimator was the first approach used in this context (Fenn et al, 1995), but this was shown to result in biased estimates of cost due to the violation of independence between the cost at event and the cost at censoring times (Lin et al, 1997; Etzioni et al, 1999). Another two estimators have also been used to provide estimates of mean cost in the presence of censoring which are referred to as "naïve" estimators in the literature because the first, referred to as the uncensored cases estimator, only uses the uncensored cases in the estimation of mean cost, while the second, referred to as the full-sample estimator, uses all cases but does not differentiate between censored and uncensored observations. Both these estimators will always be biased. The full-sample estimator is always biased downward because the costs incurred after censoring times are not accounted for whereas the uncensored-cases estimator is biased toward the costs of the patients with shorter survival times because larger survival times are more likely to be censored (Lin et al, 1997).

Lin et al (1997) acknowledge these difficulties and propose a method which attempts to resolve these issues. They introduce two estimators of mean cost under conditions of censoring which rely on the study period being partitioned into a number of subintervals such that censored observations occur at the boundaries of these intervals. Under such circumstances, the approach is shown to give consistent estimators of average cost and the associated variances are analytically derived. Hence the validity of the approach depends to an extent on the pattern of the censoring distribution being of such a form to allow censoring times to correspond to the boundaries of the intervals of the partition. There is no a priori reason however to expect censoring to conform to any such pattern and therefore in most applications consistency will be violated to some degree. This limitation has led to a further set of estimators proposed by Bang and Tsiatis (2000). Their estimators are shown to be consistent regardless of the censoring pattern and their variances are analytically derived.

All the above non-parametric estimators of cost together with their properties and underlying assumptions are presented in this chapter using a common analytical framework. Their performance is empirically assessed under extreme censoring conditions using the UKPDS data introduced in chapter 1. While the theoretical properties of these recently proposed estimators have been studied by Lin et al (1997) and by Bang and Tsiatis (2000), their performance has not been assessed under conditions of extreme censoring using real data. In this chapter the estimators are investigated using a real clinical dataset which exhibits levels of censoring of 82 per cent.

The estimators' theoretical properties have been investigated using the theory of stochastic processes as applied to the study of time-to-event data. Stochastic processes are often used to model clinical data collected over a period of time and in particular data counting the number of events over time. The standard application of the counting process approach to survival analysis is a powerful tool in deriving statistics of interest as well as in studying their properties in the presence of censoring (Gill, 1980; Fleming and Harrington, 1991; Andersen et al, 1993). As will be shown later in the chapter, the same approach is equally powerful when applied to the study of cost-to-event data under conditions of censoring as it provides the analytical framework in which the asymptotic properties of the estimators of cost statistics are being established. More specifically, use of this particular analytical approach allows the notion of the time element in the cost observations to be captured, censoring to be incorporated, variance estimators to be derived and convergence and asymptotic normality of the statistics of interest to be proven by invoking martingale convergence theorems. It is important therefore to present the general setting for the analysis viewed within the counting process framework as applied to the study of time-to-event data as the same concepts underlie the study of cost-to-event data as undertaken by the approaches of interest. Thus the following section provides a general introduction to stochastic processes, their relationship to counting processes and stochastic integration, as well as the general application of martingales to counting processes with specific reference to martingale theorems used in the study of the statistics of interest. This section draws on the work by Gill (1980), Fleming and Harrington (1991) and Andersen et al (1993). Having established the conceptual context, the set of non-

parametric estimators of cost together with the assumptions underlying their validity are then presented. The main analysis, whose aim is to assess the estimators' performance under extreme conditions, is presented in the following section which reports the results derived from the application of the cost estimators to the UKPDS dataset. A number of problems are identified within this part of the analysis which are subsequently investigated using subsets of the original data as well as an artificially generated dataset with varying degrees of censoring. The final part of the analysis derives variance estimates using the bootstrap approach as an alternative to the theoretically derived formulae for the asymptotic variance estimators in an attempt to determine the validity of the underlying assumption of asymptotic normality when the estimators are applied to the smaller sample sizes observed in real medical data.

4.2. Analytical framework

4.2.1. General setting

The aim of the approaches to be presented later in the chapter is to derive an estimate of the mean total cost $\mu = E(M)$ and its variance over a specified period when the data is right censored, where the random variable M denotes the total cost for a patient during some specified time T and E denotes expectation. The distribution of the random variable T is assumed continuous over $(0, L]$ where L denotes the upper bound of T, i.e. the maximum time for which each patient is evaluated. In that case M is the total cost incurred by a patient up to a maximum of L units of time. If all patients were observed for a minimum of L units of time then complete information on M would be available and the mean cost would be estimated by the average of the costs for each patient. In most cases however cost information is incomplete due to censoring. Defining therefore a potential time to censoring denoted by U and letting T denote the time to death, the observables from a study in the presence of censoring are $X = \min(T,U)$, i.e. the last contact date; $\delta = I(T \le U)$, where $I(\cdot)$ is the indicator function taking the value of 1 when the argument is true (i.e. if the observation is uncensored) and zero otherwise; the cost accrued up to time X and other intermediate cost history for each subject, i.e. $M^H(t) = \{M(u), u \le t\}$, where $M^H(t)$ denotes the cost history up to time t, $M = M(T)$, with $M(u)$ being the known accumulated cost up to time u and u denoting points in time at which cost information becomes available. The observable data for n individuals are then the independent and identically distributed random vectors

$$\left\{X_i = \min(T_i,U_i),\ \ \delta_i = I(T_i \le U_i),\ \ M_i^H(X_i)\right\}, i = 1,...,n$$

where i identifies an individual.

Regardless of whether censoring is present or not, when studying time to event data interest lies in the distribution of the non-negative continuous random variable T denoting the time to event with cumulative distribution function $F(t)$ given as

$$F(t) = pr(T \le t) = \int_0^t f(u)du$$

where $f(t)$ is the associated probability density function $f(t) = \lim_{\Delta t \to 0} \frac{pr(t \le T \le t + \Delta t)}{\Delta t}$.

In the absence of censoring, a non-parametric estimator for $F(t)$ would be based on the empirical cumulative distribution function. The associated survival function, that is the probability that the individual will survive at least until time *t*, is given as

$$S(t) = 1 - F(t) = pr(T > t)$$

and the related hazard rate, that is the probability that the individual will die in the next short interval Δt given that he has survived until time *t*, is given as

$$\begin{aligned}\lambda(t) &= \lim_{\Delta t \to 0} \frac{pr(t \le T \le t + \Delta t | T \ge t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{pr[(t \le T \le t + \Delta t) \cap (T \ge t)]}{pr(T \ge t)} \\ &= \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{pr(t \le T \le t + \Delta t)}{pr(T \ge t)} = \lim_{\Delta t \to 0} \frac{1}{\Delta t} \frac{F(t + \Delta t) - F(t)}{S(t)} \\ &= \frac{1}{S(t)} \frac{dF(t)}{dt} = \frac{f(t)}{S(t)}\end{aligned}$$

or

$$\lambda(t) = \frac{1}{S(t)} \frac{dF(t)}{dt} = \frac{1}{S(t)} \frac{d[1 - S(t)]}{dt} = -\frac{d[\ln S(t)]}{dt}$$

which shows the relationship between the conditional probability of death and the unconditional probability of survival. An alternative expression for the above relationship is then derived as

$$\ln S(t) = -\int_0^t \lambda(u)du \Leftrightarrow S(t) = e^{-\int_0^t \lambda(u)du} \Leftrightarrow S(t) = e^{-\Lambda(t)}$$

where $\Lambda(t) = \int_0^t \lambda(u)du$ is the integrated or cumulative hazard function for the failure time.

4.2.1.1. Stochastic processes and filtration

Adopting the counting process analytical framework allows the properties of failure time statistics to be established both in the absence of censoring and when censoring is present. Having stated the importance of considering this analytical framework more formally for the purpose of the study of cost estimators to be defined subsequently, exposition starts with the case of censored time-to-event

data. Viewed within the counting process framework, such time to event data in the presence of censoring can be modelled by the counting processes $N_i(t) = I(X_i \le t, \delta_i = 1)$ with $N(t) = \sum_{i=1}^{n} N_i(t)$ counting the number of individuals dying over time, $N_i^c(t) = I(X_i \le t, \delta_i = 0)$ with $N^c(t) = \sum_{i=1}^{n} N_i^c(t)$ counting the number of individuals censored over time, and the accumulated information over time these processes generate referred to as filtration given by $\mathcal{F}_t = \sigma\{N(u), N^c(u), 0 \le u \le t, i = 1,...,n\}$ and representing the increasing information over time on the individuals' survival or censoring up to and including time *t*.

To consider formally the processes and the associated filtration defined above, a convenient starting point is the concept of a stochastic process given that a counting process is itself a stochastic process. Following Fleming and Harrington (1991), a stochastic process is a family of random variables $X = \{X(t) : t \in \Gamma\}$ indexed by a set Γ, all defined on the same probability space $(\Omega, \mathcal{F}, \mathrm{P})$.[1] The set Γ indexes time and is usually either $\{0, 1, 2,...\}$ defining discrete time processes or $[0, \infty)$ defining continuous time processes. Given that a random variable is a function defined on a sample space of outcomes, Ω, it follows that a random process $\{X(t) : t \in \Gamma\}$ is a function of two arguments $\{X(t,\omega), t \in \Gamma, \omega \in \Omega\}$. For a fixed $t = t_k$, $X(t_k, \omega) = X_k(\omega)$ is a random variable with ω varying over the sample space Ω while for a fixed sample point $\omega_i \in \Omega$, $X(t, \omega_i) = X_i(t)$ is a single function of time *t*, called a sample path or a realisation of the process. Rather than study the properties of the random variable $X(t)$ for fixed *t*, the modern approach to the general theory of processes relies on properties of the sample path $X(t, \omega_i), t \in \Gamma$ for fixed ω. The processes studied later in the chapter are limited to the index set $\Gamma = \Re^+ = [0, \infty)$ and are denoted as $\{X(t) : t \ge 0\}$ for fixed ω. When a process is said to have a particular property

---

[1] A probability space $(\Omega, \mathcal{F}, \mathrm{P})$ consists of a space of outcomes Ω with each outcome denoted generically by ω, a selected σ-algebra of events $\mathcal{F}$ in Ω, and a measure P defined on $\mathcal{F}$ such that $\mathrm{P}(\Omega) = 1$ where the measure P is called the probability. An event A is said to occur almost surely (a.s.) whenever $\mathrm{P}(\mathrm{A}) = 1$. As noted by Gihman and Skorohod (1974) the first fundamental assumption when formalising the notions of probability theory is that the results of a collection of experiments under investigation in a given situation can be described by means of a certain set Ω. Furthermore, an experiment is completely characterised by the class of those events, subsets of Ω, such that each time one can assert whether a particular outcome occurred or not during the given experiment. Although any arbitrary subset of Ω forms an event, the class of events which characterises any experiment in the sense mentioned above is always assumed to form a σ-algebra of events. A class of events is called an algebra of events if it contains the certain event, i.e. the space Ω, the impossible event, i.e. the empty set $\varnothing$, and together with each pair of events A and B belonging to the class, i.e. A and B subsets of Ω, their sum, i.e. their union, and the contrary event of A, i.e. the complement of set A. An algebra of events which contains a sequence of events together with their sum is called a σ-algebra. The space Ω along with the σ-algebra of sets $\mathcal{F}$ defined on it is called the measurable space $\{\Omega, \mathcal{F}\}$ and the subsets of Ω belonging to $\mathcal{F}$ are called $\mathcal{F}$-measurable sets ($\mathcal{F}$-measurable events). With respect to the measurable space $\{\Omega, \mathcal{F}\}$ any given stochastic experiment is completely characterised by the class of events $\mathcal{F}_t$ observed during this experiment and as such any stochastic experiment is determined by a certain σ-algebra $\mathcal{F}_t$ of $\mathcal{F}$-measurable events where $\mathcal{F}_t$ represents the history of the experiment.

such as being continuous, or left- or right-continuous, or of bounded variation, or increasing, it means that the set of sample paths with the corresponding property has probability one. In other words, almost all of its sample paths have the particular property.

As already implied central to the theory of stochastic processes is the notion of filtration or history. Letting $(\Omega, \mathcal{F}, P)$ be a probability space, a filtration $\{\mathcal{F}_t : t \in \Gamma\}$ is an increasing right-continuous family of σ-algebras of $\mathcal{F}$, that is:

$$\mathcal{F}_s \subseteq \mathcal{F}_t \subseteq \mathcal{F} \quad \text{for all } s < t \qquad (\mathcal{F}_t \text{ is increasing})$$
$$\mathcal{F}_s = \bigcap_{t>s} \mathcal{F}_t \quad \text{for all } s \qquad (\mathcal{F}_t \text{ is right - continuous})$$
$$A \subset B \in \mathcal{F}, P(B) = 0 \Rightarrow A \in \mathcal{F}_0 \qquad (\mathcal{F}_t \text{ is complete})$$

Thus the σ-algebra $\mathcal{F}_t$ contains all events whose occurrence or not is fixed by time t. In other words, it denotes the history of the experiment under investigation, in this case the clinical study, up to and including time t. The first condition expresses the fact that as time evolves, new events may occur, whereas the other two conditions are technical ones and hold for all filtrations considered in the applications of interest with the second condition implying that $\mathcal{F}_t$ contains all $\mathcal{F}_s$ for $s < t$ and the third condition extending the filtration to time zero. There is also a pre-t σ-algebra $\mathcal{F}_{t-}$, the smallest σ-algebra containing all $\mathcal{F}_s, s < t$, which contains events fixed strictly before t. A filtration therefore models information that is increasing with time. A stochastic process X is said to be adapted to the filtration $\mathcal{F}_t$ if $X(t)$ is $\mathcal{F}_t$ measurable for each t, that is the realised value $X(t,\omega)$ of $X(t)$ for a given ω can be determined based on the accumulated information up to time t contained in $\mathcal{F}_t$. A filtration can also be described as being generated by a stochastic process X. This means that $\mathcal{F}_t$ is the σ-algebra generated by $X(s), s \le t$ and this in turn implies that $\mathcal{F}_{t-}$ is generated by $X(s), s < t$. In particular, processes all of whose paths are left-continuous or right-continuous as the processes studied in the applications of interest are measurable for each t which ensures that they are adapted to a filtration $\mathcal{F}_t$. In particular, a filtration generated by a right-continuous jump process, as is for instance the process $\{N(t) : t \ge 0\}$ defined above, is right-continuous.

A counting process $\{N(t) : t \ge 0\}$ is a stochastic process adapted to a filtration $\{\mathcal{F}_t : t \ge 0\}$ with $N(0) = 0$ and $N(t) < \infty$ almost surely and whose paths are right-continuous with probability one, piecewise constant, and have only jump discontinuities with jumps of size +1. By its definition, $N(t)$ represents the total number of events that have occurred in the interval $(0, t]$, $N(t) \ge 0$, $N(t)$ is integer valued, $N(s) \le N(t)$ if $s < t$, and $N(t) - N(s)$ equals the number of events that have occurred in the interval $(s, t]$. As such, the counting processes $\{N(t) : t \ge 0\}$ and $\{N^c(t) : t \ge 0\}$ defined in the example above count the number of deaths and the number of censored individuals in the interval $(0, t]$ respectively and being right-continuous they are adapted to the filtration

$\mathcal{F}_t = \sigma\{N(u), N^c(u), 0 \le u \le t\}$ which contains the increasing information over time on failure and censoring up to and including time t generated by the processes.

4.2.1.2. Stochastic integration

Having presented the statistical model used to express time to event data, the following sections address specific aspects of the theory of stochastic processes which are essential to the study of the properties of the statistics of interest. Counting processes as applied to survival analysis involve stochastic integration, that is the forming of the integral of one stochastic process with respect to another. Within the context of this approach, the stochastic processes X and Y entering the integrals for a given $\omega \in \Omega$ satisfy such properties that the integral $\int XdY$, which is a stochastic process itself, is an ordinary Lebesgue-Stieltjes integral over a given time interval. More specifically, to ensure that $\int_s^t X(\cdot,\omega)dY(\cdot,\omega)$ is well defined as a Lebesgue-Stieltjes integral for each ω the processes X and Y must satisfy the following measurability and sample path properties. The sample path of X for each ω must be a measurable function on interval $(s,t]$. The sample paths of Y for almost all ω must be right-continuous with left-hand limits and of locally bounded variation on $(s,t]$, i.e. $\int_s^t |dY(u)|$ must be finite for all $t \in \Gamma$, for almost all ω. Such a process Y is called a finite variation process and the process $\int |dY|$ is its (total) variation process. Any bounded variation process can be written as the difference $Y_1 - Y_2$ of two non-decreasing processes and consequently, all sample path properties of the integral $\int_s^t XdY$ (considered as a function of t) follow from properties of integrals with respect to non-decreasing functions. Most processes encountered within the context of survival analysis are of bounded variation on finite intervals. Any integral $\int_s^\infty XdY$ over an infinite interval is also well defined as Lebesgue-Stieltjes integral as it is a limit of integrals over finite intervals (Fleming and Harrington, 1991).

In the applications of interest, the stochastic integrals are of the form $\int_s^t f(u)dN(u)$ or $\int_s^\infty f(u)dN(u)$ where N is a counting process, $f(\cdot)$ is some function of time and $0 \le s \le t \le \infty$, with $N(\cdot)$ and $f(\cdot)$ satisfying the properties stated above so that the stochastic integrals are well defined as Lebesgue-Stieltjes. By the definition of the Lebesgue-Stieltjes integral and given that N as a step function will have countably many jumps at $\{u_1, u_2, ...\}$ with $\Delta N(u_k) = N(u_k) - N(u_k-) > 0$, it then follows that $\int_s^t f(u)dN(u) = \sum_{k:s<u_k\le t} f(u_k)\Delta N(u_k)$. Thus the integral $\int_s^t f(u)dN(u)$ represents the sum of the values

of $f(\cdot)$ at the jump times of N in the interval $(s,t]$. If N has a discontinuity at either s or t, it is the convention to take $\int_s^t f(u)dN(u) = \int_{(s,t]} f(u_k)\Delta N(u_k) = \sum_{k:s<u_k\le t} f(u_k)\Delta N(u_k)$.

4.2.1.3. Martingale and predictable processes

A further analytical aspect is the decomposition of the counting processes defined above into two specific types of stochastic processes referred to as martingales and predictable processes both of which play a particularly important role in establishing the estimators' asymptotic properties as will be shown below. A martingale can be viewed as a form of random noise process whereas a predictable process exhibits some kind of regular behaviour. In general, many stochastic processes can be decomposed as the sum of a martingale (random part) and a finite variation predictable process (systematic part). The latter is called the compensator of the process because when subtracted from the process, a martingale, i.e. non-systematic noise, is left.

Defined formally, a process $\mathcal{M}$ is a martingale with respect to a filtration $\mathcal{F}_t$ if it is right-continuous with left-hand limits,[2] adapted to $\mathcal{F}_t$, integrable, i.e. $E\{|\mathcal{M}(t)|\} < \infty$ for all $t \in \Gamma$, and satisfies the martingale property $E\{\mathcal{M}(t)|\mathcal{F}_s\} = \mathcal{M}(s)$ for all $s \le t$, i.e. the expected value of the process at t conditional on its history up to a previous point s will on average be same as its value at s. The martingale property then implies that for a right-continuous martingale $E\{d\mathcal{M}(t)|\mathcal{F}_{t-}\} = 0$, i.e. given prior history strictly before t represented by the filtration $\mathcal{F}_{t-}$, the increments of the process have an expected value of zero. A stochastic process H is said to be predictable with respect to $\{\mathcal{F}_t : t \ge 0\}$ if when H is adapted to $\{\mathcal{F}_t : t \ge 0\}$ then $H(t)$ is $\mathcal{F}_{t-}$-measurable, i.e. its behaviour at t is determined by the information on $[0,t)$ for all t. Predictable processes are essentially left-continuous processes adapted to a history $\mathcal{F}_t$ and thus determined at time t by the past strictly prior to t, i.e. by $\mathcal{F}_{t-}$. Predictable processes arise as compensators in martingales and as integrands in stochastic integrals. A martingale is therefore a process without any systematic behaviour in the mean: the process $\mathcal{M}(t) - \mathcal{M}(s)$ has zero mean given everything that has happened up to time s. In contrast, a predictable process is one whose value at time t is fixed given everything that has happened up to, but not including, t.

The martingale approach to statistical models for counting processes is useful in situations where the compensator is known or can be computed. This is the case in the applications of interest in this

[2] An example of a right-continuous process with left hand limits is the indicator process $I(T \le t)$ where T denotes the time of some random event. This process is equal to zero at time zero, then jumps to one at time T when the event occurs and then stays at that value. That is, the process approaches zero as t approaches T from the left, i.e. has a left hand limit zero, and its limit equals the value of the process at T as t approaches T from the right. The point T is thus a point of discontinuity.

chapter and then the martingale approach forms the basis for studying the statistical properties of the estimators and deriving explicit expressions for their variance estimators for large sample sizes with the use of the martingale version of the central limit theorem. In deriving such variance estimators the second moment of the martingale process $E\{\mathcal{M}^2(t)\}$ appears in the stochastic integrals. In general as considered in detail below, these stochastic integrals are of the form $\sum_i \int H_i d\mathcal{M}_i$, where N_i is some counting process and H_i is predictable with respect to the filtration making the processes $\mathcal{M}_i$ martingales. When the compensator for $\mathcal{M}^2$ has a simple form the decomposition of $\mathcal{M}^2$ leads to computationally appealing expressions for $E\{\mathcal{M}^2(t)\}$. The second moment of the martingale process is then calculated based on the following theorems.

If $\mathcal{M}$ is a right-continuous martingale with respect to a right-continuous filtration $\{\mathcal{F}_t : t \geq 0\}$ and $E\{\mathcal{M}^2(t)\} < \infty$ for any $t \geq 0$, then there exists a unique increasing right-continuous predictable process $\langle \mathcal{M}, \mathcal{M} \rangle$ called the predictable variation process of $\mathcal{M}$ such that $\langle \mathcal{M}, \mathcal{M} \rangle(0) = 0$ almost surely, $E\langle \mathcal{M}, \mathcal{M} \rangle(t) < \infty$ for each t, and $\{\mathcal{M}^2(t) - \langle \mathcal{M}, \mathcal{M} \rangle(t) : t \geq 0\}$ is a right-continuous martingale. Thus the predictable variation process $\langle \mathcal{M}, \mathcal{M} \rangle$ is the predictable compensator for $\mathcal{M}^2$ and it satisfies $d\langle \mathcal{M}, \mathcal{M} \rangle(t) = E\{d\{\mathcal{M}^2(t)\} | \mathcal{F}_{t-}\} = \text{var}\{d\mathcal{M}(t) | \mathcal{F}_{t-}\}$. The predictable variation process can be viewed as the "sum" of the conditional variance of $d\mathcal{M}(t)$ as time increases given information up to time t and can be used to calculate the variance $E\{\mathcal{M}^2(t)\}$ as follows: since $\mathcal{M}^2(t) - \langle \mathcal{M}, \mathcal{M} \rangle(t)$ is a martingale, when $\mathcal{M}(0) = 0$ almost surely, it follows that $E\{\mathcal{M}^2(t)\} = E\langle \mathcal{M}, \mathcal{M} \rangle(t)$.

The covariance between martingales can be calculated based on the following theorem. If $\mathcal{M}_1$ and $\mathcal{M}_2$ are two right-continuous martingales with respect to a right-continuous filtration $\{\mathcal{F}_t : t \geq 0\}$ and $E\{\mathcal{M}_i(t)\}^2 < \infty$ for $t \geq 0$ with $i = 1, 2$, then there exists a unique right-continuous predictable process $\langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ called the predictable covariation process, with $\langle \mathcal{M}_1, \mathcal{M}_2 \rangle(0) = 0$ and $E\langle \mathcal{M}_1, \mathcal{M}_2 \rangle(t) < \infty$, such that $\langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ is the difference of two increasing right-continuous predictable processes which implies that $\langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ has paths of bounded variation and therefore if it appears as the integrator in a stochastic integral the integral is well defined as a Lebesgue-Stieltjes integral with respect to this process, $\mathcal{M}_1\mathcal{M}_2 - \langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ is a martingale, and $d\langle \mathcal{M}_1, \mathcal{M}_2 \rangle(t) = E\{d\{\mathcal{M}_1(t)\mathcal{M}_2(t)\} | \mathcal{F}_{t-}\} = \text{cov}\{d\mathcal{M}_1(t), d\mathcal{M}_2(t) | \mathcal{F}_{t-}\}$. The predictable covariation process $\langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ can be used to calculate the covariance $E\{\mathcal{M}_1(t)\mathcal{M}_2(t)\}$ as follows: since $\mathcal{M}_1\mathcal{M}_2 - \langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ is a martingale, when $\mathcal{M}_i(0) = 0$ almost surely, it follows that $E\{\mathcal{M}_1(t)\mathcal{M}_2(t)\} = E\langle \mathcal{M}_1, \mathcal{M}_2 \rangle(t)$.

The predictable variation and covariation processes therefore allow calculation of the second moments of the martingale process appearing in the stochastic integrals involved in the statistics of interest. Existence and uniqueness of these processes is ensured by implication of the following theorem which primarily stipulates the conditions for the decomposition of a stochastic process into

a compensator and a martingale. The decomposition theorem (Doob-Meyer decomposition, Meyer 1966) states that a process H is the compensator of X if H is predictable right-continuous with left-hand limits and finite variation process such that the process $X - H$ is a martingale zero at time zero. Moreover, if a compensator exists, it is unique. Expressed with reference to a counting process, the theorem states that given a counting process $\{N(t): t \geq 0\}$ adapted to right-continuous filtration $\{\mathcal{F}_t: t \geq 0\}$ with $E\{N(t)\} < \infty$ for all t, then there exists a unique increasing right-continuous $\mathcal{F}_{t-}$-predictable process A such that

$A(0) = 0$ almost surely,
$E\{A(t)\} < \infty$ for all t, and
$\{\mathcal{M}(t) = N(t) - A(t): t \geq 0\}$ is a right-continuous $\mathcal{F}_{t-}$-martingale.

The compensator A of N is therefore a process that carries the predictable component of N determined at time t by the strict past, i.e. by $\mathcal{F}_{t-}$. The above theorem implies that an arbitrary adapted process N always allows a decomposition $\mathcal{M} = N - A$ with $\mathcal{M}$ being a martingale. Moreover, the theorem implies that unique predictable variation and covariation processes exist so that $\mathcal{M}^2 - \langle \mathcal{M}, \mathcal{M} \rangle$ and $\mathcal{M}_1\mathcal{M}_2 - \langle \mathcal{M}_1, \mathcal{M}_2 \rangle$ are martingales.

4.2.1.4. The $\int H d\mathcal{M}$ martingale

As noted above, in the analysis of counting process data martingales often appear when studying the statistical properties of the estimators of interest. In general, many censored data statistics are of the form $\sum_i \int H_i d\mathcal{M}_i$ where $\mathcal{M}_i = N_i - A_i$ for some counting process N_i and H_i is predictable with respect to the filtration making the processes $\mathcal{M}_i$ martingales. In addition if H_i is a bounded predictable process and $E\{N_i(t)\} < \infty$ for all t, the processes $\sum_i \int H_i d\mathcal{M}_i$ are themselves martingales. That the process $\int H_i d\mathcal{M}_i$ has the martingale property is shown as

$$
\begin{aligned}
E\left\{d\left[\int H_i(t) d\mathcal{M}_i(t)\right] \middle| \mathcal{F}_{t-}\right\} &= E\left\{H_i(t) d\mathcal{M}_i(t) \middle| \mathcal{F}_{t-}\right\} \\
&= H_i(t) E\left\{d\mathcal{M}_i(t) \middle| \mathcal{F}_{t-}\right\} && \text{(since } H_i \text{ is predictable)} \\
&= 0 && \text{(since } \mathcal{M}_i \text{ is a martingale)}
\end{aligned}
$$

If there is a common filtration $\{\mathcal{F}_t: t \geq 0\}$ with respect to which each H_i is predictable and each $\mathcal{M}_i$ is a martingale then the process $\sum_i \int H_i d\mathcal{M}_i$ will be a martingale with respect to $\{\mathcal{F}_t: t \geq 0\}$. When $\int H_i d\mathcal{M}_i$ is a martingale, $E\int H_i d\mathcal{M}_i = 0$ and when $\mathcal{M}_i = N_i - A_i$ it follows that $E\int H_i dN_i = E\int H_i dA_i$ thus enabling calculation of first moments for counting process statistics.

Second moments as stated previously require the predictable variation and covariation processes. For arbitrary counting processes[3] and locally bounded predictable H_i with $\mathscr{M}_i = N_i - A_i, i = 1, 2$, $\int H_i d\mathscr{M}_i$ is a martingale over $[0,t]$ if $E\int_0^t H_i^2 d\langle \mathscr{M}_i, \mathscr{M}_i \rangle < \infty$. This in turn implies that $E\int_0^t H_i d\mathscr{M}_i = 0$ and for $i, j \in \{1,2\}$ $E\left(\int_0^t H_i d\mathscr{M}_i \int_0^t H_j d\mathscr{M}_j\right) = E\int_0^t H_i H_j d\langle \mathscr{M}_i, \mathscr{M}_j \rangle$.

It then becomes clear that formulae for $\langle \mathscr{M}_i, \mathscr{M}_j \rangle$ are required. These are derived both for the case of continuous compensators and when the compensators may have discontinuities. In the applications to be considered later in the chapter the counting processes $\{N_1,...,N_n\}$ have continuous compensators $\{A_1,...,A_n\}$ due to the continuity of the distribution of failure time. In this case with the counting process determining whether failure has occurred and the failure time having an absolutely continuous distribution $F(t) = 1 - \exp\left\{-\int_0^t \lambda(u)du\right\}$, the compensator is absolutely continuous and is given by $A(t) = \int_0^t I(X \geq u) d\Lambda(u) = \int_0^t I(X \geq u)\lambda(u)du$, where X is the minimum of failure and censoring time, Λ is the cumulative hazard for failure and λ is the hazard rate for failure as defined on page 42. Then for statistics of the form

$$U(t) = \sum_{i=1}^{n} \int_0^t H_i d(N_i - A_i) = \sum_{i=1}^{n} \int_0^t H_i(u) d\mathscr{M}_i(u)$$

with H_i being $\mathscr{F}_t$–predictable and bounded on $[0,\infty)$ the following conditions hold:

1. The process U is a martingale over $[0,\infty)$
2. $E\{U(t)\} = 0, 0 \leq t \leq \infty$
3. $\text{var}\{U(t)\} = E\{U^2(t)\} = \sum_{i=1}^{n} \int_0^t E\{H_i^2(u) I(X_i \geq u)\}\lambda(u)du,\ 0 \leq t \leq \infty$.[4]

In other words, when the compensator A of the counting process N is continuous and $E\{A(t)\} < \infty$ or equivalently $E\{N(t)\} < \infty$ for all t, then

---

[3] For the counting process martingales $\mathscr{M}_1 = N_1 - A_1$ and $\mathscr{M}_2 = N_2 - A_2$ under stricter conditions requiring that N_i is bounded as opposed to arbitrary, with H_1 and H_2 being bounded and $E\{\mathscr{M}_i(t)\}^2 < \infty$, the following relationships hold:
$\left\langle \int H_i d\mathscr{M}_i, \int H_i d\mathscr{M}_i \right\rangle = \int H_i^2 d\langle \mathscr{M}_i, \mathscr{M}_i \rangle$ and $\left\langle \int H_1 d\mathscr{M}_1, \int H_2 d\mathscr{M}_2 \right\rangle = \int H_1 H_2 d\langle \mathscr{M}_1, \mathscr{M}_2 \rangle$

[4] In general, $\text{var}\{U(t)\} = E\{U^2(t)\} = E\sum_{i=1}^{n}\sum_{j=1}^{n} \int_0^t H_i(u) H_j(u) d\langle M_i, M_j \rangle(u),\ 0 \leq t \leq \infty$. Condition 3 follows when the martingales are orthogonal for $i \neq j$, i.e. $\langle \mathscr{M}_i, \mathscr{M}_j \rangle(t) = 0$ for $i \neq j$.

$E\{\mathscr{M}^2(t)\} = E\langle \mathscr{M}, \mathscr{M} \rangle = E\{A(t)\}$, $t \geq 0$, that is $\langle \mathscr{M}, \mathscr{M} \rangle(t) = A(t)$, and $\langle \mathscr{M}_i, \mathscr{M}_j \rangle(t) = 0$ for $i \neq j$, that is the component processes are pairwise uncorrelated i.e. orthogonal.[5]

Orthogonality of the corresponding martingales allows use of the martingale central limit theorem in establishing asymptotic distribution results for linear combinations of stochastic integrals with respect to orthogonal martingales, i.e. for statistics of the form $U(t) = \sum_{i=1}^{n} \int_0^t H_i(u) d\mathscr{M}_i(u)$.

4.2.1.5. A martingale central limit theorem

The martingale version of the central limit theorem gives the asymptotic distribution of statistics of the form $U^{(n)}(t) = \sum_{i=1}^{n} \int_0^t H_i(u) d\mathscr{M}_i(u)$ as the sample size $n \to \infty$ and uses the notion of weak convergence or convergence in distribution of stochastic processes.[6] The process $U^{(n)}$ is a sum of n orthogonal martingale transforms and the notation indicates the dependence of the process on the sample size n. Under certain conditions the process $U^{(n)}$ converges weakly to a time-transformed Wiener or Brownian motion process $W^*(t)$ as the number of summand martingales increases, where the stochastic process $W^*(t)$ satisfies the following conditions:

---

[5] In the case of a compensator A_i with discontinuities, $\langle \mathscr{M}_i, \mathscr{M}_i \rangle(t) = \int (1 - \Delta A_i) dA_i$. If for each $t \geq 0$ given $\mathscr{F}_{t-}$, the increments of the counting processes $\{\Delta N_1(t), \ldots, \Delta N_n(t)\}$ are independent 0,1 random variables, as is the case in most applications, then $\langle \mathscr{M}_i, \mathscr{M}_j \rangle(t) = 0$ almost surely for $i \neq j$, i.e. the component processes are orthogonal.

Then for statistics of the form $U_l(t) = \sum_{i=1}^{n} \int_0^t H_{i,l}(u) d\mathscr{M}_i(u), l = 1, 2$, such that $H_{i,l}$ is $\mathscr{F}_t$-predictable and bounded on $[0, \infty)$ and $A_i = \int_0^t I(X_i \geq u) d\Lambda(u)$, and for $l, l' \in \{1, 2\}$, the following conditions hold:

1. U_l is a martingale over $[0, \infty)$
2. $E\{U_l(t)\} = 0, 0 \leq t \leq \infty$
3. $\operatorname{cov}\{U_l(t), U_{l'}(t)\} = \sum_{i=1}^{n} \int_0^t E\{H_{i,l}(u) H_{i,l'}(u) I(X_i \geq u)\}\{1 - \Delta\Lambda(u)\} d\Lambda(u), \; 0 \leq t \leq \infty$.

[6] Weak convergence of stochastic processes generalises the notion of convergence in distribution of real-valued random variables. For arbitrary distribution functions on the real line F and F_n, F_n with $n = 1, .., n$ converges weakly to F as $n \to \infty$ if and only if $F_n(x) \to F(x)$ at all continuity points of F. This definition is extended to weak convergence for a sequence of random variables $\{X, X_n, n = 1, \ldots, n\}$ as follows. Assuming that the random variables all take values on the real line and letting $F_n(t) = pr\{X_n \leq t\}$ and $F(t) = pr\{X \leq t\}$, then X_n is said to converge in distribution to X, written as $X_n \xrightarrow{D} X$, if and only if F_n converges weakly to F.

1. $W^*(0) = 0$ and $E\{W^*(t)\} = 0$ for all t
2. $W^*(t)$ has independent increments, i.e. $W^*(t) - W^*(u)$ is independent of $W^*(u)$ for any $0 \le u \le t$
3. $W^*(t)$ has continuous sample paths
4. $W^*(t)$ is a Gaussian process, i.e. for any positive integer n and time points $t_1,...,t_n$ the joint distribution of $\{W^*(t_1),...,W^*(t_n)\}$ is multivariate normal
5. The standard Wiener or Brownian motion $W(t)$ has variance $\text{var}\{W(t)\} = t$. A time-transformed Brownian motion $W^*(t)$ has variance $\text{var}\{W^*(t)\} = \alpha(t)$, where $\alpha(t) = \int_0^t f^2(s)ds$ with f being a measurable non-negative function.

The process $U^{(n)}$ satisfies condition 1 for all n. By the martingale property, $U^{(n)}$ has uncorrelated increments so that condition 2 is plausible for large n. If the jumps of $U^{(n)}$ become negligible as $n \to \infty$ then the sample paths of $U^{(n)}$ become continuous for large n and then conditions 3 and 4 also hold. Also as stated above, $\text{var}\{U^{(n)}(t)\} = E\langle U^{(n)}, U^{(n)}\rangle(t)$ and if $\langle U^{(n)}, U^{(n)}\rangle(t) \xrightarrow{P} \alpha(t)$ for some integrand f^2, where the notation $\xrightarrow{P}$ denotes convergence in probability,[7] then $U^{(n)}$ should satisfy $\text{var}\{U^{(n)}(t)\} = \alpha(t)$ as $n \to \infty$.

Assuming that conditions 1 through 5 are true with f being a measurable non-negative function and $\alpha(t) = \int_0^t f^2(s)ds$ for all $t > 0$, assuming further that for all $t > 0$ as $n \to \infty$,

$$\langle U^{(n)}, U^{(n)}\rangle(t) \xrightarrow{P} \alpha(t), \qquad \text{that is} \qquad \sum_{i=1}^{n} \int_0^t \{H_i^{(n)}\}^2(s) d\langle \mathscr{M}_i^{(n)}, \mathscr{M}_i^{(n)}\rangle(s) \xrightarrow{P} \int_0^t f^2(s)ds,$$

and that the jumps of $U^{(n)}$ become negligible as $n \to \infty$, then the process $U^{(n)}$ converges weakly to a time-transformed Brownian motion $W^*(t)$. This means that the process $U^{(n)}$ converges in distribution to a multivariate normal distribution with zero mean and variance matrix given by the appropriate values of $\alpha(t)$.[8] In studying the statistical properties of the cost estimators presented later in the chapter, the martingale version of the central limit theorem given above is used to prove

---

[7] Let X_n be a sequence of random variables indexed by the size of the sample n. Then the sequence of random variables X_n converges in probability to a limit process X, written as $X_n \xrightarrow{P} X$ or $\text{plim} X_n = X$ if $\lim_{n\to\infty} pr(|X_n - X| > \varepsilon) = 0$, $\forall \varepsilon > 0$. This means that the values that the sequence X_n may take that are not close to the values of X become increasingly unlikely as n increases.

[8] For example, for any $t > 0$, $U^{(n)} \xrightarrow{D} N(0, \alpha(t))$, while because of independent increments for any two points in time $t_1, t_2 > 0$ the vector $\{U^{(n)}(t_1), U^{(n)}(t_2)\}$ converges in distribution to a bivariate normal with mean {0, 0} and variance matrix $\begin{bmatrix} \text{var}U^{(n)}(t_1) & \text{cov}\{U^{(n)}(t_1), U^{(n)}(t_2)\} \\ \text{cov}\{U^{(n)}(t_1), U^{(n)}(t_2)\} & \text{var}U^{(n)}(t_2) \end{bmatrix} = \begin{bmatrix} \alpha(t_1) & \alpha(t_1 \wedge t_2) \\ \alpha(t_1 \wedge t_2) & \alpha(t_2) \end{bmatrix}$ where $t_1 \wedge t_2 = \min(t_1, t_2)$ (Therneau and Grambsch, 2000, p.26)

asymptotic convergence of the cost statistics to a normal distribution and to derive asymptotic variance estimators as $n \to \infty$.

4.2.1.6. The independent censoring assumption

As the exposition above shows, one of the great advantages in modelling time-to-event data using the counting process and martingale framework is that censoring can be easily accommodated. The theory described above assumes that the underlying censoring mechanism is random, that is time to failure and time to censoring are independent random variables. Gill (1980) gives other examples of right-censoring mechanisms which may arise in medical studies and shows that for all these mechanisms the assumption of independence between T and U is justified. Moreover, the assumption of independence between T and U underlies all the proposed estimators of cost to be considered in this chapter. It is important therefore to investigate more closely the random censoring mechanism before presenting the estimators of cost. Investigation is undertaken using the theory presented above.

The same observed data are assumed as described earlier in section 4.2.1 and the same notation is adopted. Thus assuming a continuous non-negative failure time denoted by the random variable T and a censoring time variable with an arbitrary distribution denoted by U, where T and U are independent and λ is the hazard function for T, the observable data for n individuals are the independent and identically distributed random vectors

$\{X_i = \min(T_i, U_i),\ \ \delta_i = I(T_i \le U_i)\}, i = 1,...,n$, where i identifies an individual.

The following stochastic processes are defined. $N_i(t) = I(X_i \le t, \delta_i = 1)$ with $N(t) = \sum_{i=1}^{n} N_i(t)$ counting the number of individuals dying over time, $N_i^c(t) = I(X_i \le t, \delta_i = 0)$ with $N^c(t) = \sum_{i=1}^{n} N_i^c(t)$ counting the number of individuals censored over time, and $Y_i(t) = I(X_i \ge t)$ with $Y(t) = \sum_{i=1}^{n} Y_i(t)$ counting the number of individuals at risk over time. The associated filtration $\{\mathcal{F}_t : t \ge 0\}$ is given by $\mathcal{F}_t = \sigma\{N(u), N^c(u) : 0 \le u \le t, i = 1,...,n\}$ and provides information on the individuals who have died or have been censored up to and including time t. Due to the independence between T and U the hazard rate in the presence of censoring is

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{pr(t \le T \le t + \Delta t | T \ge t)}{\Delta t} = \lim_{\Delta t \to 0} \frac{pr(t \le T \le t + \Delta t | T \ge t, U \ge t)}{\Delta t}$$

Hence $pr(t \le T \le t + \Delta t | T \ge t, U \ge t) \approx \lambda(t)\Delta t$.

Because $N(\cdot)$ is right-continuous, i.e. $N(t-) = \lim_{s \to t} N(s)$,

$$\lambda(t)\Delta t \approx pr\{N[(t+\Delta t)-] - N(t-) = 1 | T \geq t, U \geq t\}$$

and since $N[(t+\Delta t)-] - N(t-)$ is a 0, 1-valued random variable,[9]

$$\lambda(t)\Delta t \approx E\{\{N[(t+\Delta t)-] - N(t-)\} | T \geq t, U \geq t\}.$$

The hazard rate therefore under conditions of independence between T and U, gives the average rate of change in N over $[t, t+\Delta t)$ conditional on both the survival and censoring time exceeding or being equal to t and thus specifies the conditional rate at which N jumps in small intervals. Then the process A given by

$$A(t) = \int_0^t I(X \geq u)\lambda(u)du$$

is a random variable at each fixed t and approximates the number of jumps of N over $(0, t]$. Furthermore A is the compensator of N since A is predictable (it is continuous and adapted to a history $\mathcal{F}_t$ and thus determined at time t by the past strictly prior to t, i.e. by $\mathcal{F}_{t-}$) which implies that the process defined by

$$\mathcal{M}(t) = N(t) - A(t) = I(X \leq t, \delta = 1) - \int_0^t I(X \geq u)\lambda(u)du$$

is a martingale. To show that $\mathcal{M}(t)$ is indeed a martingale, following Fleming and Harrington (1991), the following filtration is assumed

$$\mathcal{F}_s = \sigma\{N(u), I(X \leq u, \delta = 0) : 0 \leq u \leq s\}.$$

Then the filtration $\mathcal{F}_{s-} = \sigma\{N(u), I(X \leq u, \delta = 0) : 0 \leq u < s\}$ represents the accumulated information on $N(u)$ up to, but not including, time s.

As shown above, when T and U are independent, $dN(s)$ denotes the change in the process $N(s)$ over an infinitesimal interval $(s - ds, s]$ and is one if a failure occurred at s and zero otherwise, that is, $dN(s)$ is a 0, 1-valued random variable with conditional probability $I(X \geq s)\lambda(s)ds$ of being 1 given $\mathcal{F}_{s-}$ implying

---

[9] For a 0, 1-valued random variable X,
$E(X) = \sum_X Xpr(X = x) = \{(X = 0)pr(X = 0)\} + \{(X = 1)pr(X = 1)\} = pr(X = 1)$

$$E\{dN(s)|\mathcal{F}_{s-}\} = pr\{dN(s)=1|\mathcal{F}_{s-}\} = I(X \geq s)\lambda(s)ds = dA(s).$$

Also

$$E\{dA(s)|\mathcal{F}_{s-}\} = E\{I(X \geq s)\lambda(s)ds|\mathcal{F}_{s-}\} = I(X \geq s)\lambda(s)ds = dA(s).$$

The change in the process $\mathcal{M} = N - A$ over an infinitesimal interval $(s-ds, s]$ is $d\mathcal{M}(s) = dN(s) - dA(s)$. If then follows from the above that $E\{d\mathcal{M}(s)|\mathcal{F}_{s-}\} = 0$, that is, $\mathcal{M}$ is a martingale with respect to $\mathcal{F}_s$.

The exposition above implies that when T is a continuous failure time random variable and U a censoring time variable with an arbitrary distribution, with $X = \min(T,U)$, $\delta = I(T \leq U)$, λ being the hazard function for T and

$N(t) = I(X \leq t, \delta = 1)$,
$N^c(t) = I(X \leq t, \delta = 0)$,
$\mathcal{F}_t = \sigma\{N(u), N^c(u) : 0 \leq u \leq t\}$

then the process given by $\mathcal{M}(t) = N(t) - \int_0^t I(X \geq u)\lambda(u)du$ is a martingale if X and U are independent.[10] Consequently there is always an increasing process $A(t) = \int_0^t I(X \geq u)\lambda(u)du$ such that $N - A$ is a martingale with respect to the filtration $\mathcal{F}_t = \sigma\{N(u), N^c(u) : 0 \leq u \leq t\}$.

[10] Counting process and martingale theory can accommodate dependency between the variable of interest and its censoring variable as shown in Fleming and Harrington (1991). If T and U are dependent, given the counting processes and filtration as defined in the random censorship model above, the associated martingale process is given as

$$\mathcal{M}(t) = N(t) - \int_0^t I(X \geq u)\lambda^{\#}(u)du.$$

Thus, there is still a compensator $A(t) = \int_0^t I(X \geq u)\lambda^{\#}(u)du$ which makes the previous process a martingale with respect to the above mentioned filtration, but the hazard rate for failure time is now $\lambda^{\#}$ which can be very different from the hazard rate for failure λ which holds when T and U are independent. Specification of $\lambda^{\#}$ requires knowledge of the joint distribution between T and U.

4.3. Non-parametric estimators of cost under censoring

The preceding sections presented the counting process approach to the analysis of time-to-event data. As shown the approach allows such data to be modelled and analysed while accommodating the presence of censoring. The following sections will concentrate on the analysis of cost-to-event data under conditions of censoring using the same analytical framework. All non-parametric estimators referred to in the introduction to the chapter together with the assumptions underlying their validity will be presented and their statistical properties will be studied using the counting processes and martingale theory given above. The same general setting is assumed as given in section 4.2.1.

To reiterate, the aim of the approaches presented below is to derive an estimate of the mean total cost $\mu = E(M)$ and its variance over a specified period when the data is right censored, where the random variable M denotes the total cost for a patient during some specified time T and E denotes expectation. The distribution of the random variable T is assumed continuous over $(0, L]$ where L denotes the upper bound of T, i.e. the maximum time for which each patient is evaluated in which case M is the total cost incurred by a patient up to a maximum of L units of time. To accommodate censoring, a potential time to censoring denoted by U is defined and letting T denote the time to death, the observables from a study in the presence of censoring are $X = \min(T,U)$, i.e. the last contact date; $\delta = I(T \leq U)$, where $I(\cdot)$ is the indicator function taking the value of 1 when the observation is uncensored and zero otherwise; the cost accrued up to time X and other intermediate cost history for each subject, i.e. $M^H(t) = \{M(u), u \leq t\}$, where $M^H(t)$ denotes the cost history up to time t, $M = M(T)$, with $M(u)$ being the known accumulated cost up to time u and u denoting points in time at which cost information becomes available. The observable data for n individuals are then the independent and identically distributed random vectors

$\{X_i = \min(T_i, U_i),\ \delta_i = I(T_i \leq U_i),\ M_i^H(X_i)\}, i = 1,...,n$, where i identifies an individual.

4.3.1. Kaplan-Meier and "naïve" estimators

The first attempt to account for censoring in cost estimates used the Kaplan-Meier estimator (Fenn et al, 1995). Before outlining how this estimator has been applied in the estimation of the distribution of costs, it is useful to present the method as applied in the estimation of the distribution of failure times and study the statistical properties of the estimator within the framework of the theory of stochastic processes. The Kaplan-Meier estimator (Kaplan and Meier, 1958) plays a role for censored data similar to that of the empirical distribution function for uncensored data, that is, it is an estimator of the cumulative distribution function for failure $F(t)$ based on the observations $(X_i, \delta_i)\ i = 1,...,n$, which reduces to the usual distribution function based on $T_1,...,T_n$ if $\delta_i = 1$ for each i, where the T_i's are independent and identically distributed with distribution function F.

Under the assumption of independent censoring, the Kaplan-Meier estimator for the probability of survival to time t is given by

$$\hat{S}(t) = \prod_{s \le t} \left\{ 1 - \frac{\Delta N(s)}{Y(s)} \right\} \tag{4.1}$$

where the process $N(t) = I(X \le t, \delta = 1)$ counts the number of failures and $Y(t) = I(X \ge t)$ counts the number at risk. It should be noted that different versions of the Kaplan-Meier estimator have been proposed to define the estimator when the largest observed time corresponds to censoring. All versions of the Kaplan-Meier estimator equal $\prod_{s \le t} (1 - \Delta N(s))/Y(s)$ for $t \le X_{\max}$, where $X_{\max}$ denotes the largest observed time, and they are all equal to zero for $t > X_{\max}$ if the event at $X_{\max}$ is a failure. In the original paper by Kaplan and Meier, the estimator was left undefined for $t > X_{\max}$ if $X_{\max}$ is a censored observation. Efron (1967) set the estimator equal to zero for $t > X_{\max}$ even if the last observation was censored. The version adopted here was proposed by Gill (1980) and sets the estimator equal to $\hat{S}(X_{\max})$, that is equal to its value at the largest observed time, for $t > X_{\max}$ even when the last observation is censored.

The mean survival time is the area under the Kaplan-Meier curve

$$\mu = \int_0^{\infty} S(u)du$$

and the mean survival time over $(0, t]$ is estimated as

$$\hat{\mu}_t = \int_0^{t} \hat{S}(u)du \tag{4.2}$$

This estimator is shown to be consistent with asymptotic variance estimated as

$$\hat{\text{var}}(\hat{\mu}_t) = \int_0^{t} \left(\int_v^{t} \hat{S}(u)du \right)^2 d\left\{ \int_0^{v} \frac{dN(u)}{Y(u)[Y(u) - \Delta N(u)]} \right\} \tag{4.3}$$

(see, Andersen, Borgan, Gill et al, 1993, p.279). For calculation purposes, the mean survival time can be written as

$$\hat{\mu}_t = \sum_{i=1}^{m} \hat{S}(t_{i-1})(t_i - t_{i-1})$$

where t_m is the largest observed point in time, and the asymptotic variance can be expressed as

$$\hat{\text{var}}(\hat{\mu}_t) = \sum_{i=1}^{m-1} \frac{\left(\sum_{j=i}^{m-1} \hat{S}(t_j)(t_{j+1} - t_j) \right)^2 d_i}{n_i(n_i - d_i)}$$

where d_i denotes deaths and n_i denotes individuals at risk.

As noted above the Kaplan-Meier estimator was the first non-parametric estimator to be applied to cost-to-event data in an attempt to account for censoring in the cost estimates. Table 4.1 outlines the main concepts underlying the approach when this is applied to time-to-event data and contrasts these with the analogous concepts underlying the application of the same approach to cost-to-event data.

Table 4.1. Kaplan-Meier estimator applied to cost versus survival

In time to event analysis	**In cost analysis**
The random variable of interest T is time to event	The random variable of interest M is the level of cost incurred to the event, where the event is the termination of the study period
The hazard associated with a particular moment in time is the conditional probability of death at that moment, given survival until that moment	The hazard associated with a particular level of cost is the conditional probability that cost will not exceed that level, given that it has reached that level
The survival function when evaluated at time t gives the probability that the patient will survive at least until t: $S(t) = pr(T \geq t)$	The analogous cost function implied by the equation $S(c) = pr(M \geq c)$ gives the probability that the cost will be at least c
An observation is right-censored at a moment in time if it is known only that the patient survived past that moment (was alive at that moment), i.e. time of death or even whether death has occurred is not observable	Right-censoring at a particular cost level within the study period occurs if it were known only that the patient's cost within that period was at least that great, i.e. the patient's cost behaviour across the complete period is not available for analysis
The Kaplan-Meier estimator for the probability of survival to time t is $\hat{S}(t) = \prod_{s \leq t} \left\{ 1 - \frac{\Delta N(s)}{Y(s)} \right\}$ where $N(t) = \sum_{i=1}^{n} I(X_i \leq t, \delta_i = 1)$, $Y(t) = \sum_{i=1}^{n} I(X_i \geq t)$	The Kaplan-Meier estimator for the probability of cost being at least c is $\hat{S}(c) = \prod_{k \leq c} \left\{ 1 - \frac{\Delta N(k)}{Y(k)} \right\}$ where $N(c) = \sum_{i=1}^{n} I(M_i \leq c, \delta_i = 1)$, $Y(c) = \sum_{i=1}^{n} I(M_i \geq c)$
The mean survival time is estimated by $\hat{\mu}_L = \int_0^L \hat{S}(u)du$ where L is the maximum observed duration in the study	The mean cost over the study period is estimated by $\hat{\mu}_{KM} = \int_0^C \hat{S}(c)dc$ where C is the maximum observed cost in the study
The variance estimator for the mean survival over $(0, L]$ is given as $\hat{\text{var}}(\hat{\mu}_L) = \int_0^L \left(\int_v^L \hat{S}(u)du \right)^2 d\left\{ \int_0^v \frac{dN(u)}{Y(u)[Y(u) - \Delta N(u)]} \right\}$ where u and v denote points in time	The variance estimator for the mean cost over $(0, C]$ is given as $\hat{\text{var}}(\hat{\mu}_{KM}) = \int_0^C \left(\int_h^C \hat{S}(c)dc \right)^2 d\left\{ \int_0^h \frac{dN(c)}{Y(c)[Y(c) - \Delta N(c)]} \right\}$ where c and h denote levels of cost

In the application of the Kaplan-Meier approach to cost-to-event data, the hazard rate associated with a given cost level specifies the conditional probability of having "completed" that cost level, that is, it gives the probability of an individual dying having attained c units of cost given that the individual was alive after having attained c-1 units of cost. The probability that the cost will be at least c, $S(c) = pr(M \geq c)$, is then given by the Kaplan-Meier estimator as

$$\hat{S}(c) = \prod_{k \leq c} \left\{ 1 - \frac{\Delta N(k)}{Y(k)} \right\} \tag{4.4}$$

where c and k denote levels of cost, $N(c) = \sum_{i=1}^{n} I(M_i \leq c, \delta_i = 1)$, that is the counting process counts the number of complete cost observations, or stated differently, the number of individuals who die having reached a cost level of less or equal to c, $Y(c) = \sum_{i=1}^{n} I(M_i \geq c)$ is the number of individuals who have attained a cost level of at least c, M_i denotes the observed cost for individual i and $\delta_i = I(T_i \leq U_i)$. The estimator of the mean cost over the cost interval $(0, C]$ is given by the area under the Kaplan-Meier cost curve as

$$\hat{\mu}_{KM} = \int_0^C \hat{S}(c)dc \tag{4.5}$$

which is computed as

$$\hat{\mu}_{KM} = \sum_{i=1}^{m} \hat{S}(c_{i-1})(c_i - c_{i-1})$$

where c_m is the maximum observed cost in the study. The asymptotic variance of the mean cost over $(0, C]$ is estimated as

$$\hat{\text{var}}(\hat{\mu}_{KM}) = \int_0^C \left(\int_h^C \hat{S}(c)dc \right)^2 d\left\{ \int_0^h \frac{dN(c)}{Y(c)[Y(c) - \Delta N(c)]} \right\} \tag{4.6}$$

where c and h denote levels of cost and is computed as

$$\hat{\text{var}}(\hat{\mu}_{KM}) = \sum_{i=1}^{m-1} \frac{\left(\sum_{j=i}^{m-1} \hat{S}(c_j)(c_{j+1} - c_j) \right)^2 d_i}{n_i(n_i - d_i)}$$

where d_i denotes the number of individuals who die having reached a cost level of less or equal to a given value and n_i denotes the number of individuals who have attained a cost level of at least that value.

Another two estimators have also been used to estimate the mean cost in the presence of censoring which are referred to as "naïve" estimators in the literature because the first only uses the uncensored cases in the estimation of mean cost and is referred to as the uncensored cases estimator, while the second uses all cases but does not differentiate between censored and uncensored observations and is referred to as the full-sample estimator. To show explicitly how these estimators are computed, it is convenient to present first the Kaplan-Meier estimator for cost as

$$\hat{\mu}_{KM} = \sum_{h=1}^{C}\left[\prod_{k=1}^{h}\left\{1-\frac{n_k}{n-\sum_{j=0}^{k-1}(n_j+c_j)}\right\}\right]$$

where n is the total number of subjects entering in the study, n_k is the number of complete cost observations at k as defined above, n_j and c_j are uncensored and censored observations respectively and j, h, and k denote units of cost.[11]

The uncensored cases estimator $\hat{\mu}_U$, where the mean cost is calculated with reference to the uncensored data alone is given as

$$\hat{\mu}_U = \sum_{h=1}^{C}\left[\prod_{k=1}^{h}\left\{1-\frac{n_k}{n_n-\sum_{j=0}^{k-1}n_j}\right\}\right] \qquad (4.7)$$

where n_n is the total number of uncensored cases.

The full sample estimator $\hat{\mu}_{FS}$, where the mean is estimated by reference to the full sample but without distinction between censored and uncensored observations is given as

$$\hat{\mu}_{FS} = \sum_{h=1}^{C}\left[\prod_{k=1}^{h}\left\{1-\frac{n_k+c_k}{n-\sum_{j=0}^{k-1}(n_j+c_j)}\right\}\right] \qquad (4.8)$$

Fenn et al (1995) show that the latter two estimators impart bias with both the full-sample estimator and the uncensored estimator resulting in smaller estimates of mean cost than the 'true' Kaplan-

[11] In a similar manner, the Kaplan-Meier estimator of the mean survival time can be given as

$$\hat{\mu}_L = \sum_{t=1}^{L}\left[\prod_{k=1}^{t}\left\{1-\frac{n_k}{n-\sum_{j=0}^{k-1}(n_j+c_j)}\right\}\right]$$

where n is the total number of subjects entering in the study, n_k is the number of deaths at k, n_j and c_j are uncensored and censored observations respectively and j, k, and t denote units of time. Thus $\hat{\mu}_L$ represents the area under the Kaplan-Meier survival curve up to L units of time, which is the maximum observed duration in the study.

Meier estimator. They conclude that the Kaplan-Meier is to be preferred when analysing censored cost data. Lin et al (1997), Etzioni et al (1999), and Bang and Tsiatis (2000) argue however that such a conclusion is misplaced. Their argument is based on the fact that the validity of the Kaplan-Meier approach relies on the assumption of independence between the variable of interest and its censoring variable which is satisfied with respect to time-to-event but it fails with respect to cost-to-event.

In the analysis of failure time data, this assumption requires independence between time to failure (T) and time to censoring (U) which as stated above is satisfied when the censoring mechanism is random. Independence between T and U can be interpreted in the following manner. Considering the T_i's as lifetimes starting at time $t = 0$, $X_i > t$ means that individual i is still under observation just after time t. Independence means that for every t, given what has happened up to and including time t, the remaining lifetimes of the individuals who are still under observation just after time t have the same joint distribution as if there had been no censoring. In particular, the fact that individual i has not been censored in $(0, t]$ gives no information about his remaining lifetime distribution. In other words, the removal of certain observations due to censoring does not affect the joint distribution of failure time for the remaining observations. Stated differently, independent censoring implies that the probability of an individual being censored at any point in time t is not related to the individual's risk of failure. As a result, the expected survival time is the same for censored patients as for uncensored patients. Independence between the variable of interest and its censoring variable within the context of cost-to-event analysis requires independence between cost at failure time and cost at censoring time. If this was the case, patients censored at the same time with the same accumulated costs would be expected to have the same total costs if they were followed to death and then the Kaplan-Meier estimator would provide unbiased cost estimates. In other words, the Kaplan-Meier estimator is inappropriate unless all patients accumulate costs at a common rate over time yielding a one-to-one correspondence between the survival time and total cost. Commonly, however, the rate of cost accumulation varies among individuals, with those in worse health utilising higher levels of resource and costing more per unit of time. Independence is therefore violated as patients who accrue costs at higher rates tend to generate larger total costs at both the survival time and the censoring time, which implies positive correlation between the total cost at failure time and the total cost at censoring time. Consequently, the removal of certain observations due to censoring affects the joint distribution of cost for the remaining observations, that is, at any point in time future cost expectation is statistically altered (from what it would have been without censoring) by censoring. The condition of independent censoring required for the validity of the Kaplan-Meier method is thus violated and this estimator is therefore inappropriate in the analysis of censored cost data.

Furthermore, the "naïve" estimators defined above will always be biased. The full-sample estimator is always biased downward because the costs incurred after censoring times are not accounted for

whereas the uncensored-cases estimator is biased toward the costs of the patients with shorter survival times because larger survival times are more likely to be censored.

Lin et al (1997) acknowledge these difficulties and propose an alternative which attempts to deal with this bias. Their estimators are derived by partitioning the study time period into a number of subintervals and consistency is ensured if censoring occurs solely at the interval boundaries. Under such censoring conditions, the estimators are shown to be asymptotically normal and asymptotic variances are analytically derived using the martingale theory presented above. This censoring pattern essentially requires discreetness of the censoring time distribution so that censoring times can be confined to the boundaries of the subintervals of the partition. Failure to meet this requirement will result in some bias in the estimates. This limitation led to a further set of estimators introduced by Bang and Tsiatis (2000) which are shown to be consistent regardless of the censoring pattern. Asymptotic normality and consistent variance estimators are also derived using the counting process and martingale theory given above. These estimators together with the assumptions underlying their validity are considered in turn below.

4.3.2. Lin et al estimators

Lin et al (1997) present two approaches in estimating the mean total cost over the period $(0, L]$. The first requires information on a patient's intermediate cost history whereas the second only uses the observed total costs at the last contact dates. In both approaches, the entire study period $(0, L]$ is divided into K intervals $[\alpha_k, \alpha_{k+1})$, $(k = 1,..., K)$, where $\alpha_1 = 0$ and $\alpha_{K+1} = L$. The assumptions underlying both approaches are as follows. Independence between time to failure and censoring time, an extension of the independent censoring assumption to ensure that at no point in time t are patients censored because they accrue unusually high or low costs, continuous distribution of failure time and continuous or discrete distribution of censoring time. However, as stated above, consistency of the estimators requires that the pattern of the censoring distribution is such that the censoring times can be made to coincide with specific points in time corresponding to the interval boundaries of the partition of the study period $(0, L]$. This essentially imposes a discrete pattern for the distribution of censoring times.

4.3.2.1. Estimator of mean cost when cost histories are recorded

The authors' first approach (referred to as Lin1 below) can be used to estimate $\mu = E(M)$ when the cost histories are recorded in which case M may be decomposed as $(M_1,..., M_K)$, where M_k is the observed cost over $[\alpha_k, \alpha_{k+1})$. That is, $M = \sum_{k=1}^{K} M_k$ which implies that

$$\mu = \sum_{k=1}^{K} E(M_k) = \sum_{k=1}^{K} E\{E(M_k | T \geq \alpha_k)\} = \sum_{k=1}^{K} pr(T \geq \alpha_k) E(M_k | T \geq \alpha_k) = \sum_{k=1}^{K} S_k E_k$$

where $S_k = pr(T \geq \alpha_k)$ and $E_k = E(M_k | T \geq \alpha_k)$. Replacing the unknown quantities S_k and E_k by their consistent sample estimators will result in a consistent estimator for μ. The mean total cost μ is thus estimated by

$$\hat{\mu}_{LIN1} = \sum_{k=1}^{K} \hat{S}_k \hat{E}_k \tag{4.9}$$

where $\hat{S}_k = \Pr(T \geq a_k)$ is the probability of surviving to a_k and it is consistently estimated by the Kaplan-Meier method as

$$\hat{S}_k = \prod_{u \leq a_k} \left\{ 1 - \frac{dN(u)}{Y(u)} \right\} \tag{4.10}$$

where the counting processes $N(u)$ and $Y(u)$ have been defined above, and

$$\hat{E}_k = \frac{\sum_{i=1}^{n} Y_{ki} M_{ki}}{\sum_{i=1}^{n} Y_{ki}} \quad , \; k = 1,\ldots,K \tag{4.11}$$

where M_{ki} is the observed cost of individual *i* incurred in interval *k* and in this case $Y_{ki} = I(X_i \geq \alpha_k)$. That is, $\hat{E}_k$ is an estimator for mean cost E_k in interval *k* and is derived from those individuals who are under observation at the start of the interval.[12] In other words, $\hat{E}_k$ is the sample average of the observed costs over the interval $[\alpha_k, \alpha_{k+1})$ conditional on survival to the start of the interval. Thus, $\hat{E}_k$ is an unbiased estimator of the true average cost E_k in interval *k* if censoring occurs at the end of the interval, since in that case M_{ki} represents the cost of individual *i* over the whole interval *k* for all for all *i*'s with $Y_{ki} = 1$, that is for all individuals who were under observation at the start of the interval. If censoring occurs before the end of the interval, $\hat{E}_k$ will underestimate E_k since it does not take into account the costs of the censored observations from the point of censoring to the end of the interval. The authors also suggest an alternative way of estimating E_k based on the exclusion of those who are censored during $[\alpha_k, \alpha_{k+1})$ from the calculation of the sample average $\hat{E}_k$. The resulting estimator will be unbiased if all the patients who are under observation at time α_k have the same probability of being censored during $[\alpha_k, \alpha_{k+1})$. This condition, which implies that the uncensored M_{ki}'s are representative of all the M_{ki}'s in the *k*th interval so that exclusion of those censored in the interior of the interval would not impart bias in the estimates, essentially requires that censoring occur at the start of the interval on the basis that larger survival times are associated with higher probabilities of being censored.

---

[12] Assuming that extended independent censoring as defined above holds, i.e. $E(M_k | T \geq \alpha_k) = E(M_k | X \geq \alpha_k)$, implies that E_k can be estimated from those who are under observation at the start of the interval.

Clearly, the bias diminishes as the intervals of the partition shrink and as the authors note both estimators of E_k are nearly consistent for narrow time intervals regardless of the censoring pattern.

For large samples, $\hat{\mu}_{LIN1}$ is shown to be asymptotically normal and its variance estimator is derived as follows. By the law of the large numbers, the estimators $\hat{E}_k = \frac{\sum_{i=1}^{n} Y_{ki} M_{ki}}{\sum_{i=1}^{n} Y_{ki}}$ $(k = 1,\ldots,K)$ converge in probability to $E_k^* = E(M_{k1} | Y_{k1} = 1)$. It then follows from Slutsky's theorem and the consistency of the Kaplan-Meier estimator that $\hat{\mu}_{LIN1} = \sum_{k=1}^{K} \hat{S}_k \hat{E}_k$ converges in probability to $\mu_{LIN1}^* = \sum_{k=1}^{K} S_k E_k^*$. Let $Z = n^{1/2}\left(\hat{\mu}_{LIN1} - \mu_{LIN1}^*\right)$.[13] Then

$$
\begin{aligned}
Z &= n^{1/2}\left(\sum_{k=1}^{K} \hat{S}_k \hat{E}_k - \sum_{k=1}^{K} S_k E_k^*\right) \\
&= n^{1/2}\left(\sum_{k=1}^{K} \hat{S}_k \hat{E}_k - \sum_{k=1}^{K} \hat{S}_k E_k^*\right) + n^{1/2}\left(\sum_{k=1}^{K} \hat{S}_k E_k^* - \sum_{k=1}^{K} S_k E_k^*\right) \\
&= n^{1/2} \sum_{k=1}^{K} \hat{S}_k \left(\hat{E}_k - E_k^*\right) + n^{1/2} \sum_{k=1}^{K} E_k^* \left(\hat{S}_k - S_k\right) \\
&= Z_1 + Z_2, \quad \text{say.}
\end{aligned}
$$

Due to the consistency of the $\hat{S}_k$'s,

$$
\begin{aligned}
Z_1 &= n^{1/2} \sum_{k=1}^{K} S_k \left(\hat{E}_k - E_k^*\right) + o_p(1) \\
&= \sum_{k=1}^{K} S_k \frac{n^{-1/2}}{n^{-1}} \left(\frac{\sum_{i=1}^{n} Y_{ki} M_{ki}}{\sum_{i=1}^{n} Y_{ki}} - E_k^*\right) + o_p(1) \\
&= \sum_{k=1}^{K} S_k \frac{n^{-1/2} \sum_{i=1}^{n} Y_{ki} \left(M_{ki} - E_k^*\right)}{n^{-1} \sum_{i=1}^{n} Y_{ki}} + o_p(1)
\end{aligned}
$$

---

[13] The choice of the particular expression $n^{1/2}(\hat{\mu} - \mu)$ in studying the estimator's asymptotic properties is due to the a frequently observed result in asymptotic statistics, namely that in general many statistics allow approximation by an average of the form

$$n^{1/2}(\hat{\mu} - \mu) = n^{-1/2} \sum_{i=1}^{n} \psi(x_i) + o_p(1)$$

where $\psi(x_i)$ is some random variable and the notation $o_p(1)$ denotes a term that converges to zero in probability. In these circumstances, if $\psi(x_i)$ has zero mean and finite second moments then by the central limit theorem $n^{1/2}(\hat{\mu} - \mu)$ is asymptotically zero mean normally distributed.

where $o_p(1)$ denotes an asymptotically negligible term which converges in probability to zero. By the central limit theorem and the law of large numbers, the random variable $n^{-1/2}\sum_{i=1}^{n} Y_{ki}(M_{ki} - E_k^*)$ is asymptotically zero-mean normal and the random variable $n^{-1}\sum_{i=1}^{n} Y_{ki}$ converges in probability to the constant $E(Y_{k1})$. It then follows from Slutsky's theorem that

$$Z_1 = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\frac{S_k Y_{ki}(M_{ki} - E_k^*)}{E(Y_{k1})} + o_p(1) \qquad (4.12)$$

which is a sum of n independent and identically distributed (i.i.d.) zero-mean random variables.

To derive an i.i.d. representation for Z_2 it is convenient to use the counting process and martingale theory presented above. The counting processes are defined as shown in sections 4.2.1.1 and 4.2.1.6, that is,

$$N_i(t) = I(X_i \le t, \delta_i = 1) = \delta_i I(X_i \le t) \text{ with } N(t) = \sum_{i=1}^{n} N_i(t),$$

$$N_i^c(t) = I(X_i \le t, \delta_i = 0) \text{ with } N^c(t) = \sum_{i=1}^{n} N_i^c(t)$$

and the filtration these processes generate is $\mathcal{F}_t = \sigma\{N(u), N^c(u) : 0 \le u \le t\}$.

Due to the independence between failure and censoring time, as shown in section 4.2.1.6, the processes given by $\mathcal{M}_i(t) = N_i(t) - \int_0^t I(X_i \ge u)d\Lambda(u)$ are the associated subject specific martingales, where $\Lambda(\cdot)$ is the integrated hazard function for the failure time T. The Kaplan-Meier estimator $\hat{S}_k$ is asymptotically equivalent to $e^{-\hat{\Lambda}_k}$ where $\hat{\Lambda}_k$ is the Nelson-Aalen estimator[14] for $\Lambda_k = \Lambda(\alpha_k)$,

---

[14] Assuming the random censoring model and the associated counting processes, filtration and underlying distributions given above, then as stated previously the process given by $\mathcal{M}_i(t) = N_i(t) - \int_0^t Y_i(s)d\Lambda(s)$ is a martingale for each i with respect to $\mathcal{F}_t$. An estimator of the integrated hazard function $\Lambda(t) = \int_0^t \lambda(u)du$, first proposed by Nelson (1969), is $\hat{\Lambda}(t) = \int_0^t \frac{I\{Y(u) > 0\}}{Y(u)} dN(u)$, where the stochastic integrand is taken to be 0 when both the numerator and the denominator vanish. Following Fleming and Harrington (1991), when no statistical model is assumed, information is available only for $\{u : Y(u) > 0\}$ and $\hat{\Lambda}(t)$ estimates the random quantity $\Lambda^*(t) = \int_0^t I\{Y(u) > 0\}\lambda(u)du$. Then

$$\hat{\Lambda}_k = \hat{\Lambda}(\alpha_k) = \int_0^{\alpha_k} \frac{dN(t)}{Y(t)} = \sum_{i=1}^{n} \int_0^{\alpha_k} \frac{dN_i(t)}{\sum_{j=1}^{n} I(X_j \geq t)} \tag{4.13}$$

since $Y(t) = \sum_{i=1}^{n} Y_i(t) = \sum_{i=1}^{n} I(X_i \geq t)$

Furthermore as shown in Fleming and Harrington (1991),

$$\begin{aligned} n^{1/2}(\hat{\Lambda}_k - \Lambda_k) &= n^{1/2} \sum_{i=1}^{n} \int_0^{\alpha_k} \frac{I\{Y(t) > 0\}}{Y(t)} d\mathscr{M}_i(t) + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^{n} \int_0^{\alpha_k} \frac{d\mathscr{M}_i(t)}{n^{-1} \sum_{j=1}^{n} I(X_j \geq t)} + o_p(1) \end{aligned} \tag{4.14}$$

given that

$$\frac{I\{Y(t) > 0\}}{Y(t)} = \begin{cases} \dfrac{1}{Y(t)} & \text{if } Y(t) > 0 \\ 0 & \text{if } Y(t) = 0 \end{cases}$$

By the martingale central limit theorem (section 4.2.1.5) the denominator on the right side of (4.14) can be replaced by its expectation, yielding

$$n^{1/2}(\hat{\Lambda}_k - \Lambda_k) = n^{-1/2} \sum_{i=1}^{n} \int_0^{\alpha_k} \frac{d\mathscr{M}_i(t)}{pr(X \geq t)} + o_p(1) \tag{4.15}$$

which is a sum of n i.i.d. zero-mean random variables. By the Taylor series expansion[15]

$$\begin{aligned} \hat{\Lambda}(t) - \Lambda^*(t) &= \int_0^t \frac{I\{Y(u) > 0\}}{Y(u)} dN(u) - \int_0^t I\{Y(u) > 0\} \lambda(u) du \\ &= \int_0^t \frac{I\{Y(u) > 0\}}{Y(u)} \{dN(u) - Y(u)\lambda(u)du\} \\ &= \sum_{i=1}^{n} \int_0^t \frac{I\{Y(u) > 0\}}{Y(u)} d\mathscr{M}_i(u) \end{aligned}$$

[15] In general, the Taylor series expansion can be used to study the weak convergence of a function of an estimator when the estimator is known to weakly converge to a limit distribution. In particular, if $n^{1/2}(\hat{\Lambda} - \Lambda) = Q + o_p(1)$ for some variable Q, then for some known differentiable function $\varphi(\Lambda)$ the following approximation holds $n^{1/2}(\varphi(\hat{\Lambda}) - \varphi(\Lambda)) \approx \varphi'(\Lambda)Q + o_p(1)$.

In the present application, $\varphi(\Lambda) = e^{-\Lambda} = S$ and $\varphi'(\Lambda) = -e^{-\Lambda} = -S$. It follows therefore that

$$n^{1/2}(\hat{S}_k - S_k) \approx -S_k n^{-1/2} \sum_{i=1}^{n} \int_0^{\alpha_k} \frac{d\mathscr{M}_i(t)}{pr(X \geq t)} + o_p(1)$$

$$
\begin{aligned}
Z_2 &= n^{1/2}\sum_{k=1}^{K} E_k^*(\hat{S}_k - S_k) \\
&= n^{1/2}\sum_{k=1}^{K} E_k^*(-S_k)\frac{n^{-1/2}}{n^{1/2}}\sum_{i=1}^{n}\int_0^{\alpha_k}\frac{d\mathscr{M}_i(t)}{pr(X \geq t)} + o_p(1) \\
&= -n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K} E_k^* S_k \int_0^{\alpha_k}\frac{d\mathscr{M}_i(t)}{pr(X \geq t)} + o_p(1)
\end{aligned}
\tag{4.16}
$$

Combination of (4.12) and (4.16) yields

$$
Z = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K}\xi_{ki} + o_p(1)
$$

where

$$
\xi_{ki} = \frac{S_k Y_{ki}(M_{ki} - E_k^*)}{E(Y_{k1})} - E_k^* S_k \int_0^{\alpha_k}\frac{d\mathscr{M}_i(t)}{pr(X \geq t)} \tag{4.17}
$$

Since for every *i* the random elements involved in ξ_{ki} $(k = 1,...,K)$ pertain to the *i*th individual only, the random variable *Z* is a sum of *n* i.i.d. zero-mean random variables and applying the central limit theorem it follows that *Z* converges in distribution to a zero-mean normal random variable with variance $\sigma^2 = E\left(\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{k1}\xi_{l1}\right)$.[16]

A natural estimator for σ^2 is given by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K}\hat{\xi}_{ki}\hat{\xi}_{li}$, where the $\hat{\xi}_{ki}$'s are obtained from the ξ_{ki}'s by replacing the unknown quantities in (4.17) with their respective sample estimators as follows

$$
\hat{\xi}_{ki} = \frac{\hat{S}_k Y_{ki}(M_{ki} - \hat{E}_k)}{n^{-1}\sum_{j=1}^{n} Y_{kj}} - \hat{E}_k \hat{S}_k \int_0^{\alpha_k}\frac{dN_i(t) - I(X_i \geq t)d\hat{\Lambda}(t)}{n^{-1}\sum_{l=1}^{n} I(X_l \geq t)}
$$

where

---

[16] Expressing (4.17) as $\xi_{ki} = \xi_{ki}^{(1)} + \xi_{ki}^{(2)}$, where $\xi_{ki}^{(1)}$ and $\xi_{ki}^{(2)}$ denote the two terms on the right side of (4.17), then

$$
\sigma^2 = E\left(\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{k1}^{(1)}\xi_{l1}^{(1)}\right) + E\left(\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{k1}^{(2)}\xi_{l1}^{(2)}\right) + 2E\left(\sum_{k=1}^{K}\sum_{l=1}^{K}\xi_{k1}^{(1)}\xi_{l1}^{(2)}\right) \quad (**)
$$

which is a representation for $\mathrm{var}(Z) = \mathrm{var}(Z_1) + \mathrm{var}(Z_2) + 2\,\mathrm{cov}(Z_1, Z_2)$, that is the first two terms on the right side of (**) are the variances due to the variations of the $\hat{E}_k$'s and the $\hat{S}_k$'s respectively and the third term is the covariance. Each of the three terms accounts for the variations within the intervals and for the covariances among the intervals (Lin et al, 1997).

$$\int_0^{\alpha_k} \frac{dN_i(t) - I(X_i \geq t)d\hat{\Lambda}(t)}{n^{-1}\sum_{l=1}^{n} I(X_l \geq t)} = \int_0^{\alpha_k} \frac{dN_i(t)}{n^{-1}\sum_{l=1}^{n} I(X_l \geq t)} - \int_0^{\alpha_k} \frac{I(X_i \geq t)}{n^{-1}\sum_{l=1}^{n} I(X_l \geq t)} \frac{dN(t)}{\sum_{l=1}^{n} I(X_l \geq t)} \quad \text{(because of 4.13)}$$

$$= \int_0^{\alpha_k} \frac{dN_i(t)}{n^{-1}\sum_{l=1}^{n} I(X_l \geq t)} - \sum_{j=1}^{n} \int_0^{\alpha_k} \frac{I(X_i \geq t)dN_j(t)}{n^{-1}\left\{\sum_{l=1}^{n} I(X_l \geq t)\right\}^2}$$

$$= \frac{I(X_i \leq \alpha_k)\delta_i}{n^{-1}\sum_{l=1}^{n} I(X_l \geq X_i)} - \sum_{j=1}^{n} \frac{I(X_i \geq X_j)I(X_j \leq \alpha_k)\delta_j}{n^{-1}\left\{\sum_{l=1}^{n} I(X_l \geq X_j)\right\}^2}$$

Thus the variance estimator for $\hat{\mu}_{LIN1}$ is $\hat{\sigma}^2 / n$ and is given as

$$\hat{\text{var}}(\hat{\mu}_{LIN1}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} W_{ki}W_{li} \tag{4.18}$$

where

$$W_{ki} = \frac{\hat{S}_k Y_{ki}(M_{ki} - \hat{E}_k)}{\sum_{j=1}^{n} Y_{kj}} - \hat{S}_k \hat{E}_k \left\{ \frac{I(X_i \leq a_k)\delta_i}{R_i} - \sum_{j:X_j \leq \min(a_k, X_i)} \frac{\delta_j}{R_j^2} \right\} \tag{4.19}$$

and

$$R_i = \sum_{l=1}^{n} I(X_l \geq X_i) \tag{4.20}$$

4.3.2.2. Estimator of mean cost when cost histories are not recorded

The approach for estimating mean cost when individual cost histories are not recorded (referred to as Lin2 below) again entails partitioning the duration of the study into subintervals $[\alpha_k, \alpha_{k+1})$, but now only observed total costs are being used in the estimation process. In this case the mean total cost μ can be given as

$$\mu = \sum_{k=1}^{K} E\left(M \middle| \alpha_k \leq T < \alpha_{k+1}\right) pr\left(\alpha_k \leq T < \alpha_{k+1}\right) + E\left(M \middle| T \geq L\right) pr(T \geq L)$$

$$= \sum_{k=1}^{K+1} E\left(M \middle| \alpha_k \leq T < \alpha_{k+1}\right) pr(\alpha_k \leq T < \alpha_{k+1})$$

$$= \sum_{k=1}^{K+1} A_k (S_k - S_{k+1})$$

with $\alpha_{K+2} = \infty$ and $A_k = E\left(M \middle| \alpha_k \leq T < \alpha_{k+1}\right)$. This leads to the following estimator for μ

$$\hat{\mu}_{LIN2} = \sum_{k=1}^{K+1} \hat{A}_k (\hat{S}_k - \hat{S}_{k+1}) \text{ with } \alpha_{K+2} = \infty \tag{4.21}$$

where the survival probabilities S_k are consistently estimated by the Kaplan-Meier method, with $\hat{S}_k - \hat{S}_{k+1}$ being the estimated Kaplan-Meier probability of death over the interval $[\alpha_k, \alpha_{k+1})$, and

$$\hat{A}_k = \frac{\sum_{i=1}^{n} Y_{ki} M_i}{\sum_{i=1}^{n} Y_{ki}} \;,\; k = 1,\ldots,K \tag{4.22}$$

where now $Y_{ki} = I(\alpha_k \leq X_i < \alpha_{k+1}, \delta_i = 1)$ and M_i is the observed total cost of individual *i*. That is, $\hat{A}_k$ is an estimator for mean cost for interval *k* and is derived from those individuals who are observed to die in the interval $[\alpha_k, \alpha_{k+1})$.[17] If censoring occurs at the end of the interval, $\hat{A}_k$ is a consistent estimator of the mean cost A_k for interval *k* since $Y_{ki} = 1$ implies that M_i represents the cost of individual *i* until the point of his death. If censoring occurs at the start of the interval then, given $\{X_i \geq \alpha_k\}$, the failure times have the same probability of being censored in the interval $[\alpha_k, \alpha_{k+1})$, and the observed deaths in $[\alpha_k, \alpha_{k+1})$ are thus a random subset of all deaths in the same interval which implies that $\hat{A}_k$ is still a consistent estimator of the mean cost for interval *k*. If censoring occurs in the interior of the interval, $\hat{A}_k$ is going to be biased towards the costs of those who die early in the interval because given the same censoring distribution larger survival times are associated with a higher probability of being censored.

With respect to the estimator of mean cost for the last interval of the partition $[\alpha_{K+1}, \alpha_{K+2})$, this is defined as $A_{K+1} = E(M|T \geq L)$ and involves the observed total costs of the patients who are censored at *L*. The assumption of extended independent censoring implies $A_{K+1} = E(M|X \geq L)$ and hence A_{K+1} is estimated as

$$\hat{A}_{K+1} = \frac{\sum_{i=1}^{n} Y_{K+1,i} M_i}{\sum_{i=1}^{n} Y_{K+1,i}} \text{ where } Y_{K+1,i} = I(X_i \geq L) \tag{4.23}$$

As it can be seen from the above expressions estimation of the interval cost A_k $(k = 1,\ldots,K+1)$ does not require cost information on those individuals who are censored before the largest observed time *L*.

For large *n*, the estimator for the variance of $\hat{\mu}_{LIN2}$ is derived using the same theoretical framework as for the previous estimator as follows. By the law of the large numbers, the estimators

[17] Again assuming the condition of extended independent censoring, as this implies that $A_k = E\left(M|\alpha_k \leq T < \alpha_{k+1}, U \geq \alpha_k\right) = E\left(M|X \geq \alpha_k, T < \alpha_{k+1}\right)$.

$\hat{A}_k = \dfrac{\sum_{i=1}^{n} Y_{ki} M_i}{\sum_{i=1}^{n} Y_{ki}} (k = 1,..., K+1)$ converge in probability to $A_k^* = E(M | \alpha_k \le X < \alpha_{k+1}, \delta = 1)$, $(k = 1,..., K)$. It then follows from Slutsky's theorem and the consistency of the Kaplan-Meier estimator that $\hat{\mu}_{LIN2} = \sum_{k=1}^{K+1} \hat{A}_k (\hat{S}_k - \hat{S}_{k+1})$ converges in probability to $\mu_{LIN2}^* = \sum_{k=1}^{K+1} A_k^* (S_k - S_{k+1})$ where $A_{K+1}^* = A_{K+1}$. Letting $Z = n^{1/2}\left(\hat{\mu}_{LIN2} - \mu_{LIN2}^*\right)$, it follows that

$$
\begin{aligned}
Z &= n^{1/2}\left(\sum_{k=1}^{K+1} \hat{A}_k \left(\hat{S}_k - \hat{S}_{k+1}\right) - \sum_{k=1}^{K+1} A_k^* \left(S_k - S_{k+1}\right)\right) \\
&= n^{1/2} \sum_{k=1}^{K+1} \hat{A}_k \left(\hat{S}_k - \hat{S}_{k+1}\right) - n^{1/2} \sum_{k=1}^{K+1} A_k^* \left(\hat{S}_k - \hat{S}_{k+1}\right) + n^{1/2} \sum_{k=1}^{K+1} A_k^* \left(\hat{S}_k - \hat{S}_{k+1}\right) - n^{1/2} \sum_{k=1}^{K+1} A_k^* \left(S_k - S_{k+1}\right) \\
&= n^{1/2} \sum_{k=1}^{K+1} \left(\hat{S}_k - \hat{S}_{k+1}\right)\left(\hat{A}_k - A_k^*\right) + n^{1/2} \sum_{k=1}^{K+1} A_k^* \left(\hat{S}_k - S_k\right) - n^{1/2} \sum_{k=1}^{K+1} A_k^* \left(\hat{S}_{k+1} - S_{k+1}\right) \\
&= Z_1 + Z_2 - Z_3, \quad \text{say.}
\end{aligned}
$$

Due to the consistency of the $\hat{S}_k$'s,

$$
\begin{aligned}
Z_1 &= n^{1/2} \sum_{k=1}^{K+1} \left(S_k - S_{k+1}\right)\left(\hat{A}_k - A_k^*\right) + o_p(1) \\
&= \sum_{k=1}^{K+1} \left(S_k - S_{k+1}\right) \frac{n^{-1/2}}{n^{-1}} \left(\frac{\sum_{i=1}^{n} Y_{ki} M_i}{\sum_{i=1}^{n} Y_{ki}} - A_k^* \right) + o_p(1) \\
&= \sum_{k=1}^{K+1} \left(S_k - S_{k+1}\right) \frac{n^{-1/2} \sum_{i=1}^{n} Y_{ki} \left(M_i - A_k^*\right)}{n^{-1} \sum_{i=1}^{n} Y_{ki}} + o_p(1)
\end{aligned}
$$

By the central limit theorem and the law of large numbers, the random variable $n^{-1/2} \sum_{i=1}^{n} Y_{ki} (M_i - A_k^*)$ is asymptotically zero-mean normal and the random variable $n^{-1} \sum_{i=1}^{n} Y_{ki}$ converges in probability to the constant $E(Y_{k1})$. It then follows from Slutsky's theorem that

$$
Z_1 = n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K+1} \frac{\left(S_k - S_{k+1}\right) Y_{ki} (M_i - A_k^*)}{E(Y_{k1})} + o_p(1) \tag{4.24}
$$

which is a sum of n independent and identically distributed (i.i.d.) zero-mean random variables. By the same arguments as given in the study of the large-sample properties of $\hat{\mu}_{LIN1}$, an i.i.d. representation for Z_2 is given as

$$
Z_2 = -n^{-1/2} \sum_{i=1}^{n} \sum_{k=1}^{K+1} A_k^* S_k \int_0^{\alpha_k} \frac{d\mathcal{M}_i(t)}{pr(X \ge t)} + o_p(1) \tag{4.25}
$$

and for Z_3

$$Z_3 = -n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K+1} A_k^* S_{k+1}\int_0^{\alpha_{k+1}} \frac{d\mathscr{M}_i(t)}{pr(X \geq t)} + o_p(1) \tag{4.26}$$

where $\mathscr{M}_i(t)\,(i = 1,...,n)$ are the martingale processes introduced above. Combination of (4.24), (4.25) and (4.26) yields

$$Z = n^{-1/2}\sum_{i=1}^{n}\sum_{k=1}^{K+1}\xi_{ki} + o_p(1)$$

where

$$\xi_{ki} = \frac{(S_k - S_{k+1})Y_{ki}(M_i - A_k^*)}{E(Y_{k1})} + A_k^*\left\{S_{k+1}\int_0^{\alpha_{k+1}} \frac{d\mathscr{M}_i(t)}{pr(X \geq t)} - S_k\int_0^{\alpha_k} \frac{d\mathscr{M}_i(t)}{pr(X \geq t)}\right\} \tag{4.27}$$

Since for every *i* the random elements involved in $\xi_{ki}\ \ (k = 1,...,K)$ pertain to the *i*th individual only, the random variable *Z* is a sum of *n* i.i.d. zero-mean random variables and applying the central limit theorem it follows that *Z* converges in distribution to a zero-mean normal random variable with variance $\sigma^2 = E\left(\sum_{k=1}^{K+1}\sum_{l=1}^{K+1}\xi_{k1}\xi_{l1}\right)$.

As stated previously, a natural estimator for σ^2 is given by $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\sum_{k=1}^{K+1}\sum_{l=1}^{K+1}\hat{\xi}_{ki}\hat{\xi}_{li}$, where the $\hat{\xi}_{ki}$'s are obtained from the ξ_{ki}'s by replacing the unknown quantities in (4.27) with their respective sample estimators as follows

$$\hat{\xi}_{ki} = \frac{(\hat{S}_k - \hat{S}_{k+1})Y_{ki}(M_i - \hat{A}_k)}{n^{-1}\sum_{j=1}^{n} Y_{kj}} + n\hat{A}_k\left(\hat{S}_{k+1}D_{k+1,i} - \hat{S}_k D_{ki}\right)$$

where

$$D_{ki} = \frac{I(X_i \leq \alpha_k)\delta_i}{R_i} - \sum_{j:X_j \leq \min(\alpha_k, X_i)} \frac{\delta_j}{R_j^2}$$

and

$$R_i = \sum_{l=1}^{n} I(X_l \geq X_i).$$

Hence for large *n*, $\hat{\mu}_{LIN2}$ is approximately normal with variance estimator given as $\hat{\sigma}^2 / n$, i.e.

$$\hat{\text{var}}(\hat{\mu}_{LIN2}) = \sum_{i=1}^{n}\sum_{k=1}^{K+1}\sum_{l=1}^{K+1} W_{ki}W_{li} \tag{4.28}$$

where

$$W_{ki} = \frac{(\hat{S}_k - \hat{S}_{k+1})Y_{ki}(M_i - \hat{A}_k)}{\sum_{j=1}^{n} Y_{kj}} + \hat{A}_k(\hat{S}_{k+1}D_{k+1,i} - \hat{S}_k D_{ki}) \tag{4.29}$$

$$D_{ki} = \frac{I(X_i \le a_k)\delta_i}{R_i} - \sum_{j:X_j \le \min(a_k, X_i)} \frac{\delta_j}{R_j^2} \tag{4.30}$$

$$R_t = \sum_{l=1}^{n} I(X_l \ge X_t) \tag{4.31}$$

Due to the consistency of the Kaplan-Meier estimator, the estimators $\hat{\mu}_{LIN1}$ and $\hat{\mu}_{LIN2}$ are consistent as long as the $\hat{E}_k$'s and $\hat{A}_k$'s are consistent. Their consistency as shown above is dependent on the censoring pattern and is ensured if censoring occurs at the boundaries of the intervals of the partition. If the censoring distribution is discrete, the boundaries α_k's can in theory be chosen to correspond to the possible censoring times and therefore the estimators are still going to be consistent. If the censoring distribution is continuous, the shorter the interval length, that is the finer the partition of the study period, the more unbiased the estimators. There is however a constraint associated with this point with reference to Lin et al's second approach (Lin2), which requires that the length of the intervals of the partition is such that allows a reasonable number of deaths to be observed in each subinterval. It may not be possible however to meet this requirement while simultaneously ensuring that the censoring times are confined to the boundaries of the intervals of the partition as required for consistency.

4.3.3. Bang and Tsiatis estimators

The set of estimators proposed by Bang and Tsiatis (2000) do not impose any restrictions on the distribution of censoring times. The idea underlying this class of estimators is the use of an inverse probability weight in the estimating equations through which censoring is appropriately accounted for. The first estimator uses cost information from only the uncensored cases while the second uses intermediate cost history from all study subjects. The last two estimators build on the work of Robins & Rotnitzky (1992) and Robins, Rotnitzky & Zhao (1994) and attempt to improve efficiency by recovering information lost due to censoring.

The same notation as above is adopted here and the assumptions underlying the Bang and Tsiatis estimators are as follows. The distribution of failure time T is assumed continuous from 0 to L, the censoring distribution is assumed continuous with the random variable U denoting time to censoring having survivor function $K(u) = pr(U > u)$, i.e. the survivor function $K(u)$ evaluated at a point in time u gives the probability of an individual not being censored at u, and censoring is assumed to arise completely at random. A further assumption is that $pr(U_i \ge L) > 0$ which ensures

that $K(u)$ is bounded away from zero and that a number of patients are still under observation at L to enable calculation of the cost over the defined period $(0, L]$.

4.3.3.1. Simple weighted estimator

All the proposed estimators originate from the weighted complete-case estimator. If complete costs were available for each patient, then an obvious estimator for the mean cost μ would be $\frac{1}{n}\sum_{i=1}^{n} M_i$.

Under conditions of independent censoring, an estimator accounting for censoring using cost information from uncensored individuals only is given as $\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{K(T_i)}$.

The idea underlying the use of this specific probability weight to adjust for censoring is that under conditions of independent censoring, at time T_i, $K(T_i) = pr(U > T_i)$ is the probability that individual i has survived to T_i without being censored. Therefore, if individual i is observed to die at T_i, then he represents $1/K(T_i)$ individuals who might have been observed if there was no censoring. This is an unbiased estimator of μ as is shown below.

$$
\begin{aligned}
E\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{K(T_i)}\right\} &= E\left[\frac{1}{n}\sum_{i=1}^{n} E\left\{\frac{\delta_i M_i}{K(T_i)}\middle| T_i, M_i(\cdot)\right\}\right] \\
&= E\left[\frac{1}{n}\sum_{i=1}^{n}\frac{M_i}{K(T_i)} E\{I(U_i \geq T_i)| T_i, M_i(\cdot)\}\right] \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n} M_i\right) = \mu
\end{aligned}
$$

The unknown survivor function $K(\cdot)$ is estimated by the Kaplan-Meier estimator based on the data $\{X_i = \min(T_i, U_i), 1-\delta_i, i = 1,\ldots,n\}$ as

$$\hat{K}(t) = \prod_{u \leq t}\left\{1 - \frac{dN^c(u)}{Y(u)}\right\} \tag{4.32}$$

where $N^c(u) = \sum I(X_i \leq u, \delta_i = 0)$ and $Y(u) = \sum I(X_i \geq u)$.

The simple weighted complete-case estimator is then defined as

$$\hat{\mu}_{WT} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{\hat{K}(T_i)} \tag{4.33}$$

and its consistency is shown in Appendix A.4.1. Proof of asymptotic normality for this estimator and derivation of its variance for large samples are based on the theory of counting processes and the associated martingale framework. The filtration adopted here is given by

$$\mathcal{F}_u = \sigma\{I(U_i \le t),\ t \le u;\ I(T_i \le x),\ M_i(x),\ 0 \le x < \infty,\ i = 1,...,n\}$$

and represents the increasing information over time on the censoring times up to time u and survival times and cost histories over all non-negative times. The counting process $N^c(u)$ counts the number of individuals censored over time, and given that censoring is independent (see section 4.2.1.6) the associated martingale is given as

$$\mathcal{M}_i^c(u) = N_i^c(u) - \int_0^u \lambda^c(t) Y_i(t) dt$$

where $N_i^c(u) = I(X_i \le u, \delta_i = 0)$, $Y_i(u) = I(X_i \ge u)$ and $\lambda^c(u)$ is the hazard function for the censoring distribution. In addition, $\mathcal{M}^c(u) = \sum \mathcal{M}_i^c(u)$, $N^c(u) = \sum N_i^c(u)$ and $Y(u) = \sum Y_i(u)$.

To study the large sample properties of the simple weighted estimator, the authors show that it can be expanded as follows

$$\begin{aligned} n^{1/2}(\hat{\mu}_{WT} - \mu) &= n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i M_i}{\hat{K}(T_i)} - n^{1/2}\mu \\ &= n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i M_i}{K(T_i)} - n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i M_i}{K(T_i)} + n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i M_i}{\hat{K}(T_i)} - n^{1/2}\mu \\ &= n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i M_i}{K(T_i)} + n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i M_i}{K(T_i)} \left\{ \frac{K(T_i) - \hat{K}(T_i)}{\hat{K}(T_i)} \right\} - n^{1/2}\mu \end{aligned}$$

Using the following identity from Robins & Rotnitzky (1992, p. 313)

$$\frac{\delta_i}{K(X_i)} = 1 - \int_0^{\infty} \frac{d\mathcal{M}_i^c(u)}{K(u)}$$

the above expression becomes

$$n^{1/2}(\hat{\mu}_{WT} - \mu) = n^{-1/2} \sum_{i=1}^{n} M_i \left\{ 1 - \int_0^{L} \frac{d\mathcal{M}_i^c(u)}{K(u)} \right\} - \frac{n^{1/2}}{n} n\mu - n^{-1/2} \sum_{i=1}^{n} \frac{\delta_i M_i}{\hat{K}(T_i)} \left\{ \frac{\hat{K}(T_i) - K(T_i)}{K(T_i)} \right\}$$

Using the martingale integral representation

$$\frac{\hat{K}(t)-K(t)}{K(t)} = -\int_0^t \frac{\hat{K}(u^-)}{K(u)} \frac{d\mathscr{M}^c(u)}{Y(u)}$$

(Gill, 1980, p. 37), where $\hat{K}(u^-)$ is the left-continuous version of the Kaplan-Meier estimator for censoring and $n^{-1}Y(u) = \hat{K}(u^-)\hat{S}(u^-)$, with $\hat{S}(u)$ being the Kaplan-Meier estimator for $S(u) = pr(T > u)$, it follows that

$$\begin{aligned}
n^{1/2}(\hat{\mu}_{WT} - \mu) &= n^{-1/2}\sum_{i=1}^{n} M_i - n^{-1/2}\sum_{i=1}^{n}\mu - n^{-1/2}\sum_{i=1}^{n} M_i \int_0^L \frac{d\mathscr{M}_i^c(u)}{K(u)} + n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i M_i}{\hat{K}(T_i)}\left\{\int_0^{T_i}\frac{\hat{K}(u^-)}{K(u)}\frac{d\mathscr{M}^c(u)}{\hat{K}(u^-)\hat{S}(u^-)n}\right\} \\
&= n^{-1/2}\sum_{i=1}^{n}(M_i - \mu) - n^{-1/2}\sum_{i=1}^{n} M_i \int_0^L \frac{d\mathscr{M}_i^c(u)}{K(u)} + n^{-1/2}\int_0^{T_i}\frac{d\mathscr{M}^c(u)}{K(u)}\frac{1}{n}\frac{1}{\hat{S}(u^-)}\sum_{i=1}^{n}\frac{\delta_i M_i}{\hat{K}(T_i)} \\
&= n^{-1/2}\sum_{i=1}^{n}(M_i - \mu) - n^{-1/2}\sum_{i=1}^{n} M_i \int_0^L \frac{d\mathscr{M}_i^c(u)}{K(u)} + n^{-1/2}\sum_{i=1}^{n}\int_0^L \frac{d\mathscr{M}_i^c(u)}{K(u)}\hat{G}(M,u) \\
&= n^{-1/2}\sum_{i=1}^{n}(M_i - \mu) - n^{-1/2}\sum_{i=1}^{n}\int_0^L \frac{d\mathscr{M}_i^c(u)}{K(u)}\{M_i - \hat{G}(M,u)\} \\
&= n^{-1/2}\sum_{i=1}^{n}(M_i - \mu) - n^{-1/2}\sum_{i=1}^{n}\int_0^L \frac{d\mathscr{M}_i^c(u)}{K(u)}\{M_i - G(M,u)\} + o_P(1)
\end{aligned} \tag{4.34}$$

where

$$G(M,u) = \frac{1}{S(u)}E\{M_i I(T_i \geq u)\}$$

and

$$\hat{G}(M,u) = \frac{1}{n}\frac{1}{\hat{S}(u)}\sum_{i=1}^{n}\frac{\delta_i M_i I(T_i \geq u)}{\hat{K}(T_i)} \tag{4.35}$$

The above expression is the sum of n elements, each one pertaining to individual i and hence they are identically and independently (i.i.d.) distributed. Using the martingale theorem presented on page 49 used in deriving the second moments for statistics of the form $U(t) = \sum_{i=1}^{n}\int_0^t H_i d(N_i - A_i) = \sum_{i=1}^{n}\int_0^t H_i(u)d\mathscr{M}_i(u)$ for the case of continuous compensators, the variance of the expression (4.34) is derived as

$$\text{var}\{n^{1/2}(\hat{\mu}_{WT} - \mu)\} = \text{var}(M_i - \mu) + E\left[\int_0^L \{M_i - G(M,u)\}^2 I(T_i \geq u)\frac{\lambda^c(u)}{K(u)}du\right]$$

For large samples, the martingale version of the central limit theorem can be used to show that $\hat{\mu}_{WT}$ is asymptotically normal with consistent variance estimator given by

$$\hat{\text{var}}(\hat{\mu}_{WT}) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(M_i - \hat{\mu}_{WT})^2}{\hat{K}(T_i)} + \frac{1}{n}\int_0^L \frac{dN^c(u)}{\hat{K}^2(u)}\left\{\hat{G}(M^2,u) - \hat{G}^2(M,u)\right\}\right] \tag{4.36}$$

where $\hat{G}(M^2,u)$ and $\hat{G}^2(M,u)$ are defined according to (4.35).

4.3.3.2. Partitioned estimator

The authors also propose a partitioned version of the simple weighted complete-case estimator which makes use of the cost history for the censored observations that are not used by the simple weighted estimator. The idea underlying the partitioned estimator is similar to that proposed by Lin et al (1997) but the advantage of their method is that the consistency and asymptotic normality of the proposed estimator, unlike Lin et al, does not depend on the choice of the partition or the discreteness of the censoring times.[18]

The duration of analysis $(0, L]$ is partitioned into K subintervals $(t_j, t_{j+1}], (j = 0,...,K-1)$, the simple weighted estimator is then used to derive the estimated cost incurred in each of these K subintervals and the final estimate of mean cost is derived by summing across these intervals. The partitioned estimator is therefore given as

$$\hat{\mu}_P = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{\delta_i^j\left\{M_i(t_j) - M_i(t_{j-1})\right\}}{\hat{K}_j(T_i^j)} \tag{4.37}$$

where for individual i: $\delta_i^j = I\left\{\min(T_i, t_j) \le U_i\right\}$, $M_i(t_j)$ is the cumulative cost up to time t_j, $\hat{K}_j(T_i^j)$ is the Kaplan-Meier estimator for the probability of not being censored based on the dataset $\left\{X_i^j, \delta_i^j, i = 1,...,n\right\}$ where $X_i^j = \min(T_i^{t_j}, U_i)$ and $T_i^{t_j} = \min(T_i, t_j)$.

The advantage of this method over the simple weighted estimator is that individual i is considered uncensored in the jth interval whenever $U_i > \min(T_i, t_j)$. Consequently, there is an increase in the cost information being used by this estimator, as individuals who were treated as censored in the simple weighted estimator having $U_i < T_i$ and whose cost information was thus not used in the

[18] Recently in an unpublished paper O'Hagan and Stevens (2002) purport to show formally the link between the Lin et al (1997) estimator not using cost histories and the Bang and Tsiatis (2000) simple weighted estimator as well as the link between the Lin et al (1997) using cost histories estimator and the Bang and Tsiatis (2000) partitioned estimator. The authors' aim is to "make these methods more accessible by clarifying the relationships between them and to facilitate the take-up of more sophisticated techniques". However their paper has a number of weaknesses. First, in establishing the links between the two methodologies the authors essentially remove the assumption of a discrete censoring pattern which underlies the Lin et al estimators. Secondly, in deriving an alternative form for the Bang and Tsiatis partitioned estimator the authors express uncertainty over the equivalence of their estimator with regards to the original. Finally, their conclusion that "parametric modelling is more appropriate for cost-effectiveness decision making" does not follow from the content of their paper.

estimation process will be now uncensored in some of the intervals of the partition in which their costs will contribute to the estimates.

Consistency follows by an argument similar to that used for the simple weighted estimator and proof of asymptotic normality for this estimator and derivation of its variance for large samples are again based on the theory of counting processes and the associated martingale framework. For asymptotic normality, the partitioned estimator is expanded as

$$n^{1/2}(\hat{\mu}_P - \mu) = \sum_{j=1}^{K}\left[n^{-1/2}\sum_{i=1}^{n}\left(M_{ij} - \mu_j\right) - n^{-1/2}\sum_{i=1}^{n}\int_0^{t_j}\frac{d\mathscr{M}_i^c(u)}{K(u)}\left\{M_{ij} - G_j(M_j,u)\right\}\right] + o_p(1)$$

$$= n^{-1/2}\sum_{i=1}^{n}(M_i - \mu) - n^{-1/2}\sum_{i=1}^{n}\int_0^{L}\frac{d\mathscr{M}_i^c(u)}{K(u)}\left[\sum_{j=1}^{K}\left\{M_{ij} - G_j(M_j,u)\right\}I(u < t_j)\right]$$

where μ_j is the true mean cost in interval j, $M_{ij} = M_i(t_j) - M_i(t_{j-1})$,

$G_j(M_l,u) = \dfrac{1}{S_j(u)}E\left\{M_{il}I(T_i^{t_j} \geq u)\right\}$ and $S_j(u) = pr\left\{\min(T_i,t_j) \geq u\right\}$.

Martingale theory then gives the variance of $n^{1/2}(\hat{\mu}_P - \mu)$ as

$$\text{var}(M_i - \mu) + E\int_0^{L}\left[\sum_{j=1}^{K}\left\{M_{ij} - G_j(M_j,u)\right\}I(u < t_j)\right]^2 I(T_i \geq u)\frac{\lambda^c(u)}{K(u)}du$$

For large n, $\hat{\mu}_P$ is approximately normal with variance estimator given by (also see Appendix A.4.2)

$$\hat{\text{var}}(\hat{\mu}_P) = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(M_i - \hat{\mu}_P)^2}{\hat{K}(T_i)} + \int_0^{L}\sum_{j=1}^{K}\sum_{l=1}^{K}\hat{S}_{j\wedge l}(u)\left\{\hat{G}_{j\wedge l}(M_jM_l,u) - \hat{G}_{j\wedge l}(M_j,u)\hat{G}_{j\wedge l}(M_l,u)\right\}\frac{dN^c(u)}{Y(u)\hat{K}(u)}\right] \tag{4.38}$$

where

$$\hat{G}_{j\wedge l}(M_l,u) = \frac{1}{n}\frac{1}{\hat{S}_{j\wedge l}(u)}\sum_{i=1}^{n}\frac{\delta_i^{j\vee l}M_{il}I(T_i^{j\wedge l} \geq u)}{\hat{K}_{j\vee l}(T_i^{j\vee l})} \tag{4.39}$$

$$\hat{G}_{j\wedge l}(M_jM_l,u) = \frac{1}{n}\frac{1}{\hat{S}_{j\wedge l}(u)}\sum_{i=1}^{n}\frac{\delta_i^{j\vee l}M_{ij}M_{il}I(T_i^{j\wedge l} \geq u)}{\hat{K}_{j\vee l}(T_i^{j\vee l})} \tag{4.40}$$

$j \vee l = \max(j,l)$, $j \wedge l = \min(j,l)$, $T_i^{t_j} = T_i^j$ and $\hat{S}_j(u)$ is the Kaplan-Meier estimator of $pr\{\min(T_i,t_j) \geq u\}$.

4.3.3.3. Simple Improved estimator

In an attempt to improve the efficiency of the simple weighted and partitioned estimators, the authors use the theory for missing data processes given by Robins and Rotnizky (1992), and Robins et al. (1994). The idea is that efficiency will be improved through use of some functional of the cost history which will allow recovery of information lost due to censoring.

Estimation and the study of efficiency of the proposed estimators are based on the general theory for semiparametric models when data are missing at random. The development of such models, which consist of both parametric and non parametric components, has been motivated mainly to address the problem of misspecification of econometric and statistical models in a number of applications. In addressing this issue, the semiparametric approach allows the functional form of some components of the model to be unknown and therefore unrestricted. Given that part of the model is completely unspecified, estimation of the parameters of interest requires that some assumptions be made or restrictions be imposed on the statistical relationship between what is observed and what is not observed. Assessment of the asymptotic efficiency of any given semiparametric estimator is performed by comparing the estimator's asymptotic variance with a standard variance measure referred to as the asymptotic variance bound or the semiparametric efficiency bound. Efficiency bounds therefore provide a standard against which the semiparametric estimator's asymptotic efficiency can be assessed and thus provide a means for measuring the loss of efficiency resulting from adopting a semiparametric rather than a parametric model. To ensure the existence of a semiparametric efficiency bound the estimators must be regular.[19] The class of regular estimators excludes both superefficient[20] estimators and estimators that make use of more information that is contained in the semiparametric model. Regularity conditions lead to the following definition of efficiency. An estimator for the parameters of interest of a semiparametric model is said to be efficient if it is regular and its limiting distribution is zero mean normal with the asymptotic variance attaining the semiparametric efficiency bound. Regularity conditions can be easily derived for a particular class of estimators referred to as asymptotically linear estimators. Furthermore establishing regularity conditions for asymptotically linear estimators not only ensures

---

[19] An estimator $\hat{\alpha}$ is said to be regular if it is regular in every parametric submodel and its limiting distribution does not depend on the parametric submodel. Assuming that the data are generated by a parametric model that satisfies the semiparametric assumptions and contains the truth, such a model is referred to as a parametric submodel where the "sub" prefix indicates that the model is a subset of the model consisting of all distributions satisfying the assumptions of the semiparametric model. Assuming further that the data generating process is one where for each sample size *n* the data are distributed according to a parameter θ_n with $\sqrt{n}(\theta_n - \theta_0)$ bounded, an estimator $\hat{\alpha}$ is regular in a parametric submodel if $\forall \theta_0$ the limiting distribution of $\sqrt{n}(\hat{\alpha} - \alpha(\theta_n))$ does not depend on the data generating process (Newey 1990).

[20] This condition is required to ensure that convergence of the estimator in distribution is uniform in the true parameter values which in turn implies that the limiting distribution is continuous in the parameters. Typically the limiting distribution of superefficient estimators is discontinuous in the parameters (Newey 1990).

the existence of a semiparametric efficiency bound but also allows calculation of the bound (Newey 1990). An estimator $\hat{\alpha}$ of α_0 is asymptotically linear with influence function D if

$$n^{1/2}(\hat{\alpha}-\alpha_0)=n^{-1/2}\sum_i D_i+o_p(1)$$

with $E(D)=0$ and $E(D'D)<\infty$. If $\hat{\alpha}$ is asymptotically linear, then by the central limit theorem and Slutsky's theorem, $n^{1/2}(\hat{\alpha}-\alpha_0)$ is asymptotically normal with mean zero and variance $E(D'D)$. Furthermore, asymptotically linear estimators $\hat{\alpha}^{(1)}$ and $\hat{\alpha}^{(2)}$ with the same influence function are asymptotically equivalent in the sense that $n^{1/2}(\hat{\alpha}^{(1)}-\hat{\alpha}^{(2)})=o_p(1)$. Conversely, two asymptotically linear estimators that are asymptotically equivalent must have the same influence function. Hence, the asymptotic properties of such asymptotical linear estimators are directly related to their influence function (Newey 1990).

On the basis of the above exposition and on the premise that most estimators are asymptotically linear, the issue in the application of interest becomes to identify the class of influence functions for regular asymptotically linear estimators when the data may be censored. When there is no censoring, i.e. when complete cost information is available for each individual, given that the parameter $\mu=E(M)$ is an explicit function of the distribution of the random variable M, and that the distribution of M is left unrestricted, there exists only one such influence function for regular asymptotically linear estimators of μ, namely $M_i-\mu$ (Newey 1990). This corresponds to the influence function of the sample average. Introduction of random censoring with unspecified distribution makes the class of influence functions infinite. Following Robins and Rotnitzky (1992), Bang and Tsiatis show that the entire class of influence functions in the presence of censoring is

$$M_i-\mu-\int_0^\infty \frac{d\mathscr{M}_i^C(u)}{K(u)}\{M_i-G(M,u)\}+\int_0^\infty \frac{d\mathscr{M}_i^C(u)}{K(u)}[e\{M_i^H(u)\}-G(e\{M^H(u)\},u)] \qquad (4.41)$$

where $e\{M_i^H(u)\}$ is an arbitrary functional of the cost history and

$$G(e\{M^H(u)\},u)=\frac{1}{S(u)}E\left[e\{M_i^H(u)\}I(T_i\geq u)\right]$$

The estimator of mean cost whose influence function is given by (4.41) is then of the form

$$\hat{\mu}_{gen}=\frac{1}{n}\sum_{i=1}^n\frac{\delta_i M_i}{\hat{K}(T_i)}+\frac{1}{n}\sum_{i=1}^n\int_0^\infty\frac{d\mathscr{M}_i^C(u)}{\hat{K}(u)}[e\{M_i^H(u)\}-\hat{G}^*(e\{M^H(u)\},u)]$$

where

$$\hat{G}^*(e\{M^H(u)\},u)=\frac{\sum_{i=1}^n e\{M_i^H(u)\}Y_i(u)}{Y(u)} \qquad (4.42)$$

Hence all consistent asymptotically normal regular estimators for mean cost are asymptotically equivalent to some estimator of this form. Determining therefore the optimal set of functionals of cost history $e_{opt}\{M_i^H(u)\}$ will result in deriving the most efficient estimator within this class. The optimal vector of these functionals has been shown to be $e_{opt}\{M_i^H(u)\} = E\{M_i | M_i^H(u)\}$ (Robins et al, 1994; Laan and Hubbard, 1998). Bang and Tsiatis argue that as it is practically impossible to estimate this conditional expectation without imposing specific assumptions on the cost histories, an alternative approach to improve efficiency is to specify a fixed number of functionals $[e_1\{M^H(u)\},...,e_J\{M^H(u)\}]$ thus restricting the class of influence functions to

$$M_i - \mu - \int_0^\infty \frac{d\mathscr{M}_i^c(u)}{K(u)}\{M_i - G(M,u)\} + \sum_{j=1}^{J}\gamma_j \int_0^\infty \frac{d\mathscr{M}_i^c(u)}{K(u)}[e_j\{M_i^H(u)\} - G(e_j\{M^H(u)\},u)]$$

and determine the set of constants γ_j, $j = 1,...,J$ which minimise the variance of the above expression. Provided that the vector of prespecified functionals makes the restricted class of influence functions a good approximation to the entire class, the constants $[\gamma_1,...,\gamma_J]$ which minimise the variance above will result in the identification of estimators close to efficient. As $(M_i - \mu)$ is independent of the other terms, the set of optimal constants $\gamma_1^{opt},...,\gamma_J^{opt}$ are the solution to minimising the variance of $(y_i - \gamma_1 z_{i1} - ... - \gamma_J z_{iJ})$ for individual *i*, where

$$y_i = \int_0^\infty \frac{d\mathscr{M}_i^c(u)}{K(u)}\{M_i - G(M,u)\},$$

$$z_{ij} = \int_0^\infty \frac{d\mathscr{M}_i^c(u)}{K(u)}[e_j\{M_i^H(u)\} - G(e_j\{M^H(u)\},u)] \quad , \quad j = 1,...,J .$$

The optimal set of constants are then derived by formalising the above as a multiple regression problem as $\gamma^{opt} = \mathrm{cov}(y_i, Z_i)\,\mathrm{var}(Z_i)^{-1}$, where Z_i is a $1\times J$ vector and z_{ij} and y_i are scalars. The simple improved estimator is then given by

$$\hat{\mu}_{imp} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{\hat{K}(T_i)} + \frac{1}{n}\hat{\gamma}\sum_{i=1}^{n}\int_0^\infty \frac{dN_i^c(u)}{\hat{K}(u)}[e\{M_i^H(u)\} - \hat{G}^*(e\{M^H(u)\},u)] \qquad (4.43)$$

where $\hat{\gamma} = \hat{\mathrm{cov}}(y_i, Z_i)\,\hat{\mathrm{var}}(Z_i)^{-1}$, $e\{M_i^H(u)\}$ is the $J\times 1$ vector of the prespecified functionals $e_j\{M_i^H(u)\}$ which the authors suggest taking as $e_j\{M_i^H(u)\} = M_{ij}(u)$ if $u > t_j$ and zero otherwise, where $M_{ij}(u)$ is the cost incurred in subinterval $(t_{j-1}, \min(t_j,u)]$ and $\hat{G}^*(e\{M^H(u)\},u)$ is the $J\times 1$ vector of $\hat{G}^*(e_j\{M^H(u)\},u)$ where $\hat{G}^*(\cdot)$ is defined by (4.42).

For large samples the variance is estimated by

$$\hat{\text{var}}(\hat{\mu}_{imp}) =$$

$$\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\left(M_i-\hat{\mu}_{imp}\right)^2}{\hat{K}(T_i)}+\frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left\{\hat{G}(M^2,u)-\hat{G}^2(M,u)\right\}-\hat{\text{cov}}(y_i,Z_i)\,\hat{\text{var}}(Z_i)^{-1}\,\hat{\text{cov}}(y_i,Z_i)'\right] \tag{4.44}$$

The jth element in the 1xJ vector of the estimator of $\text{cov}(y_i,Z_i)$ is

$$\frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left(\frac{1}{n}\frac{1}{\hat{S}(u)}\sum_{i=1}^{n}\left[\frac{\delta_i}{\hat{K}(T_i)}\{M_i-\hat{G}(M,u)\}\{M_{ij}(u)-\hat{G}(M_j(u),u)\}I(T_i\geq u)\right]\right) \tag{4.45}$$

and a consistent estimator of $\text{var}(Z_i)$ has its (j, l)th element as

$$\frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left(\frac{1}{n}\frac{1}{\hat{S}(u)}\sum_{i=1}^{n}\left[\frac{\delta_i}{\hat{K}(T_i)}\{M_{ij}(u)-\hat{G}(M_j(u),u)\}\{M_{il}(u)-\hat{G}(M_l(u),u)\}I(T_i\geq u)\right]\right) \tag{4.46}$$

for $j,l=1,\ldots J$.

4.3.3.4. Improved partitioned estimator

When the same methodology is applied to improve the efficiency of the partitioned estimator the resultant improved partitioned estimator is given by

$$\hat{\mu}_{Pimp}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{\delta_i^j\{M_i(t_j)-M_i(t_{j-1})\}}{\hat{K}_j(T_i^j)}+\frac{1}{n}\hat{\gamma}\sum_{i=1}^{n}\int_0^{\infty}\frac{dN_i^c(u)}{\hat{K}(u)}[e\{M_i^H(u)\}-\hat{G}*(e\{M^H(u)\},u)] \tag{4.47}$$

where $\hat{\gamma}=\hat{\text{cov}}(y_i,Z_i)\,\hat{\text{var}}(Z_i)^{-1}$, the vector $\hat{\text{var}}(Z_i)$ and the set of functionals are as defined for the improved simple estimator and the jth element of the 1xJ vector of $\text{cov}(y_i,Z_i)$ is

$$\frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left(\sum_{l=1}^{K}\frac{1}{n}\frac{1}{\hat{S}_l(u)}\sum_{i=1}^{n}\left[\frac{\delta_i^{l\vee j}}{\hat{K}_{l\vee j}(T_i^{l\vee j})}\{M_{il}-\hat{G}_l(M_l,u)\}\{M_{ij}(u)-\hat{G}(M_j(u),u)\}I(T_i^l\geq u)\right]\right) \tag{4.48}$$

The asymptotic variance is estimated by

$$\hat{\text{var}}(\hat{\mu}_{Pimp})=$$

$$\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i\left(M_i-\hat{\mu}_{Pimp}\right)^2}{\hat{K}(T_i)}+\int_0^{L}\sum_{j=1}^{K}\sum_{l=1}^{K}\hat{S}_{j\wedge l}(u)\left\{\hat{G}_{j\wedge l}(M_jM_l,u)-\hat{G}_{j\wedge l}(M_j,u)\hat{G}_{j\wedge l}(M_l,u)\right\}\frac{dN^c(u)}{Y(u)\hat{K}(u)}\right]$$

$$-n^{-1}\,\hat{\text{cov}}(y_i,Z_i)\,\hat{\text{var}}(Z_i)^{-1}\,\hat{\text{cov}}(y_i,Z_i)' \tag{4.49}$$

4.4. Methods and results

The previous sections have presented a number of approaches to estimating mean cost from censored data. All these estimators were applied to the UKPDS data described in chapter 1 and a subset were also applied to a simulated data set, described below, to test various aspects of their performance. The most important feature of the UKPDS data for the purposes of this analysis is the presence of heavy censoring in both trial arms. As a consequence, the results of the present analysis reflect an assessment of the various estimators under extreme censoring conditions which is the issue they all attempt to address. A number of secondary analyses are subsequently performed, some of which attempt to assess the impact of the level of censoring on the performance of the various estimators.

4.4.1. The UKPDS data

As stated in chapter 1, the UKPDS was a randomised controlled clinical trial whose main randomisation involved a type 2 diabetic population of 3867 individuals allocated either to conventional policy (1138) or intensive policy (2729) with the aim of assessing the effectiveness of improved blood glucose control. The trial started in 1978 and ended in 1998 with a median follow-up period to death, the last date at which clinical status was known, or to the end of the trial period of 10 years. For each individual in the study the trial collected information on both clinical effectiveness and resource use. The unit costs of hospitalisation and treatment medication were attached to the volume of resources to calculate the total cost per patient per year directly from the trial data and these were then aggregated to give a total cost per patient for the whole trial period. As noted in chapter 1 costs associated with the non-hospital resource use were excluded from the analysis undertaken here as these were not available from the trial on a patient level basis but had been estimated based on a regression approach.[21] The analysis here aims at deriving an estimate of average total cost over the trial period adjusting for censoring where an observation was defined as censored if the patient was not observed for the full time to death.

A brief description of the data is given in Table 4.2. As can be seen there is no difference in the average duration of follow-up between the conventional and intensive policy population. There is no significant difference in the average cost when the estimates are not adjusted for censoring between the conventional and intensive arms but when only the uncensored cases are considered the conventional group incurs higher costs on average compared to the intensive population. The failure event was all-cause mortality, resulting in 925 censored patients [81.3% censoring] and 213 failures in the conventional group and 2240 censored patients [82% censoring] and 489 failures in the intensive group by the end of the trial. Average follow-up time was equal to 9.9 years reaching a

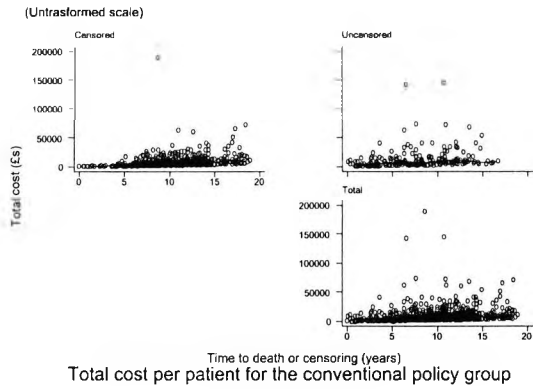
[21] As such, the cost estimates reported in the thesis are not directly comparable to the reported UKPDS economic evaluation results.

maximum of 18.934 years for the conventional group and 10.01 years reaching a maximum of 19.463 years for the intensive group. The assumption of independent censoring is valid in this data as censoring was not related to any cost or medical reasons.

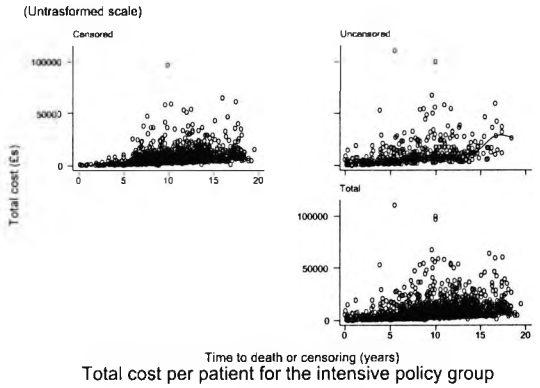
Table 4.2. Descriptive statistics of the UKPDS data

	Conventional	Intensive
Sample size (*n*)	1138	2729
Censored	925 (81.3%)	2240 (82%)
Time duration of analysis (years): mean [range]	9.9 [0.01 to 18.934]	10.01 [0.05 to 19.463]
Time duration of analysis for failures (years): mean [range]	7.66 [0.12 to 16.73]	7.71 [0.05 to 18.41]
Time duration of analysis for censored (years): mean [range]	10.41 [0.01 to 18.934]	10.51 [0.197 to 19.463]
Total cost of all individuals: mean [range]	8348 [119 to 189242]	8070 [20 to 110,921]
Total cost of failures: mean [range]	12586 [220 to 145549]	10857 [20 to 110921]
Total cost of censored: mean [range]	7373 [119 to 189242]	7462 [149 to 97121]

Figures 4.1 and 4.2 give an overview of the observed total cost data for each randomisation group. Figure 4.1 plots the data as a function of time to failure or censoring, while Figure 4.2 plots the observed distribution of costs on the untransformed scale and on a log transformed scale. The graphs in the latter figure are overlain with a normal probability density curve and show that the costs in all groups are positively skewed with a very small number of high outlying costs. As can be seen in Table 4.2 in the conventional population the uncensored individuals have a mean cost of 12,586 ranging from 220 to 145,549. More detailed descriptive statistics showed that 75% of individuals have costs under 14,000; 90% have costs under 30,000; 95% have costs under 42,000; and 99% have costs under 75,000. The censored population in the conventional arm have a mean cost of 7,373 ranging from 119 to 189,242. Again the outliers are small in number: 75% of individuals have costs under 8,500; 90% have costs under 14,500; 95% have costs under 21,000; and 99% have costs under 36,500. In the intensive uncensored population the average cost incurred was 10,857 ranging from 20 to 110,921. Again 75% had costs under 14,000; 90% under 24,500; 95% under 32,500; and 99% under 58,500. In the censored population the mean cost was 7,462 in a range 149 to 97,121, with 75% under 8,900; 90% under 14,000; 95% under 18,000; and 99% under 35,500. Appendix A.4.3 gives some descriptive statistics of the observed annual costs for the two randomisation groups which reveal again a wide spread of costs within each year of follow-up.

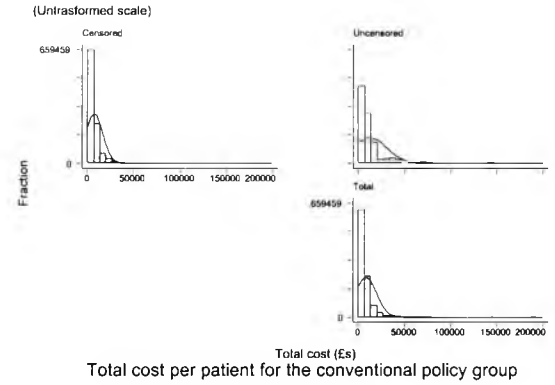


Total cost per patient for the conventional policy group

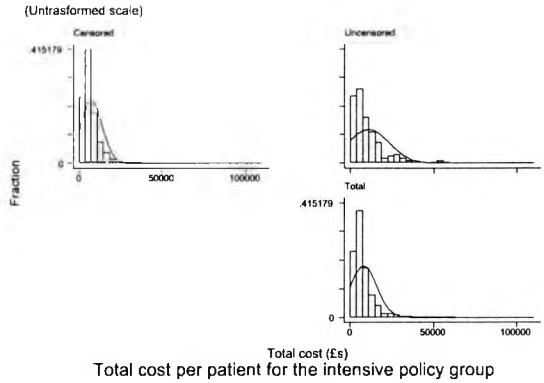


Total cost per patient for the intensive policy group

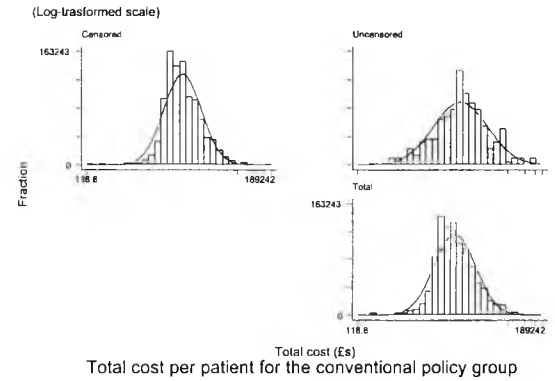
Figure 4.1. Total cost per patient over the study period for conventional and intensive policy groups



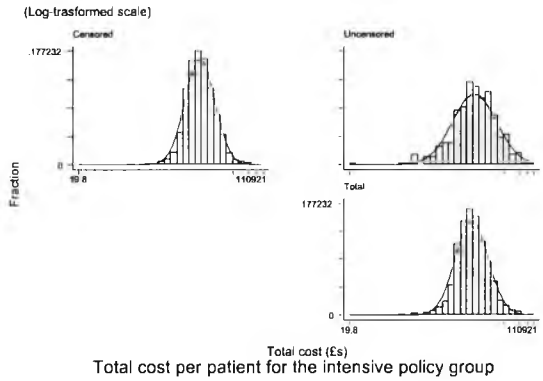
Total cost per patient for the conventional policy group



Total cost per patient for the intensive policy group



Total cost per patient for the conventional policy group



Total cost per patient for the intensive policy group

Figure 4.2. Total cost per patient on the untransformed and on the log-transformed scale for conventional and intensive policy groups

4.4.2. Main analysis

The purpose of the main analysis is to estimate the average total cost per patient over the UKPDS study period. For each individual the observables were time to death or last contact, a variable taking the values of 0 or 1 indicating censoring or failure respectively, the annual costs and the total cost from the start of follow-up to death or the last contact date. All non-parametric estimators considered in this chapter and summarised in Table 4.3 below were applied to these trial data within each arm of the main randomisation, i.e. n=1138 for the conventional policy over a period of (0, 18.934] years and n=2729 for the intensive policy over a period of (0, 19.463] years. Before presenting the results obtained from the various estimators some methodological points with regards to the main analysis follow.

Table 4.3. Non-parametric estimators of mean cost

Kaplan-Meier estimator

$$\hat{\mu}_{KM} = \int_0^C \hat{S}(c)dc$$

where $\hat{S}(c) = \prod_{k \le c}\left\{1 - \frac{\Delta N(k)}{Y(k)}\right\}$, $N(c) = \sum_{i=1}^{n} I(M_i \le c, \delta_i = 1)$, and $Y(c) = \sum_{i=1}^{n} I(M_i \ge c)$

$$\hat{\text{var}}(\hat{\mu}_{KM}) = \int_0^C \left(\int_h^C \hat{S}(c)dc\right)^2 d\left\{\int_0^h \frac{dN(c)}{Y(c)[Y(c) - \Delta N(c)]}\right\}$$

Uncensored cases estimator

$\hat{\mu}_U$ same as in Kaplan-Meier but only using the uncensored data

$\hat{\text{var}}(\hat{\mu}_U)$ same as in Kaplan-Meier but only using the uncensored data

Full-sample estimator

$\hat{\mu}_{FS}$ same as in Kaplan-Meier but treating time of censoring as time of failure for the censored individuals

$\hat{\text{var}}(\hat{\mu}_{FS})$ same as in Kaplan-Meier but treating time of censoring as time of failure for the censored individuals

Lin1 (Using cost histories)

$$\hat{\mu}_{LIN1} = \sum_{k=1}^{K} \hat{S}_k \hat{E}_k$$

where $\hat{S}_k = \prod_{u \le \alpha_k}\left\{1 - \frac{dN(u)}{Y(u)}\right\}$ and $\hat{E}_k = \frac{\sum_{i=1}^{n} Y_{ki} M_{ki}}{\sum_{i=1}^{n} Y_{ki}}$, $k = 1,\ldots,K$, with $Y_{ki} = I(X_i \ge \alpha_k)$

$$\hat{\text{var}}(\hat{\mu}_{LIN1}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\sum_{l=1}^{K} W_{ki} W_{li}$$

where $W_{ki} = \frac{\hat{S}_k Y_{ki}(M_{ki} - \hat{E}_k)}{\sum_{j=1}^{n} Y_{kj}} - \hat{S}_k \hat{E}_k \left\{\frac{I(X_i \le a_k)\delta_i}{R_i} - \sum_{j:X_j \le \min(a_k, X_i)} \frac{\delta_j}{R_j^2}\right\}$ and $R_i = \sum_{l=1}^{n} I(X_l \ge X_i)$

Table 4.3. Non-parametric estimators of mean cost (Contd.)

Lin2 (Not using cost histories)

$$\hat{\mu}_{LIN2} = \sum_{k=1}^{K+1} \hat{A}_k (\hat{S}_k - \hat{S}_{k+1}) \text{ with } \alpha_{K+2} = \infty$$

where $\hat{A}_k = \dfrac{\sum_{i=1}^{r} Y_{ki} M_i}{\sum_{i=1}^{n} Y_{ki}}$, $k = 1,...,K$, with $Y_{ki} = I(\alpha_k \le X_i < \alpha_{k+1}, \delta_i = 1)$

and $\hat{A}_{K+1} = \dfrac{\sum_{i=1}^{n} Y_{K+1,i} M_i}{\sum_{i=1}^{n} Y_{K+1,i}}$ with $Y_{K+1,i} = I(X_i \ge L)$

$$\hat{\text{var}}(\hat{\mu}_{LIN2}) = \sum_{i=1}^{n} \sum_{k=1}^{K+1} \sum_{l=1}^{K+1} W_{ki} W_{li}$$

where $W_{ki} = \dfrac{(\hat{S}_k - \hat{S}_{k+1}) Y_{ki} (M_i - \hat{A}_k)}{\sum_{j=1}^{n} Y_{kj}} + \hat{A}_k (\hat{S}_{k+1} D_{k+1,i} - \hat{S}_k D_{ki})$,

$D_{ki} = \dfrac{I(X_i \le a_k)\delta_i}{R_i} - \sum_{j: X_j \le \min(a_k, X_i)} \dfrac{\delta_j}{R_j^2}$, and $R_i = \sum_{l=1}^{n} I(X_l \ge X_i)$

Simple weighted estimator

$$\hat{\mu}_{WT} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i M_i}{\hat{K}(T_i)}$$

where $\hat{K}(t) = \prod_{u \le t} \left\{ 1 - \dfrac{dN^c(u)}{Y(u)} \right\}$

$$\hat{\text{var}}(\hat{\mu}_{WT}) = \frac{1}{n} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i (M_i - \hat{\mu}_{WT})^2}{\hat{K}(T_i)} + \frac{1}{n} \int_0^L \frac{dN^c(u)}{\hat{K}^2(u)} \left\{ \hat{G}(M^2, u) - \hat{G}^2(M, u) \right\} \right]$$

where $\hat{G}(M,u) = \dfrac{1}{n} \dfrac{1}{\hat{S}(u)} \sum_{i=1}^{n} \dfrac{\delta_i M_i I(T_i \ge u)}{\hat{K}(T_i)}$

Partitioned estimator

$$\hat{\mu}_P = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K} \frac{\delta_i^j \{ M_i(t_j) - M_i(t_{j-1}) \}}{\hat{K}_j(T_i^j)}$$

$$\hat{\text{var}}(\hat{\mu}_P) = \frac{1}{n} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i (M_i - \hat{\mu}_P)^2}{\hat{K}(T_i)} + \int_0^L \sum_{j=1}^{K} \sum_{l=1}^{K} \hat{S}_{j \wedge l}(u) \left\{ \hat{G}_{j \wedge l}(M_j M_l, u) - \hat{G}_{j \wedge l}(M_j, u) \hat{G}_{j \wedge l}(M_l, u) \right\} \frac{dN^c(u)}{Y(u)\hat{K}(u)} \right]$$

where

$$\hat{G}_{j \wedge l}(M_l, u) = \frac{1}{n} \frac{1}{\hat{S}_{j \wedge l}(u)} \sum_{i=1}^{n} \frac{\delta_i^{j \vee l} M_{il} I(T_i^{j \wedge l} \ge u)}{\hat{K}_{j \vee l}(T_i^{j \vee l})}$$

$$\hat{G}_{j \wedge l}(M_j M_l, u) = \frac{1}{n} \frac{1}{\hat{S}_{j \wedge l}(u)} \sum_{i=1}^{n} \frac{\delta_i^{j \vee l} M_{ij} M_{il} I(T_i^{j \wedge l} \ge u)}{\hat{K}_{j \vee l}(T_i^{j \vee l})}$$

Table 4.3. Non-parametric estimators of mean cost (Contd.)

Simple improved estimator

$$\hat{\mu}_{imp} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{\hat{K}(T_i)} + \frac{1}{n}\hat{\gamma}\sum_{i=1}^{n}\int_0^{\infty}\frac{dN_i^c(u)}{\hat{K}(u)}[e\{M_i^H(u)\} - \hat{G}*(e\{M^H(u)\},u)]$$

where

$\hat{\gamma} = \hat{\text{cov}}(y_i, Z_i)\,\hat{\text{var}}(Z_i)^{-1}$

$e\{M_i^H(u)\}$ is the $J\times 1$ vector of $e_j\{M_i^H(u)\}$ with $e_j\{M_i^H(u)\} = M_{ij}(u)$ if $u > t_j$ and zero otherwise, where $M_{ij}(u)$ is the cost incurred in subinterval $(t_{j-1}, \min(t_j, u)]$,

$\hat{G}*(e\{M^H(u)\},u)$ is the $J\times 1$ vector of $\hat{G}*(e_j\{M^H(u)\},u)$ where

$$\hat{G}*(e\{M^H(u)\},u) = \frac{\sum_{i=1}^{n} e\{M_i^H(u)\}Y_i(u)}{Y(u)}$$

$\hat{\text{var}}(\hat{\mu}_{imp}) =$

$$\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(M_i - \hat{\mu}_{imp})^2}{\hat{K}(T_i)} + \frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left\{\hat{G}(M^2,u) - \hat{G}^2(M,u)\right\} - \hat{\text{cov}}(y_i,Z_i)\,\hat{\text{var}}(Z_i)^{-1}\,\hat{\text{cov}}(y_i,Z_i)'\right]$$

The jth element in the $1\text{x}J$ vector of the estimator of $\text{cov}(y_i, Z_i)$ is

$$\frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left(\frac{1}{n}\frac{1}{\hat{S}(u)}\sum_{i=1}^{n}\left[\frac{\delta_i}{\hat{K}(T_i)}\{M_i - \hat{G}(M,u)\}\{M_{ij}(u) - \hat{G}(M_j(u),u)\}I(T_i \geq u)\right]\right)$$

and a consistent estimator of $\text{var}(Z_i)$ has its (j, l)th element as

$$\frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left(\frac{1}{n}\frac{1}{\hat{S}(u)}\sum_{i=1}^{n}\left[\frac{\delta_i}{\hat{K}(T_i)}\{M_{ij}(u) - \hat{G}(M_j(u),u)\}\{M_{il}(u) - \hat{G}(M_l(u),u)\}I(T_i \geq u)\right]\right)$$

for $j, l = 1,\ldots J$

Improved partitioned estimator

$$\hat{\mu}_{Pimp} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{K}\frac{\delta_i^j\{M_i(t_j) - M_i(t_{j-1})\}}{\hat{K}_j(T_i^j)} + \frac{1}{n}\hat{\gamma}\sum_{i=1}^{n}\int_0^{\infty}\frac{dN_i^c(u)}{\hat{K}(u)}[e\{M_i^H(u)\} - \hat{G}*(e\{M^H(u)\},u)]$$

where $\hat{\gamma} = \hat{\text{cov}}(y_i, Z_i)\,\hat{\text{var}}(Z_i)^{-1}$, the vector $\hat{\text{var}}(Z_i)$ and the set of functionals are as defined for the improved simple estimator and the jth element of the $1\text{x}J$ vector of $\text{cov}(y_i, Z_i)$ is

$$\frac{1}{n}\int_0^{\infty}\frac{dN^c(u)}{\hat{K}^2(u)}\left(\sum_{l=1}^{K}\frac{1}{n}\frac{1}{\hat{S}_l(u)}\sum_{i=1}^{n}\left[\frac{\delta_i^{l\vee j}}{\hat{K}_{l\vee j}(T_i^{l\vee j})}\{M_{il} - \hat{G}_l(M_l,u)\}\{M_{ij}(u) - \hat{G}(M_j(u),u)\}I(T_i^l \geq u)\right]\right)$$

$\hat{\text{var}}(\hat{\mu}_{Pimp}) =$

$$\frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i(M_i - \hat{\mu}_{Pimp})^2}{\hat{K}(T_i)} + \int_0^{L}\sum_{j=1}^{K}\sum_{l=1}^{K}\hat{S}_{j\wedge l}(u)\left\{\hat{G}_{j\wedge l}(M_jM_l,u) - \hat{G}_{j\wedge l}(M_j,u)\hat{G}_{j\wedge l}(M_l,u)\right\}\frac{dN^c(u)}{Y(u)\hat{K}(u)}\right]$$

$$- n^{-1}\,\hat{\text{cov}}(y_i,Z_i)\,\hat{\text{var}}(Z_i)^{-1}\,\hat{\text{cov}}(y_i,Z_i)'$$

For each of the Lin et al estimators, two sets of results are presented. The first was obtained when the study duration was partitioned into yearly intervals, given that individual costs were available from the trial on an annual basis, while the second was obtained assuming a monthly interval length with the individual's monthly cost calculated as the annual cost divided by twelve. This was undertaken to assess the impact that the interval length of the partition has on the estimates as the validity of the Lin et al approach relies on the pattern of the censoring distribution being such that censoring times can be confined to the boundaries of the intervals of the partition.

The Bang and Tsiatis partitioned and improved partitioned estimators are based on yearly subintervals for the same reason as stated above, that is, because intermediate cost history for each subject was available on an annual basis. Also for this reason, for both simple improved and partitioned improved estimators, the class of influence functions was restricted by defining a fixed number of prespecified functionals $[e_1\{M^H(u)\},...,e_J\{M^H(u)\}]$, where J=19 for conventional and J=20 for intensive, i.e. annual subintervals were assumed in the recovery of cost information lost due to censoring, and the set of prespecified functionals were defined in accordance to the authors' suggestion as $e_j\{M_i^H(u)\} = M_{ij}(u)$ if $u > t_j$ and zero otherwise, where $M_{ij}(u)$ is the cost incurred in subinterval $(t_{j-1}, \min(t_j, u)]$.

With regards to the stochastic integrals appearing in the Bang and Tsiatis estimators presented above, these are of the form $\int_0^L f(u)dN^c(u) = \int_{0,L]} f(u)dN^c(u)$ (or $\int_0^\infty f(u)dN^c(u)$) where $f(\cdot)$ is some function of time and $0 \le L \le \infty$. Both $f(\cdot)$ and $N^c(\cdot)$ satisfy the required properties stated in section 4.2.1.2 to ensure that the above integrals are well defined as Lebesgue-Stieltjes integrals. Moreover, given that $N^c(u) = \sum_{i=1}^n N_i^c(u) = \sum_{i=1}^n I(X_i \le u, \delta_i = 0)$ and as a step function has countably many jumps at $\{u_1, u_2, ...\}$ with $\Delta N^c(u_k) = N^c(u_k) - N^c(u_k-) > 0$, the above integrals were evaluated as follows.

$$\int_0^L f(u)dN^c(u) = \sum_{i=1}^n \int_0^L f(u)dN_i^c(u)$$

$$= \int_0^L f(u)dN_1^c(u) + \int_0^L f(u)dN_2^c(u) + ... + \int_0^L f(u)dN_n^c(u)$$

$$= \sum_{k:0<u_k\le L} f(u_k)\Delta N_1^c(u_k) + \sum_{k:0<u_k\le L} f(u_k)\Delta N_2^c(u_k) + ... + \sum_{k:0<u_k\le L} f(u_k)\Delta N_n^c(u_k)$$

$$= f(u)\Big|_{u=X_1 \text{ and } \delta_1=0} + f(u)\Big|_{u=X_2 \text{ and } \delta_2=0} + ... + f(u)\Big|_{u=X_n \text{ and } \delta_n=0}$$

$$= \sum_{i=1}^n f(u)\Big|_{u=X_i \text{ and } \delta_i=0}$$

where for each individual i, the Stieltjes integral $\int_0^L f(u)dN_i^c(u) = \sum_{k:0<u_k\le L} f(u_k)\Delta N_i^c(u_k)$ represents the sum of the values of $f(\cdot)$ at the jump times (u_k) of $N_i^c(u)$ in the interval $(0, L]$ (or $(0,\infty]$) where the jumps of the paths of the process $N_i^c(u)$ are of size +1 at the time of censoring for individual i, i.e. at $u = X_i$ with $\delta_i = 0$.

Finally, despite the fact that the Kaplan-Meier, the uncensored cases estimator and the full-sample estimators are all known to be biased as concluded previously in the theoretical section, these were still applied to the data for purposes of comparison.

4.4.3. Results of the main analysis

Table 4.4 presents the estimates of mean cost and the associated variances for the conventional and the intensive policy groups over the study period as derived from application of the above non-parametric estimators to the UKPDS data. Programming was undertaken in Stata 7.0 and the programs are presented in Appendix A.4.4. for the Kaplan-Meier, the full sample and the uncensored cases estimators, in Appendix A.4.5. for the Lin estimators and in Appendix A.4.6. for the Bang and Tsiatis estimators.

Table 4.4. Results of the main analysis

	Conventional		**Intensive**	
Estimator	*Mean*	*Standard error*	*Mean*	*Standard error*
Kaplan-Meier ($\hat{\mu}_{KM}$)	38770.74	5312.02	31620.59	2034.89
Uncensored cases only ($\hat{\mu}_U$)	11901.01	1061.36	10629.97	510.00
Full sample ($\hat{\mu}_{FS}$)	8181.581	305.62	8029.867	146.79
Lin et al				
Subintervals in years				
Lin1 ($\hat{\mu}_{LIN1}$)	14006.2	897.73	13172	340.55
Lin2 ($\hat{\mu}_{LIN2}$)	12428	636.93	16910.64	1010.87
Subintervals in months				
Lin1 ($\hat{\mu}_{LIN1}$)	13771.35	1025.60	13078.02	365.95
Lin2 ($\hat{\mu}_{LIN2}$)	12530.39	668.21	16926.22	1012.53
Bang and Tsiatis				
Simple weighted ($\hat{\mu}_{WT}$)	5732.735	840.8	9737.65	3043.5
Partitioned ($\hat{\mu}_P$)	14639.48	1219.4	13839.67	445.6
Simple improved ($\hat{\mu}_{imp}$)	3668.924	398.1	1620.974	1634
Partitioned improved ($\hat{\mu}_{Pimp}$)	334563.3	variance<0	-326298.3	variance<0

As expected the Kaplan-Meier estimator returns very high values of the average total cost for both arms, that are of a different order of magnitude compared to all others (excluding the partitioned improved estimator which displays results outside the permitted range of values), while the full sample estimator gives lower estimates for both groups, also as expected, since it does not take into account the costs incurred past the censoring times. The uncensored cases estimator is also known to be biased towards the costs of the complete cases who are likely to have shorter survival times since it is based on this subset alone, which in the UKPDS data represent a very small proportion of the total number of subjects.

Concentrating on the estimates for mean cost obtained with each of the two methods proposed by Lin et al, the results show that the length of the intervals of the partition does not have an impact on the estimates returned by either the Lin1 or the Lin2 estimators. This finding holds for both trial arms.

All estimators apart from Lin2 and the Bang and Tsiatis simple weighted display higher estimates for the conventional group compared to the intensive. One could argue that the conventional policy group incurring higher costs on average is probably indicative of the "true" result, as the intensive policy group were known to have significantly lower hospitalisation rates. In addition, the "naïve" estimators resulted in the same direction of difference in mean cost between the two groups which could be interpreted in support of the previous argument in the following manner. Despite the fact that the uncensored cases estimator is biased toward the costs of the individuals with shorter survival times as longer survival times are more likely to be censored, the trial data has not shown a significant difference in survival, both with respect to the proportion dying and the length of survival time, between the two groups and therefore one may assume that the degree of bias imparted in the uncensored cases estimator is similar between the two groups. Along similar lines, although the full sample estimator is known to be biased downward as the costs incurred after censoring times are not accounted for, it could again be argued that the degree of bias in the estimates is similar between the two arms on the basis that the trial data show the same proportion of censoring in the two groups and that this similarity is also maintained over time. All information from the trial is suggestive therefore of the conventional policy group incurring higher costs than the intensive policy population.

On this basis, the fact that the Lin2 and the Bang and Tsiatis simple weighted estimators display lower estimates for the conventional group compared to the intensive group in direct contrast to the results obtained from all other estimators, gives a first indication of poor performance. This statement may be supported by the following observations. First, there is a similarity between the Lin2 and the Bang and Tsiatis simple weighted estimators in that they both use only the complete cost observations in estimating mean cost. Lin et al explicitly state that their second approach relies on a "reasonable" number of deaths in each sub-interval of the partition and suggest a minimum number of 5 deaths in each subinterval. The number of deaths in the UKPDS data is small in the

majority of subintervals and decreases markedly towards the end of the trial, resulting in a number of deaths below 5 or even zero in the last intervals of the partition. The Bang and Tsiatis simple weighted estimator not only displays the same pattern as stated above with respect to the direction of the difference in mean costs between the two arms, but it also results in low values of mean cost for both arms which are totally unlikely to be true since they are even lower than the respective mean costs estimated by the full sample estimator. Although the Bang and Tsiatis simple weighted estimator does not rely on the pattern of the censoring distribution and therefore the small number of complete cost observations does not affect the estimates in the same manner as in the Lin et al second approach, the authors state however that caution should be exercised when applying all their estimators in circumstances where there is heavy censoring in the tails of the distribution with small sample sizes which is precisely the case in the UKPDS data as can also be seen from Figure 4.3.

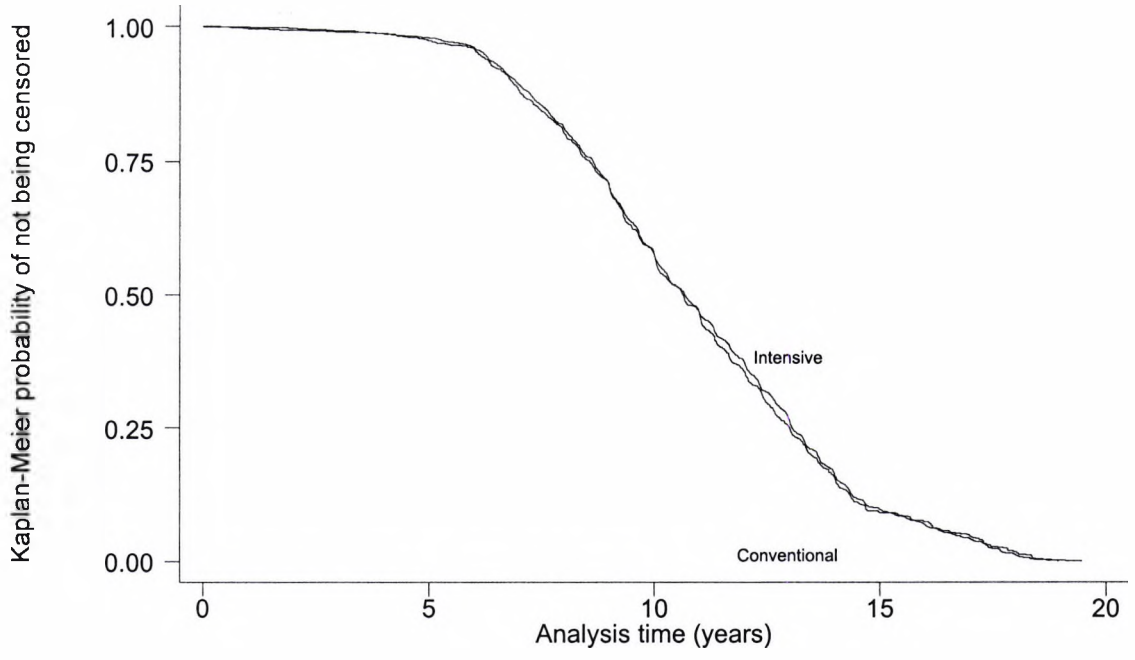


Figure 4.3. Kaplan-Meier estimates for the probability of an individual not being censored

The Bang and Tsiatis simple improved estimator gives even lower estimates of average cost for both arms than the simple weighted estimator. Once again this probably reflects the heavy censoring experienced at the tails of the distribution. The Bang and Tsiatis partitioned improved estimator performs very poorly resulting in mean cost estimates of extreme magnitude including negative values for mean cost for the intensive arm and for variances in both arms. Furthermore, although the same level of censoring affects both improved estimators, the improved simple does not result in such extreme values as observed in the improved partitioned. This finding suggests that the high degree of censoring in particular at the tails of the distribution, makes the improved partitioned much more unstable than the simple improved. A direct consequence of heavy censoring

at the tails is that the probability of an individual not being censored reaches very small values some of which approach zero. Thus any quantity weighted by the inverse of such probabilities will be of extremely large absolute value. Partitioning the study period could amplify this problem. Noting that the covariance vector is the major difference between the two improved estimators, the most likely explanation for the observed pattern of results is that the degree of censoring especially at the tails leads to extremely inflated quantities within this vector and leads to the final estimator being extremely unstable. Given that the problem cannot be located precisely, further investigation is undertaken below using artificially generated data.

This leaves two estimators which may be said on first indication to perform adequately in this particular dataset; the Lin1 estimator and the Bang and Tsiatis partitioned estimator. Not only do they both give estimates of a similar sensible magnitude with accompanying reasonable standard errors, but they also display the anticipated direction of difference in mean cost between the two groups, with the conventional arm having higher average cost than the intensive arm. The similarities between these two estimators are the partitioning of the study duration into subintervals, the use of intermediate cost history for each subject and the use of a probability weight to adjust cost in interval for censoring. The difference lies in the choice of this weight and in the interval cost adjusted by it. In the Lin1 estimator the weight is defined as the probability of surviving to the beginning of each interval and this is used to adjust estimates of mean cost in the interval. The consistency of this estimator, as stated above, requires appropriate censoring conditions, so that censoring times correspond to the interval boundaries of the partition. By contrast, in the Bang and Tsiatis partitioned estimator the weight is defined as the inverse of the probability of an individual not being censored at a given point in time and this is used to adjust individual observed costs in the interval. Moreover, consistency and asymptotic normality of the partitioned estimator does not depend on the choice of the intervals of the partition or the distribution of the censoring times, that is the asymptotic properties of this estimator are independent of the censoring pattern.

Generally the results of the main analysis support - under conditions of extreme censoring - the findings reported by Lin et al (1997) and Bang and Tsiatis (2000). Even under moderate censoring conditions as are considered in these studies, the Lin1 estimator is reported to perform better than Lin2 and is clearly preferred to Lin2 at higher levels of censoring if intermediate cost histories are available as it uses more cost information and requires smaller sample sizes. Bang and Tsiatis show that the partitioned estimator performed better than their other proposed estimators with increasing censoring. The results of the main analysis here, however, indicate a number of potential difficulties which may arise when applying the estimators considered above to data with heavy censoring. Consequently, a number of additional analyses were undertaken to determine whether these difficulties arose because of the characteristics of the specific dataset or the intrinsic properties of the estimators and thus empirically identify conditions under which the estimators perform as expected from the theory.

4.4.4. Secondary analysis and results

The additional analyses presented below investigate further the Lin et al and Bang and Tsiatis estimators concentrating on the specific problems raised above. The estimators are thus assessed under the following circumstances. First, using the same trial data but excluding the highest observed total costs. Secondly, using the same clinical trial data but varying the durations of analysis. Thirdly, using an "artificial" dataset constructed by randomly generating costs and survival times and varying the levels of censoring. Finally, using the bootstrap method to obtain estimates of the standard error for the estimators as an alternative to the analytically derived asymptotic variance estimators.

4.4.4.1. Sensitivity to high cost outliers

As stated previously and shown in Figure 4.2, in both trial arms the distribution of cost was positively skewed with a very small number of observations having extremely high values. To assess whether these high cost outliers influence the estimates, the extreme high cost observations in each arm were excluded from the analysis and the Lin estimates of mean cost based on these data are presented in Table 4.5.

Table 4.5. Lin estimators excluding the highest observed costs from each group

	Conventional		Intensive	
Estimator	*Mean*	*Standard error*	*Mean*	*Standard error*
Lin 1	13583.07	865.32	13058.34	334.46
Lin 2	12078.4	580.88	16821.76	1009.15

The resultant estimates are naturally slightly lower than the respective ones derived in the main analysis, but the differences are not significant and the overall pattern of results is not altered. The conclusions drawn from the main analysis results hold therefore regardless of whether these extreme cost values are included in the analysis or not. The pattern of a positively skewed cost distribution with a small number of high outliers is also observed in the administrative dataset used by Lin et al and is likely to be a characteristic of any medical dataset. The relevant results in Lin et al give no indication that such a characteristic of cost has an impact on the performance of their estimators which is not inconsistent with the finding reported above.

4.4.4.2. Impact of varying the duration of analysis

As mentioned previously, the main analysis results indicated that the Lin et al second approach – not using individual cost histories - gave inconsistent estimates with respect to the direction of the difference in average cost between the two trial arms. Given the reliance of this method on the

number of uncensored individuals in each subinterval and on the number who are censored at the largest observed time, the duration of analysis was restricted to 18, 17, 16, 15 and 12 years and both Lin et al estimators were applied to these data.

Table 4.6. Total number of individuals and number of uncensored cases in each interval of the partition

	Conventional		Intensive	
Interval	*Number entering interval*	*Number dying within interval*	*Number entering interval*	*Number dying within interval*
1	1138	8	2729	24
2	1125	10	2700	22
3	1112	11	2673	28
4	1097	18	2632	23
5	1076	12	2596	37
6	1050	16	2539	51
7	1017	19	2442	35
8	917	19	2233	43
9	816	20	1985	39
10	695	12	1681	37
11	561	22	1347	36
12	433	9	1062	33
13	323	16	818	30
14	214	9	556	18
15	129	7	326	13
16	69	3	187	8
17	53	2	127	8
18	32	0	70	3
19	12	0	18	1
20			2	0

As shown in Table 4.6 the number of uncensored individuals decreases towards the end of the study and is equal to zero for the last two intervals in the conventional group and the last interval in the intensive group, thus falling below the minimum of five deaths in each interval of the partition suggested by Lin et al (1997). In addition in both groups there is only one individual censored at the maximum observed time which implies that the estimate of average cost in the $K+1$ interval (with $\alpha_{K+2} = \infty$) is determined on the basis of this one individual. Restricting the duration of analysis effectively results in increasing the number of individuals who are censored at the upper bound of the analysis time, and more importantly eliminates the impact of the last intervals of the partition on the estimates in which the number of uncensored individuals is very small. Table 4.7 reports the impact of differing durations of analysis on the two Lin et al estimators.

Table 4.7. Lin et al estimators for different durations of analysis

	Conventional		Intensive	
Estimator	*Mean*	*Standard error*	*Mean*	*Standard error*
L=18 years				
Lin 1	13564.66	798.29	12752.36	340.01
Lin 2	15409.63	2930.63	12597.07	1118.68
L=17 years				
Lin 1	12831.22	665.14	12295.81	324.57
Lin 2	14785.19	1661.96	14031.57	842.77
L=16 years				
Lin 1	11884.79	583.97	11434.42	245.7
Lin 2	13683.87	1215.49	12206.42	583.56
L=15 years				
Lin 1	11258.03	512.54	10750.22	210.09
Lin 2	12381.43	987.25	11434.13	480.74
L=12 years				
Lin 1	8869.467	357.95	8642.844	163.97
Lin 2	9230.879	458.12	8858.892	220.52

The initial point to be made here is that the first estimator by Lin et al (Lin1) remains stable for all different durations of analysis. That is, its absolute magnitude decreases as duration decreases since it is estimating average costs over a shorter time period and the rate of decrease appears to be reasonable in both trial arms. More importantly, as in the main analysis results, the conventional group is shown to incur higher costs on average than the intensive group for all time periods of analysis.

With respect to the second estimator by Lin et al (Lin2), the results show that when duration of analysis was restricted to 17 years or less, this estimator became stable resulting in the expected estimates, that is the estimator resulted in conventional policy having a higher mean cost than intensive policy. These results indicate that Lin2 is indeed sensitive to the number of deaths in the intervals of the partition and to the number of individuals censored at the largest observed time. More specifically, increasing these numbers to a "reasonable" level results in obtaining less biased estimates of mean cost in each of the subintervals, as the greater the number of individuals who contribute cost information in each interval the more representative are the estimates of mean cost in the corresponding interval.

With regards to the Bang and Tsiatis estimators, the problems identified in the main analysis were the very low estimates resulting from the simple weighted and the improved simple estimators and the extreme estimates resulting from the improved partitioned estimator. As the authors point out, very heavy censoring in the tails of the distribution may render the estimators unstable with small

sample sizes. As already stated, the trial data were very heavily censored reaching 82% in both conventional and intensive policy groups by the trial end. In addition, as shown in Table 4.6, towards the end of the study the number of individuals still under observation decreases substantially falling to twelve in the conventional group at the last year of follow-up and to two in the intensive group at the last year of follow-up. To assess the impact that an increase in the number of individuals being under observation at the end of the analysis time has on the Bang and Tsiatis estimators, the duration of analysis was restricted to 18, 17, 16, 15 and 12 years and the estimators were computed for the conventional policy group.

Table 4.8. Bang & Tsiatis estimators for different durations of analysis for the conventional policy group

Estimator	*Mean*	*Standard error*
L=18 years (censoring 81.3%)		
Simple	5732.735	840.7795
Partitioned	14639.48	1219.374
Simple improved	3668.924	398.1052
Partitioned improved	334562.5	variance<0
L=17 years (censoring 81.3%)		
Simple	5732.735	840.7795
Partitioned	13410.59	731.8333
Simple improved	3668.924	398.1052
Partitioned improved	mean<0 (mean=-15906.16)	variance<0
L=16 years (censoring 81.5%)		
Simple	5481.435	829.8462
Partitioned	12519.31	683.966
Simple improved	5064.389	384.9361
Partitioned improved	mean<0 (mean=-25535.95)	variance<0
L=15 years (censoring 81.7%)		
Simple	5261.124	826.0686
Partitioned	11832.58	599.1149
Simple improved	5139.35	383.5506
Partitioned improved	13705.7	variance<0
L=12 years (censoring 84.5%)		
Simple	2577.509	388.2635
Partitioned	9113.796	307.1622
Simple improved	4308.985	285.7729
Partitioned improved	8667.038	variance<0

The results are presented in Table 4.8 and show the same pattern as observed in the main analysis. That is, the partitioned estimator gives estimates of average cost very similar to the Lin1 estimator for the various time durations of analysis, the simple weighted and the improved simple estimators still give low estimates compared to the partitioned and Lin1 and the improved partitioned again results in extreme values. In other words, the problems identified in the main analysis do not appear

to be resolved by restricting total analysis time. This is probably due to the fact that although the total number of individuals under observation towards the end of the analysis period slightly increases as the period is restricted, the proportion of patients who are censored remains high (between 81% and 82% until 15 years), even increasing slightly as duration decreases (84.5% at 12 years). This gives a strong indication that the issue of heavy censoring especially in the tails of the distribution is primarily responsible for the estimators' poor performance.

4.4.4.3. Simulation

As the level of censoring is directly related to the performance of all estimators considered and more specifically as Bang and Tsiatis state very heavy censoring in the tails of the distribution could result in their estimators becoming unstable, the performance of the various estimators was assessed for different levels of censoring. Both Lin et al and Bang and Tsiatis construct artificial datasets and vary the levels of censoring by up to 45%. In a similar manner a simulated dataset was constructed here to explicitly test the impact that the degree of censoring has on the estimators of interest while at the same time ensuring that individual costs vary in a predefined manner. As well as having the advantage that different levels of censoring can be set and the impact of censoring can be isolated, an artificial dataset also allows estimation of the "true" mean cost, that is the mean cost if censoring was not present in the data. A direct assessment of the performance of the various estimators is thus achieved through comparison of the estimated means to the "true" mean.

The approach adopted in the construction of this dataset is similar to the one described in the Lin et al and Bang and Tsiatis simulation experiment, although the estimates obtained here are not based on replications of the data but have resulted from a single application of each approach to the artificial dataset once this was generated as follows. A sample size of 1138 individuals was chosen for this "artificial" dataset to equal the smaller sample size of the clinical trial data used in the main analysis - since one of the concerns for the validity of the methods is related to the sample size. Survival times were generated from a uniform distribution on [0, 10] years. The average 10-year cost is the parameter of interest, with the total cost for individual *i* being

$$M_i = M_i(0) + b_i T_i^L + \sum_{j=1}^{10} \tau_{ij} (\min[\{T_i^L - (j-1)\}^+, 1]) + d_i I(T_i \leq 10)$$

where $M_i(0)$ is the initial diagnostic cost, b_i is the deterministic annual cost, τ_{ij} is the random annual cost for the jth year, d_i is the terminal death cost and $\alpha^+ = \max(0, \alpha)$. For the distribution of each cost element, $M_i(0)$, b_i, τ_{ij}, d_i are assumed uniformly distributed on [5000, 15000], [1000, 2600], [0, 400] and [10000, 30000] respectively. Various levels of censoring were considered with the censoring times being uniformly distributed on [0, 20] years, i.e. 25% censoring, [0, 12.5] years, i.e. 41% censoring, [0, 10] years, i.e. 51% censoring, [0, 9.5] years, i.e. 55% censoring and [0, 9] years, i.e. 57.5% censoring. Bang and Tsiatis impose similar levels of

censoring to Lin et al who refer to "light" censoring as censoring set at 25 to 30% and "heavy" censoring as censoring set at 40 to 45%. Programming was again undertaken in Stata 7.0. as shown in Appendix A.4.7. The components were generated independently and the simple weighted estimator, the partitioned estimator, the improved simple, the improved partitioned and the Lin1 estimator (which uses individual cost history) were calculated for all levels of censoring. The Bang and Tsiatis partitioned and improved partitioned estimators were based on yearly subintervals and for both simple improved and partitioned improved estimators, annual subintervals were assumed in the recovery of cost information lost due to censoring, and the set of prespecified e-functionals were defined as in the analysis of the real trial data. The Lin1 estimator based on annual subintervals was also estimated using this artificial dataset as under all circumstances considered in the real data analyses it remained stable and generally performed well.

Table 4.9. Estimates based on the "artificial dataset"

Estimator	*Mean*	*Standard error*
Censoring 25%		
Simple weighted	41348.2	475
Simple improved	40452.9	433.8
Partitioned	41654.9	342.1
Improved partitioned	40876	316.2
Lin 1	39545.6	311
Censoring 41%		
Simple weighted	37228.3	1713.2
Simple improved	34724.4	854.7
Partitioned	40000.2	734.4
Improved partitioned	38575.4	366.3
Lin 1	37367.4	355.8
Censoring 51%		
Simple weighted	29284.3	3340.4
Simple improved	25334.2	1627.8
Partitioned	37242.8	1306.4
Improved partitioned	34683.7	514
Lin 1	35456.3	354
Censoring 55%		
Simple weighted	21037.6	684.6
Simple improved	15922.8	536.2
Partitioned	33839.1	315.2
Improved partitioned	32446.4	variance<0
Lin 1	34280	296.1
Censoring 57.5%		
Simple weighted	18921.7	605
Simple improved	13361.8	549.4
Partitioned	33048	271.6
Improved partitioned	32787.7	variance<0
Lin 1	33686	271.9

The true mean cost (no censoring) is 41144.5.

The resultant estimates of the average 10-year cost and its asymptotic variance based on the artificial data are reported in Table 4.9. The "true" average cost obtained when complete information was assumed on all individuals in these data was equal to 41144.50 and serves as the reference cost to be compared with all other estimates under different levels of censoring.
As expected, as the level of censoring increases all estimators generally exhibit higher degrees of bias. The Lin1 and the Bang and Tsiatis partitioned estimators performed well at all levels of censoring. The Bang and Tsiatis improved partitioned estimator performed equally well up to a level of censoring of 51%. It resulted in negative estimates for the variance when censoring reached 55%. The simple weighted and simple improved estimators appear to give increasingly lower estimates as censoring increases, with the estimates being close to the others only up to 41% censoring. At 55% censoring the average cost derived by the simple weighted estimator was approximately half the "true" mean cost value and the one derived by the simple improved estimator was even lower. Overall, the findings from this analysis support the findings of the analysis based on the real clinical data. The Lin1 and Bang and Tsiatis partitioned estimators appear stable at all levels of censoring whereas the simple weighted and both improved estimators appear extremely sensitive to the level of censoring reflecting a similar pattern, only less extreme, to the one observed in the trial data.

4.4.4.4. Bootstrap estimates of the variances

The derivation of the standard errors for all the estimators proposed by Lin et al and Bang and Tsiatis is based on the large sample properties of these estimators. Study of their asymptotic properties as presented previously has shown that the estimators converge to a normal distribution and use of the martingale version of the central limit theorem allows estimators for their variances to be formulated. While efficiency is therefore shown to hold conceptually a potential problem relates to the validity of the underlying assumption of asymptotic normality when the approaches are applied to any particular dataset. Although asymptotic statistics is of both theoretical and practical importance, it is a theory of approximations. Such approximations are particularly useful as shown in the preceding analysis in studying theoretically the efficiency of the statistics of interest but are of questionable value if the statistical procedure which has been shown to function for $n \to \infty$ is to be applied to a finite sample. In most situations the theory itself does not provide a means for assessing the magnitude of the approximation errors and it is usually the case that the accuracy of the asymptotic results is judged by simulation studies.

To test the validity of the estimators' asymptotic results, empirical standard errors for both Lin et al estimators and for the Bang and Tsiatis simple weighted and partitioned estimators were derived using the bootstrap method. The bootstrap estimates were obtained by drawing random samples of size n=1138 from the observed distribution for the conventional group and n=2729 for the intensive group and calculating the Lin 1, Lin 2 and the Bang & Tsiatis simple weighted and simple

partitioned estimates of average cost across a large number of replications. All sets of bootstrap estimates were obtained for 200 and 1000 bootstrap replications which are deemed adequate for the calculation of standard errors and the relevant Stata programs are presented in Appendix A.4.8. The standard errors derived from the bootstrap method are reported in Table 4.10.

Table 4.10. Bootstrap estimates of the standard error

Lin et al: When the time of analysis was the complete follow-up period (L=18.9 years for conventional and L=19.5 years for intensive)

	Conventional	**Intensive**
Replications 200		
Lin 1	823.4994	333.5156
Lin 2	8085.001	3784.319
Replications 1000		
Lin 1	927.3038	343.5167
Lin 2	7392.851	3837.986

Lin et al: When the time of analysis was 17 years for both conventional and intensive

	Conventional	**Intensive**
Replications 200		
Lin 1	628.704	307.9546
Lin 2	1541.102	724.1773
Replications 1000		
Lin 1	670.0437	322.9126
Lin 2	1789.984	763.5229

Bang & Tsiatis: when the time of analysis was the complete follow-up period (L=18.9 years for conventional and L=19.5 years for intensive)

	Conventional	**Intensive**
Replications 200		
Simple weighted	786.1384	3098.452
Partitioned	1207.907	439.8706
Replications 1000		
Simple weighted	830.2415	3132.548
Partitioned	1379.021	451.2829

With respect to the Lin et al estimators the bootstrap estimates were also derived for a duration of analysis of 17 years as this was the point at which the Lin2 estimator became stable. Comparison of the empirically derived variance estimates using the bootstrap method with their respective asymptotic variance estimates reported in Table 4.4 and Table 4.7 shows that for all estimators the bootstrap estimates of the standard error confirm those obtained from the formulae (this being the case for Lin2 under the conditions where this became stable as expected). This finding therefore supports the validity of the assumptions underlying the estimators' asymptotic properties. Conversely, the bootstrap method gives reasonable approximation to the theoretically derived variances.

4.5. Discussion

This chapter has concentrated on non-parametric estimators of cost statistics under conditions of right censoring. As such estimators are free of assumptions regarding the distribution of cost and can easily incorporate the presence of censoring in the cost observations they can be particularly appealing. The Kaplan-Meier estimator has been proven inappropriate in the analysis of cost-to-event data due to the violation of independence between the random variable of interest and its censoring variable. Consequently a number of alternative non-parametric estimators have been proposed recently that require independence between time-to-event and time-to-censoring but not independence between cost-at-event and cost-at-censoring. Although these estimators are free of assumptions with respect to the distribution of cost, they are not entirely free of restrictions. More specifically, consistency of the estimators proposed by Lin et al depends on the pattern of the distribution of censoring times and although the asymptotic properties of the estimators proposed by Bang and Tsiatis are independent of the censoring pattern, the estimators can become unstable under conditions of heavy censoring at the tails of the distribution. In theory, provided that their respective assumptions are valid, each of the Lin et al and Bang and Tsiatis estimators considered in this chapter will provide consistent estimators of average cost. From the theory it is also expected that the degree of censoring will have a direct impact on the estimators' performance with this deteriorating as censoring increases although this impact will vary among the approaches. While the estimators' desirable properties, that is consistency and efficiency, have been shown to hold conceptually the degree to which these properties are retained in practice will depend on the particular application. That is, establishing that an estimator is asymptotically efficient or asymptotically more efficient than a competing estimator does not ensure that this property holds for finite samples. This is the reason why simulation studies are commonly undertaken as a means of assessing the accuracy of the asymptotic results in a practical setting before use of the estimator is recommended within the specific analytic context.

Within the context of the analysis presented here the first estimator proposed by Lin et al which uses information on intermediate individual cost histories appeared stable under a wide variety of conditions as opposed to their second estimator which only uses information on total costs from individuals who are either observed for the full time to event or are censored at the upper bound of analysis time and was shown to be sensitive to the number of individuals contributing cost information. With respect to the set of estimators proposed by Bang and Tsiatis, the simple weighted estimator using only complete cost information and both improved estimators appeared extremely sensitive to the level of censoring and became increasingly unstable as censoring increased. In contrast, their partitioned estimator which uses information on intermediate individual cost histories performed well under all circumstances. Concentrating on the two most stable estimators, these are similar in that they both partition the study period into subintervals and make use of individual intermediate cost history within each subinterval and in that they both use a weight to adjust interval costs for censoring. They are different both in the choice of this weight and in the

interval costs that are adjusted by it. In Lin1 the weight is the Kaplan-Meier probability of survival to the start of the interval that adjusts estimates of mean cost in the interval, whereas the Bang and Tsiatis partitioned estimator uses the inverse of the probability of an individual not being censored evaluated at a given point in time to adjust individual observed costs in the interval. On the basis that both approaches require the same amount of cost information, but the second approach is not restricted by the pattern of the censoring distribution and is therefore more general, it might be preferred.

There is a long history related to the use of the inverse of the probability of inclusion in adjusting estimates for missingness. The same inverse probability weight was first used by Horvitz and Thompson (1952) in the context of sample surveys, by Koul et al (1981) in studying censored failure times using a linear regression methodology, by Robins and Rotnitzky (1992) in the context of recovering information missing due to censoring, by Lin and Ying (1993) in non-parametric estimation of the bivarate survival function under univariate censoring, by Robins, Rotnitzky and Zhao (1994) in adjusting estimates of regression coefficients for missingness in the data, by Rotnitzky and Robins (1995) in studying semiparametric regression models in the presence of censoring dependent on covariates, by Robins, Rotnitzky and Zhao (1995) in studying semiparametric regression models for repeated outcomes in the presence of censoring dependent on covariates, by Zhao and Tsiatis (1997) in deriving a consistent estimator for the distribution of quality adjusted survival time under conditions of censoring, and recently by Lin (2000) in adjusting cost estimates for censoring using a linear regression approach as shown in the next chapter. In all these applications use of this weight results in consistent estimators for the statistics of interest while adjusting for missingness. The same general finding emerges from the analysis undertaken in this chapter but at the same time the performance of the corresponding estimators appears to be subject to the amount of cost history information entering the estimating equations. This is why the simple weighted and the partitioned estimators yield such different estimates of mean cost. That is, although the same general definition of the probability weight underlies both estimators, the points in time at which the individual probabilities are evaluated differ between the approaches in a manner that is determined by the points at which information on individual cost histories becomes available. The implication is that the weight alone is not sufficient to adjust the estimates for the loss of information when the level of missingness is too high.

Nevertheless despite the limitations associated with the assumptions underlying the estimators' validity and their dependence on the data under consideration, the present analysis has identified estimators whose performance is deemed satisfactory under extreme censoring conditions. Consequently, their application to the analysis of censored cost data is appropriate when estimates of mean cost over the study period are sought. When interest extends however beyond the maximum time for which data is available or when questions regarding the effect of covariates on cost arise, parametric models become a necessary alternative. It is clearly important that such parametric models make adjustment for censoring. Provided that censoring is appropriately

accounted for and that the distributional assumptions imposed by a specific parametric approach are justified, the within study estimates derived by such a model could be compared to non-parametric estimates as a means of assessing the validity of the parametric approach before this was used to extrapolate beyond the end of the study period or to a different population setting.

Chapter 5

Cost analysis: Parametric estimators of treatment cost under conditions of censoring

5.1. Introduction

The primary advantage of non-parametric models is that they are free of assumptions concerning the distribution of cost. There are circumstances however where parametric methods may be the preferred or necessary alternative. Some investigators including Mullahy and Manning (1996) as noted in Chapter 2, even suggest that parametric modelling is generally preferable given the inherent characteristics embodied in trial data. More specifically, while clinical trials attempt to standardise for population characteristics through randomisation, there may remain systematic differences in the treatment costs across subgroups of the population defined by different covariate values. Information on the pattern of cost accumulation may then be gained by assessing covariate effects on cost using a parametric approach. Furthermore, although focus in this thesis is on within trial estimates of average cost, it may also be desirable to derive cost estimates associated with benefits continuing beyond the end of the study period. Parametric models can provide an instrument for extrapolating estimates of costs over the study period to points in time exceeding the duration of the study. Censoring is again the main concern in the analysis presented in this chapter. Consequently the parametric approaches to be considered here all attempt to derive estimates of cost accounting for the presence of censoring. An additional concern common to all these approaches relates to the specific functional form the parametric model assumes between cost and the explanatory variables especially given that cost distributions are generally complex and therefore difficult to parameterise. It is natural to expect the difficulty of appropriately specifying this relationship to be increased due to the information loss induced by censoring.

As in the case of non-parametric models the earliest attempts to account for censoring in deriving estimates of mean cost using a parametric approach involved direct application of the classical survival techniques to censored cost data. The Cox proportional hazards model and the Weibull and exponential models were applied for example by Dudley et al (1993) and Fenn et al (1996) in estimating within study average cost. However, these approaches are biased for the same reason as the Kaplan-Meier estimator, that is due to dependent censoring between cost at event and cost at censoring times. The classical linear regression is also always biased when the outcome variable is subject to censoring as shown analytically below. A naïve alternative would be to estimate the classical linear regression model using the complete cases only but this is always going to be biased as it discards the censored observations completely with the degree of bias increasing as censoring increases. Failure of these approaches to account for censoring in the cost estimates led to two

recently proposed alternatives. The first adopts a regression approach where cost is modelled as a function of failure time and adjustment for censoring is achieved in the cost estimates through adjusting failure time for censoring. The second uses a linear regression methodology in which adjustment for censoring in the cost estimates is performed through use of the inverse of the probability of an individual not being censored in the estimating equations. All these estimators of cost together with their properties and underlying assumptions are considered below. When necessary, counting process and martingale theory again provides the analytical framework in which the statistical properties of the estimators are studied. When the estimators considered are based on a methodology originally used to analyse time to event data, the approach is first considered within this context and the extension to cost to event data follows. Although typically in regression problems the important inference questions are about the conditional distribution of the outcome variable given the covariates, the aim of the present analysis is to assess the estimators' relative performance with respect to the resultant mean cost estimates over the study period under extreme censoring conditions using the UKPDS data. Given that the analysis in the previous chapter identified estimators whose performance is deemed adequate under these conditions, assessment of the estimators considered below is undertaken by comparison to the most adequately performing non-parametric estimators considered earlier.

The chapter proceeds as follows. The general setting for the analysis is first outlined and the set of parametric estimators for cost together with the assumptions underlying their validity are then presented. The proposed semiparametric regression methodology is considered first and includes the Cox proportional hazards regression and a proportional means regression model in which the mean cumulative cost forms the outcome variable. The fully parametric Weibull and exponential regression models are presented next and these are followed by alternative least squares regression approaches starting with the classical linear regression model. Extensions to the naïve ordinary least squares approach where adjustment for censoring enters the estimating equations are then investigated. This methodology allows the analysis to be undertaken both when the cost data are available at the individual's death or last contact date and when these data are available at multiple points in time over the study duration. An alternative regression approach models cost as a function of time and attempts to account for censoring in cost through accounting for censoring in failure time. The resultant cost estimates from application of the alternative regression methodologies to the UKPDS data follow.

5.2. Parametric estimators of cost under censoring

5.2.1. General setting

As in the previous chapter the basic aim of the approaches presented below is to derive an estimate of the mean total cost $\mu = E(M)$ and its variance over a specified period when the data is right

censored, where the random variable M denotes the total cost for a patient during some specified time T and E denotes expectation. Again the distribution of the random variable T is assumed continuous over $(0, L]$ where L denotes the upper bound of T and M is the total cost incurred by a patient up to a maximum of L units of time. The main difference between the approaches considered in the previous chapter and the ones considered below is that the latter attempt to derive mean cost estimates using a parametric model which relates cost to a set of covariates and as such they make specific assumptions about the distribution of cost. To accommodate censoring, a potential time to censoring denoted by U is defined and letting T denote time to death, the observables from a study in the presence of censoring are $X = \min(T,U)$, i.e. the last contact date; $\delta = I(T \leq U)$, where $I(\cdot)$ is the indicator function taking the value of 1 when the argument is true (i.e. if the observation is uncensored) and zero otherwise; the cost accrued up to time X and other intermediate cost history for each subject, i.e. $M^H(t) = \{M(u), u \leq t\}$, where $M^H(t)$ denotes the cost history up to time t, $M = M(T)$, with $M(u)$ being the known accumulated cost up to time u and u denoting points in time at which cost information becomes available. Letting $Z = (Z_1,...,Z_p)'$ denote a $p \times 1$ vector of the covariates of interest, the observable data for n individuals are then the independent and identically distributed random vectors

$$\{X_i = \min(T_i,U_i),\ \ \delta_i = I(T_i \leq U_i),\ \ M_i^H(X_i), Z_i\}, i = 1,...,n$$

where i identifies an individual.

5.2.2. Cox proportional hazards regression

The Cox proportional hazards model falls into the category of semiparametric models. Due to its semiparametric nature, it allows the functional form of part of the model to be unknown and therefore unrestricted. The proportional hazards model assumes that the hazard of an individual having an event is a function of a set of individual covariates and an underlying arbitrary baseline hazard. Given that part of the model is completely unspecified, as stated previously in section 4.3.3.3, estimation of the parameters of interest requires that some assumptions be made or restrictions be imposed on the statistical relationship between what is observed and what is not observed. The assumption imposed by this model is that the hazard functions for any two individuals are proportional with a ratio determined by the covariates that is constant over time. Clearly if one is unsure as to the functional form of the hazard function, adopting a semiparametric approach could be a preferred alternative to imposing specific parametric assumptions on the distribution of the hazard function. The usefulness of the particular semiparametric specification is due to a number of reasons such as the easily understood interpretation of the idea that the effect of a given covariate, for instance a treatment, is to multiply the hazard by a constant factor; the empirical evidence in certain areas that supports the assumption of proportionality of hazards in distinct treatment groups; the fact that censoring and the occurrence of several types of failure are

relatively easily accommodated within this model specification and the technical problems of statistical inference associated with the unknown part of the model, that is the arbitrary baseline hazard, have a simple solution (Cox and Oakes, 1984).

The observed data in regression problems when the time to failure is subject to right censoring are the independent observations of the quantities (X, δ, Z) as defined above. The same counting processes can be used to model such data as defined in the previous chapter, that is,

$N(t) = I(X \le t, \delta = 1)$ with $N(t) = \sum_{i=1}^{n} N_i(t)$ where $N_i(t) = I(X_i \le t, \delta_i = 1)$,

$N^c(t) = I(X \le t, \delta = 0)$ with $N^c(t) = \sum_{i=1}^{n} N_i^c(t)$ where $N_i^c(t) = I(X_i \le t, \delta_i = 0)$, and

$Y(t) = I(X \ge t)$ with $Y(t) = \sum_{i=1}^{n} Y_i(t)$ where $Y_i(t) = I(X_i \ge t)$.

The filtration $\{\mathcal{F}_t : t \ge 0\}$ generated by these processes is given by

$$\mathcal{F}_t = \sigma\{Z, N(u), N^c(u) : 0 \le u \le t, i = 1,...,n\}$$

and provides information on the individuals' covariates and failure or censoring status up to and including time *t*. However interest now lies in the conditional distribution of failure time given the set of covariates. The conditional survival function is then $S(t|Z) = pr(T > t|Z)$ and the conditional hazard function is given by

$$\lambda(t|Z) = \lim_{\Delta t \to 0} \frac{pr(t \le T < t + \Delta t | T \ge t, Z)}{\Delta t}$$

The proportional hazards regression proposed by Cox (1972) studies the relationship between the set of covariates *Z* and the distribution of censored failure times using a model in which the hazard function is

$$\lambda(t|Z) = \lambda_0(t) e^{\beta' Z} \qquad (5.1)$$

where $\beta = (\beta_1,..., \beta_p)'$ is a $p \times 1$ vector of unknown regression coefficients and $\lambda_0(t)$ is an unknown arbitrary nonnegative function of time giving the hazard function when Z=0. As such the model assumes a parametric form only for the covariate effect while the baseline hazard is treated nonparametrically. The term proportional hazards refers to the fact that the hazard functions for different individuals defined according to (5.1) are multiplicatively related with a ratio that is constant over time. For small values of Δt, the conditional hazard rate satisfies

$$\lambda(t|Z)\Delta t \approx pr(t \le T, t+\Delta t|T \ge t, Z)$$

and can thus be interpreted as the conditional probability of a failure occurring in the interval $[t, t+\Delta t)$, given Z and no failure before t. Censoring plays a similar role in the proportional hazards model as in the case of the non-parametric hazard and the condition of independence between T and U now required in the presence of covariates is expressed as

$$pr(t \le T < t+\Delta t|T \ge t, U \ge t, Z) = pr(t \le T < t+\Delta t|T \ge t, Z)$$

Allowing the covariates to be time-dependent the hazard function for individual i is

$$\lambda_i(t|Z_i) = \lambda_0(t)e^{\beta' Z_i(t)}$$

Assuming a continuous distribution for failure time, a censoring mechanism that does not depend on β and independence between failure and censoring time, inferences about the regression parameters β are based on the partial likelihood introduced by Cox (1972, 1975) as

$$L(\beta) = \prod_{i=1}^{n}\prod_{t\ge 0}\left\{\frac{Y_i(t)e^{\beta' Z_i(t)}}{\sum_j Y_j(t)e^{\beta' Z_j(t)}}\right\}^{dN_i(t)}$$

where the processes $N(\cdot)$ and $Y(\cdot)$ have been defined above. The term partial likelihood was used because the likelihood expression is not dependent on the unknown baseline hazard $\lambda_0(t)$ but only on the parameters β. The product is over all uncensored failure times and each term represents the conditional probability that individual i fails at time t given that one individual among those at risk at time t fails at time t. Cox argued that the resulting parameter estimates from the partial likelihood function would have the same distributional properties as the ones derived from full maximum likelihood estimators. Thus he suggested treating this likelihood function as an ordinary likelihood function for the purpose of large sample inference about β. The log partial likelihood evaluated at time t is given as

$$\sum_{i=1}^{n}\int_0^t\left\{\beta' Z_i(s) - \log\left(\sum_j Y_j(s)e^{\beta' Z_i(s)}\right)\right\}dN_i(s)$$

and differentiating this expression with respect to β results in

$$U(\beta, t) = \sum_{i=1}^{n}\int_0^t\left\{Z_i(s) - \frac{\sum_i Y_i(s)Z_i(s)e^{\beta' Z_i(s)}}{\sum_i Y_i(s)e^{\beta' Z_i(s)}}\right\}dN_i(s) \qquad (5.2)$$

The solution to the partial likelihood equation $U(\hat{\beta},\cdot)=0$ yields the maximum partial likelihood estimator.[1] Estimators for the hazard function, as this is not estimated by solving the likelihood equations, have been suggested by Cox (1972), Breslow (1972, 1974, 1975) Oakes (1972) and others. The estimator $\hat{\Lambda}_0$ for $\Lambda_0(t)=\int_0^t \lambda_0(u)du$ proposed by Breslow (1974) is most commonly used and is given by

$$\hat{\Lambda}_0(t)=\int_0^t \frac{\sum_{i=1}^{n} dN_i(u)}{\sum_{i=1}^{n} Y_i(u)e^{\hat{\beta}'Z_i(u)}} \tag{5.3}$$

which reduces to the Nelson-Aalen estimator given in chapter 4 when $\hat{\beta}=0$. The survival function conditional upon the covariates is given as

$$S(t|Z)=e^{-\int_0^t \lambda(u|Z)du}=e^{-\int_0^t \lambda_0(u)e^{\beta'Z}du}=S_0(t)^{\exp(\beta'Z)}$$

and can be estimated by

$$\hat{S}(t|Z)=e^{-\hat{\Lambda}_0(t)\exp(\hat{\beta}'Z)}$$

where $\hat{\Lambda}_0$ is the estimator for the integrated hazard. Under circumstances in which there are covariates whose effect on the hazard is not proportional, a stratified variant of the proportional hazards model is adopted. This extension leads to stratum specific hazard functions and for stratum *s* the proportional hazard function is given as

$$\lambda_s(t|Z)=\lambda_{0s}(t)\exp(\beta'Z) \tag{5.4}$$

Under this specification, the subjects in the *s*th stratum have an arbitrary baseline hazard function $\lambda_{0s}(t)$ and the effect of other explanatory variables on the hazard is represented by a proportional hazards model in that stratum as given by (5.4). Hazard ratios are then computed within each stratum but the regression coefficients are assumed to be the same across all strata although the

---

[1] The score statistic given by (5.2) can also be written as

$$U(\beta,t)=\sum_{i=1}^{n}\int_0^t\left\{Z_i(s)-\frac{\sum_i Y_i(s)e^{\beta'Z_i(s)}Z_i(s)}{\sum_i Y_i(s)e^{\beta'Z_i(s)}}\right\}d\mathscr{M}_i(s)$$

where $\mathscr{M}_i(t)=N_i(t)-\int_0^t Y_i(u)e^{\beta'Z_i(u)}\lambda_0(u)du$ is the associated martingale process.

baseline hazard functions may be different and completely unrelated. Each stratum contributes a stratum specific partial likelihood and the full stratified partial likelihood is obtained by multiplying the contributions to the likelihood with the overall log likelihood being given by

$$l(\beta) = \sum_{s=1}^{S} l_s(\beta)$$

where $l_s(\beta)$ is the log partial likelihood in stratum s $(s = 1,...,S)$. Stratum specific survival functions can then be estimated using the same methods as for the non-stratified model. Hence the Breslow estimate for $\Lambda_{0s}(t) = \int_0^t \lambda_{0s}(u)du$ is

$$\hat{\Lambda}_{0s}(t) = \int_0^t \frac{\sum_{i=1}^{n_s} dN_{si}(u)}{\sum_{i=1}^{n_s} Y_{si}(u) e^{\hat{\beta}' Z_{si}(u)}} \qquad (5.5)$$

Application to cost

Adoption of the proportional hazards model in assessing covariate effects on cost could be appealing on the basis that the model is free of distributional assumptions concerning the hazard rate for cost. It could be useful therefore in modelling censored cost data which typically have complex distributions. In this setting the proportional hazards model relates the hazard of attaining a particular cost level to a set of covariates under the following specification

$$\lambda(c|Z) = \lambda_0(c) e^{\beta' Z} \qquad (5.6)$$

where $\lambda(c|Z)$ is the hazard of attaining a given cost level c conditional upon the set of covariates Z, $\beta = (\beta_1,...,\beta_p)'$ is a $p \times 1$ vector of unknown regression coefficients and $\lambda_0(c)$ is an unknown arbitrary nonnegative function of cost giving the hazard function when $Z = 0$. Although there are no assumptions about how the hazard rates vary with time underlying this model, the hazards rates for different levels of covariates must be proportional with a constant ratio over cost levels. An estimate of mean cost over the study period can then be derived as

$$\hat{M} = \int_0^C \{\hat{S}_0(c)\}^{\exp(\hat{\beta}'\tilde{Z})} dc$$

where $S(c) = pr(M \geq c|Z) = \{S_0(c)\}^{\exp(\beta' Z)}$ is the probability that the cost will be at least c given the covariates, $\tilde{Z}$ denotes the covariates vector evaluated at the mean values of the covariates and

an estimate $\hat{S}_0(c) = e^{-\hat{\Lambda}_0(c)}$ for $S_0(c)$ could be derived by applying the Breslow estimator given by (5.3) to cost. However due to the positive correlation between cost at failure and cost at censoring time the resultant estimates are biased unless individuals accumulate costs at a common rate over time. A solution to the problem of dependent censoring suggested by Lipscomb et al (1998) is to apply the stratified version of the proportional hazards model with time as the stratification variable. The hazard function for cost in time period *t* is then given as

$$\lambda_t(c|Z) = \lambda_{0t}(c)e^{\beta'Z} \tag{5.7}$$

where $\lambda_{0t}(c)$ is an unknown nonnegative baseline hazard function for cost in the *t*th stratum, that is in time period *t*. The cumulative cost function gives the probability that the cost for the time period *t* will be at least *c* given the covariates and is

$$S_t(c) = pr(M_t \geq c|t, Z_t) = \{S_{0t}(c)\}^{\exp(\beta'Z_t)}$$

where $S_{0t}(c) = \exp\left[-\int_0^c \lambda_{0t}(k)dk\right]$ is the baseline cumulative cost function defined for each time period *t* giving the level of cost when all explanatory variables are set to zero with *k* and *c* denoting levels of cost. An estimate for the mean cost in time period *t* is then given as

$$\hat{M}_t = \int_0^\infty \{\hat{S}_{0t}(c)\}^{\exp(\hat{\beta}'\bar{Z}_t)} dc$$

where an estimate $\hat{S}_{0t}(c) = e^{-\hat{\Lambda}_{0t}(c)}$ for $S_{0t}(c)$ could be derived by applying the Breslow estimator given by (5.5) to cost. Lipscomb et al (1998) recommend use of this model under censoring conditions on the basis that stratification by time circumvents the problem of dependent censoring between cost at failure and cost at censoring, as this specification imposes no constraint as to how cost varies over time within a given time period which implies that the one-to-one mapping between time to failure and cost is no longer an issue.

Etzioni et al (1999) however criticise the use of the Cox proportional hazards model in analysing censored cost data and argue that its use within this context will generally lead to biased estimates on the following basis. For Cox regression to be unbiased, independent censoring is required within each group formed by each level of covariate ensuring that individuals who are still under observation are representative of the population at risk in each group. When Cox regression is applied to cost analysis the accrual of costs at different rates will result in dependent censoring within the subgroups defined by the covariate levels. Covariates that affect the rate of cost accrual will lead to differential dependent censoring across groups defined by different covariate values. As a result estimates of cost statistics will be biased. An additional concern relates to the

proportionality assumption underlying the validity of the Cox regression model. Although the model does not specify the underlying cost hazard, it assumes proportionality of the cost hazards defined by different levels of covariates. The same assumption underlies the stratified variant for each stratum. Etzioni et al (1999) show that the proportionality assumption will not generally hold in circumstances when individuals accumulate costs at different rates as follows.

Assuming a binary covariate Z taking the values 0 or 1 and assuming that given $Z = z$ survival is exponential with mean $1/\lambda_z$ and costs accumulate at a rate of a_z per unit of time (or over a fixed time period) with probability p_z and at a rate b_z with probability $1 - p_z$, the cost at event $c(t)$ for an individual with $Z = z$ would be $c(t) = a_z t$ with probability p_z and $c(t) = b_z t$ with probability $1 - p_z$. Under this model the probability density function for failure time t given $Z = z$ is

$f_z(t) = \lambda_z \exp\{-\lambda_z t\}$ with survivor function $S_z(t) = \exp\{-\lambda_z t\}$

and the probability density function for cost given $Z = z$ is

$$f_z(c(t)) = \frac{\lambda_z}{a_z}\exp\left\{-\frac{\lambda_z}{a_z}t\right\}p_z + \frac{\lambda_z}{b_z}\exp\left\{-\frac{\lambda_z}{b_z}t\right\}(1-p_z)$$

with survivor function

$$S_z(c(t)) = \exp\left\{-\frac{\lambda_z}{a_z}t\right\}p_z + \exp\left\{-\frac{\lambda_z}{b_z}t\right\}(1-p_z)$$

The hazard function for cost given $Z = z$ is

$$h_z(c(t)) = \frac{f_z(c(t))}{S_z(c(t))}$$

and the hazard ratio for cost for the different values of the covariate Z is then

$$\frac{h_{Z=1}(c(t))}{h_{Z=0}(c(t))} = \frac{\frac{\lambda_1}{a_1}\exp\left\{-\frac{\lambda_1}{a_1}t\right\}p_1 + \frac{\lambda_1}{b_1}\exp\left\{-\frac{\lambda_1}{b_1}t\right\}(1-p_1)}{\exp\left\{-\frac{\lambda_1}{a_1}t\right\}p_1 + \exp\left\{-\frac{\lambda_1}{b_1}t\right\}(1-p_1)} \Bigg/ \frac{\frac{\lambda_0}{a_0}\exp\left\{-\frac{\lambda_0}{a_0}t\right\}p_0 + \frac{\lambda_0}{b_0}\exp\left\{\frac{\lambda_0}{b_0}t\right\}(1-p_0)}{\exp\left\{-\frac{\lambda_0}{a_0}t\right\}p_0 + \exp\left\{\frac{\lambda_0}{b_0}t\right\}(1-p_0)}$$

It becomes clear from the above expression that the assumption of proportional hazards for different values of the covariates depends on the values of a_z, b_z, λ_z and p_z which as the authors argue are not generally expected to have the values required to ensure the validity of the proportionality assumption. The authors conclude therefore that even if the bias imparted by dependent censoring within the subgroups defined by the covariate levels due to the differential rate of cost accrual is not severe or is eliminated, use of the model will not be justified as the assumption of proportional cost hazards across different covariate levels is unlikely to be valid when the rate of cost accumulation varies among individuals. Essentially whether the assumption of proportionality is valid or not is an empirical question and depends on the particular application. This assumption can be tested as presented below both for the non-stratified model and for its stratified variant.

Assessing the proportionality assumption in the proportional hazards model

There are a number of graphical approaches for assessing the proportional hazards assumption (see Fleming and Harrington, 1991). An approach that provides a test statistic is due to Grambsch and Therneau (1994). Their approach is a generalisation of the approach by Shoenfeld (1982) who considered departures from proportionality with respect to one covariate only. In the context of time to failure data analysis, under the proportional hazards model, the intensity process for individual *i* is given as

$$Y_i(t)e^{\beta' Z_i(t)}d\Lambda_0(t)$$

In general, the assumption of proportionality with respect to covariate *j* means that

$$\beta_j(t) = \beta \text{ for all } t$$

which in turn implies that a plot of $\beta_j(t)$ against time will have a zero slope. Under the alternative assumption of time-varying coefficients, the intensity process for individual *i* is

$$Y_i(t)e^{\beta'(t)Z_i(t)}d\Lambda_0(t)$$

Grambsch and Therneau (1994) have shown that

$$E(r_k^*) + \hat{\beta} \approx \beta(t_k)$$

where $\hat{\beta}$ is the estimated coefficient vector from the proportional hazards model and $r_k^* = [\hat{\text{var}}(r_k)]^{-1} r_k$ is referred to as the scaled Schoenfeld residual derived by scaling the Schoenfeld

residual by an estimator of its variance.[2] Expressing $\beta(t)$ as a regression on some function of time $g(t)$ as

$$\beta_j(t) = \beta_j + \theta_j g_j(t) \tag{5.8}$$

where j indexes covariates $j = 1,..., p$, Grambsch and Therneau (1994) propose testing the null hypothesis of proportional hazards, that is, $\theta_j = 0$, against the alternative of coefficients which vary over time through testing the null hypothesis of zero slope in the generalised linear regression of the scaled Schoenfeld residuals on functions of time. The test of zero slope is equivalent to testing that the log hazard function is constant over time and rejection of the null hypothesis would imply deviation from proportionality.

In applying the approach to testing the assumption of proportionality in the cost hazards, the analogous regression for covariate j is given as

$$\beta_j(c) = \beta_j + \theta_j g_j(c) \tag{5.9}$$

where $g_j(c)$ is some function of cost. Testing the null hypothesis of proportional hazards $\theta_j = 0$, is now based on testing the null hypothesis of zero slope in the generalised linear regression of the scaled Schoenfeld residuals defined at cost levels c on functions of cost under the analogous relationship $E(r_c^*) + \hat{\beta} \approx \beta(c)$. If the stratified version is adopted, it is recommended that this test be performed on each individual stratum, the reason being that the test described above assumes homogeneity of variance across risk sets, an assumption which may not be justified across different strata.

5.2.3. Proportional means regression

A related methodology has been proposed by Lin (2000) in which the mean cumulative cost is modelled as a function of an unspecified baseline mean function and a set of covariates as

$$\mu(t|Z) = \mu_0(t)e^{\beta'Z} \tag{5.10}$$

---

[2] The Schoenfeld residual at event time t_k is defined as

$r_k(\beta) = Z_{(k)} - \dfrac{\sum_{i=1}^{n} Y_i(t)e^{\beta' Z_i(t)} Z_i(t)}{\sum_{i=1}^{n} Y_i(t)e^{\beta' Z_i(t)}}$ where $Z_{(k)}$ denotes the covariate vector of individual who has an event at time t_k. These residuals first proposed by Schoenfeld are based on the individual contributions to the derivative of the log partial likelihood as can be seen from equation (5.2) above.

where $\mu(t|Z) = E\{N^*(t)|Z)\}$ is the mean cumulative cost at time t given the covariates with $N^*(t)$ denoting the cumulative cost up to time t and $\mu_0(t)$ is an unspecified baseline mean function of cost. The model is referred to as the proportional means regression model and is similar to the semiparametric Cox proportional hazards in that the baseline function is left unspecified and the covariates have a multiplicative effect on the mean cost. The model assumes that the underlying cost function is a process with positive jumps of arbitrary sizes. In the case where the process represents the cumulative number of some cost generating events, for example hospital admissions, the approach models the marginal mean for recurrent events and is in the same vein as the approach of modelling the marginal hazard functions for recurrent events or multivariate failure times studied by Lin et al (2000) and Andersen and Gill (1982) among others.

Lin argues that this specification avoids the problem of dependent censoring between cost at failure and cost at censoring as it models the mean cost at a point in time without assuming any dependence between failure time and cumulative cost or between the increments of the cumulative cost function $N^*(\cdot)$. The underlying assumption is now the proportionality of the mean costs across groups defined by different levels of covariates. Under random censoring the author presents the corresponding likelihood function incorporating adjustment for censoring, maximisation of which through iteration techniques leads to the estimation of the regression parameters. The estimation process is shown to be valid even in the case where the underlying process models accumulated cost expressed in monetary terms where the increments of the underlying cost function have non-negative arbitrary values. However, deriving an estimator of average cost over the duration of interest requires estimation of the unknown baseline mean function $\mu_0(t)$. A consistent estimator for $\mu_0(t)$ is given by

$$\hat{\mu}_0(t) = \sum_{i=1}^{n} \int_0^t \frac{dN_i^*(s)}{\hat{K}(s)\sum_{j=1}^{n} e^{\hat{\beta}'Z_j}} \tag{5.11}$$

where $\hat{K}(t)$ is the Kaplan-Meier estimator for $K(t) = pr(U > t)$ defined in the previous chapter in section 4.3.3.1. As can be seen from the above expression, deriving an estimator for $\mu_0(t)$ is only feasible if all sample paths of $N^*(\cdot)$ are known which in turn requires knowledge of the amount of cost accrual for all individuals at every point in time s over $(0, t]$ in order to determine the jumps of the process $dN^*(s)$ for all i's.[3] As the author points out this is unlikely to be the case when $N^*(t)$ represents charges because in most applications the accumulated cost is only recorded at given points in time, for example at the individual's death or last contact date, or at intermediate points in time corresponding to the time of the individual's cost generating event. Under these circumstances and on the assumption that the mean costs across groups defined by different levels of covariates are proportional the approach can only be used to estimate the regression parameters.

[3] The stochastic integrals appearing in the estimator for $\mu_0(t)$ are of the same form as shown in the previous chapter on page 87 from where it can be seen that their evaluation requires that all sample paths of $N^*(\cdot)$ be observed.

5.2.4. Weibull and exponential regression

Weibull regression

A fully parametric approach to assessing the effects of covariates on cumulative cost has been based on the on the Weibull and exponential regression models (e.g. Fenn et al, 1996). Viewed within the context of failure time analysis, in contrast with the proportional hazards model, such models specify a form for the hazard function over time. The Weibull distribution specifies the hazard function as

$$\lambda(t) = \lambda p(\lambda t)^{p-1}$$

for $\lambda, p > 0$. Thus the hazard is time dependent and is monotonically increasing for $p > 1$, monotonically decreasing for $p < 1$ and is constant for $p = 1$. By changing the parameter p and the scale parameter λ a variety of hazard functions are obtained. The density function of the Weibull distribution is given by $f(t) = \lambda p(\lambda t)^{p-1} e^{-(\lambda t)^p}$ and the survivor function is then $S(t) = e^{-(\lambda t)^p}$.

In general, the jth moment of the Weibull distribution is given by

$$E(T^j) = \lambda^{-j}\Gamma(1 + j/p)$$

Thus, the mean and variance for the failure time T under the parameterisation above are

$$E(T) = \lambda^{-1}\Gamma(1+1/p)$$
$$\text{var}(T) = \lambda^{-2}\{\Gamma(1+2/p) - \Gamma^2(1+1/p)\}$$

where $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$, $\alpha > 0$ is the Gamma function whose value is positive for all $\alpha > 0$ and $\Gamma(1) = 1$. This model can be generalised to incorporate the effect of covariates on failure time by allowing the hazard rate to be a function of both time and covariates. Assuming that time-invariant regressors enter λ in the form $\lambda = \exp\{-\beta'Z\}$ the hazard function given a set of covariates Z becomes

$$\lambda(t|Z) = pt^{p-1}(e^{-\beta'Z})^p$$

This form for the hazard function is also referred to as the accelerated failure time hazard on the basis that making λ a function of a set of regressors as shown above is equivalent to changing the units of measurement on the time axis. The corresponding density function is $f(t|Z) = pt^{p-1}(e^{-\beta'Z})^p e^{-t^p(e^{-\beta'Z})^p}$ and the survivor function is $S(t|Z) = e^{-t^p(e^{-\beta'Z})^p}$.

The mean and variance for the failure time T are then given as

$$E(T|Z) = \{\exp(-\beta'Z)\}^{-1}\Gamma(1+1/p)$$
$$\text{var}(T|Z) = \{\exp(-\beta'Z)\}^{-2}\{\Gamma(1+2/p) - \Gamma^2(1+1/p)\}$$

Exponential regression

The exponential distribution can be viewed as a special case of the Weibull distribution for $p = 1$. The corresponding hazard function is $\lambda(t) = \lambda$, that is the hazard is constant over time and the associated density function is $f(t) = \lambda e^{-\lambda t}$ with survivor function given as $S(t) = e^{-\lambda t}$. The mean and variance for the failure time T are

$$E(T) = \lambda^{-1}\Gamma(1+1) = \lambda^{-1}1! = \lambda^{-1}$$
$$\text{var}(T) = \lambda^{-2}\{\Gamma(1+2) - \Gamma^2(1+1)\} = \lambda^{-2}(2! - 1!) = \lambda^{-2}$$

Assuming the same parameterisation $\lambda = \exp\{-\beta'Z\}$ for incorporating covariate effects results in a hazard function conditional upon a set of covariates Z given by

$$\lambda(t|Z) = te^{-\beta'Z}$$

The corresponding density function is $f(t|Z) = te^{-\beta'Z}e^{-te^{-\beta'Z}}$ and the survivor function is $S(t|Z) = e^{-te^{-\beta'Z}}$. The mean and variance for the failure time T are then

$$E(T|Z) = \{\exp(-\beta'Z)\}^{-1}\Gamma(1+1) = e^{\beta'Z}$$
$$\text{var}(T|Z) = \{\exp(-\beta'Z)\}^{-2}\{\Gamma(1+2) - \Gamma^2(1+1)\} = (e^{\beta'Z})^2$$

Application to cost

Both the Weibull and exponential regression models have been applied to study the effects of covariates on cumulative cost and to provide an estimate of average cost over the duration of interest (Fenn et al 1996). In the application of such models to censored cost data, the hazard rate specifies the conditional probability of having "completed" a given cost conditional upon a set of covariates. That is, the hazard rate gives the probability of dying having attained c units of cost given the covariates and that the individual was alive after having attained c-1 units of cost. Under the Weibull regression model the hazard function for cost given a set of time independent covariates Z is

$$\lambda(c|Z) = pc^{p-1}(e^{-\beta'Z})^p \tag{5.12}$$

with density function $f(c|Z) = pc^{p-1}(e^{-\beta'Z})^p e^{-c^p(e^{-\beta'Z})^p}$ and survivor function $S(c|Z) = e^{-c^p(e^{-\beta'Z})^p}$. If the hazard function is increasing, this implies that the likelihood of completing a given cost conditional upon having reached that cost level and the covariates is increasing in cost, whereas if the hazard function is decreasing this implies that the likelihood of completing a given cost conditional upon having reached that cost level and the covariates is decreasing in cost. The mean and variance for the random variable *M* denoting cost are

$$E(M|Z) = \{\exp(-\beta'Z)\}^{-1}\Gamma(1+1/p) \tag{5.13}$$

$$\text{var}(M|Z) = \{\exp(-\beta'Z)\}^{-2}\{\Gamma(1+2/p) - \Gamma^2(1+1/p)\} \tag{5.14}$$

Similarly the cost hazard under the exponential regression model is given by

$$\lambda(c|Z) = ce^{-\beta'Z} \tag{5.15}$$

where the corresponding density function is $f(c|Z) = ce^{-\beta'Z}e^{-ce^{-\beta'Z}}$ and the survivor function is $S(c|Z) = e^{-ce^{-\beta'Z}}$. This specification implies that the likelihood of completing a particular cost given that the individual has reached that cost level and the covariates is independent of the cost level, i.e. is constant over cost levels. The mean cost and its variance are

$$E(M|Z) = \{\exp(-\beta'Z)\}^{-1}\Gamma(1+1) = e^{\beta'Z} \tag{5.16}$$

$$\text{var}(M|Z) = \{\exp(-\beta'Z)\}^{-2}\{\Gamma(1+2) - \Gamma^2(1+1)\} = (e^{\beta'Z})^2 \tag{5.17}$$

As in the non-parametric approach to the analysis of time to event data, the central concept in the semiparametric and parametric approaches considered above is the conditional probability of an event occurring at a given point in time given that it has not occurred until that point in time as modelled through the hazard functions. The additional element here is that the hazard function is also a function of covariates. For all these models independent censoring requires that individuals who are censored at time *t* after allowing for covariates be representative of all individuals who are still under observation at *t*. When applying these approaches to modelling cost to event data individuals who are censored having attained a particular cost level must be representative of all individuals who are still under observation having attained that cost level. This is not the case when the rate of cost accumulation varies across individuals and therefore all these approaches are generally inappropriate for the analysis of censored cost data.

5.2.5. Least squares regression

5.2.5.1. The classical linear regression model

Under the classical linear regression model the relationship between the outcome variable Y and a set of covariates Z is

$$Y = \beta' Z + \varepsilon$$

where β is a $p \times 1$ vector of unknown regression parameters and ε is a zero-mean error term assumed normally distributed with constant variance σ^2. This model is known to give biased estimates when applied to censored data (Lancaster 1990, Greene 1997). Following Green (1997), the reason why this bias arises can be shown in the following manner. The relevant distribution theory for a censored random variable is similar to that for a truncated one.[4] In general, for the moments of a truncated normal distribution the following theorem holds for the random variable x (Green 1977, p.951).

If $x \sim N[\mu, \sigma^2]$ and the truncation point is w where w is a constant,

$$E(x|\text{truncation}) = \mu + \sigma\lambda(\alpha)$$
$$\text{Var}(x|\text{truncation}) = \sigma^2[1 - \delta(\alpha)]$$

where

$\alpha = (w - \mu)/\sigma$,
$\lambda(\alpha) = \phi(\alpha)/[1 - \Phi(\alpha)]$ if truncation is $x > w$,
$\lambda(\alpha) = -\phi(\alpha)/\Phi(\alpha)$ if truncation is $x < w$,
$\delta(\alpha) = \lambda(\alpha)[\lambda(\alpha) - \alpha]$ with $0 < \delta(\alpha) < 1$,
$\phi(\alpha) = \frac{1}{\sqrt{2\pi}} e^{-\alpha^2/2}$ is the density function of the standard normal distribution $N[0,1]$ and
$\Phi(\alpha) = \int_{-\infty}^{\alpha} \phi(x)dx$ is the cumulative distribution function.

This theorem is used to derive the moments of the censored random distribution as follows. Assuming that the censored random variable y^* follows a normal distribution with censoring occurring at the upper part of the distribution at point w, a new random variable y is defined from the original y^* as

[4] In the econometrics literature the distinction between truncation and censoring with respect to the regression model is that a censored regression model is one where the dependent variable is not fully observed whereas a truncated regression model is one where both the dependent and independent variables are not fully observed.

$y = w \quad \text{if } y^* \geq w$

$y = y^* \quad \text{if } y^* < w$

It follows that if $y^* \sim N[\mu, \sigma^2]$ then using the above relations, the mean of the censored random variable y is

$$\begin{aligned} E(y) &= pr(y = w) * E(y|y = w) + pr(y < w) * E(y|y < w) \\ &= pr(y^* \geq w) * w + pr(y^* < w) * E(y^*|y^* < w) \\ &= (1 - \Phi) * w + \Phi * (\mu + \sigma\lambda) \end{aligned}$$

and the variance is shown to be

$$\text{var}(y) = \sigma^2 \Phi[(1 - \delta) + (\alpha - \lambda)^2 (1 - \Phi)]$$

If censoring occurs at the lower part of the distribution, the same expressions apply with the roles of Φ and $(1 - \Phi)$ reversed and λ defined as shown in the above theorem.

If censoring occurs from above at $w = 0$ and because the distribution is symmetric which implies $\phi(\alpha) = \phi(-\alpha)$ and $\Phi(-\alpha) = 1 - \Phi(\alpha)$, the mean of the right-censored random variable is

$$\begin{aligned} E(y) &= \Phi(-\mu/\sigma) * [\mu + \sigma\lambda(-\mu/\sigma)] \\ &= \Phi(-\mu/\sigma)] * \left[\mu - \sigma \frac{\phi(-\mu/\sigma)}{\Phi(-\mu/\sigma)}\right] \\ &= [1 - \Phi(\mu/\sigma)]\mu - \sigma\phi(\mu/\sigma) \end{aligned}$$

showing that the mean of the right-censored random variable will be lower than the mean of the original random variable.

Similarly, if censoring occurs from below at $w = 0$ the mean of the left-censored random variable is

$$\begin{aligned} E(y) &= [1 - \Phi(-\mu/\sigma)] * [\mu + \sigma\lambda(-\mu/\sigma)] \\ &= \Phi(\mu/\sigma)] * \left[\mu + \sigma \frac{\phi(-\mu/\sigma)}{1 - \Phi(-\mu/\sigma)}\right] \\ &= \Phi(\mu/\sigma)\mu + \sigma\phi(\mu/\sigma) \end{aligned}$$

showing that the mean of the censored random variable with censoring from below will be higher than the mean of the original random variable.

Extending the above to the case of the censored regression model, also referred to as the Tobit model (Tobin, 1958),[5] this would be obtained by making the mean correspond to a classical regression model and for the case of right-censoring at w would be given as

$$y_i^* = \beta' Z_i + \varepsilon_i$$

$$y_i = w \quad \text{if } y_i^* \geq w$$
$$y_i = y^* \quad \text{if } y_i^* < w$$

$$\varepsilon_i \sim N[0, \sigma^2] \text{ and } \mu = \beta' Z$$

Based on the preceding results, for an observation y_i which might be right-censored,

$$E(y_i|Z_i) = \{1 - \Phi[(w - \beta' Z_i)/\sigma]\}w + \Phi[(w - \beta' Z_i)/\sigma]\left[\beta' Z_i - \frac{\phi[(w - \beta' Z_i)/\sigma]}{\Phi[(w - \beta' Z_i)/\sigma]}\right]$$

as opposed to $E(y_i|Z_i) = \beta' Z_i$ which would be the case if there was no censoring.

Application to cost

When the classical linear regression model is adopted to study the effect of covariates on cost, the relationship between cost and the set of covariates Z is

$$M = \beta' Z + \varepsilon \tag{5.18}$$

where β is a $p \times 1$ vector of unknown regression parameters and ε is a zero-mean error term assumed normally distributed with constant variance σ^2. Setting the first component of Z equal to 1 makes the first component of β correspond to the intercept. In the absence of censoring β is estimated by the least-squares normal equation

$$\sum_{i=1}^{n} (M_i - \beta' Z_i) Z_i = 0 \tag{5.19}$$

---

[5] The Tobit model was originally defined for censoring of the lower part of the distribution with censoring occurring at point zero as

$$y_i^* = \beta' Z_i + \varepsilon_i$$
$$y_i = 0 \quad \text{if } y_i^* \leq 0$$
$$y_i = y^* \quad \text{if } y_i^* > 0$$

Here, the analysis is modified to reflect the case of censoring of the upper part of the distribution with censoring occurring at point w.

In the presence of censoring, estimation by the above equation will lead to biased estimates for the regression parameters as shown above. A naïve approach is then to estimate the model by including only the uncensored cases in the estimation process. The regression parameters are again estimated by the least-squares normal equation but now only individuals with complete cost observations contribute information to the estimation process. As is the case in any similar missing data situation such an analysis, referred to as complete case analysis, which totally discards the cases with missing values leads to loss of information which could be a substantial problem if the proportion of cases with missing values is high. On this basis the approach has been deemed useful only for providing a baseline method for comparisons. The mean cost over $(0, L]$ is then estimated as $\hat{M} = \hat{\beta}'\widetilde{Z}$, where $\widetilde{Z}$ denotes the covariates vector evaluated at the mean values of the covariates and $\hat{\beta}$ is the vector of the estimated regression parameters where only the uncensored cases have been used in the estimation process.

5.2.5.2. Least squares regression analysis with randomly right-censored data

In the context of time to event data

A number of alternative approaches have been proposed to handle regression problems when the dependent variable is subject to censoring. In the context of failure time data analysis, a number of such models have been proposed to study the effect of covariates on censored failure time some of which assume specific parametric families for the failure time distribution such as the Weibull and exponential while others are free of such assumptions. Of the regression techniques that do not assume specific parametric families for the failure time distribution, one of which is the Cox proportional hazards discussed above, the ones following a least squares approach are described below.

As stated previously in the presence of right censoring the observables are $X_i = \min(T_i, U_i)$, $\delta_i = I(T_i \leq U_i)$ and a $p \times 1$ vector of covariates Z_i. The general form of the linear model when the dependent variable is failure time T is given as

$$T_i = \beta' Z_i + \varepsilon_i \quad i = 1,...,n \tag{5.20}$$

where Z_i is the covariate vector and i identifies individuals. Under random censorship, the error terms are assumed to be independent and identically distributed random variables with zero mean and the censoring variables $U_i's$ are independent and identically distributed random variables which are independent of the error terms.

Miller (1976) introduced an estimator for the unknown parameters β which is derived by minimising the weighted sum of squares of the residuals with the weights determined by the

Kaplan-Meier estimator of the error distribution based on the residuals. Assuming one covariate only, model (5.20) above is

$$F(t|Z) = F(t - \beta_0 - \beta_1 Z)$$

with

$$E(T|Z) = \beta_0 + \beta_1 Z \tag{5.21}$$

The estimator proposed by Miller is derived by minimising the sum of squares

$$n\int e^2 d\hat{F}(e;\beta_0,\beta_1) \tag{5.22}$$

with respect to β_0, β_1 where $\hat{F}(e;\beta_0,\beta_1)$ is the Kaplan-Meier estimator based on the data $\{\delta_1, \hat{e}_i = X_i - \beta_0 - \beta_1 Z, i = 1,...,n\}$, that is,

$$\hat{F}(e;\beta_0,\beta_1) = 1 - \prod_{\hat{e}_i \le e}(1 - \frac{d_{(\hat{e}_i)}}{n_{(\hat{e}_i)}})$$

where $\hat{e}_1 < \hat{e}_2 < ...$ are the distinct ordered values of $\hat{e}_i$, $d_{(\hat{e}_i)}$ are the number dying at $\hat{e}_i$ and $n_{(\hat{e}_i)}$ are the number at risk strictly prior to $\hat{e}_i$, i.e. at $\hat{e}_i -$. Expression (5.22) can be viewed as a generalisation of the usual sum of squares $\sum(X_i - \beta_0 - \beta_1 Z)^2$ for uncensored data. Because (5.22) is a discontinuous function of β_1, it is difficult to locate the infimum point and hence Miller proposed an iterative sequence for estimating the regression parameter β_1 (Miller, 1976).

Buckley and James (1979) modified Miller's approach by basing their estimation on the censored-data analogue of the normal equations rather than the least squares criterion of minimising the sum of squares. Their estimator for the regression coefficients is based on the following relationship.

$$E\{\delta_i X_i + (1-\delta_i)E(T_i|T_i > X_i)|Z_i\} = \beta_0 + \beta_1 Z_i \tag{5.23}$$

By replacing the conditional expectation $E(T_i|T_i > X_i)$ with an estimate based on the Kaplan-Meier estimator in the expression above, estimates of the coefficients are derived by solving the usual least squares normal equations iteratively. Thus, the censored observations are replaced by their expectations and then the sum of squares is minimised.

Koul et al (1981) introduced an estimator that does not require iteration methods and applies the adjustment for censoring to the original observations rather than the estimated residuals. Their estimator is based on the following relationship

$$E(\delta_i X_i) = -\int tK(t)dF(t - \beta_0 - \beta_1 Z_i) \quad \text{for all } t \tag{5.24}$$

where $F_i(t) = pr(X_i > t) = F(t - \beta_0 - \beta_1 Z_i)$ and $K(t) = pr(U > t)$, that is, $K(t)$ is the survivor function of the censoring distribution. For $K(t) > 0$, this yields

$$E[\delta_i X_i \{K(X_i)\}^{-1}] = -\int tdF(t - \beta_0 - \beta_1 Z_i) = \beta_0 + \beta_1 Z_i$$

that is, the variables $\{\delta_i X_i \{K(X_i)\}^{-1}, i = 1,...,n\}$ follow a linear regression model which has the same parameters as model (5.20) but the error terms here need not be identically distributed. Koul et al propose an estimator for the unknown survivor function $K(t)$ given as

$$\hat{K}(t) = \prod_{j=1}^{n} \left(\frac{1 + \sum_{j=1}^{n} I(X_i > X_j)}{2 + \sum_{j=1}^{n} I(X_i > X_j)} \right)^{[\delta_j = 0, X_j \le t]} \quad \text{for all } t$$

Replacing $K(t)$ with its estimator in the expectation $E[\delta_i X_i \{K(X_i)\}^{-1}]$ allows estimation of the regression coefficients β using standard least squares methods. As such, the great advantage of this technique is that the regression parameters are estimated without requiring iteration procedures.

Consistency of the various estimators presented in this section has also been considered and estimators for the covariance matrix of the regression parameters have been derived in each case. The idea of weighting the uncensored observations by the inverse of their probabilities of not being censored within the context of regression analysis which underlies the approach by Koul et al (1981) has been also used by Lin et al (2000) in deriving estimators of cost adjusting for covariate effects under conditions of censoring as will be shown below.

In the context of cost to event data

Lin regression methodology

Assuming the general setting as defined in section 5.2.1 and defining $T^* = \min(T, L)$ with Z being a $p \times 1$ vector of covariates whose effect on the cumulative cost at T^* one wishes to study, the methodology presented in this section introduced by Lin (2000) attempts to adjust the estimates derived by the linear model given as

$$M = \beta' Z + \varepsilon$$

where β is a $p \times 1$ vector of unknown regression parameters and ε is a zero-mean error term with an unspecified distribution for censoring. The first component of Z is set equal to 1 so that the first component of β corresponds to the intercept. As stated above, in the absence of censoring β is estimated by the least-squares normal equation

$$\sum_{i=1}^{n}(M_i - \beta' Z_i)Z_i = 0$$

In the presence of censoring, under the assumption of a continuous distribution for failure time over $(0, L]$ and a continuous distribution of censoring time with censoring arising in a completely random manner, time to censoring has survivor function $K(u) = pr(U > u)$, i.e. the survivor function $K(u)$ evaluated at a point in time u gives the probability of an individual not being censored at u. Defining $\delta_i^* = I(U \geq T^*)$ under random censoring conditions the estimating equation for β is modified as

$$\sum_{i=1}^{n}\frac{\delta_i^*}{K(T_i^*)}(M_i - \beta' Z_i)Z_i = 0$$

which implies that only individuals with complete cost observations over the duration of interest contribute cost information to the estimation process. The unknown survivor function $K(\cdot)$ is estimated by the Kaplan-Meier estimator based on the data $\{X_i = \min(T_i, U_i), 1 - \delta_i, i = 1,...,n\}$ as

$$\hat{K}(t) = \prod_{u \leq t}\left\{1 - \frac{dN^c(u)}{Y(u)}\right\} \tag{5.25}$$

where the counting processes $N^c(u)$ and $Y(u)$ have been defined in section 4.1.2.6. Replacing the survivor function $K(\cdot)$ with its consistent Kaplan-Meier estimator results in the following estimating equation for β.

$$\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*)}(M_i - \beta' Z_i)Z_i = 0 \tag{5.26}$$

whose solution is given as

$$\hat{\beta} = \left\{\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*)}Z_i^{\otimes 2}\right\}^{-1}\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*)}M_i Z_i \tag{5.27}$$

where $\alpha^{\otimes 0} = 1, \alpha^{\otimes 1} = \alpha, \alpha^{\otimes 2} = \alpha\alpha'$.

Lin (2000) studies the asymptotic properties of this estimator and derives estimates for its covariance matrix for large samples as outlined below. The left hand of (5.26) is written as

$$\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*)}(M_i-\beta' Z_i)Z_i =$$

$$\sum_{i=1}^{n}\frac{\delta_i^*}{K(T_i^*)}(M_i-\beta' Z_i)Z_i+\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*)}(M_i-\beta' Z_i)Z_i-\sum_{i=1}^{n}\frac{\delta_i^*}{K(T_i^*)}(M_i-\beta' Z_i)Z_i =$$

$$\sum_{i=1}^{n}\frac{\delta_i^*}{K(T_i^*)}(M_i-\beta' Z_i)Z_i+\sum_{i=1}^{n}\frac{K(T_i^*)-\hat{K}(T_i^*)}{K(T_i^*)\hat{K}(T_i^*)}\delta_i^*(M_i-\beta' Z_i)Z_i$$

$U(\beta)=U_1(\beta)+U_2(\beta)$, say.

Given that $\{T_i,U_i,Z_i\},(i=1,..,n)$ are assumed independent and identically distributed and failure time and censoring time are independent conditional on Z_i implying $E(\delta_i^*|M_i,Z_i,T_i^*)=E(\delta_i^*|T_i^*)=K(T_i^*)$, the term $U_1(\beta)$ consists of n independent zero mean random vectors being the sum of n vectors each one pertaining to one individual. To derive an independent and identically distributed representation for the term $U_2(\beta)$, Lin uses the martingale process associated with the process counting censored individuals defined in section 4.3.3.1 as

$$\mathscr{M}_i^c(u)=N_i^c(u)-\int_0^u\lambda^c(t)Y_i(t)dt$$

where $N_i^c(u)=I(X_i\le u,\delta_i=0)$, $Y_i(u)=I(X_i\ge u)$ and $\lambda^c(u)$ is the hazard function for the censoring distribution with $\mathscr{M}^c(u)=\sum\mathscr{M}_i^c(u)$, $N^c(u)=\sum N_i^c(u)$ and $Y(u)=\sum Y_i(u)$. Modifying the expression (4.15) shown in section 4.3.2.1 to relate to the censoring time, expression (4.15) becomes

$$\begin{aligned}n^{1/2}(\hat{\Lambda}^c-\Lambda^c)&=n^{1/2}\sum_{i=1}^{n}\int_0^t\frac{I\{Y(u)>0\}}{Y(u)}d\mathscr{M}_i^c(u)+o_p(1)\\&=n^{1/2}\sum_{i=1}^{n}\int_0^t\frac{d\mathscr{M}_i^c(u)}{\sum_{j=1}^{n}I(X_j\ge u)}+o_p(1)\end{aligned}\qquad(5.28)$$

where $\hat{\Lambda}^c(t)=\int_0^t\frac{I\{Y(u)>0\}}{Y(u)}dN^c(u)$ is the Nelson-Aalen estimator for the integrated hazard function $\Lambda^c(t)=\int_0^t\lambda^c(u)du$ for censoring. As shown in the same section (section 4.3.2.1), by the Taylor series expansion

$$n^{1/2}(\hat{K}(t)-K(t))\approx -K(t)n^{-1/2}\sum_{i=1}^{n}\int_0^t\frac{d\mathscr{M}_i^c(u)}{n^{-1}\sum_{j=1}^{n}I(X_j\ge u)}$$

Thus

$$n^{1/2}\frac{K(t)-\hat{K}(t)}{K(t)} = n^{-1/2}\sum_{i=1}^{n}\int_0^t \frac{d\mathscr{M}_i^c(u)}{n^{-1}\sum_{j=1}^{n} I(X_j \geq u)} + o_p(1)$$

which leads to the following expression for term $U_2(\beta)$:

$$n^{-1/2}U_2(\beta) = n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty}\frac{1}{n}\sum_{i=1}^{n}\frac{I(T_i^* > t)\delta_i^*(M_i - \beta' Z_i)Z_i}{n^{-1}\sum_{j=1}^{n} I(X_j \geq t)\hat{K}(T_i^*)} d\mathscr{M}_i^c(t) + o_p(1)$$

By the law of large numbers and due to the consistency of $\hat{K}(\cdot)$

$\lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\frac{I(T_i^* > t)\delta_i^*(M_i - \beta' Z_i)Z_i}{n^{-1}\sum_{j=1}^{n} I(X_j \geq t)\hat{K}(T_i^*)} = q(t)$, where $q(t)$ is well-defined.

Thus

$$n^{-1/2}U(\beta) = n^{-1/2}\sum_{i=1}^{n}\frac{\delta_i^*(M_i - \beta' Z_i)Z_i}{\hat{K}(T_i^*)} + n^{-1/2}\sum_{i=1}^{n}\int_0^{\infty} q(t)d\mathscr{M}_i^c(t) + o_p(1)$$

$$= n^{-1/2}\sum_{i=1}^{n}\xi_i + o_p(1)$$

Because ξ_i $(n = 1,...,n)$ are n independent zero-mean random matrices, the central limit theorem implies that $n^{-1/2}U(\beta)$ converges in distribution to a zero mean normal random matrix with limiting covariance matrix given as $B = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\xi_i\xi_i'$. By the Taylor series expansion, $n^{1/2}(\hat{\beta} - \beta) = \tilde{A}^{-1}n^{-1/2}U(\beta)$, where $\tilde{A} = n^{-1}\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*)}Z_i^{\otimes 2}$ which converges in probability to $A \equiv \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Z_i^{\otimes 2}$. From the asymptotic normality of $n^{-1/2}U(\beta)$, it follows that $n^{1/2}(\hat{\beta} - \beta)$ converges in distribution to a zero mean normal random matrix and the limiting covariance is $A^{-1}BA^{-1}$.

Replacing the unknown quantities in the expressions above with their respective sample estimators, a consistent estimator for the covariance matrix is given as

$$\hat{A}^{-1}\hat{B}\hat{A}^{-1} \qquad (5.29)$$

where

$$\hat{A} = n^{-1}\sum_{i=1}^{n} Z_i^{\otimes 2} \tag{5.30}$$

$$\hat{B} = n^{-1}\sum_{i=1}^{n}\left[\frac{\delta_i^*(M_i - \hat{\beta}'Z_i)Z_i}{\hat{K}(T_i^*)} + \bar{\delta}_i Q(X_i) - \sum_{j=1}^{n}\frac{\bar{\delta}_j I(X_j \le X_i)Q(X_j)}{\sum_{l=1}^{n} I(X_l \ge X_j)}\right]^{\otimes 2} \tag{5.31}$$

with $\bar{\delta}_i = 1 - \delta_i$ and

$$Q(t) = \sum_{i=1}^{n}\frac{I(T_i^* > t)\delta_i^*(M_i - \hat{\beta}'Z_i)Z_i}{\hat{K}(T_i^*)} \Big/ \sum_{j=1}^{n} I(X_j \ge t) \tag{5.32}$$

The mean cost over $(0, L]$ is then estimated as $\hat{M} = \hat{\beta}'\tilde{Z}$, where $\tilde{Z}$ denotes the covariates vector evaluated at the mean values of the covariates.

Multiple time intervals

The second approach presented by Lin (2000) extends the previous idea in situations where information on individual cost histories is available at various point in time over the duration of interest. The main purpose of this method is to increase efficiency by allowing use of cost information not being used by the preceding estimator. The approach draws on the methods for analysing longitudinal data using generalised linear models proposed by Liang et al (1986). The general setting for these approaches is outlined below.

Such data consist of an outcome variable and a $p \times 1$ vector of covariates observed at various points in time for each individual *i*. An issue inherent in longitudinal data, that is data consisting of multiple observations for each subject, is the dependency among the repeated measurements for any given subject. In such circumstances ordinary least squares is not an appropriate estimation procedure as the assumptions concerning the error terms are no longer valid. The general procedure to analysing such data in the econometric and statistical literature is to adopt an alternative model to ordinary least squares, referred to as the generalised linear regression model, which accommodates more general patterns for the distribution of the disturbances.

In general in the analysis of longitudinal data, interest lies either in studying the change over time or in assessing the dependence of the outcome variable on the covariates. Liang et al (1986) proposed a class of generalised estimating equations for the regression parameters which result in consistent estimates of the regression parameters and of their variance without requiring specification of the joint distribution of a subject's observations. This approach has wide application if interest is in modelling the dependence of the outcome variable on the covariates and not in the pattern of change of the outcome variable over time. Under these circumstances, the approach models the marginal

expectation of the outcome variable as a function of the covariates at each point in time whilst accounting for the correlation among the repeated measurements for a given subject by treating the time dependence among repeated measurements for an individual as a nuisance. When the time dependence is of primary importance, models for the conditional distribution of the outcome variable given its past values would be more appropriate and then the joint distribution of a subject's observations would need to be specified. The authors argue that if observations gained from different subjects are independent, the estimates of the regression parameters will be consistent, provided that the model for the marginal means of the outcome variable at each time is correctly specified even if the correlation structure, that is, the time dependence among repeated observations for a given subject, is misspecified. More importantly, this approach can also be applied in the event of some observations being missing, in which case the same results hold provided that data are missing completely at random in the sense of Rubin (1976).[6] This type of model, that only assumes a functional form for the marginal distribution of the outcome variable at each time and treats the correlation structure over time as a nuisance, describes how the average response across individuals changes with the covariates and as a result consistency of the estimators for the regression parameters depends only on the correct specification of the functional form for the marginal expectation of the outcome variable. Using this approach, whereby the marginal expectation of the outcome variable is modelled as a function of covariates, the estimated regression coefficients have an interpretation for the population on average rather than for any individual in particular.

Adopting the same framework, Lin (2000) models the marginal expectation of cost at each point in time for which cost information is available as a function of the covariates as follows. The duration of analysis $(0, L]$ is partitioned into K subintervals $(t_{k-1}, t_k], (k = 1,...,K)$, with $t_0 = 0$ and $t_K = L$, and for each subinterval k the following linear model is assumed.

$$M_{ki} = \beta_k' Z_i + \varepsilon_{ki} \qquad k = 1,...,K \qquad i = 1,...,n$$

where for individual i $M_{ik} = M_i(t_k) - M_i(t_{k-1})$ is the cost incurred over subinterval $(t_{k-1}, t_k]$, β_k $(k = 1,...,K)$ are $p \times 1$ vectors of unknown regression parameters and the error terms ε_{ki}'s are assumed to be independent among different subjects but allowed to be correlated within the same subject. By summing over all k subintervals, the linear model for the cost over the whole duration of interest becomes

$$M_i = \beta' Z_i + \varepsilon_i \qquad i = 1,...,n$$

[6] The missing data are missing at random if for each parameter value, the conditional probability of the observed pattern of missing data given the missing data and the value of the observed data is free of the missing data (see chapter 2, section 2.4.4)

where $M_i = \sum_{k=1}^{K} M_{ki}$, $\beta = \sum_{k=1}^{K} \beta_k$, and $\varepsilon_i = \sum_{k=1}^{K} \varepsilon_{ki}$. Defining $T_{ki}^* = \min(T_i, t_k)$ and $\delta_{ki}^* = I(U_i \geq T_{ki}^*)$, i.e. $\delta_{ki}^* = I\{\min(T_i, t_k) \leq U_i\}$, the estimating equation for β_k $(k = 1,...,K)$ is given as

$$\sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{K}(T_{ki}^*)}(M_{ki} - \beta_k' Z_i)Z_i = 0 \tag{5.33}$$

where $\hat{K}(T_{ki}^*)$ is the Kaplan-Meier estimator for the probability of not being censored based on the dataset $\{X_{ki}, \delta_{ki}^*, i = 1,...,n\}$ where $X_{ki} = \min(T_{ki}^*, U_i)$.

The solution to the above estimating equation is then given as

$$\hat{\beta}_k = \left\{\sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{K}(T_{ki}^*)} Z_i^{\otimes 2}\right\}^{-1} \sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{K}(T_{ki}^*)} M_{ki} Z_i \tag{5.34}$$

with

$$\hat{\beta} = \sum_{k=1}^{K}\left[\left\{\sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{K}(T_{ki}^*)} Z_i^{\otimes 2}\right\}^{-1} \sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{K}(T_{ki}^*)} M_{ki} Z_i\right] \tag{5.35}$$

Comparing this estimator with its counterpart from the previous approach, the gain in cost information is due to the fact that here a subject contributes cost information to the estimating equations over all time intervals for which the individual is not censored, i.e. over all k's for which $U_i > \min(T_i^*, t_k)$. By contrast, in equation (5.26) an individual only contributes cost information to the estimates if the individual's censoring time exceeds the maximum observed time in the study.

In studying the asymptotic properties of this estimator, the same methodology as above is adopted and the left side of (5.33) becomes

$$\begin{aligned} n^{-1/2} U_k(\beta_k) &= n^{-1/2} \sum_{i=1}^{n} \frac{\delta_{ki}^*(M_{ki} - \beta_k' Z_i)Z_i}{\hat{K}(T_{ki}^*)} + n^{-1/2} \sum_{i=1}^{n} \int_0^{\infty} q_k(t) d\mathcal{M}_i^c(t) + o_p(1) \\ &= n^{-1/2} \sum_{i=1}^{n} \xi_{ki} + o_p(1) \end{aligned}$$

Because ξ_{ki} $(n = 1,...,n)$ are n independent zero-mean random matrices, the central limit theorem implies that $n^{-1/2}\{U_1(\beta_1),...,U_K(\beta_K)\}$ converges in distribution to a zero mean normal random matrix with limiting covariance matrix between $n^{-1/2}U_k(\beta_k)$ and $n^{-1/2}U_l(\beta_l)$ given as $B_{kl} = \lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} \xi_{ki}\xi_{li}'$ $(k,l = 1,..,K)$. By the Taylor series expansion,

$n^{1/2}(\hat{\beta}_k - \beta_k) = \tilde{A}_k^{-1} n^{-1/2} U_k(\beta_k)$, where $\tilde{A}_k = n^{-1}\sum_{i=1}^{n} \frac{\delta_{ki}^*}{\hat{K}(T_{ki}^*)} Z_i^{\otimes 2}$ which converges in probability to $A \equiv \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n} Z_i^{\otimes 2}$. From the asymptotic normality of $n^{-1/2}\{U_1(\beta_1),...,U_K(\beta_K)\}$, it follows that $n^{1/2}(\hat{\beta}_1 - \beta_1,...,\hat{\beta}_K - \beta_K)$ converges in distribution to a zero mean normal random matrix and the limiting covariance between $n^{1/2}(\hat{\beta}_k - \beta_k)$ and $n^{1/2}(\hat{\beta}_l - \beta_l)$ is $A^{-1}B_{kl}A^{-1}$.

Replacing the unknown quantities in the expressions above with their respective sample estimators, a consistent estimator for the covariance matrix is given as

$$\hat{A}^{-1}\hat{B}\hat{A}^{-1} \tag{5.36}$$

where

$$\hat{B} = \sum_{k=1}^{K}\sum_{l=1}^{K} \hat{B}_{kl} \tag{5.37}$$

$$\hat{A} = n^{-1}\sum_{i=1}^{n} Z_i^{\otimes 2} \tag{5.38}$$

$$\hat{B}_{kl} = n^{-1}\sum \hat{\xi}_{ki}\hat{\xi}_{li}' \tag{5.39}$$

$$\hat{\xi}_{ki} = \frac{\delta_{ki}^*(M_{ki} - \hat{\beta}_k' Z_i)Z_i}{\hat{K}(T_{ki}^*)} + \bar{\delta}_i Q_k(X_i) - \sum_{j=1}^{n} \frac{\bar{\delta}_j I(X_j \le X_i) Q_k(X_j)}{\sum_{l=1}^{n} I(X_l \ge X_j)} \tag{5.40}$$

$$Q_k(t) = \sum_{i=1}^{n} \frac{I(T_{ki}^* > t)\delta_{ki}^*(Y_{ki} - \hat{\beta}_k' Z_i)Z_i}{\hat{K}(T_{ki}^*)} \Big/ \sum_{j=1}^{n} I(X_j \ge t) \tag{5.41}$$

The mean cost over $(0, L]$ is then estimated as $\hat{M} = \hat{\beta}'\tilde{Z}$, where $\tilde{Z}$ denotes the covariates vector evaluated at the mean values of the covariates.

Censoring dependent on covariates

Both approaches described above are generalised to the case of covariate-dependent censoring. In the context of censored time failure data the issue of covariate dependent censoring has been addressed by Robins and Rotnizky (1992), Rotnitzky and Robins (1995) and Robins, Rotnizky and Zhao (1995) using a semiparametric regression methodology. The proposed estimating equations use the inverse of the probability of an individual not being censored as the weight to adjust for missingness due to censoring and the resultant estimators are consistent when the data are missing at random in the sense of Rubin (1976). The model is specified based on the assumption that the probability of censoring at any given time *t* is independent of the outcome variable conditional on

the history up to time t of a vector of time dependent covariates that are correlated with the outcome variable.

To accommodate covariate dependent censoring, Lin (2000) proposes using the proportional hazards specification (Cox, 1972) to model the effect of covariates on the censoring distribution allowing formulation of the dependence of censoring both on discrete covariates, which might be used as stratification variables, and on continuous covariates as

$$\lambda(t|V,W) = e^{\gamma' W(t)}\lambda_V(t)$$

where V represents potential stratification variables and W the remaining covariates, $\lambda(t|V,W)$ is the conditional hazard function for censoring given V and W, $\lambda_V(\cdot)$ is an unspecified baseline hazard function for stratum V and γ is a set of unknown regression parameters. Censoring is assumed independent of all other random variables given (V, W) and the set of covariates V and W are allowed to include a subset of Z.

The survivor function of the censoring distribution is estimated as

$$\hat{K}(T_i^*|V_i,W_i) = \exp\left\{-\sum_{j=1}^{n}\frac{\bar{\delta}_j I(V_j = V_i, X_j < T_i^*)e^{\hat{\gamma}' W_i(X_j)}}{S^{(0)}(X_j;V_i,\hat{\gamma})}\right\} \tag{5.42}$$

where $\hat{\gamma}$ is the partial maximum likelihood estimator of γ and

$$S^{(\rho)}(t;V,\gamma) = \sum_{i=1}^{n} I(V_i = V, X_i \geq t)e^{\gamma' W_i(t)}W_i^{\otimes\rho}(t)$$

The estimating equation for β is

$$\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*|V_i,W_i)}(M_i - \beta' Z_i)Z_i = 0 \tag{5.43}$$

whose solution is given as

$$\hat{\beta} = \left\{\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*|V_i,W_i)}Z_i^{\otimes 2}\right\}^{-1}\sum_{i=1}^{n}\frac{\delta_i^*}{\hat{K}(T_i^*|V_i,W_i)}M_i Z_i \tag{5.44}$$

It can be seen that these expressions reduce to the respective expressions for covariate independent censoring for $V_i = 1$ and $W_i = 0$ for all $i = 1,...,n$.

When multiple time intervals are considered the corresponding estimating equation for β_k $(k = 1,...,K)$ is given as

$$\sum_{i=1}^{n} \frac{\delta_{ki}^{*}}{\hat{K}(T_{ki}^{*}|V_i, W_i)} (M_{ki} - \beta_k' Z_i) Z_i = 0 \tag{5.45}$$

whose solution is

$$\hat{\beta}_k = \left\{ \sum_{i=1}^{n} \frac{\delta_{ki}^{*}}{\hat{K}(T_{ki}^{*}|V_i, W_i)} Z_i^{\otimes 2} \right\}^{-i} \sum_{i=1}^{n} \frac{\delta_{ki}^{*}}{\hat{K}(T_{ki}^{*}|V_i, W_i)} M_{ki} Z_i$$

with $\hat{\beta} = \sum_{k=1}^{K} \hat{\beta}_k$, that is,

$$\hat{\beta} = \sum_{k=1}^{K} \left[\left\{ \sum_{i=1}^{n} \frac{\delta_{ki}^{*}}{\hat{K}(T_{ki}^{*}|V_i, W_i)} Z_i^{\otimes 2} \right\}^{-1} \sum_{i=1}^{n} \frac{\delta_{ki}^{*}}{\hat{K}(T_{ki}^{*}|V_i, W_i)} M_{ki} Z_i \right] \tag{5.46}$$

The asymptotic properties for both these estimators and the expressions for the limiting covariance matrices reported by Lin (2000) are derived adopting the same analytical framework as for the case of covariate independent censoring and follow as a direct generalisation of the results presented above for the covariate independent censoring case. The main reason for developing all the approaches presented in this section is to allow incorporation of a number of discrete and continuous covariates in modelling costs under conditions of censoring. In addition, the methods are not restricted by the censoring pattern or by the number of covariates.

5.2.6. Two-stage regression (Carides et al methodology)

Carides et al (2000) proposed a parametric estimator for mean cost in which the total cumulative cost is modelled as a function of failure time. Their method was introduced as an attempt to overcome the limitation of the second Lin et al (1997) non-parametric approach, presented in the previous chapter, imposed by the requirement of a discrete censoring pattern to ensure the estimator's consistency. Their estimator is referred to as a two-stage estimator because in the first stage of the estimation process the expected cost at any given point in time is estimated as a function of failure time and in the second stage the estimated expected costs at given points in time are weighted by the Kaplan–Meier probability of death at these points in time. The estimate of mean total cost is then derived as the sum over time of these weighted individual cost estimates. Under this model the mean cost is therefore given by

$$\mu = \int_0^{\infty} g(t)|dS(t)|$$

where $g(t) = E(M|T = t)$ is the expected cost of an individual with survival time T and $S(t) = pr(T \geq t)$. The first stage involves deriving an estimator $\hat{g}(t)$ for $g(t) = E(M|T = t)$ using a regression approach. The authors suggest that the regression be performed only on the uncensored observations on the basis that the treatment costs of censored individuals typically differ from the treatment costs of uncensored individuals at the same point in time and inclusion of censored observations will therefore impart bias into the estimate of $g(t)$. The second stage of the estimation process involves the weighting of the estimated regression function $\hat{g}(t)$ by the Kaplan-Meier estimate of the probability of death at time *t*. The two-stage estimator of the mean cost over $(0, L]$ is then given as

$$\hat{\mu}_{TS} = \int_0^{L} \hat{g}(t)\left|d\hat{S}(t)\right| \tag{5.47}$$

where $\hat{g}(t)$ is an estimator for $g(t) = E(M|T = t)$, $\hat{S}(t) = \prod_{s \leq t}\left\{1 - \frac{\Delta N(s)}{Y(s)}\right\}$, that is, $\hat{S}(t)$ is the Kaplan-Meier estimator for $S(t) = pr(T \geq t)$. If the last observed time corresponds to censoring in which case the Kaplan-Meier estimator is undefined (see section 4.3.1), to ensure consistency the estimator can be expressed as

$$\hat{\mu}_{TS} = \int_0^{L} \hat{g}(t)\left|d\hat{S}(t)\right| + \overline{M}_{u \geq L}\hat{S}(L)$$

where $\overline{M}_{u \geq L}$ is an estimate of cost accumulated over $(0, L]$ for patients who survive beyond *L*. The choice of the functional form for $g(t)$ depends mainly on the data under consideration and the authors suggest use of either a parametric regression model or a non-parametric smoother. In the case of a parametric regression model the authors consider models which are, with or without some transformation of the data, linear in the coefficients thus allowing use of the ordinary least squares regression technique to derive estimates for the regression parameters.

If the parametric model for $g(t)$ is a monotonic non-decreasing function of failure time the two-stage estimator can be given as

$$\hat{\mu}_{TS} = \int_0^{\hat{g}(L)} \hat{S}_{g(T)}\{\hat{g}(t)\}d\hat{g}(t) \tag{5.48}$$

where $\hat{S}_{g(T)}\{\hat{g}(t)\}$ is the Kaplan-Meier estimator for $S_{g(T)}\{\hat{g}(t)\} = pr\{\hat{g}(t) \geq c\}$, that is, the Kaplan-Meier estimator for the probability of the expected cost, as estimated by the model, being at least *c*.

Following the definition of the Kaplan-Meier estimator for cost given in section 4.3.1 of the previous chapter, $\hat{S}_{g(T)}\{\hat{g}(t)\}$ is given by

$$\hat{S}_{g(T)}\{\hat{g}(t)\} = \prod_{\hat{g}(u)\leq\hat{g}(t)} \left\{1 - \frac{\Delta N\{\hat{g}(u)\}}{Y\{\hat{g}(u)\}}\right\}$$

where $N\{\hat{g}(u)\} = \sum_{i=1}^{n} I\{\hat{g}(X_i) \leq \hat{g}(u), \delta_i = 1\}$ and $Y\{\hat{g}(u)\} = \sum_{i=1}^{n} I\{\hat{g}(X_i) \geq \hat{g}(u)\}$.

Thus, the difference between (5.47) and (5.48) is that in (5.47) the estimate for mean cost is derived as the sum of the expected costs as estimated from the model weighted by the Kaplan-Meier probability of death at the respective points in time, whereas in (5.48), the estimate for mean cost is derived as the area under the Kaplan-Meier cost curve as defined by equation (4.5) given in section 4.3.1. In other words, if the assumed parametric model is monotonically non-decreasing, a one-to-one correspondence between cost and failure time is ensured which implies that the assumption of independent censoring between cost at censoring and cost at failure time is satisfied thus allowing use of the Kaplan-Meier approach in deriving an estimate for the mean cost.

Due to the consistency of the Kaplan-Meier estimator, consistency of the two-stage estimator is ensured if the parametric model $g(t)$ is consistently estimated. Although under specific parametric assumptions the two-stage estimator is asymptotically normal with variance estimator directly following from the specific statistical distribution, the authors recommend that for practical purposes the bootstrap method be used to derive standard error estimates for the mean, as they argue that the assumption of asymptotic normality is unlikely to be valid in most applications. The issue then becomes to choose a functional form for $g(t)$. Many different alternative models could be adopted in estimating this functional form. The models considered in this analysis follow the author's suggestions and are presented below.

The first model assumes a linear relationship between total costs and failure times specified as

$$M_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

where the error terms are normally distributed with zero mean and finite variance, so that the two-stage estimator for mean cost is

$$\hat{M} = \hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_t \qquad (5.49)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are estimated from ordinary least squares regression using only the uncensored cost observations and $\hat{\mu}_t = \int_0^t \hat{S}(u)du$ is the Kaplan-Meier estimator for mean survival time over

$(0,t]$ where $\hat{S}(t) = \prod_{s \le t} \left\{ 1 - \frac{\Delta N(s)}{Y(s)} \right\}$.[7]

Given that the distribution of costs is commonly positively skewed, the authors consider a model in which the relationship between cost and survival time is specified by transforming the total costs on the natural logarithm scale. The regression model is then given as

$$\ln M_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

where the error term is assumed lognormal distributed. Ordinary least squares regression is again used on the uncensored cost observations to derive estimates for β_0 and β_1. The mean cost is given as

$$\hat{M} = e^{\hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_t} \qquad (5.50)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates from the ordinary least squares regression on the uncensored observations only and $\hat{\mu}_t$ is the Kaplan-Meier estimate for mean survival time.

Although not considered by Carides et al (2000), under specifications where there has been some transformation of the dependent variable estimation of the untransformed scale expectation requires that retransformation to the untransformed scale be performed. As mentioned in chapter 2 a number of investigators including Duan (1983) have argued that when retransformation of the dependent variable to the untransformed scale takes place, an incorrect normality assumption with regards to the error distribution will lead to inconsistent estimates for the untransformed scale expectation although the ordinary least squares estimate for the regression coefficients is consistent and unbiased with minimum variance regardless of whether the true error distribution is normal or not. Given therefore that the error distribution in the untransformed scale is unknown, Duan (1983)

---

[7] Although not shown by the authors this expression for the estimator of mean cost can be derived as follows. The two-stage estimator for the unrestricted mean is $\hat{\mu}_{TS} = \int_0^\infty \hat{g}(t) \left| d\hat{S}(t) \right|$. As shown in section 4.2.1,

$$\lambda(t) = -\frac{d \ln S(t)}{dt} \Rightarrow S(t) = e^{-\int \lambda(t) dt} \Rightarrow \frac{dS(t)}{dt} = [-\lambda(t)] e^{-\int \lambda(t) dt} \Rightarrow \frac{dS(t)}{dt} = [-\lambda(t)] S(t)$$

and because $\lambda(t) = \frac{f(t)}{S(t)}$, it follows that $\frac{dS(t)}{dt} = \left(-\frac{f(t)}{S(t)} \right) S(t) \Rightarrow |dS(t)| = f(t)dt$.

Assuming $g(t) = \beta_0 + \beta_1 t$, the two-stage estimator is

$$\mu_{TS} = \int_0^\infty g(t) |dS(t)| = \int_0^\infty \beta_0 |dS(t)| + \int_0^\infty \beta_1 t |dS(t)| = \beta_0 \int_0^\infty f(t)dt + \beta_1 \int_0^\infty t f(t)dt = \beta_0 + \beta_1 \mu_t$$

Replacing the unknown parameters with their sample estimators results in the estimator for mean cost as given in (5.49).

suggested a non-parametric estimator for the untransformed scale expectation referred to as the smearing estimator derived as shown below.

Denoting the observations for the dependent variable on the untransformed scale by $Y_i, i = 1,...,n$ and letting Z be a covariate of interest, a linear regression model relating the dependent variable on the transformed scale to the covariate is given by

$$\varphi(Y_i) = \beta Z_i + \varepsilon_i$$

where $\eta_i = \varphi(Y_i)$ are the observations on the transformed scale and $Y_i = h(\eta_i)$ denotes the observations after the retransformation to the untransformed scale has taken place, that is, $h = \varphi^{-1}$ with φ and h being monotonic and continuously differentiable. For the model $\eta_i = \beta Z_i + \varepsilon_i$ on the transformed scale it is also assumed that the error terms have zero mean and constant variance without being necessarily normally distributed. The smearing estimate for the untransformed scale expectation, that is after the retransformation, is then given by

$$\hat{E}(Y_i) = \frac{1}{n}\sum_{i=1}^{n} h(\hat{\beta} Z_i + \hat{\varepsilon}_i)$$

where $\hat{\beta}$ is the ordinary least squares regression estimate on the transformed scale, that is $\hat{\beta} = (Z'Z)^{-1}Z'\eta$ and $\hat{\varepsilon}_i = \eta_i - \hat{\beta} Z_i$ are the estimated ordinary least squares residuals. This estimator is shown to be consistent even when the error distribution in the above model is normal.

The estimate for the mean cost from the model $\ln M_i = \beta_0 + \beta_1 T_i + \varepsilon_i$ after smearing is therefore

$$\hat{M} = e^{\hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_t} \frac{1}{n}\sum_{i=1}^{n} e^{\hat{\varepsilon}_i} \qquad (5.51)$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates from the ordinary least squares regression on the uncensored observations only, $\hat{\mu}_t$ is the Kaplan-Meier estimate for mean survival time and $\hat{\varepsilon}_i$ are the ordinary least squares residuals.

Following Carides et al (2000), a third parameterisation for the relationship between cost and survival time was obtained by transforming the natural logarithm of the total costs on the log scale once more. The regression model is then given as

$$\ln(\ln M_i) = \beta_0 + \beta_1 \ln T_i + \varepsilon_i$$

Ordinary least squares regression is again used on the uncensored cost observations to derive estimates for β_0 and β_1. The mean cost without smearing is given as

$$\hat{M} = e^{e^{\hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_t}} \tag{5.52}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates from the ordinary least squares regression on the uncensored observations only and $\hat{\mu}_t$ is the Kaplan-Meier estimate for mean survival time. The estimator for mean cost with smearing is

$$\hat{M} = \frac{1}{n} \sum_{i=1}^{n} e^{e^{(\hat{\beta}_0 + \hat{\beta}_1 \hat{\mu}_t)} e^{\hat{\varepsilon}_i}} \tag{5.53}$$

The authors conclude that such a regression based approach where the relationship between cost and failure time is specified through a parametric model is advantageous compared to a non-parametric approach due to efficiency gains resulting from the use of such a relationship. On the other hand, this is only going to be the case if the parameterisation reflects the true functional form of cost and failure time. In the event of model misspecification, a non-parametric approach for estimating the relationship between cost and failure time will be preferred. Under either case however the degree of censoring in the data is expected to have a direct impact on the estimates of mean cost and the methodology presented here cannot incorporate information on individual cost histories which as shown by the analysis in the previous chapter could be of significant advantage under extreme censoring conditions.

5.3. Methods and results

5.3.1. Methods

The parametric models discussed above were applied to the UKPDS trial data as the non-parametric models of the previous chapter. All estimates were derived separately for each randomisation group. The covariates of interest were time independent and represented measurements obtained on each individual at the start of the study on age, body mass index (bmi), fasting plasma glucose level (fpg), race and sex. The descriptive statistics for each of the covariates are shown in Table 5.1.

Table 5.1. Baseline covariate values in conventional and intensive policy groups

	Mean	*Standard deviation*	*Minimum*	*Maximum*
Conventional (n=1138)				
Age (years)	53.40	8.69	25.62	72
Bmi (kg/m^2)	27.80	5.46	17.57	55.68
Fpg (mmol/l)	8.48	2.03	5.5	17.5
Race*	1.32	0.72	1	5
Sex**	1.38	0.49	1	2
Intensive (n=2729)				
Age (years)	53.21	8.62	24.69	72
Bmi (kg/m^2)	27.49	5.07	16.59	60.61
Fpg (mmol/l)	8.61	2.14	5.4	19.9
Race*	1.31	0.70	1	5
Sex**	1.39	0.49	1	2

Frequency of categorical covariates

	Conventional	***Intensive***
Race*		
1 = Caucasian	927	2215
2 = Afro-Caribbean	76	216
3 = Indian Asian	126	276
4 = Other	2	8
5 = Other	7	14
Sex**		
1 = male	701	1664
2 = female	437	1065

As can be seen from Table 5.1 there are no differences in the baseline covariate values between the two groups. These covariates were deemed clinically meaningful given that the fasting plasma glucose level provides the means of defining diabetes and body mass index gives an indication of obesity which is highly positively correlated with the risk of diabetes as is age. There is also evidence of racial differences in the incidence and prevalence of diabetes with for example higher rates in the Asian population. The fact that these covariates were deemed important explanatory variables for diabetes progression and complications does not imply that they will necessarily explain cost especially as they were only measured at the start of the study. However this represents the most frequent pattern of covariate measurements within a clinical trial setting where interest lies

in recording disease specific predictive factors at the time of the individual's entry to the study and in certain cases at various points in time over the follow-up period. Before presenting the results derived from the various estimators some specific methodological points follow.

With respect to the semiparametric methodology it is clear from the arguments presented earlier that the proportional hazards model will generally return biased estimates of cost for a number of reasons. Under the strong assumption that adopting the stratified variant of the model with time as the stratification variable overcomes the bias arising from dependent censoring between cost at failure and cost at censoring time, the issue then becomes to determine whether the assumption of proportional hazards holds within each individual stratum. Although based on the conceptual arguments given above this is not likely to be the case, the proportionality assumption was empirically assessed for the stratified model using the Gramsch and Therneau approach given on pages 112-113. The results of these tests are presented in Appendix A.5.1 and as expected they show violation of the proportional hazards assumption across all strata. Consequently the proportional hazards approach is not pursued further. The proportional means model suggested by Lin (2000) is not pursued either primarily because as stated above a consistent estimator of average cost over the study period requires derivation of a consistent estimator for the baseline mean hazard function. The latter can only be derived if all sample paths of the cost process are known which would require knowledge of the amount of cost accrual for all individuals at every point in time over the duration of interest. Given that this information was not available it was not possible to derive an estimator of mean cost using this approach.

The Weibull and exponential regression models although known to be biased were estimated for comparison purposes to the Kaplan-Meier estimator, the rationale being that they all share the same source of bias but they differ in the underlying distributional assumptions about cost, with the first two models imposing specific cost distributions and the latter being free of such assumptions. Estimation was based on equations (5.12), (5.13) and (5.14) for the Weibull and (5.15), (5.16) and (5.17) for the exponential regressions.

The three models proposed by Carides et al (2000) presented in section 5.2.6 were applied in an attempt to estimate cost conditional on failure time. All these models involve the Kaplan-Meier estimate of mean survival time. This was estimated using equations (4.1), (4.2) and (4.3) as 15.65 years (se=0.21) for the conventional group and 15.96 (se=0.18) for the intensive group. Although the authors do not consider smearing estimators when the outcome variable was transformed, the analysis undertaken here derived mean estimates with and without smearing. The mean cost estimates were based on (5.49) for the untransformed model, on (5.50) and (5.51) for the first transformation and on (5.52) and (5.53) for the second transformation. An indication of the underlying relationship between treatment cost and study time is given by Figure 4.1 (section 4.4.1) which plots the observed costs against time for the UKPDS censored and uncensored populations for both trial arms.

The regression methodology proposed by Lin (2000) was applied to the data both when individual costs were available at the last contact date or death and when multiple observations at different points in time were available for each individual. In the second model annual time intervals were assumed for each individual because as stated in the previous chapter intermediate cost history was available for each subject on an annual basis. The approach relating to covariate dependent censoring was not explored as this was not applicable given the data. Under the first model estimation was based on equations (5.25) and (5.27) for the regression parameters and on (5.29) to (5.32) for the coefficient standard errors. Under the second approach the respective estimating equations are given by (5.33) and (5.35) for the coefficients and by (5.36) to (5.41) for their standard errors.

The classical linear regression model given by (5.18) was estimated using the uncensored cases only as a baseline means for comparison to the alternative linear regression methodologies. All regression models aside from those proposed by Carides et al which used failure time as the independent variable, were based on the set of covariates described above.

Estimates of the variance associated with the mean estimators resulting from the above models were derived using the bootstrap approach with the exception of the Weibull and exponential regressions where the variance estimators were derived using equations (5.14) and (5.17) respectively. For the remaining models the bootstrap estimates were obtained from 1000 replications. The reason for using the bootstrap approach is that the asymptotic variance estimators for the mean cost have not been defined. With respect to the regression parameter standard errors for the Lin regression models, these were derived both using the bootstrap approach and analytically using equations (5.29) to (5.32) for the first approach and equations (5.36) to (5.41) for the second as described above.

5.3.2. Results

The results derived from the parametric approaches are shown in Table 5.2 while Table 5.3 reports the best non-parametric estimates (with the standard errors as derived from the asymptotic variance estimators) from the previous chapter as a means of assessing the parametric estimators' performance. Based on the conclusions drawn in the previous chapter the first approach by Lin et al (1997), using information on intermediate cost histories, and the Bang and Tsiatis partitioned estimator were deemed to perform adequately under all circumstances considered. Given that these two estimators remained stable even under the extreme censoring conditions arising in the UKPDS data, it can be reasonably confidently asserted that the resultant cost estimates are reflecting the true cost values.

Table 5.2. Parametric estimators of mean cost

Estimator	Conventional		Intensive	
	Mean	*Standard error**	*Mean*	*Standard error**
Weibull regression	43090.52	37577.80	36502.85	28279.72
Exponential regression	54151.55	54151.55	56793.82	56793.82
Carides et al regression models				
total cost on time	20353.71	2551.99	19548.07	1228.00
ln(total cost) on time without smearing	18086.73	2599.06	21096	1927.38
ln(total cost) on time with smearing	16070.78	1914.10	17939.50	1368.74
ln(ln(total cost)) on time without smearing	19080.38	3155.30	23132.24	2680.18
ln(ln(total cost) on time with smearing	18959.67	3152.28	21626.47	2545.91
Lin regression methodology				
Complete costs	14015.82	3588.94	17573.79	1961.70
Multiple intervals	14941.14	1274.07	13789.33	452.70
Naïve OLS	11708.78	1268.10	10845.21	693.58

*The standard errors were estimated using the bootstrap (1000 replications) with the exception of the Weibull and exponential regression models were the analytic formulae were used.

Table 5.3. Best non-parametric estimators of mean cost

Estimator	Conventional		Intensive	
	Mean	*Standard error*	*Mean*	*Standard error*
Lin1 (Lin et al 1997) ($\hat{\mu}_{LIN1}$)	14006.2	897.73	13172	340.55
Bang and Tsiatis Partitioned ($\hat{\mu}_P$)	14639.48	1219.4	13839.67	445.6

Reported in Table 4.4

With respect to the Weibull and exponential regression models the resultant estimates are high for both groups confirming the estimators' bias as expected given the arguments above (estimated regression coefficients shown in Appendix A.5.2). Comparison with the Kaplan Meier estimates of 38770.74 (se=5312.02) for the conventional arm and 31620.59 (se=2034.89) for the intensive arm reveals that the bias resulting from these parametric regressions is even greater than that of the non-parametric approach. The most likely explanation is that the bias in the parametric models can arise from both violation of the independent censoring assumption as well as model misspecification. These empirical results confirm once more the predictions of the theory that standard survival analysis techniques requiring independence between the variable of interest and its censoring variable are not suitable for the analysis of censored cost data and at the same time highlight the effect of inappropriate parameterisation.

Turning to the Carides et al two-stage estimator, the resultant mean cost estimates are high relative to the non-parametric estimates for both groups (estimated regression coefficients for the various models and the programs for the bootstrap estimates for the standard error of the mean are shown in Appendix A.5.3). Moreover the difference in average cost between the treatment arms is generally of the wrong expected direction. Although the approach has initial appeal given that it attempts to model the time pattern of costs and is not restricted by assumptions concerning the censoring distribution, the analysis reveals the estimator's inadequate performance in all the parameterisations considered. This finding holds even when smearing estimates were obtained following a logarithmic transformation to account for positive skewness in the cost data. While model misspecification is liable to be a contributory factor, the estimators' inadequate performance is also likely due to the high degree of censoring present in the data. As the regression parameters are estimated using information from the uncensored cases alone, which in this case amounts to a mere 18% of the total number of observations and will reflect the bias imparted from a complete case analysis, it is to be expected that the estimated coefficients will not reflect the true parameter values even assuming the relationship between cost and failure time is correctly specified.

This postulate is supported further by the results obtained when the expected costs were estimated by a non-parametric regression approach. Carides et al recommend use of such a regression when there is not enough confidence in a specific parametric relationship between cost and survival time. The method adopted provides smoothed estimates of cost using locally weighted scatterplot smoothing (lowess) according to which the smoothed values of the dependent variable are derived by running a regression of the dependent variable on the independent variable using for each estimate the data at the estimation point and a small amount of data near the point. In lowess the regression is weighted so that the central point each time receives the highest weight and points farther away receive less. A separate weighted regression is estimated for each point in the data in order to provide the smoothed estimates. Applying this approach resulted in estimates of mean cost of 5674.92 (se=853.24) for the conventional group and 9407.87 (se= 3230.63) for the intensive group where the standard errors were obtained from 1000 bootstrap replications. Such an approach for deriving expected cost estimates, being free of assumptions about the functional form between cost and failure time gives a strong indication that an equally important, if not more important, source of bias aside from model misspecification in the Carides et al estimator is the high level of censoring. This was to be expected based on the results obtained from the non-parametric estimators of the previous chapter which only used cost information from the complete cases. Both the respective Lin et al and Bang and Tsiatis non-parametric estimators performed inadequately when only complete costs were included in the estimation process and both techniques showed dramatic improvement when information was increased by incorporating individual cost histories into the estimating equations.

Before considering the set of parametric estimators proposed by Lin (2000) the estimates derived from the naïve ordinary least squares regression are considered.[8] The estimates derived from this approach are known to be biased as they are based on a complete cases analysis which ignores all censored observations, but as stated above, they provide a means for baseline comparisons to the alternative linear regression methodologies and in particular to Lin's (2000) regression models which use the same set of covariates. Although the naïve least squares regression resulted in the expected direction of the difference between the two arms of the trial with the conventional group incurring higher costs on average than the intensive, the estimates of mean cost are low. This was anticipated as the information from censored observations is not used in the estimation process and it is known that the bias increases as the level of censoring increases. Comparison of the ordinary least squares cost estimates with the non-parametric uncensored cases estimates reported in chapter 4 – which were 11901.01(se=1061.36) and 10629.97 (se=510.00) for the conventional and intensive arms respectively – reveals a close similarity. This may indicate that model parameterisation does not provide additional information with respect to the distribution of cost and results in the same degree of bias as that imparted by censoring in the non-parametric naïve estimator.

With respect to Lin's (2000) parametric approach that uses information only on the complete total costs, the resultant difference in mean cost between the trial arms is of the wrong expected direction. In addition the estimated mean cost for the intensive group is much higher than expected.[9] This pattern alters when the regression uses information on multiple cost observations on each patient obtained at a number of points in time.[10] The latter approach results in estimates that are very close to the non-parametric counterparts derived from the first Lin non-parametric method using individual cost histories and even closer to the Bang and Tsiatis partitioned estimator which again uses individual cost histories. Thus the regression model which uses cost history information from all individuals results in a significant improvement compared to the parametric model which discards cost information from the censored cases. This was anticipated and confirms Lin's argument that the multiple time intervals approach improves efficiency by using information which is ignored by the complete costs approach. However the reason why the second regression methodology performs adequately does not appear to be related to model specification as the estimates of the regression parameters reported in Table 5.4 indicate.

[8] Appendix A.5.4. presents the resultant coefficients and the programs for estimating the standard error of the mean using the bootstrap.

[9] Appendix A.5.5. presents the programs for estimating the coefficients and their standard errors using the analytically derived formulae as well as the program for deriving the bootstrap estimates of the standard errors for the coefficients and the mean.

[10] Appendix A.5.6. presents the programs for estimating the coefficients and their standard errors using the analytically derived formulae as well as the program for deriving the bootstrap estimates of the standard errors for the coefficients and the mean.

Table 5.4. Estimated regression parameters for the naïve OLS and the Lin regression models

	Conventional			*Intensive*		
	Regression coefficients	Standard error**1	Standard error**2	Regression coefficients	Standard error**1	Standard error**2
Naïve OLS						
const	-23980.37	16190.03	13813.83	5647.46	7148.85	8007.90
age	262.47	197.58	161.68	57.55	85.66	100.61
bmi	454.01	270.10	258.15	34.84	123.74	108.39
fpg	537.43	611.44	654.39	-176.55	267.84	231.98
race	1783.51	2204.65	1990.84	146.71	954.35	998.54
sex	1545.94	2646.61	2627.59	1802.75	1250.54	1318.99
Lin complete costs						
const	-21043.55	223315.92	25522.02	32901.42	370405.18	24882.84
age	141.61	2388.67	337.58	-211.60	3100.51	315.09
bmi	596.27	4586.22	610.46	208.61	5442.36	267.38
fpg	1099.66	10597.17	1012.11	-979.99	12255.61	753.35
race	1424.06	44620.62	4309.11	1740.02	49061.19	2739.86
sex	-206.40	41170.38	8907.59	-2619.76	66535.70	4142.53
Lin multiple intervals						
const	-217.49	260748.92	8723.92	12170.94**	256027.34	4883.95
age	-16.99	3653.66	141.88	1.86	1830.22	41.83
bmi	127.08	6275.02	211.20	23.84	4818.33	108.31
fpg	1493.37**	13190.91	634.84	148.15	9463.72	187.03
race	-247.16	45221.68	1711.94	-863.83	27533.62	630.64
sex	139.01	109954.54	3271.34	517.88	38910.15	806.28

**1 Standard errors analytically derived by the normal equations for OLS and by expressions (5.29), (5.30), (5.31) for Lin using complete costs and (5.32) (5.36), (5.37), (5.38), (5.39), (5.40) and (5.41) for Lin using multiple intervals
**2 Standard errors derived from 1000 bootstrap replications
**significant

The coefficient estimates resulting from all these regressions indicate that the covariates have low explanatory power. With respect to Lin's regressions all are insignificant in the complete costs approach and significant only for fasting plasma glucose in the conventional group in the multiple time intervals approach. In the case of the naïve ordinary least squares regression all coefficients are insignificant. In other words, the multiple time intervals approach does not exhibit a significantly improved model specification compared to the complete costs regression. Nevertheless the mean cost estimates derived from the multiple time intervals regression model are very close to the comparative non-parametric estimates. As both regressions use an inverse probability weight in attempting to account for censoring, the most likely explanation for this result is therefore the increased cost information used in conjunction with the particular weight by the multiple intervals regression. This appears to be confirmed by the results of a secondary analysis which was undertaken for the Lin regression models where only fasting plasma glucose was used as a covariate. Fasting plasma glucose was chosen on the basis that this was the only covariate

associated with a significant coefficient even though this was only the case for one of the regression models. The results together with the naïve ordinary least squares estimates are shown in Table 5.5 for the regression parameters and in Table 5.6 for the mean costs.

Table 5.5. Estimated regression parameters for the naïve OLS and the Lin regression models using fasting plasma glucose as the only covariate

	Conventional			*Intensive*		
	Regression coefficients	Standard error[**1]	Standard error[**2]	Regression coefficients	Standard error[**1]	Standard error[**2]
Naïve OLS						
const	7367.78	5369.69	5391.06	12142.42**	2403.76	2171.18
fpg	602.40	603.33	624.86	-144.15	262.42	232.21
Lin complete costs						
const	-466.98	85058.80	9367.33	24497.55**	142302.49	8912.96
fpg	1690.44	10737.78	968.63	-891.05	12645.16	996.83
Lin multiple intervals						
const	4263.38	97544.86	3963.33	13220.31**	96850.40	1948.28
fpg	1270.71**	12934.45	575.85	99.13	9973.46	203.60

[**1] Standard errors analytically derived by the normal equations for OLS and by expressions (5.29), (5.30), (5.31) for Lin using complete costs and (5.32) (5.36), (5.37), (5.38), (5.39), (5.40) and (5.41) for Lin using multiple intervals
[**2] Standard errors derived from 1000 bootstrap replications
**significant

Table 5.6. Estimated mean costs from regression models using fasting plasma glucose as the only covariate

	Conventional		**Intensive**	
Estimator	*Mean*	*Standard error**	*Mean*	*Standard error**
Lin complete costs	13870.84	7060.85	16821.34	2419.14
Lin multiple intervals	15041.21	1578.42	14074.33	454.88
Naïve OLS	12477.19	1212.05	10900.58	560.30

*The standard errors were estimated using the bootstrap (1000 replications)

Although the coefficient on fasting plasma glucose did not become significant in any other model, the mean estimates are very similar to their respective counterparts derived in the analysis based on the complete set of covariates. In this particular application therefore the choice of the set of covariates does not appear to have an impact on the resultant mean cost estimates. The inverse of the probability of an individual not being censored entering the estimating equations seems to be primarily responsible for the resultant mean estimates. However this particular weight alone is incapable of adequately adjusting for the loss in information when the level of censoring is too high

as indicated by the poor performance of the complete costs regression. As was the case in the non-parametric analysis, the amount of available information on the cost history process proves as important as the probability weight which adjusts the estimates for the information loss due to censoring.

The fact that there was no gain associated with incorporating covariate information in the estimation process should not be interpreted as a general criticism of the particular regression methodology. It is rather the case that the covariates available in the dataset used in the present study did not provide any additional information with respect to cost and as such the approach did not provide further insight into the distribution of costs when compared to the non-parametric approaches of the previous chapter.

5.4. Discussion

Parametric approaches provide a necessary alternative in deriving estimates of cost statistics in a number of circumstances, such as when interest lies in the assessment of individual covariate effects on cost or in extrapolation of estimates beyond the observed study duration or to different patient subpopulations. Inherent in all parametric approaches is the specification of a functional form for the relationship between the outcome variable and the covariates of interest. When the outcome variable is subject to censoring classical least squares estimation is biased and alternative models have been proposed, a number of which make specific distributional assumptions whereas others are free of assumptions regarding the underlying distribution of cost. This chapter considered a number of parametric methodologies which attempt to account for the presence of censoring within the context of cost analysis. The performance of the proposed estimators of cost was assessed under extreme censoring conditions using the same trial data as were used in the non-parametric analysis and the main findings are as follows.

The standard parametric techniques for analysing censored failure time data such as the Weibull and exponential regression are inappropriate for the analysis of censored cost data due to dependent censoring between cost at event and cost at censoring. The semiparametric proportional hazards approach although it could be potentially better suited for modelling complex distributions as it allows the functional form of part of the model to be unknown and therefore unrestricted, is also subject to the bias induced by dependent censoring. Ordinary least squares regression based on the complete cases alone is biased with the degree of bias increasing as censoring increases. The regression methodology proposed by Carides et al (2000) which models cost as a function of failure time is also sensitive to the level of censoring because although the weight providing adjustment for censoring is consistently estimated, the cost estimates being adjusted by it are based on regression parameter estimates which are in turn obtained using information on complete cases alone. Consequently, aside from a potential misspecification of the functional form for the relationship

between cost and failure time, bias in the estimates is also going to arise due to the bias attributed to the complete cases analysis involved in the estimation process.

The regression methodology proposed by Lin (2000) adjusts the estimates for censoring through weighting the cost observations by the inverse of the probability of an individual not being censored as was the case in the Bang and Tsiatis non-parametric approach. Of the two models constituting the regression methodology, the one using cost information from the uncensored individuals alone was shown to result in biased estimates at the levels of censoring present in the particular data, while the second using information on individual cost histories from all individuals in the study was shown to perform adequately in deriving estimates of mean cost. The observed performance patterns however appeared to be unrelated to model specification indicating that in this particular application incorporation of covariate information did not improve upon the cost estimates derived from the best performing non-parametric estimators.

Based on these findings, it might be concluded that for the purposes of the analysis undertaken here whose aim was to assess the various estimators' performance when estimates of mean cost over the study period are sought, there is no gain in adopting a parametric approach over a consistent non-parametric estimator. Moreover a non-parametric methodology is usually preferable on the basis that it involves fewer assumptions compared to a parametric alternative. In situations however where interest lies in assessing covariate effects on cost and in extrapolation beyond the study duration or to a different population setting, a parametric methodology becomes the necessary alternative. Of the proposed methodologies considered here, whose assessment was undertaken under extreme censoring conditions, the multiple time intervals regression proposed by Lin (2000) performed adequately in providing estimates of mean cost compared to the best non-parametric estimators even though there was no gain associated from incorporating covariate information.

This finding aside from providing a regression methodology that performs well under extreme censoring conditions, also confirms the general result of the previous chapter by reaching the same conclusion from a different analytical perspective. As was the case when the non-parametric estimators were considered, the present analysis established that censoring in the cost estimates is most successfully accounted for through weighting the complete observations by the inverse of the probability of non-missingness although the degree of retrieval of information lost due to censoring will also be determined by the amount of available information on the cost history process.

Chapter 6

Conclusions

This thesis has been set against a background of increasing interest in economic evaluation of health care technologies and a related gradual expansion in economic analysis conducted alongside clinical trials. Despite the difficulties associated with its theoretical justification and implementation on occasions, this form of evaluation has been increasingly gaining acceptance as a useful means of assisting the decision-making process in the choice concerning the allocation of resources among competing health care interventions. It is implicit in adopting this method that efficiency in the health care sector is an underlying objective. Although alternative definitions of efficiency lead to different analytical perspectives, the objective of welfare maximisation is nevertheless maintained. A prerequisite condition in pursuing this objective is to ensure that the outcome of such an analysis truly represents a relative valuation of the alternative resource allocation states under consideration. Fulfilment of this condition in turn requires that the methodology employed in deriving relative valuations is theoretically justified and that the measures it incorporates in deriving these valuations are appropriately specified and quantified. Assessment of alternative health care resource allocation patterns is then undertaken through the evaluation of the respective competing health care interventions by investigating the intervention specific resource costs incurred in achieving a given health outcome. On this basis the importance of deriving appropriate and accurate valuations of both cost and outcome is self-evident. It is within this context that this thesis has addressed specific problems relating to the collection and analysis of treatment cost data. Concentration on this particular subject has been motivated both by the limited consideration of the measurement of direct treatment costs relative to other areas of economic evaluation as well as by the recognition that the adoption of statistical methodology within the analysis of cost data is necessary especially in situations where particular data problems arise.

Having established the general setting, the analysis was preceded by an overview of the existing literature which served the purpose of identifying commonly encountered measurement problems relating to cost data, indicating their importance and revealing the current state of the development of solutions to these problems. Of the identified problems, the limited availability of cost data due to the data collection process and the incompleteness of cost information for analysis due to censoring were considered in detail in subsequent parts of the thesis with censoring constituting the major issue of concern.

Viewed within the context of a clinical trial setting the problem of limited availability of cost data due to the collection process arises mainly because of data constraints imposed on the economic

variables by the trial design given that questions regarding the cost-effectiveness of a given intervention are normally of secondary importance relative to the testing of the clinical hypothesis concerning treatment efficacy that the trial considers. As a consequence of these priorities the trial typically records information prospectively on resource use at the patient level and leaves the unit costs of the resources to be determined retrospectively. Within a multi-centre trial setting this implies that centre specific unit cost information is not normally collected alongside the trial and some alternative source is thus required to provide the necessary unit cost information in calculating the total treatment cost. This alternative information source usually provides an average unit cost for each resource element. Deriving an estimate of treatment cost by combining centre specific resource volumes with an average unit cost estimate for each resource component results in the calculated treatment costs encompassing the variation in resource use across the participating centres but not the variation in the unit costs of the resources. The question of whether this matters was addressed in this thesis by a simulation experiment which assessed the difference in the estimates of treatment cost between two alternative estimation methods, the first using an average unit cost for each resource component and the second using centre specific unit cost information. The alternative approaches were considered within the context of economic theory assuming an underlying production function in specifying the relationship between inputs and health outcome while the relationship with the cost of producing a given health outcome was investigated by considering two distinct scenarios. The first assumed that treatment centres operate as dictated by economic theory and upheld by economic evaluation and therefore respond to changes in the relative prices of inputs, resulting for instance from the introduction of a new health care technology, by substituting the relatively less expensive inputs for the more expensive ones in a predictable manner, whereas the second assumed that treatment centres operate on the production function but do not respond to changes in relative input prices.

The analysis shows that if treatment centres respond to unit cost changes as expected from the theory, then the difference in the estimates resulting from the two methods of cost calculation is statistically significant and this result holds for a wide range of values of the elasticity of substitution representing conditions of near perfect substitutability to near perfect complementarity of the inputs entering the production process. If on the other hand treatment centres are not responsive to changes in input unit costs, the differences in the resultant estimates are not statistically significant. These results held when input unit costs were drawn from a number of alternative statistical distributions and under circumstances where the response to relative changes of input unit costs was assumed to have a stochastic component. The implication of this finding is that under the assumption that treatment centres operate in a rational economic manner in producing a given level of health outcome, as assessed by the study of a substitution effect on the production process, the method of cost data collection has an impact on the estimates of treatment cost. In these circumstances, everything else being equal, lack of centre specific unit cost information will lead to biased cost estimates because potential substitution effects on the production process are completely ignored.

The subsequent analysis concentrated on the issue of primary concern in the thesis, namely the derivation of unbiased and consistent estimates of cost statistics when the data are subject to censoring. The objective was to investigate the theoretical justification, underlying assumptions, statistical properties and empirical performance of a number of alternative estimators of cost which attempt to account for the loss of information imparted by censoring. A distinction was made between non-parametric and parametric approaches reflecting the difference in the assumptions underlying the two groups of methodologies. Regardless of whether the estimators are of a parametric or a non-parametric nature, the majority are directly related to or originate from the theory underlying the statistical analysis of time to event data under conditions of censoring. Investigation of the theoretical properties of the estimators of cost has been undertaken using the theory of stochastic processes as applied to the study of time to event data. This analytical approach allows the notion of the time element in the cost observations to be captured, censoring to be incorporated, variance estimators to be derived and convergence and asymptotic normality of the statistics of interest to be proven by invoking martingale convergence theorems.

Given the existence of consistent estimators of failure time statistics in the presence of censoring, initial attempts to adjust estimates of cost statistics for censoring were based on application of traditional survival analysis techniques to cost data. The assumption underlying the validity of these approaches is the one of independence between the variable of interest and its censoring variable. This implies independence between time to event and time to censoring when failure time data are considered and independence between cost at event and cost at censoring when cost data are analysed. In the former case the assumption is valid under the random censoring mechanism but in the latter case it is normally violated due to the lack of a common rate of cost accrual over time among individuals, as patients who are in poorer health states generate higher costs per unit of time and consequently are expected to generate higher cumulative costs at both the failure time and the censoring time. This positive correlation implies that removal of certain observations from the sample due to censoring affects the joint distribution of cost for the remaining observations in the sense that at any point in time future cost expectation is statistically altered (from what it would have been in the absence of censoring) by censoring. As a result any analysis that does not model this dependency will lead to erroneous inferences. Consequently cost estimators based on survival analysis approaches such as the non-parametric Kaplan-Meier, the semiparametric proportional hazards regression or parametric models assuming distribution families such as the Weibull and the exponential are all inappropriate.

Alternative methodologies have been recently introduced by a number of investigators both on the non-parametric and on the parametric side. Concentrating on the former, Lin et al (1997) proposed two estimators of average cost under conditions of censoring which under appropriate censoring conditions are theoretically shown to be consistent and asymptotically normal with analytically derived consistent variance estimators. More specifically, under both methodologies the study period is partitioned into a number of subintervals, an estimate of average cost in the interval is

derived and the estimator of cost for the whole duration of analysis is obtained by summing over the subintervals the interval cost estimates weighted by interval specific Kaplan-Meier probability estimates of time to event. The difference between the two alternatives is that the first only uses information on the total costs of uncensored individuals incurred up to the point of the individual's death in the estimating process with the weight being defined by an estimate of the Kaplan-Meier probability of death in each interval of the partition, while the second uses information on intermediate individual cost history from all individuals and the weight is the Kaplan-Meier probability of survival to the start of each interval of the partition. Under the assumption of independent censoring, an extension of this assumption to require that the censoring mechanism is unrelated to cost levels, continuous distribution of failure time and appropriate censoring conditions both estimators are shown to be consistent and martingale theory enables consistent asymptotic variance estimators to be derived. By appropriate censoring conditions, it is meant that the censoring distribution is of such a form that individual censoring times can be made to correspond to the boundaries of the intervals of the partition, a condition which essentially requires a discrete pattern for the censoring times if consistency is to be ensured.

By contrast, the approach proposed by Bang and Tsiatis (2000) allows arbitrary censoring distribution patterns and in addition attempts to improve efficiency of the proposed estimators by recovering information lost due to censoring through incorporation of some functionals of the cost history process in the estimating equations. The idea underlying all estimators in this class is the use of an inverse probability weight in the estimating equations through which censoring is appropriately accounted for. The first estimator uses cost information from the uncensored individuals alone while the second estimator also incorporates information on intermediate cost history from censored individuals. Under the first approach, the estimate of mean cost is derived as the average of the complete individual costs weighted by the inverse of the Kaplan-Meier probability of an individual not being censored evaluated at the point of the individual's death. Under the second approach, the duration of analysis is partitioned into a number of subintervals, the first estimator is used to derive the estimated cost incurred in each of these subintervals and the final estimate of mean cost is derived by summing over these intervals. The advantage of the latter method over the former is that an individual is considered uncensored in a given interval whenever the individual's censoring time exceeds the end of the interval. Consequently, there is an increase in the cost information being used by this estimator, as individuals who were treated as censored by the first approach not having failed by the end of the study and whose cost information was thus not used in the estimation process will be now uncensored in some of the intervals of the partition in which their costs will contribute to the estimates. Each of these estimators is accompanied by an improved alternative that attempts to increase efficiency in the estimates through use of some functional of the cost history process that allows recovery of information lost due to censoring. Under independent censoring all four estimators are shown to be consistent and variance estimators are analytically derived by invoking martingale convergence theorems. In addition to the theory of stochastic processes and martingales which provides the mathematical framework for studying the

statistical properties of all the above estimators, the study of efficiency of the improved Bang and Tsiatis estimators is also based on the general theory of semiparametric models when data are missing at random.

Although the theoretical investigation of the estimators resulted in general justification for their use in deriving estimates of medical costs in the presence of censoring, varying performance patterns emerged when these were applied to heavily censored data. Specifically, the Lin et al and the Bang and Tsiatis estimators which only use cost information from the uncensored individuals displayed poor performance at high levels of censoring. In both cases this is due to the degree of censoring which has the following consequences. In the first case the high degree of censoring restricts the number of individuals who contribute cost information to each interval of the partition resulting in an estimated average interval cost that is not representative of the true expected cost in the interval. In the second case, the heavy censoring observed towards the end of the study period results in the estimated probability of an individual not being censored reaching extremely low values. Consequently the inverse of these probability values, which enter the estimating equations as the weights attempting to account for censoring, result in extremely inflated values of weighted costs whose impact on the final cost estimate is distortionary.

In contrast, the Lin et al and the Bang and Tsiatis estimators that use intermediate cost history from all individuals in the study performed adequately under the same censoring conditions. With respect to the improved set of estimators proposed by Bang and Tsiatis, contrary to what was anticipated from the theory, they both exhibited very poor performance and were completely unstable at the levels of censoring considered. Additional analysis investigated the impact of various levels of censoring on the estimators' performance controlling for other factors and the results confirmed the above findings with the two adequately performing estimators remaining stable under all circumstances and the remaining estimators becoming increasingly unstable as censoring increased.

The idea underlying both best performing estimators is the partitioning of the study period into subintervals to allow incorporation of individual intermediate cost histories in the estimating equations, which are subsequently weighted by an estimated probability that accounts for the presence of censoring. The estimators differ both in the choice of this weight and in the interval costs that are adjusted by it. In the estimator by Lin et al the weight is the Kaplan-Meier probability of survival to the start of the interval that adjusts estimates of mean cost in the interval, whereas the Bang and Tsiatis partitioned estimator uses the inverse of the probability of an individual not being censored evaluated at a given point in time to adjust individual observed costs in the interval. On the basis that both approaches require the same amount of cost information but consistency of the second estimator is independent of the pattern of the censoring distribution, the latter estimator becomes the preferred alternative.

The analysis therefore identified estimators of mean cost whose performance is deemed satisfactory under extreme censoring conditions. Under these circumstances and given that even though such estimators are not assumption free they involve fewer assumptions compared to parametric alternatives, they are likely to be the preferred estimation technique. When interest extends however beyond the maximum time for which data is available or when questions regarding the effect of covariates on cost arise, parametric models become a necessary alternative. Clearly for such models to provide an appropriate alternative, censoring must be accounted for.

Naturally a first candidate in this category would be the classical linear regression model with cost forming the response variable but such an approach is known to yield biased estimates when the outcome variable is drawn from a censored distribution regardless of the application of interest. The naïve solution of estimating the regression parameters by completely discarding the censored cases from the estimation process is also biased with the degree of bias increasing as the proportion of censored observations increases. This together with the failure of parametric regression models traditionally used in the analysis of time to event data to account for censoring in the cost estimates due to informative censoring has led to two alternative regression methodologies within the context of parametric censored cost analysis.

The first of these methodologies introduced by Carides et al (2000) assumes a relationship between cost and failure time and involves two stages in deriving estimates of mean cost. In the first stage of the estimation process the expected cost at any given point in time is estimated as a function of failure time and in the second stage the estimated expected costs at given points in time are weighted by the Kaplan–Meier probability of death at these points in time. The estimate of mean total cost over the duration of interest is then derived as the sum over time of these weighted individual cost estimates. A regression approach is used to derive the expected costs where only uncensored individuals contribute cost information in order to avoid the bias in the regression parameter estimates imparted by censoring. Alternative parametric assumptions can be made regarding the relationship between cost and survival time depending on the data under consideration. Due to the consistency of the Kaplan-Meier estimator, consistency of the proposed estimator is ensured if the regression model specifying the relationship between cost and failure time is consistently estimated.

The second parametric alternative was introduced by Lin (2000) and assumes a regression model in which cost is linearly related to a set of covariates of interest. The method derives estimates of the regression parameters accounting for the presence of censoring and is not restricted by the censoring pattern. Two estimators result from this approach. The first uses the individual total accumulated costs at the individual's point of death or censoring while the second makes use of multiple cost observations on each subject obtained at various points in time over the study period. The main advantage of the latter estimator is an increase in efficiency by allowing use of cost information that is not used by the preceding estimator. In both cases the estimates of the regression

parameters are adjusted for censoring by incorporating the inverse of the probability of an individual not being censored evaluated at the point of the individual observed cost in the estimating equations. The approach derives consistent estimates for the regression parameters and martingale theory provides asymptotic covariance matrix estimates.

Assessment of the estimators' performance under the censoring conditions described above was based on comparing the resultant estimates with the respective estimates derived from the best non-parametric estimators. The Carides et al estimator resulted in biased estimates for all parameterisations considered for the relationship between cost and failure time. The results indicated that the major source of bias was the high degree of censoring rather than a potential misspecification of the regression model. Given that under this approach bias in the cost estimates arises from bias in the estimates of the regression parameters, it is not surprising that the estimated coefficients do not reflect the true parameter values when their derivation was based on only 18% of the observed data which constituted the uncensored subset. Therefore, although such an approach is appealing on the basis that it attempts to model the time pattern of cost, it is of limited value at high levels of censoring. Given the potential value of methods that allow extrapolation of cost beyond the study period development of parametric models that successfully do so under conditions of heavy censoring appears to be a fruitful area for future research.

Concentrating on the Lin regression methodology, the approach using cost information solely from the complete cases yielded biased estimates of cost as expected given the limited amount of cost information entering the estimation process, while the approach using information on individual cost histories resulted in estimates that were very close to the ones derived from the best performing non-parametric methods which also use information on the individual cost history process. This result indicates that incorporation of covariate information in the estimation did not improve upon the cost estimates when these were compared to the ones derived from the best performing non-parametric estimators. Such a finding however is not meant to undermine the validity and usefulness of this particular regression methodology in modelling censored costs. It is rather the case that the covariates considered in the particular application did not provide any additional information in explaining cost which in turn implies the lack of any additional gain in adopting this methodology in deriving cost estimates compared to a non-parametric alternative. In general however assessment of the impact of individual covariates on cost is likely to be of major importance not least because it allows generalising the study results to different patient populations defined by different covariate values. The implication for the data collection process alongside a clinical trial is then that at the design stage of such a study identification of covariates likely to explain cost should be pursued.

Although within the context of this particular application there is no gain in adopting a parametric approach over a consistent non-parametric estimator especially given that a parametric methodology by definition involves a greater number of assumptions compared to a non-parametric

alternative, the findings reached from the investigation of the Lin regression methodology provide further insight into the general issue of cost estimation in the presence of censoring in the following manner. Aside from identifying a regression methodology which performs well under extreme censoring conditions, the analysis strengthens the validity of the main conclusion reached in the non-parametric analysis. That is, in general weighting the complete observations by the inverse of the probability of an observation not being censored in deriving cost estimates provides an effective means for handling the presence of censoring. Nevertheless under conditions of heavy censoring the success of such a method will also be subject to the amount of available information on the cost history process as this will in turn determine the degree of retrieval of cost information missing due to censoring. The implication for the design of the clinical study is that regardless of whether the statistical methodology to be employed in the analysis of cost data under conditions of censoring is of a parametric or non-parametric nature, effort should be made to record cost information at intermediate points in time over the study duration. The findings derived from the preceding analysis provide conclusive evidence in support of this requirement with the value of the available information on the cost history process increasing as the degree of censoring increases.

To conclude, all aspects of the investigation undertaken in this thesis signify the importance that the level of cost information imparts to the study of the distribution of treatment cost. Normally cost information will be incomplete to differing degrees and for a variety of reasons. The issue then becomes to identify the most appropriate methodology for deriving consistent estimates of cost when information is missing. The choice between alternative methods narrows as the degree of incompleteness becomes higher. At the levels of incompleteness considered in the preceding analysis when cost information is missing due to censoring, the general conclusion is consistent with that reached in similar investigations of censored data within different analytical contexts. However additional concerns were raised by the present analysis mainly due to the extreme censoring levels which were not an issue in the few existing studies of censored cost data. More specifically, the effectiveness of the inverse of the probability of non-missingness in adjusting estimates of the statistics of interest for censoring is confirmed within the context of censored cost data analysis. The same probability weight has been used in numerous applications in an attempt to adjust estimates for missingness including the study of censored failure times, adjusting regression coefficients for missingness in the data, studying semiparametric regression models in the presence of covariate dependent censoring, and estimating the distribution of quality adjusted survival time under conditions of censoring. In all these applications use of the inverse of the probability of inclusion in the estimating equations results in consistent estimators for the statistics of interest while adjusting for missingness. The same general finding emerges from the analysis undertaken in this thesis but under conditions. When the level of censoring is too high the specific weight is necessary but not sufficient in adjusting the estimates for the particular type of missingness. In these circumstances knowledge of the history of the process under study proves a determining factor in the performance of the estimator. The proposed methodology then becomes both necessary and

sufficient and derives consistent estimates of medical cost accounting for the missingness in the data due to censoring even when censoring reaches extreme levels.

References

Andersen, P.K., Borgan, Ø., Gill, R.D., Keiding N., 1993. Statistical Models Based on Counting Processes. Springer-Verlag, New York.

Andersen, P.K., Gill, R.D., 1982. Cox's regression model for counting processes: A large sample study. The Annals of Statistics 10, 1100-1120.

Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Identification of causal effects using instrumental variables. Journal of the American Statistical Association 91, 444-455.

Bang, H., Tsiatis, A.A., 2000. Estimating medical costs with censored data. Biometrika 87, 329-343.

Barber, J., Thompson, S., 1998. Analysis and interpretation of cost data in randomised controlled trials: review of the published literature. British Medical Journal 317, 1195-1200.

Beale, E.M.L., Little, R.J.A., 1975. Missing values in multivariate analysis. Journal of Royal Statistical Society B 37, 129-145.

Breslow, N.E., 1972. Contribution to discussion of paper by D.R. Cox. Journal of the Royal Statistical Society B 34, 216-217.

Breslow, N.E., 1974. Covariance analysis of censored survival data. Biometrics 30, 89-99.

Breslow, N.E., 1975. Analysis of survival data under the proportional hazards model. International Statistical Review 43, 55-58.

Briggs, A., Gray, A., 1998. Handling Uncertainty when Performing Economic Evaluation of Health Care Technologies. Health Technology Assessment Review 3, 2.

Briggs, A., 2001. Handling uncertainty in economic evaluation and presenting the results. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Brouwer, W., Rutten, F., and Koopmanschap, M., 2001. Costing in economic evaluation. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Buckley, J., James, I., 1979. Linear regression with censored data. Biometrika 66, 429-436.

Buxton, M., Drummond, M., Hout, van B., et al, 1997. Modelling in economic evaluation: an unavoidable fact of life. Health Economics 6, 217-227.

Carides, G.W., Heyse, J., F., Iglewicz, B., 2000. A regression-based method for estimating mean treatment cost in the presence of right-censoring. Biostatistics 1, 299-313.

Carroll, R.J., Wand, M.P., 1991. Semiparametric estimation in logistic measurement error models. Journal of the Royal Statistical Society B 53, 573-587.

Chiang, A.C., 1984. Fundamental Methods of Mathematical Economics. McGraw-Hill International Editions.

Commonwealth of Australia, 1995. Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Advisory Committee. Australian Government Print Office, Canberra.

Cox, D.R., 1972. Regression models and life-tables (with discussion). Journal of the Royal Statistical Society B 34, 187-220.

Cox, D.R., 1975. Partial likelihood. Biometrika 62, 269-276.

Cox, D.R., Oakes, D., 1984. Analysis of Survival Data. Monographs on Statistics and Applied Probability 21. Chapman & Hall, London.

Coyle, D., 1996. Statistical analysis in pharmacoeconomic studies. Pharmacoeconomics 9, 506-516.

Coyle, D., Drummond, M., 1996. Analysing differences in costs of treatment across centres within economic evaluations. Paper presented to iHEA, Vancouver.

Csörgö, S., Horvath, L., 1983. The rate of strong uniform consistency for the product limit estimator. Z. Wahrsch. Verw. Geb. 62, 411-426.

Dagenais, M.G., 1973. The use of incomplete observations and multiple regression analysis: A generalised least squares approach. Journal of Econometrics 1, 317-328.

Deiner, A., O'Brien, B., Gafni, A., 1998. Health care contingent valuation: a review and classification of the literature. Health Economics 7, 313-326.

Dolan, P., 2000. The measurement of health related quality of life. In Handbook of Health Economics, Eds. Culyer, A.J., Newhouse, J.P. North-Holland, Amsterdam.

Dolan, P., 2001. Output measures and valuation in health. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Dranove, D., 1996. Measuring costs. In Valuing Health Care, Ed. Sloan, F.A., 61-75. Cambridge University Press.

Duan, N., 1983. Smearing estimate: A nonparametric retransformation method. Journal of the American Statistical Association 78, 605-610.

Duan, N., Manning, W.G., Morris, C.N., 1983. A comparison of alternative models for the demand for medical care. Journal of Bussiness and Economic Statistics 1, 115-126.

Dudley, R.A., Harrell, F.E., Smith, L.R., Mark, D.B., Califf, R.M., Pryor, D.B., Glower, D., Lipscomb, J., Hlatky, M., 1993. Comparison of analytic models for estimating the effect of clinical factors on the cost of coronary artery bypass graft surgery. Journal of Clinical Epidemiology 46, 261-271.

Drummond, M., Stoddart, G., 1984. Economic analysis and clinical trials. Controlled Clinical Trials 5, 115-128.

Drummond, M., Davies, L., 1991. Economic analysis alongside clinical trials: revisiting the methodological issues. International Journal of Health Technology Assessment 7, 561-573.

Drummond, M., Bloom, B., Carrin, G., et al, 1992. Issues in the cross-national assessment of health care technology. International journal of Health Technology Assessment 8, 671-682.

Drummond, M., 1994. Economic Analysis Alongside Clinical Trials. Department of Health, Leeds.

Drummond, M., O'Brien, B., Stoddart, G., Torrance, G., 1998. Methods for the Evaluation of Health Care. Oxford University Press, Oxford.

Drummond, M., McGuire, A., Editors, 2001. Economic Evaluation in Health Care: Merging theory with practice. Oxford University Press.

Drummond, M., Pang, F., 2001. Transferability of economic evaluation results. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Ellwein, L., Drummond, M., 1996. Economic analysis alongside clinical trials: bias in the assessment of economic outcomes. International Journal of Health Technology Assessment 12, 691-697.

Efron, B., Tibshirani, R., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Efron, B., 1967. The two sample problem with censored data. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 4, 831-853. Prentice-Hall, New York.

Etzioni, R., Urban, N., Baker, M., 1996. Estimating the costs attributable to a disease with application to ovarian cancer. Journal of Clinical Epidemiology 49, 95-103.

Etzioni, R., Feuer, E., Sullivan, S., Lin, D., Hu, C., Ramsey, S.D., 1999. On the use of survival analysis techniques to estimate medical care costs, Journal of Health Economics, 18, 365-380.

Fenn, P., McGuire, A., Phillips, V., Backhouse, M., Jones, D., 1995. The analysis of censored treatment cost data in economic evaluation. Medical Care 33, 851-863.

Fenn, P., McGuire, A., Backhouse, M., Jones, D., 1996. Modelling programme costs in economic evaluation. Journal of Health Economics 15, 115-125.

Fleming, T.R., Harrington, D.P., 1991. Counting Processes and Survival Analysis. Wiley, New York.

Garber, A., 2000. Advances in cost-effectiveness analysis. In Handbook of Health Economics, Eds. Culyer, A.J., Newhouse, J.P. North-Holland, Amsterdam.

Gihman, I.I., Skorohod, A.V., 1974. The theory of Stochastic Processes I. Springer-Verlag, Berlin Heidelberg New York.

Gill, R.D., 1980. Censoring and Stochastic integrals, Mathematical Centre Tracts 124. Mathematish Centrum, Amsterdam.

Glick, H., Polsky, D., Sculman, K., 2001. Trial-based economic evaluations: an overview of design and analysis. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Gold, M., Seigal, J.E., Russell, L.B., Weinstein, M.C., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press, New York.

Gourieroux, C., Montfort, A., 1981. On the problem of missing data in linear models. Review of Econometric Study xlviii, 579-586.

Grambsch, P.M., Therneau, T.M., 1994. Proportional hazards tests and diagnostics based on weighted residuals. Biometrika 81, 515-526.

Graves, N., Walker, D., Raine, R., et al, 2002. Cost data for individual patients included in clinical studies: no amount of statistical analysis can compensate for inadequate costing methods. Health Economics Letters (in press)

Green, W.H., 1997. Econometric Analysis. Prentice-Hall International.

Griliches, Z., 1986. Economic data issues. In Handbook of Econometrics vol.3, Eds. Griliches, Z., Intriligator, M. Amsterdam, North Holland.

Hay, J.W., 1989. Econometric issues in modelling the costs of AIDS. Health Policy 11, 125-145.

Heathfield. D.F., 1971. Production Functions. Macmillan Studies in Economics. The Macmillan Press, London.

Heyse, J.F., Cook, J.R., Carides, G.W., 2001. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Horvitz, D.G., Thompson, D.J., 1952. A generalisation of sampling without replacement from a finite universe. Journal of the American Statistical Association 47, 663-685.

Johannesson, M., 1996. Theory and methods of economic evaluation of health care. Kluwer, Amsterdam.

Johnston, K., Buxton, M., Jones, D.R., Fitzpatrick, R., 1999. Assessing the Costs of Healthcare Technologies in Clinical Trials. Health Technology Assessment Review 6, 3.

Jonsson, B., Weinstein, M.C., 1997. Economic evaluation alongside clinical trials: study considerations for GUSTOIIb. International Journal of Health Technology Assessment 13, 49-58.

Kaplan, E.L., Meier, P., 1958. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association 53, 457-481.

Knapp, M., Beecham, J., 1993. Costing mental health services. Psychological Medicine 20, 892-908.

Koul, H., Susarla, V., van Ryzin, J., 1981. Regression analysis with randomly right-censored data. The Annals of Statistics 9, 1276-1288.

Kuntz, K., Weinstein, M., 2001. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Laan, M.J, Van Der, Hubbard, A.E., 1998. Locally efficient estimation of the survival distribution with right-censored data and covariates when collection of data is delayed. Biometrika 85, 771-783.

Lancaster, T., 1990. The Econometric Analysis of Transition Data. Econometric Society Monographs 17. Cambridge University Press.

Liang, K.-Y., Zeger, S.L., 1986. Longitudinal data analysis using generalised linear models. Biometrika 73, 13-22.

Lin, D.Y., Feuer, E.J., Etzioni, R., Wax, Y., 1997. Estimating medical costs from incomplete follow-up data. Biometrics 53, 113-128.

Lin, D.Y., Ying, Z., 1993. A simple nonparametric estimator of the bivariate survival function under univariate censoring. Biometrika 80, 573-581.

Lin, D.Y., Ying, Z., 1993. Cox regression with incomplete covariate measurements. Journal of the American Statistical Association 88, 1341-1349.

Lin, D.Y., Wei, L.J., Yang, I., Ying, Z., 2000. Semiparametric regression for the mean and rate functions of recurrent events. Journal of the Royal Statistical Society B 62, 711-730.

Lin, D.Y., 2000. Linear regression analysis of censored medical costs. Biostatistics 1, 35-47.

Lin, D.Y., 2000. Proportional means regression for censored medical costs. Biometrics 56, 775-778.

Lipscomb, J., Ancukiewicz, M., Parmigiani, G., Hasselblad, V., Samsa, G., Matchar, D., 1998. Predicting the cost of illness: A comparison of alternative models applied to stroke. Medical Decision Making 18 Supplement, S39-S56.

Little, R.J.A., 1988. A test of missing completely at random for multivariate data with missing values. Journal of the American Statistical Association 83, 1198-1202.

Little, R.J.A., 1992. Regression with missing X's: A review. Journal of the American Statistical Association 87, 1227-1237.

Manning, W.G., 1998. The logged dependent variable, heteroscedasticity, and the retransformation problem. Journal of Health Economics 17, 283-295.

Meyer, P.A., 1966. Probability and Potentials. Waltham MA: Blaisdell.

Miller, R.G., 1976. Least squares regression with censored data. Biometrika 63, 449-464.

Morris, S., 1997. A comparison of economic modelling and clinical trials in the economic evaluation of cholesterol-modifying pharmacotherapy. Health Economics 6, 589-601.

Mullahy, J., Manning, W., 1996. Statistical issues in cost-effectiveness analyses. In Valuing Health Care, Ed. Sloan, F.A., 149-184. Cambridge University Press.

Nelson, W., 1969. Hazard plotting for incomplete failure data. Journal of Qualitative Technology 1, 27-52.

Newey, W.K., 1990. Semiparametric efficiency bounds. Journal of Applied Econometrics 5, 99-135.

Oakes, D., 1972. Contribution to discussion of paper by D.R. Cox. Journal of the Royal Statistical Society B 34, 208.

O'Brien, B., 1996. Economic evaluation of pharmaceuticals: Frankenstein's monster or vampire of trials. Medical Care 34, DS99-DS108.

O'Brien, B., Drummond, M., Labell, R., Willan, A., 1994. In search of power and significance: issues in the design and analysis of stochastic cost-effectiveness studies in health care. Medical Care 32, 150-163.

O'Hagan, A., Stevens, J., 2002. On estimators of medical costs with censored data. Unpublished mimeo, University of Sheffield.

Pepe, M.S., Fleming, T.R., 1991. A nonparametric method for dealing with mismeasured covariate data. Journal of the American Statistical Association 86, 108-113.

Raikou, M., Briggs, A., Gray, A., McGuire, A., 2000. Centre-specific or average unit costs in multi-centre studies? Some theory and simulation. Health Economics 9, 191-198.

Robins, J.M, Rotnitzky, A., 1992. Recovery of information and adjustment for dependent censoring using surrogate markers. In AIDS Epidemiology-Methodological Issues, Eds. Jewell, J., Dietz, K., Farewell, V., 297-331. Birkhäuser, Boston.

Robins, J.M, Rotnitzky, A., Zhao, L.P., 1994. Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89, 846-866.

Robins, J.M., Rotnitzky, A., Zhao, L.P., 1995. Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. Journal of the American Statistical Association 90, 106-121.

Robins, J.M., Greenland, S., 1996. Comment on Angrist, J.D., Imbens, G.W., Rubin, D.B., 1996. Journal of the American Statistical Association 91, 456-458.

Rotnitzky, A., Robins, J.M., 1995. Semiparametric regression estimation in the presence of dependent censoring. Biometrika 82, 805-820.

Rubin, D.B., 1976. Inference and missing data. Biometrika 63, 581-592.

Rubin, D.B., 1978. Multiple imputations in sample surveys – A phenomenological Bayesian approach. In Proceedings of the Survey Research Methods Section, American Statistical Association 20-34.

Rubin, D.B., 1978a. Multiple Imputation for Nonresponse in Surveys. John Wiley, New York.

Schoenfeld, D., 1982. Partial residuals for the proportional hazards regression model with censored observations. Biometrika 69, 239-241.

Schulman, K., Burke, J., Drummond, M., et al, 1998. Resource costing for multinational neurologic clinical trials: methods and results. Health Economics 7, 629-638.

Sculpher, M., 2001. The role and estimation of productivity costs in economic evaluation. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

Sen, A., 1982. Choice, Welfare and Measurement. Harvard University Press, Cambridge, Mass.

Spiegelhalter, F.A, Jones, D.R., Parmar, M.K.B., et al, 1996, Being economical with the costs. Department of Epidemiology and Public Health Technical Paper 96-06. University of Leicester, Leicester.

Therneau, T. M., Grambsch, P. M., 2000. Modelling Survival Data: Extending the Cox Model. Springer-Verlag, New York.

Tobin, J., 1958. Estimation of relationships for limited dependent variables. Econometrica 26, 24-36.

Tsuchiya, A., Williams, A., 2001. Welfare economics and economic evaluation. In Economic Evaluation in Health Care: Merging theory with practice, Eds. Drummond, M., McGuire, A. Oxford University Press.

UKPDS 33: UK Prospective Diabetes Study Group, 1998. Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes. The Lancet 352, 837-853.

UKPDS 41: Gray, A., Raikou, M., McGuire, A., et al, 2000. Cost-effectiveness of an intensive blood glucose control policy in patients with type 2 diabetes: economic analysis alongside randomised controlled trial. British Medical Journal 320, 1373-1378.

Weinstein, M.C., Zeckhauser, R., 1973. Critical ratios and efficient allocation. Journal of Public Economics 2, 147-157.

Weinstein, M. C., Stason, W.B., 1976. Hypertension: A Policy Perspective. Harvard University Press, Cambridge.

Whynes, D., Walker, A.R, 1995. On approximations in treatment costing. Health Economics 4, 31-40.

Williams, A., Cookson, R., 2000. Equity in health care. In Handbook of Health Economics, Eds. Culyer, A.J., Newhouse, J.P. North-Holland, Amsterdam.

Willke, R., Glick, H., Polsky, D. Schulman, K., 1998. Estimating country-specific cost-effectiveness from multinational clinical trials. Health Economics 7, 481- 494.

Zhao, H., Tsiatis, A.A., 1997. A consistent estimator for the distribution of quality adjusted survival time. Biometrika 84, 339-348.

Appendix A.2.1. Review of the economic literature on type 2 diabetes

As stated in chapter 1 the empirical analysis in chapters 4 and 5 was based on the UKPDS trial data which assessed whether intensive policy delayed diabetes related mortality and the onset and progression of diabetic complications compared to conventional policy in a population of newly diagnosed type 2 diabetic patients. The aim of the review chapter was to provide general background information on a number of methodological issues arising in the collection and analysis of cost data which are considered in detail in the following chapters. Given that interest in the thesis is in methodological aspects of general applicability that are not restricted to a particular disease area, discussion of the economic literature in the area of diabetes was deemed to be of secondary importance. An overview of this literature is thus presented below for completeness purposes. In fact, as will be shown, there have been relatively few economic analyses undertaken in this clinical area even though this is a relatively prevalent disease. Moreover in accordance with the findings of Briggs and Gray (1998) discussed above, the limited degree of statistical analysis undertaken within economic evaluation appears to be true within this disease area. Given the overriding concern with censored cost data it is also important to note that none of the studies reviewed below account for censoring in their analyses.

The economics of screening and treatment in type 2 diabetes

Over a relatively recent period there has been a vast improvement in the understanding of the basic aetiology, epidemiology and treatment of diabetes mellitus. This chronic metabolic disorder which occurs when the body is unable to control blood glucose levels efficiently is classified into two main types. Type 1 diabetes, arising from chronic failure of insulin secretion and requiring insulin therapy, is commonly associated with children and individuals under the age of 40. Type 2 diabetes, common in the elderly and certain ethnic populations, is caused by defective insulin secretion and action and is treated through dietary control and drugs. A third and less common form is gestational diabetes and results from complications during pregnancy. Type 2 diabetes has the highest prevalence with over 75 per cent of all diabetic patients being classified as type 2 (Harris, 1996). Information on all types of manifestation is difficult to acquire because of the chronic, long-term nature of the disease, the large range of risk factors and the many associated complications. Knowledge has improved with the collection of basic survey data on a range of populations (Alberti, 1993; Alwin and King, 1995; King and Rewers, 1993), by greater awareness of the attributable risks of comorbidities resulting from diabetes and through a number of clinical trials which have followed individuals over time to defined clinical end-points (UKPDS 1998; Ohkubo et al, 1995). Such trials have typically defined diabetes as a fasting plasma blood glucose level greater than 6mmol/l on at least two occasions.

As epidemiological and clinical information has improved, there has been a greater emphasis placed on the economic consequences of the disease. This is not surprising given the magnitude of the economic impact that this particular disease has on individuals and health systems. As the illness progresses the range and prevalence of complications also increase. Any therapy whose aim is to slow disease progression, thereby delaying the onset of complications, is therefore likely to be of major clinical and economic benefit. Consequently one of the main research questions is whether intensified management of type 2 diabetes is effective and, if so, cost-effective.

The aim here is to present a systematic review of the literature on the economics of screening and treatment in type 2 diabetes with emphasis on cost of illness and cost-effectiveness studies. This systematic review is supplemented by an overview of some of the main findings in the literature relating to the cost-effectiveness of treating type 2 diabetic complications. The review is structured as follows. A discussion of the methods employed in identifying the relevant literature and the

criteria used for selecting the studies for inclusion is presented first. Subsequently the main findings of the individual studies under review are reported and these are followed by a discussion of the main issues identified within this literature. Finally there is a short discussion of the literature related to the economics of type 2 diabetic complications.

Methods

A search of the literature dating back to 1995 was undertaken. The justification for the choice of this date is that recent epidemiological and clinical data have had a dramatic impact on the general knowledge of diabetes and treatment patterns have subsequently been changing significantly. Moreover the incidence and prevalence of diabetes, in particular type 2 diabetes, appear to have been increasing over the recent past. Finally the evidence on cost-effectiveness of type 2 diabetes treatment interventions prior to this date was covered in a review by Gulliford (1997), specifically within the UK by Marks (1995) and more generally by Jonsson et al (1995). Computerised searches were used to identify articles on the economics of diabetes. MEDLINE and the Social Sciences Citation Index were the primary source databases. Search terms included "diabetes mellitus", "diabetes", "economics", "economic evaluation", "cost-effectiveness", "cost" and "cost-benefit". The web site http://www.pitt.edu/~tjs/costrefs.html, dedicated to referencing economic literature relating to diabetes, was also searched. In addition the references of each individual article were hand searched. While the initial search identified over 300 studies the vast majority of these were excluded as the analysis was not specific to type 2 diabetes or its complications. The 58 remaining articles were then examined for duplication, lack of analytical content and general suitability leaving less than 20 studies forming the core of this review. The inclusion criteria were that all studies undertake an economic analysis of type 2 diabetes alone or that there was a clear distinction of the type 2 population from the general diabetic population. Exclusion criteria were based on whether or not the individual studies were based solely on secondary data, whether findings relating to type 2 diabetes could be distinguished from general results (i.e. results relating to diabetes mellitus in general), whether budgetary considerations alone were discussed and whether an economic analysis was reported. Articles were then classified into two main categories: those dealing with the cost/burden of type 2 diabetes and those concerned with economic evaluations of type 2 diabetes interventions.

The review of the economic studies relating to diabetic complications is not considered to be systematic as a number of articles were identified which presented results on diabetes mellitus generally without differentiating between type 1 and type 2 diabetes. Where a clear distinction was made, the findings have been included here, but this does not mean that the present overview of the economics of diabetic complication in type 2 diabetics is fully comprehensive.

According to the general format recommended by the U.K. National Health Service Economic Evaluation Database the following information was extracted, where possible and appropriate, from each study:

Author (s)
Year of publication
Year used for cost valuation
Country of analysis
Currency used for cost valuation
Alternative considered for evaluation (if relevant)
Cost-effectiveness measure (where relevant)
Patient population
Effectiveness data sources

Cost elements
Cost data sources
Time horizon
Discount rate
Variables included in sensitivity analysis
Baseline results
Results from sensitivity analysis
Author(s) conclusions

Information was extracted in this manner for ease of exposition. All costs were converted into 1999 prices using a domestic deflator and converted into US$ using the prevailing exchange rate to allow comparison.

Results

The main findings are reported in Tables A.2.1.1 and A.2.1.2 with textual commentary highlighting methodological or additional results. The burden and cost-of-illness studies in type 2 diabetes are presented first (results in Table A.2.1.1). Given that the information gathered is both country- and time-specific the results are reported largely in a descriptive manner. Following this set of results, the cost-effectiveness literature is discussed (results in Table A.2.1.2). Finally a brief discussion of the literature on the economics of complications arising from type 2 diabetes is presented in the text but not in a separate table as it only attempts to be informative rather than comprehensive. More extensive reviews of cost studies on a number of specific diabetic complications are given by Ragnarson-Tennvall and Apelqvist (1997), Deerochanawong (1992), Waugh (1989) and Wood (1990).

Burden of illness studies

It has been reported that type 2 diabetes affected an estimated 110 million individuals worldwide in 1995 and that this would more than double by the year 2010 due to demographic influences (Zimmet and McCarthy, 1995; Alberti, 1997). Such estimates however rely on susceptible prevalence figures. It has been estimated that at any given time up to 50 per cent of type 2 diabetic cases in the population are undiagnosed (Alberti, 1995). The limited although continuously improving prevalence and incidence data make it difficult to estimate the true patient population.

Even if disease prevalence is established, calculation of the burden of illness requires estimation of the attributable cost arising from associated complications. It is known that diabetes increases the relative risk of other diseases. The range of complications is well known: retinopathy, nephropathy, neuropathy and macrovascular disease are all common. The precise attributable risks of these diseases are however not fully established although estimates for a number of complications exist. The aetiology of the disease therefore makes it difficult to assess the true cost of illness of type 2 diabetes for at least two reasons. First, costs of illness studies based solely on a primary diagnosis of type 2 diabetes will considerably underestimate the true resource cost given that they do not take account of the resource cost associated with the treatment of complications. The acknowledged under-reporting of type 2 diabetes in hospital and mortality statistics exacerbates such underestimation. Secondly and related to this, the true resource costs associated with the treatment of complications are best defined as the excess treatment costs incurred by the diabetic population over and above the costs incurred by a matched cohort population without diabetes.

Bearing these problems in mind the most straightforward studies in terms of methodology are those by Henriksson et al (2000) and Evans et al (2000). The first of these studies calculated costs based on questionnaires concerning resource utilisation over a six-month period within nine primary care centres in Sweden covering a total of 777 type 2 diabetic patients. These data were factored up to the Swedish population through the use of prevalence data to give an annual estimate of the cost of type 2 diabetes in Sweden. Prevalence and incidence data on complications were obtained from the questionnaires and used to estimate the cost of treating the associated complications. These rates were compared with rates gained from a review of the literature. Generally the study prevalence rates were smaller than the literature based figures making it difficult to verify the study rates. Overall, the results showed that hospitalisation costs dominated treatment costs amounting to 42 per cent of the total. Drug costs were 27 per cent of total costs. Drug costs for insulin treated individuals were twice as high as in those treated with oral antidiabetic agents. Complications had a varying impact on cost. For individuals with microvascular complications alone the annual cost was of the same order of magnitude as for those with no complications. For the ones with macrovascular complications alone the annual cost was approximately double the cost of those individuals without complications. Costs for individuals with both microvascular and macrovascular complications were approximately three fold higher than the costs in those without complications. In a similar vein the study by Evans et al (2000) used a patient register in a region of Scotland to identify both type 1 and type 2 diabetic patients. Using extrapolation the authors suggest that patients with diabetes account for approximately 8 per cent of the UK drugs budget and of this 90 per cent of the expenditure is attributable to type 2 diabetic patients. The relative risk of drug usage was higher for type 1 diabetes (1.70 for type 2, 2.07 for type 1) but given their lower prevalence the budget impact was greater for type 2 diabetes.

O'Brien et al (1998) used resource utilisation and unit costs from a wide variety of sources to derive their estimates. Resource use profiles were designed for a number of complications over lifetime with the first year costs defined explicitly and subsequent years defined as static states such that subsequent years were allocated the same cost for each of the complications if this was applicable. This is obviously a simplification likely to lead to a conservative estimate of costs since it assumes a constant rate of complication treatment cost over time. Five US state discharge databases were used to determine the prevalence of the complications based on primary diagnosis with diabetes recorded as a comorbidity. The event cost was dominated by the hospitalisation cost, apart from the case of ischaemic stroke where hospitalisation cost represented a quarter of the event total cost. The formulation of the subsequent annual costs was based on treatment guidelines due to inadequate published observational studies. The most obvious application of this analysis would be to the population of the standard therapy arm of a cost-effectiveness model. The authors point out however that their approach has a number of limitations arising from the lack of observational studies on the treatment of type 2 diabetes and its complications, their assumption that the annual costs attributed to complications are constant over time and that as treatment technology improves their estimates of cost will date.

A similar method is used in two related studies (Brown et al, 1999; Brown et al, 1999) that are based on observational data. The first study considers newly diagnosed type 2 HMO patients matched with a control cohort over 8 years of follow-up. Observational data were based on clinical records and limited to the HMO scheme. As outpatient and other ancillary service use were not directly recorded, resource utilisation of these services was estimated by a regression model. Broadly speaking the annual costs of the diabetic population were approximately double those of the matched case controls. After years one and two costs fell to their lowest level and then increased steadily over the remaining study period by 65 per cent at the end of the study compared to the initial year. Costs were dominated by in-patient treatment (46 per cent of the total). Cardiac and cerebrovascular complications were 23 per cent of total incremental costs where the increment was defined as costs attributable to the diabetic population over and above the matched control costs.

An interesting finding was that hospital admissions unrelated to diabetes accounted for more than half of the incremental cost over the study period and consequently for most of the growth in these costs over time. The authors concluded that the more costly stages of treatment for diabetic complications occur after an 8-year lead-time. The authors also state that their cost figures are lower than other reported estimates and attribute this to the use of costs rather than charges and the use of cost-effective strategies utilised by the particular HMO.

In the second study undertaken by Brown et al (1999) a similar population was used and incremental costs over and above a matched control population were calculated for the differing stages of treatment for cardiovascular and renal complications. The estimates were based on regression analysis. Type 2 diabetic patients without complications had an average annual treatment cost of $2,263. The incremental cost for cardiovascular disease was approximately $1,210, making total average annual treatment cost equal to $3,472 if the patient was on drug therapy alone for cardiovascular disease and rising to approximately an annual total treatment costs of $8,235 if a major cardiovascular event was suffered. Type 2 diabetic patients with renal disease were estimated to have annual total treatment costs of $3,750, $4,428 and $17,445 (all 1999 prices) for early onset, advanced renal dysfunction and chronic renal failure respectively. These were higher estimates than earlier studies had reported which according to the authors was due to a better representation of incremental costs through the modelling approach than reliance purely on observational data on the basis that observational studies only attribute costs after clinical identification of the complications. The modelling approach, according to the authors, is better able to identify the increased costs incurred prior to clinical (or labelled) identification as a model provides more accurate estimates across the full term of treatment. Clinical understanding of the disease however did not support some of the model predictions. Females for instance had slightly higher predicted treatment costs in the model than males which is not what clinical evidence would suggest.

Economic evaluation of treatment interventions for type 2 diabetes

There are relatively few studies on the cost-effectiveness of interventions for type 2 diabetes as shown in Table A.2.2. This reflects as stated above the long observation time required to track disease progression and complications, the difficulties in establishing optimal standard therapies and the relatively few long-term clinical studies that assess interventions in terms of final outcome. The economic evaluation literature mainly addresses the following questions: is screening for type 2 diabetes cost-effective; are there primary prevention strategies which are cost-effective; is intensive therapy cost-effective; is early initiation of insulin therapy cost-effective. In general, each one of these questions has been addressed in one study only reflecting the limited extent of the literature in this area.

Given the under-reporting of type 2 diabetes and the extensive range of complications, screening and primary prevention clearly become policy options. The Centre for Disease Control and Prevention of Diabetes Cost-effectiveness Group ran a Monte Carlo simulation model to estimate the lifetime cost-effectiveness of a one-year opportunistic screening programme (Engelgau et al, 1998). Cost-effectiveness was estimated for a cohort of individuals aged 25 years and over. This opportunistic programme was compared with the current US guidelines which recommend that screening be initiated at 45 years of age. The model predicted that screening in the younger cohorts reduced microvascular complications. The health benefits were large in terms of life years gained, and more than doubled when measured in terms of quality adjusted life years (QALYs). Screening was more cost-effective when applied to the youngest age groups as they had the most QALYs to gain and to ethnic minorities as these have a higher incidence of the disease. The results were sensitive to the assumptions made and should therefore be taken as indicative rather than authoritative.

Segal et al (1998) used a similar modelling approach in the analysis of primary prevention programmes in Australia. Given that particular populations are at higher risk of type 2 diabetes and that incidence increases with age this study considered whether particular prevention programmes would be cost-effective from a health service provider perspective. The programmes ranged from intensive dietary therapy and behavioural change in obese men through surgical intervention for obesity to a media campaign aimed at informing the general population of disease symptoms and progression. Transition matrices were formed across three states: normal glucose tolerance, impaired glucose tolerance and non-insulin dependent diabetes. The transition probabilities were based on a Swedish study of intensive weight loss and fitness enhancement programme for overweight individuals and mortality rates were obtained from epidemiological literature. Under a range of assumptions all primary prevention programmes considered were stated to be "highly cost-effective". This result arose because the low provider incurred cost in the majority of the prevention programmes was recouped in cost-offsets due to delay or avoidance of complications. Even when success rates were reduced to relatively low levels the cost-effectiveness of the majority of programmes was retained. As the authors state the cost-effectiveness of these programmes would be improved greatly if quality of life was incorporated into the outcome measure. The model is characterised as relatively simple by the authors with single transition matrices used to progress each cohort to different diabetic states. There are only three transition probabilities which influence the outcomes and this does not adequately reflect the observed complex progression of the disease. Long-term clinical studies show that glucose intolerance is an inherently dynamic property associated with increasing baseline levels over time.

Until recently the standard clinical practice of treating type 2 diabetes has been based on dietary control initiated at diagnosis and altered to drug therapy as glucose intolerability increases and complications manifest. This has recently been challenged and the effectiveness and cost-effectiveness of intensive therapy and early introduction of insulin have been studied. Given the importance of dietary advice, Franz et al (1995) compared basic nutritional advice to advice based on formal guidelines. The cost-effectiveness of this study was based on a trial that failed to show a statistical difference in the clinical outcomes which were assessed in terms of fasting plasma glucose levels and HbA_{1C} levels. Nevertheless, the insignificant difference in the clinical outcomes resulted in similar cost-effectiveness ratios. The authors suggested that the lack of any conclusive finding might have been a reflection of the short time period (six months) considered.

At the other end of the treatment spectrum is the evaluation of insulin therapy in the type 2 diabetic population. The Kumamoto clinical study (Wake et al, 2000) investigated whether intensive glycaemic control based on multiple insulin injections (MIT) reduces the frequency or severity of microvascular complications compared to conventional insulin therapy (CIT) in this population. On entry, patients were classified into a primary prevention group who had no evidence of retinopathy or microalbuminuria and to a secondary prevention group with mild retinopathy and microalbuminuria. Over a six-year period the cumulative percentages of the development and progression of retinopathy and nephropathy were 7.7 per cent for the MIT group and 32 per cent for the CIT group in the primary prevention sub-population, and 19.2 per cent for the MIT group and 44 per cent for the CIT group in the secondary prevention sub-population. The population was subsequently followed-up for a total of ten years. Generally the clinical trial results were maintained over time. MIT prolonged the number of years free of diabetic complications (for example 2 years for progression of retinopathy and 2.2 years for progression of clinical neuropathy). Associated total treatment costs over the ten-year period were also calculated and the MIT cost was shown to be $1,233 (1999 prices) less expensive than the CIT therapy due to the greater cost offsets achieved. On this basis the authors conclude that intensive treatment appears to be justified.

A study of a more comprehensive treatment package is given by Eastman et al (1997, 1997). This is based on a simulation model which considered an incident population cohort of type 2 diabetic

patients with disease progression modelled over lifetime focussing on fourteen disease stages encompassing microvascular and macrovascular disease. The data used to populate the model were taken from various US epidemiological studies. Glycaemic control was introduced through a one-off reduction in incidence rates and then risk gradients, taken from the Diabetes Control and Complications Trial study in type 1 diabetes (DCCT), were applied over time. Standard care costs were based on treatment patterns prior to the DCCT study and comprehensive care was based on the treatment patterns associated with two previously undertaken clinical trials. The measure of effectiveness was defined to be a QALY with life years estimated by the model and weighted by quality of life weights obtained from the literature. The results of the model suggest that comprehensive care reduces end-stage microvascular and neuropathic conditions by 67 to 87 per cent. Under baseline assumptions cardiovascular disease increased by 3 per cent as glycaemic control was assumed not to affect this disease. The incremental cost per QALY under comprehensive therapy was estimated to be $17,809 (1999 prices). This was sensitive to age at onset of disease but remained below $55,000 per QALY if age at onset was less than 50. For the cohort developing diabetes at 75 years of age the cost per QALY rose considerably to $248,844. Ethnicity also had a marked impact on the incremental cost-effectiveness ratio and when the rate of renal failure in the standard care arm decreased the cost-effectiveness ratio increased as cost-offsets in this arm were lower (a 25 per cent reduction increased the cost per QALY to $23,535). Consequently, the use of ACE inhibitors in reducing renal failure (with no concomitant impact on cardiovascular disease) while effective and inexpensive increased the cost per QALY further to $26,142. If glycaemic control is assumed to affect cardiovascular disease (cardiovascular disease attributable to diabetes falling by 20 per cent per 10 per cent reduction in HbA_{1C}) the cost per QALY falls to $13,097. The general conclusions reached are that cost-effectiveness is greatest in the younger age groups even though treatment duration is longer and in minority populations. These conclusions reflect the level of effectiveness and cost-offsets to be gained over time through early detection and targeting of high risk sub-populations. The authors stress the conservative nature of their results particularly with regards to cardiovascular disease. Notwithstanding these conservative assumptions most sensitivity analyses show comprehensive therapy to be "in the range of interventions generally considered to be cost-effective" (Eastman et al, 1997).

By far the greatest improvement in knowledge concerning the treatment and progression of type 2 diabetes has come from the UK Prospective Diabetes Study (UKPDS 33, 1998). As described in chapter 1, this was a prospective randomised control trial in which 5102 newly diagnosed type 2 diabetic patients were followed over a median follow-up of ten years. After initial dietary treatment 4209 patients with baseline fasting plasma glucose concentrations of 6.1 to 15 mmol/l who had no symptoms of hyperglycaemia entered the trial. Of these, 342 overweight patients were randomised to metformin, with the remainder (3867) entering the main randomisation and allocated either to conventional policy (mainly diet, 1138) or intensive policy with insulin (1156) or sulphonylureas (1573). Conventional policy had the aim of maintaining patients free of diabetic symptoms and with fasting plasma glucose concentration below 15 mmol/l, while intensive policy had the aim of maintaining fasting plasma glucose concentration below 6 mmol/l. The trial results showed that although baseline levels of fasting plasma glucose concentration increased over time in all groups, intensive therapy significantly reduced the risk of any diabetes related end-point by 12 per cent. The trial did not show a statistically significant difference in diabetes related deaths or all cause mortality over the study period.

The results of the economic evaluation based on the main randomisation data reported the incremental cost per diabetes event free year gained, as mortality differences were not statistically significant (UKPDS 41, 2000). Time to first diabetes related endpoint formed the basis of this measure. A competing risk simulation model was used to predict the time to first event for those individuals who had not experienced a diabetes related endpoint over the trial period. Measurements of direct provider cost were based on the trial data and these were supplemented with predicted

costs for non-inpatient resource use derived from regression analysis based on a cross-sectional questionnaire obtained from trial participants. The resultant cost effectiveness ratios were accompanied by confidence interval estimates based on Fieller's method and were compared with bootstrap estimates for consistency given the skewed nature of the cost data. While some of the individual components of cost were found to be statistically significantly different between the intensive and conventional policies, the overall difference in total treatment costs was not statistically significant. When trial protocol driven costs were replaced with estimates of costs likely to be incurred in a standard clinical practice setting, the incremental cost effectiveness ratio was estimated to be $1,458 (1999 prices) with both costs and effects discounted at the 6 per cent rate as recommended by the UK government. Overall the intensive therapy appeared to be cost-effective over a range of assumptions. Acceptability curves reported that there was a 10 per cent probability that the intensive therapy was cost saving, a 50 per cent probability that the cost per diabetes related endpoint free year lay below (or above) $1,458, and an 80 per cent probability that the ratio was less than $3,132.

As noted above the UKPDS included a number of overweight patients, 342 of whom were further randomised to metformin and 411 were allocated to conventional policy achieved primarily through diet (UKPDS 51, 2000). In the conventional policy group the aim was to achieve the lowest possible fasting plasma glucose level attainable with diet alone and in the intensive policy group the aim was a fasting plasma glucose level of less than 6 mmol/l achieved by increasing dosage of metformin (from 500 to 2550 mg per day) as required. If fasting plasma glucose concentration became greater than 15mmol/l or hyperglycaemic symptoms developed on the maximum tolerated dosage of metformin then glibenclamide was administered. If hyperglycaemia persisted, insulin therapy was initiated. Resource use data were taken from the trial but missing in-patient records for approximately 16 per cent of the cases meant that length of hospital stay was imputed for these patients. As in the main UKPDS economic evaluation non-inpatient resource use was estimated from a regression model based on a cross-sectional questionnaire data. The outcome measure was life years gained based on the differences in mortality recorded within the trial and on a model that simulated life expectancy in those individuals who were still alive at the end of the trial assuming that hazard rates were the same between the groups. While the metformin group patients were associated with higher treatment costs they also experienced greater cost-offsets mainly due to shorter length of hospital stay. This coupled with a gain in life expectancy of one year in the metformin group, resulted in this therapy being effective and cost saving on average. The results also reported the incremental cost-effectiveness ratio using acceptability curves and showed that for the metformin policy there was a 70 per cent probability that it is cost-saving and a greater than 95 per cent probability that the cost-effectiveness ratio is less than $2,259. The reduction in costs of complications will therefore in most cases outweigh the higher treatment costs.

The UKPDS had a further randomisation trial embedded within it. The Hypertension in Diabetes Study (HDS) randomised a total of 1148 type 2 diabetic hypertensive patients to tight or less tight blood pressure control. The aim of the tight control policy was to achieve blood pressure <150/<85 mm Hg using the ACE inhibitor captopril (25 mg twice daily increasing to 50 mg twice daily if required) or the β-blocker atenolol (50 mg daily increasing to 100 mg daily if required). The aim of the less tight control policy was initially to achieve blood pressure ≤200/105 mm Hg, which was modified in 1992 to <180/<105mm Hg after publication of other clinical trial findings in non-diabetic hypertensive patients. Clinical data were analysed as in the main economic evaluation of the UKPDS and years free of diabetes related endpoint was the defined outcome measure (UKPDS 40, 1998). Life years gained were also calculated based on a parametric model which predicted the hazard rate for fatal and non-fatal cardiac events, the hazard for fatal and non-fatal strokes and the hazard for all other deaths. Incremental cost-effectiveness ratios and associated confidence intervals, calculated in the same manner as the main UKPDS analysis, were reported and acceptability curves were presented. There was no significant difference in the total cost between

the tight and the less tight blood pressure control policies even when standard practice resource use patterns replaced trial protocol driven patterns. The incremental cost per additional year free of diabetes related endpoints was $1,312 (costs and effects discounted at 6 per cent; 1999 prices) and the incremental cost per life year gained was $900. The acceptability curve associated with the incremental cost per additional year free of diabetes related endpoints showed a 33 per cent probability that tight blood pressure control was cost saving, a 50 per cent probability that the ratio lay below $1,312, and a two per cent probability that it was not cost effective. Similar results were reached for the cost per life year gained.

A related study by the same investigators assessed the incremental cost effectiveness of the ACE inhibitor captopril versus the β-blocker atenolol as the two competing tight blood pressure control policies in hypertensive type 2 diabetic patients (UKPDS 54, 2001). This study was based on the main Hypertension in Diabetes Study and performed the secondary comparisons between the two blood pressure control agents assessing patients on the ACE inhibitor (400 patients) and those on β-blocker (358 patients). A similar methodological approach as adopted in the main hypertension study was undertaken with costs calculated in a similar manner and life years gained estimated through the simulation model mentioned above. The results showed that 66 per cent of patients receiving the β-blocker received additional glucose lowering treatment four years after randomisation compared to 53 per cent in the ACE inhibitor group and that a difference was maintained over time. This led to the mean cost of antidiabetic drugs in the β-blocker group being higher over time offsetting the higher cost of the ACE inhibitor. There was also a higher cost of hospitalisations in the ACE inhibitor group leading to overall treatment costs being statistically significantly higher for ACE inhibitors than for β-blocker. There was no statistical difference in outcomes with the β-blocker therapy performing slightly better than the ACE inhibitor. Given no difference in effect but a significant difference in costs of $1201 per patient (1999 prices), the results of the cost-effectiveness analysis favoured β-blocker over ACE inhibitor as the preferred tight blood pressure control policy.

Economic evaluation of complications associated with type 2 diabetes

As stated previously the literature concerned with the cost-effectiveness of interventions aimed at diabetic complications in general has not differentiated between type 1 and type 2 diabetes. A broad conclusion is that screening and treatment of complications is cost-effective. The results reported below relate to studies which have either specifically targeted the type 2 diabetic population or have reported results within this population. As mentioned earlier, this part of the review is not comprehensive but rather highlights general findings using specific studies as examples.

There has been some controversy over the cost-effectiveness of screening for diabetic retinopathy. Early studies were criticised for utilising sub-optimal screening technologies and not using opportunistic screening as a comparator (Buxton et al, 1991). A recent UK study assessed the incremental cost-effectiveness of a systematic screening programme compared to an opportunistic programme (James et al, 2000). This analysis was based on an earlier clinical study and the incremental cost-effectiveness was calculated as the additional cost required to generate each additional true positive case identified after replacing the opportunistic programme with the systematic one. The base case results report an incremental cost-effectiveness of complete replacement of the opportunistic programme as being $39 (1999 prices). One of the controversies underlying the studies is the baseline prevalence. This study assumed a prevalence of approximately 14 per cent while earlier studies had assumed prevalence rates to be about one third of this. This is important as prevalence will fall to the incidence rate over time with a systematic screening programme. The study reported that while the incremental cost-effectiveness ratios rose over time, the systematic programme was always more cost-effective than the opportunistic programme. These

results were similar to the results of an earlier US study (Lairson et al, 1992). Javitt and Aiello (1996) used an epidemiological model to extrapolate screening results into treatment effectiveness and subsequently into the cost per QALY. Ophthalmic screening and treatment for type 2 diabetic patients ranged from $3,198 to $3,849 (1999 prices) per QALY compared to no treatment depending on whether insulin was used or not. A similar criticism may be levelled at this study, that is that the true alternative ought to be opportunistic screening. These results suggest that systematic screening alone as well as a combined screening and treatment programme are relatively cost-effective. A Dutch study by Crijns et al, 1999 considered the optimal timing of screening intervals using a simulation model to calculate the marginal cost of no screening versus screening once a year if diabetic retinopathy is diagnosed, twice a year if macular oedema is diagnosed and four times a year if proliferative diabetic retinopathy is diagnosed. Three further scenarios decreased the frequency of screening. Direct and indirect costs were considered for both type 1 and type 2 diabetic patients. The results relating to type 2 diabetes showed that the indirect costs were of little consequence while the direct costs per year of realised sight gained were considered low compared to other interventions with optimal initiation of screening associated with the youngest age cohort considered (35 years of age). Generally the results were similar to those gained by Javitt and Aiello (1996) and support that screening is cost-effective relative to other interventions.

It is estimated that between 5 and 15 per cent of diabetic patients have ulceration of their feet with approximately 1 to 20 per cent requiring amputation (Ragnarson-Tennvall et al, 1997; Krentz et al, 1997; Ollendorf et al, 1998). Difficulty is encountered in defining precisely some of the conditions that lead to diabetic foot disorders, neuropathy in particular. It has been estimated that some 50 per cent of amputations are avoidable and one study has suggested that considerable savings may be attained through the adoption of prevention programmes (Ollendorf et al, 1998). The costs of treatment, particularly for amputation, (van Houtum, 1995) are known to increase considerably as complexity increases. Panayiotopoulos et al (1997) showed that there was no statistically significant difference in surgical outcome between diabetic and non-diabetic patients, although the diabetic patients tended to perform worse. In contrast, the diabetic population undergoing surgery had significantly higher costs compared to the non-diabetic population. Total cost for the diabetic population was $12,823 (1999 prices) compared to $8,868 for the non-diabetic population. With regards to less severe diabetic foot infections Eckman et al (1995) found little difference in the cost-effectiveness across a range of strategies.

Recently interest has turned to the analysis of treatment of diabetic patients at risk of coronary heart disease. This interest stems not only from the high risk of CHD in the diabetic population but also from the sub-group analysis of the diabetic population undertaken within the Scandinavian Simvastatin Survival Study (Herman et al, 1999). In this analysis three sub-groups were identified: those with normal fasting glucose, those with impaired fasting glucose and those with diabetes. Each group was analysed comparing the simvastatin arm to a placebo with direct health care resource utilisation defining the outcome measure. While the study did not differentiate between type 1 and type 2 diabetes the impaired fasting glucose group can be considered representative of a mild to moderate type 2 diabetic population. Concentrating on the findings for this group, the study reports that cardiovascular hospitalisations were reduced by 30 per cent in the simvastatin group compared to placebo (comparable reduction was 23 per cent in the normal fasting glucose group), and length of hospital stay was reduced by 38 per cent (compared to 28 per cent in the normal fasting glucose group). On average the impaired fasting glucose group showed a decrease in hospital costs of $4,600 (1999 prices) which offset 74 per cent of the cost of simvastatin. Indeed there was a net cost saving for the diabetic sub-group in this respect. Building on this study, but using a modelling approach, Grover et al (2000) considered the cost-effectiveness of simvastatin use in primary prevention by comparing a diabetic population without symptomatic CVD to a non-diabetic population who had CVD. Interest was in whether the treatment of primary prevention patients at high risk was as cost-effective as secondary prevention. The cost-effectiveness ratio was

similar in male diabetic patients without CVD to CVD males without diabetes. These results held across several countries, largely because the cost offsets were low relative to the treatment costs. Grover et al (2001) also showed similar results when they stratified by different LDL cholesterol levels.

Banz et al (1998) analysed the impact that bodyweight gain in type 2 diabetic patients had on the development of CHD. They used data from a clinical trial within a decision analytic approach to predict the rate of CHD in individuals with weight gain of less than two kilograms over a ten-year period and with weight gain greater than two kilograms over the same period. The results predicted a significantly lower rate of CHD (30.3%) in individuals with relatively stable bodyweight compared to those with bodyweight gain greater than two kilograms over the study period (CHD rate was predicted to be 38.2%). The study compared further first-line therapy use of glibenclamide, which has the side-effect of increasing bodyweight, with acarbose, which does not increase bodyweight. Their findings suggest that although acarbose is four times more expensive than glibenclamide in the environment studied (Germany), approximately one-third of this increase was offset by the lower CHD event rate associated with acarbose. The authors state that this is a tentative finding as all bodyweight changes were attributed to the drug therapy.

There is limited evidence of cost-effectiveness in the treatment of type 2 diabetic patients for nephropathy. One study has extended the modelling of type 1 diabetic patients treated with captopril to type 2 diabetic patients (Rodby et al, 1996). The study reported that this treatment resulted in direct cost savings over lifetime. A study by Golan et al (1999) assessed the cost-effectiveness of screening for gross proteinuria and microalbuminuria, both assumed to be predictors of diabetic nephropathy, as well as the use of ACE-inhibitors. The study used a Markov model which simulated the progression of diabetic nephropathy. The screening strategies were shown to have higher cost and lower benefit compared to treating all patients with ACE-inhibitors.

Conclusions

As shown in the papers reviewed the costs associated with the treatment of type 2 diabetes increase over the lifetime of any individual patient. Consequently the cost-offsets arising from delaying the progression of the disease and the development of diabetic complications are substantial. Even though the more costly diabetic complications such as chronic renal failure are rare, even the less costly complications incur substantial expenditure given their intrusive nature. Although all these studies have contributed to the general knowledge of disease progression and associated complications to differing degrees with the UKPDS providing the most recent and extensive information, there is still substantial ground to be covered in reaching perfect understanding of the disease itself and in identifying optimal treatment patterns. Even the UKPDS after 11 years of average follow-up did not show a statistically significant difference in mortality between the two randomisation groups. A direct consequence of the lack of adequate clinical information is the lack of adequate information on the cost impact of the disease which will only be accentuated by adopting inappropriate analytical techniques or by ignoring specific data problems in the analysis.

Table A.2.1.1. Cost of illness studies

Author	Brown, Nichols, Glauber et al	Brown, Pedula and Bakst	Evans, MacDonald, Leese et al	Henriksson, Agardh, Berne, et al	O'Brien, Shomphe, Kavanagh et al
Year of Publication	1999	1999	2000	2000	1998
Year used for cost valuation	1993	1993	1995	1998	1996
Country where analysis occurred	USA	USA	UK	Sweden	USA
Currency used for cost valuation	US $	US $	UK £	Swedish Kroner	US $
Methodology	Excess cost-of-illness	Cohort study	Cost-of-illness	Cost-of-illness	Cost-of-illness
Alternatives considered	n.a.	No alternatives	n.a	n.a	n.a.
Cost-effectiveness measure	n.a.	n.a.	n.a	n.a	n.a.
Patient population	HMO matched newly diagnosed Type 2 diabetic and non-diabetic patients	HMO Type 2 diabetics over 30 years of age	Prevalent diabetic population	Type 2 diabetics from selected Swedish health care centres	Type 2 diabetics suffering complications
Effectiveness data sources	n.a.	n.a.	n.a	n.a	n.a.
Cost elements	Direct health care costs	Annual treatment costs; costs of treating cardiovascular and renal complications	Drug prescribing costs	Direct medical costs	Direct health care costs
Cost data sources	Internal HMO cost calculations. Charges for out-of-HMO direct care. Discharge data bases for prevalence of treatment	HMO expenditures on in-patient stays, out-patient stays and procedures.	Internal Health Authority costs	Health care expenditures and questionnaires from patients and practitioners on resource utilisation	Discharge databases, clinical guidelines, government reports, fee schedules, and literature review. Cost-to-charge ratios used.
Time horizon	up to 8 years	average 5.3 years follow-up	one year	6 months factored up to 1 year	Defined by the episodic event
Discount rate	n.a.	n.a.	n.a	n.a	not reported
Variables considered in the sensitivity analysis	n.a.	n.a.	n.a	n.a	

Table A.2.1.1. Cost of illness studies (Contd.)

Author	Brown, Nichols, Glauber et al	Brown, Pedula and Bakst	Evans, MacDonald, Leese et al	Henriksson, Agardh, Berne, et al	O'Brien, Shomphe, Kavanagh et al
Baseline results	Incremental per annum costs averaged $2,930 higher for the Type 2 diabetic group than the non-diabetic group over the 8 year period. Hospital costs were the highest component (46%). Primary and out-patient care accounted for an average of 26% of the costs and drugs 20%. Total costs for the diabetic population were approximately double those for the control population throughout the period.	$2,263 total average annual treatment cost with no complications; increasing to $3,472 with minor cardiovascular treatment; and to $8,235 with major cardiovascular complication. Total average annual treatment cost for Type 2 patients with abnormal renal function was $3,750; $4,428 for those with advanced renal disease; and $17,445 with end-stage renal disease	Patients with Type 2 diabetes accounted for 6.6% of total prescriptions dispensed, representing 7.1% of the cost in the Health Region (5.5% excluding antidiabetic prescriptions). Higher proportionate drug costs in nearly all drug categories ranging from 2.6% higher (endocrine system) to 10.8% higher (cardiovascular). Type 2 diabetics were 1.70 times more likely to be dispensed a drug item.	Overall direct costs of treatment were $613 million. 42% were borne by the hospital sector; ambulatory care was 31%; drug costs 27%; insulin was approximately 4% of the total cost	AMI event cost $28,920 (subsequent annual costs $2,983); Angina event cost $2,592 (subsequent annual costs $1,133); Ischemic stroke event cost $42,513 (sub. annual cost $9,687); TIA event $6,494 (sub. annual $47); microalbuminuria event $65 (sub. annual cost $15); gross proteinuria event $72 (sub. annual cost $24); end-stage renal dialysis annual cost $56,164;Background retinopathy annual cost $59; macular edema event $1,151 (sub. annual cost $59); proliferative diabetic retinopathy event $1,092 (sub. annual cost $59); blindness annual cost $3,684; Symptomatic neuropathy event $228; 1st LEA event cost $28,150 (sub. annual cost $1,820); 2nd LEA event cost $28,389)
Results from sensitivity analysis				Patient with both macro- and micro-vascular complications had approximately double the costs of those without complications	n.a.

Table A.2.1.1. Cost of illness studies (Contd.)

Author	Brown, Nichols, Glauber et al	Brown, Pedula and Bakst	Evans, MacDonald, Leese et al	Henriksson, Agardh, Berne, et al	O'Brien, Shomphe, Kavanagh et al
Author(s) conclusions	More than 60% of the total cost in hospital admissions and most of the annual growth in costs over the 8 year period was attributable to admissions not normally associated with diabetes. Hospital admissions for acute complications of diabetes and for renal, lower-extremity, ophthalmic, hypoglycaemic and infectious complications accounted for 13% of incremental hospital costs. Cardiac disease contributed 17% to total incremental hospital costs. The low level of costs compared to other studies was taken as representative of using costs rather than charges. It was anticipated that costs would increase dramatically over time.	Renal complications most expensive, but occur in only 23% of Type 2 population compared to cardiovascular complications (75% of the population). Women have significantly higher costs. Age does not affect the cost of treating complications. If co-morbidity of diabetes is not taken into account there is an under-estimation of cost.	Increased drug utilisation is high in Type 1 diabetics compared to Type 2 diabetics. Type 2 diabetics are higher absolute users as they are older, and the prevalence is greater. Increased drug utilisation even in drug categories unrelated to diabetes.		

Table A.2.1.2. Cost consequence and cost effectiveness of screening and treatments for type 2 diabetes

Author	Eastman, Javitt, Herman et al.	Engelau, Narayan, Thompson et al	Franz, Splett, Monk et al	Golan, Birkmeyer and Welch	Segal, Dalton and Richardson	UKPDS 40	UKPDS 41	UKPDS 51	UKPDS 54	Wake, Hisashige, Katayama et al
Year of Publication	1997	1998	1995	1999	1998	1998	2000	2001	2001	2000
Year used for cost valuation	1994	1995	1993	1996	1997	1997	1997	1997	1997	1998
Country where analysis occurred	USA	USA	USA	USA	Australia	UK	UK	UK	UK	Japan
Currency used for cost valuation	US $	US $	US $	US $	Aus $	UK £	UK £	UK £	UK£	US $
Methodology	Model	Monte Carlo simulation model	Clinical trial	Markov model	Markov model	Economic evaluation alongside clinical trial	Economic evaluation alongside trial	Economic evaluation alongside trial	Economic evaluation alongside clinical trial	Cost consequence analysis
Alternatives considered	Conventional therapy versus intensive therapy	Opportunistic screening for diabetes versus current practice	Basic guidance on dietary advice versus practice guideline advice on diet	Screen for gross proteinuria; Screen for microalbuminuria; treat all with ACEi	Intensive diet and behavioural modification; surgery for severe obesity; group behavioural modification for men; General practitioner advice; media campaign with community support	Tight and less tight control of hypertensive Type 2 diabetics	Conventional versus intensive glucose control	Conventional therapy (primarily diet) versus intensive therapy with metformin in overweight Type 2 diabetics	Use of Atenolol (beta-blocker) versus captopril (ACEi) in Type 2 diabetic patients with hypertension	Multiple insulin injection therapy; Conventional insulin injection therapy
Cost-effectiveness measure	Incremental cost per QALY	Cost per life years gained; QALYs	Change from baseline fasting plasma glucose level and glycated haemoglobin	Cost per QALY	Life years gained	incremental cost per extra year of life free from diabetic end point, incremental cost per life year gained	incremental cost per event free (of diabetic end point) year gained	Incremental cost per life year gained	Life expectancy and mean cost per patient	n.a.

Table A.2.1.2. Cost consequence and cost effectiveness of screening and treatments for type 2 diabetes (Contd.)

Author	Eastman, Javitt, Herman et al.	Engelau, Narayan, Thompson et al	Franz, Splett, Monk et al	Golan, Birkmeyer and Welch	Segal, Dalton and Richardson	UKPDS 40	UKPDS 41	UKPDS 51	UKPDS 54	Wake, Hisashige, Katayama et al
Patient population	Incident Type 2 diabetics	hypothetical cohort of 10,000 incident Type 2 diabetics	Type 2 diabetics	Incident Type 2 diabetics over 50 years old at risk of renal failure	Hypothetical cohort	Randomised clinical trial population	Randomised clinical trial population	Randomised clinical trial population who had >120% of ideal body weight	Randomised clinical trial population with Type 2 diabetes and hypertension	Randomised clinical trial population
Effectiveness data sources	Incidence data from National Health Interview Survey: complications from Wisconsin Epidemiologic Study of Diabetic Retinopathy and Rochester Epidemiology Study		Clinical trial run in conjunction with the economic evaluation	Transition probabilities: RCT (for Type 1 diabetics) US Renal Data System; Clinical Guidelines. Utilities; Beaver Dam Health Outcomes Study	Survey of clinical trials, observational, epidemiological and intervention studies	Clinical trial outcomes, based on frequency of diabetic related end-points	Clinical trial outcomes, based on frequency of diabetic related end-points	Clinical trial outcomes, based on frequency of diabetic related end-points	Clinical trial outcomes, based on frequency of diabetic related end-points	Clinical trial outcomes, based on frequency of diabetic related end-points
Cost elements	Standard direct therapy costs: pharmacy costs, hospital costs,self-monitoring blood glucose	Screening and treatment cost	Total cost of nutritional care	ACE inhibitor therapy, screening, treatment of end-stage renal disease	Direct health care programme costs	Direct health care programme costs	Direct health care programme costs	Direct health care programme costs	Direct health care programme costs	Direct medical care costs

Table A.2.1.2. Cost consequence and cost effectiveness of screening and treatments for type 2 diabetes (Contd.)

Author	Eastman, Javitt, Herman et al.	Engelau, Narayan, Thompson et al	Franz, Splett, Monk et al	Golan, Birkmeyer and Welch	Segal, Dalton and Richardson	UKPDS 40	UKPDS 41	UKPDS 51	UKPDS 54	Wake, Hisashige, Katayama et al
Cost data sources	National Medical Expenditure Survey; National Health Interview Survey; Three studies (DCCT, Veterns Administration Cooperative Study & Metformin Cooperative Trial)		Clinical trial	Medicare Clinical Diagnostic Fee Schedule; U.S. Renal Database System	Literature survey material, internal Health Authority costs, National Health Survey costs	Resource data from trial, unit costs from national statistics and participating units	Resource data from trial, unit costs from national statistics and participating units	Resource data from trial, unit costs from national statistics and participating units.	Resource data from trial, unit costs from national statistics and participating units.	Internal National Health care costs
Time horizon	Lifetime	lifetime	6-month study period	Lifetime	lifetime	11 years and lifetime	11 years	Lifetime	Lifetime	10 years
Discount rate	3%	3%	n.a.	3%	5%	6%, 3% for costs and benefits and 0% for benefits reported	6%	6%	6%	3%
Variables considered in the sensitivity analysis	Risk of complications; discounting	Risk factors	95% c.i. for outcomes; inclusion/exclusion of laboratory test	Age at diagnosis; relative risk of disease progression;costs; quality of life; screening adherence; treatment discontinuation	Effectiveness rates; discount rate; impact of preventative programme on incidence of NIDDM; life expectancy; high risk group	Protocol driven resourcing changed to likely standard practice. Standard practice pattern and related unit costs of visits and tests. Confidence intervals and acceptability curves reported.	Protocol driven resourcing changed to likely standard practice. Discount rate varied. Confidence intervals and acceptability curves reported.	Increase in therapy costs. Different regression model applied to non-inpatient costs. Confidence intervals and acceptability curve reported.	Cost of standard practice. Cost of captopril. Non-hospital costs.	Relative risks in progression, costs, discount rate

Table A.2.1.2. Cost consequence and cost effectiveness of screening and treatments for type 2 diabetes (Contd.)

Author	Eastman, Javitt, Herman et al.	Engelau, Narayan, Thompson et al	Franz, Splett, Monk et al	Golan, Birkmeyer and Welch	Segal, Dalton and Richardson	UKPDS 40	UKPDS 41	UKPDS 51	UKPDS 54	Wake, Hisashige, Katayama et al
Baseline results	Baseline cost per QALY $17,809. Sensitive to age of onset of diabetes; ICER < $55,000 if age of onset less than 50. If age of onset >75 the ICER is $248,844. Ethnicity had a major impact as rate of complications differ. Sensitive to rate of renal failure. Sensitive to compliance rates.	Incremental cost of opportunistic screening of individuals aged 25 or older was estimated to be $252,192 per life-year gained and $60,420 per QALY.	Average cost per mg/dl change in fasting blood glucose level is $5.92 in the basic dietary information group and $4.67 in the practice guidelines nutrition care group. Difference statistically insignificant.	$7,850 per QALY	All primary prevention programmes had low cost-effectiveness ratios. Surgery yielded the largest reduction in the number of diabetic years, but was the most expensive intervention at $5,894 per life year gained; certain programmes were cost saving (intensive diet & behavioural therapy in the seriously obese impaired glucose tolerance population; behavioural therapy for overweight men, & the media campaign plus community support in a mixed population).	Based on standard practice resource use incremental cost per extra year free from end points was $1312 (6% discounted costs and benefits); $543 (costs discounted at 6% and effects undiscounted). Incremental life year gained was $900 (6% discounted costs and benefits); $364 (6% discounted costs and benefits).	Based on standard practice resource use incremental cost per event free year was $1,458 (6% discounted costs and benefits) and $703 (costs discounted at 6% and effects undiscounted).	Therapy with metformin is cost-saving under a range of assumptions	No statistical difference in life expectancy. Beta-blocker (atenolol) was less expensive with an average treatment cost 14% lower than ACEi.	Multiple insulin injection therapy (MIT) reduced the relative risk of progression to a diabetic end-point for retinopathy (by 76%); photocoagulation (by 77%); nephropathy (by 66%); albuminuria (100%) and clinical neuropathy (64%). MIT also prolonged complication free time. MIT had a lower ten-year cost of therapy than conventional insulin therapy due to the reduced costs of treating complications - MIT $1,233 less expensive

Table A.2.1.2. Cost consequence and cost effectiveness of screening and treatments for type 2 diabetes (Contd.)

Author	Eastman, Javitt, Herman et al.	Engelau, Narayan, Thompson et al	Franz, Splett, Monk et al	Golan, Birkmeyer and Welch	Segal, Dalton and Richardson	UKPDS 40	UKPDS 41	UKPDS 51	UKPDS 54	Wake, Hisashige, Katayama et al
Results from sensitivity analysis		Opportunistic screening is most cost-effective in younger age groups and ethnic minorities as these groups have a higher lifetime risk of major diabetic complications.	Results extremely sensitive to outcome level assumed to be achieved	> $21,000 per QALY if age at diagnosis 55 or over; or if ACEi treatment cost increased by one-third; or relative risk of microalbiminuria increased	Net cost per life year saved is sensitive to assumptions of programme success and discount rate. However prevention remains below $5,512 per life year gained in all cases.	Acceptability curves presented	Acceptability curves presented	Cost-saving findings robust to a range of sensitivity analyses	Results robust to sensitivity analysis	
Author(s) conclusions	Cost per QALY is lowest for those at greatest risk of complication; and therefore for ethnic minorities and those with higher HbA1c. Cost-effectiveness of comprehensive care of diabetes appears similar to other preventative treatments.	Early diagnosis & treatment by opportunistic screening of type 2 diabetes increases costs but could reduce the lifetime incidence of major microvascular complications generating gains in health benefits. The selection of target populations for opportunistic screening should consider risk factors as well as disease prevalence.			Programmes for the prevention of NIDDM are stated to be highly cost-effective relative to other funded health care programmes. Access to such programmes should be increased for the sub-populations at highest risk of NIDDM	Evidence that tight control of blood pressure for Type 2 hypertensive patients is a cost-effective means of reducing complications and improving health outcomes.	Intensive blood glucose control in patients with Type 2 diabetes increased treatment costs but these were offset by reduced costs of treating complications and increased time free of complications. Intensive therapy of Type 2 diabetics is feasible and economically supportable.	Cost-saving is induced largely through lower hospital in-patient costs.	UKPDS supports the use of either captopril or atenolol on clinical grounds with the latter being the less expensive therapy.	Intensive glycaemic control can delay onset and progression of the early stages of microvascular complications in Type 2 diabetics. This reduces treatment costs and, therefore, MIT is a recommended strategy in those Type 2 patients requiring insulin.

References relating to the review of the economic literature on type 2 diabetes

Alberti, K., 1993. Problems related to definitions and epidemiology of type 2 (non-insulin dependent) diabetes mellitus: studies throughout the world. Diabetologia 36, 978-984.

Alberti, K., 1997. The costs of NIDDM. Diabetic Medicine 14, 7-9.

Alwin, A., King, H., 1995. Diabetes in the East Mediterranean region. Diabetic Medicine 12, 1057-1058.

Banz, K., Dinkel, R., Hanefeld, M. et al, 1998. Evaluation of the potential clinical and economic effects of bodyweight stabilisation with acarbose in patients with type 2 diabetes mellitus. Pharmacoeconomics 13, 449-459.

Brown, J., Nichols, G., Glauber, H., et al, 1999. Type 2 diabetes: incremental medical care costs during the first 8 years after diagnosis. Diabetes Care 22, 1116-1124.

Brown, J., Pedula, K., Bakst, A., 1999. The progressive cost of complications in Type 2 diabetes mellitus. Archives of Internal Medicine 159, 1873-1880.

Buxton, M., Sculpher, M., Ferguson, B., et al., 1991. Screening for treatable diabetic retinopathy: a comparison of different methods. Diabetic Medicine 8, 371-7.

Clarke, P., Gray, A., Adler, A., et al, 2001. Cost-effectiveness analysis of intensive blood-glucose control with metformin in overweight patients with Type II diabetes (UKPDS 51). Diabetologia 44, 298-304.

Crijns, H., Casparie, A., and Hendrikse, F., 1999. Continuous computer simulation analysis of the cost-effectiveness of screening and treating diabetic retinopathy. International Journal of Technology Assessment in Health Care 15, 198-206.

Deerochanawong, C., 1992. A survey of lower limb amputation in diabetic patients. Diabetic Medicine 9, 942-964.

Eastman, R., Javitt, J., Herman, W., et al, 1997. Model of complications of NIDDM: I Model construction and assumptions. Diabetes Care 20, 725-734.

Eastman, R., Javitt, J., Herman, W., et al, 1997. Model of complications of NIDDM: II Analysis of the health benefits and cost-effectiveness of treating NIDDM with the goal of normoglycemia. Diabetes Care 20, 735-744.

Eckman, M., Greenfield, S., Mackey, W., et al, 1995. Foot infections in diabetic patients: decision and cost-effectiveness analysis. Journal of the American Medical Association 273, 712-720.

Engelgau, M., Venkat Narayan, K., Thomson, T., et al, 1998. The cost-effectiveness of screening for Type 2 diabetes. Journal of the American Medical Association 280, 1757-1763.

Evans, J., MacDonald, T., Leese, G., et al., 2000. Impact of Type 1 and Type 2 diabetes on patterns and costs of drug prescribing. Diabetes Care 23, 770-774.

Franz, M., Splett, P., Monk, A., et al, 1995. Cost-effectiveness of medical nutrition therapy provided by dieticians for persons with non-insulin-dependent diabetes mellitus. Journal of the American Dietetic Association 95, 1018-1024.

Golan, L., Birkmeyer, J., and Welch, G., 1999. The cost-effectiveness of treating all patients with Type 2 diabetes with angiotensin-converting enzyme inhibitors. Annals of Internal Medicine 131, 660-667.

Gray, A., Raikou, M., McGuire, A., et al, 2000. Cost-effectiveness of an intensive blood glucose control policy in patients with type 2 diabetes: economic analysis alongside randomised controlled trial (UKPDS 41). British Medical Journal 320, 1373-1378.

Gray, A., Clarke, P., Raikou, M., et al, 2001. An economic evaluation of atenolol versus captopril in patients with Type 2 diabetes (UKPDS 54). Diabetic Medicine 18, 438-444.

Grover, S., Coupal, L., Zowall, H., et al, 2000. Cost-effectiveness of treating hyperlipidemia in the presence of diabetes. Who should be treated? Circulation 102, 722-727.

Grover, S., Coupal, L., Zowall, H., et al, 2001. How cost-effective is the treatment of dyslipidemia in patients with diabetes but without cardiovascular disease. Diabetes Care 24, 45-50.

Gulliford, M., 1997. Design of cost-effective packages of care for non-insulin dependent diabetes mellitus. International Journal of Technology Assessment in Health Care 13, 395-410.

Harris, M., 1996. Medical Care for patients with diabetes: Epidemiological aspects. Annals of Internal Medicine 124, 117-127.

Henriksson, F., Agardh, C., Berne, C., et al, 2000. Direct medical costs for patients with type 2 diabetes in Sweden. Journal of Internal Medicine 248, 387-396.

Herman, W., Alexander, C., Cook, J., et al, 1999. Effect of simvastatin treatment on cardiovascular resource utilisation in impaired fasting glucose and diabetes. Diabetes Care 22, 1771-1778.

Houtum, van W., Lavery, L., and Harkless, L., 1995. The costs of diabetes related lower extremity amputations in the Netherlands. Diabetic Medicine 12, 777-781.

James, M., Turner, D., Broadbent, D., et al, 2000. Cost-effectiveness analysis of screening for sight threatening diabetic eye disease. British Medical Journal 320, 1627-1631.

Javitt, J. and Aiello, L., 1996. Cost-effectiveness of detecting and treating diabetic retinopathy. Annals of Internal Medicine 124, 164-169.

Jonsson, B., Krans, H., 1995. The Social and Cost Implications of Type II Diabetes. PharmacoEconomics (Supplement) 8, s1-s94.

King, H., Rewers, M., 1993. Global estimates for prevalence of diabetes mellitus and impaired glucose tolerance in adults. Diabetes Care 16, 157-177.

Krentz, A., Acheson, P., Basu, A., et al, 1997. Morbidity and mortality associated with diabetic foot disease: a 12 month prospective survey of hospital admissions in a single UK centre. The Foot 7, 144-147.

Lairson, D., Puigh, J., Kapadia, et al, 1992. Cost-effectiveness of alternative methods for diabetic retinopathy screening. Diabetes Care 15, 1369-1377.

Marks, L., 1995. Counting the Cost: the Real Impact of Non-Insulin Dependent Diabetes. King's Fund Policy Institute, London.

O'Brien, J., Shomphe, L., Kavanagh, P. et al, 1998. Direct medical costs of complications resulting from Type 2 diabetes in the U.S. Diabetes Care 21, 1122-1128.

Ohkubo, Y., Kishikawa, H., Araki, E., et al, 1995. Intensive insulin therapy prevents the progression of diabetic microvascular complications in Japanese patients with NIDDM: a randomised prospective 6-year study. Diabetes Research in Clinical Practice 28, 103-117.

Ollendorf, D., Kotsanos, J., Wishner, W., et al, 1998. Potential economic benefits of lower extremity amputation prevention strategies in diabetes. Diabetes Care 21, 1240-1245.

Panayiotopoulos, Y., Tyrell, M., Arnold, F., et al, 1997. Results and cost analysis of distal (crural/pedal) arterial revascularisation for limb salvage in diabetic and non-diabetic patients. Diabetic Medicine 14, 214-220.

Ragnarson-Tennvall, G., Apelqvist, J., 1997. Cost-effective management of diabetic foot ulcers. Pharmacoeconomics 12, 42-53.

Rodby, R., Firth, L., Lewis, E., et al, 1996. An economic analysis of captopril in the treatment of diabetic nephropathy. Diabetes Care 19, 1051-1061.

Segal, L., Dalton, A., Richardson, J., 1998. Cost-effectiveness of the primary prevention of non-insulin dependent diabetes mellitus. Health Promotion International 13, 197-208.

UK Prospective Diabetes Study Group, 1998. Intensive blood glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). The Lancet 352, 837-853.

UKPDS Group, 1998. Cost-effectiveness analysis of improved blood pressure control in hypertensive patients with type 2 diabetes (UKPDS 40). British Medical Journal 317, 720-726.

Wake, N., Hisahige, A., Katayama, et al, 2000. Cost-effectiveness of intensive insulin therapy for type 2 diabetes: a 10-year follow-up of the Kumamoto study. Diabetes Research in Clinical Practice 48, 201-210.

Waugh, N., 1989. Amputations in diabetic patients – a review of rates, relative risks and resource use. Community Medicine 6, 346-350.

Wood, J., 1990. A review of diabetes initiatives in primary care settings. Health Trends 22, 39-43.

Zimmet, P., McCarthy, D., 1995. The NIDDM epidemic: global estimates and projections – a look into the crystal ball. IDF Bulletin 40, 8-16.

Appendix A.4.1. Consistency of the Bang and Tsiatis simple weighted estimator

$$\hat{\mu}_{WT} = \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{\hat{K}(T_i)}\frac{K(T_i)}{K(T_i)} + \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{K(T_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{K(T_i)}\frac{\hat{K}(T_i)}{\hat{K}(T_i)}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{K(T_i)} + \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i M_i}{K(T_i)}\left\{\frac{K(T_i)-\hat{K}(T_i)}{\hat{K}(T_i)}\right\}$$

The second term is bounded from above by

$$\sup_{u<L}\frac{\left|K(u)-\hat{K}(u)\right|}{K(L)\hat{K}(L)}\max\left|M_i\right|$$

Assuming that total cost is bounded, the upper bound above converges to zero in probability due to the fact that $\sup_{u<L}\left|K(u)-\hat{K}(u)\right| = o(n^{-1/2+\varepsilon})$ with probability one (Csörgö and Horvath, 1983). Hence, consistency follows by the law of large numbers.

Appendix A.4.2. Variance of the Bang and Tsiatis partitioned estimator

$$\mathrm{var}(M_i-\mu) + E\int_0^L\left[\sum_{j=1}^{K}\left\{M_{ij}-G_j(M_j,u)\right\}I(u<t_j)\right]^2 I(T_i\geq u)\frac{\lambda^c(u)}{K(u)}du$$
$$= \mathrm{var}(M_i-\mu) + \int_0^L E\left[\sum_{j=1}^{K}\left\{M_{ij}-G_j(M_j,u)\right\}I(T_i^{t_j}\geq u)\right]^2\frac{\lambda^c(u)}{K(u)}du$$
$$= \mathrm{var}(M_i-\mu) + \int_0^L\sum_{j=1}^{K}\sum_{l=1}^{K}S_{j\wedge l}(u)\left\{G_{j\wedge l}(M_jM_l,u)-G_{j\wedge l}(M_j,u)G_{j\wedge l}(M_l,u)\right\}\frac{\lambda^c(u)}{K(u)}du$$

where the integrand in the second term corresponds to the K variance terms and the pairwise covariance terms among them. Its expansion is shown below.

$$E\left[\sum_{j=1}^{K}\left\{M_{ij}-G_j(M_j,u)\right\}I(T_i^{t_j}\geq u)\right]^2 = \sum_{j=1}^{K}E\left\{M_{ij}-G_j(M_j,u)\right\}^2$$
$$+2\sum_{j=1}^{K}\sum_{l=j}^{K}E\left[\left\{M_{ij}-G_j(M_j,u)\right\}I(T_i^{t_j}\geq u)\times\left\{M_{il}-G_l(M_l,u)\right\}I(T_i^{t_l}\geq u)\right]$$

For $j<l$, the cross-product terms in the summand reduce to

$$E\left[\left\{M_{ij}-G_j(M_j,u)\right\}\left\{M_{il}-G_l(M_l,u)\right\}I(T_i^{t_j}\geq u)\right] = S_j(u)\left\{G_j(M_jM_l,u)-G_j(M_j,u)G_j(M_l,u)\right\}$$

Appendix A.4.3. Descriptive statistics of the UKPDS annual cost data

Table A.4.3.1. Annual cost per patient by year of randomisation for the conventional policy group

Year	*Number of observations*	*Mean annual cost*	*Standard deviation*	*Minimum annual cost*	*Maximum annual cost*
1	1138	580.7277	1477.233	19.8	29020.4
2	1125	573.3312	1150.522	19.8	11602.4
3	1112	774.554	2847.989	19.8	74200.4
4	1097	856.332	4397.077	19.8	131010.7
5	1076	730.7807	1740.947	19.8	24692.35
6	1050	872.7527	2790.088	19.8	64387.83
7	1017	839.9952	2657.49	16.15	60169.1
8	917	832.7916	2225.708	19.8	36255.14
9	816	747.0578	1838.934	19.8	30230.81
10	695	976.0969	3139.52	19.8	57920
11	561	894.3764	1943.828	19.8	22582.6
12	433	934.465	1978.162	77.8	29271.91
13	323	1054.579	3044.994	77.8	35017.7
14	214	852.0495	1931.715	77.8	23846.85
15	129	1264.868	4680.558	77.8	45530.36
16	69	946.3403	2847.115	220.4	23856.83
17	53	1500.837	3253.058	149.1	22122.4
18	32	1222.613	3099.91	243.35	17772.29
19	12	736.0087	951.7124	80.55	3668.684

Table A.4.3.2. Annual cost per patient by year of randomisation for the intensive policy group

Year	*Number of observations*	*Mean annual cost*	*Standard deviation*	*Minimum annual cost*	*Maximum annual cost*
1	2729	567.4993	1375.245	19.8	45151.02
2	2700	622.5739	1241.575	16.88	19759
3	2673	660.4301	1368.677	19.8	30435.51
4	2632	715.3494	1705.962	19.8	34714.98
5	2596	681.6874	1389.174	19.8	30483.65
6	2539	750.5341	1339.22	17.975	22638.41
7	2442	891.7456	2326.112	17.975	43403.86
8	2233	802.6076	1772.513	17.975	44092.96
9	1985	914.5527	2353.804	19.8	45555.75
10	1681	948.5564	2393.213	19.8	39735.35
11	1347	906.1987	1839.467	16.15	24701.2
12	1062	890.46	1724.105	14.325	25616.9
13	818	857.7013	1564.98	16.15	22035.3
14	556	868.8157	1424.465	19.8	14481
15	326	1018.485	2224.572	33.1	28933.1
16	187	977.7186	1679.65	33.1	12498.17
17	127	1296.673	3432.144	19.8	29200.56
18	70	749.9228	854.7124	210.1895	5563.155
19	18	454.16	191.4103	233.7	1048.372
20	2	350.337	164.9496	233.7	466.9739

Appendix A.4.4. Program for the Kaplan-Meier estimator

Based on equations (4.4), (4.5) and (4.6)

Kaplan-Meier estimator

```
**mean**

stset  Mi, failure(censorig==1)
sts gen stcost_KM=s

gsort - stcost_KM Mi
quietly gen lagMi=Mi[_n-1]
replace lagMi=0 if lagMi==.

gen areai= stcost_KM*(Mi- lagMi)
egen meanKMtrue=sum(areai)

**variance**

gsort -Mi
gen Aix=sum(areai)
egen Ai=max( Aix), by(Mi)

gen int const=1
gsort -Mi
gen Rix=sum(const)
egen Ri=max(Rix), by(Mi)

egen di=sum(censorig), by(Mi)
gen termi=((Ai^2)*censorig)/(Ri*(Ri-di))

egen varKMtrue =sum(termi)
gen seKMtrue=sqrt(varKMtrue)
```

Full sample estimator
Same as in Kaplan-Meier but treating time of censoring as time of failure for the censored individuals

Uncensored cases estimator
Same as in Kaplan-Meier but only using the data for the uncensored individuals

**

Appendix A.4.5. Programs for the Lin et al estimators (Lin et al, 1997)

Lin1: Cost histories recorded

Based on equations (4.9), (4.10) and (4.11) for the mean and on equations (4.18), (4.19) and (4.20) for the variance

For Conventional (similarly for intensive)

```
** Mean **

gen ak=year-1
gen ak1=year

egen Xi=min(timallde), by(ukno)
egen di=min(censorig), by(ukno)

stset Xi if year==1, failure(di==1)

sts gen surv1138=s

gen int inttime=int(Xi)+1 if year==1
egen mintime=min(Xi), by(inttime)
gen st=surv1138 if mintime==Xi
egen Skx=min(st), by(inttime)

gen s1x=Skx if inttime==1
gen s2x=Skx if inttime==2
gen s3x=Skx if inttime==3
gen s4x=Skx if inttime==4
gen s5x=Skx if inttime==5
gen s6x=Skx if inttime==6
gen s7x=Skx if inttime==7
gen s8x=Skx if inttime==8
gen s9x=Skx if inttime==9
gen s10x=Skx if inttime==10
gen s11x=Skx if inttime==11
gen s12x=Skx if inttime==12
gen s13x=Skx if inttime==13
gen s14x=Skx if inttime==14
gen s15x=Skx if inttime==15
gen s16x=Skx if inttime==16
gen s17x=Skx if inttime==17
gen s18x=Skx if inttime==18
gen s19x=Skx if inttime==19

egen s1=min(s1x)
egen s2=min(s2x)
egen s3=min(s3x)
egen s4=min(s4x)
egen s5=min(s5x)
egen s6=min(s6x)
egen s7=min(s7x)
egen s8=min(s8x)
egen s9=min(s9x)
egen s10=min(s10x)
egen s11=min(s11x)
egen s12=min(s12x)
egen s13=min(s13x)
egen s14=min(s14x)
egen s15=min(s15x)
egen s16=min(s16x)
```

```
egen s17=min(s17x)
egen s18=min(s18x)
egen s19=min(s19x)

gen Sk=s1 if ak==0
replace Sk=s2 if ak==1
replace Sk=s3 if ak==2
replace Sk=s4 if ak==3
replace Sk=s5 if ak==4
replace Sk=s6 if ak==5
replace Sk=s7 if ak==6
replace Sk=s8 if ak==7
replace Sk=s9 if ak==8
replace Sk=s10 if ak==9
replace Sk=s11 if ak==10
replace Sk=s12 if ak==11
replace Sk=s13 if ak==12
replace Sk=s14 if ak==13
replace Sk=s15 if ak==14
replace Sk=s16 if ak==15
replace Sk=s17 if ak==16
replace Sk=s18 if ak==17
replace Sk=s19 if ak==18


gen Yki=1 if Xi>=ak
replace Yki=0 if Yki==.

gen Cki=costyr
replace Cki=0 if Cki==.

egen sYki=sum(Yki), by(ak)
gen YkiCki= Yki* Cki
egen sYkiCki=sum( YkiCki), by(ak)

gen Ek= sYkiCki/ sYki
gen SkEk= Sk* Ek
egen meanlin1=sum( SkEk), by(ukno)


**Variance**

gen term1= (Sk* Yki*( Cki- Ek))/ sYki
gen Xi_le_ak=1 if Xi<=ak
replace Xi_le_ak=0 if Xi>ak

gen int const=1
gsort ak -Xi
by ak: gen Rix=sum(const)
egen Ri=max(Rix), by(ak Xi)
drop Rix

gen term2a=( Xi_le_ak* di)/Ri
gen diRj2=di/(Ri^2)
gen minakXi=min(ak, Xi)

sort ak Xi
by ak: gen S1x=sum( diRj2) if Xi<= minakXi
egen S1=max(S1x), by(ak Xi)
egen S2=max(S1), by(ak)
replace S1=S2 if S1==.
replace S1=0 if S1==.

gen term2b=S1
```

```
gen term2=Sk*Ek*(term2a-term2b)
gen wki=term1-term2
egen sumwki=sum(wki), by(ukno)
gen wkiwli=wki*sumwki

egen varlin1=sum(wkiwli)

**
```

Lin2: Cost histories not recorded

Based on equations (4.21), (4.22) and (4.23) for the mean and on equations (4.28), (4.29), (4.30) and (4.31) for the variance

For Conventional (similarly for intensive)

```
** Mean**

egen Ci=min(Mi), by(ukno)
egen Xi=min(timallde), by(ukno)
egen di=min(censorig), by(ukno)

sort ukno ak

quietly by ukno: gen Sk1=Sk[_n+1]

replace Sk1=0 if ak==19

egen tmax=max(Xi)

gen Yki=1 if ((ak<=Xi & Xi<ak1) & di==1) | (Xi>=tmax & ak==19)
replace Yki=0 if Yki==.


egen sYki=sum(Yki), by(ak)

gen YkiCi= Yki* Ci

egen sYkiCi=sum(YkiCi), by(ak)

gen Ak= sYkiCi/ sYki
replace Ak=0 if Ak==.

gen AkS= Ak*(Sk-Sk1)

egen meanlin2=sum(AkS), by(ukno)


** Variance **

gen term1= ((Sk- Sk1)* Yki*( Ci- Ak))/ sYki
replace term1=0 if term1==.

**Ri**

gen int const=1

gsort ak -Xi
by ak: gen Rix=sum(const)
egen Ri=max(Rix), by(ak Xi)
drop Rix
```

```
** For Dki **

gen Xi_le_ak=1 if Xi<=ak
replace Xi_le_ak=0 if Xi>ak

gen term2a=( Xi_le_ak* di)/Ri

gen diRj2=di/(Ri^2)

gen minakXi=min(ak, Xi)

sort ak Xi

by ak: gen S1x=sum( diRj2) if Xi<= minakXi
egen S1=max(S1x), by(ak Xi)
egen S2=max(S1), by(ak)
replace S1=S2 if S1==.
replace S1=0 if S1==.

gen term2b=S1


**For Dk+1,i**

gen Xi_le_ak1=1 if Xi<=ak1
replace Xi_le_ak1=0 if Xi>ak1

gen term2ak1=( Xi_le_ak1* di)/Ri

gen minak1Xi=min(ak1, Xi)

sort ak1 Xi

by ak1: gen S1xk1=sum( diRj2) if Xi<= minak1Xi
egen S1k1=max(S1xk1), by(ak1 Xi)
egen S2k1=max(S1k1), by(ak1)
replace S1k1=S2k1 if S1k1==.
replace S1k1=0 if S1k1==.

gen term2bk1=S1k1

gen Dki= term2a- term2b

gen Dk1i= term2ak1- term2bk1

gen term2=Ak*( Sk1* Dk1i- Sk* Dki)

gen wki=term1+term2

egen sumwki=sum(wki), by(ukno)

gen wkiwli=wki*sumwki

egen varlin2=sum(wkiwli)


**
```

- Similarly for monthly interval partitions
- Same programs when the highest cost outliers were excluded and when different durations of analysis were considered

```
**
```

Appendix A.4.6. Programs for the Bang and Tsiatis estimators (Bang and Tsiatis, 2000)

Simple weighted estimator

Based on equations (4.32) and (4.33) for the mean and on equations (4.35) and (4.36) for the variance

For Conventional (similarly for intensive)

```
** Simple weighted estimator: Mean and Variance_ Conventional **

stset timallde, failure(censorig==1)

sts gen surv1138=s

gen censor1=1 if censorig==0
replace censor1=0 if censor1==.

label var censor1 "1:censored; 0:dead"

stset  timallde, failure(censor1==1)

sts gen cens1138=s

gen diMi_KTi= (censorig* Mi)/ cens1138
replace  diMi_KTi=0 if diMi_KTi==.

egen sumalli=sum(diMi_KTi)

gen meansimp=(1/1138)* sumalli


**Variance**

gen MiMiu=( censorig* Mi* Mi)/ cens1138
replace MiMiu=0 if MiMiu==.

gsort -timallde

gen sMiMiu=sum(MiMiu)
egen sumMiMiu=max(sMiMiu), by(timallde)
drop sMiMiu

gen gMiMiu=(1/1138)*(1/ surv1138)*sumMiMiu

gen Miu=( censorig* Mi)/ cens1138
replace Miu=0 if Miu==.

gsort -timallde

gen sMiu=sum(Miu)
egen sumMiu=max(sMiu), by(timallde)

drop sMiu

gen gMiu_2=((1/1138)*(1/surv1138)*sumMiu)^2

gen intern2=(censor1/(cens1138^2))*(gMiMiu- gMiu_2)
replace  intern2=0 if intern2==.

egen sumint2=sum(intern2)
```

```
gen term2=(1/1138)* sumint2

gen intern1= censorig*((Mi- meansimp)^2)/ cens1138
replace intern1=0 if intern1==.

egen sumint1=sum(intern1)

gen term1=(1/1138)*sumint1

gen varsimp=(1/1138)*(term1+term2)

gen sesimp=sqrt(varsimp)

** To calculate Yu **

gen int const=1

gsort -timallde

gen Yux=sum(const)

egen Yu=max(Yux), by(timallde)

drop Yux

label var Yu "no. at risk"

**
```

Partitioned estimator

Based on equations (4.37) for the mean and on equations (4.38), (4.39), and (4.40) for the variance

For Conventional (similarly for intensive)

```
** Partitioned mean and variance: Conventional**

gen tj_1=year-1
gen tj=year

gen Mij=costyr
replace Mij=0 if Mij==.

egen Xi=min(timallde), by(ukno)

egen di=min(censorig), by(ukno)

gen minTitj=min(Xi, tj)

gen Xij=min(minTitj, Xi)

gen dij=1 if (minTitj==tj | (minTitj==Xi & di==1))
replace dij=0 if dij==.

stset Xij, failure(dij==0)

sts gen KjTij=s, by(tj)

gen intern=(dij*Mij)/KjTij
replace intern=0 if intern==.
```

```
egen sumintjs=sum(intern), by(ukno)

** Variance: conventional**

gen l1=1
gen l2=2
gen l3=3
gen l4=4
gen l5=5
gen l6=6
gen l7=7
gen l8=8
gen l9=9
gen l10=10
gen l11=11
gen l12=12
gen l13=13
gen l14=14
gen l15=15
gen l16=16
gen l17=17
gen l18=18
gen l19=19


gen jmaxl1=max(tj, l1)
gen jmaxl2=max(tj, l2)
gen jmaxl3=max(tj, l3)
gen jmaxl4=max(tj, l4)
gen jmaxl5=max(tj, l5)
gen jmaxl6=max(tj, l6)
gen jmaxl7=max(tj, l7)
gen jmaxl8=max(tj, l8)
gen jmaxl9=max(tj, l9)
gen jmaxl10=max(tj, l10)
gen jmaxl11=max(tj, l11)
gen jmaxl12=max(tj, l12)
gen jmaxl13=max(tj, l13)
gen jmaxl14=max(tj, l14)
gen jmaxl15=max(tj, l15)
gen jmaxl16=max(tj, l16)
gen jmaxl17=max(tj, l17)
gen jmaxl18=max(tj, l18)
gen jmaxl19=max(tj, l19)

gen jminl1=min(tj, l1)
gen jminl2=min(tj, l2)
gen jminl3=min(tj, l3)
gen jminl4=min(tj, l4)
gen jminl5=min(tj, l5)
gen jminl6=min(tj, l6)
gen jminl7=min(tj, l7)
gen jminl8=min(tj, l8)
gen jminl9=min(tj, l9)
gen jminl10=min(tj, l10)
gen jminl11=min(tj, l11)
gen jminl12=min(tj, l12)
gen jminl13=min(tj, l13)
gen jminl14=min(tj, l14)
gen jminl15=min(tj, l15)
gen jminl16=min(tj, l16)
gen jminl17=min(tj, l17)
gen jminl18=min(tj, l18)
```

```
gen jminl19=min(tj, l19)

gen Tijmin1=min(Xi, jminl1)
gen Tijmin2=min(Xi, jminl2)
gen Tijmin3=min(Xi, jminl3)
gen Tijmin4=min(Xi, jminl4)
gen Tijmin5=min(Xi, jminl5)
gen Tijmin6=min(Xi, jminl6)
gen Tijmin7=min(Xi, jminl7)
gen Tijmin8=min(Xi, jminl8)
gen Tijmin9=min(Xi, jminl9)
gen Tijmin10=min(Xi, jminl10)
gen Tijmin11=min(Xi, jminl11)
gen Tijmin12=min(Xi, jminl12)
gen Tijmin13=min(Xi, jminl13)
gen Tijmin14=min(Xi, jminl14)
gen Tijmin15=min(Xi, jminl15)
gen Tijmin16=min(Xi, jminl16)
gen Tijmin17=min(Xi, jminl17)
gen Tijmin18=min(Xi, jminl18)
gen Tijmin19=min(Xi, jminl19)


gen Tijmax1=min(Xi, jmaxl1)
gen Tijmax2=min(Xi, jmaxl2)
gen Tijmax3=min(Xi, jmaxl3)
gen Tijmax4=min(Xi, jmaxl4)
gen Tijmax5=min(Xi, jmaxl5)
gen Tijmax6=min(Xi, jmaxl6)
gen Tijmax7=min(Xi, jmaxl7)
gen Tijmax8=min(Xi, jmaxl8)
gen Tijmax9=min(Xi, jmaxl9)
gen Tijmax10=min(Xi, jmaxl10)
gen Tijmax11=min(Xi, jmaxl11)
gen Tijmax12=min(Xi, jmaxl12)
gen Tijmax13=min(Xi, jmaxl13)
gen Tijmax14=min(Xi, jmaxl14)
gen Tijmax15=min(Xi, jmaxl15)
gen Tijmax16=min(Xi, jmaxl16)
gen Tijmax17=min(Xi, jmaxl17)
gen Tijmax18=min(Xi, jmaxl18)
gen Tijmax19=min(Xi, jmaxl19)


gen dijmax1=1 if Tijmax1==jmaxl1 | (Tijmax1==Xi & di==1)
replace dijmax1=0 if dijmax1==.

gen dijmax2=1 if Tijmax2==jmaxl2 | (Tijmax2==Xi & di==1)
replace dijmax2=0 if dijmax2==.

gen dijmax3=1 if Tijmax3==jmaxl3 | (Tijmax3==Xi & di==1)
replace dijmax3=0 if dijmax3==.

gen dijmax4=1 if Tijmax4==jmaxl4 | (Tijmax4==Xi & di==1)
replace dijmax4=0 if dijmax4==.

gen dijmax5=1 if Tijmax5==jmaxl5 | (Tijmax5==Xi & di==1)
replace dijmax5=0 if dijmax5==.

gen dijmax6=1 if Tijmax6==jmaxl6 | (Tijmax6==Xi & di==1)
replace dijmax6=0 if dijmax6==.

gen dijmax7=1 if Tijmax7==jmaxl7 | (Tijmax7==Xi & di==1)
replace dijmax7=0 if dijmax7==.
```

```
gen dijmax8=1 if Tijmax8==jmax18 | (Tijmax8==Xi & di==1)
replace dijmax8=0 if dijmax8==.

gen dijmax9=1 if Tijmax9==jmax19 | (Tijmax9==Xi & di==1)
replace dijmax9=0 if dijmax9==.

gen dijmax10=1 if Tijmax10==jmax110 | (Tijmax10==Xi & di==1)
replace dijmax10=0 if dijmax10==.

gen dijmax11=1 if Tijmax11==jmax111 | (Tijmax11==Xi & di==1)
replace dijmax11=0 if dijmax11==.

gen dijmax12=1 if Tijmax12==jmax112 | (Tijmax12==Xi & di==1)
replace dijmax12=0 if dijmax12==.

gen dijmax13=1 if Tijmax13==jmax113 | (Tijmax13==Xi & di==1)
replace dijmax13=0 if dijmax13==.

gen dijmax14=1 if Tijmax14==jmax114 | (Tijmax14==Xi & di==1)
replace dijmax14=0 if dijmax14==.

gen dijmax15=1 if Tijmax15==jmax115 | (Tijmax15==Xi & di==1)
replace dijmax15=0 if dijmax15==.

gen dijmax16=1 if Tijmax16==jmax116 | (Tijmax16==Xi & di==1)
replace dijmax16=0 if dijmax16==.

gen dijmax17=1 if Tijmax17==jmax117 | (Tijmax17==Xi & di==1)
replace dijmax17=0 if dijmax17==.

gen dijmax18=1 if Tijmax18==jmax118 | (Tijmax18==Xi & di==1)
replace dijmax18=0 if dijmax18==.

gen dijmax19=1 if Tijmax19==jmax119 | (Tijmax19==Xi & di==1)
replace dijmax19=0 if dijmax19==.


stset Tijmax1, failure(dijmax1==0)
sts gen Kjmax11=s, by(tj)

stset Tijmax2, failure(dijmax2==0)
sts gen Kjmax12=s, by(tj)

stset Tijmax3, failure(dijmax3==0)
sts gen Kjmax13=s, by(tj)

stset Tijmax4, failure(dijmax4==0)
sts gen Kjmax14=s, by(tj)

stset Tijmax5, failure(dijmax5==0)
sts gen Kjmax15=s, by(tj)

stset Tijmax6, failure(dijmax6==0)
sts gen Kjmax16=s, by(tj)

stset Tijmax7, failure(dijmax7==0)
sts gen Kjmax17=s, by(tj)

stset Tijmax8, failure(dijmax8==0)
sts gen Kjmax18=s, by(tj)

stset Tijmax9, failure(dijmax9==0)
sts gen Kjmax19=s, by(tj)
```

```
stset Tijmax10, failure(dijmax10==0)
sts gen Kjmaxl10=s, by(tj)

stset Tijmax11, failure(dijmax11==0)
sts gen Kjmaxl11=s, by(tj)

stset Tijmax12, failure(dijmax12==0)
sts gen Kjmaxl12=s, by(tj)

stset Tijmax13, failure(dijmax13==0)
sts gen Kjmaxl13=s, by(tj)

stset Tijmax14, failure(dijmax14==0)
sts gen Kjmaxl14=s, by(tj)

stset Tijmax15, failure(dijmax15==0)
sts gen Kjmaxl15=s, by(tj)

stset Tijmax16, failure(dijmax16==0)
sts gen Kjmaxl16=s, by(tj)

stset Tijmax17, failure(dijmax17==0)
sts gen Kjmaxl17=s, by(tj)

stset Tijmax18, failure(dijmax18==0)
sts gen Kjmaxl18=s, by(tj)

stset Tijmax19, failure(dijmax19==0)
sts gen Kjmaxl19=s, by(tj)

*

gen dijmin1=1 if Tijmin1==jminl1 | (Tijmin1==Xi & di==1)
replace dijmin1=0 if dijmin1==.

gen dijmin2=1 if Tijmin2==jminl2 | (Tijmin2==Xi & di==1)
replace dijmin2=0 if dijmin2==.

gen dijmin3=1 if Tijmin3==jminl3 | (Tijmin3==Xi & di==1)
replace dijmin3=0 if dijmin3==.

gen dijmin4=1 if Tijmin4==jminl4 | (Tijmin4==Xi & di==1)
replace dijmin4=0 if dijmin4==.

gen dijmin5=1 if Tijmin5==jminl5 | (Tijmin5==Xi & di==1)
replace dijmin5=0 if dijmin5==.

gen dijmin6=1 if Tijmin6==jminl6 | (Tijmin6==Xi & di==1)
replace dijmin6=0 if dijmin6==.

gen dijmin7=1 if Tijmin7==jminl7 | (Tijmin7==Xi & di==1)
replace dijmin7=0 if dijmin7==.

gen dijmin8=1 if Tijmin8==jminl8 | (Tijmin8==Xi & di==1)
replace dijmin8=0 if dijmin8==.

gen dijmin9=1 if Tijmin9==jminl9 | (Tijmin9==Xi & di==1)
replace dijmin9=0 if dijmin9==.

gen dijmin10=1 if Tijmin10==jminl10 | (Tijmin10==Xi & di==1)
replace dijmin10=0 if dijmin10==.

gen dijmin11=1 if Tijmin11==jminl11 | (Tijmin11==Xi & di==1)
```

```
replace dijmin11=0 if dijmin11==.

gen dijmin12=1 if Tijmin12==jmin112 | (Tijmin12==Xi & di==1)
replace dijmin12=0 if dijmin12==.

gen dijmin13=1 if Tijmin13==jmin113 | (Tijmin13==Xi & di==1)
replace dijmin13=0 if dijmin13==.

gen dijmin14=1 if Tijmin14==jmin114 | (Tijmin14==Xi & di==1)
replace dijmin14=0 if dijmin14==.

gen dijmin15=1 if Tijmin15==jmin115 | (Tijmin15==Xi & di==1)
replace dijmin15=0 if dijmin15==.

gen dijmin16=1 if Tijmin16==jmin116 | (Tijmin16==Xi & di==1)
replace dijmin16=0 if dijmin16==.

gen dijmin17=1 if Tijmin17==jmin117 | (Tijmin17==Xi & di==1)
replace dijmin17=0 if dijmin17==.

gen dijmin18=1 if Tijmin18==jmin118 | (Tijmin18==Xi & di==1)
replace dijmin18=0 if dijmin18==.

gen dijmin19=1 if Tijmin19==jmin119 | (Tijmin19==Xi & di==1)
replace dijmin19=0 if dijmin19==.


stset Tijmin1, failure(dijmin1==1)
sts gen Sjl1u=s, by(tj)

stset Tijmin2, failure(dijmin2==1)
sts gen Sjl2u=s, by(tj)

stset Tijmin3, failure(dijmin3==1)
sts gen Sjl3u=s, by(tj)

stset Tijmin4, failure(dijmin4==1)
sts gen Sjl4u=s, by(tj)

stset Tijmin5, failure(dijmin5==1)
sts gen Sjl5u=s, by(tj)

stset Tijmin6, failure(dijmin6==1)
sts gen Sjl6u=s, by(tj)

stset Tijmin7, failure(dijmin7==1)
sts gen Sjl7u=s, by(tj)

stset Tijmin8, failure(dijmin8==1)
sts gen Sjl8u=s, by(tj)

stset Tijmin9, failure(dijmin9==1)
sts gen Sjl9u=s, by(tj)

stset Tijmin10, failure(dijmin10==1)
sts gen Sjl10u=s, by(tj)

stset Tijmin11, failure(dijmin11==1)
sts gen Sjl11u=s, by(tj)

stset Tijmin12, failure(dijmin12==1)
sts gen Sjl12u=s, by(tj)

stset Tijmin13, failure(dijmin13==1)
```

```
sts gen Sjl13u=s, by(tj)

stset Tijmin14, failure(dijmin14==1)
sts gen Sjl14u=s, by(tj)

stset Tijmin15, failure(dijmin15==1)
sts gen Sjl15u=s, by(tj)

stset Tijmin16, failure(dijmin16==1)
sts gen Sjl16u=s, by(tj)

stset Tijmin17, failure(dijmin17==1)
sts gen Sjl17u=s, by(tj)

stset Tijmin18, failure(dijmin18==1)
sts gen Sjl18u=s, by(tj)

stset Tijmin19, failure(dijmin19==1)
sts gen Sjl19u=s, by(tj)

**

sort ukno tj

gen Mi1lead=Mij

quietly by ukno: gen Mi2lead=Mi1lead[_n+1]
quietly by ukno: gen Mi3lead=Mi2lead[_n+1]
quietly by ukno: gen Mi4lead=Mi3lead[_n+1]
quietly by ukno: gen Mi5lead=Mi4lead[_n+1]
quietly by ukno: gen Mi6lead=Mi5lead[_n+1]
quietly by ukno: gen Mi7lead=Mi6lead[_n+1]
quietly by ukno: gen Mi8lead=Mi7lead[_n+1]
quietly by ukno: gen Mi9lead=Mi8lead[_n+1]
quietly by ukno: gen Mi10lead=Mi9lead[_n+1]
quietly by ukno: gen Mi11lead=Mi10lead[_n+1]
quietly by ukno: gen Mi12lead=Mi11lead[_n+1]
quietly by ukno: gen Mi13lead=Mi12lead[_n+1]
quietly by ukno: gen Mi14lead=Mi13lead[_n+1]
quietly by ukno: gen Mi15lead=Mi14lead[_n+1]
quietly by ukno: gen Mi16lead=Mi15lead[_n+1]
quietly by ukno: gen Mi17lead=Mi16lead[_n+1]
quietly by ukno: gen Mi18lead=Mi17lead[_n+1]
quietly by ukno: gen Mi19lead=Mi18lead[_n+1]


replace  Mi1lead=-9 if year~=1
replace  Mi2lead=-9 if year~=1
replace  Mi3lead=-9 if year~=1
replace  Mi4lead=-9 if year~=1
replace  Mi5lead=-9 if year~=1
replace  Mi6lead=-9 if year~=1
replace  Mi7lead=-9 if year~=1
replace  Mi8lead=-9 if year~=1
replace  Mi9lead=-9 if year~=1
replace  Mi10lead=-9 if year~=1
replace  Mi11lead=-9 if year~=1
replace  Mi12lead=-9 if year~=1
replace  Mi13lead=-9 if year~=1
replace  Mi14lead=-9 if year~=1
replace  Mi15lead=-9 if year~=1
replace  Mi16lead=-9 if year~=1
replace  Mi17lead=-9 if year~=1
replace  Mi18lead=-9 if year~=1
```

```
replace  Mi19lead=-9 if year~=1


egen Mi1=max(Mi1lead), by(ukno)
egen Mi2=max(Mi2lead), by(ukno)
egen Mi3=max(Mi3lead), by(ukno)
egen Mi4=max(Mi4lead), by(ukno)
egen Mi5=max(Mi5lead), by(ukno)
egen Mi6=max(Mi6lead), by(ukno)
egen Mi7=max(Mi7lead), by(ukno)
egen Mi8=max(Mi8lead), by(ukno)
egen Mi9=max(Mi9lead), by(ukno)
egen Mi10=max(Mi10lead), by(ukno)
egen Mi11=max(Mi11lead), by(ukno)
egen Mi12=max(Mi12lead), by(ukno)
egen Mi13=max(Mi13lead), by(ukno)
egen Mi14=max(Mi14lead), by(ukno)
egen Mi15=max(Mi15lead), by(ukno)
egen Mi16=max(Mi16lead), by(ukno)
egen Mi17=max(Mi17lead), by(ukno)
egen Mi18=max(Mi18lead), by(ukno)
egen Mi19=max(Mi19lead), by(ukno)


gen gMjMl1= (dijmax1* Mij* Mi1)/ Kjmaxl1
replace  gMjMl1=0 if gMjMl1==.

gen gMjMl2= (dijmax2* Mij* Mi2)/ Kjmaxl2
replace  gMjMl2=0 if gMjMl2==.

gen gMjMl3= (dijmax3* Mij* Mi3)/ Kjmaxl3
replace  gMjMl3=0 if gMjMl3==.

gen gMjMl4= (dijmax4* Mij* Mi4)/ Kjmaxl4
replace  gMjMl4=0 if gMjMl4==.

gen gMjMl5= (dijmax5* Mij* Mi5)/ Kjmaxl5
replace  gMjMl5=0 if gMjMl5==.

gen gMjMl6= (dijmax6* Mij* Mi6)/ Kjmaxl6
replace  gMjMl6=0 if gMjMl6==.

gen gMjMl7= (dijmax7* Mij* Mi7)/ Kjmaxl7
replace  gMjMl7=0 if gMjMl7==.

gen gMjMl8= (dijmax8* Mij* Mi8)/ Kjmaxl8
replace  gMjMl8=0 if gMjMl8==.

gen gMjMl9= (dijmax9* Mij* Mi9)/ Kjmaxl9
replace  gMjMl9=0 if gMjMl9==.

gen gMjMl10= (dijmax10* Mij* Mi10)/ Kjmaxl10
replace  gMjMl10=0 if gMjMl10==.

gen gMjMl11= (dijmax11* Mij* Mi11)/ Kjmaxl11
replace  gMjMl11=0 if gMjMl11==.

gen gMjMl12= (dijmax12* Mij* Mi12)/ Kjmaxl12
replace  gMjMl12=0 if gMjMl12==.

gen gMjMl13= (dijmax13* Mij* Mi13)/ Kjmaxl13
replace  gMjMl13=0 if gMjMl13==.

gen gMjMl14= (dijmax14* Mij* Mi14)/ Kjmaxl14
```

```
replace  gMjMl14=0 if gMjMl14==.

gen gMjMl15= (dijmax15* Mij* Mi15)/ Kjmaxl15
replace  gMjMl15=0 if gMjMl15==.

gen gMjMl16= (dijmax16* Mij* Mi16)/ Kjmaxl16
replace  gMjMl16=0 if gMjMl16==.

gen gMjMl17= (dijmax17* Mij* Mi17)/ Kjmaxl17
replace  gMjMl17=0 if gMjMl17==.

gen gMjMl18= (dijmax18* Mij* Mi18)/ Kjmaxl18
replace  gMjMl18=0 if gMjMl18==.

gen gMjMl19= (dijmax19* Mij* Mi19)/ Kjmaxl19
replace  gMjMl19=0 if gMjMl19==.


gen gMjl1= (dijmax1* Mij)/ Kjmaxl1
replace  gMjl1=0 if gMjl1==.

gen gMjl2= (dijmax2* Mij)/ Kjmaxl2
replace  gMjl2=0 if gMjl2==.

gen gMjl3= (dijmax3* Mij)/ Kjmaxl3
replace  gMjl3=0 if gMjl3==.

gen gMjl4= (dijmax4* Mij)/ Kjmaxl4
replace  gMjl4=0 if gMjl4==.

gen gMjl5= (dijmax5* Mij)/ Kjmaxl5
replace  gMjl5=0 if gMjl5==.

gen gMjl6= (dijmax6* Mij)/ Kjmaxl6
replace  gMjl6=0 if gMjl6==.

gen gMjl7= (dijmax7* Mij)/ Kjmaxl7
replace  gMjl7=0 if gMjl7==.

gen gMjl8= (dijmax8* Mij)/ Kjmaxl8
replace  gMjl8=0 if gMjl8==.

gen gMjl9= (dijmax9* Mij)/ Kjmaxl9
replace  gMjl9=0 if gMjl9==.

gen gMjl10= (dijmax10* Mij)/ Kjmaxl10
replace  gMjl10=0 if gMjl10==.

gen gMjl11= (dijmax11* Mij)/ Kjmaxl11
replace  gMjl11=0 if gMjl11==.

gen gMjl12= (dijmax12* Mij)/ Kjmaxl12
replace  gMjl12=0 if gMjl12==.

gen gMjl13= (dijmax13* Mij)/ Kjmaxl13
replace  gMjl13=0 if gMjl13==.

gen gMjl14= (dijmax14* Mij)/ Kjmaxl14
replace  gMjl14=0 if gMjl14==.

gen gMjl15= (dijmax15* Mij)/ Kjmaxl15
replace  gMjl15=0 if gMjl15==.

gen gMjl16= (dijmax16* Mij)/ Kjmaxl16
```

```
replace  gMjl16=0 if gMjl16==.

gen gMjl17= (dijmax17* Mij)/ Kjmaxl17
replace  gMjl17=0 if gMjl17==.

gen gMjl18= (dijmax18* Mij)/ Kjmaxl18
replace  gMjl18=0 if gMjl18==.

gen gMjl19= (dijmax19* Mij)/ Kjmaxl19
replace  gMjl19=0 if gMjl19==.

*

gen gMl1= (dijmax1* Mi1)/ Kjmaxl1
replace  gMl1=0 if  gMl1==.

gen gMl2= (dijmax2* Mi2)/ Kjmaxl2
replace  gMl2=0 if  gMl2==.

gen gMl3= (dijmax3* Mi3)/ Kjmaxl3
replace  gMl3=0 if  gMl3==.

gen gMl4= (dijmax4* Mi4)/ Kjmaxl4
replace  gMl4=0 if  gMl4==.

gen gMl5= (dijmax5* Mi5)/ Kjmaxl5
replace  gMl5=0 if  gMl5==.

gen gMl6= (dijmax6* Mi6)/ Kjmaxl6
replace  gMl6=0 if  gMl6==.

gen gMl7= (dijmax7* Mi7)/ Kjmaxl7
replace  gMl7=0 if  gMl7==.

gen gMl8= (dijmax8* Mi8)/ Kjmaxl8
replace  gMl8=0 if  gMl8==.

gen gMl9= (dijmax9* Mi9)/ Kjmaxl9
replace  gMl9=0 if  gMl9==.

gen gMl10= (dijmax10* Mi10)/ Kjmaxl10
replace  gMl10=0 if  gMl10==.

gen gMl11= (dijmax11* Mi11)/ Kjmaxl11
replace  gMl11=0 if  gMl11==.

gen gMl12= (dijmax12* Mi12)/ Kjmaxl12
replace  gMl12=0 if  gMl12==.

gen gMl13= (dijmax13* Mi13)/ Kjmaxl13
replace  gMl13=0 if  gMl13==.

gen gMl14= (dijmax14* Mi14)/ Kjmaxl14
replace  gMl14=0 if  gMl14==.

gen gMl15= (dijmax15* Mi15)/ Kjmaxl15
replace  gMl15=0 if  gMl15==.

gen gMl16= (dijmax16* Mi16)/ Kjmaxl16
replace  gMl16=0 if  gMl16==.

gen gMl17= (dijmax17* Mi17)/ Kjmaxl17
replace  gMl17=0 if  gMl17==.
```

```
gen gMl18= (dijmax18* Mi18)/ Kjmaxl18
replace  gMl18=0 if  gMl18==.

gen gMl19= (dijmax19* Mi19)/ Kjmaxl19
replace  gMl19=0 if  gMl19==.
*

gsort tj -Tijmin1
by tj: gen gMjMl1x=sum(gMjMl1)
replace gMjMl1x=0 if Tijmin1<Xi
egen sgMjMl1=max(gMjMl1x), by(tj Tijmin1)
drop gMjMl1x

gsort tj -Tijmin2
by tj: gen gMjMl2x=sum(gMjMl2)
replace gMjMl2x=0 if Tijmin2<Xi
egen sgMjMl2=max(gMjMl2x), by(tj Tijmin2)
drop gMjMl2x

gsort tj -Tijmin3
by tj: gen gMjMl3x=sum(gMjMl3)
replace gMjMl3x=0 if Tijmin3<Xi
egen sgMjMl3=max(gMjMl3x), by(tj Tijmin3)
drop gMjMl3x

gsort tj -Tijmin4
by tj: gen gMjMl4x=sum(gMjMl4)
replace gMjMl4x=0 if Tijmin4<Xi
egen sgMjMl4=max(gMjMl4x), by(tj Tijmin4)
drop gMjMl4x

gsort tj -Tijmin5
by tj: gen gMjMl5x=sum(gMjMl5)
replace gMjMl5x=0 if Tijmin5<Xi
egen sgMjMl5=max(gMjMl5x), by(tj Tijmin5)
drop gMjMl5x

gsort tj -Tijmin6
by tj: gen gMjMl6x=sum(gMjMl6)
replace gMjMl6x=0 if Tijmin6<Xi
egen sgMjMl6=max(gMjMl6x), by(tj Tijmin6)
drop gMjMl6x

gsort tj -Tijmin7
by tj: gen gMjMl7x=sum(gMjMl7)
replace gMjMl7x=0 if Tijmin7<Xi
egen sgMjMl7=max(gMjMl7x), by(tj Tijmin7)
drop gMjMl7x

gsort tj -Tijmin8
by tj: gen gMjMl8x=sum(gMjMl8)
replace gMjMl8x=0 if Tijmin8<Xi
egen sgMjMl8=max(gMjMl8x), by(tj Tijmin8)
drop gMjMl8x

gsort tj -Tijmin9
by tj: gen gMjMl9x=sum(gMjMl9)
replace gMjMl9x=0 if Tijmin9<Xi
egen sgMjMl9=max(gMjMl9x), by(tj Tijmin9)
drop gMjMl9x

gsort tj -Tijmin10
by tj: gen gMjMl10x=sum(gMjMl10)
replace gMjMl10x=0 if Tijmin10<Xi
```

```
egen sgMjMl10=max(gMjMl10x), by(tj Tijmin10)
drop gMjMl10x

gsort tj -Tijmin11
by tj: gen gMjMl11x=sum(gMjMl11)
replace gMjMl11x=0 if Tijmin11<Xi
egen sgMjMl11=max(gMjMl11x), by(tj Tijmin11)
drop gMjMl11x

gsort tj -Tijmin12
by tj: gen gMjMl12x=sum(gMjMl12)
replace gMjMl12x=0 if Tijmin12<Xi
egen sgMjMl12=max(gMjMl12x), by(tj Tijmin12)
drop gMjMl12x

gsort tj -Tijmin13
by tj: gen gMjMl13x=sum(gMjMl13)
replace gMjMl13x=0 if Tijmin13<Xi
egen sgMjMl13=max(gMjMl13x), by(tj Tijmin13)
drop gMjMl13x

gsort tj -Tijmin14
by tj: gen gMjMl14x=sum(gMjMl14)
replace gMjMl14x=0 if Tijmin14<Xi
egen sgMjMl14=max(gMjMl14x), by(tj Tijmin14)
drop gMjMl14x

gsort tj -Tijmin15
by tj: gen gMjMl15x=sum(gMjMl15)
replace gMjMl15x=0 if Tijmin15<Xi
egen sgMjMl15=max(gMjMl15x), by(tj Tijmin15)
drop gMjMl15x

gsort tj -Tijmin16
by tj: gen gMjMl16x=sum(gMjMl16)
replace gMjMl16x=0 if Tijmin16<Xi
egen sgMjMl16=max(gMjMl16x), by(tj Tijmin16)
drop gMjMl16x

gsort tj -Tijmin17
by tj: gen gMjMl17x=sum(gMjMl17)
replace gMjMl17x=0 if Tijmin17<Xi
egen sgMjMl17=max(gMjMl17x), by(tj Tijmin17)
drop gMjMl17x

gsort tj -Tijmin18
by tj: gen gMjMl18x=sum(gMjMl18)
replace gMjMl18x=0 if Tijmin18<Xi
egen sgMjMl18=max(gMjMl18x), by(tj Tijmin18)
drop gMjMl18x

gsort tj -Tijmin19
by tj: gen gMjMl19x=sum(gMjMl19)
replace gMjMl19x=0 if Tijmin19<Xi
egen sgMjMl19=max(gMjMl19x), by(tj Tijmin19)
drop gMjMl19x


gsort tj -Tijmin1
by tj: gen gMjl1x=sum(gMjl1)
replace gMjl1x=0 if Tijmin1<Xi
egen sgMjl1=max(gMjl1x), by(tj Tijmin1)
drop gMjl1x
```

```
gsort tj -Tijmin2
by tj: gen gMjl2x=sum(gMjl2)
replace gMjl2x=0 if Tijmin2<Xi
egen sgMjl2=max(gMjl2x), by(tj Tijmin2)
drop gMjl2x

gsort tj -Tijmin3
by tj: gen gMjl3x=sum(gMjl3)
replace gMjl3x=0 if Tijmin3<Xi
egen sgMjl3=max(gMjl3x), by(tj Tijmin3)
drop gMjl3x

gsort tj -Tijmin4
by tj: gen gMjl4x=sum(gMjl4)
replace gMjl4x=0 if Tijmin4<Xi
egen sgMjl4=max(gMjl4x), by(tj Tijmin4)
drop gMjl4x

gsort tj -Tijmin5
by tj: gen gMjl5x=sum(gMjl5)
replace gMjl5x=0 if Tijmin5<Xi
egen sgMjl5=max(gMjl5x), by(tj Tijmin5)
drop gMjl5x

gsort tj -Tijmin6
by tj: gen gMjl6x=sum(gMjl6)
replace gMjl6x=0 if Tijmin6<Xi
egen sgMjl6=max(gMjl6x), by(tj Tijmin6)
drop gMjl6x

gsort tj -Tijmin7
by tj: gen gMjl7x=sum(gMjl7)
replace gMjl7x=0 if Tijmin7<Xi
egen sgMjl7=max(gMjl7x), by(tj Tijmin7)
drop gMjl7x

gsort tj -Tijmin8
by tj: gen gMjl8x=sum(gMjl8)
replace gMjl8x=0 if Tijmin8<Xi
egen sgMjl8=max(gMjl8x), by(tj Tijmin8)
drop gMjl8x

gsort tj -Tijmin9
by tj: gen gMjl9x=sum(gMjl9)
replace gMjl9x=0 if Tijmin9<Xi
egen sgMjl9=max(gMjl9x), by(tj Tijmin9)
drop gMjl9x

gsort tj -Tijmin10
by tj: gen gMjl10x=sum(gMjl10)
replace gMjl10x=0 if Tijmin10<Xi
egen sgMjl10=max(gMjl10x), by(tj Tijmin10)
drop gMjl10x

gsort tj -Tijmin11
by tj: gen gMjl11x=sum(gMjl11)
replace gMjl11x=0 if Tijmin11<Xi
egen sgMjl11=max(gMjl11x), by(tj Tijmin11)
drop gMjl11x

gsort tj -Tijmin12
by tj: gen gMjl12x=sum(gMjl12)
replace gMjl12x=0 if Tijmin12<Xi
egen sgMjl12=max(gMjl12x), by(tj Tijmin12)
```

```
drop gMjl12x

gsort tj -Tijmin13
by tj: gen gMjl13x=sum(gMjl13)
replace gMjl13x=0 if Tijmin13<Xi
egen sgMjl13=max(gMjl13x), by(tj Tijmin13)
drop gMjl13x

gsort tj -Tijmin14
by tj: gen gMjl14x=sum(gMjl14)
replace gMjl14x=0 if Tijmin14<Xi
egen sgMjl14=max(gMjl14x), by(tj Tijmin14)
drop gMjl14x

gsort tj -Tijmin15
by tj: gen gMjl15x=sum(gMjl15)
replace gMjl15x=0 if Tijmin15<Xi
egen sgMjl15=max(gMjl15x), by(tj Tijmin15)
drop gMjl15x

gsort tj -Tijmin16
by tj: gen gMjl16x=sum(gMjl16)
replace gMjl16x=0 if Tijmin16<Xi
egen sgMjl16=max(gMjl16x), by(tj Tijmin16)
drop gMjl16x

gsort tj -Tijmin17
by tj: gen gMjl17x=sum(gMjl17)
replace gMjl17x=0 if Tijmin17<Xi
egen sgMjl17=max(gMjl17x), by(tj Tijmin17)
drop gMjl17x

gsort tj -Tijmin18
by tj: gen gMjl18x=sum(gMjl18)
replace gMjl18x=0 if Tijmin18<Xi
egen sgMjl18=max(gMjl18x), by(tj Tijmin18)
drop gMjl18x

gsort tj -Tijmin19
by tj: gen gMjl19x=sum(gMjl19)
replace gMjl19x=0 if Tijmin19<Xi
egen sgMjl19=max(gMjl19x), by(tj Tijmin19)
drop gMjl19x

*
gsort tj -Tijmin1
by tj: gen gMl1x=sum(gMl1)
replace gMl1x=0 if Tijmin1<Xi
egen sgMl1=max(gMl1x), by(tj Tijmin1)
drop gMl1x

gsort tj -Tijmin2
by tj: gen gMl2x=sum(gMl2)
replace gMl2x=0 if Tijmin2<Xi
egen sgMl2=max(gMl2x), by(tj Tijmin2)
drop gMl2x

gsort tj -Tijmin3
by tj: gen gMl3x=sum(gMl3)
replace gMl3x=0 if Tijmin3<Xi
egen sgMl3=max(gMl3x), by(tj Tijmin3)
drop gMl3x

gsort tj -Tijmin4
```

```
by tj: gen gMl4x=sum(gMl4)
replace gMl4x=0 if Tijmin4<Xi
egen sgMl4=max(gMl4x), by(tj Tijmin4)
drop gMl4x

gsort tj -Tijmin5
by tj: gen gMl5x=sum(gMl5)
replace gMl5x=0 if Tijmin5<Xi
egen sgMl5=max(gMl5x), by(tj Tijmin5)
drop gMl5x

gsort tj -Tijmin6
by tj: gen gMl6x=sum(gMl6)
replace gMl6x=0 if Tijmin6<Xi
egen sgMl6=max(gMl6x), by(tj Tijmin6)
drop gMl6x

gsort tj -Tijmin7
by tj: gen gMl7x=sum(gMl7)
replace gMl7x=0 if Tijmin7<Xi
egen sgMl7=max(gMl7x), by(tj Tijmin7)
drop gMl7x

gsort tj -Tijmin8
by tj: gen gMl8x=sum(gMl8)
replace gMl8x=0 if Tijmin8<Xi
egen sgMl8=max(gMl8x), by(tj Tijmin8)
drop gMl8x

gsort tj -Tijmin9
by tj: gen gMl9x=sum(gMl9)
replace gMl9x=0 if Tijmin9<Xi
egen sgMl9=max(gMl9x), by(tj Tijmin9)
drop gMl9x

gsort tj -Tijmin10
by tj: gen gMl10x=sum(gMl10)
replace gMl10x=0 if Tijmin10<Xi
egen sgMl10=max(gMl10x), by(tj Tijmin10)
drop gMl10x

gsort tj -Tijmin11
by tj: gen gMl11x=sum(gMl11)
replace gMl11x=0 if Tijmin11<Xi
egen sgMl11=max(gMl11x), by(tj Tijmin11)
drop gMl11x

gsort tj -Tijmin12
by tj: gen gMl12x=sum(gMl12)
replace gMl12x=0 if Tijmin12<Xi
egen sgMl12=max(gMl12x), by(tj Tijmin12)
drop gMl12x

gsort tj -Tijmin13
by tj: gen gMl13x=sum(gMl13)
replace gMl13x=0 if Tijmin13<Xi
egen sgMl13=max(gMl13x), by(tj Tijmin13)
drop gMl13x

gsort tj -Tijmin14
by tj: gen gMl14x=sum(gMl14)
replace gMl14x=0 if Tijmin14<Xi
egen sgMl14=max(gMl14x), by(tj Tijmin14)
drop gMl14x
```

```
gsort tj -Tijmin15
by tj: gen gMl15x=sum(gMl15)
replace gMl15x=0 if Tijmin15<Xi
egen sgMl15=max(gMl15x), by(tj Tijmin15)
drop gMl15x

gsort tj -Tijmin16
by tj: gen gMl16x=sum(gMl16)
replace gMl16x=0 if Tijmin16<Xi
egen sgMl16=max(gMl16x), by(tj Tijmin16)
drop gMl16x

gsort tj -Tijmin17
by tj: gen gMl17x=sum(gMl17)
replace gMl17x=0 if Tijmin17<Xi
egen sgMl17=max(gMl17x), by(tj Tijmin17)
drop gMl17x

gsort tj -Tijmin18
by tj: gen gMl18x=sum(gMl18)
replace gMl18x=0 if Tijmin18<Xi
egen sgMl18=max(gMl18x), by(tj Tijmin18)
drop gMl18x

gsort tj -Tijmin19
by tj: gen gMl19x=sum(gMl19)
replace gMl19x=0 if Tijmin19<Xi
egen sgMl19=max(gMl19x), by(tj Tijmin19)
drop gMl19x

*
gen sxjl1=((1/1138)* sgMjMl1-(1/1138)* sgMjl1*(1/1138)*(1/ Sjl1u)* sgMl1)
replace sxjl1=0 if sxjl1==.

gen sxjl2=((1/1138)* sgMjMl2-(1/1138)* sgMjl2*(1/1138)*(1/ Sjl2u)* sgMl2)
replace sxjl2=0 if sxjl2==.

gen sxjl3=((1/1138)* sgMjMl3-(1/1138)* sgMjl3*(1/1138)*(1/ Sjl3u)* sgMl3)
replace sxjl3=0 if sxjl3==.

gen sxjl4=((1/1138)* sgMjMl4-(1/1138)* sgMjl4*(1/1138)*(1/ Sjl4u)* sgMl4)
replace sxjl4=0 if sxjl4==.

gen sxjl5=((1/1138)* sgMjMl5-(1/1138)* sgMjl5*(1/1138)*(1/ Sjl5u)* sgMl5)
replace sxjl5=0 if sxjl5==.

gen sxjl6=((1/1138)* sgMjMl6-(1/1138)* sgMjl6*(1/1138)*(1/ Sjl6u)* sgMl6)
replace sxjl6=0 if sxjl6==.

gen sxjl7=((1/1138)* sgMjMl7-(1/1138)* sgMjl7*(1/1138)*(1/ Sjl7u)* sgMl7)
replace sxjl7=0 if sxjl7==.

gen sxjl8=((1/1138)* sgMjMl8-(1/1138)* sgMjl8*(1/1138)*(1/ Sjl8u)* sgMl8)
replace sxjl8=0 if sxjl8==.

gen sxjl9=((1/1138)* sgMjMl9-(1/1138)* sgMjl9*(1/1138)*(1/ Sjl9u)* sgMl9)
replace sxjl9=0 if sxjl9==.

gen sxjl10=((1/1138)* sgMjMl10-(1/1138)* sgMjl10*(1/1138)*(1/ Sjl10u)* sgMl10)
replace sxjl10=0 if sxjl10==.

gen sxjl11=((1/1138)* sgMjMl11-(1/1138)* sgMjl11*(1/1138)*(1/ Sjl11u)* sgMl11)
replace sxjl11=0 if sxjl11==.
```

```
gen sxjl12=((1/1138)* sgMjMl12-(1/1138)* sgMjl12*(1/1138)*(1/ Sjl12u)* sgMl12)
replace sxjl12=0 if sxjl12==.

gen sxjl13=((1/1138)* sgMjMl13-(1/1138)* sgMjl13*(1/1138)*(1/ Sjl13u)* sgMl13)
replace sxjl13=0 if sxjl13==.

gen sxjl14=((1/1138)* sgMjMl14-(1/1138)* sgMjl14*(1/1138)*(1/ Sjl14u)* sgMl14)
replace sxjl14=0 if sxjl14==.

gen sxjl15=((1/1138)* sgMjMl15-(1/1138)* sgMjl15*(1/1138)*(1/ Sjl15u)* sgMl15)
replace sxjl15=0 if sxjl15==.

gen sxjl16=((1/1138)* sgMjMl16-(1/1138)* sgMjl16*(1/1138)*(1/ Sjl16u)* sgMl16)
replace sxjl16=0 if sxjl16==.

gen sxjl17=((1/1138)* sgMjMl17-(1/1138)* sgMjl17*(1/1138)*(1/ Sjl17u)* sgMl17)
replace sxjl17=0 if sxjl17==.

gen sxjl18=((1/1138)* sgMjMl18-(1/1138)* sgMjl18*(1/1138)*(1/ Sjl18u)* sgMl18)
replace sxjl18=0 if sxjl18==.

gen sxjl19=((1/1138)* sgMjMl19-(1/1138)* sgMjl19*(1/1138)*(1/ Sjl19u)* sgMl19)
replace sxjl19=0 if sxjl19==.


gen sumsxls= sxjl1+sxjl2+sxjl3+sxjl4+sxjl5+sxjl6+sxjl7+sxjl8+sxjl9+sxjl10+
sxjl11+sxjl12+sxjl13+sxjl14+sxjl15+sxjl16+sxjl17+sxjl18+sxjl19

egen sumsall=sum(sumsxls), by(ukno)

collapse sumintjs sumsall di censorig Mi cens1138 surv1138 censor1 Yu, by(ukno)

label var di "censorig"

**

egen mpx=sum(sumintjs)
gen meanpart=(1/1138)*mpx

*
gen intern1=di*(( Mi- meanpart)^2)/ cens1138
replace intern1=0 if intern1==.

egen sumint1=sum(intern1)

gen term1=(1/1138)* sumint1

gen intern2=(censor1/ (Yu* cens1138))* sumsall
replace intern2=0 if intern2==.

egen sumint2=sum( intern2)

gen term2=sumint2

gen varpart=(1/1138)*( term1+ term2)

gen separt=sqrt(varpart)


**
```

Simple improved estimator

Based on equations (4.42), (4.43), (4.45) and (4.46) for the mean and on equations (4.44), (4.45) and (4.46) for the variance

For Conventional (similarly for intensive)

```
**
gen tj_1=year-1
gen tj=year

gen Mij=costyr
replace Mij=0 if Mij==.

egen di=min(censorig), by(ukno)
egen Xi=min( timallde), by(ukno)

gen ejMi=Mij if Xi>tj
replace ejMi=0 if ejMi==.

gsort tj -Xi
by tj: gen Sejx=sum( ejMi)
egen Sej=max( Sejx), by(tj Xi)
drop  Sejx

gen gejM=Sej/Yu

gen difejg= ejMi- gejM

gen internj=(censor1/cens1138)* difejg
replace internj=0 if internj==.

egen sinternj=sum(internj), by(tj)


**cov vector**

gen cova1=(di*Mi)/cens1138
replace cova1=0 if cova1==.

gsort tj -Xi
by tj: gen scova1x=sum( cova1)
egen scova1=max( scova1x), by(tj Xi)
*drop scova1x*

gen gMu=(1/1138)*(1/surv1138)* scova1
replace gMu=0 if gMu==.

gen cova2= Mi-gMu

gen cova3=(di*Mij)/cens1138
replace cova3=0 if cova3==.

gsort tj -Xi
by tj: gen scova3x=sum( cova3)
egen scova3=max( scova3x), by(tj Xi)
*drop scova3x*

gen gMju=(1/1138)*(1/surv1138)* scova3
replace gMju=0 if gMju==.

gen cova4= Mij-gMju
```

```
gen cova5=( di/ cens1138)* cova2* cova4
replace cova5=0 if cova5==.

**ties**

gen int const=1
sort tj Xi
by tj Xi: gen ties=sum(const)
egen maxties=max(ties), by(tj Xi)

**

gsort tj -Xi ties
by tj: gen scova5x=sum( cova5)
gen scova5=scova5x if ties==maxties
egen scova5xx=min( scova5), by(tj Xi maxties)
replace scova5=scova5xx if scova5==.
*drop scova5x scova5xx*

gen cova6=(1/1138)*(1/surv1138)* scova5

gen cova7=( censor1/(cens1138^2))* cova6
replace cova7=0 if cova7==.

egen scova7=sum(cova7), by(tj)

gen covj=(1/1138)*scova7


**Variance vector**

sort ukno tj

gen a4lead1=cova4
quietly by ukno: gen a4lead2=a4lead1[_n+1]
quietly by ukno: gen a4lead3=a4lead2[_n+1]
quietly by ukno: gen a4lead4=a4lead3[_n+1]
quietly by ukno: gen a4lead5=a4lead4[_n+1]
quietly by ukno: gen a4lead6=a4lead5[_n+1]
quietly by ukno: gen a4lead7=a4lead6[_n+1]
quietly by ukno: gen a4lead8=a4lead7[_n+1]
quietly by ukno: gen a4lead9=a4lead8[_n+1]
quietly by ukno: gen a4lead10=a4lead9[_n+1]
quietly by ukno: gen a4lead11=a4lead10[_n+1]
quietly by ukno: gen a4lead12=a4lead11[_n+1]
quietly by ukno: gen a4lead13=a4lead12[_n+1]
quietly by ukno: gen a4lead14=a4lead13[_n+1]
quietly by ukno: gen a4lead15=a4lead14[_n+1]
quietly by ukno: gen a4lead16=a4lead15[_n+1]
quietly by ukno: gen a4lead17=a4lead16[_n+1]
quietly by ukno: gen a4lead18=a4lead17[_n+1]
quietly by ukno: gen a4lead19=a4lead18[_n+1]

replace  a4lead1=. if tj~=1
replace  a4lead2=. if tj~=1
replace  a4lead3=. if tj~=1
replace  a4lead4=. if tj~=1
replace  a4lead5=. if tj~=1
replace  a4lead6=. if tj~=1
replace  a4lead7=. if tj~=1
replace  a4lead8=. if tj~=1
replace  a4lead9=. if tj~=1
replace  a4lead10=. if tj~=1
replace  a4lead11=. if tj~=1
```

```
replace  a4lead12=. if tj~=1
replace  a4lead13=. if tj~=1
replace  a4lead14=. if tj~=1
replace  a4lead15=. if tj~=1
replace  a4lead16=. if tj~=1
replace  a4lead17=. if tj~=1
replace  a4lead18=. if tj~=1
replace  a4lead19=. if tj~=1

egen a41=min( a4lead1), by(ukno)
egen a42=min( a4lead2), by(ukno)
egen a43=min( a4lead3), by(ukno)
egen a44=min( a4lead4), by(ukno)
egen a45=min( a4lead5), by(ukno)
egen a46=min( a4lead6), by(ukno)
egen a47=min( a4lead7), by(ukno)
egen a48=min( a4lead8), by(ukno)
egen a49=min( a4lead9), by(ukno)
egen a410=min( a4lead10), by(ukno)
egen a411=min( a4lead11), by(ukno)
egen a412=min( a4lead12), by(ukno)
egen a413=min( a4lead13), by(ukno)
egen a414=min( a4lead14), by(ukno)
egen a415=min( a4lead15), by(ukno)
egen a416=min( a4lead16), by(ukno)
egen a417=min( a4lead17), by(ukno)
egen a418=min( a4lead18), by(ukno)
egen a419=min( a4lead19), by(ukno)

drop a4lead*

gen vara41=cova4*a41
gen vara42=cova4*a42
gen vara43=cova4*a43
gen vara44=cova4*a44
gen vara45=cova4*a45
gen vara46=cova4*a46
gen vara47=cova4*a47
gen vara48=cova4*a48
gen vara49=cova4*a49
gen vara410=cova4*a410
gen vara411=cova4*a411
gen vara412=cova4*a412
gen vara413=cova4*a413
gen vara414=cova4*a414
gen vara415=cova4*a415
gen vara416=cova4*a416
gen vara417=cova4*a417
gen vara418=cova4*a418
gen vara419=cova4*a419


gen vara51=(di/cens1138)*vara41
replace vara51=0 if vara51==.

gen vara52=(di/cens1138)*vara42
replace vara52=0 if vara52==.

gen vara53=(di/cens1138)*vara43
replace vara53=0 if vara53==.

gen vara54=(di/cens1138)*vara44
replace vara54=0 if vara54==.
```

```
gen vara55=(di/cens1138)*vara45
replace vara55=0 if vara55==.

gen vara56=(di/cens1138)*vara46
replace vara56=0 if vara56==.

gen vara57=(di/cens1138)*vara47
replace vara57=0 if vara57==.

gen vara58=(di/cens1138)*vara48
replace vara58=0 if vara58==.

gen vara59=(di/cens1138)*vara49
replace vara59=0 if vara59==.

gen vara510=(di/cens1138)*vara410
replace vara510=0 if vara510==.

gen vara511=(di/cens1138)*vara411
replace vara511=0 if vara511==.

gen vara512=(di/cens1138)*vara412
replace vara512=0 if vara512==.

gen vara513=(di/cens1138)*vara413
replace vara513=0 if vara513==.

gen vara514=(di/cens1138)*vara414
replace vara514=0 if vara514==.

gen vara515=(di/cens1138)*vara415
replace vara515=0 if vara515==.

gen vara516=(di/cens1138)*vara416
replace vara516=0 if vara516==.

gen vara517=(di/cens1138)*vara417
replace vara517=0 if vara517==.

gen vara518=(di/cens1138)*vara418
replace vara518=0 if vara518==.

gen vara519=(di/cens1138)*vara419
replace vara519=0 if vara519==.

*
gsort tj -Xi ties
by tj: gen svara51x=sum( vara51)
gen svara51=svara51x if ties==maxties
egen svara51xx=min( svara51), by(tj Xi maxties)
replace svara51=svara51xx if svara51==.
drop svara51x svara51xx

gsort tj -Xi ties
by tj: gen svara52x=sum( vara52)
gen svara52=svara52x if ties==maxties
egen svara52xx=min( svara52), by(tj Xi maxties)
replace svara52=svara52xx if svara52==.
drop svara52x svara52xx

gsort tj -Xi ties
by tj: gen svara53x=sum( vara53)
gen svara53=svara53x if ties==maxties
egen svara53xx=min( svara53), by(tj Xi maxties)
```

```
replace svara53=svara53xx if svara53==.
drop svara53x svara53xx

gsort tj -Xi ties
by tj: gen svara54x=sum( vara54)
gen svara54=svara54x if ties==maxties
egen svara54xx=min( svara54), by(tj Xi maxties)
replace svara54=svara54xx if svara54==.
drop svara54x svara54xx

gsort tj -Xi ties
by tj: gen svara55x=sum( vara55)
gen svara55=svara55x if ties==maxties
egen svara55xx=min( svara55), by(tj Xi maxties)
replace svara55=svara55xx if svara55==.
drop svara55x svara55xx

gsort tj -Xi ties
by tj: gen svara56x=sum( vara56)
gen svara56=svara56x if ties==maxties
egen svara56xx=min( svara56), by(tj Xi maxties)
replace svara56=svara56xx if svara56==.
drop svara56x svara56xx

gsort tj -Xi ties
by tj: gen svara57x=sum( vara57)
gen svara57=svara57x if ties==maxties
egen svara57xx=min( svara57), by(tj Xi maxties)
replace svara57=svara57xx if svara57==.
drop svara57x svara57xx

gsort tj -Xi ties
by tj: gen svara58x=sum( vara58)
gen svara58=svara58x if ties==maxties
egen svara58xx=min( svara58), by(tj Xi maxties)
replace svara58=svara58xx if svara58==.
drop svara58x svara58xx

gsort tj -Xi ties
by tj: gen svara59x=sum( vara59)
gen svara59=svara59x if ties==maxties
egen svara59xx=min( svara59), by(tj Xi maxties)
replace svara59=svara59xx if svara59==.
drop svara59x svara59xx

gsort tj -Xi ties
by tj: gen svara510x=sum( vara510)
gen svara510=svara510x if ties==maxties
egen svara510xx=min( svara510), by(tj Xi maxties)
replace svara510=svara510xx if svara510==.
drop svara510x svara510xx

gsort tj -Xi ties
by tj: gen svara511x=sum( vara511)
gen svara511=svara511x if ties==maxties
egen svara511xx=min( svara511), by(tj Xi maxties)
replace svara511=svara511xx if svara511==.
drop svara511x svara511xx

gsort tj -Xi ties
by tj: gen svara512x=sum( vara512)
gen svara512=svara512x if ties==maxties
egen svara512xx=min( svara512), by(tj Xi maxties)
replace svara512=svara512xx if svara512==.
```

```
drop svara512x svara512xx

gsort tj -Xi ties
by tj: gen svara513x=sum( vara513)
gen svara513=svara513x if ties==maxties
egen svara513xx=min( svara513), by(tj Xi maxties)
replace svara513=svara513xx if svara513==.
drop svara513x svara513xx

gsort tj -Xi ties
by tj: gen svara514x=sum( vara514)
gen svara514=svara514x if ties==maxties
egen svara514xx=min( svara514), by(tj Xi maxties)
replace svara514=svara514xx if svara514==.
drop svara514x svara514xx

gsort tj -Xi ties
by tj: gen svara515x=sum( vara515)
gen svara515=svara515x if ties==maxties
egen svara515xx=min( svara515), by(tj Xi maxties)
replace svara515=svara515xx if svara515==.
drop svara515x svara515xx

gsort tj -Xi ties
by tj: gen svara516x=sum( vara516)
gen svara516=svara516x if ties==maxties
egen svara516xx=min( svara516), by(tj Xi maxties)
replace svara516=svara516xx if svara516==.
drop svara516x svara516xx

gsort tj -Xi ties
by tj: gen svara517x=sum( vara517)
gen svara517=svara517x if ties==maxties
egen svara517xx=min( svara517), by(tj Xi maxties)
replace svara517=svara517xx if svara517==.
drop svara517x svara517xx

gsort tj -Xi ties
by tj: gen svara518x=sum( vara518)
gen svara518=svara518x if ties==maxties
egen svara518xx=min( svara518), by(tj Xi maxties)
replace svara518=svara518xx if svara518==.
drop svara518x svara518xx

gsort tj -Xi ties
by tj: gen svara519x=sum( vara519)
gen svara519=svara519x if ties==maxties
egen svara519xx=min( svara519), by(tj Xi maxties)
replace svara519=svara519xx if svara519==.
drop svara519x svara519xx


gen vara61=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara51
replace vara61=0 if vara61==.

gen vara62=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara52
replace vara62=0 if vara62==.

gen vara63=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara53
replace vara63=0 if vara63==.

gen vara64=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara54
replace vara64=0 if vara64==.
```

```
gen vara65=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara55
replace vara65=0 if vara65==.

gen vara66=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara56
replace vara66=0 if vara66==.

gen vara67=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara57
replace vara67=0 if vara67==.

gen vara68=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara58
replace vara68=0 if vara68==.

gen vara69=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara59
replace vara69=0 if vara69==.

gen vara610=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara510
replace vara610=0 if vara610==.

gen vara611=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara511
replace vara611=0 if vara611==.

gen vara612=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara512
replace vara612=0 if vara612==.

gen vara613=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara513
replace vara613=0 if vara613==.

gen vara614=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara514
replace vara614=0 if vara614==.

gen vara615=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara515
replace vara615=0 if vara615==.

gen vara616=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara516
replace vara616=0 if vara616==.

gen vara617=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara517
replace vara617=0 if vara617==.

gen vara618=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara518
replace vara618=0 if vara618==.

gen vara619=(censor1/(cens1138^2))*(1/1138)*(1/surv1138)*svara519
replace vara619=0 if vara619==.

egen var71=sum( vara61), by(tj)
egen var72=sum( vara62), by(tj)
egen var73=sum( vara63), by(tj)
egen var74=sum( vara64), by(tj)
egen var75=sum( vara65), by(tj)
egen var76=sum( vara66), by(tj)
egen var77=sum( vara67), by(tj)
egen var78=sum( vara68), by(tj)
egen var79=sum( vara69), by(tj)
egen var710=sum( vara610), by(tj)
egen var711=sum( vara611), by(tj)
egen var712=sum( vara612), by(tj)
egen var713=sum( vara613), by(tj)
egen var714=sum( vara614), by(tj)
egen var715=sum( vara615), by(tj)
egen var716=sum( vara616), by(tj)
egen var717=sum( vara617), by(tj)
egen var718=sum( vara618), by(tj)
egen var719=sum( vara619), by(tj)
```

```
gen varjl1=(1/1138)* var71
gen varjl2=(1/1138)* var72
gen varjl3=(1/1138)* var73
gen varjl4=(1/1138)* var74
gen varjl5=(1/1138)* var75
gen varjl6=(1/1138)* var76
gen varjl7=(1/1138)* var77
gen varjl8=(1/1138)* var78
gen varjl9=(1/1138)* var79
gen varjl10=(1/1138)* var710
gen varjl11=(1/1138)* var711
gen varjl12=(1/1138)* var712
gen varjl13=(1/1138)* var713
gen varjl14=(1/1138)* var714
gen varjl15=(1/1138)* var715
gen varjl16=(1/1138)* var716
gen varjl17=(1/1138)* var717
gen varjl18=(1/1138)* var718
gen varjl19=(1/1138)* var719

*
** Matrix calculations for Simple Improved for conventional**

keep if ukno=="00011D"
keep  ukno year sinternj covj varj*
mkmat varjl*,matrix(var)
matrix invvar=syminv(var)
svmat invvar
mkmat covj, matrix(covs)
matrix Gs=covs'*invvar
svmat Gs
mkmat sinternj, matrix(A)
matrix ws=Gs*A
svmat ws
gen wsterm= ws1/1138

matrix varsterm=Gs*covs
svmat varsterm

**

gen msimpimp= meansimp-xxx
gen intern1imp= censorig*((Mi- msimpimp)^2)/ cens1138
replace intern1imp=0 if intern1imp==.
egen sumint1imp=sum(intern1imp)
gen term1imp=(1/1138)*sumint1imp
gen term3imp=xxxx

gen varsimpimp=(1/1138)*(term1imp+term2- term3imp)
gen sesimpimp=sqrt( varsimpimp)

**
```

Improved Partitioned estimator

Based on equations (4.47), (4.42) (4.46) and (4.48) for the mean and on equations (4.49), (4.46) and (4.48) for the variance

For Conventional (similarly for intensive)

```
** Improved Partitioned: Covariance vector: Conventional**

gen tj_1=year-1
gen tj=year

egen Xi=min(timallde), by(ukno)
egen di=min(censorig), by(ukno)

gen Mij=costyr
replace Mij=0 if Mij==.

gen minTitj=min(Xi, tj)

gen Xij=min(minTitj, Xi)

gen dij=1 if (minTitj==tj | (minTitj==Xi & di==1))
replace dij=0 if dij==.

stset Xij, failure(dij==0)
sts gen KjTij=s, by(tj)

gen intern=(dij*Mij)/KjTij
replace intern=0 if intern==.

gen dijsurv=1 if (minTitj==Xi & di==1)
replace dijsurv=0 if dijsurv==.

stset Xij, failure(dij==1)
sts gen Slu=s, by(tj)

gsort tj -Xi
by tj: gen glMlx=sum(intern)
replace glMlx=0 if Xij<Xi
egen sglMl=max(glMlx), by(tj Xi)
replace sglMl=0 if Xij<Xi

gen glMl=(1/1138)*(1/Slu)*sglMl
replace glMl=0 if glMl==.

gen Mil_gl=Mij-glMl

gen l1=1
gen l2=2
gen l3=3
gen l4=4
gen l5=5
gen l6=6
gen l7=7
gen l8=8
gen l9=9
gen l10=10
gen l11=11
gen l12=12
gen l13=13
gen l14=14
```

```
gen l15=15
gen l16=16
gen l17=17
gen l18=18
gen l19=19

gen jmaxl1=max(tj, l1)
gen jmaxl2=max(tj, l2)
gen jmaxl3=max(tj, l3)
gen jmaxl4=max(tj, l4)
gen jmaxl5=max(tj, l5)
gen jmaxl6=max(tj, l6)
gen jmaxl7=max(tj, l7)
gen jmaxl8=max(tj, l8)
gen jmaxl9=max(tj, l9)
gen jmaxl10=max(tj, l10)
gen jmaxl11=max(tj, l11)
gen jmaxl12=max(tj, l12)
gen jmaxl13=max(tj, l13)
gen jmaxl14=max(tj, l14)
gen jmaxl15=max(tj, l15)
gen jmaxl16=max(tj, l16)
gen jmaxl17=max(tj, l17)
gen jmaxl18=max(tj, l18)
gen jmaxl19=max(tj, l19)


gen Tijmax1=min(Xi, jmaxl1)
gen Tijmax2=min(Xi, jmaxl2)
gen Tijmax3=min(Xi, jmaxl3)
gen Tijmax4=min(Xi, jmaxl4)
gen Tijmax5=min(Xi, jmaxl5)
gen Tijmax6=min(Xi, jmaxl6)
gen Tijmax7=min(Xi, jmaxl7)
gen Tijmax8=min(Xi, jmaxl8)
gen Tijmax9=min(Xi, jmaxl9)
gen Tijmax10=min(Xi, jmaxl10)
gen Tijmax11=min(Xi, jmaxl11)
gen Tijmax12=min(Xi, jmaxl12)
gen Tijmax13=min(Xi, jmaxl13)
gen Tijmax14=min(Xi, jmaxl14)
gen Tijmax15=min(Xi, jmaxl15)
gen Tijmax16=min(Xi, jmaxl16)
gen Tijmax17=min(Xi, jmaxl17)
gen Tijmax18=min(Xi, jmaxl18)
gen Tijmax19=min(Xi, jmaxl19)


gen dijmax1=1 if Tijmax1==jmaxl1 | (Tijmax1==Xi & di==1)
replace dijmax1=0 if dijmax1==.

gen dijmax2=1 if Tijmax2==jmaxl2 | (Tijmax2==Xi & di==1)
replace dijmax2=0 if dijmax2==.

gen dijmax3=1 if Tijmax3==jmaxl3 | (Tijmax3==Xi & di==1)
replace dijmax3=0 if dijmax3==.

gen dijmax4=1 if Tijmax4==jmaxl4 | (Tijmax4==Xi & di==1)
replace dijmax4=0 if dijmax4==.

gen dijmax5=1 if Tijmax5==jmaxl5 | (Tijmax5==Xi & di==1)
replace dijmax5=0 if dijmax5==.

gen dijmax6=1 if Tijmax6==jmaxl6 | (Tijmax6==Xi & di==1)
```

```
replace dijmax6=0 if dijmax6==.

gen dijmax7=1 if Tijmax7==jmaxl7 | (Tijmax7==Xi & di==1)
replace dijmax7=0 if dijmax7==.

gen dijmax8=1 if Tijmax8==jmaxl8 | (Tijmax8==Xi & di==1)
replace dijmax8=0 if dijmax8==.

gen dijmax9=1 if Tijmax9==jmaxl9 | (Tijmax9==Xi & di==1)
replace dijmax9=0 if dijmax9==.

gen dijmax10=1 if Tijmax10==jmaxl10 | (Tijmax10==Xi & di==1)
replace dijmax10=0 if dijmax10==.

gen dijmax11=1 if Tijmax11==jmaxl11 | (Tijmax11==Xi & di==1)
replace dijmax11=0 if dijmax11==.

gen dijmax12=1 if Tijmax12==jmaxl12 | (Tijmax12==Xi & di==1)
replace dijmax12=0 if dijmax12==.

gen dijmax13=1 if Tijmax13==jmaxl13 | (Tijmax13==Xi & di==1)
replace dijmax13=0 if dijmax13==.

gen dijmax14=1 if Tijmax14==jmaxl14 | (Tijmax14==Xi & di==1)
replace dijmax14=0 if dijmax14==.

gen dijmax15=1 if Tijmax15==jmaxl15 | (Tijmax15==Xi & di==1)
replace dijmax15=0 if dijmax15==.

gen dijmax16=1 if Tijmax16==jmaxl16 | (Tijmax16==Xi & di==1)
replace dijmax16=0 if dijmax16==.

gen dijmax17=1 if Tijmax17==jmaxl17 | (Tijmax17==Xi & di==1)
replace dijmax17=0 if dijmax17==.

gen dijmax18=1 if Tijmax18==jmaxl18 | (Tijmax18==Xi & di==1)
replace dijmax18=0 if dijmax18==.

gen dijmax19=1 if Tijmax19==jmaxl19 | (Tijmax19==Xi & di==1)
replace dijmax19=0 if dijmax19==.


*

stset Tijmax1, failure(dijmax1==0)
sts gen Kjmaxl1=s, by(tj)

stset Tijmax2, failure(dijmax2==0)
sts gen Kjmaxl2=s, by(tj)

stset Tijmax3, failure(dijmax3==0)
sts gen Kjmaxl3=s, by(tj)

stset Tijmax4, failure(dijmax4==0)
sts gen Kjmaxl4=s, by(tj)

stset Tijmax5, failure(dijmax5==0)
sts gen Kjmaxl5=s, by(tj)

stset Tijmax6, failure(dijmax6==0)
sts gen Kjmaxl6=s, by(tj)

stset Tijmax7, failure(dijmax7==0)
sts gen Kjmaxl7=s, by(tj)
```

```
stset Tijmax8, failure(dijmax8==0)
sts gen Kjmaxl8=s, by(tj)

stset Tijmax9, failure(dijmax9==0)
sts gen Kjmaxl9=s, by(tj)

stset Tijmax10, failure(dijmax10==0)
sts gen Kjmaxl10=s, by(tj)

stset Tijmax11, failure(dijmax11==0)
sts gen Kjmaxl11=s, by(tj)

stset Tijmax12, failure(dijmax12==0)
sts gen Kjmaxl12=s, by(tj)

stset Tijmax13, failure(dijmax13==0)
sts gen Kjmaxl13=s, by(tj)

stset Tijmax14, failure(dijmax14==0)
sts gen Kjmaxl14=s, by(tj)

stset Tijmax15, failure(dijmax15==0)
sts gen Kjmaxl15=s, by(tj)

stset Tijmax16, failure(dijmax16==0)
sts gen Kjmaxl16=s, by(tj)

stset Tijmax17, failure(dijmax17==0)
sts gen Kjmaxl17=s, by(tj)

stset Tijmax18, failure(dijmax18==0)
sts gen Kjmaxl18=s, by(tj)

stset Tijmax19, failure(dijmax19==0)
sts gen Kjmaxl19=s, by(tj)

*

sort ukno tj

gen Mi1lead=Mij

quietly by ukno: gen Mi2lead=Mi1lead[_n+1]
quietly by ukno: gen Mi3lead=Mi2lead[_n+1]
quietly by ukno: gen Mi4lead=Mi3lead[_n+1]
quietly by ukno: gen Mi5lead=Mi4lead[_n+1]
quietly by ukno: gen Mi6lead=Mi5lead[_n+1]
quietly by ukno: gen Mi7lead=Mi6lead[_n+1]
quietly by ukno: gen Mi8lead=Mi7lead[_n+1]
quietly by ukno: gen Mi9lead=Mi8lead[_n+1]
quietly by ukno: gen Mi10lead=Mi9lead[_n+1]
quietly by ukno: gen Mi11lead=Mi10lead[_n+1]
quietly by ukno: gen Mi12lead=Mi11lead[_n+1]
quietly by ukno: gen Mi13lead=Mi12lead[_n+1]
quietly by ukno: gen Mi14lead=Mi13lead[_n+1]
quietly by ukno: gen Mi15lead=Mi14lead[_n+1]
quietly by ukno: gen Mi16lead=Mi15lead[_n+1]
quietly by ukno: gen Mi17lead=Mi16lead[_n+1]
quietly by ukno: gen Mi18lead=Mi17lead[_n+1]
quietly by ukno: gen Mi19lead=Mi18lead[_n+1]

replace  Mi1lead=-9 if tj~=1
replace  Mi2lead=-9 if tj~=1
```

```
replace  Mi3lead=-9 if tj~=1
replace  Mi4lead=-9 if tj~=1
replace  Mi5lead=-9 if tj~=1
replace  Mi6lead=-9 if tj~=1
replace  Mi7lead=-9 if tj~=1
replace  Mi8lead=-9 if tj~=1
replace  Mi9lead=-9 if tj~=1
replace  Mi10lead=-9 if tj~=1
replace  Mi11lead=-9 if tj~=1
replace  Mi12lead=-9 if tj~=1
replace  Mi13lead=-9 if tj~=1
replace  Mi14lead=-9 if tj~=1
replace  Mi15lead=-9 if tj~=1
replace  Mi16lead=-9 if tj~=1
replace  Mi17lead=-9 if tj~=1
replace  Mi18lead=-9 if tj~=1
replace  Mi19lead=-9 if tj~=1

egen Mi1=max(Mi1lead), by(ukno)
egen Mi2=max(Mi2lead), by(ukno)
egen Mi3=max(Mi3lead), by(ukno)
egen Mi4=max(Mi4lead), by(ukno)
egen Mi5=max(Mi5lead), by(ukno)
egen Mi6=max(Mi6lead), by(ukno)
egen Mi7=max(Mi7lead), by(ukno)
egen Mi8=max(Mi8lead), by(ukno)
egen Mi9=max(Mi9lead), by(ukno)
egen Mi10=max(Mi10lead), by(ukno)
egen Mi11=max(Mi11lead), by(ukno)
egen Mi12=max(Mi12lead), by(ukno)
egen Mi13=max(Mi13lead), by(ukno)
egen Mi14=max(Mi14lead), by(ukno)
egen Mi15=max(Mi15lead), by(ukno)
egen Mi16=max(Mi16lead), by(ukno)
egen Mi17=max(Mi17lead), by(ukno)
egen Mi18=max(Mi18lead), by(ukno)
egen Mi19=max(Mi19lead), by(ukno)


gen mu1=(di*Mi1)/cens1138
replace mu1=0 if mu1==.

gen mu2=(di*Mi2)/cens1138
replace mu2=0 if mu2==.

gen mu3=(di*Mi3)/cens1138
replace mu3=0 if mu3==.

gen mu4=(di*Mi4)/cens1138
replace mu4=0 if mu4==.

gen mu5=(di*Mi5)/cens1138
replace mu5=0 if mu5==.

gen mu6=(di*Mi6)/cens1138
replace mu6=0 if mu6==.

gen mu7=(di*Mi7)/cens1138
replace mu7=0 if mu7==.

gen mu8=(di*Mi8)/cens1138
replace mu8=0 if mu8==.

gen mu9=(di*Mi9)/cens1138
```

```
replace mu9=0 if mu9==.

gen mu10=(di*Mi10)/cens1138
replace mu10=0 if mu10==.

gen mu11=(di*Mi11)/cens1138
replace mu11=0 if mu11==.

gen mu12=(di*Mi12)/cens1138
replace mu12=0 if mu12==.

gen mu13=(di*Mi13)/cens1138
replace mu13=0 if mu13==.

gen mu14=(di*Mi14)/cens1138
replace mu14=0 if mu14==.

gen mu15=(di*Mi15)/cens1138
replace mu15=0 if mu15==.

gen mu16=(di*Mi16)/cens1138
replace mu16=0 if mu16==.

gen mu17=(di*Mi17)/cens1138
replace mu17=0 if mu17==.

gen mu18=(di*Mi18)/cens1138
replace mu18=0 if mu18==.

gen mu19=(di*Mi19)/cens1138
replace mu19=0 if mu19==.


*

gsort tj -Xi
by tj: gen smu1x=sum(mu1)
egen smu1=max(smu1x), by(tj Xi)
drop smu1x

gsort tj -Xi
by tj: gen smu2x=sum(mu2)
egen smu2=max(smu2x), by(tj Xi)
drop smu2x

gsort tj -Xi
by tj: gen smu3x=sum(mu3)
egen smu3=max(smu3x), by(tj Xi)
drop smu3x

gsort tj -Xi
by tj: gen smu4x=sum(mu4)
egen smu4=max(smu4x), by(tj Xi)
drop smu4x

gsort tj -Xi
by tj: gen smu5x=sum(mu5)
egen smu5=max(smu5x), by(tj Xi)
drop smu5x

gsort tj -Xi
by tj: gen smu6x=sum(mu6)
egen smu6=max(smu6x), by(tj Xi)
drop smu6x
```

```
gsort tj -Xi
by tj: gen smu7x=sum(mu7)
egen smu7=max(smu7x), by(tj Xi)
drop smu7x

gsort tj -Xi
by tj: gen smu8x=sum(mu8)
egen smu8=max(smu8x), by(tj Xi)
drop smu8x

gsort tj -Xi
by tj: gen smu9x=sum(mu9)
egen smu9=max(smu9x), by(tj Xi)
drop smu9x

gsort tj -Xi
by tj: gen smu10x=sum(mu10)
egen smu10=max(smu10x), by(tj Xi)
drop smu10x

gsort tj -Xi
by tj: gen smu11x=sum(mu11)
egen smu11=max(smu11x), by(tj Xi)
drop smu11x

gsort tj -Xi
by tj: gen smu12x=sum(mu12)
egen smu12=max(smu12x), by(tj Xi)
drop smu12x

gsort tj -Xi
by tj: gen smu13x=sum(mu13)
egen smu13=max(smu13x), by(tj Xi)
drop smu13x

gsort tj -Xi
by tj: gen smu14x=sum(mu14)
egen smu14=max(smu14x), by(tj Xi)
drop smu14x

gsort tj -Xi
by tj: gen smu15x=sum(mu15)
egen smu15=max(smu15x), by(tj Xi)
drop smu15x

gsort tj -Xi
by tj: gen smu16x=sum(mu16)
egen smu16=max(smu16x), by(tj Xi)
drop smu16x

gsort tj -Xi
by tj: gen smu17x=sum(mu17)
egen smu17=max(smu17x), by(tj Xi)
drop smu17x

gsort tj -Xi
by tj: gen smu18x=sum(mu18)
egen smu18=max(smu18x), by(tj Xi)
drop smu18x

gsort tj -Xi
by tj: gen smu19x=sum(mu19)
egen smu19=max(smu19x), by(tj Xi)
```

```
drop smu19x


gen gmu1=(1/1138)*(1/surv1138)* smu1
replace gmu1=0 if gmu1==.

gen gmu2=(1/1138)*(1/surv1138)* smu2
replace gmu2=0 if gmu2==.

gen gmu3=(1/1138)*(1/surv1138)* smu3
replace gmu3=0 if gmu3==.

gen gmu4=(1/1138)*(1/surv1138)* smu4
replace gmu4=0 if gmu4==.

gen gmu5=(1/1138)*(1/surv1138)* smu5
replace gmu5=0 if gmu5==.

gen gmu6=(1/1138)*(1/surv1138)* smu6
replace gmu6=0 if gmu6==.

gen gmu7=(1/1138)*(1/surv1138)* smu7
replace gmu7=0 if gmu7==.

gen gmu8=(1/1138)*(1/surv1138)* smu8
replace gmu8=0 if gmu8==.

gen gmu9=(1/1138)*(1/surv1138)* smu9
replace gmu9=0 if gmu9==.

gen gmu10=(1/1138)*(1/surv1138)* smu10
replace gmu10=0 if gmu10==.

gen gmu11=(1/1138)*(1/surv1138)* smu11
replace gmu11=0 if gmu11==.

gen gmu12=(1/1138)*(1/surv1138)* smu12
replace gmu12=0 if gmu12==.

gen gmu13=(1/1138)*(1/surv1138)* smu13
replace gmu13=0 if gmu13==.

gen gmu14=(1/1138)*(1/surv1138)* smu14
replace gmu14=0 if gmu14==.

gen gmu15=(1/1138)*(1/surv1138)* smu15
replace gmu15=0 if gmu15==.

gen gmu16=(1/1138)*(1/surv1138)* smu16
replace gmu16=0 if gmu16==.

gen gmu17=(1/1138)*(1/surv1138)* smu17
replace gmu17=0 if gmu17==.

gen gmu18=(1/1138)*(1/surv1138)* smu18
replace gmu18=0 if gmu18==.

gen gmu19=(1/1138)*(1/surv1138)* smu19
replace gmu19=0 if gmu19==.


gen m1gm1= Mi1-gmu1
gen m2gm2= Mi2-gmu2
gen m3gm3= Mi3-gmu3
```

```
gen m4gm4= Mi4-gmu4
gen m5gm5= Mi5-gmu5
gen m6gm6= Mi6-gmu6
gen m7gm7= Mi7-gmu7
gen m8gm8= Mi8-gmu8
gen m9gm9= Mi9-gmu9
gen m10gm10= Mi10-gmu10
gen m11gm11= Mi11-gmu11
gen m12gm12= Mi12-gmu12
gen m13gm13= Mi13-gmu13
gen m14gm14= Mi14-gmu14
gen m15gm15= Mi15-gmu15
gen m16gm16= Mi16-gmu16
gen m17gm17= Mi17-gmu17
gen m18gm18= Mi18-gmu18
gen m19gm19= Mi19-gmu19

*

gen covlj1=Mil_gl*m1gm1
gen covlj2=Mil_gl*m2gm2
gen covlj3=Mil_gl*m3gm3
gen covlj4=Mil_gl*m4gm4
gen covlj5=Mil_gl*m5gm5
gen covlj6=Mil_gl*m6gm6
gen covlj7=Mil_gl*m7gm7
gen covlj8=Mil_gl*m8gm8
gen covlj9=Mil_gl*m9gm9
gen covlj10=Mil_gl*m10gm10
gen covlj11=Mil_gl*m11gm11
gen covlj12=Mil_gl*m12gm12
gen covlj13=Mil_gl*m13gm13
gen covlj14=Mil_gl*m14gm14
gen covlj15=Mil_gl*m15gm15
gen covlj16=Mil_gl*m16gm16
gen covlj17=Mil_gl*m17gm17
gen covlj18=Mil_gl*m18gm18
gen covlj19=Mil_gl*m19gm19


gen covaj1=(dijmax1/Kjmaxl1)*covlj1
replace covaj1=0 if covaj1==.

gen covaj2=(dijmax2/Kjmaxl2)*covlj2
replace covaj2=0 if covaj2==.

gen covaj3=(dijmax3/Kjmaxl3)*covlj3
replace covaj3=0 if covaj3==.

gen covaj4=(dijmax4/Kjmaxl4)*covlj4
replace covaj4=0 if covaj4==.

gen covaj5=(dijmax5/Kjmaxl5)*covlj5
replace covaj5=0 if covaj5==.

gen covaj6=(dijmax6/Kjmaxl6)*covlj6
replace covaj6=0 if covaj6==.

gen covaj7=(dijmax7/Kjmaxl7)*covlj7
replace covaj7=0 if covaj7==.

gen covaj8=(dijmax8/Kjmaxl8)*covlj8
replace covaj8=0 if covaj8==.
```

```
gen covaj9=(dijmax9/Kjmaxl9)*covlj9
replace covaj9=0 if covaj9==.

gen covaj10=(dijmax10/Kjmaxl10)*covlj10
replace covaj10=0 if covaj10==.

gen covaj11=(dijmax11/Kjmaxl11)*covlj11
replace covaj11=0 if covaj11==.

gen covaj12=(dijmax12/Kjmaxl12)*covlj12
replace covaj12=0 if covaj12==.

gen covaj13=(dijmax13/Kjmaxl13)*covlj13
replace covaj13=0 if covaj13==.

gen covaj14=(dijmax14/Kjmaxl14)*covlj14
replace covaj14=0 if covaj14==.

gen covaj15=(dijmax15/Kjmaxl15)*covlj15
replace covaj15=0 if covaj15==.

gen covaj16=(dijmax16/Kjmaxl16)*covlj16
replace covaj16=0 if covaj16==.

gen covaj17=(dijmax17/Kjmaxl17)*covlj17
replace covaj17=0 if covaj17==.

gen covaj18=(dijmax18/Kjmaxl18)*covlj18
replace covaj18=0 if covaj18==.

gen covaj19=(dijmax19/Kjmaxl19)*covlj19
replace covaj19=0 if covaj19==.


**ties**

gen int const=1
sort tj Xi
by tj Xi: gen ties=sum(const)
egen maxties=max(ties), by(tj Xi)

*

gsort tj -Xi ties
by tj: gen scvaj1x=sum(covaj1)
replace scvaj1x=0 if Xij<Xi
gen scvaj1=scvaj1x if ties==maxties
egen scvaj1xx=min(scvaj1) if Xij==Xi, by(tj Xi maxties)
replace scvaj1=scvaj1xx if scvaj1==.
replace scvaj1=0 if Xij<Xi
*drop scvaj1x scvaj1xx*

gsort tj -Xi ties
by tj: gen scvaj2x=sum(covaj2)
replace scvaj2x=0 if Xij<Xi
gen scvaj2=scvaj2x if ties==maxties
egen scvaj2xx=min(scvaj2) if Xij==Xi, by(tj Xi maxties)
replace scvaj2=scvaj2xx if scvaj2==.
replace scvaj2=0 if Xij<Xi
*drop scvaj2x scvaj2xx*

gsort tj -Xi ties
by tj: gen scvaj3x=sum(covaj3)
replace scvaj3x=0 if Xij<Xi
```

```
gen scvaj3=scvaj3x if ties==maxties
egen scvaj3xx=min(scvaj3) if Xij==Xi, by(tj Xi maxties)
replace scvaj3=scvaj3xx if scvaj3==.
replace scvaj3=0 if Xij<Xi
*drop scvaj3x scvaj3xx*

gsort tj -Xi ties
by tj: gen scvaj4x=sum(covaj4)
replace scvaj4x=0 if Xij<Xi
gen scvaj4=scvaj4x if ties==maxties
egen scvaj4xx=min(scvaj4) if Xij==Xi, by(tj Xi maxties)
replace scvaj4=scvaj4xx if scvaj4==.
replace scvaj4=0 if Xij<Xi
*drop scvaj4x scvaj4xx*

gsort tj -Xi ties
by tj: gen scvaj5x=sum(covaj5)
replace scvaj5x=0 if Xij<Xi
gen scvaj5=scvaj5x if ties==maxties
egen scvaj5xx=min(scvaj5) if Xij==Xi, by(tj Xi maxties)
replace scvaj5=scvaj5xx if scvaj5==.
replace scvaj5=0 if Xij<Xi
*drop scvaj5x scvaj5xx*

gsort tj -Xi ties
by tj: gen scvaj6x=sum(covaj6)
replace scvaj6x=0 if Xij<Xi
gen scvaj6=scvaj6x if ties==maxties
egen scvaj6xx=min(scvaj6) if Xij==Xi, by(tj Xi maxties)
replace scvaj6=scvaj6xx if scvaj6==.
replace scvaj6=0 if Xij<Xi
*drop scvaj6x scvaj6xx*

gsort tj -Xi ties
by tj: gen scvaj7x=sum(covaj7)
replace scvaj7x=0 if Xij<Xi
gen scvaj7=scvaj7x if ties==maxties
egen scvaj7xx=min(scvaj7) if Xij==Xi, by(tj Xi maxties)
replace scvaj7=scvaj7xx if scvaj7==.
replace scvaj7=0 if Xij<Xi
*drop scvaj7x scvaj7xx*

gsort tj -Xi ties
by tj: gen scvaj8x=sum(covaj8)
replace scvaj8x=0 if Xij<Xi
gen scvaj8=scvaj8x if ties==maxties
egen scvaj8xx=min(scvaj8) if Xij==Xi, by(tj Xi maxties)
replace scvaj8=scvaj8xx if scvaj8==.
replace scvaj8=0 if Xij<Xi
*drop scvaj8x scvaj8xx*

gsort tj -Xi ties
by tj: gen scvaj9x=sum(covaj9)
replace scvaj9x=0 if Xij<Xi
gen scvaj9=scvaj9x if ties==maxties
egen scvaj9xx=min(scvaj9) if Xij==Xi, by(tj Xi maxties)
replace scvaj9=scvaj9xx if scvaj9==.
replace scvaj9=0 if Xij<Xi
*drop scvaj9x scvaj9xx*

gsort tj -Xi ties
by tj: gen scvaj10x=sum(covaj10)
replace scvaj10x=0 if Xij<Xi
gen scvaj10=scvaj10x if ties==maxties
```

```
egen scvaj10xx=min(scvaj10) if Xij==Xi, by(tj Xi maxties)
replace scvaj10=scvaj10xx if scvaj10==.
replace scvaj10=0 if Xij<Xi
*drop scvaj10x scvaj10xx*

gsort tj -Xi ties
by tj: gen scvaj11x=sum(covaj11)
replace scvaj11x=0 if Xij<Xi
gen scvaj11=scvaj11x if ties==maxties
egen scvaj11xx=min(scvaj11) if Xij==Xi, by(tj Xi maxties)
replace scvaj11=scvaj11xx if scvaj11==.
replace scvaj11=0 if Xij<Xi
*drop scvaj11x scvaj11xx*

gsort tj -Xi ties
by tj: gen scvaj12x=sum(covaj12)
replace scvaj12x=0 if Xij<Xi
gen scvaj12=scvaj12x if ties==maxties
egen scvaj12xx=min(scvaj12) if Xij==Xi, by(tj Xi maxties)
replace scvaj12=scvaj12xx if scvaj12==.
replace scvaj12=0 if Xij<Xi
*drop scvaj12x scvaj12xx*

gsort tj -Xi ties
by tj: gen scvaj13x=sum(covaj13)
replace scvaj13x=0 if Xij<Xi
gen scvaj13=scvaj13x if ties==maxties
egen scvaj13xx=min(scvaj1) if Xij==Xi, by(tj Xi maxties)
replace scvaj13=scvaj13xx if scvaj13==.
replace scvaj13=0 if Xij<Xi
*drop scvaj13x scvaj13xx*

gsort tj -Xi ties
by tj: gen scvaj14x=sum(covaj14)
replace scvaj14x=0 if Xij<Xi
gen scvaj14=scvaj14x if ties==maxties
egen scvaj14xx=min(scvaj14) if Xij==Xi, by(tj Xi maxties)
replace scvaj14=scvaj14xx if scvaj14==.
replace scvaj14=0 if Xij<Xi
*drop scvaj14x scvaj14xx*

gsort tj -Xi ties
by tj: gen scvaj15x=sum(covaj15)
replace scvaj15x=0 if Xij<Xi
gen scvaj15=scvaj15x if ties==maxties
egen scvaj15xx=min(scvaj15) if Xij==Xi, by(tj Xi maxties)
replace scvaj15=scvaj15xx if scvaj15==.
replace scvaj15=0 if Xij<Xi
*drop scvaj15x scvaj15xx*

gsort tj -Xi ties
by tj: gen scvaj16x=sum(covaj16)
replace scvaj16x=0 if Xij<Xi
gen scvaj16=scvaj16x if ties==maxties
egen scvaj16xx=min(scvaj16) if Xij==Xi, by(tj Xi maxties)
replace scvaj16=scvaj16xx if scvaj16==.
replace scvaj16=0 if Xij<Xi
*drop scvaj16x scvaj16xx*

gsort tj -Xi ties
by tj: gen scvaj17x=sum(covaj17)
replace scvaj17x=0 if Xij<Xi
gen scvaj17=scvaj17x if ties==maxties
egen scvaj17xx=min(scvaj17) if Xij==Xi, by(tj Xi maxties)
```

```
replace scvaj17=scvaj17xx if scvaj17==.
replace scvaj17=0 if Xij<Xi
*drop scvaj17x scvaj17xx*

gsort tj -Xi ties
by tj: gen scvaj18x=sum(covaj18)
replace scvaj18x=0 if Xij<Xi
gen scvaj18=scvaj18x if ties==maxties
egen scvaj18xx=min(scvaj18) if Xij==Xi, by(tj Xi maxties)
replace scvaj18=scvaj18xx if scvaj18==.
replace scvaj18=0 if Xij<Xi
*drop scvaj18x scvaj18xx*

gsort tj -Xi ties
by tj: gen scvaj19x=sum(covaj19)
replace scvaj19x=0 if Xij<Xi
gen scvaj19=scvaj19x if ties==maxties
egen scvaj19xx=min(scvaj19) if Xij==Xi, by(tj Xi maxties)
replace scvaj19=scvaj19xx if scvaj19==.
replace scvaj19=0 if Xij<Xi
*drop scvaj19x scvaj19xx*

*
gen covb1=(1/1138)*(1/Slu)*scvaj1
replace covb1=0 if covb1==.

gen covb2=(1/1138)*(1/Slu)*scvaj2
replace covb2=0 if covb2==.

gen covb3=(1/1138)*(1/Slu)*scvaj3
replace covb3=0 if covb3==.

gen covb4=(1/1138)*(1/Slu)*scvaj4
replace covb4=0 if covb4==.

gen covb5=(1/1138)*(1/Slu)*scvaj5
replace covb5=0 if covb5==.

gen covb6=(1/1138)*(1/Slu)*scvaj6
replace covb6=0 if covb6==.

gen covb7=(1/1138)*(1/Slu)*scvaj7
replace covb7=0 if covb7==.

gen covb8=(1/1138)*(1/Slu)*scvaj8
replace covb8=0 if covb8==.

gen covb9=(1/1138)*(1/Slu)*scvaj9
replace covb9=0 if covb9==.

gen covb10=(1/1138)*(1/Slu)*scvaj10
replace covb10=0 if covb10==.

gen covb11=(1/1138)*(1/Slu)*scvaj11
replace covb11=0 if covb11==.

gen covb12=(1/1138)*(1/Slu)*scvaj12
replace covb12=0 if covb12==.

gen covb13=(1/1138)*(1/Slu)*scvaj13
replace covb13=0 if covb13==.

gen covb14=(1/1138)*(1/Slu)*scvaj14
replace covb14=0 if covb14==.
```

```
gen covb15=(1/1138)*(1/Slu)*scvaj15
replace covb15=0 if covb15==.

gen covb16=(1/1138)*(1/Slu)*scvaj16
replace covb16=0 if covb16==.

gen covb17=(1/1138)*(1/Slu)*scvaj17
replace covb17=0 if covb17==.

gen covb18=(1/1138)*(1/Slu)*scvaj18
replace covb18=0 if covb18==.

gen covb19=(1/1138)*(1/Slu)*scvaj19
replace covb19=0 if covb19==.

*
egen covc1=sum(covb1), by(ukno)
egen covc2=sum(covb2), by(ukno)
egen covc3=sum(covb3), by(ukno)
egen covc4=sum(covb4), by(ukno)
egen covc5=sum(covb5), by(ukno)
egen covc6=sum(covb6), by(ukno)
egen covc7=sum(covb7), by(ukno)
egen covc8=sum(covb8), by(ukno)
egen covc9=sum(covb9), by(ukno)
egen covc10=sum(covb10), by(ukno)
egen covc11=sum(covb11), by(ukno)
egen covc12=sum(covb12), by(ukno)
egen covc13=sum(covb13), by(ukno)
egen covc14=sum(covb14), by(ukno)
egen covc15=sum(covb15), by(ukno)
egen covc16=sum(covb16), by(ukno)
egen covc17=sum(covb17), by(ukno)
egen covc18=sum(covb18), by(ukno)
egen covc19=sum(covb19), by(ukno)

*


gen covd1=(censor1/(cens1138^2))*covc1
replace covd1=0 if covd1==.

gen covd2=(censor1/(cens1138^2))*covc2
replace covd2=0 if covd2==.

gen covd3=(censor1/(cens1138^2))*covc3
replace covd3=0 if covd3==.

gen covd4=(censor1/(cens1138^2))*covc4
replace covd4=0 if covd4==.

gen covd5=(censor1/(cens1138^2))*covc5
replace covd5=0 if covd5==.

gen covd6=(censor1/(cens1138^2))*covc6
replace covd6=0 if covd6==.

gen covd7=(censor1/(cens1138^2))*covc7
replace covd7=0 if covd7==.

gen covd8=(censor1/(cens1138^2))*covc8
replace covd8=0 if covd8==.
```

```
gen covd9=(censor1/(cens1138^2))*covc9
replace covd9=0 if covd9==.

gen covd10=(censor1/(cens1138^2))*covc10
replace covd10=0 if covd10==.

gen covd11=(censor1/(cens1138^2))*covc11
replace covd11=0 if covd11==.

gen covd12=(censor1/(cens1138^2))*covc12
replace covd12=0 if covd12==.

gen covd13=(censor1/(cens1138^2))*covc13
replace covd13=0 if covd13==.

gen covd14=(censor1/(cens1138^2))*covc14
replace covd14=0 if covd14==.

gen covd15=(censor1/(cens1138^2))*covc15
replace covd15=0 if covd15==.

gen covd16=(censor1/(cens1138^2))*covc16
replace covd16=0 if covd16==.

gen covd17=(censor1/(cens1138^2))*covc17
replace covd17=0 if covd17==.

gen covd18=(censor1/(cens1138^2))*covc18
replace covd18=0 if covd18==.

gen covd19=(censor1/(cens1138^2))*covc19
replace covd19=0 if covd19==.

*
egen cove1=sum(covd1), by(tj)
egen cove2=sum(covd2), by(tj)
egen cove3=sum(covd3), by(tj)
egen cove4=sum(covd4), by(tj)
egen cove5=sum(covd5), by(tj)
egen cove6=sum(covd6), by(tj)
egen cove7=sum(covd7), by(tj)
egen cove8=sum(covd8), by(tj)
egen cove9=sum(covd9), by(tj)
egen cove10=sum(covd10), by(tj)
egen cove11=sum(covd11), by(tj)
egen cove12=sum(covd12), by(tj)
egen cove13=sum(covd13), by(tj)
egen cove14=sum(covd14), by(tj)
egen cove15=sum(covd15), by(tj)
egen cove16=sum(covd16), by(tj)
egen cove17=sum(covd17), by(tj)
egen cove18=sum(covd18), by(tj)
egen cove19=sum(covd19), by(tj)

gen covj1=(1/1138)*cove1
gen covj2=(1/1138)*cove2
gen covj3=(1/1138)*cove3
gen covj4=(1/1138)*cove4
gen covj5=(1/1138)*cove5
gen covj6=(1/1138)*cove6
gen covj7=(1/1138)*cove7
gen covj8=(1/1138)*cove8
gen covj9=(1/1138)*cove9
```

```
gen covj10=(1/1138)*cove10
gen covj11=(1/1138)*cove11
gen covj12=(1/1138)*cove12
gen covj13=(1/1138)*cove13
gen covj14=(1/1138)*cove14
gen covj15=(1/1138)*cove15
gen covj16=(1/1138)*cove16
gen covj17=(1/1138)*cove17
gen covj18=(1/1138)*cove18
gen covj19=(1/1138)*cove19


** Matrix calculations for Improved partitioned for conventional**

keep if ukno=="00011D"
keep if tj==1
mkmat covj*, matrix(covp)
svmat covp
matrix Gp= covp*invvar
svmat Gp
matrix wp=Gp*A
svmat wp
gen wpterm= wp1/1138

matrix varpterm=Gp*covp'
svmat varpterm


**

gen intern1pimp= di*((Mi- mpartimp)^2)/ cens1138
replace intern1pimp=0 if intern1pimp==.

egen sumint1pimp=sum(intern1pimp)
gen term1pimp=(1/1138)*sumint1pimp
gen vartermp3=1540000000000
gen varpartimp=(1/1138)*( term1pimp+ term2)-(1/1138)* vartermp3

gen separtimp=sqrt( varpartimp)

**
```

Appendix A.4.7. Generation of the artificial dataset (as described in section 4.4.4.3)

```
For 25% censoring (Similarly for all other levels of censoring)

**Generating the artificial dataset for 25% censoring**

drop _all

set seed 1001

set obs 1138

gen const=1
gen ukno=sum(const)
drop const

sort ukno

gen timallde=10*uniform()+0

gen Ci=20*uniform()+0

gen censorig=1 if  timallde<=Ci
replace censorig=0 if censorig==.

gen Mi0=10000*uniform()+5000

gen bi=1600*uniform()+1000

gen ti1=400*uniform()
gen ti2=400*uniform()
gen ti3=400*uniform()
gen ti4=400*uniform()
gen ti5=400*uniform()
gen ti6=400*uniform()
gen ti7=400*uniform()
gen ti8=400*uniform()
gen ti9=400*uniform()
gen ti10=400*uniform()

gen deathci=20000*uniform()+10000

sort ukno

gen censor1=1 if censorig==0
replace censor1=0 if censorig==1

stset  timallde, failure(censorig==1)

sts gen surv1138=s

stset  timallde, failure(censor1==1)

sts gen cens1138=s

drop _t0 _t _d _st

gen int const=1
gsort  - timallde
gen Yux=sum(const)
egen Yu=max(Yux), by(timallde)
drop const Yux

sort ukno
```

```
save "C:\WINDOWS\Desktop\artificialdata\n1138_cens6\n1138cens6.dta", replace

use "C:\WINDOWS\Desktop\artificialdata\n1138_cens1\n1138_yrs10.dta", clear

keep ukno year

sort ukno year

merge ukno using "C:\WINDOWS\Desktop\artificialdata\n1138_cens6\n1138cens6.dta"

drop _merge

gen cyr=bi+ti1 if year==1
replace cyr=bi+ti2 if year==2
replace cyr=bi+ti3 if year==3
replace cyr=bi+ti4 if year==4
replace cyr=bi+ti5 if year==5
replace cyr=bi+ti6 if year==6
replace cyr=bi+ti7 if year==7
replace cyr=bi+ti8 if year==8
replace cyr=bi+ti9 if year==9
replace cyr=bi+ti10 if year==10

gen x=1 if  timallde>=year-1

gen truecyr=cyr if x==1
replace truecyr=0 if x==.

egen struecyr=sum(truecyr), by(ukno)

gen Mitrue=Mi0+ struecyr+deathci

gen Mi=Mi0+ struecyr+censorig*deathci

label var Mitrue "NO censoring"

egen maxyear=max(year) if x~=., by(ukno)

gen costyr=truecyr+Mi0 if year==1
replace costyr=truecyr+censorig*deathci if year==maxyear & costyr==.
replace costyr=truecyr+censorig*deathci+Mi0  if maxyear==1
replace costyr=truecyr  if costyr==. & x~=.

sort ukno year

collapse Mitrue Mi, by(ukno)

label var Mitrue "NO censoring"

sort ukno

merge ukno using "C:\WINDOWS\Desktop\artificialdata\n1138_cens6\n1138cens6.dta"

drop _merge

save "C:\WINDOWS\Desktop\artificialdata\n1138_cens6\n1138cens6.dta", replace

*
```

Appendix A.4.8. Programs for estimating the standard errors of Lin1, Lin2 and Bang and Tsiatis simple weighted and partitioned estimators using the bootstrap

```
**For Lin 1: Cost histories recorded**

use "C:\WINDOWS\Desktop\lin1convyrs19test.dta", clear
rename ukno ukno0
do "C:\WINDOWS\Desktop\bsLin1conv.txt"

where "C:\WINDOWS\Desktop\bsLin1conv.txt" is:

**Lin1: Cost histories recorded**

**CONVENTIONAL** Mean **

program define lin1sim

    if "`1'"=="?" {
           global S_1 "mean"
           exit
    }

egen Xi=min(timallde), by(ukno)
egen di=min(censorig), by(ukno)

stset Xi if year==1, failure(di==1)

sts gen surv1138=s

gen int inttime=int(Xi)+1 if year==1
egen mintime=min(Xi), by(inttime)
gen st=surv1138 if mintime==Xi
egen Skx=min(st), by(inttime)

gen ak=year-1
gen ak1=year

gen s1x=Skx if inttime==1
gen s2x=Skx if inttime==2
gen s3x=Skx if inttime==3
gen s4x=Skx if inttime==4
gen s5x=Skx if inttime==5
gen s6x=Skx if inttime==6
gen s7x=Skx if inttime==7
gen s8x=Skx if inttime==8
gen s9x=Skx if inttime==9
gen s10x=Skx if inttime==10
gen s11x=Skx if inttime==11
gen s12x=Skx if inttime==12
gen s13x=Skx if inttime==13
gen s14x=Skx if inttime==14
gen s15x=Skx if inttime==15
gen s16x=Skx if inttime==16
gen s17x=Skx if inttime==17
gen s18x=Skx if inttime==18
gen s19x=Skx if inttime==19

egen s1=min(s1x)
egen s2=min(s2x)
egen s3=min(s3x)
egen s4=min(s4x)
egen s5=min(s5x)
```

```
egen s6=min(s6x)
egen s7=min(s7x)
egen s8=min(s8x)
egen s9=min(s9x)
egen s10=min(s10x)
egen s11=min(s11x)
egen s12=min(s12x)
egen s13=min(s13x)
egen s14=min(s14x)
egen s15=min(s15x)
egen s16=min(s16x)
egen s17=min(s17x)
egen s18=min(s18x)
egen s19=min(s19x)

gen Sk=s1 if ak==0
replace Sk=s2 if ak==1
replace Sk=s3 if ak==2
replace Sk=s4 if ak==3
replace Sk=s5 if ak==4
replace Sk=s6 if ak==5
replace Sk=s7 if ak==6
replace Sk=s8 if ak==7
replace Sk=s9 if ak==8
replace Sk=s10 if ak==9
replace Sk=s11 if ak==10
replace Sk=s12 if ak==11
replace Sk=s13 if ak==12
replace Sk=s14 if ak==13
replace Sk=s15 if ak==14
replace Sk=s16 if ak==15
replace Sk=s17 if ak==16
replace Sk=s18 if ak==17
replace Sk=s19 if ak==18

gen Yki=1 if Xi>=ak
replace Yki=0 if Yki==.

gen Cki=costyr
replace Cki=0 if Cki==.

egen sYki=sum(Yki), by(ak)

gen YkiCki= Yki* Cki
egen sYkiCki=sum( YkiCki), by(ak)

gen Ek= sYkiCki/ sYki

gen SkEk= Sk* Ek
egen sSkEk=sum( SkEk), by(ukno)

rename sSkEk meanlin1

sum meanlin1

post `1' (r(mean))

end

*
set seed 1001
bstrap lin1sim, reps(1000) dots cluster(ukno0) idcluster(ukno)
saving(C:\Windows\desktop\bslin1conv1000.dta)
**
```

For Lin 2: Cost histories not recorded

```
use "C:\WINDOWS\Desktop\lin2convyrs19.dta", clear
rename ukno ukno0
do "C:\WINDOWS\Desktop\bsLin2conv.txt"

where "C:\WINDOWS\Desktop\bsLin2conv.txt" is:

** Lin 2: Cost histories NOT recorded **

** CONVENTIONAL ** Mean**

program define lin2sim

    if "`1'"=="?" {
           global S_1 "mean"
           exit
    }

egen Xi=min(timallde), by(ukno)
egen di=min(censorig), by(ukno)

stset Xi if year==1, failure(di==1)

sts gen surv1138=s

gen int inttime=int(Xi)+1 if year==1
egen mintime=min(Xi), by(inttime)
gen st=surv1138 if mintime==Xi
egen Skx=min(st), by(inttime)

gen ak=year-1
gen ak1=year

gen s1x=Skx if inttime==1
gen s2x=Skx if inttime==2
gen s3x=Skx if inttime==3
gen s4x=Skx if inttime==4
gen s5x=Skx if inttime==5
gen s6x=Skx if inttime==6
gen s7x=Skx if inttime==7
gen s8x=Skx if inttime==8
gen s9x=Skx if inttime==9
gen s10x=Skx if inttime==10
gen s11x=Skx if inttime==11
gen s12x=Skx if inttime==12
gen s13x=Skx if inttime==13
gen s14x=Skx if inttime==14
gen s15x=Skx if inttime==15
gen s16x=Skx if inttime==16
gen s17x=Skx if inttime==17
gen s18x=Skx if inttime==18
gen s19x=Skx if inttime==19

egen s1=min(s1x)
egen s2=min(s2x)
egen s3=min(s3x)
egen s4=min(s4x)
egen s5=min(s5x)
egen s6=min(s6x)
egen s7=min(s7x)
egen s8=min(s8x)
egen s9=min(s9x)
egen s10=min(s10x)
```

```
egen s11=min(s11x)
egen s12=min(s12x)
egen s13=min(s13x)
egen s14=min(s14x)
egen s15=min(s15x)
egen s16=min(s16x)
egen s17=min(s17x)
egen s18=min(s18x)
egen s19=min(s19x)

gen Sk=s1 if ak==0
replace Sk=s2 if ak==1
replace Sk=s3 if ak==2
replace Sk=s4 if ak==3
replace Sk=s5 if ak==4
replace Sk=s6 if ak==5
replace Sk=s7 if ak==6
replace Sk=s8 if ak==7
replace Sk=s9 if ak==8
replace Sk=s10 if ak==9
replace Sk=s11 if ak==10
replace Sk=s12 if ak==11
replace Sk=s13 if ak==12
replace Sk=s14 if ak==13
replace Sk=s15 if ak==14
replace Sk=s16 if ak==15
replace Sk=s17 if ak==16
replace Sk=s18 if ak==17
replace Sk=s19 if ak==18 | ak==19

egen Ci=min(Mi), by(ukno)

sort ukno ak
quietly by ukno: gen Sk1=Sk[_n+1]

replace Sk1=0 if ak==19
egen tmax=max(Xi)

gen Yki=1 if ((ak<=Xi & Xi<ak1) & di==1) | (Xi>=tmax & ak==19)
replace Yki=0 if Yki==.

egen sYki=sum(Yki), by(ak)

gen YkiCi= Yki* Ci
egen sYkiCi=sum(YkiCi), by(ak)

gen Ak= sYkiCi/ sYki
replace Ak=0 if Ak==.

gen AkS= Ak*(Sk-Sk1)
egen sAkS=sum(AkS), by(ukno)
rename sAkS meanlin2

sum meanlin2

post `1' (r(mean))

end

*
set seed 1001
bstrap lin2sim, reps(1000) dots cluster(ukno0) idcluster(ukno)
saving(C:\Windows\desktop\bslin2conv1000.dta)
**
```

For Bang and Tsiatis simple weighted estimator

```
use "D:\Chapter1_COST\Original_ukpds_data\KM_conv1138.dta", clear
do "D:\Chapter1_COST\tsiatis_new\Partitioned_simul\sesimplesim_conv.txt"

where sesimplesim_conv.txt is:

** Simple mean: conventional **

program define sesimpsim

      if "`1'"=="?" {
              global S_1 "meansimp"
              exit
      }

stset timallde, failure(censorig==1)

sts gen surv1138=s

gen censor1=1 if censorig==0
replace censor1=0 if censor1==.

label var censor1 "1:censored; 0:dead"

stset  timallde, failure(censor1==1)

sts gen cens1138=s

gen diMi_KTi= (censorig* Mi)/ cens1138
replace  diMi_KTi=0 if diMi_KTi==.

egen sumalli=sum(diMi_KTi)

gen meansimp=(1/1138)* sumalli

sum meansimp, meanonly

post `1' (r(mean))

end

*

set seed 1001

bstrap sesimpsim, reps(1000) dots saving(C:\windows\desktop\conv_sesimp1000.dta)

*
```

For Bang and Tsiatis partitioned estimator

```
use "C:\……..\conv_partyrs.dta", clear
rename ukno ukno0
do "C:\…….\separtsim_conv.txt"

where separtsim_conv.txt is:

** Partitioned mean:conventional **

program define separtsim

      if "`1'"=="?" {
              global S_1 "meanpart"
              exit
      }

gen tj_1=year-1
gen tj=year

gen Mij=costyr
replace Mij=0 if Mij==.

egen Xi=min(timallde), by(ukno)
egen di=min(censorig), by(ukno)

gen minTitj=min(Xi, tj)
gen Xij=min(minTitj, Xi)

gen dij=1 if (minTitj==tj | (minTitj==Xi & di==1))
replace dij=0 if dij==.

stset Xij, failure(dij==0)

sts gen KjTij=s, by(tj)

gen intern=(dij*Mij)/KjTij
replace intern=0 if intern==.

egen sumintjs=sum(intern), by(ukno)

collapse sumintjs, by(ukno)

egen sumintjn=sum(sumintjs)
gen meanpart=(1/1138)*sumintjn

sum meanpart, meanonly

post `1' (r(mean))

end

*

set seed 1001
bstrap separtsim, reps(1000) dots cluster(ukno0) idcluster(ukno)
saving(C:\…….\conv_separt1000.dta)

**
```

Appendix A.5.1. Assessing the proportionality assumption in the stratified Cox model

The test for assessing the assumption of proportional hazards proposed by Grambsch and Therneau (1994) results in a χ^2 distribution for the statistic and whenever Prob<0.05 the null hypothesis of proportional hazards is rejected.

CONVENTIONAL

year==1

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.00730	0.58	1	0.4455
fpg	-0.04028	18.69	1	0.0000
bmi	0.03453	13.77	1	0.0002
race	-0.00716	0.54	1	0.4604
sex	0.01358	2.02	1	0.1556
global test		0.41	5	0.9949

year==2

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01343	1.98	1	0.1595
fpg	-0.01023	1.18	1	0.2765
bmi	0.02807	9.16	1	0.0025
race	-0.03575	13.71	1	0.0002
sex	0.00576	0.36	1	0.5464
global test		0.32	5	0.9972

year==3

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01347	1.99	1	0.1584
fpg	0.03543	14.30	1	0.0002
bmi	0.00308	0.11	1	0.7398
race	0.00315	0.11	1	0.7422
sex	-0.00634	0.44	1	0.5074
global test		0.18	5	0.9993

year==4

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01939	4.13	1	0.0422
fpg	-0.00331	0.12	1	0.7241
bmi	0.00173	0.03	1	0.8525
race	0.00205	0.05	1	0.8298
sex	0.01866	3.80	1	0.0512
global test		0.09	5	0.9999

year==5

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.02979	9.67	1	0.0019
fpg	-0.01945	4.33	1	0.0375
bmi	0.01885	4.14	1	0.0418
race	0.01245	1.66	1	0.1977
sex	0.02159	5.10	1	0.0239
global test		0.27	5	0.9982

year==6

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.00316	0.11	1	0.7418
fpg	-0.02637	7.86	1	0.0051
bmi	0.01008	1.16	1	0.2807
race	-0.02485	6.67	1	0.0098
sex	0.03614	14.22	1	0.0002
global test		0.29	5	0.9978

year==7

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01230	1.61	1	0.2050
fpg	0.02570	7.17	1	0.0074
bmi	-0.00058	0.00	1	0.9504
race	0.01507	2.43	1	0.1193
sex	0.02397	6.12	1	0.0133
global test		0.15	5	0.9995

year==8

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.02564	6.86	1	0.0088
fpg	0.01777	3.33	1	0.0681
bmi	-0.01460	2.33	1	0.1266
race	0.02067	4.55	1	0.0328
sex	-0.00348	0.13	1	0.7177
global test		0.12	5	0.9998

year==9

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01398	2.08	1	0.1489
fpg	0.03390	11.49	1	0.0007
bmi	0.00712	0.55	1	0.4566
race	0.01743	3.41	1	0.0648
sex	0.02573	7.14	1	0.0075
global test		0.13	5	0.9997

year==10

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.00196	0.04	1	0.8411
fpg	0.06275	37.97	1	0.0000
bmi	-0.00398	0.17	1	0.6820
race	0.00419	0.20	1	0.6536
sex	-0.03594	13.80	1	0.0002
global test		0.13	5	0.9997

year==11

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01868	3.69	1	0.0549
fpg	-0.02439	5.67	1	0.0172
bmi	-0.04956	24.92	1	0.0000
race	0.01462	2.55	1	0.1105
sex	0.02038	4.41	1	0.0358
global test		0.06	5	0.9999

year==12

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.05852	37.55	1	0.0000
fpg	-0.00362	0.13	1	0.7192
bmi	0.00480	0.23	1	0.6311
race	-0.00193	0.05	1	0.8264
sex	0.06320	43.20	1	0.0000
global test		0.08	5	0.9999

year==13

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.04364	20.01	1	0.0000
fpg	-0.01616	2.53	1	0.1118
bmi	-0.01559	2.67	1	0.1024
race	0.02144	6.55	1	0.0105
sex	0.04384	19.91	1	0.0000
global test		0.02	5	1.0000

year==14

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.03388	12.70	1	0.0004
fpg	-0.02012	3.62	1	0.0570
bmi	-0.05235	27.10	1	0.0000
race	0.03085	14.59	1	0.0001
sex	-0.02536	6.82	1	0.0090
global test		0.01	5	1.0000

year==15

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.08021	67.07	1	0.0000
fpg	-0.02630	5.37	1	0.0205
bmi	0.09872	94.27	1	0.0000
race	-0.06248	35.79	1	0.0000
sex	-0.04451	19.97	1	0.0000
global test		0.01	5	1.0000

year==16

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.01978	3.93	1	0.0473
fpg	0.01225	1.02	1	0.3132
bmi	-0.00341	0.10	1	0.7473
race	0.01193	1.25	1	0.2642
sex	-0.01670	2.74	1	0.0977
global test		0.00	5	1.0000

year==17

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.37451	1249.91	1	0.0000
fpg	0.26504	325.63	1	0.0000
bmi	0.08720	75.29	1	0.0000
race	0.29140	424.58	1	0.0000
sex	-0.07251	52.53	1	0.0000
global test		0.02	5	1.0000

year==18

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.07142	37.37	1	0.0000
fpg	0.15217	49.09	1	0.0000
bmi	0.06861	23.44	1	0.0000
race	0.22924	43.23	1	0.0000
sex	-0.04373	15.29	1	0.0001
global test		0.00	5	1.0000

INTENSIVE

year==1 Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.02339	14.94	1	0.0001
fpg	0.02470	17.02	1	0.0000
bmi	-0.00468	0.62	1	0.4326
race	-0.00544	0.82	1	0.3640
sex	0.01778	8.33	1	0.0039
global test		0.46	5	0.9934

year==2

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.03174	27.49	1	0.0000
fpg	0.03792	40.31	1	0.0000
bmi	0.02688	20.26	1	0.0000
race	-0.01710	8.18	1	0.0042
sex	0.00133	0.05	1	0.8294
global test		1.11	5	0.9536

year==3

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.00214	0.12	1	0.7244
fpg	0.04231	49.36	1	0.0000
bmi	0.00963	2.60	1	0.1070
race	0.00951	2.49	1	0.1149
sex	0.01871	9.26	1	0.0023
global test		0.74	5	0.9808

year==4

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01373	5.13	1	0.0236
fpg	0.02216	13.61	1	0.0002
bmi	-0.00807	1.79	1	0.1813
race	-0.00219	0.13	1	0.7151
sex	-0.00605	0.97	1	0.3250
global test		0.22	5	0.9989

year==5

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01167	3.71	1	0.0541
fpg	0.04401	53.40	1	0.0000
bmi	0.01653	7.46	1	0.0063
race	0.00777	1.70	1	0.1927
sex	-0.01166	3.60	1	0.0577
global test		0.67	5	0.9843

year==6

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.00127	0.04	1	0.8341
fpg	0.04362	53.22	1	0.0000
bmi	0.01400	5.36	1	0.0206
race	0.00728	1.42	1	0.2332
sex	0.01966	10.23	1	0.0014
global test		0.72	5	0.9821

year==7

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.01623	6.98	1	0.0083
fpg	0.03346	30.85	1	0.0000
bmi	0.01618	7.11	1	0.0077
race	-0.04225	46.28	1	0.0000
sex	0.03045	24.56	1	0.0000
global test		0.99	5	0.9636

year==8

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.00899	2.11	1	0.1461
fpg	0.03152	27.18	1	0.0000
bmi	0.01654	7.01	1	0.0081
race	0.01202	3.59	1	0.0583
sex	0.01533	6.23	1	0.0126
global test		0.30	5	0.9977

year==9

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.00443	0.51	1	0.4731
fpg	0.03666	37.57	1	0.0000
bmi	0.02703	18.12	1	0.0000
race	0.00584	0.82	1	0.3665
sex	-0.02640	18.56	1	0.0000
global test		0.28	5	0.9979

year==10

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.00324	0.27	1	0.6019
fpg	0.01661	7.01	1	0.0081
bmi	0.01453	5.24	1	0.0221
race	0.00151	0.06	1	0.8145
sex	0.01302	4.49	1	0.0342
global test		0.06	5	1.0000

year==11

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.02087	11.31	1	0.0008
fpg	0.03978	39.39	1	0.0000
bmi	0.00763	1.49	1	0.2224
race	0.00880	1.84	1	0.1754
sex	0.00057	0.01	1	0.9263
global test		0.11	5	0.9998

year==12

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.00987	2.43	1	0.1187
fpg	0.01510	5.52	1	0.0188
bmi	-0.03257	26.59	1	0.0000
race	-0.01875	8.02	1	0.0046
sex	0.00951	2.38	1	0.1225
global test		0.04	5	1.0000

year==13

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.03706	32.90	1	0.0000
fpg	0.06124	91.79	1	0.0000
bmi	-0.01886	9.03	1	0.0027
race	-0.01051	2.71	1	0.0998
sex	-0.02676	18.95	1	0.0000
global test		0.08	5	0.9999

year==14

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.00878	1.83	1	0.1766
fpg	-0.01685	6.58	1	0.0103
bmi	0.00248	0.16	1	0.6861
race	-0.02739	18.38	1	0.0000
sex	0.04579	54.17	1	0.0000
global test		0.01	5	1.0000

year==15

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	0.05050	61.02	1	0.0000
fpg	0.04553	47.49	1	0.0000
bmi	-0.03090	22.07	1	0.0000
race	-0.10985	334.91	1	0.0000
sex	0.09169	214.26	1	0.0000
global test		0.05	5	1.0000

year==16

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.05392	63.18	1	0.0000
fpg	-0.05905	92.40	1	0.0000
bmi	-0.11303	172.69	1	0.0000
race	0.05597	79.46	1	0.0000
sex	0.06123	95.89	1	0.0000
global test		0.01	5	1.0000

year==17

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.07907	133.29	1	0.0000
fpg	-0.01417	6.06	1	0.0138
bmi	-0.00769	0.74	1	0.3885
race	-0.04379	42.12	1	0.0000
sex	0.08005	169.73	1	0.0000
global test		0.00	5	1.0000

year==18

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.07637	94.12	1	0.0000
fpg	0.04809	33.82	1	0.0000
bmi	-0.05181	22.06	1	0.0000
race	0.00330	0.21	1	0.6461
sex	-0.03560	30.49	1	0.0000
global test		0.00	5	1.0000

year==19

Test of proportional hazards assumption

	rho	chi2	df	Prob>chi2
age	-0.44672	5529.79	1	0.0000
fpg	-0.55970	1358.31	1	0.0000
bmi	-0.03993	18.46	1	0.0000
race	0.45140	1145.20	1	0.0000
sex	0.14766	733.09	1	0.0000
global test		0.00	5	1.0000

Appendix A.5.2. Weibull and Exponential regression models on total cost

CONVENTIONAL

Weibull regression

```
Weibull regression -- accelerated failure-time form

No. of subjects =         1138                     Number of obs    =      1138
No. of failures =          213
Time at risk    =  9500509.408
                                                   LR chi2(5)       =     41.86
Log likelihood  =   -575.43213                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   -.051336   .0098131    -5.23   0.000    -.0705694   -.0321026
         bmi |   .0032357   .0126375     0.26   0.798    -.0215333    .0280048
         fpg |   .0046819   .0304612     0.15   0.878    -.0550209    .0643848
        race |   .0592436   .1076293     0.55   0.582    -.1517061    .2701932
         sex |   .2290909   .1317603     1.74   0.082    -.0291545    .4873363
       _cons |   12.93738   .7914047    16.35   0.000     11.38626    14.48851
-------------+----------------------------------------------------------------
      /ln_p |   .1394609   .0429129     3.25   0.001     .0553531    .2235687
-------------+----------------------------------------------------------------
           p |   1.149654    .049335                      1.056914    1.250532
         1/p |    .869827   .0373268                        .79966     .946151
------------------------------------------------------------------------------
```

Exponential regression

```
Exponential regression -- accelerated failure-time form

No. of subjects =         1138                     Number of obs    =      1138
No. of failures =          213
Time at risk    =  9500509.408
                                                   LR chi2(5)       =     47.79
Log likelihood  =   -580.15453                     Prob > chi2      =    0.0000

------------------------------------------------------------------------------
          _t |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |  -.0622891   .0105939    -5.88   0.000    -.0830527   -.0415254
         bmi |   .0002449   .0142341     0.02   0.986    -.0276535    .0281432
         fpg |  -.0064317   .0342738    -0.19   0.851    -.0736071    .0607437
        race |   .0894702   .1232363     0.73   0.468    -.1520685    .3310088
         sex |   .2737993   .1502651     1.82   0.068    -.0207148    .5683134
       _cons |   13.77712   .8571194    16.07   0.000     12.09719    15.45704
------------------------------------------------------------------------------
```

INTENSIVE

Weibull regression

Weibull regression -- accelerated failure-time form

No. of subjects =	2729	Number of obs =	2729
No. of failures =	489		
Time at risk =	22024713.02		
		LR chi2(5) =	133.63
Log likelihood =	-1305.9663	Prob > chi2 =	0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.048417	.005711	-8.48	0.000	-.0596104	-.0372237
bmi	.0011463	.0074388	0.15	0.878	-.0134335	.015726
fpg	-.0453473	.0160413	-2.83	0.005	-.0767877	-.0139068
race	.0236118	.0650777	0.36	0.717	-.1039381	.1511616
sex	.496793	.0800094	6.21	0.000	.3399775	.6536085
_cons	12.79873	.4562066	28.05	0.000	11.90458	13.69288
/ln_p	.2637249	.0301343	8.75	0.000	.2046629	.322787
p	1.30177	.0392279			1.227111	1.380971
1/p	.7681848	.0231487			.7241281	.814922

Exponential regression

Exponential regression -- accelerated failure-time form

No. of subjects =	2729	Number of obs =	2729
No. of failures =	489		
Time at risk =	22024713.02		
		LR chi2(5) =	155.10
Log likelihood =	-1337.773	Prob > chi2 =	0.0000

_t	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age	-.0691696	.0069622	-9.94	0.000	-.0828153	-.055524
bmi	-.0022877	.0097351	-0.23	0.814	-.021368	.0167927
fpg	-.064018	.0204434	-3.13	0.002	-.1040863	-.0239497
race	.0542264	.0846879	0.64	0.522	-.1117589	.2202117
sex	.6271029	.1021918	6.14	0.000	.4268106	.8273951
_cons	14.29921	.5686191	25.15	0.000	13.18474	15.41368

Appendix A.5.3. Carides regression models (Carides et al, 2000)

Carides regression models for conventional

Total cost against time-to-failure using the uncensored cases only

```
      Source |       SS       df       MS              Number of obs =     213
-------------+------------------------------           F(  1,   211) =   10.03
       Model |  3.1175e+09     1  3.1175e+09           Prob > F      =  0.0018
    Residual |  6.5604e+10   211   310917115           R-squared     =  0.0454
-------------+------------------------------           Adj R-squared =  0.0408
       Total |  6.8721e+10   212   324155825           Root MSE      =   17633

------------------------------------------------------------------------------
          Mi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    timallde |   972.4001   307.0877     3.17   0.002     367.0471    1577.753
       _cons |   5137.195   2644.424     1.94   0.053    -75.68007    10350.07
------------------------------------------------------------------------------
```

Log transformed total cost against time-to-failure using the uncensored cases only

```
      Source |       SS       df       MS              Number of obs =     213
-------------+------------------------------           F(  1,   211) =   46.26
       Model |  45.0008132     1  45.0008132           Prob > F      =  0.0000
    Residual |  205.262457   211  .972807853           R-squared     =  0.1798
-------------+------------------------------           Adj R-squared =  0.1759
       Total |   250.26327   212  1.18048712           Root MSE      =  .98631

------------------------------------------------------------------------------
        lnMi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    timallde |   .1168289   .0171773     6.80   0.000     .0829679    .1506899
       _cons |   7.974748   .1479184    53.91   0.000      7.68316    8.266335
------------------------------------------------------------------------------
```

Log-Log transformed total cost against time-to-failure using the uncensored cases only

```
      Source |       SS       df       MS              Number of obs =     213
-------------+------------------------------           F(  1,   211) =   50.64
       Model |  .663258557     1  .663258557           Prob > F      =  0.0000
    Residual |  2.76336055   211  .013096496           R-squared     =  0.1936
-------------+------------------------------           Adj R-squared =  0.1897
       Total |  3.42661911   212  .016163298           Root MSE      =  .11444

------------------------------------------------------------------------------
      lnlnMi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    timallde |   .0141834    .001993     7.12   0.000     .0102546    .0181123
       _cons |   2.066175   .0171627   120.39   0.000     2.032342    2.100007
------------------------------------------------------------------------------
```

Carides regression models for Intensive

Total cost against time-to-failure using the uncensored cases only

```
      Source |       SS       df       MS              Number of obs =     489
-------------+------------------------------           F(  1,   487) =   70.99
       Model |  9.1749e+09     1  9.1749e+09           Prob > F      =  0.0000
    Residual |  6.2939e+10   487   129238492           R-squared     =  0.1272
-------------+------------------------------           Adj R-squared =  0.1254
       Total |  7.2114e+10   488   147774693           Root MSE      =   11368

------------------------------------------------------------------------------
          Mi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    timallde |   1053.654   125.0527     8.43   0.000     807.9443    1299.363
       _cons |   2731.592   1092.836     2.50   0.013     584.3366    4878.848
------------------------------------------------------------------------------
```

Log transformed total cost against time-to-failure using the uncensored cases only

```
      Source |       SS       df       MS              Number of obs =     489
-------------+------------------------------           F(  1,   487) =  205.11
       Model |  156.708772     1  156.708772           Prob > F      =  0.0000
    Residual |  372.078572   487  .764021708           R-squared     =  0.2964
-------------+------------------------------           Adj R-squared =  0.2949
       Total |  528.787344   488  1.08358062           Root MSE      =  .87408

------------------------------------------------------------------------------
        lnMi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    timallde |   .1377031    .009615    14.32   0.000     .1188111    .1565952
       _cons |   7.759076   .0840257    92.34   0.000     7.593978    7.924174
------------------------------------------------------------------------------
```

Log-Log transformed total cost against time-to-failure using the uncensored cases only

```
      Source |       SS       df       MS              Number of obs =     489
-------------+------------------------------           F(  1,   487) =  196.21
       Model |  2.31441056     1  2.31441056           Prob > F      =  0.0000
    Residual |  5.74446578   487  .011795618           R-squared     =  0.2872
-------------+------------------------------           Adj R-squared =  0.2857
       Total |  8.05887634   488  .016514091           Root MSE      =  .10861

------------------------------------------------------------------------------
      lnlnMi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    timallde |   .0167347   .0011947    14.01   0.000     .0143873    .0190821
       _cons |   2.040383   .0104405   195.43   0.000     2.019869    2.060897
------------------------------------------------------------------------------
```

Programs for obtaining bootstrap estimates for the standard errors of the mean for the Carides et al models

```
**Carides Models: Bootstrap for the standard errors **


program define pcarides
      if "`1'"=="?" {
             global S_1 "mclinear mclnMi mclnMism mclnlnMi mclnlnMism"
             exit
      }


**KM mean survival**

stset   timallde, failure(censorig==1)

sts gen survt_KM=s
gsort - survt_KM  timallde
quietly gen lagXi= timallde[_n-1]
replace lagXi=0 if lagXi==.
gen areaiXi= survt_KM*( timallde- lagXi)
egen meansurvtKM=sum(areaiXi)


**Regression models**

**Mi against time-to-failure**

regress Mi timallde if censorig==1

matrix beta=e(b)
svmat beta
gen b0x=beta2
egen b0=min(b0x)
gen b1x=beta1
egen b1=min(b1x)

gen meanclinear= b0+b1* meansurvtKM


**lnMi against time-to-failure**

gen lnMi=ln(Mi)

regress lnMi timallde if censorig==1

predict residlnMi, residuals

matrix betaln=e(b)
svmat betaln
gen b0lnx=betaln2
egen b0ln=min(b0lnx)
gen b1lnx=betaln1
egen b1ln=min(b1lnx)

gen eresidlnMi=exp( residlnMi)

egen seresidlnMi=sum( eresidlnMi)

gen smearlnMi=(1/1138)* seresidlnMi
```

```
gen  meanclnMi=exp(b0ln+ b1ln* meansurvtKM)

gen  meanclnMism=(exp(b0ln+ b1ln* meansurvtKM))* smearlnMi


**lnlnMi against time-to-failure**

gen lnlnMi=ln(lnMi)

regress lnlnMi timallde if censorig==1

predict residlnlnMi, residuals


matrix betalnln=e(b)
svmat betalnln
gen b0lnlnx=betalnln2
egen b0lnln=min(b0lnlnx)
gen b1lnlnx=betalnln1
egen b1lnln=min(b1lnlnx)

gen eresidlnlnMi=exp( residlnlnMi)

gen meanclnlnMi=exp(exp(b0lnln+b1lnln* meansurvtKM))

gen eelnlnMism=exp((exp(b0lnln+b1lnln* meansurvtKM))* eresidlnlnMi)

egen seelnlnMism=sum( eelnlnMism)

gen meanclnlnMism=(1/1138)* seelnlnMism

**

tempname y1
summarize meanclinear, meanonly
scalar `y1'=r(mean)

tempname y2
summarize meanclnMi, meanonly
scalar `y2'=r(mean)

tempname y3
summarize meanclnMism, meanonly
scalar `y3'=r(mean)

tempname y4
summarize meanclnlnMi, meanonly
scalar `y4'=r(mean)

summarize meanclnlnMism, meanonly

      post `1' (`y1') (`y2') (`y3') (`y4') (r(mean))
end

**

end of do-file

set seed 1001

bstrap pcarides, reps(1000) dots
saving(C:\WINDOWS\Desktop\regressions_new\Carides\bsconv1000.dta)

**
```

Appendix A.5.4. Ordinary least squares regression using the uncensored cases only

Using the complete set of covariates

CONVENTIONAL

```
      Source |       SS       df       MS              Number of obs =     213
-------------+------------------------------           F(  5,   207) =    1.21
       Model |  1.9587e+09     5   391745120           Prob > F      =  0.3034
    Residual |  6.6762e+10   207   322523233           R-squared     =  0.0285
-------------+------------------------------           Adj R-squared =  0.0050
       Total |  6.8721e+10   212   324155825           Root MSE      =   17959

------------------------------------------------------------------------------
          Mi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   262.4724   197.5792     1.33   0.185    -127.0531     651.998
         bmi |   454.0052   270.1044     1.68   0.094    -78.50296    986.5134
         fpg |   537.4346   611.4397     0.88   0.380    -668.0129    1742.882
        race |   1783.516   2204.647     0.81   0.419    -2562.924    6129.956
         sex |   1545.937   2646.605     0.58   0.560    -3671.818    6763.693
       _cons |  -23980.37   16190.03    -1.48   0.140    -55898.86    7938.124
------------------------------------------------------------------------------
```

INTENSIVE

```
      Source |       SS       df       MS              Number of obs =     489
-------------+------------------------------           F(  5,   483) =    0.66
       Model |   492715435     5  98543086.9           Prob > F      =  0.6505
    Residual |  7.1621e+10   483   148284337           R-squared     =  0.0068
-------------+------------------------------           Adj R-squared = -0.0034
       Total |  7.2114e+10   488   147774693           Root MSE      =   12177

------------------------------------------------------------------------------
          Mi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   57.55115   85.66172     0.67   0.502    -110.7645    225.8668
         bmi |   34.83646   123.7422     0.28   0.778     -208.303    277.9759
         fpg |  -176.5466   267.8438    -0.66   0.510    -702.8295    349.7363
        race |   146.7123   954.3455     0.15   0.878    -1728.469    2021.894
         sex |   1802.747   1250.537     1.44   0.150     -654.417    4259.911
       _cons |   5647.455   7148.846     0.79   0.430    -8399.223    19694.13
------------------------------------------------------------------------------
```

Using fasting plasma glucose (fpg) as the only covariate

CONVENTIONAL

```
      Source |       SS       df       MS              Number of obs =     213
-------------+------------------------------           F(  1,   211) =    1.00
       Model |   323160914     1   323160914           Prob > F      =  0.3192
    Residual |  6.8398e+10   211   324160540           R-squared     =  0.0047
-------------+------------------------------           Adj R-squared = -0.0000
       Total |  6.8721e+10   212   324155825           Root MSE      =   18004

------------------------------------------------------------------------------
          Mi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fpg |   602.4029   603.3339     1.00   0.319    -586.9315    1791.737
       _cons |   7367.779   5369.693     1.37   0.171    -3217.339     17952.9
------------------------------------------------------------------------------
```

INTENSIVE

```
      Source |       SS       df       MS              Number of obs =     489
-------------+------------------------------           F(  1,   487) =    0.30
       Model |  44656082.7     1  44656082.7           Prob > F      =  0.5830
    Residual |  7.2069e+10   487   147986436           R-squared     =  0.0006
-------------+------------------------------           Adj R-squared = -0.0014
       Total |  7.2114e+10   488   147774693           Root MSE      =   12165

------------------------------------------------------------------------------
          Mi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         fpg |  -144.1522   262.4169    -0.55   0.583    -659.7613    371.4569
       _cons |   12142.42   2403.761     5.05   0.000     7419.398    16865.44
------------------------------------------------------------------------------
```

Programs for obtaining bootstrap estimates for the standard errors of the coefficients and the mean for the OLS regression model

```
**OLS naive on total costs: Bootstrap estimates for the standard errors**

program define polsnaive
      if "`1'"=="?" {
            global S_1 "b0 b1 b2 b3 b4 b5 mclinear b0fpg b1fpg mcfpglinear"
            exit
      }

egen meanage=mean(age)
egen meanbmi=mean(bmi)
egen meanfpg=mean(fpg)
egen meanrace=mean(race)
egen meansex=mean(sex)

regress Mi age bmi fpg race sex if censorig==1

matrix betaallZ=e(b)

svmat betaallZ

gen b0x= betaallZ6
egen b0=min( b0x)

gen b1x= betaallZ1
egen b1=min( b1x)

gen b2x= betaallZ2
egen b2=min( b2x)

gen b3x= betaallZ3
egen b3=min( b3x)

gen b4x= betaallZ4
egen b4=min( b4x)

gen b5x= betaallZ5
egen b5=min( b5x)

drop  b0x b1x b2x b3x b4x b5x

gen meanclinear=b0+b1* meanage+b2* meanbmi+b3* meanfpg+b4* meanrace+b5* meansex
```

```
regress Mi fpg if censorig==1

matrix betafpg=e(b)
svmat betafpg

gen b0fpgx= betafpg2
egen b0fpg=min( b0fpgx)

gen b1fpgx= betafpg1
egen b1fpg=min( b1fpgx)

drop  b0fpgx b1fpgx

gen meancfpglinear=b0fpg+b1fpg* meanfpg

tempname y1
summarize b0, meanonly
scalar `y1'=r(mean)

tempname y2
summarize b1, meanonly
scalar `y2'=r(mean)

tempname y3
summarize b2, meanonly
scalar `y3'=r(mean)

tempname y4
summarize b3, meanonly
scalar `y4'=r(mean)

tempname y5
summarize b4, meanonly
scalar `y5'=r(mean)

tempname y6
summarize b5, meanonly
scalar `y6'=r(mean)

tempname y7
summarize meanclinear, meanonly
scalar `y7'=r(mean)

tempname y8
summarize b0fpg, meanonly
scalar `y8'=r(mean)

tempname y9
summarize b1fpg, meanonly
scalar `y9'=r(mean)

summarize meancfpglinear, meanonly

      post `1' (`y1') (`y2') (`y3') (`y4') (`y5') (`y6') (`y7') (`y8') (`y9')
(r(mean))

end

**
end of do-file

set seed 1001
bstrap polsnaive, reps(1000) dots
saving(C:\WINDOWS\Desktop\regressions_new\OLS_naive\bsconv1000.dta)
```

Appendix A.5.5. Programs for the Lin (2000) regression methodology using the total costs at the last contact dates or at the point of death

```
Based on equations (5.25) and (5.27) for the regression parameters and on
equations (5.29), (5.30), (5.31) and (5.32) for the coefficient standard errors


**Lin 2000: Total costs: Conventional (similarly for intensive)**

**Mean and coefficients' standard errors**

egen maxtimeL=max( timallde)

gen di=1 if censorig==1
replace di=0 if di==.

gen di_star=1 if (di==1 |  maxtimeL== timallde)
replace  di_star=0 if di_star==.

gen di_censor=1-di

stset  timallde, failure(di_star==0)

sts gen G_Tistar=s

egen meanage=mean(age)
egen meanbmi=mean( bmi)
egen meanfpg=mean(fpg)
egen meanrace=mean(race)
egen meansex=mean(sex)

gen int replicate=6

expand replicate

gen int const=1

sort ukno

by ukno: gen constx=sum(const)

sort ukno constx

gen Zi=const if constx==1
replace Zi=age if constx==2
replace Zi=bmi if constx==3
replace Zi=fpg if constx==4
replace Zi=race if constx==5
replace Zi=sex if constx==6

move  Zi age

gen Zi0_Zi0p=Zi*const
gen Zi1_Zi1p=Zi*age
gen Zi2_Zi2p=Zi*bmi
gen Zi3_Zi3p=Zi*fpg
gen Zi4_Zi4p=Zi*race
gen Zi5_Zi5p=Zi*sex

gen wZi0_Zi0p= (di_star/  G_Tistar)* Zi0_Zi0p
gen wZi1_Zi1p= (di_star/  G_Tistar)* Zi1_Zi1p
gen wZi2_Zi2p= (di_star/  G_Tistar)* Zi2_Zi2p
gen wZi3_Zi3p= (di_star/ G_Tistar)* Zi3_Zi3p
```

```
gen wZi4_Zi4p= (di_star/  G_Tistar)* Zi4_Zi4p
gen wZi5_Zi5p= (di_star/  G_Tistar)* Zi5_Zi5p

egen swZi0_Zi0p=sum(wZi0_Zi0p), by(constx)
egen swZi1_Zi1p=sum(wZi1_Zi1p), by(constx)
egen swZi2_Zi2p=sum(wZi2_Zi2p), by(constx)
egen swZi3_Zi3p=sum(wZi3_Zi3p), by(constx)
egen swZi4_Zi4p=sum(wZi4_Zi4p), by(constx)
egen swZi5_Zi5p=sum(wZi5_Zi5p), by(constx)

gen wYiZi=( di_star/ G_Tistar)*Mi*Zi

egen swYiZi=sum(wYiZi), by(constx)

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
ukno=="00011D", matrix(bterm1)

mkmat swYiZi if ukno=="00011D", matrix(bterm2)

matrix beta=syminv(bterm1)*bterm2

svmat beta

matrix list beta

egen b0x=min(beta1) if constx==1
egen b0=min(b0x)
egen b1x=min(beta1) if constx==2
egen b1=min(b1x)
egen b2x=min(beta1) if constx==3
egen b2=min(b2x)
egen b3x=min(beta1) if constx==4
egen b3=min(b3x)
egen b4x=min(beta1) if constx==5
egen b4=min(b4x)
egen b5x=min(beta1) if constx==6
egen b5=min(b5x)

drop b0x b1x b2x b3x b4x b5x

gen meancost=b0+b1* meanage+b2* meanbmi+b3* meanfpg+b4* meanrace+b5* meansex

sum meancost
```

****Standard errors for the coefficients****

```
gen beta_Zi=b0*const+b1* age+b2* bmi+b3* fpg+b4* race+b5* sex

gen Yi_bZi=Mi-beta_Zi

gen Bterm1i= (di_star/  G_Tistar)* Yi_bZi*Zi

gen Xi= timallde

**ties**
sort constx Xi
by constx Xi: gen ties=sum(const)
egen maxties=max(ties), by(constx Xi)

**

**For I(Ti_star>t in Q(t)**
```

```
gsort  constx -Xi ties
by constx: gen QXinumx=sum( Bterm1i)
gen QXinum=QXinumx if ties==maxties
egen QXinumxx=min(QXinum), by(constx Xi maxties)
replace QXinum=QXinumxx if QXinum==.

gsort  constx -Xi
by constx: gen QXidenomx=sum(const)
egen QXidenom=max(QXidenomx), by(constx Xi)

gsort constx -Xi ties

egen QXxx=sum(Bterm1i), by(constx Xi)
gen QXinumtest=QXinum-QXxx

gen QXitest=QXinumtest/ QXidenom

gen Bterm2itest=di_censor*QXitest
gen  djcens_QXjsXltest= di_censor* QXitest/ QXidenom

sort  constx Xi ties
by constx: gen Bterm3ixtest=sum(djcens_QXjsXltest)
gen Bterm3itest=Bterm3ixtest if ties==maxties
egen Bterm3ixxtest=min(Bterm3itest), by(constx Xi maxties)
replace Bterm3itest=Bterm3ixxtest if Bterm3itest==.

gen Bterm123itest= Bterm1i+Bterm2itest- Bterm3itest

gen  Bterm123i_1xtest=Bterm123itest if constx==1
egen  Bterm123i_1test=min( Bterm123i_1xtest), by(ukno)

gen  Bterm123i_2xtest=Bterm123itest if constx==2
egen  Bterm123i_2test=min( Bterm123i_2xtest), by(ukno)

gen  Bterm123i_3xtest=Bterm123itest if constx==3
egen  Bterm123i_3test=min( Bterm123i_3xtest), by(ukno)

gen  Bterm123i_4xtest=Bterm123itest if constx==4
egen  Bterm123i_4test=min( Bterm123i_4xtest), by(ukno)

gen  Bterm123i_5xtest=Bterm123itest if constx==5
egen  Bterm123i_5test=min( Bterm123i_5xtest), by(ukno)

gen  Bterm123i_6xtest=Bterm123itest if constx==6
egen  Bterm123i_6test=min( Bterm123i_6xtest), by(ukno)

gen Btermp_p1test= Bterm123itest* Bterm123i_1test
gen Btermp_p2test= Bterm123itest* Bterm123i_2test
gen Btermp_p3test= Bterm123itest* Bterm123i_3test
gen Btermp_p4test= Bterm123itest* Bterm123i_4test
gen Btermp_p5test= Bterm123itest* Bterm123i_5test
gen Btermp_p6test= Bterm123itest* Bterm123i_6test

egen sBtermp_p1test=sum( Btermp_p1test), by(constx)
egen sBtermp_p2test=sum( Btermp_p2test), by(constx)
egen sBtermp_p3test=sum( Btermp_p3test), by(constx)
egen sBtermp_p4test=sum( Btermp_p4test), by(constx)
egen sBtermp_p5test=sum( Btermp_p5test), by(constx)
egen sBtermp_p6test=sum( Btermp_p6test), by(constx)


mkmat   sBtermp_p1test sBtermp_p2test sBtermp_p3test sBtermp_p4test
sBtermp_p5test sBtermp_p6test if ukno=="00011D", matrix(seBxtest)
```

```
matrix seBtest=(1/1138)*seBxtest

svmat seBtest

matrix list seBtest

matrix covbetamattest=seinvA*seBtest*seinvA

svmat covbetamattest

matrix list covbetamattest

**
```

Bootstrap estimates of the standard errors for the coefficients and the mean for Lin (2000) using the total costs at the last contact dates or death

```
use "C:\Desktop\Lin2000\conv_orig.dta", clear
rename ukno ukno0
do "C:\Desktop\Lin2000\Total_costs\bs_total.txt"

where bs_total.txt is:

**Lin2000 on total costs: Conventional and/or Intensive**

**Bootstrap Estimation of standard errors for coefficients and mean**


program define ptotal
      if "`1'"=="?" {
            global S_1 "b0 b1 b2 b3 b4 b5 meanc"
            exit
      }

egen maxtimeL=max( timallde)

gen di=1 if censorig==1
replace di=0 if di==.

gen di_star=1 if (di==1 |  maxtimeL== timallde)
replace  di_star=0 if di_star==.

gen di_censor=1-di

stset  timallde, failure(di_star==0)

sts gen G_Tistar=s

egen meanage=mean(age)
egen meanbmi=mean( bmi)
egen meanfpg=mean(fpg)
egen meanrace=mean(race)
egen meansex=mean(sex)

gen int replicate=6

expand replicate

gen int const=1
```

```
sort ukno

by ukno: gen constx=sum(const)

sort ukno constx

gen Zi=const if constx==1
replace Zi=age if constx==2
replace Zi=bmi if constx==3
replace Zi=fpg if constx==4
replace Zi=race if constx==5
replace Zi=sex if constx==6

move  Zi age

gen Zi0_Zi0p=Zi*const
gen Zi1_Zi1p=Zi*age
gen Zi2_Zi2p=Zi*bmi
gen Zi3_Zi3p=Zi*fpg
gen Zi4_Zi4p=Zi*race
gen Zi5_Zi5p=Zi*sex

gen wZi0_Zi0p= (di_star/  G_Tistar)* Zi0_Zi0p
gen wZi1_Zi1p= (di_star/  G_Tistar)* Zi1_Zi1p
gen wZi2_Zi2p= (di_star/  G_Tistar)* Zi2_Zi2p
gen wZi3_Zi3p= (di_star/ G_Tistar)* Zi3_Zi3p
gen wZi4_Zi4p= (di_star/  G_Tistar)* Zi4_Zi4p
gen wZi5_Zi5p= (di_star/  G_Tistar)* Zi5_Zi5p

egen swZi0_Zi0p=sum(wZi0_Zi0p), by(constx)
egen swZi1_Zi1p=sum(wZi1_Zi1p), by(constx)
egen swZi2_Zi2p=sum(wZi2_Zi2p), by(constx)
egen swZi3_Zi3p=sum(wZi3_Zi3p), by(constx)
egen swZi4_Zi4p=sum(wZi4_Zi4p), by(constx)
egen swZi5_Zi5p=sum(wZi5_Zi5p), by(constx)

gen wYiZi=( di_star/ G_Tistar)*Mi*Zi

egen swYiZi=sum(wYiZi), by(constx)


collapse meanage meanbmi meanfpg meanrace meansex swYiZi swZi0_Zi0p swZi1_Zi1p
swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p, by(constx)

sort constx

mkmat swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p,
matrix(bterm1)

mkmat swYiZi, matrix(bterm2)

matrix beta=syminv(bterm1)*bterm2

svmat beta

gen b0x=beta1 if constx==1
egen b0=min(b0x)
gen b1x=beta1 if constx==2
egen b1=min(b1x)
gen b2x=beta1 if constx==3
egen b2=min(b2x)
gen b3x=beta1 if constx==4
egen b3=min(b3x)
```

```
gen b4x=beta1 if constx==5
egen b4=min(b4x)
gen b5x=beta1 if constx==6
egen b5=min(b5x)

gen meancost=b0+b1* meanage+b2* meanbmi+b3* meanfpg+b4* meanrace+b5* meansex

tempname y1
summarize b0, meanonly
scalar `y1'=r(mean)

tempname y2
summarize b1, meanonly
scalar `y2'=r(mean)

tempname y3
summarize b2, meanonly
scalar `y3'=r(mean)

tempname y4
summarize b3, meanonly
scalar `y4'=r(mean)

tempname y5
summarize b4, meanonly
scalar `y5'=r(mean)

tempname y6
summarize b5, meanonly
scalar `y6'=r(mean)

summarize meancost, meanonly
      post `1' (`y1') (`y2') (`y3') (`y4') (`y5') (`y6') (r(mean))
end

**
end of do-file

set seed 1001

bstrap ptotal, reps(1000) dots cluster(ukno0) idcluster(ukno)
saving(C:\Documents and
Settings\raikou\Desktop\Lin2000\Total_costs\Conventional\bsconv1000.dta)

**
```

Appendix A.5.6. Programs for the Lin (2000) regression methodology using multiple time intervals

Based on equations are given by (5.33) and (5.35) for the coefficients and by (5.36), (5.37), (5.38), (5.39), (5.40) and (5.41) for their standard errors.

```
**Lin 2000 on Annual Costs: Conventional (similarly for intensive) **

gen tk_1=year-1
gen tk=year

gen Xik=min(Xi, tk)

egen maxtimeL=max(Xi)

gen dik_star=1 if (Xik==tk | (Xik==Xi & di==1) | (Xik==Xi & maxtimeL==Xi))
replace  dik_star=0 if dik_star==.

stset Xik, failure( dik_star==0)
sts gen GTik_star=s, by(tk)

drop _st _d _t _t0

egen meanage=mean(age)
egen meanbmi=mean(bmi)
egen meanfpg=mean(fpg)
egen meanrace=mean(race)
egen meansex=mean(sex)

gen int replicate=6
expand replicate

gen int const=1

sort ukno year

by ukno year: gen constx=sum(const)

sort ukno year constx

gen Zi=const if constx==1
replace Zi=age if constx==2
replace Zi=bmi if constx==3
replace Zi=fpg if constx==4
replace Zi=race if constx==5
replace Zi=sex if constx==6

move Zi age
move constx age_entr
move const  age
drop age_entr maxyear gender

sort  ukno year constx

gen Zi0_Zi0p=Zi*const
gen Zi1_Zi1p=Zi*age
gen Zi2_Zi2p=Zi*bmi
gen Zi3_Zi3p=Zi*fpg
gen Zi4_Zi4p=Zi*race
gen Zi5_Zi5p=Zi*sex

gen wZi0_Zi0p= ( dik_star/  GTik_star)* Zi0_Zi0p
```

```
gen wZi1_Zi1p= ( dik_star/  GTik_star)* Zi1_Zi1p
gen wZi2_Zi2p= ( dik_star/  GTik_star)* Zi2_Zi2p
gen wZi3_Zi3p= ( dik_star/  GTik_star)* Zi3_Zi3p
gen wZi4_Zi4p= ( dik_star/  GTik_star)* Zi4_Zi4p
gen wZi5_Zi5p= ( dik_star/  GTik_star)* Zi5_Zi5p

egen swZi0_Zi0p=sum(wZi0_Zi0p), by(tk constx)
egen swZi1_Zi1p=sum(wZi1_Zi1p), by(tk constx)
egen swZi2_Zi2p=sum(wZi2_Zi2p), by(tk constx)
egen swZi3_Zi3p=sum(wZi3_Zi3p), by(tk constx)
egen swZi4_Zi4p=sum(wZi4_Zi4p), by(tk constx)
egen swZi5_Zi5p=sum(wZi5_Zi5p), by(tk constx)

gen wYikZi=( dik_star/ GTik_star)* Mik*Zi
egen swYikZi=sum(wYikZi), by(tk constx)


sort ukno tk constx

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==1), matrix(bk1term1)

mkmat  swYikZi if (ukno=="00011D" & tk==1), matrix(bk1term2)

matrix betak1=syminv(bk1term1)*bk1term2

svmat betak1

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==2), matrix(bk2term1)

mkmat  swYikZi if (ukno=="00011D" & tk==2), matrix(bk2term2)

matrix betak2=syminv(bk2term1)*bk2term2

svmat betak2

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==3), matrix(bk3term1)

mkmat  swYikZi if (ukno=="00011D" & tk==3), matrix(bk3term2)

matrix betak3=syminv(bk3term1)*bk3term2

svmat betak3

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==4), matrix(bk4term1)

mkmat  swYikZi if (ukno=="00011D" & tk==4), matrix(bk4term2)

matrix betak4=syminv(bk4term1)*bk4term2

svmat betak4

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==5), matrix(bk5term1)

mkmat  swYikZi if (ukno=="00011D" & tk==5), matrix(bk5term2)
```

```
matrix betak5=syminv(bk5term1)*bk5term2

svmat betak5

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==6), matrix(bk6term1)

mkmat  swYikZi if (ukno=="00011D" & tk==6), matrix(bk6term2)

matrix betak6=syminv(bk6term1)*bk6term2

svmat betak6

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==7), matrix(bk7term1)

mkmat  swYikZi if (ukno=="00011D" & tk==7), matrix(bk7term2)

matrix betak7=syminv(bk7term1)*bk7term2

svmat betak7

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==8), matrix(bk8term1)

mkmat  swYikZi if (ukno=="00011D" & tk==8), matrix(bk8term2)

matrix betak8=syminv(bk8term1)*bk8term2

svmat betak8

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==9), matrix(bk9term1)

mkmat  swYikZi if (ukno=="00011D" & tk==9), matrix(bk9term2)

matrix betak9=syminv(bk9term1)*bk9term2

*
svmat betak9

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==10), matrix(bk10term1)

mkmat  swYikZi if (ukno=="00011D" & tk==10), matrix(bk10term2)

matrix betak10=syminv(bk10term1)*bk10term2

*
svmat betak10

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==11), matrix(bk11term1)

mkmat  swYikZi if (ukno=="00011D" & tk==11), matrix(bk11term2)

matrix betak11=syminv(bk11term1)*bk11term2

svmat betak11
```

```
*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==12), matrix(bk12term1)

mkmat  swYikZi if (ukno=="00011D" & tk==12), matrix(bk12term2)

matrix betak12=syminv(bk12term1)*bk12term2

svmat betak12

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==13), matrix(bk13term1)

mkmat  swYikZi if (ukno=="00011D" & tk==13), matrix(bk13term2)

matrix betak13=syminv(bk13term1)*bk13term2

svmat betak13

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==14), matrix(bk14term1)

mkmat  swYikZi if (ukno=="00011D" & tk==14), matrix(bk14term2)

matrix betak14=syminv(bk14term1)*bk14term2

svmat betak14

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==15), matrix(bk15term1)

mkmat  swYikZi if (ukno=="00011D" & tk==15), matrix(bk15term2)

matrix betak15=syminv(bk15term1)*bk15term2

svmat betak15

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==16), matrix(bk16term1)

mkmat  swYikZi if (ukno=="00011D" & tk==16), matrix(bk16term2)

matrix betak16=syminv(bk16term1)*bk16term2

svmat betak16

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==17), matrix(bk17term1)

mkmat  swYikZi if (ukno=="00011D" & tk==17), matrix(bk17term2)

matrix betak17=syminv(bk17term1)*bk17term2

svmat betak17

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==18), matrix(bk18term1)
```

```
mkmat  swYikZi if (ukno=="00011D" & tk==18), matrix(bk18term2)

matrix betak18=syminv(bk18term1)*bk18term2

svmat betak18

*
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
(ukno=="00011D" & tk==19), matrix(bk19term1)

mkmat  swYikZi if (ukno=="00011D" & tk==19), matrix(bk19term2)

matrix betak19=syminv(bk19term1)*bk19term2

svmat betak19

*
matrix
beta=betak1+betak2+betak3+betak4+betak5+betak6+betak7+betak8+betak9+betak10+beta
k11+betak12+betak13+betak14+betak15+betak16+betak17+betak18+betak19

svmat beta

*
gen b0x=beta1 if constx==1
egen b0=min(b0x)

gen b1x=beta1 if constx==2
egen b1=min(b1x)

gen b2x=beta1 if constx==3
egen b2=min(b2x)

gen b3x=beta1 if constx==4
egen b3=min(b3x)

gen b4x=beta1 if constx==5
egen b4=min(b4x)

gen b5x=beta1 if constx==6
egen b5=min(b5x)

matrix list beta

sum b0 b1 b2 b3 b4 b5

drop  b0x b1x b2x b3x b4x b5x

gen meancost= b0+b1* meanage+b2* meanbmi+b3* meanfpg+b4* meanrace+b5* meansex

sum meancost
```

****For the standard errors of the coefficients****

```
gen b0k1x=betak11 if (constx==1)
egen b0k1=min(b0k1x)

gen b1k1x=betak11 if (constx==2)
egen b1k1=min(b1k1x)

gen b2k1x=betak11 if (constx==3)
egen b2k1=min(b2k1x)
```

```
gen b3k1x=betak11 if (constx==4)
egen b3k1=min(b3k1x)

gen b4k1x=betak11 if (constx==5)
egen b4k1=min(b4k1x)

gen b5k1x=betak11 if (constx==6)
egen b5k1=min(b5k1x)

gen b0k2x=betak21 if (constx==1)
egen b0k2=min(b0k2x)

gen b1k2x=betak21 if (constx==2)
egen b1k2=min(b1k2x)

gen b2k2x=betak21 if (constx==3)
egen b2k2=min(b2k2x)

gen b3k2x=betak21 if (constx==4)
egen b3k2=min(b3k2x)

gen b4k2x=betak21 if (constx==5)
egen b4k2=min(b4k2x)

gen b5k2x=betak21 if (constx==6)
egen b5k2=min(b5k2x)

gen b0k3x=betak31 if (constx==1)
egen b0k3=min(b0k3x)

gen b1k3x=betak31 if (constx==2)
egen b1k3=min(b1k3x)

gen b2k3x=betak31 if (constx==3)
egen b2k3=min(b2k3x)

gen b3k3x=betak31 if (constx==4)
egen b3k3=min(b3k3x)

gen b4k3x=betak31 if (constx==5)
egen b4k3=min(b4k3x)

gen b5k3x=betak31 if (constx==6)
egen b5k3=min(b5k3x)

gen b0k4x=betak41 if (constx==1)
egen b0k4=min(b0k4x)

gen b1k4x=betak41 if (constx==2)
egen b1k4=min(b1k4x)

gen b2k4x=betak41 if (constx==3)
egen b2k4=min(b2k4x)

gen b3k4x=betak41 if (constx==4)
egen b3k4=min(b3k4x)

gen b4k4x=betak41 if (constx==5)
egen b4k4=min(b4k4x)

gen b5k4x=betak41 if (constx==6)
egen b5k4=min(b5k4x)

gen b0k5x=betak51 if (constx==1)
```

```
egen b0k5=min(b0k5x)

gen b1k5x=betak51 if (constx==2)
egen b1k5=min(b1k5x)

gen b2k5x=betak51 if (constx==3)
egen b2k5=min(b2k5x)

gen b3k5x=betak51 if (constx==4)
egen b3k5=min(b3k5x)

gen b4k5x=betak51 if (constx==5)
egen b4k5=min(b4k5x)

gen b5k5x=betak51 if (constx==6)
egen b5k5=min(b5k5x)

gen b0k6x=betak61 if (constx==1)
egen b0k6=min(b0k6x)

gen b1k6x=betak61 if (constx==2)
egen b1k6=min(b1k6x)

gen b2k6x=betak61 if (constx==3)
egen b2k6=min(b2k6x)

gen b3k6x=betak61 if (constx==4)
egen b3k6=min(b3k6x)

gen b4k6x=betak61 if (constx==5)
egen b4k6=min(b4k6x)

gen b5k6x=betak61 if (constx==6)
egen b5k6=min(b5k6x)

gen b0k7x=betak71 if (constx==1)
egen b0k7=min(b0k7x)

gen b1k7x=betak71 if (constx==2)
egen b1k7=min(b1k7x)

gen b2k7x=betak71 if (constx==3)
egen b2k7=min(b2k7x)

gen b3k7x=betak71 if (constx==4)
egen b3k7=min(b3k7x)

gen b4k7x=betak71 if (constx==5)
egen b4k7=min(b4k7x)

gen b5k7x=betak71 if (constx==6)
egen b5k7=min(b5k7x)

gen b0k8x=betak81 if (constx==1)
egen b0k8=min(b0k8x)

gen b1k8x=betak81 if (constx==2)
egen b1k8=min(b1k8x)

gen b2k8x=betak81 if (constx==3)
egen b2k8=min(b2k8x)

gen b3k8x=betak81 if (constx==4)
egen b3k8=min(b3k8x)
```

```
gen b4k8x=betak81 if (constx==5)
egen b4k8=min(b4k8x)

gen b5k8x=betak81 if (constx==6)
egen b5k8=min(b5k8x)

gen b0k9x=betak91 if (constx==1)
egen b0k9=min(b0k9x)

gen b1k9x=betak91 if (constx==2)
egen b1k9=min(b1k9x)

gen b2k9x=betak91 if (constx==3)
egen b2k9=min(b2k9x)

gen b3k9x=betak91 if (constx==4)
egen b3k9=min(b3k9x)

gen b4k9x=betak91 if (constx==5)
egen b4k9=min(b4k9x)

gen b5k9x=betak91 if (constx==6)
egen b5k9=min(b5k9x)

gen b0k10x=betak101 if (constx==1)
egen b0k10=min(b0k10x)

gen b1k10x=betak101 if (constx==2)
egen b1k10=min(b1k10x)

gen b2k10x=betak101 if (constx==3)
egen b2k10=min(b2k10x)

gen b3k10x=betak101 if (constx==4)
egen b3k10=min(b3k10x)

gen b4k10x=betak101 if (constx==5)
egen b4k10=min(b4k10x)

gen b5k10x=betak101 if (constx==6)
egen b5k10=min(b5k10x)

gen b0k11x=betak111 if (constx==1)
egen b0k11=min(b0k11x)

gen b1k11x=betak111 if (constx==2)
egen b1k11=min(b1k11x)

gen b2k11x=betak111 if (constx==3)
egen b2k11=min(b2k11x)

gen b3k11x=betak111 if (constx==4)
egen b3k11=min(b3k11x)

gen b4k11x=betak111 if (constx==5)
egen b4k11=min(b4k11x)

gen b5k11x=betak111 if (constx==6)
egen b5k11=min(b5k11x)

gen b0k12x=betak121 if (constx==1)
egen b0k12=min(b0k12x)
```

```
gen b1k12x=betak121 if (constx==2)
egen b1k12=min(b1k12x)

gen b2k12x=betak121 if (constx==3)
egen b2k12=min(b2k12x)

gen b3k12x=betak121 if (constx==4)
egen b3k12=min(b3k12x)

gen b4k12x=betak121 if (constx==5)
egen b4k12=min(b4k12x)

gen b5k12x=betak121 if (constx==6)
egen b5k12=min(b5k12x)

gen b0k13x=betak131 if (constx==1)
egen b0k13=min(b0k13x)

gen b1k13x=betak131 if (constx==2)
egen b1k13=min(b1k13x)

gen b2k13x=betak131 if (constx==3)
egen b2k13=min(b2k13x)

gen b3k13x=betak131 if (constx==4)
egen b3k13=min(b3k13x)

gen b4k13x=betak131 if (constx==5)
egen b4k13=min(b4k13x)

gen b5k13x=betak131 if (constx==6)
egen b5k13=min(b5k13x)

gen b0k14x=betak141 if (constx==1)
egen b0k14=min(b0k14x)

gen b1k14x=betak141 if (constx==2)
egen b1k14=min(b1k14x)

gen b2k14x=betak141 if (constx==3)
egen b2k14=min(b2k14x)

gen b3k14x=betak141 if (constx==4)
egen b3k14=min(b3k14x)

gen b4k14x=betak141 if (constx==5)
egen b4k14=min(b4k14x)

gen b5k14x=betak141 if (constx==6)
egen b5k14=min(b5k14x)

gen b0k15x=betak151 if (constx==1)
egen b0k15=min(b0k15x)

gen b1k15x=betak151 if (constx==2)
egen b1k15=min(b1k15x)

gen b2k15x=betak151 if (constx==3)
egen b2k15=min(b2k15x)

gen b3k15x=betak151 if (constx==4)
egen b3k15=min(b3k15x)

gen b4k15x=betak151 if (constx==5)
```

```
egen b4k15=min(b4k15x)

gen b5k15x=betak151 if (constx==6)
egen b5k15=min(b5k15x)

gen b0k16x=betak161 if (constx==1)
egen b0k16=min(b0k16x)

gen b1k16x=betak161 if (constx==2)
egen b1k16=min(b1k16x)

gen b2k16x=betak161 if (constx==3)
egen b2k16=min(b2k16x)

gen b3k16x=betak161 if (constx==4)
egen b3k16=min(b3k16x)

gen b4k16x=betak161 if (constx==5)
egen b4k16=min(b4k16x)

gen b5k16x=betak161 if (constx==6)
egen b5k16=min(b5k16x)

gen b0k17x=betak171 if (constx==1)
egen b0k17=min(b0k17x)

gen b1k17x=betak171 if (constx==2)
egen b1k17=min(b1k17x)

gen b2k17x=betak171 if (constx==3)
egen b2k17=min(b2k17x)

gen b3k17x=betak171 if (constx==4)
egen b3k17=min(b3k17x)

gen b4k17x=betak171 if (constx==5)
egen b4k17=min(b4k17x)

gen b5k17x=betak171 if (constx==6)
egen b5k17=min(b5k17x)

gen b0k18x=betak181 if (constx==1)
egen b0k18=min(b0k18x)

gen b1k18x=betak181 if (constx==2)
egen b1k18=min(b1k18x)

gen b2k18x=betak181 if (constx==3)
egen b2k18=min(b2k18x)

gen b3k18x=betak181 if (constx==4)
egen b3k18=min(b3k18x)

gen b4k18x=betak181 if (constx==5)
egen b4k18=min(b4k18x)

gen b5k18x=betak181 if (constx==6)
egen b5k18=min(b5k18x)

gen b0k19x=betak191 if (constx==1)
egen b0k19=min(b0k19x)

gen b1k19x=betak191 if (constx==2)
egen b1k19=min(b1k19x)
```

```
gen b2k19x=betak191 if (constx==3)
egen b2k19=min(b2k19x)

gen b3k19x=betak191 if (constx==4)
egen b3k19=min(b3k19x)

gen b4k19x=betak191 if (constx==5)
egen b4k19=min(b4k19x)

gen b5k19x=betak191 if (constx==6)
egen b5k19=min(b5k19x)


**

gen b0k=b0k1 if tk==1
replace b0k=b0k2 if tk==2
replace b0k=b0k3 if tk==3
replace b0k=b0k4 if tk==4
replace b0k=b0k5 if tk==5
replace b0k=b0k6 if tk==6
replace b0k=b0k7 if tk==7
replace b0k=b0k8 if tk==8
replace b0k=b0k9 if tk==9
replace b0k=b0k10 if tk==10
replace b0k=b0k11 if tk==11
replace b0k=b0k12 if tk==12
replace b0k=b0k13 if tk==13
replace b0k=b0k14 if tk==14
replace b0k=b0k15 if tk==15
replace b0k=b0k16 if tk==16
replace b0k=b0k17 if tk==17
replace b0k=b0k18 if tk==18
replace b0k=b0k19 if tk==19

*
gen b1k=b1k1 if tk==1
replace b1k=b1k2 if tk==2
replace b1k=b1k3 if tk==3
replace b1k=b1k4 if tk==4
replace b1k=b1k5 if tk==5
replace b1k=b1k6 if tk==6
replace b1k=b1k7 if tk==7
replace b1k=b1k8 if tk==8
replace b1k=b1k9 if tk==9
replace b1k=b1k10 if tk==10
replace b1k=b1k11 if tk==11
replace b1k=b1k12 if tk==12
replace b1k=b1k13 if tk==13
replace b1k=b1k14 if tk==14
replace b1k=b1k15 if tk==15
replace b1k=b1k16 if tk==16
replace b1k=b1k17 if tk==17
replace b1k=b1k18 if tk==18
replace b1k=b1k19 if tk==19

*
gen b2k=b2k1 if tk==1
replace b2k=b2k2 if tk==2
replace b2k=b2k3 if tk==3
replace b2k=b2k4 if tk==4
replace b2k=b2k5 if tk==5
replace b2k=b2k6 if tk==6
```

```
replace b2k=b2k7 if tk==7
replace b2k=b2k8 if tk==8
replace b2k=b2k9 if tk==9
replace b2k=b2k10 if tk==10
replace b2k=b2k11 if tk==11
replace b2k=b2k12 if tk==12
replace b2k=b2k13 if tk==13
replace b2k=b2k14 if tk==14
replace b2k=b2k15 if tk==15
replace b2k=b2k16 if tk==16
replace b2k=b2k17 if tk==17
replace b2k=b2k18 if tk==18
replace b2k=b2k19 if tk==19

*
gen b3k=b3k1 if tk==1
replace b3k=b3k2 if tk==2
replace b3k=b3k3 if tk==3
replace b3k=b3k4 if tk==4
replace b3k=b3k5 if tk==5
replace b3k=b3k6 if tk==6
replace b3k=b3k7 if tk==7
replace b3k=b3k8 if tk==8
replace b3k=b3k9 if tk==9
replace b3k=b3k10 if tk==10
replace b3k=b3k11 if tk==11
replace b3k=b3k12 if tk==12
replace b3k=b3k13 if tk==13
replace b3k=b3k14 if tk==14
replace b3k=b3k15 if tk==15
replace b3k=b3k16 if tk==16
replace b3k=b3k17 if tk==17
replace b3k=b3k18 if tk==18
replace b3k=b3k19 if tk==19

*
gen b4k=b4k1 if tk==1
replace b4k=b4k2 if tk==2
replace b4k=b4k3 if tk==3
replace b4k=b4k4 if tk==4
replace b4k=b4k5 if tk==5
replace b4k=b4k6 if tk==6
replace b4k=b4k7 if tk==7
replace b4k=b4k8 if tk==8
replace b4k=b4k9 if tk==9
replace b4k=b4k10 if tk==10
replace b4k=b4k11 if tk==11
replace b4k=b4k12 if tk==12
replace b4k=b4k13 if tk==13
replace b4k=b4k14 if tk==14
replace b4k=b4k15 if tk==15
replace b4k=b4k16 if tk==16
replace b4k=b4k17 if tk==17
replace b4k=b4k18 if tk==18
replace b4k=b4k19 if tk==19

*
gen b5k=b5k1 if tk==1
replace b5k=b5k2 if tk==2
replace b5k=b5k3 if tk==3
replace b5k=b5k4 if tk==4
replace b5k=b5k5 if tk==5
replace b5k=b5k6 if tk==6
replace b5k=b5k7 if tk==7
```

```
replace b5k=b5k8 if tk==8
replace b5k=b5k9 if tk==9
replace b5k=b5k10 if tk==10
replace b5k=b5k11 if tk==11
replace b5k=b5k12 if tk==12
replace b5k=b5k13 if tk==13
replace b5k=b5k14 if tk==14
replace b5k=b5k15 if tk==15
replace b5k=b5k16 if tk==16
replace b5k=b5k17 if tk==17
replace b5k=b5k18 if tk==18
replace b5k=b5k19 if tk==19

*
gen betak_Zi=b0k+b1k*age+b2k*bmi+b3k*fpg+b4k*race+b5k*sex

gen Yik_bkZi=Mik-betak_Zi

gen ksiki_term1= (dik_star/ GTik_star)* Yik_bkZi*Zi

**ties**
sort constx tk Xi
by constx tk Xi: gen ties=sum(const)
egen maxties=max(ties), by(constx tk Xi)

**

**For I(Ti_star>t) in Q(t)**

gsort  constx tk -Xik -Xi ties
by constx tk: gen QkXinumx=sum(ksiki_term1)
gen QkXinum=QkXinumx if ties==maxties
egen QkXinumxx=min(QkXinum), by(constx tk Xi maxties)
replace QkXinum=QkXinumxx if QkXinum==.
replace QkXinum=0 if Xik<Xi

gsort  constx tk -Xi
by constx tk: gen QkXidenomx=sum(const)
egen QkXidenom=max(QkXidenomx), by(constx tk Xi)

gen di_censor=1-di

gsort constx -Xik -Xi ties

egen QkXxx=sum(ksiki_term1), by(constx tk Xik)
replace QkXxx=0 if Xik<Xi
gen QkXinumtest=QkXinum-QkXxx
replace QkXinumtest=0 if Xik<Xi

gen QkXitest=QkXinumtest/ QkXidenom

gen ksiki_term2test=di_censor*QkXitest

gen  djcens_QkXjsXltest= di_censor* QkXitest/ QkXidenom

sort  constx tk Xi ties
by constx tk: gen ksiki_term3xtest=sum(djcens_QkXjsXltest)
gen ksiki_term3test=ksiki_term3xtest if ties==maxties
egen ksiki_term3xxtest=min(ksiki_term3test), by(constx tk Xi maxties)
replace ksiki_term3test=ksiki_term3xxtest if ksiki_term3test==.

gen ksikitest=ksiki_term1+ksiki_term2test-ksiki_term3test

**
```

```
**l=1**
gen ksikl1_1x= ksikitest if (tk==1 & constx==1)
egen ksikl1_1=min( ksikl1_1x), by(ukno)

gen ksikl1_2x= ksikitest if (tk==1 & constx==2)
egen ksikl1_2=min( ksikl1_2x), by(ukno)

gen ksikl1_3x= ksikitest if (tk==1 & constx==3)
egen ksikl1_3=min( ksikl1_3x), by(ukno)

gen ksikl1_4x= ksikitest if (tk==1 & constx==4)
egen ksikl1_4=min( ksikl1_4x), by(ukno)

gen ksikl1_5x= ksikitest if (tk==1 & constx==5)
egen ksikl1_5=min( ksikl1_5x), by(ukno)

gen ksikl1_6x= ksikitest if (tk==1 & constx==6)
egen ksikl1_6=min( ksikl1_6x), by(ukno)

drop ksikl1_1x ksikl1_2x ksikl1_3x ksikl1_4x ksikl1_5x ksikl1_6x

**

**l=2**
gen ksikl2_1x= ksikitest if (tk==2 & constx==1)
egen ksikl2_1=min( ksikl2_1x), by(ukno)

gen ksikl2_2x= ksikitest if (tk==2 & constx==2)
egen ksikl2_2=min( ksikl2_2x), by(ukno)

gen ksikl2_3x= ksikitest if (tk==2 & constx==3)
egen ksikl2_3=min( ksikl2_3x), by(ukno)

gen ksikl2_4x= ksikitest if (tk==2 & constx==4)
egen ksikl2_4=min( ksikl2_4x), by(ukno)

gen ksikl2_5x= ksikitest if (tk==2 & constx==5)
egen ksikl2_5=min( ksikl2_5x), by(ukno)

gen ksikl2_6x= ksikitest if (tk==2 & constx==6)
egen ksikl2_6=min( ksikl2_6x), by(ukno)

drop ksikl2_1x ksikl2_2x ksikl2_3x ksikl2_4x ksikl2_5x ksikl2_6x

**

**l=3**
gen ksikl3_1x= ksikitest if (tk==3 & constx==1)
egen ksikl3_1=min( ksikl3_1x), by(ukno)

gen ksikl3_2x= ksikitest if (tk==3 & constx==2)
egen ksikl3_2=min( ksikl3_2x), by(ukno)

gen ksikl3_3x= ksikitest if (tk==3 & constx==3)
egen ksikl3_3=min( ksikl3_3x), by(ukno)

gen ksikl3_4x= ksikitest if (tk==3 & constx==4)
egen ksikl3_4=min( ksikl3_4x), by(ukno)

gen ksikl3_5x= ksikitest if (tk==3 & constx==5)
egen ksikl3_5=min( ksikl3_5x), by(ukno)

gen ksikl3_6x= ksikitest if (tk==3 & constx==6)
```

```
egen ksikl3_6=min( ksikl3_6x), by(ukno)

drop ksikl3_1x ksikl3_2x ksikl3_3x ksikl3_4x ksikl3_5x ksikl3_6x

**

**l=4**
gen ksikl4_1x= ksikitest if (tk==4 & constx==1)
egen ksikl4_1=min( ksikl4_1x), by(ukno)

gen ksikl4_2x= ksikitest if (tk==4 & constx==2)
egen ksikl4_2=min( ksikl4_2x), by(ukno)

gen ksikl4_3x= ksikitest if (tk==4 & constx==3)
egen ksikl4_3=min( ksikl4_3x), by(ukno)

gen ksikl4_4x= ksikitest if (tk==4 & constx==4)
egen ksikl4_4=min( ksikl4_4x), by(ukno)

gen ksikl4_5x= ksikitest if (tk==4 & constx==5)
egen ksikl4_5=min( ksikl4_5x), by(ukno)

gen ksikl4_6x= ksikitest if (tk==4 & constx==6)
egen ksikl4_6=min( ksikl4_6x), by(ukno)

drop ksikl4_1x ksikl4_2x ksikl4_3x ksikl4_4x ksikl4_5x ksikl4_6x

**
**l=5**
gen ksikl5_1x= ksikitest if (tk==5 & constx==1)
egen ksikl5_1=min( ksikl5_1x), by(ukno)

gen ksikl5_2x= ksikitest if (tk==5 & constx==2)
egen ksikl5_2=min( ksikl5_2x), by(ukno)

gen ksikl5_3x= ksikitest if (tk==5 & constx==3)
egen ksikl5_3=min( ksikl5_3x), by(ukno)

gen ksikl5_4x= ksikitest if (tk==5 & constx==4)
egen ksikl5_4=min( ksikl5_4x), by(ukno)

gen ksikl5_5x= ksikitest if (tk==5 & constx==5)
egen ksikl5_5=min( ksikl5_5x), by(ukno)

gen ksikl5_6x= ksikitest if (tk==5 & constx==6)
egen ksikl5_6=min( ksikl5_6x), by(ukno)

drop ksikl5_1x ksikl5_2x ksikl5_3x ksikl5_4x ksikl5_5x ksikl5_6x

**
**l=6**
gen ksikl6_1x= ksikitest if (tk==6 & constx==1)
egen ksikl6_1=min( ksikl6_1x), by(ukno)

gen ksikl6_2x= ksikitest if (tk==6 & constx==2)
egen ksikl6_2=min( ksikl6_2x), by(ukno)

gen ksikl6_3x= ksikitest if (tk==6 & constx==3)
egen ksikl6_3=min( ksikl6_3x), by(ukno)

gen ksikl6_4x= ksikitest if (tk==6 & constx==4)
egen ksikl6_4=min( ksikl6_4x), by(ukno)

gen ksikl6_5x= ksikitest if (tk==6 & constx==5)
```

```
egen ksikl6_5=min( ksikl6_5x), by(ukno)

gen ksikl6_6x= ksikitest if (tk==6 & constx==6)
egen ksikl6_6=min( ksikl6_6x), by(ukno)

drop ksikl6_1x ksikl6_2x ksikl6_3x ksikl6_4x ksikl6_5x ksikl6_6x

**
**l=7**
gen ksikl7_1x= ksikitest if (tk==7 & constx==1)
egen ksikl7_1=min( ksikl7_1x), by(ukno)

gen ksikl7_2x= ksikitest if (tk==7 & constx==2)
egen ksikl7_2=min( ksikl7_2x), by(ukno)

gen ksikl7_3x= ksikitest if (tk==7 & constx==3)
egen ksikl7_3=min( ksikl7_3x), by(ukno)

gen ksikl7_4x= ksikitest if (tk==7 & constx==4)
egen ksikl7_4=min( ksikl7_4x), by(ukno)

gen ksikl7_5x= ksikitest if (tk==7 & constx==5)
egen ksikl7_5=min( ksikl7_5x), by(ukno)

gen ksikl7_6x= ksikitest if (tk==7 & constx==6)
egen ksikl7_6=min( ksikl7_6x), by(ukno)

drop ksikl7_1x ksikl7_2x ksikl7_3x ksikl7_4x ksikl7_5x ksikl7_6x

**
**l=8**
gen ksikl8_1x= ksikitest if (tk==8 & constx==1)
egen ksikl8_1=min( ksikl8_1x), by(ukno)

gen ksikl8_2x= ksikitest if (tk==8 & constx==2)
egen ksikl8_2=min( ksikl8_2x), by(ukno)

gen ksikl8_3x= ksikitest if (tk==8 & constx==3)
egen ksikl8_3=min( ksikl8_3x), by(ukno)

gen ksikl8_4x= ksikitest if (tk==8 & constx==4)
egen ksikl8_4=min( ksikl8_4x), by(ukno)

gen ksikl8_5x= ksikitest if (tk==8 & constx==5)
egen ksikl8_5=min( ksikl8_5x), by(ukno)

gen ksikl8_6x= ksikitest if (tk==8 & constx==6)
egen ksikl8_6=min( ksikl8_6x), by(ukno)

drop ksikl8_1x ksikl8_2x ksikl8_3x ksikl8_4x ksikl8_5x ksikl8_6x

**
**l=9**
gen ksikl9_1x= ksikitest if (tk==9 & constx==1)
egen ksikl9_1=min( ksikl9_1x), by(ukno)

gen ksikl9_2x= ksikitest if (tk==9 & constx==2)
egen ksikl9_2=min( ksikl9_2x), by(ukno)

gen ksikl9_3x= ksikitest if (tk==9 & constx==3)
egen ksikl9_3=min( ksikl9_3x), by(ukno)

gen ksikl9_4x= ksikitest if (tk==9 & constx==4)
egen ksikl9_4=min( ksikl9_4x), by(ukno)
```

```
gen ksikl9_5x= ksikitest if (tk==9 & constx==5)
egen ksikl9_5=min( ksikl9_5x), by(ukno)

gen ksikl9_6x= ksikitest if (tk==9 & constx==6)
egen ksikl9_6=min( ksikl9_6x), by(ukno)

drop ksikl9_1x ksikl9_2x ksikl9_3x ksikl9_4x ksikl9_5x ksikl9_6x

**
**l=10**
gen ksikl10_1x= ksikitest if (tk==10 & constx==1)
egen ksikl10_1=min( ksikl10_1x), by(ukno)

gen ksikl10_2x= ksikitest if (tk==10 & constx==2)
egen ksikl10_2=min( ksikl10_2x), by(ukno)

gen ksikl10_3x= ksikitest if (tk==10 & constx==3)
egen ksikl10_3=min( ksikl10_3x), by(ukno)

gen ksikl10_4x= ksikitest if (tk==10 & constx==4)
egen ksikl10_4=min( ksikl10_4x), by(ukno)

gen ksikl10_5x= ksikitest if (tk==10 & constx==5)
egen ksikl10_5=min( ksikl10_5x), by(ukno)

gen ksikl10_6x= ksikitest if (tk==10 & constx==6)
egen ksikl10_6=min( ksikl10_6x), by(ukno)

drop ksikl10_1x ksikl10_2x ksikl10_3x ksikl10_4x ksikl10_5x ksikl10_6x

**
**l=11**
gen ksikl11_1x= ksikitest if (tk==11 & constx==1)
egen ksikl11_1=min( ksikl11_1x), by(ukno)

gen ksikl11_2x= ksikitest if (tk==11 & constx==2)
egen ksikl11_2=min( ksikl11_2x), by(ukno)

gen ksikl11_3x= ksikitest if (tk==11 & constx==3)
egen ksikl11_3=min( ksikl11_3x), by(ukno)

gen ksikl11_4x= ksikitest if (tk==11 & constx==4)
egen ksikl11_4=min( ksikl11_4x), by(ukno)

gen ksikl11_5x= ksikitest if (tk==11 & constx==5)
egen ksikl11_5=min( ksikl11_5x), by(ukno)

gen ksikl11_6x= ksikitest if (tk==11 & constx==6)
egen ksikl11_6=min( ksikl11_6x), by(ukno)

drop ksikl11_1x ksikl11_2x ksikl11_3x ksikl11_4x ksikl11_5x ksikl11_6x

**
**l=12**
gen ksikl12_1x= ksikitest if (tk==12 & constx==1)
egen ksikl12_1=min( ksikl12_1x), by(ukno)

gen ksikl12_2x= ksikitest if (tk==12 & constx==2)
egen ksikl12_2=min( ksikl12_2x), by(ukno)

gen ksikl12_3x= ksikitest if (tk==12 & constx==3)
egen ksikl12_3=min( ksikl12_3x), by(ukno)
```

```
gen ksikl12_4x= ksikitest if (tk==12 & constx==4)
egen ksikl12_4=min( ksikl12_4x), by(ukno)

gen ksikl12_5x= ksikitest if (tk==12 & constx==5)
egen ksikl12_5=min( ksikl12_5x), by(ukno)

gen ksikl12_6x= ksikitest if (tk==12 & constx==6)
egen ksikl12_6=min( ksikl12_6x), by(ukno)

drop ksikl12_1x ksikl12_2x ksikl12_3x ksikl12_4x ksikl12_5x ksikl12_6x

**
**l=13**
gen ksikl13_1x= ksikitest if (tk==13 & constx==1)
egen ksikl13_1=min( ksikl13_1x), by(ukno)

gen ksikl13_2x= ksikitest if (tk==13 & constx==2)
egen ksikl13_2=min( ksikl13_2x), by(ukno)

gen ksikl13_3x= ksikitest if (tk==13 & constx==3)
egen ksikl13_3=min( ksikl13_3x), by(ukno)

gen ksikl13_4x= ksikitest if (tk==13 & constx==4)
egen ksikl13_4=min( ksikl13_4x), by(ukno)

gen ksikl13_5x= ksikitest if (tk==13 & constx==5)
egen ksikl13_5=min( ksikl13_5x), by(ukno)

gen ksikl13_6x= ksikitest if (tk==13 & constx==6)
egen ksikl13_6=min( ksikl13_6x), by(ukno)

drop ksikl13_1x ksikl13_2x ksikl13_3x ksikl13_4x ksikl13_5x ksikl13_6x

**
**l=14**
gen ksikl14_1x= ksikitest if (tk==14 & constx==1)
egen ksikl14_1=min( ksikl14_1x), by(ukno)

gen ksikl14_2x= ksikitest if (tk==14 & constx==2)
egen ksikl14_2=min( ksikl14_2x), by(ukno)

gen ksikl14_3x= ksikitest if (tk==14 & constx==3)
egen ksikl14_3=min( ksikl14_3x), by(ukno)

gen ksikl14_4x= ksikitest if (tk==14 & constx==4)
egen ksikl14_4=min( ksikl14_4x), by(ukno)

gen ksikl14_5x= ksikitest if (tk==14 & constx==5)
egen ksikl14_5=min( ksikl14_5x), by(ukno)

gen ksikl14_6x= ksikitest if (tk==14 & constx==6)
egen ksikl14_6=min( ksikl14_6x), by(ukno)

drop ksikl14_1x ksikl14_2x ksikl14_3x ksikl14_4x ksikl14_5x ksikl14_6x

**
**l=15**
gen ksikl15_1x= ksikitest if (tk==15 & constx==1)
egen ksikl15_1=min( ksikl15_1x), by(ukno)

gen ksikl15_2x= ksikitest if (tk==15 & constx==2)
egen ksikl15_2=min( ksikl15_2x), by(ukno)

gen ksikl15_3x= ksikitest if (tk==15 & constx==3)
```

```
egen ksikl15_3=min( ksikl15_3x), by(ukno)

gen ksikl15_4x= ksikitest if (tk==15 & constx==4)
egen ksikl15_4=min( ksikl15_4x), by(ukno)

gen ksikl15_5x= ksikitest if (tk==15 & constx==5)
egen ksikl15_5=min( ksikl15_5x), by(ukno)

gen ksikl15_6x= ksikitest if (tk==15 & constx==6)
egen ksikl15_6=min( ksikl15_6x), by(ukno)

drop ksikl15_1x ksikl15_2x ksikl15_3x ksikl15_4x ksikl15_5x ksikl15_6x

**
**l=16**
gen ksikl16_1x= ksikitest if (tk==16 & constx==1)
egen ksikl16_1=min( ksikl16_1x), by(ukno)

gen ksikl16_2x= ksikitest if (tk==16 & constx==2)
egen ksikl16_2=min( ksikl16_2x), by(ukno)

gen ksikl16_3x= ksikitest if (tk==16 & constx==3)
egen ksikl16_3=min( ksikl16_3x), by(ukno)

gen ksikl16_4x= ksikitest if (tk==16 & constx==4)
egen ksikl16_4=min( ksikl16_4x), by(ukno)

gen ksikl16_5x= ksikitest if (tk==16 & constx==5)
egen ksikl16_5=min( ksikl16_5x), by(ukno)

gen ksikl16_6x= ksikitest if (tk==16 & constx==6)
egen ksikl16_6=min( ksikl16_6x), by(ukno)

drop ksikl16_1x ksikl16_2x ksikl16_3x ksikl16_4x ksikl16_5x ksikl16_6x

**
**l=17**
gen ksikl17_1x= ksikitest if (tk==17 & constx==1)
egen ksikl17_1=min( ksikl17_1x), by(ukno)

gen ksikl17_2x= ksikitest if (tk==17 & constx==2)
egen ksikl17_2=min( ksikl17_2x), by(ukno)

gen ksikl17_3x= ksikitest if (tk==17 & constx==3)
egen ksikl17_3=min( ksikl17_3x), by(ukno)

gen ksikl17_4x= ksikitest if (tk==17 & constx==4)
egen ksikl17_4=min( ksikl17_4x), by(ukno)

gen ksikl17_5x= ksikitest if (tk==17 & constx==5)
egen ksikl17_5=min( ksikl17_5x), by(ukno)

gen ksikl17_6x= ksikitest if (tk==17 & constx==6)
egen ksikl17_6=min( ksikl17_6x), by(ukno)

drop ksikl17_1x ksikl17_2x ksikl17_3x ksikl17_4x ksikl17_5x ksikl17_6x

**
**l=18**
gen ksikl18_1x= ksikitest if (tk==18 & constx==1)
egen ksikl18_1=min( ksikl18_1x), by(ukno)

gen ksikl18_2x= ksikitest if (tk==18 & constx==2)
egen ksikl18_2=min( ksikl18_2x), by(ukno)
```

```
gen ksikl18_3x= ksikitest if (tk==18 & constx==3)
egen ksikl18_3=min( ksikl18_3x), by(ukno)

gen ksikl18_4x= ksikitest if (tk==18 & constx==4)
egen ksikl18_4=min( ksikl18_4x), by(ukno)

gen ksikl18_5x= ksikitest if (tk==18 & constx==5)
egen ksikl18_5=min( ksikl18_5x), by(ukno)

gen ksikl18_6x= ksikitest if (tk==18 & constx==6)
egen ksikl18_6=min( ksikl18_6x), by(ukno)

drop ksikl18_1x ksikl18_2x ksikl18_3x ksikl18_4x ksikl18_5x ksikl18_6x

**
**l=19**
gen ksikl19_1x= ksikitest if (tk==19 & constx==1)
egen ksikl19_1=min( ksikl19_1x), by(ukno)

gen ksikl19_2x= ksikitest if (tk==19 & constx==2)
egen ksikl19_2=min( ksikl19_2x), by(ukno)

gen ksikl19_3x= ksikitest if (tk==19 & constx==3)
egen ksikl19_3=min( ksikl19_3x), by(ukno)

gen ksikl19_4x= ksikitest if (tk==19 & constx==4)
egen ksikl19_4=min( ksikl19_4x), by(ukno)

gen ksikl19_5x= ksikitest if (tk==19 & constx==5)
egen ksikl19_5=min( ksikl19_5x), by(ukno)

gen ksikl19_6x= ksikitest if (tk==19 & constx==6)
egen ksikl19_6=min( ksikl19_6x), by(ukno)

drop ksikl19_1x ksikl19_2x ksikl19_3x ksikl19_4x ksikl19_5x ksikl19_6x

**
sort ukno tk constx

gen xl1_1= ksikitest* ksikl1_1
gen xl1_2= ksikitest* ksikl1_2
gen xl1_3= ksikitest* ksikl1_3
gen xl1_4= ksikitest* ksikl1_4
gen xl1_5= ksikitest* ksikl1_5
gen xl1_6= ksikitest* ksikl1_6

gen xl2_1= ksikitest* ksikl2_1
gen xl2_2= ksikitest* ksikl2_2
gen xl2_3= ksikitest* ksikl2_3
gen xl2_4= ksikitest* ksikl2_4
gen xl2_5= ksikitest* ksikl2_5
gen xl2_6= ksikitest* ksikl2_6

gen xl3_1= ksikitest* ksikl3_1
gen xl3_2= ksikitest* ksikl3_2
gen xl3_3= ksikitest* ksikl3_3
gen xl3_4= ksikitest* ksikl3_4
gen xl3_5= ksikitest* ksikl3_5
gen xl3_6= ksikitest* ksikl3_6

gen xl4_1= ksikitest* ksikl4_1
gen xl4_2= ksikitest* ksikl4_2
gen xl4_3= ksikitest* ksikl4_3
```

```
gen xl4_4= ksikitest* ksikl4_4
gen xl4_5= ksikitest* ksikl4_5
gen xl4_6= ksikitest* ksikl4_6

gen xl5_1= ksikitest* ksikl5_1
gen xl5_2= ksikitest* ksikl5_2
gen xl5_3= ksikitest* ksikl5_3
gen xl5_4= ksikitest* ksikl5_4
gen xl5_5= ksikitest* ksikl5_5
gen xl5_6= ksikitest* ksikl5_6

gen xl6_1= ksikitest* ksikl6_1
gen xl6_2= ksikitest* ksikl6_2
gen xl6_3= ksikitest* ksikl6_3
gen xl6_4= ksikitest* ksikl6_4
gen xl6_5= ksikitest* ksikl6_5
gen xl6_6= ksikitest* ksikl6_6

gen xl7_1= ksikitest* ksikl7_1
gen xl7_2= ksikitest* ksikl7_2
gen xl7_3= ksikitest* ksikl7_3
gen xl7_4= ksikitest* ksikl7_4
gen xl7_5= ksikitest* ksikl7_5
gen xl7_6= ksikitest* ksikl7_6

gen xl8_1= ksikitest* ksikl8_1
gen xl8_2= ksikitest* ksikl8_2
gen xl8_3= ksikitest* ksikl8_3
gen xl8_4= ksikitest* ksikl8_4
gen xl8_5= ksikitest* ksikl8_5
gen xl8_6= ksikitest* ksikl8_6

gen xl9_1= ksikitest* ksikl9_1
gen xl9_2= ksikitest* ksikl9_2
gen xl9_3= ksikitest* ksikl9_3
gen xl9_4= ksikitest* ksikl9_4
gen xl9_5= ksikitest* ksikl9_5
gen xl9_6= ksikitest* ksikl9_6

gen xl10_1= ksikitest* ksikl10_1
gen xl10_2= ksikitest* ksikl10_2
gen xl10_3= ksikitest* ksikl10_3
gen xl10_4= ksikitest* ksikl10_4
gen xl10_5= ksikitest* ksikl10_5
gen xl10_6= ksikitest* ksikl10_6

gen xl11_1= ksikitest* ksikl11_1
gen xl11_2= ksikitest* ksikl11_2
gen xl11_3= ksikitest* ksikl11_3
gen xl11_4= ksikitest* ksikl11_4
gen xl11_5= ksikitest* ksikl11_5
gen xl11_6= ksikitest* ksikl11_6

gen xl12_1= ksikitest* ksikl12_1
gen xl12_2= ksikitest* ksikl12_2
gen xl12_3= ksikitest* ksikl12_3
gen xl12_4= ksikitest* ksikl12_4
gen xl12_5= ksikitest* ksikl12_5
gen xl12_6= ksikitest* ksikl12_6

gen xl13_1= ksikitest* ksikl13_1
gen xl13_2= ksikitest* ksikl13_2
gen xl13_3= ksikitest* ksikl13_3
gen xl13_4= ksikitest* ksikl13_4
```

```
gen xl13_5= ksikitest* ksikl13_5
gen xl13_6= ksikitest* ksikl13_6

gen xl14_1= ksikitest* ksikl14_1
gen xl14_2= ksikitest* ksikl14_2
gen xl14_3= ksikitest* ksikl14_3
gen xl14_4= ksikitest* ksikl14_4
gen xl14_5= ksikitest* ksikl14_5
gen xl14_6= ksikitest* ksikl14_6

gen xl15_1= ksikitest* ksikl15_1
gen xl15_2= ksikitest* ksikl15_2
gen xl15_3= ksikitest* ksikl15_3
gen xl15_4= ksikitest* ksikl15_4
gen xl15_5= ksikitest* ksikl15_5
gen xl15_6= ksikitest* ksikl15_6

gen xl16_1= ksikitest* ksikl16_1
gen xl16_2= ksikitest* ksikl16_2
gen xl16_3= ksikitest* ksikl16_3
gen xl16_4= ksikitest* ksikl16_4
gen xl16_5= ksikitest* ksikl16_5
gen xl16_6= ksikitest* ksikl16_6

gen xl17_1= ksikitest* ksikl17_1
gen xl17_2= ksikitest* ksikl17_2
gen xl17_3= ksikitest* ksikl17_3
gen xl17_4= ksikitest* ksikl17_4
gen xl17_5= ksikitest* ksikl17_5
gen xl17_6= ksikitest* ksikl17_6

gen xl18_1= ksikitest* ksikl18_1
gen xl18_2= ksikitest* ksikl18_2
gen xl18_3= ksikitest* ksikl18_3
gen xl18_4= ksikitest* ksikl18_4
gen xl18_5= ksikitest* ksikl18_5
gen xl18_6= ksikitest* ksikl18_6

gen xl19_1= ksikitest* ksikl19_1
gen xl19_2= ksikitest* ksikl19_2
gen xl19_3= ksikitest* ksikl19_3
gen xl19_4= ksikitest* ksikl19_4
gen xl19_5= ksikitest* ksikl19_5
gen xl19_6= ksikitest* ksikl19_6

**
drop ksikl*

**

egen sxl1_1=sum(xl1_1), by(tk constx)
egen sxl1_2=sum(xl1_2), by(tk constx)
egen sxl1_3=sum(xl1_3), by(tk constx)
egen sxl1_4=sum(xl1_4), by(tk constx)
egen sxl1_5=sum(xl1_5), by(tk constx)
egen sxl1_6=sum(xl1_6), by(tk constx)

egen sxl2_1=sum(xl2_1), by(tk constx)
egen sxl2_2=sum(xl2_2), by(tk constx)
egen sxl2_3=sum(xl2_3), by(tk constx)
egen sxl2_4=sum(xl2_4), by(tk constx)
egen sxl2_5=sum(xl2_5), by(tk constx)
egen sxl2_6=sum(xl2_6), by(tk constx)
```

```
egen sxl3_1=sum(xl3_1), by(tk constx)
egen sxl3_2=sum(xl3_2), by(tk constx)
egen sxl3_3=sum(xl3_3), by(tk constx)
egen sxl3_4=sum(xl3_4), by(tk constx)
egen sxl3_5=sum(xl3_5), by(tk constx)
egen sxl3_6=sum(xl3_6), by(tk constx)

egen sxl4_1=sum(xl4_1), by(tk constx)
egen sxl4_2=sum(xl4_2), by(tk constx)
egen sxl4_3=sum(xl4_3), by(tk constx)
egen sxl4_4=sum(xl4_4), by(tk constx)
egen sxl4_5=sum(xl4_5), by(tk constx)
egen sxl4_6=sum(xl4_6), by(tk constx)

egen sxl5_1=sum(xl5_1), by(tk constx)
egen sxl5_2=sum(xl5_2), by(tk constx)
egen sxl5_3=sum(xl5_3), by(tk constx)
egen sxl5_4=sum(xl5_4), by(tk constx)
egen sxl5_5=sum(xl5_5), by(tk constx)
egen sxl5_6=sum(xl5_6), by(tk constx)

egen sxl6_1=sum(xl6_1), by(tk constx)
egen sxl6_2=sum(xl6_2), by(tk constx)
egen sxl6_3=sum(xl6_3), by(tk constx)
egen sxl6_4=sum(xl6_4), by(tk constx)
egen sxl6_5=sum(xl6_5), by(tk constx)
egen sxl6_6=sum(xl6_6), by(tk constx)

egen sxl7_1=sum(xl7_1), by(tk constx)
egen sxl7_2=sum(xl7_2), by(tk constx)
egen sxl7_3=sum(xl7_3), by(tk constx)
egen sxl7_4=sum(xl7_4), by(tk constx)
egen sxl7_5=sum(xl7_5), by(tk constx)
egen sxl7_6=sum(xl7_6), by(tk constx)

egen sxl8_1=sum(xl8_1), by(tk constx)
egen sxl8_2=sum(xl8_2), by(tk constx)
egen sxl8_3=sum(xl8_3), by(tk constx)
egen sxl8_4=sum(xl8_4), by(tk constx)
egen sxl8_5=sum(xl8_5), by(tk constx)
egen sxl8_6=sum(xl8_6), by(tk constx)

egen sxl9_1=sum(xl9_1), by(tk constx)
egen sxl9_2=sum(xl9_2), by(tk constx)
egen sxl9_3=sum(xl9_3), by(tk constx)
egen sxl9_4=sum(xl9_4), by(tk constx)
egen sxl9_5=sum(xl9_5), by(tk constx)
egen sxl9_6=sum(xl9_6), by(tk constx)

egen sxl10_1=sum(xl10_1), by(tk constx)
egen sxl10_2=sum(xl10_2), by(tk constx)
egen sxl10_3=sum(xl10_3), by(tk constx)
egen sxl10_4=sum(xl10_4), by(tk constx)
egen sxl10_5=sum(xl10_5), by(tk constx)
egen sxl10_6=sum(xl10_6), by(tk constx)

egen sxl11_1=sum(xl11_1), by(tk constx)
egen sxl11_2=sum(xl11_2), by(tk constx)
egen sxl11_3=sum(xl11_3), by(tk constx)
egen sxl11_4=sum(xl11_4), by(tk constx)
egen sxl11_5=sum(xl11_5), by(tk constx)
egen sxl11_6=sum(xl11_6), by(tk constx)

egen sxl12_1=sum(xl12_1), by(tk constx)
```

```
egen sxl12_2=sum(xl12_2), by(tk constx)
egen sxl12_3=sum(xl12_3), by(tk constx)
egen sxl12_4=sum(xl12_4), by(tk constx)
egen sxl12_5=sum(xl12_5), by(tk constx)
egen sxl12_6=sum(xl12_6), by(tk constx)

egen sxl13_1=sum(xl13_1), by(tk constx)
egen sxl13_2=sum(xl13_2), by(tk constx)
egen sxl13_3=sum(xl13_3), by(tk constx)
egen sxl13_4=sum(xl13_4), by(tk constx)
egen sxl13_5=sum(xl13_5), by(tk constx)
egen sxl13_6=sum(xl13_6), by(tk constx)

egen sxl14_1=sum(xl14_1), by(tk constx)
egen sxl14_2=sum(xl14_2), by(tk constx)
egen sxl14_3=sum(xl14_3), by(tk constx)
egen sxl14_4=sum(xl14_4), by(tk constx)
egen sxl14_5=sum(xl14_5), by(tk constx)
egen sxl14_6=sum(xl14_6), by(tk constx)

egen sxl15_1=sum(xl15_1), by(tk constx)
egen sxl15_2=sum(xl15_2), by(tk constx)
egen sxl15_3=sum(xl15_3), by(tk constx)
egen sxl15_4=sum(xl15_4), by(tk constx)
egen sxl15_5=sum(xl15_5), by(tk constx)
egen sxl15_6=sum(xl15_6), by(tk constx)

egen sxl16_1=sum(xl16_1), by(tk constx)
egen sxl16_2=sum(xl16_2), by(tk constx)
egen sxl16_3=sum(xl16_3), by(tk constx)
egen sxl16_4=sum(xl16_4), by(tk constx)
egen sxl16_5=sum(xl16_5), by(tk constx)
egen sxl16_6=sum(xl16_6), by(tk constx)

egen sxl17_1=sum(xl17_1), by(tk constx)
egen sxl17_2=sum(xl17_2), by(tk constx)
egen sxl17_3=sum(xl17_3), by(tk constx)
egen sxl17_4=sum(xl17_4), by(tk constx)
egen sxl17_5=sum(xl17_5), by(tk constx)
egen sxl17_6=sum(xl17_6), by(tk constx)

egen sxl18_1=sum(xl18_1), by(tk constx)
egen sxl18_2=sum(xl18_2), by(tk constx)
egen sxl18_3=sum(xl18_3), by(tk constx)
egen sxl18_4=sum(xl18_4), by(tk constx)
egen sxl18_5=sum(xl18_5), by(tk constx)
egen sxl18_6=sum(xl18_6), by(tk constx)

egen sxl19_1=sum(xl19_1), by(tk constx)
egen sxl19_2=sum(xl19_2), by(tk constx)
egen sxl19_3=sum(xl19_3), by(tk constx)
egen sxl19_4=sum(xl19_4), by(tk constx)
egen sxl19_5=sum(xl19_5), by(tk constx)
egen sxl19_6=sum(xl19_6), by(tk constx)

**

egen sZi0_Zi0p=sum(Zi0_Zi0p), by(tk constx)
egen sZi1_Zi1p=sum(Zi1_Zi1p), by(tk constx)
egen sZi2_Zi2p=sum(Zi2_Zi2p), by(tk constx)
egen sZi3_Zi3p=sum(Zi3_Zi3p), by(tk constx)
egen sZi4_Zi4p=sum(Zi4_Zi4p), by(tk constx)
egen sZi5_Zi5p=sum(Zi5_Zi5p), by(tk constx)
```

```
keep if ukno=="00011D"

sort tk constx

mkmat  sZi0_Zi0p sZi1_Zi1p sZi2_Zi2p sZi3_Zi3p sZi4_Zi4p sZi5_Zi5p if tk==1,
matrix(seAx)

matrix seA=(1/1138)*seAx

svmat seA
matrix list seA

matrix seinvA=syminv(seA)

svmat seinvA
matrix list seinvA

**

mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==1, matrix(Bk111x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==1, matrix(Bk112x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==1, matrix(Bk113x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==1, matrix(Bk114x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==1, matrix(Bk115x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==1, matrix(Bk116x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==1, matrix(Bk117x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==1, matrix(Bk118x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==1, matrix(Bk119x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==1, matrix(Bk1110x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==1, matrix(Bk1111x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==1, matrix(Bk1112x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==1, matrix(Bk1113x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==1, matrix(Bk1114x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==1, matrix(Bk1115x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==1, matrix(Bk1116x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==1, matrix(Bk1117x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==1, matrix(Bk1118x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==1, matrix(Bk1119x)

**
matrix Bk1_lall=(1/1138)*( Bk111x+ Bk112x+ Bk113x+ Bk114x+ Bk115x+ Bk116x+
Bk117x+ Bk118x+ Bk119x+ Bk1110x+ Bk1111x+ Bk1112x+ Bk1113x+ Bk1114x+ Bk1115x+
Bk1116x+ Bk1117x+ Bk1118x+ Bk1119x)

svmat Bk1_lall

matrix list Bk1_lall

**
**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==2, matrix(Bk211x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==2, matrix(Bk212x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==2, matrix(Bk213x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==2, matrix(Bk214x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==2, matrix(Bk215x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==2, matrix(Bk216x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==2, matrix(Bk217x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==2, matrix(Bk218x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==2, matrix(Bk219x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==2, matrix(Bk2110x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==2, matrix(Bk2111x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==2, matrix(Bk2112x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==2, matrix(Bk2113x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==2, matrix(Bk2114x)
```

```
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==2, matrix(Bk2l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==2, matrix(Bk2l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==2, matrix(Bk2l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==2, matrix(Bk2l18x)

mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==2, matrix(Bk2l19x)

**
matrix Bk2_lall=(1/1138)*( Bk2l1x+ Bk2l2x+ Bk2l3x+ Bk2l4x+ Bk2l5x+ Bk2l6x+
Bk2l7x+ Bk2l8x+ Bk2l9x+ Bk2l10x+ Bk2l11x+ Bk2l12x+ Bk2l13x+ Bk2l14x+ Bk2l15x+
Bk2l16x+ Bk2l17x+ Bk2l18x+ Bk2l19x)

svmat Bk2_lall

matrix list Bk2_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==3, matrix(Bk3l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==3, matrix(Bk3l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==3, matrix(Bk3l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==3, matrix(Bk3l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==3, matrix(Bk3l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==3, matrix(Bk3l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==3, matrix(Bk3l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==3, matrix(Bk3l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==3, matrix(Bk3l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==3, matrix(Bk3l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==3, matrix(Bk3l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==3, matrix(Bk3l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==3, matrix(Bk3l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==3, matrix(Bk3l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==3, matrix(Bk3l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==3, matrix(Bk3l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==3, matrix(Bk3l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==3, matrix(Bk3l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==3, matrix(Bk3l19x)

**
matrix Bk3_lall=(1/1138)*( Bk3l1x+ Bk3l2x+ Bk3l3x+ Bk3l4x+ Bk3l5x+ Bk3l6x+
Bk3l7x+ Bk3l8x+ Bk3l9x+ Bk3l10x+ Bk3l11x+ Bk3l12x+ Bk3l13x+ Bk3l14x+ Bk3l15x+
Bk3l16x+ Bk3l17x+ Bk3l18x+ Bk3l19x)

svmat Bk3_lall

matrix list Bk3_lall

**

mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==4, matrix(Bk4l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==4, matrix(Bk4l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==4, matrix(Bk4l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==4, matrix(Bk4l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==4, matrix(Bk4l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==4, matrix(Bk4l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==4, matrix(Bk4l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==4, matrix(Bk4l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==4, matrix(Bk4l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==4, matrix(Bk4l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==4, matrix(Bk4l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==4, matrix(Bk4l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==4, matrix(Bk4l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==4, matrix(Bk4l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==4, matrix(Bk4l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==4, matrix(Bk4l16x)
```

```
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==4, matrix(Bk4l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==4, matrix(Bk4l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==4, matrix(Bk4l19x)

**
matrix Bk4_lall=(1/1138)*( Bk4l1x+ Bk4l2x+ Bk4l3x+ Bk4l4x+ Bk4l5x+ Bk4l6x+
Bk4l7x+ Bk4l8x+ Bk4l9x+ Bk4l10x+ Bk4l11x+ Bk4l12x+ Bk4l13x+ Bk4l14x+ Bk4l15x+
Bk4l16x+ Bk4l17x+ Bk4l18x+ Bk4l19x)

svmat Bk4_lall

matrix list Bk4_lall

**

mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==5, matrix(Bk5l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==5, matrix(Bk5l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==5, matrix(Bk5l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==5, matrix(Bk5l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==5, matrix(Bk5l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==5, matrix(Bk5l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==5, matrix(Bk5l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==5, matrix(Bk5l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==5, matrix(Bk5l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==5, matrix(Bk5l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==5, matrix(Bk5l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==5, matrix(Bk5l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==5, matrix(Bk5l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==5, matrix(Bk5l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==5, matrix(Bk5l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==5, matrix(Bk5l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==5, matrix(Bk5l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==5, matrix(Bk5l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==5, matrix(Bk5l19x)

**
matrix Bk5_lall=(1/1138)*( Bk5l1x+ Bk5l2x+ Bk5l3x+ Bk5l4x+ Bk5l5x+ Bk5l6x+
Bk5l7x+ Bk5l8x+ Bk5l9x+ Bk5l10x+ Bk5l11x+ Bk5l12x+ Bk5l13x+ Bk5l14x+ Bk5l15x+
Bk5l16x+ Bk5l17x+ Bk5l18x+ Bk5l19x)

svmat Bk5_lall

matrix list Bk5_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==6, matrix(Bk6l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==6, matrix(Bk6l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==6, matrix(Bk6l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==6, matrix(Bk6l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==6, matrix(Bk6l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==6, matrix(Bk6l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==6, matrix(Bk6l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==6, matrix(Bk6l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==6, matrix(Bk6l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==6, matrix(Bk6l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==6, matrix(Bk6l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==6, matrix(Bk6l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==6, matrix(Bk6l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==6, matrix(Bk6l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==6, matrix(Bk6l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==6, matrix(Bk6l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==6, matrix(Bk6l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==6, matrix(Bk6l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==6, matrix(Bk6l19x)
```

```
**
matrix Bk6_lall=(1/1138)*( Bk6l1x+ Bk6l2x+ Bk6l3x+ Bk6l4x+ Bk6l5x+ Bk6l6x+
Bk6l7x+ Bk6l8x+ Bk6l9x+ Bk6l10x+ Bk6l11x+ Bk6l12x+ Bk6l13x+ Bk6l14x+ Bk6l15x+
Bk6l16x+ Bk6l17x+ Bk6l18x+ Bk6l19x)

svmat Bk6_lall

matrix list Bk6_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==7, matrix(Bk7l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==7, matrix(Bk7l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==7, matrix(Bk7l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==7, matrix(Bk7l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==7, matrix(Bk7l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==7, matrix(Bk7l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==7, matrix(Bk7l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==7, matrix(Bk7l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==7, matrix(Bk7l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==7, matrix(Bk7l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==7, matrix(Bk7l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==7, matrix(Bk7l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==7, matrix(Bk7l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==7, matrix(Bk7l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==7, matrix(Bk7l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==7, matrix(Bk7l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==7, matrix(Bk7l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==7, matrix(Bk7l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==7, matrix(Bk7l19x)

**
matrix Bk7_lall=(1/1138)*( Bk7l1x+ Bk7l2x+ Bk7l3x+ Bk7l4x+ Bk7l5x+ Bk7l6x+
Bk7l7x+ Bk7l8x+ Bk7l9x+ Bk7l10x+ Bk7l11x+ Bk7l12x+ Bk7l13x+ Bk7l14x+ Bk7l15x+
Bk7l16x+ Bk7l17x+ Bk7l18x+ Bk7l19x)

svmat Bk7_lall

matrix list Bk7_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==8, matrix(Bk8l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==8, matrix(Bk8l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==8, matrix(Bk8l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==8, matrix(Bk8l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==8, matrix(Bk8l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==8, matrix(Bk8l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==8, matrix(Bk8l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==8, matrix(Bk8l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==8, matrix(Bk8l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==8, matrix(Bk8l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==8, matrix(Bk8l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==8, matrix(Bk8l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==8, matrix(Bk8l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==8, matrix(Bk8l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==8, matrix(Bk8l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==8, matrix(Bk8l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==8, matrix(Bk8l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==8, matrix(Bk8l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==8, matrix(Bk8l19x)

**
```

```
matrix Bk8_lall=(1/1138)*( Bk8l1x+ Bk8l2x+ Bk8l3x+ Bk8l4x+ Bk8l5x+ Bk8l6x+
Bk8l7x+ Bk8l8x+ Bk8l9x+ Bk8l10x+ Bk8l11x+ Bk8l12x+ Bk8l13x+ Bk8l14x+ Bk8l15x+
Bk8l16x+ Bk8l17x+ Bk8l18x+ Bk8l19x)

svmat Bk8_lall

matrix list Bk8_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==9, matrix(Bk9l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==9, matrix(Bk9l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==9, matrix(Bk9l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==9, matrix(Bk9l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==9, matrix(Bk9l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==9, matrix(Bk9l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==9, matrix(Bk9l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==9, matrix(Bk9l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==9, matrix(Bk9l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==9, matrix(Bk9l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==9, matrix(Bk9l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==9, matrix(Bk9l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==9, matrix(Bk9l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==9, matrix(Bk9l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==9, matrix(Bk9l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==9, matrix(Bk9l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==9, matrix(Bk9l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==9, matrix(Bk9l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==9, matrix(Bk9l19x)

**
matrix Bk9_lall=(1/1138)*( Bk9l1x+ Bk9l2x+ Bk9l3x+ Bk9l4x+ Bk9l5x+ Bk9l6x+
Bk9l7x+ Bk9l8x+ Bk9l9x+ Bk9l10x+ Bk9l11x+ Bk9l12x+ Bk9l13x+ Bk9l14x+ Bk9l15x+
Bk9l16x+ Bk9l17x+ Bk9l18x+ Bk9l19x)

svmat Bk9_lall

matrix list Bk9_lall

**

mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==10, matrix(Bk10l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==10, matrix(Bk10l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==10, matrix(Bk10l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==10, matrix(Bk10l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==10, matrix(Bk10l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==10, matrix(Bk10l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==10, matrix(Bk10l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==10, matrix(Bk10l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==10, matrix(Bk10l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==10,
matrix(Bk10l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==10,
matrix(Bk10l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==10,
matrix(Bk10l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==10,
matrix(Bk10l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==10,
matrix(Bk10l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==10,
matrix(Bk10l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==10,
matrix(Bk10l16x)
```

```
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==10,
matrix(Bk10l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==10,
matrix(Bk10l18x)

mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==10,
matrix(Bk10l19x)

**
matrix Bk10_lall=(1/1138)*( Bk10l1x+ Bk10l2x+ Bk10l3x+ Bk10l4x+ Bk10l5x+
Bk10l6x+ Bk10l7x+ Bk10l8x+ Bk10l9x+ Bk10l10x+ Bk10l11x+ Bk10l12x+ Bk10l13x+
Bk10l14x+ Bk10l15x+ Bk10l16x+ Bk10l17x+ Bk10l18x+ Bk10l19x)

svmat Bk10_lall

matrix list Bk10_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==11, matrix(Bk11l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==11, matrix(Bk11l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==11, matrix(Bk11l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==11, matrix(Bk11l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==11, matrix(Bk11l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==11, matrix(Bk11l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==11, matrix(Bk11l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==11, matrix(Bk11l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==11, matrix(Bk11l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==11,
matrix(Bk11l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==11,
matrix(Bk11l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==11,
matrix(Bk11l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==11,
matrix(Bk11l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==11,
matrix(Bk11l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==11,
matrix(Bk11l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==11,
matrix(Bk11l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==11,
matrix(Bk11l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==11,
matrix(Bk11l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==11,
matrix(Bk11l19x)

**
matrix Bk11_lall=(1/1138)*( Bk11l1x+ Bk11l2x+ Bk11l3x+ Bk11l4x+ Bk11l5x+
Bk11l6x+ Bk11l7x+ Bk11l8x+ Bk11l9x+ Bk11l10x+ Bk11l11x+ Bk11l12x+ Bk11l13x+
Bk11l14x+ Bk11l15x+ Bk11l16x+ Bk11l17x+ Bk11l18x+ Bk11l19x)

svmat Bk11_lall

matrix list Bk11_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==12, matrix(Bk12l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==12, matrix(Bk12l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==12, matrix(Bk12l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==12, matrix(Bk12l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==12, matrix(Bk12l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==12, matrix(Bk12l6x)
```

```
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==12, matrix(Bk12l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==12, matrix(Bk12l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==12, matrix(Bk12l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==12,
matrix(Bk12l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==12,
matrix(Bk12l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==12,
matrix(Bk12l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==12,
matrix(Bk12l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==12,
matrix(Bk12l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==12,
matrix(Bk12l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==12,
matrix(Bk12l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==12,
matrix(Bk12l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==12,
matrix(Bk12l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==12,
matrix(Bk12l19x)

**
matrix Bk12_lall=(1/1138)*( Bk12l1x+ Bk12l2x+ Bk12l3x+ Bk12l4x+ Bk12l5x+
Bk12l6x+ Bk12l7x+ Bk12l8x+ Bk12l9x+ Bk12l10x+ Bk12l11x+ Bk12l12x+ Bk12l13x+
Bk12l14x+ Bk12l15x+ Bk12l16x+ Bk12l17x+ Bk12l18x+ Bk12l19x)

svmat Bk12_lall

matrix list Bk12_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==13, matrix(Bk13l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==13, matrix(Bk13l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==13, matrix(Bk13l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==13, matrix(Bk13l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==13, matrix(Bk13l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==13, matrix(Bk13l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==13, matrix(Bk13l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==13, matrix(Bk13l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==13, matrix(Bk13l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==13,
matrix(Bk13l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==13,
matrix(Bk13l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==13,
matrix(Bk13l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==13,
matrix(Bk13l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==13,
matrix(Bk13l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==13,
matrix(Bk13l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==13,
matrix(Bk13l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==13,
matrix(Bk13l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==13,
matrix(Bk13l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==13,
matrix(Bk13l19x)
```

```
**
matrix Bk13_lall=(1/1138)*( Bk13l1x+ Bk13l2x+ Bk13l3x+ Bk13l4x+ Bk13l5x+
Bk13l6x+ Bk13l7x+ Bk13l8x+ Bk13l9x+ Bk13l10x+ Bk13l11x+ Bk13l12x+ Bk13l13x+
Bk13l14x+ Bk13l15x+ Bk13l16x+ Bk13l17x+ Bk13l18x+ Bk13l19x)

svmat Bk13_lall

matrix list Bk13_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==14, matrix(Bk14l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==14, matrix(Bk14l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==14, matrix(Bk14l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==14, matrix(Bk14l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==14, matrix(Bk14l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==14, matrix(Bk14l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==14, matrix(Bk14l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==14, matrix(Bk14l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==14, matrix(Bk14l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==14,
matrix(Bk14l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==14,
matrix(Bk14l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==14,
matrix(Bk14l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==14,
matrix(Bk14l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==14,
matrix(Bk14l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==14,
matrix(Bk14l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==14,
matrix(Bk14l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==14,
matrix(Bk14l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==14,
matrix(Bk14l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==14,
matrix(Bk14l19x)

**
matrix Bk14_lall=(1/1138)*( Bk14l1x+ Bk14l2x+ Bk14l3x+ Bk14l4x+ Bk14l5x+
Bk14l6x+ Bk14l7x+ Bk14l8x+ Bk14l9x+ Bk14l10x+ Bk14l11x+ Bk14l12x+ Bk14l13x+
Bk14l14x+ Bk14l15x+ Bk14l16x+ Bk14l17x+ Bk14l18x+ Bk14l19x)

svmat Bk14_lall

matrix list Bk14_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==15, matrix(Bk15l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==15, matrix(Bk15l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==15, matrix(Bk15l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==15, matrix(Bk15l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==15, matrix(Bk15l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==15, matrix(Bk15l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==15, matrix(Bk15l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==15, matrix(Bk15l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==15, matrix(Bk15l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==15,
matrix(Bk15l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==15,
matrix(Bk15l11x)
```

```
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==15,
matrix(Bk15l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==15,
matrix(Bk15l13x)

mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==15,
matrix(Bk15l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==15,
matrix(Bk15l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==15,
matrix(Bk15l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==15,
matrix(Bk15l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==15,
matrix(Bk15l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==15,
matrix(Bk15l19x)

**
matrix Bk15_lall=(1/1138)*( Bk15l1x+ Bk15l2x+ Bk15l3x+ Bk15l4x+ Bk15l5x+
Bk15l6x+ Bk15l7x+ Bk15l8x+ Bk15l9x+ Bk15l10x+ Bk15l11x+ Bk15l12x+ Bk15l13x+
Bk15l14x+ Bk15l15x+ Bk15l16x+ Bk15l17x+ Bk15l18x+ Bk15l19x)

svmat Bk15_lall

matrix list Bk15_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==16, matrix(Bk16l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==16, matrix(Bk16l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==16, matrix(Bk16l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==16, matrix(Bk16l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==16, matrix(Bk16l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==16, matrix(Bk16l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==16, matrix(Bk16l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==16, matrix(Bk16l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==16, matrix(Bk16l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==16,
matrix(Bk16l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==16,
matrix(Bk16l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==16,
matrix(Bk16l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==16,
matrix(Bk16l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==16,
matrix(Bk16l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==16,
matrix(Bk16l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==16,
matrix(Bk16l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==16,
matrix(Bk16l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==16,
matrix(Bk16l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==16,
matrix(Bk16l19x)

**
matrix Bk16_lall=(1/1138)*( Bk16l1x+ Bk16l2x+ Bk16l3x+ Bk16l4x+ Bk16l5x+
Bk16l6x+ Bk16l7x+ Bk16l8x+ Bk16l9x+ Bk16l10x+ Bk16l11x+ Bk16l12x+ Bk16l13x+
Bk16l14x+ Bk16l15x+ Bk16l16x+ Bk16l17x+ Bk16l18x+ Bk16l19x)

svmat Bk16_lall
```

```
matrix list Bk16_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==17, matrix(Bk17l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==17, matrix(Bk17l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==17, matrix(Bk17l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==17, matrix(Bk17l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==17, matrix(Bk17l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==17, matrix(Bk17l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==17, matrix(Bk17l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==17, matrix(Bk17l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==17, matrix(Bk17l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==17,
matrix(Bk17l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==17,
matrix(Bk17l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==17,
matrix(Bk17l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==17,
matrix(Bk17l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==17,
matrix(Bk17l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==17,
matrix(Bk17l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==17,
matrix(Bk17l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==17,
matrix(Bk17l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==17,
matrix(Bk17l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==17,
matrix(Bk17l19x)

**
matrix Bk17_lall=(1/1138)*( Bk17l1x+ Bk17l2x+ Bk17l3x+ Bk17l4x+ Bk17l5x+
Bk17l6x+ Bk17l7x+ Bk17l8x+ Bk17l9x+ Bk17l10x+ Bk17l11x+ Bk17l12x+ Bk17l13x+
Bk17l14x+ Bk17l15x+ Bk17l16x+ Bk17l17x+ Bk17l18x+ Bk17l19x)

svmat Bk17_lall

matrix list Bk17_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==18, matrix(Bk18l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==18, matrix(Bk18l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==18, matrix(Bk18l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==18, matrix(Bk18l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==18, matrix(Bk18l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==18, matrix(Bk18l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==18, matrix(Bk18l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==18, matrix(Bk18l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==18, matrix(Bk18l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==18,
matrix(Bk18l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==18,
matrix(Bk18l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==18,
matrix(Bk18l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==18,
matrix(Bk18l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==18,
matrix(Bk18l14x)
```

```
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==18,
matrix(Bk18l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==18,
matrix(Bk18l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==18,
matrix(Bk18l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==18,
matrix(Bk18l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==18,
matrix(Bk18l19x)

**
matrix Bk18_lall=(1/1138)*( Bk18l1x+ Bk18l2x+ Bk18l3x+ Bk18l4x+ Bk18l5x+
Bk18l6x+ Bk18l7x+ Bk18l8x+ Bk18l9x+ Bk18l10x+ Bk18l11x+ Bk18l12x+ Bk18l13x+
Bk18l14x+ Bk18l15x+ Bk18l16x+ Bk18l17x+ Bk18l18x+ Bk18l19x)

svmat Bk18_lall

matrix list Bk18_lall

**
mkmat sxl1_1 sxl1_2 sxl1_3 sxl1_4 sxl1_5 sxl1_6 if tk==19, matrix(Bk19l1x)
mkmat sxl2_1 sxl2_2 sxl2_3 sxl2_4 sxl2_5 sxl2_6 if tk==19, matrix(Bk19l2x)
mkmat sxl3_1 sxl3_2 sxl3_3 sxl3_4 sxl3_5 sxl3_6 if tk==19, matrix(Bk19l3x)
mkmat sxl4_1 sxl4_2 sxl4_3 sxl4_4 sxl4_5 sxl4_6 if tk==19, matrix(Bk19l4x)
mkmat sxl5_1 sxl5_2 sxl5_3 sxl5_4 sxl5_5 sxl5_6 if tk==19, matrix(Bk19l5x)
mkmat sxl6_1 sxl6_2 sxl6_3 sxl6_4 sxl6_5 sxl6_6 if tk==19, matrix(Bk19l6x)
mkmat sxl7_1 sxl7_2 sxl7_3 sxl7_4 sxl7_5 sxl7_6 if tk==19, matrix(Bk19l7x)
mkmat sxl8_1 sxl8_2 sxl8_3 sxl8_4 sxl8_5 sxl8_6 if tk==19, matrix(Bk19l8x)
mkmat sxl9_1 sxl9_2 sxl9_3 sxl9_4 sxl9_5 sxl9_6 if tk==19, matrix(Bk19l9x)
mkmat sxl10_1 sxl10_2 sxl10_3 sxl10_4 sxl10_5 sxl10_6 if tk==19,
matrix(Bk19l10x)
mkmat sxl11_1 sxl11_2 sxl11_3 sxl11_4 sxl11_5 sxl11_6 if tk==19,
matrix(Bk19l11x)
mkmat sxl12_1 sxl12_2 sxl12_3 sxl12_4 sxl12_5 sxl12_6 if tk==19,
matrix(Bk19l12x)
mkmat sxl13_1 sxl13_2 sxl13_3 sxl13_4 sxl13_5 sxl13_6 if tk==19,
matrix(Bk19l13x)
mkmat sxl14_1 sxl14_2 sxl14_3 sxl14_4 sxl14_5 sxl14_6 if tk==19,
matrix(Bk19l14x)
mkmat sxl15_1 sxl15_2 sxl15_3 sxl15_4 sxl15_5 sxl15_6 if tk==19,
matrix(Bk19l15x)
mkmat sxl16_1 sxl16_2 sxl16_3 sxl16_4 sxl16_5 sxl16_6 if tk==19,
matrix(Bk19l16x)
mkmat sxl17_1 sxl17_2 sxl17_3 sxl17_4 sxl17_5 sxl17_6 if tk==19,
matrix(Bk19l17x)
mkmat sxl18_1 sxl18_2 sxl18_3 sxl18_4 sxl18_5 sxl18_6 if tk==19,
matrix(Bk19l18x)
mkmat sxl19_1 sxl19_2 sxl19_3 sxl19_4 sxl19_5 sxl19_6 if tk==19,
matrix(Bk19l19x)

**
matrix Bk19_lall=(1/1138)*( Bk19l1x+ Bk19l2x+ Bk19l3x+ Bk19l4x+ Bk19l5x+
Bk19l6x+ Bk19l7x+ Bk19l8x+ Bk19l9x+ Bk19l10x+ Bk19l11x+ Bk19l12x+ Bk19l13x+
Bk19l14x+ Bk19l15x+ Bk19l16x+ Bk19l17x+ Bk19l18x+ Bk19l19x)

svmat Bk19_lall

matrix list Bk19_lall

**
matrix B= Bk1_lall+ Bk2_lall+ Bk3_lall+ Bk4_lall+ Bk5_lall+ Bk6_lall+ Bk7_lall+
Bk8_lall+ Bk9_lall+ Bk10_lall+ Bk11_lall+ Bk12_lall+ Bk13_lall+ Bk14_lall+
Bk15_lall+ Bk16_lall+ Bk17_lall+ Bk18_lall+ Bk19_lall
```

```
svmat B

matrix list B

matrix covbeta= seinvA*B* seinvA

svmat covbeta

matrix list covbeta

**
```

Bootstrap estimates of the standard errors for the coefficients and the mean for Lin (2000) using multiple time intervals

```
For conventional (Similarly for intensive)

**
use "C:\Desktop\Lin2000\convyrs19_orig.dta", clear
rename ukno ukno0
do "C:\Desktop\Lin2000\bs_annualconv.txt"

where bs_annualconv.txt is:

**Lin Annual: Conventional**

program define pannual
      if "`1'"=="?" {
            global S_1 "b0 b1 b2 b3 b4 b5 meanc"
            exit
      }


gen tk_1=year-1
gen tk=year

gen Xik=min(Xi, tk)

egen maxtimeL=max(Xi)

gen dik_star=1 if (Xik==tk | (Xik==Xi & di==1) | (Xik==Xi & maxtimeL==Xi))
replace  dik_star=0 if dik_star==.

stset Xik, failure( dik_star==0)
sts gen GTik_star=s, by(tk)

drop _st _d _t _t0

egen meanage=mean(age)
egen meanbmi=mean(bmi)
egen meanfpg=mean(fpg)
egen meanrace=mean(race)
egen meansex=mean(sex)

gen int replicate=6

expand replicate

gen int const=1
```

```
sort ukno year

by ukno year: gen constx=sum(const)

sort ukno year constx

gen Zi=const if constx==1
replace Zi=age if constx==2
replace Zi=bmi if constx==3
replace Zi=fpg if constx==4
replace Zi=race if constx==5
replace Zi=sex if constx==6

move Zi age

move constx age_entr

move const  age

drop age_entr maxyear gender

sort  ukno year constx

gen Zi0_Zi0p=Zi*const
gen Zi1_Zi1p=Zi*age
gen Zi2_Zi2p=Zi*bmi
gen Zi3_Zi3p=Zi*fpg
gen Zi4_Zi4p=Zi*race
gen Zi5_Zi5p=Zi*sex

gen wZi0_Zi0p= ( dik_star/  GTik_star)* Zi0_Zi0p
gen wZi1_Zi1p= ( dik_star/  GTik_star)* Zi1_Zi1p
gen wZi2_Zi2p= ( dik_star/  GTik_star)* Zi2_Zi2p
gen wZi3_Zi3p= ( dik_star/  GTik_star)* Zi3_Zi3p
gen wZi4_Zi4p= ( dik_star/  GTik_star)* Zi4_Zi4p
gen wZi5_Zi5p= ( dik_star/  GTik_star)* Zi5_Zi5p

egen swZi0_Zi0p=sum(wZi0_Zi0p), by(tk constx)
egen swZi1_Zi1p=sum(wZi1_Zi1p), by(tk constx)
egen swZi2_Zi2p=sum(wZi2_Zi2p), by(tk constx)
egen swZi3_Zi3p=sum(wZi3_Zi3p), by(tk constx)
egen swZi4_Zi4p=sum(wZi4_Zi4p), by(tk constx)
egen swZi5_Zi5p=sum(wZi5_Zi5p), by(tk constx)

gen wYikZi=( dik_star/ GTik_star)* Mik*Zi
egen swYikZi=sum(wYikZi), by(tk constx)


sort ukno tk constx

collapse meanage meanbmi meanfpg meanrace meansex swYikZi swZi0_Zi0p swZi1_Zi1p
swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p, by(tk constx)

sort tk constx

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==1, matrix(bk1term1)

mkmat  swYikZi if tk==1, matrix(bk1term2)

matrix betak1=syminv(bk1term1)*bk1term2

svmat betak1
```

```
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==2, matrix(bk2term1)

mkmat  swYikZi if tk==2, matrix(bk2term2)

matrix betak2=syminv(bk2term1)*bk2term2

svmat betak2

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==3, matrix(bk3term1)

mkmat  swYikZi if tk==3, matrix(bk3term2)

matrix betak3=syminv(bk3term1)*bk3term2

svmat betak3

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==4, matrix(bk4term1)

mkmat  swYikZi if tk==4, matrix(bk4term2)

matrix betak4=syminv(bk4term1)*bk4term2

svmat betak4

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==5, matrix(bk5term1)

mkmat  swYikZi if tk==5, matrix(bk5term2)

matrix betak5=syminv(bk5term1)*bk5term2

svmat betak5

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==6, matrix(bk6term1)

mkmat  swYikZi if tk==6, matrix(bk6term2)

matrix betak6=syminv(bk6term1)*bk6term2

svmat betak6

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==7, matrix(bk7term1)

mkmat  swYikZi if tk==7, matrix(bk7term2)

matrix betak7=syminv(bk7term1)*bk7term2

svmat betak7

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==8, matrix(bk8term1)

mkmat  swYikZi if tk==8, matrix(bk8term2)

matrix betak8=syminv(bk8term1)*bk8term2

svmat betak8
```

```
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==9, matrix(bk9term1)

mkmat  swYikZi if tk==9, matrix(bk9term2)

matrix betak9=syminv(bk9term1)*bk9term2

svmat betak9

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==10, matrix(bk10term1)

mkmat  swYikZi if tk==10, matrix(bk10term2)

matrix betak10=syminv(bk10term1)*bk10term2

svmat betak10

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==11, matrix(bk11term1)

mkmat  swYikZi if tk==11, matrix(bk11term2)

matrix betak11=syminv(bk11term1)*bk11term2

svmat betak11

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==12, matrix(bk12term1)

mkmat  swYikZi if tk==12, matrix(bk12term2)

matrix betak12=syminv(bk12term1)*bk12term2

svmat betak12

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==13, matrix(bk13term1)

mkmat  swYikZi if tk==13, matrix(bk13term2)

matrix betak13=syminv(bk13term1)*bk13term2

svmat betak13

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==14, matrix(bk14term1)

mkmat  swYikZi if tk==14, matrix(bk14term2)

matrix betak14=syminv(bk14term1)*bk14term2

svmat betak14

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==15, matrix(bk15term1)

mkmat  swYikZi if tk==15, matrix(bk15term2)

matrix betak15=syminv(bk15term1)*bk15term2

svmat betak15
```

```
mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==16, matrix(bk16term1)

mkmat  swYikZi if tk==16, matrix(bk16term2)

matrix betak16=syminv(bk16term1)*bk16term2

svmat betak16

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==17, matrix(bk17term1)

mkmat  swYikZi if tk==17, matrix(bk17term2)

matrix betak17=syminv(bk17term1)*bk17term2

svmat betak17

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==18, matrix(bk18term1)

mkmat  swYikZi if tk==18, matrix(bk18term2)

matrix betak18=syminv(bk18term1)*bk18term2

svmat betak18

mkmat   swZi0_Zi0p swZi1_Zi1p swZi2_Zi2p swZi3_Zi3p swZi4_Zi4p swZi5_Zi5p if
tk==19, matrix(bk19term1)

mkmat  swYikZi if tk==19, matrix(bk19term2)

matrix betak19=syminv(bk19term1)*bk19term2

svmat betak19

matrix
beta=betak1+betak2+betak3+betak4+betak5+betak6+betak7+betak8+betak9+betak10+beta
k11+betak12+betak13+betak14+betak15+betak16+betak17+betak18+betak19

svmat beta

matrix list beta

gen b0x=beta1 if tk==1 & constx==1
egen b0=min(b0x)

gen b1x=beta1 if tk==1 & constx==2
egen b1=min(b1x)

gen b2x=beta1 if tk==1 & constx==3
egen b2=min(b2x)

gen b3x=beta1 if tk==1 & constx==4
egen b3=min(b3x)

gen b4x=beta1 if tk==1 & constx==5
egen b4=min(b4x)

gen b5x=beta1 if tk==1 & constx==6
egen b5=min(b5x)


gen meancost= b0+b1* meanage+b2* meanbmi+b3* meanfpg+b4* meanrace+b5* meansex
```

```
tempname y1
summarize b0, meanonly
scalar `y1'=r(mean)

tempname y2
summarize b1, meanonly
scalar `y2'=r(mean)

tempname y3
summarize b2, meanonly
scalar `y3'=r(mean)

tempname y4
summarize b3, meanonly
scalar `y4'=r(mean)

tempname y5
summarize b4, meanonly
scalar `y5'=r(mean)

tempname y6
summarize b5, meanonly
scalar `y6'=r(mean)

summarize meancost, meanonly
      post `1' (`y1') (`y2') (`y3') (`y4') (`y5') (`y6') (r(mean))
end

**

end of do-file

set seed 1001

bstrap pannual, reps(1000) dots cluster(ukno0) idcluster(ukno) saving
(C:\Desktop\Lin2000\bs_convannual1000.dta)

**
```