The Analysis of Data where Response or Selection is Dependent on the Variable of Interest

This work is presented as a thesis for the degree of

DOCTOR OF PHILOSOPHY

in

Statistical Science

at the

Faculty of Mathematical and Physical Sciences

University College London

by

Andrew John Copas

March 1999

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

ACKNOWLEDGMENTS

I would like to thank the following people:

Anne Johnson, Julia Field, Kaye Wellings and the late Jane Wadsworth who allowed access to data from the National Survey of Sexual Attitudes and Lifestyles, and helped with the interpretation of my work.

Two referees who made comments on the published version of chapter two.

The Steering Committee of the UK Register of HIV Seroconverters, who allowed access to their data.

The UCL Department of STDs, and in particular Anne Johnson and Ian Weller, who gave me the opportunity to study for this thesis part-time.

Vern Farewell for his encouragement and supervision.

Kavita for her support and love.

ABSTRACT

In surveys of sensitive subjects non-response may be dependent on the variable of interest, both at the unit and item levels. In some clinical and epidemiological studies, units are selected for entry on the basis of the outcome variable of interest. Both of these scenarios pose problems for statistical analysis, and standard techniques may be invalid or inefficient, except in some special cases.

A new approach to the analysis of surveys of sensitive topics is developed, central to which is at least one variable which represents the enthusiasm to participate. This variable is included along with demographic variables in the calculation of a response propensity score. The score is derived as the fitted probabilities of item non-response to the question of interest. The distribution of the score for the unit non-responders is assumed equal to that of item non-responders. Response is assumed independent of the variable of interest, conditional on the score. Weights based on the score can be used to derive unbiased estimates of the distribution of the variable of interest. The bootstrap is recommended for confidence interval construction. The technique is applied to data from the National Survey of Sexual Attitudes and Lifestyles. A simplification of the technique is developed that does not use the bootstrap, and which enables users to analyse the data without knowledge of the factors affecting non-response, and using standard statistical software.

To analyse the time from an initiating event to illness, a prospective study may be regarded as the optimal design. However, additional data from those already with the illness and still alive may also be available. A standard technique would be to ignore the additional data, and left-truncate the times to illness at study entry. We develop a full likelihood approach, and a weighted pseudolikelihood approach, and compare these with the standard truncated data approach. The techniques are used to fit simple models of time to illness based on data from a study of time to AIDS from HIV seroconversion.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Description of the General Problem

In many studies the units available for analysis are not a random sample of the population of interest, either overall or within identifiable subsets. In some studies this is an undesirable consequence of non-response or drop-out, in others it is a deliberate feature of the selection mechanism employed for the study. There is a natural distinction between these two scenarios, and hence we describe each in more detail in the following subsections. In general, standard techniques of analysis for the distribution of the variables of interest will lead to bias or a loss of efficiency in these situations, though there are some exceptions. This thesis is concerned with techniques of analysis which reduce such bias and/or increase efficiency. The focus is on the development of new techniques in the context of survey non-response, and the analysis of time to illness in certain cohort studies.

1.1.1 Non-response in Surveys and Ignorability

In survey analysis there may be several types of bias. Firstly there may be bias due to the limitations of the sampling frame, e.g. those without telephones are excluded from a sampling frame based on a telephone directory. A related bias arises if whilst some units are included in the sampling frame, their lifestyle determines that they would never actually be invited to participate even if many attempts are made to approach them, e.g. shift workers might never be approached in a survey based on calling at residential addresses

during the weekday evenings. An important potential source of bias, particularly in surveys of sensitive issues is misreporting. This may arise through respondent memory error, or through an unwillingness to report certain behaviours or experiences. Such an unwillingness might be expected particularly with respect to behaviours or experiences which are socially unacceptable and/or illegal in that country. The source of bias that will be primarily addressed in this thesis is that from non-response, considered to be refusals, and tacit refusals, such as breaking pre-arranged appointments with the interviewer.

In surveys, non-response may occur at the unit or item level. By unit non-response we mean that no interview takes place with the sampled unit. By item non-response we mean that whilst the interview is partially completed, no response is obtained for the question of interest. Unit response rates in surveys are rarely close to 100%, and item non-response results in yet more missing data. Non-response may be ignorable or non-ignorable. At the unit level, ignorability means that non-response is independent of the variable of interest, at least within classes defined by information available for all units. At the item level, ignorability means that non-response is independent of the variable of interest, at least within classes defined by information available for all unit responders. Since the variable of interest is unobserved for both item and unit non-responders (NRs), it is not possible to test the degree to which non-response is ignorable, except where extra information is available (e.g. from interviewing a random sample of the initial NRs). Hence techniques of estimation may only proceed under assumptions about non-response, and the issue of the sensitivity of estimates to these assumptions is naturally of central interest.

It is commonly assumed that both unit and item non-response are ignorable. Under this assumption, techniques of imputation ('filling-in' missing values with data observed for other units) and weighting provide simple unbiased estimation of population parameters (see section 1.2.1). These techniques are often used where ignorability is thought to be roughly true, i.e. where the degree of non-ignorability is considered small. However the use of these techniques fails to remove bias where non-response is non-ignorable i.e. dependent

on the unobserved answer even within classes defined by information available for all units. Furthermore the bias present may be as large or larger than the bias in an analysis based only on item responders, and the standard errors of estimates may be greater.

In this thesis we address the issue of missing data where unit and item non-response are thought to be non-ignorable. In surveys, non-ignorable unit and item non-response is likely whenever the topic of interest is sensitive. Our work is motivated by the need to estimate behavioural prevalences on the basis of the National Survey of Sexual Attitudes and Lifestyles (NATSSAL). Details of this survey are given in Johnson *et al.* (1994) and Wadsworth *et al.* (1993).

1.1.2 Response-based Selection and Analysis of Time to Illness

In many fields of research, and notably in epidemiology, studies are designed in which units are selected with probability dependent on the value of the variable of interest. A commonly used example is the epidemiological case-control study, in which 'cases' are selected with a certain probability, and 'controls' with another. Other examples include the case-cohort, prevalent cohort, and various two phase sampling designs. Such designs have been developed because they offer greater efficiency than alternative designs. For example, when a condition is rare, then a study to examine differences between those who acquire the condition and those who do not may only be practical if cases are more likely to be sampled than controls.

For some studies where selection is dependent on the outcome variable, e.g. the case-control study, techniques of analysis have been developed which ignore the selection mechanism. This however is not possible with all such study designs. Techniques of analysis for all of the established designs are well developed, see section 1.2.2.

Consider an illness-death model in which there is an initial state e.g. HIV seroconversion or birth, an illness state e.g. AIDS or diabetes, and death. Units may experience death with or without prior illness. See Figure 1.1 for clarification. The hazards of illness and of death without illness can be considered to be 'competing risks' (Kalbfleisch and Lawless,

1980). A model such as the proportional hazards model may be specified for the hazard of illness, and a semi-parametric version is available (Cox, 1972).

To examine the effect of factors on the illness process, a prevalent cohort study may be designed. In such a study a sample of those still in the initial state is recruited. Selection then is dependent on the time to illness, the variable of interest. Often units that have died are not recruited because it is uncertain how many such units there are, or data may not be reliably recorded for such units, or because interest is in the effect of a covariate which was not recorded in the past. For example if interest centres on the effect of a genotype on time to illness, then unless appropriate samples have been taken and stored from those who have since died, then only those alive can be recruited to the study.

Where interest centres on both time to illness and time from illness to death, then a natural study design consists of recruiting patients who are alive. Such a design could be termed an augmented prevalent cohort (APC) design, since it consists of a prevalent cohort design augmented by units who are ill but alive. Typically time to illness would be analysed using only data from those units who are not yet ill, i.e. from the prevalent cohort that is 'nested' within an APC study. The units that are ill at entry, which could be termed 'retrospective' cases, would be used in the analysis of time from illness to death. Chapter 4 of this thesis will consider whether there may be a benefit to the researcher from additionally including these retrospective cases in the analysis of time to illness, or equivalently whether the APC design may be preferable to the prevalent cohort analysis even when interest is primarily or exclusively in the time to illness. Recruitment to an APC study, as to a prevalent cohort study, is clearly related to the time to illness, the outcome of interest, and could be described as 'selection by virtue of survival' (Hoem, 1985).

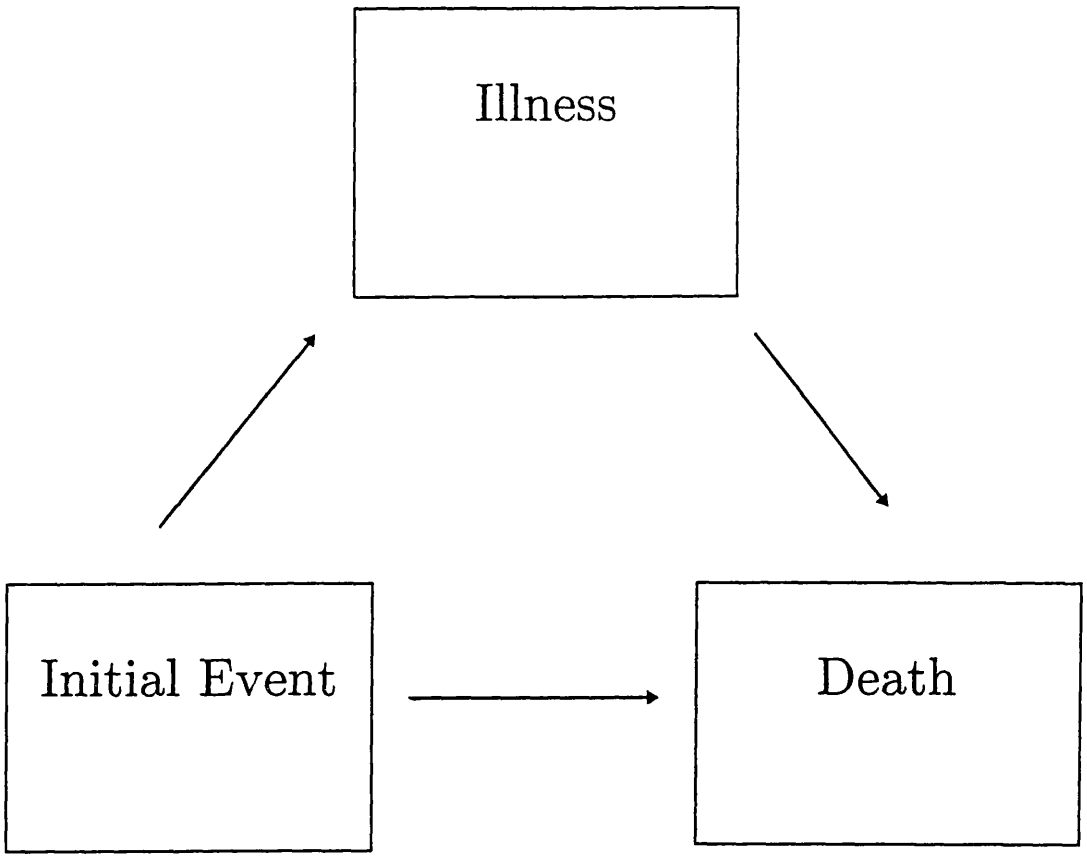


Figure 1.1 The illness death model

Whilst in theory there must be a gain to the researcher from the recruitment of additional units, the issue of whether there is a practical benefit relates to the development of suitable analytical techniques. In this thesis alternative methods of analysis of an APC study are developed and compared. These techniques require information concerning time to death (with and without illness).

1.2 Previous Work

There is some published work that explicitly brings together the fields of non-ignorable non-response and selection dependent on the outcome variable. This work includes that of Lawless *et al.* (1999), and some work in the field of two-phase sampling designs, which are described in section 1.2.2. However other previous work has developed techniques for application to only one of the two scenarios. Hence we describe the work in the two areas separately, with an emphasis on those topics most relevant to the methodology developed in later chapters. In the field of non-response we briefly describe previous work related to ignorable non-response in surveys, to non-ignorable non-response in categorical outcome variables, and to some alternative approaches. In the field of selection dependent on the outcome variable we consider primarily work relating to the case-control study and various cohort studies, and to the use of pseudolikelihood techniques.

Overviews of techniques for dealing with missing data are provided by Little and Rubin (1987) and more recently by Schafer (1997). Apart from work in the areas of missing data mentioned previously, some authors have focused on non-ignorable missing data in repeated measures studies either due to drop-out (Diggle and Kenward, 1994, Little, 1995) or more general patterns of missingness (Conaway, 1994, Baker, 1995). Heitjan (1993) introduces the concept of coarse data, where data is not exactly observed, e.g. censored, or grouped. Sometimes the coarsening can be ignored in analysis and sometimes not, as with missing data.

Previous work in the analysis of studies where selection is based on the outcome variable has in general focused on particular study designs. The description of previous work here focuses on those study designs commonly applied in epidemiology, economics, and the social sciences.

1.2.1 Previous Work in the Field of Non-response

Reviews of the methods to adjust for ignorable non-response are provided by Chapman (1976), Kalton (1983), Kalton and Kasprzyk (1986), Little and Rubin (1987), and Little (1988). Weighting and imputation are the two most commonly used methods. Often weighting is used to account for unit non-response, and imputation for item non-response, though either method can be used for non-response at either level. Techniques of imputation are described by Rubin (1987). When weighting is used to adjust for unit non-response, generally weighting classes are defined and then a weight for each class is calculated in some way.

There are several methods by which to define the weighting classes. In general, whilst classes are defined in terms of factors known to be associated with response, informally factors may be chosen from amongst those which are also associated with the variables of interest. In many surveys classes are based solely on data from that survey, a technique known as sample weighting. Such classes would be defined in terms of variables recorded for all units. The class weight would be simply the inverse of the observed unit response rate. A second approach would be to define classes in terms of information known for the target population (e.g. census data), a technique known as population weighting or informally as 'post-stratification' (Holt and Smith, 1979). In population weighting, the variables used to define the weighting classes may include some recorded for item responders only. The aim of this technique is to standardise or calibrate the estimates from the survey to the target population, and so the class weights would be the population proportions. The class weights from population weighting are typically treated as fixed, since they are based on

large sample data.

In some cases the number of variables selected for possible use in defining weighting classes is small, e.g. in population weighting where the number of variables recorded in both the census and the survey itself is limited. In that case classes may simply be defined as all cells defined by the values of the variables, categorised if necessary. However often this approach would lead to many classes, which in turn would lead to inefficiency due to small numbers of item responders in some cells. Possible approaches to forming a small number of classes from many variables include the use of so-called 'automatic interaction detection' software where response is considered to be the outcome variable (Goksel *et al.*, 1991), categorising on the basis of a response propensity score (Little, 1986, Czajka *et al.*, 1992) and the formation of classes based solely on those variables found important in a regression model of non-response. See Bloch and Segal (1989) for a comparison of these techniques applied in a different context. Other possibilities are based around examining the association of factors with the variable of interest, rather than non-response (Little, 1986). Where there are two categorical variables available for weighting class definition, an alternative technique known as raking may be used (Deville *et al.*, 1993). Under this technique the weighted marginal distribution of the item responders will match the distribution in the whole sample or that observed in the census, without ensuring matching on the joint distribution of the two variables. Whilst the number of weighting classes is large, being the number of cells in the cross-classification, under this technique the variability in the class weights is controlled.

A method of implementing weighting adjustment which will feature heavily in chapter two is the use of a response propensity model. This is a modification of a technique originally developed for the analysis of observational studies (Rosenbaum and Rubin, 1983, 1984). Under this approach a model of response such as a logistic regression model is fitted, and the fitted probabilities are considered to be response propensity scores. The possibility of inverting the fitted probabilities of response from this model to calculate weights has been discussed by Little (1986, 1988) and Cassel *et al.* (1983). Czajka *et al.* (1992) apply this

method to some income and tax data. Iannacchione *et al.* (1991) also propose a method based on fitting a response model, and inverting the fitting probabilities to form weights. The fitting of the model is however constrained so that the weighted sum of the explanatory variables across the responders is equal to the unweighted sum across both responders and non-responders. This approach can be regarded as an extension of 'raking'. However most authors have suggested that the scores be used to form weighting classes. The key advantage of formulating classes rather than applying the inverse score directly would be considered to be a reduction in the variance of the estimator, since the inverse scores may sometimes be very large. A second advantage may be considered to be that the method puts less reliance on specifying a correct form for the response model. These arguments are described by Little (1986), and Czajka *et al.* (1992) apply this approach. Goksel *et al.* (1991) also apply a weighting class approach, but the classes are formed directly from a categorical response propensity model, fitted by automatic interaction detection software. In many surveys, those variables available for all units are very limited, and hence the propensity score technique would more naturally be applied to item non-response than unit non-response.

Previous authors have suggested that in some surveys, in order for item response to be considered independent of the variables of interest within classes defined in terms of information available for all units, then this information must include measures of 'survey hostility' in addition to other variables. Goksel *et al.* (1991) analyse data from a two-phase survey, using the number of address moves, the number of visits to the address before the interview occurred, and whether the interviewee supplied their phone number or not as measures of 'survey hostility' in their formation of weighting classes. These variables are typically unavailable for unit NRs, so obviously this approach could not be used directly to adjust for unit non-response.

A related approach is based on an assumption that missing units have the same distribution of the variable(s) of interest as those item responders of low availability. Bartholomew (1961) proposes, for example, that only two attempts be made to interview units, and that

the distributions of the variables of interest in the missing units then be assumed to be the same as that observed amongst item responders for whom two attempts were needed to secure an interview. Politz and Simmons (1949) propose a related approach based on only one attempt to secure an interview, in which unit responders are asked to state how many of the previous six evenings they had been at home, and therefore available for interview. Estimation can then clearly proceed by weighting up the item responders of low availability. This work was designed to address 'not at home' bias, rather than the effect of refusal, and to limit the number of times attempts are made to secure an interview. However these methods could be adapted to deal with non-response. On such adaptation to deal with non-ignorable non-response, for example, might be to assume that unit and/or item NRs have the same distribution of the variables of interest as those item responders who initially asked the interviewer to come back on another day, so that they could think about whether to participate or not.

In the absence of additional information (e.g. brief interviews with a random subset of initial unit and item NRs) it cannot be determined whether non-response is ignorable or non-ignorable. Several possible methods of analysis to deal with potentially non-ignorable unit and item non-response in surveys have been proposed. In general, these methods include approaches to assessing the sensitivity of estimates to the assumptions which are made about non-responders. Perhaps the most obvious method proposed in the literature is that of formulating twin log-linear models of response (unit and item) and of the variable of interest. Baker and Laird (1988), Chambers and Welsh (1993), Park and Brown (1994), and Ibrahim and Lipsitz (1996) describe variations on this approach, with useful practical advice. Baker and Laird (1988) focus on a categorical outcome variable and categorical explanatory variables, and suggest the use of the E-M algorithm for model fitting and of a profile likelihood for confidence interval construction. Park and Brown (1994) suggest a Bayesian modification where informative priors are used to avoid the boundary solutions that frequently arise from the approach of Baker and Laird. Chambers and Welsh (1993) ad-

dress the more general situation where, whilst all variables are categorical, several outcome variables may be considered simultaneously, and explanatory variables are not necessarily observed for all units. They suggest the use of a Newton-Raphson algorithm to fit the models. Ibrahim and Lipsitz (1996) focus on the case of a binary response variable. The E-M algorithm is suggested for model fitting, with an extension to estimate the observed information matrix. However model fitting under these techniques is a slow process. For a description of fitting such log-linear models to our example dataset, following the approach of Baker and Laird, see section 3.4. A sensitivity analysis based on these techniques would consist of examining the range of estimates derived from those twin models which give an adequate fit to the data from the responders.

Alternative formulations of the problem are presented by Fay (1986) and Little (1993). Fay develops what are called 'causal models', which can generally be represented pictorially, which may include non-ignorable models. Little (1993), and later Little and Wang (1996), develop 'pattern-mixture' models, where twin models are specified of response level, and of the variable of interest conditional on response level. By response level we refer to the missing data pattern, which in the survey context could naturally mean item responder, item NR and unit NR. Whilst it might in general be considered more natural to represent response as depending on explanatory variables and the variable of interest, the 'pattern mixture' modelling approach allows a wide range of possible models to be fitted, and also makes the assumptions required more transparent. A Bayesian approach to deal with non-ignorable non-response in a categorical outcome variable has been developed by Paulino and Pereira (1995).

Forster and Smith (1998) and Copas and Li (1997) have developed techniques of specifying models for the data in which the strength of non-ignorability is represented by one or more parameters. Forster and Smith (1998) address the issue of non-response to categorical variables, and propose a model in which a group of parameters determine the extent of non-ignorability. They express a plausible distribution of the extent of non-ignorability

through prior distributions and assumptions about the parameters. For the case of item non-response to a continuous variable, Copas and Li (1997) propose a model in which a single parameter determines the degree of non-ignorability. They suggest exploring the effect on estimates as this parameter varies over what might be considered to be a plausible symmetric range, centred on the case of zero non-ignorability (i.e. ignorable non-response). They also suggest that in a 'well-designed and well-executed survey' that the degree of non-ignorability would not be large. However, as is pointed out by Raab (1997) in the discussion of this paper, in fact when a survey is properly carried out, and the response rate very high, then the strength of the dependence of response on the variable of interest may be very high, since the only non-responders will be the 'die-hard' refusers. Such refusers may be very different to the rest of the population.

1.2.2 Previous Work in the Field of Response-based Selection

Some of the earliest work in this field relates to the analysis of the case-control study. The explicit use of logistic disease incidence models in the analysis of case-control studies was developed by Farewell (1979) and Prentice and Pyke (1979), after earlier work by Anderson (1972). This study design has become so commonly used in part because prospective models can be applied, and estimation proceeds in a relatively straightforward manner. Issues arising from stratification of the case-control study are addressed by Scott and Wild (1991, 1997). Weinberg and Wacholder (1993) demonstrate that prospective models more general than the logistic can be used to analyse case-control data, ignoring the selection process. The class of models that can be applied in this way to case-control studies is termed multiplicative-intercept risk models.

The case-control design is equivalent to choice based sampling in economics and the social sciences. Hausman and Wise (1981) describe analysis of a choice based sampling study of income, where income is grouped and then units sampled at different rates according to their group membership. Hsieh *et al.* (1985) and Manski and McFadden (1981) consider

how one may estimate prospective probabilities from case-control or choice based sampling data when one has auxiliary data, or one is prepared to make structural assumptions.

The prevalent cohort study can be analysed using prospective regression models such as the proportional hazards model (Cox, 1972). This is indeed very natural when using hazard based methods since units are simply truncated at their point of entry and then supply wholly prospective information relating to the hazard function from that time point onwards. Both fully and semi-parametric approaches are available for proportional hazards models as with standard prospective cohort designs. The methods of analysis are clearly presented by Wang *et al.* (1993), and details are also given in chapter 4, where such models are applied to simulation and example data.

For the class of designs that could be described as sampled cohort designs which typically consist of sampling all cases and a proportion of the controls, e.g. nested case-control, partial likelihood techniques based on the proportional hazards model can be applied. Prentice and Breslow (1978) demonstrate this for the nested case-control study, and Borgan *et al.* (1995) and Langholz and Goldstein (1996) develop a general approach for sampled cohort designs.

A pseudolikelihood is essentially an estimate of the full likelihood for a finite population. Typically nuisance parameters are replaced in the full likelihood by estimates, and then the reduced set of equations maximised. Gong and Samaniego (1981) present some theoretical work in the area of pseudolikelihood. Hu and Lawless (1997) present a range of possible approaches to forming pseudolikelihoods, as do Lawless *et al.* (1999) whose focus is on semiparametric regression. Pseudolikelihood based regression methods have been proposed for a variety of other outcome dependent selection scenarios and also for missing data problems including potentially non-ignorable non-response. For case-control studies, Wild (1991) and Scott and Wild (1997) present and compare full likelihood and pseudolikelihood techniques to apply prospective regression models. Scott and Wild (1997) develop an algorithm based on a pseudolikelihood but which leads finally to the maximum full likelihood estimates. Prentice (1986) and Kalbfleisch and Lawless (1988) present pseudolikelihood

based approaches to the analysis of time to illness from a case-cohort study. Samuelsen (1997) develops a pseudolikelihood based technique to analyse time to illness in a nested case-control study. In the context of complex survey data, Skinner (1989, 1996) presents pseudolikelihood based regression methods to deal with missing data that may arise through non-ignorable non-response.

Outcome dependent selection is also used in a variety of two-phase designs. At the first phase limited data is collected on the entire sample, and then in the second phase complete data is collected on a subset of the first phase units. The selection of second phase units may depend on any of the variables collected at the first phase, including in the case-control setting, that defining whether the unit is a case or control. Overviews of these designs and their analysis are given by Carroll *et al.* (1995), and also by Reilly and Pepe (1995). Such studies may consist of collecting cheaper error prone covariate measurements at phase one, or collecting a surrogate at phase one, and then the covariate of real interest at phase two. Alternatively studies arise from missing data at phase two. The pseudolikelihood is often recommended in the literature for such designs (Flanders and Greenland, 1991, Schill *et al.*, 1993). A related technique described as a mean score method is suggested by Reilly and Pepe (1995). Where full likelihood is recommended, its implementation is generally through an iterative procedure, such as the E-M algorithm (Wacholder and Weinberg, 1994), or techniques based on pseudolikelihood (Breslow and Holubkov, 1997).

In the work reported later relating to the analysis of the time to illness in augmented prevalent cohort studies the most relevant papers are Kalbfleisch and Lawless (1988) and Samuelsen (1997), though they deal with different study designs. Their development of pseudolikelihood based techniques is relatively similar, and indeed the development presented in chapter 4 can be regarded as an extension of these methods to a different study design. Kalbfleisch and Lawless (1988) present a very clear description of their pseudolikelihood approach and develop a variance estimator for their parametric pseudolikelihood. Samuelsen (1997) develops a variance estimator for his semiparametric pseudolikelihood

approach which forms a basis for our own. A key distinction between the study types considered by these authors (the case-cohort and the nested case-control) and the APC design is that in the APC design the recruitment probability of a unit conditional on the variable of interest needs to be estimated, but in the other designs it is known.

Keiding *et al.* (1989) analyse diabetes incidence from prevalent cohort data together with a separate source of historical mortality data. This has some similarity with the analysis of an APC study in that both prospective and retrospective data are analysed together. The focus of their research, however, was on presenting smoothed incidence rates, rather than on regression analysis of the time to illness, which forms the focus of this work.

APC studies have been used in many contexts where interest has focused on both the time to illness and the time to death. Where the times of the initiating event are known for the study units, then analysis of time to illness has proceeded solely on the basis of the prospective cases, using standard prevalent cohort techniques (Wang *et al.*, 1993). The time of initiating event is unknown in some studies, for example in the UK MRC Collaborative Study in HIV Infection in Women, where the ideal initiating event would be HIV seroconversion, but this is unknown for most women. The authors therefore analysed time from study entry to AIDS (Study Group, 1998), using the prospective cases only. This study provided the initial motivation for the work reported in chapter 4, but the methods developed there are primarily applicable to studies where the initial event is easily ascertained e.g. birth, or HIV seroconversion within a seroconverter cohort.

Chapter 2

Dealing with Non-ignorable Non-response using an 'Enthusiasm to Respond' Variable

2.1 Introduction

In surveys, non-ignorable unit and item non-response may be likely whenever the topic of interest is sensitive. Our work is motivated by the need to construct unbiased estimates of population parameters in analysis of data from the British National Survey of Sexual Attitudes and Lifestyles (NATSSAL). The level of virginity (defined as no heterosexual intercourse) will be our example variable of interest, and background information about the survey and reasons to suspect non-ignorable unit and item non-response are provided in the next section.

To estimate population parameters from survey data where non-response is thought to be non-ignorable, previous authors have suggested a variety of techniques, see section 1.2.1. These include fitting twin log-linear models of response and the variable of interest (Baker and Laird, 1988, Chambers and Welsh, 1993, Park and Brown, 1994, and Ibrahim and Lipsitz, 1996). This technique however has limitations which are discussed in section 3.4, when we apply the method to our example data. Other authors have suggested that models should be formulated where the strength of non-ignorability is determined by one or more parameters, and then the estimates of interest be computed over a plausible range or distribution of these parameters. Forster and Smith (1998) address the problem of non-response in a categorical variable, and Copas and Li (1997) that of item non-response in a

continuous variable.

Our approach is motivated by the desire to use as much of the information in a survey dataset as possible, under certain special assumptions. These assumptions centre on a response propensity score, and link the unit NRs to the item NRs. In our approach, all useful information can be used, even if only recorded for unit responders. However, at least one variable representing the enthusiasm to respond is required to make one of the assumptions reasonable. The approach can be thought of as weighting the item responders, and this leads to simple calculation of estimates.

In section 2.3 we outline the assumptions of the method and the form of the estimator. In section 2.4 we compare methods of confidence interval calculation. In section 2.5 we present results from the application of the method to our example problem, and in section 2.6 we examine sensitivity to our assumptions.

2.2 The Survey Data

NATSSAL was conducted in 1990-91, and full details of the sampling are published elsewhere (Wadsworth *et al.*, 1993). See Johnson *et al.* (1994) for full questionnaires and extensive results. A total of 50010 addresses were issued, and 29807 were found potentially eligible, from which one person was selected and interviewed at 18876 (63%). At many of those addresses where no interview was completed, age and sexes were supplied, and one person selected, before refusal occurred. Unit non-response was found higher amongst the old and amongst men (see Johnson *et al.*, 1994, for details). The analysis for this thesis is restricted to those 25505 units with recorded age and sex.

The interview combined face-to-face questioning and a self-completion booklet containing the more sensitive questions. A proportion were not offered the booklet based on responses to questions about past sexual behaviour in the face-to-face interview. All those who reported no sexual experience were not offered the booklet. In addition all those aged 16 or 17

who reported only sexual experience with the opposite sex and no intercourse since 13, were not offered the booklet. In total 2.9% of the unit responders were not offered the booklet and of those offered, 3.8% refused the booklet. The answers to questions in the booklet can be deduced in almost all cases for those not offered it, due to their lack of experience. Item non-response for most questions of interest is less than 5%, for virginity (which is a question asked before booklet offer) it is only 0.9%.

Without additional data, it is impossible to test whether unit and item non-response in a survey are ignorable or non-ignorable. However one approach which may provide a useful hypothesis about non-response is to compare subgroups of the item responders. These subgroups could be willing and reluctant item responders, defined in some way, under the assumption that any reported differences in the question of interest between the two subgroups may also reflect differences between item responders and NRs (unit and item). Of course, the possibility that any such differences between the two subgroups have arisen through misreporting by the unwilling cannot be discounted, and in this scenario the assumption is unreasonable. An investigation of this type was carried out for the NATSSAL; for details see Copas *et al.* (1997). Those embarrassed during the interview were compared to those not embarrassed (as recorded by the interviewer), and those who refused the self-completion booklet to those who accepted. After controlling for demographic variables, willingness to respond was still associated with sexual behaviour. Assuming that unit and item NRs have sexual behaviour closer to that reported by the booklet refusers and the embarrassed than to that reported by the other item responders, the findings indicate likely non-ignorable non-response. The main analyses of the survey were performed assuming ignorable unit and item non-response, but the possibility of non-ignorable non-response related to one particular question has been considered (see Wadsworth *et al.*, 1996).

Sampling of households in NATSSAL followed a two stage design, and from each household one person was selected without replacement. For simplicity the data are analysed in this chapter as if the sampling method were simple random sampling. An extension of the

method proposed in this chapter to complex sampling schemes is presented in section 3.3.

2.3 The Response-Propensity Model

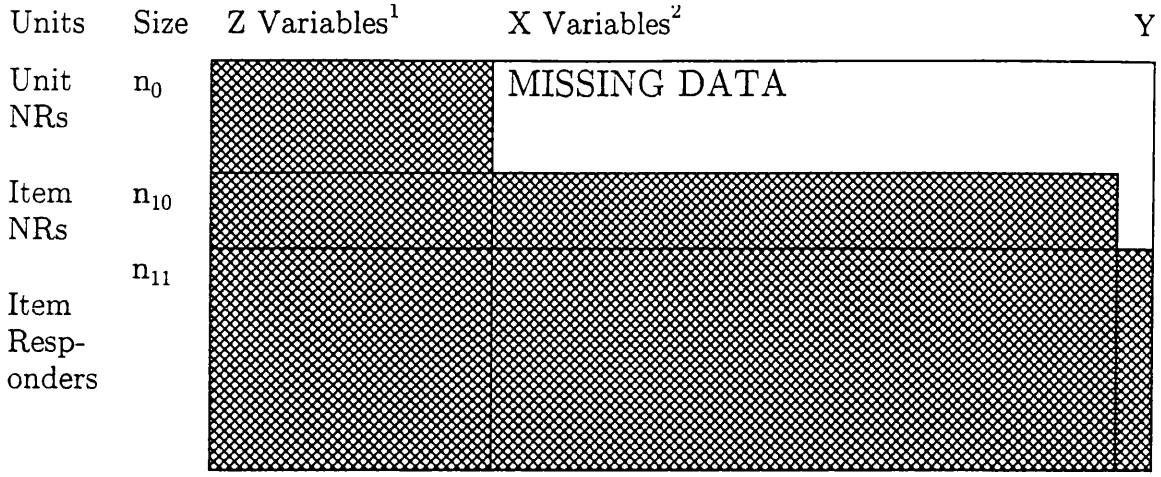
For any given variable of interest, Y, the NATSSAL units can be divided into three response categories: item responders, item NRs, and unit NRs. Equally the explanatory variables can be divided into those limited, Z, variables observed for all units and those, X, variables observed only for unit responders. We define Z classes to represent all combinations of the Z variables (categorised as appropriate). Ignoring item non-response amongst X variables, this pattern is displayed in Figure 2.1.

The broad aim of the model we propose is firstly to impute the Y distribution of the item NRs using all available X and Z variables, including in particular X variables thought to represent the enthusiasm to respond (unit and item). This is achieved by weighting based on a response propensity score, as discussed in Little (1986) and Little (1988). Secondly we propose that the unit NRs be assigned the same Y distribution as item NRs, within Z classes.

2.3.1 The Role and Calculation of the Propensity Score

The assumptions that underlie the model we propose centre on a score variable, S, representing the propensity to respond (unit and item), a measure which is conditional on the Z and X variables.

- *Assumption 1*: Conditional on Z variables and S, unit and item response are independent of Y.
- *Assumption 2*: The S distribution of the unit NRs is the same as that of the item NRs, within Z classes.



1. age, sex, urban/rural
2. marital status, occupational class, ethnicity, problems of understanding, embarrassment etc.

Figure 2.1 The response pattern for variables in NATSSAL

Pattern-mixture models (see Little, 1993) provide a framework to consider these assumptions. In the language of these models, Assumptions 1 and 2 are identifying restrictions. Assumption 2 provides a link between different response patterns, and Assumption 1 is a complete-case missing-variable restriction. Weighting based on a response propensity score and on assumptions similar to Assumption 1 is now a relatively common practice. For applications of this method see Goksel *et al.* (1991) and Czajka *et al.* (1992), as described in section 1.2.1. In surveys of sensitive topics such as NATSSAL, Assumption 1 is only reasonable if S incorporates a measure of the enthusiasm to respond, in addition to demographic information. Previous authors have similarly included measures of 'survey hostility' in their adjustments for missing data (Goksel *et al.*, 1991).

Suppose for unit responders that R represents item response, $R = 0$ denoting non-response, and $R = 1$ response. We propose to calculate S for unit responders as the fitted probabilities of item response from a logistic regression model, regressing R on those Z and X variables thought important. In our NATSSAL example the variables ultimately included in the model are age (continuous), household occupational class (categorised as skilled non-manual, unskilled, and unclassifiable), problems of understanding (yes/no), and interviewee embarrassment (very, somewhat, only slightly and not at all). Interactions with age were tested, and none found significant. Interestingly interviewee gender did not contribute significantly to the model, and since 90% of unit responders were interviewed by women, the interaction between interviewee and interviewer genders was not considered. The fitted odds ratios from the model are presented in Table 2.1. Embarrassment is the dominant term in this model. In model fitting, the reduction in deviance due to the embarrassment term was only increased by 4.5% on addition of the other terms.

Factor		Fitted O.R.	95% Conf. Interval
Embarrassment	Very	0.011	0.007 - 0.018
	Somewhat	0.051	0.033 - 0.081
	Only Slightly	0.340	0.204 - 0.566
	Not at all	1	-
Age	Increase of 1 yr	0.981	0.967 - 0.994
Problems of Understanding	No	1	-
	Yes	1.81	1.06 - 3.09
Household Occupational Class	Skilled non-manual	1	-
	Unskilled or manual	0.70	0.48 - 1.01
	Unclassifiable*	0.62	0.36 - 1.06

Table 2.1 Logistic regression model of the odds of item response

*This arises where no-one in the household has ever been in employment

Let T represent unit response, $T = 0$ denotes non-response and $T = 1$ response. The three combinations of T and R that are logically present are $T = 1$ and $R = 1$, $T = 1$ and $R = 0$, and $T = 0$. Then symbolically, Assumption 1 states that

$$Pr(Y|R,T,S,Z) = Pr(Y|S,Z) \tag{2.1}$$

2.3.2 Adjustment for Item Non-Response

Consider $\theta = Pr(Y = y)$ to be the parameter we wish to estimate within one Z class. Consider first the adjustment required to account for item non-response. Under Assumption 1, obvious possibilities would be weighting in classes defined by categorising S (see Czajka *et al.*, 1992), or some form of 'hot-deck' imputation (Rubin, 1987).

However, for our example, we propose to weight item responders by the inverse fitted probability of item response for that individual, taken from the model defining S, as discussed in Little (1986), Little (1988) and Cassel *et al.* (1983). We do however calculate weights subject to marginal constraints, viewing the approach as distributing the observed number of item NRs across the levels of Y. The key characteristic of this approach is that it uses Y variable information from all item responders in the Z class, whereas the alternative methods use only information from the (potentially few) item responders with similar values of S to the item NRs. This may be an advantage where (as in our example) item

non-response is low and the number of Z classes large so that there are very few item NRs in each class. This is particularly important when we deal with unit non-response in the next section, and indeed our approach will impute the Y distribution of unit NRs in a Z class even when there are no item NRs. In our example the minimum fitted probability of item response across the item responders is 0.65. Where item non-response is higher, this strategy will assign to some item responders weights greater than any assigned under alternative methods. This is recognised as a potential disadvantage by Little (1986). A brief discussion of the advantages of weighting in classes is provided in section 1.2.1.

Suppose the explanatory variables (including all Z variables) in the model defining S are all categorical, and all higher order interactions involving these variables are included in the model. In this case, our approach is equivalent to weighting by classes defined by the levels of S, within each Z class. More generally, S may be continuous, though since the number of item responders will be finite, it can be regarded as discrete with units taking unique values. It is convenient to assume that the S distribution among item and unit NRs can be adequately represented as a distribution over those values found amongst the item responders. Bearing this in mind our approach is based on the following summation over the levels, $\{s\}$, of S:

$$Pr(Y = y|T = 1) = \sum_s \Pr(Y = y|T = 1, S = s) Pr(S = s|T = 1) \qquad (2.2)$$

where

$$\begin{aligned} & Pr(S = s|T = 1) \qquad (2.3) \\ = \; & Pr(S = s|R = 1)Pr(R = 1) + Pr(S = s|R = 0)Pr(R = 0) \\ = \; & Pr(S = s|R = 1)Pr(R = 1)[1 + \{Pr(S = s|R = 0)Pr(R = 0)\}/ \\ & \qquad \{Pr(S = s|R = 1)Pr(R = 1)\}] \\ = \; & Pr(S = s|R = 1)Pr(R = 1)[1 + \{Pr(R = 0|S = s)/Pr(R = 1|S = s)\}] \end{aligned}$$

Let n represent the total number of units in the Z class. Let n_0 represent the number of unit NRs ($T = 0$), and n_1 the number of unit responders ($T = 1$), so that $n = n_0 + n_1$. Furthermore let n_{10} represent the number of item NRs ($T = 1, R = 0$), and n_{11} the number of item responders ($T = 1, R = 1$), so that $n_1 = n_{10} + n_{11}$. Let n_{11}^s represent the number of item responders with $S = s$, and let n_{11}^{sy} denote the number of item responders with $S = s$ and $Y = y$.

In using the value of S to estimate $Pr(R = 0|S = s)/Pr(R = 1|S = s)$, the inflation factor to represent item non-response, the natural estimate is $s^{-1} - 1$. However we fix the marginal ratio $Pr(R = 0)/Pr(R = 1)$ to be the observed ratio n_{10}/n_{11}, by introducing a scaling constraint a_1, such that

$$\sum_s n_{11}^s a_1(s^{-1} - 1) = n_{10} \tag{2.4}$$

and we estimate $Pr(R = 0|S = s)/Pr(R = 1|S = s)$ by $a_1(s^{-1} - 1)$.

Under Assumption 1,

$$Pr(Y = y|T = 1, S = s) = Pr(Y = y|R = 1, S = s).$$

Substituting direct estimates into (2.3) we estimate $Pr(Y = y|T = 1)$ by

$$\sum_s (n_{11}^{sy}/n_{11}^s)(n_{11}^s/n_1)[1 + a_1(s^{-1} - 1)] = (1/n_1)\sum_s n_{11}^{sy}[1 + a_1(s^{-1} - 1)] \tag{2.5}$$

Denote the set of item responders by IR. Then (2.5) can be seen as a weighted mean over the set IR, of the indicator variable $I[Y = y]$, with the individual weight $w_{1i} = 1 + a_1(s_i^{-1} - 1)$, where s_i is the value of S for item responder i. From (2.4) it is clear that $\sum_{IR} a_1(s_i^{-1} - 1) = n_{10}$, and hence $\sum_{IR} w_{1i} = n_1$

Note that $a_1(s_i^{-1} - 1)$ can be thought of as the number of item NRs 'represented' by item responder i.

2.3.3 Adjustment for Unit Non-Response

Removing the conditioning on T from (2.2) we see that within one Z class:

$$\theta = Pr(Y = y) = \sum_s Pr(Y = y|S = s)Pr(S = s) \tag{2.6}$$

where

$$\begin{aligned} Pr(S = s) &= Pr(S = s|T = 1)Pr(T = 1) + Pr(S = s|T = 0)Pr(T = 0) \\ &= Pr(S = s|R = 1)Pr((T, R) = (1, 1)) + Pr(S = s|R = 0)Pr((T, R) = (1, 0)) \\ &\quad + Pr(S = s|T = 0)Pr(T = 0). \end{aligned} \tag{2.7}$$

Under Assumption 2, $Pr(S = s|T = 0) = Pr(S = s|R = 0)$, and so it is clear how adjustment for unit non-response can be made. The weights assigned to the responders to account for item non-response need to be inflated to account for unit non-response also. So the overall weight to correct for item and unit non-response is:

$$w_{2i} = 1 + a_1(s_i^{-1} - 1) + a_2(s_i^{-1} - 1) = 1 + (a_1 + a_2)(s_i^{-1} - 1) \tag{2.8}$$

where a_2 is calculated such that

$$\sum_{IR} a_2(s_i^{-1} - 1) = n_0$$

Hence

$$\sum_{IR} (a_1 + a_2)(s_i^{-1} - 1) = n_0 + n_{10}$$

and

$$\sum_{IR} w_{2i} = n$$

2.3.4 Presentation of Estimates

Suppose we index the Z classes by $\{Z_h : h = 1, 2, ...\}$, and wish to present an overall estimate of $\theta = Pr(Y = y)$ for the domain of interest (which we shall assume does not cross Z class boundaries) defined by the set of Z classes $\{Z_h : h \in H_0\}$. First the estimates are worked out within each Z_h, $h \in H_0$. Then if n_h represents the total number of units

sampled from Z_h, and if θ_h represents the estimate of θ in Z_h, then the overall estimate of θ for the domain is given by the following ratio of sums over $h \in H_0$:

$$\sum_h n_h \theta_h / \sum_h n_h$$

2.4 Variance and Confidence Interval Estimation

The bootstrap provides one potential method of variance and confidence interval calculation. With NATSSAL (see following section) we have taken all units with known Z variables (age, sex, urbanicity), and sampled from these units with replacement. From the new sample the estimate of interest is calculated for the desired Z class. One thousand samples are taken, and hence one thousand estimates calculated. The variance of these thousand can be calculated, and a confidence interval derived by excluding the highest and lowest 2.5% of cases.

In the case where S is categorical, the delta method provides an alternative technique. We have carried out a small simulation study comparing the variance estimates from these two methods with 'true' figures, determined by a separate and larger simulation. We also compared the coverage of the confidence intervals. A single Z class was taken in each case. S was derived from a simple model for item response, containing just one four-level 'enthusiasm to respond' variable, equivalent to the embarrassment variable of NATSSAL. The distributions of the four-level variable amongst item responders and amongst item and unit NRs were also chosen to reflect the range of embarrassment distributions found in NATSSAL among Z classes. The response variable, Y, is taken to be binary with value 0 or 1, and the proportions of $Y = 1$ by propensity class were chosen to reflect how virginity and embarrassment vary in NATSSAL.

Eight separate scenarios were considered, all possible combinations of low and high item non-response, low and high item NRs' enthusiasm, and low and high proportion of $Y = 1$. High item non-response rate was taken as 8% and low as 2%. Item responders' enthusiasm

was characterised as 75% level 1 (highest), 18% level 2, 5% level 3 and 2% level 4 (lowest). High item non-responders' enthusiasm was taken as 40% level 1, 30% level 2, 20% level 3 and 10% level 4, as compared to low enthusiasm, taken as 20%, 20%, 30%, 30% respectively. The low Y rate was taken as 2% in the level 1 enthusiasm group, 3% in level 2, 6% in level 3, and 10% in level 4, as compared to the high rate taken as 15%, 25%, 30%, 30% respectively. The unit non-response rate was set at 25% for all simulations, and the initial sample size set at 1500.

Let β represent the vector of the simple probabilities from which θ is constructed in (2.6) and (2.7), e.g. $Pr(Y = y|S = s)$, and $Pr(S = s|T = 0)$. The variance-covariance matrix of these probabilities, Σ, can be simply calculated, since they are independent of each other, except for the enthusiasm proportions within item responders and within unit and item NRs which are dependent on each other. Then if we write $\theta = f(\beta)$ then the delta method leads to the expression for the variance of $f'(\beta)^T \Sigma f'(\beta)$, where the prime denotes differentiation with respect to β. This estimate was calculated using symbolic differentiation in S-PLUS, and through substitution of the MLEs for the components of β. Symmetric confidence intervals were created using the variance estimate in the usual way based on an assumption of a Normal distribution for the estimator.

From each simulation a bootstrap variance estimate was created by sampling 500 times from the simulated proportions. A confidence interval was also created by removing the highest and lowest 2.5% of cases. We have defined confidence intervals from 500 samples to save computing time, and recommend taking at least 1000 for a single application of the methodology, as we have done with NATSSAL.

Table 2.2 Variance estimates from the delta and bootstrap methods: a simulation study

Item non-response	*Item NRs' enthusiasm*	*Y*	*'True' variance*	*Delta estimate*	*Bootstrap estimate*	*% coverage delta CI*	*% coverage bootstrap CI*	*'True' 95% interval*	*CI mean end points - delta*	*CI mean end points - bootstrap*	*% bootstrap CIs - true value below*	*% bootstrap CIs - true value above*	*% delta CIs - true value below*	*% delta CIs - true value above*
low	*high*	*low*	3.59×10^{-5}	6.67×10^{-5}	4.06×10^{-5}	96.9	94.9	1.83 - 4.18	1.42 - 4.45	1.85 - 4.29	1.4	3.7	0.2	2.9
high	*high*	*low*	3.85×10^{-5}	5.19×10^{-5}	4.00×10^{-5}	95.3	93.6	1.83 - 4.28	1.61 - 4.36	1.87 - 4.30	1.9	4.5	0.3	4.4
low	*low*	*low*	7.25×10^{-5}	9.08×10^{-5}	7.66×10^{-5}	92.4	92.5	1.94 - 5.30	1.59 - 5.15	1.94 - 5.22	0.7	6.8	0.1	7.5
high	*low*	*low*	8.57×10^{-5}	9.65×10^{-5}	9.07×10^{-5}	91.8	92.1	1.91 - 5.55	1.74 - 5.45	2.01 - 5.58	2.1	5.8	0.6	7.6
low	*high*	*high*	1.77×10^{-4}	2.32×10^{-4}	1.87×10^{-4}	96.4	94.3	16.6 - 21.8	16.2 - 22.0	16.6 - 21.9	3.3	2.4	1.4	2.2
high	*high*	*high*	1.81×10^{-4}	2.03×10^{-5}	1.86×10^{-5}	95.8	95.0	16.8 - 22.1	16.6 - 22.2	16.8 - 22.1	2.5	2.5	1.5	2.7
low	*low*	*high*	2.45×10^{-4}	2.88×10^{-5}	2.73×10^{-5}	95.4	94.8	17.1 - 23.3	16.8 - 23.4	17.1 - 23.5	2.7	2.5	1.4	3.2
high	*low*	*high*	2.84×10^{-4}	2.97×10^{-4}	2.93×10^{-4}	95.0	93.6	17.2 - 23.8	17.1 - 23.8	17.3 - 23.9	3.5	2.9	2.1	2.9

For both methods, each simulation was based on an initial sample of 1500, and the mean of 1000 such simulations is displayed in Table 2.2, together with the 'true' variance, estimated from 10000 simulations. The coverage of the confidence intervals for the estimate of $Pr(Y = 1)$ is also presented, based on the 1000 simulations, together with the mean lower and upper bounds. The 'true' 95% interval presented is simply the range from the 2.5 percentile to the 97.5 percentile of the 10000 simulations.

The bootstrap method is seen to estimate the variance of the estimator more accurately than the delta method in all cases. However the coverage of the bootstrap confidence interval is further removed from 95% than the delta interval in two of our eight scenarios. Nevertheless the bootstrap confidence intervals in every scenario mimic the distribution of the estimator better, in the sense that the cases where the true value is outside the interval are more evenly divided between below the interval and above. In four of the eight scenarios the number of cases where the true value lies above the delta confidence interval is more than ten times the number of cases where the true value lies below. This reflects the lack of symmetry of the distribution of the estimator. In certain scenarios where the numbers of people from which the parameters of the model are to be estimated are large, e.g. high item non-response, low enthusiasm and high Y rate, the delta method performs relatively well in variance estimation, and could be preferred for confidence interval construction also. However, given its drastically worse performance in some scenarios e.g. when item NRs' enthusiasm is high and the Y rate is low, we would in general recommend the use of the bootstrap. As mentioned previously we expect that the bootstrap confidence interval would perform somewhat better if based on more samples.

2.5 Estimates for the NATSSAL Data

For data from the NATSSAL, Table 2.3 shows the unadjusted and new adjusted estimates of the proportion of virgins in each of eight age/sex domains. The unadjusted estimate

Age	Sex	Unadjusted % virgins	Bootstrap 95% confidence interval	Adjusted % virgins	Bootstrap 95% confidence interval
16-24	Male	18.2	16.2 - 20.2	20.4	17.6 - 23.1
25-34		3.14	2.42 - 3.89	4.09	2.93 - 5.44
35-44		3.15	2.40 - 3.94	4.15	2.79 - 5.93
45-59		2.49	1.79 - 3.22	4.22	2.31 - 6.38
16-24	Female	16.3	14.6 - 18.0	18.2	15.9 - 20.7
25-34		2.07	1.58 - 2.54	2.73	1.94 - 3.62
35-44		1.36	0.92 - 1.80	1.63	1.01 - 2.35
45-59		2.49	1.95 - 3.10	2.86	2.05 - 3.77

Table 2.3 Estimated proportion of virgins, ignoring non-response and under the model

of the proportion of virgins is an average of the observed proportions amongst the urban and the rural respondents, weighted by the proportions of units sampled from the domain that are urban and rural. The adjusted estimates are higher in all age and sex groups, since increased virginity is associated with lower propensity scores. In six of the age/sex categories the adjusted estimates are above the upper limit of the unadjusted confidence interval. This demonstrates that adjustment has a substantial impact. The confidence intervals are appreciably wider for the adjusted estimates relative to the unadjusted, which is to be expected. The confidence intervals are calculated from 1000 bootstrap samples.

2.6 Sensitivity

Since the assumptions of the model are untestable, the question of how sensitive the adjusted estimators are to departures from these assumptions is of great importance. Of particular interest is how sensitive estimators are to deviations from Assumptions 1 and 2.

Assumption 1 asserts ignorability given S and Z. However one might reasonably suppose that within each Z class the probabilities of virginity given propensity score amongst item responders, and those amongst item and unit NRs differ. The relationship between the two may be conveniently represented for any given distribution of S by

$$logit(pr(Y = 1|T = 0 \cup R = 0)) = logit(pr(Y = 1|R = 1)) + k$$

The transform can be applied to all Z classes, holding k constant. Since all the item and

unit NRs may be virgins or all not, the data provide no indication as to the value of k. Note that $k = 0$ represents Assumption 1.

Assumption 2 asserts that the distribution of S in the unit NRs is the same as that in the item NRs. In the notation of section 2.3.3,

$$Pr(S = s|R = 0) = Pr(S = s|T = 0).$$

One might suppose that unit NRs have lower/higher propensity scores than item NRs and so want to give extra/less weight to item responders of lower propensity when accounting for unit non-response. Then we can consider estimating $Pr(S = s|T = 0)$ by p_{cs}, where p_{cs} satisfies the relationship:

$$odds(S \leq s|T = 0)/odds(S \leq s|R = 0) = c$$

for $0 \leq s \leq 1$. Note that $c = 1$ represents Assumption 2.

Using NATSSAL data from women aged 16-24 and women aged 25-34, we calculate the estimates of virginity obtained from the proposed method under a set of pairs of k and c, and present them in Table 2.4. The parameters k and c can be seen to determine the strength of non-ignorability. The range -0.4 to +0.4 was chosen as plausible for k, if 3% and 18% virginity are taken to correspond to typical values for those aged over 25 and those aged 16-24. The value k=-0.4 transforms 3% and 18% to 2.0% and 12.8% respectively, and k=+0.4 to 4.4% and 24.7%. The range 0.5 to 2 was chosen as plausible for c. The value c=0.5 corresponds to 14% of unit NRs having S below the 25th percentile of the item NRs, and 40% above the 75th percentile, and c=2 conversely to 40% below the 25th percentile and 14% above the 75th percentile.

The estimates based on the assumptions used in our example in Table 2.3, corresponding to k=0 and c=1, are found in the centre of Table 2.4. The ranges of estimates from our sensitivity analysis are 16.3 - 20.3% for women aged 16-24 and 2.26 - 3.34% for women aged 25-34. For women aged 16-24 this range just includes the unadjusted estimate, 16.3%, but for women aged 25-34 this range does not include the unadjusted estimate, 2.07%.

k	c	Women aged 16-24, %	Women aged 25-34, %
-0.4	0.5	16.3	2.26
-0.4	1	16.7	2.38
-0.4	2	17.0	2.45
0	0.5	17.7	2.55
0	1	18.2	2.73
0	2	18.5	2.82
0.4	0.5	19.3	2.96
0.4	1	20.0	3.21
0.4	2	20.3	3.34

Table 2.4 Estimated proportion of virgins in two age/sex groups under different assumptions

2.7 Extensions and Limitations

In the analysis of the NATSSAL, note that the method can be extended to any sexual behaviour variable, including those with many levels such as number of lifetime partners. However applying the same distribution of S to the item NRs and the unit NRs is somewhat less satisfactory when considering those sexual behaviour variables asked in the self-completion booklet. Almost all the answers to questions in the booklet can be deduced for those not offered the booklet on the basis of answers to earlier questions. Hence the item NRs will have been offered the booklet, and so will not be sexually completely inexperienced. Since the unit NRs may well include such people, they may be expected to have a different distribution of S.

Equally in the analysis of two or more variables, it would be unsatisfactory to apply the distributions of S derived for each set of item NRs to the unit NRs. For consistency the unit NRs should have the same distributions of S when analysing different variables. For example, one possible approach would be to define S on the basis of item non-response to a single key question, or non-response to any or all of a group of key questions. The difficulties in using weighting methods with complex patterns of non-response are discussed by Little (1988). Further work is needed here into techniques of generating a distribution of S for the unit NRs that could be applied across all questions, and into how these can be applied in practice.

Typically only limited Z variables are known for the unit NRs in a survey, and the method can only be directly applied to estimate parameters within Z classes. Furthermore it may arise, as in the NATSSAL, that this limited information is only known for a proportion of the unit NRs, and the analysis excludes those for whom it was not available i.e. implicitly it is assumed that those for whom the information was not available are similar to the others with regard to the variables of interest.

2.8 Discussion and Final Recommendations

In the analysis of survey data, the method of estimation we propose seems to have promise, although suitable enthusiasm variable(s) are required. Furthermore, the method enables the simple calculation, with the bootstrap, of confidence intervals. The bootstrap, however, is a computer intensive approach, and this along with other complexities of the method have led to the development of a simplification in section 3.2. The extension of the method to the analysis of complex surveys and to regression analysis is developed in section 3.3, using a pseudolikelihood based approach.

The assumptions that underlie our proposed method are untestable, and in practical situations there will be a need to consider the range of estimates obtained through a sensitivity analysis. The two parameters proposed for the sensitivity analysis are readily interpretable, however the selection of plausible ranges for them is entirely subjective. Even where suitable enthusiasm variables are available, the estimates and intervals obtained from the approaches of Park and Brown (1994), and Forster and Smith (1998) over a range of models may also be of interest. Estimates from the approach of Baker and Laird (1988) will be of interest in some scenarios, but in others the approach may be considered *a priori* inappropriate. For further discussion of this point see section 3.4.

The inclusion of enthusiasm variables could be a useful tool in future surveys, and may lead to great gains at no extra cost to the surveyor. Further work is needed into the extension

of the method to several questions. Where item non-response is high, large weights can be generated through low predicted probability of item response in our response propensity model. This may be considered inappropriate and can lead to high variance of estimates. Further work is needed to determine the best method to deal with non-response in this scenario. The possibilities of weighting in categories derived from the propensity score or perhaps 'truncating' the weights generated by our approach need to be examined and compared.

Chapter 3

Further Developments in Dealing with Non-ignorable Non-response

3.1 Introduction

In chapter two a method to estimate population parameters, adjusting for possible non-ignorable non-response (NINR), is proposed. Also described is an approach to the analysis of sensitivity to the assumptions required. However there are limitations to the application of the approach due to its complexity and the type of analysis considered. A simplification may be desirable for situations where an analysis requiring programming skills, adequate computing time, and extensive knowledge of the survey is impractical. Furthermore the method needs to be extended to allow for sampling schemes more complex than simple random sampling (SRS), and also to incorporate regression analysis. Such a simplification and an extension of the approach form the second and third sections of this chapter.

The method of chapter two also needs to be compared with the alternative, more general, approach of Baker and Laird (1988), described briefly in section 1.2.2. In the fourth section of this chapter the method of Baker and Laird is described in detail, and the application of the methods to the example UK sexual behaviour survey forms the focus of a comparison.

3.2 A Simplified Method to Deal with Non-ignorable Non-response

3.2.1 Introduction

The response propensity method proposed in chapter two requires a full understanding of the survey and factors affecting non-response, and also considerable computing time and programming skills. This arises from the need to fit a model of item non-response for each variable of interest and the need to use bootstrap interval estimation. Whilst this may not be considered a problem for the analysis when the number of variables of interest is small, often surveys address many issues and the range of analyses required is broad. Therefore the analysis of a survey is often performed by several researchers of differing statistical expertise. Furthermore if a survey dataset is later placed in a data archive, then other researchers with no specialist knowledge of the survey may obtain the data and wish to perform analyses. For these reasons it is desirable to develop methods which adjust for NINR but which also allow both the rapid and simple calculation of estimates and confidence intervals, and for analysis to be performed without a thorough understanding of the pattern of non-response. In the analysis of many surveys, where ignorable non-response is assumed, 'multi-purpose datasets' are created. These consist of data for the unit responders only, together with a single weight variable. Such datasets can be easily supplied to those who wish to analyse the data, and also placed into data archives. The simplification here will focus on how to produce, and subsequently use, such datasets when NINR is suspected.

This simplified response propensity approach is conceptually similar to the full approach, and is based on an assumption of ignorable non-response conditional on a response propensity score, and that the distribution of the score among unit NRs is linked to that of item NRs. In this score, a central role is again played by at least one variable that represents the enthusiasm to respond. The aim is to produce estimates and confidence intervals as close as possible to the full method. In this section, for ease of presentation, the data are

analysed as if the sampling scheme is SRS. In the third section of this chapter the approach is extended to complex sampling schemes.

3.2.2 The Simplified Response Propensity Score

The simplification centres on the nature of the propensity score. For the simplified method, in place of the continuous response propensity score proposed under the full approach of chapter two, a discrete score will be used. Weighting based on the score will be used as under the full approach, but since the score is discrete, weighting will be applied through weighting cells or classes, defined by the levels of the score and key demographic variables. Common weighting classes can be used in all analyses, and this removes the need to fit a complex model of item response for each variable of interest. The selection of classes, and the class membership of all units would be regarded as fixed in advance, so that any variability in the estimates due to class selection or membership would be ignored. Following the approach of chapter two, under this simplified method to deal with NINR the unit NRs could be assigned the same weighting class distribution as the item NRs to one key question. Alternatively, as discussed in section 2.7, where there is more than one variable of interest one might want to use item non-response information on some or all of these variables.

There are several different approaches to formulating weighting classes (equivalently the propensity score). The approaches to the formation of classes on the basis of the variation in unit non-response rates described in section 1.2.1 may be applied here, though the classes would be based rather on item non-response. Whilst some research has been directed to the selection of classes, no precise theory is available (Little and Rubin, 1987). Following the previous chapter, we denote variables collected for all survey units as Z variables, and those recorded only for unit responders as X variables. Classes may be based on X and/or Z variables, but the variable(s) representing the enthusiasm to participate must be included as part of the basis of weighting classes, since non-response will be assumed ignorable given class membership. To follow the full approach of chapter two as closely as possible, classes

may be defined by fitting a model (e.g. logistic regression model) of item non-response. Weighting classes could be defined to be the quantiles of the predicted probabilities of response among the item responders. It may be convenient to categorise any continuous variables before their inclusion in the model. The classes would be defined as subsets of the domains of interest. Weighting classes may in practice be defined by the domain of interest and a small number of additional variables that predict item non-response, including the enthusiasm variable(s). See subsection 3.2.6, where the simplified approach is applied to an example dataset, for clarification of how the domains of interest and weighting classes are derived in practice.

The selection of classes, in particular the number of classes, would however be restricted by the need to have enough item responders in each class to estimate the distribution of any variable of interest within each weighting class. In another context, other authors suggest that five classes within each domain of interest may suffice (Rosenbaum and Rubin, 1984).

3.2.3 Estimation through Weighting Classes

Let K denote weighting class membership (equivalently the discrete propensity score), and index the set of weighting classes by $\{k : k = 1, 2, ...\}$. Let D denote domain membership, and index the set of domains of interest by $\{d : d = 1, 2, ...\}$. In practice $\{k\}$ may be chosen so that the classes do not cross domain boundaries, and form mutually exhaustive subsets of the domains. Domains of interest would most naturally be defined in terms of Z variables alone. For example, $\{d\}$ may be defined as all cross-classifications of the Z variables (categorised if necessary), e.g. sex and age groups, and hence equivalent to the 'Z classes' described in chapter two. Denote the variable of interest as Y. Consider first estimation within one domain of interest.

Consider $\theta_y = \Pr(Y = y)$ to be the parameter we wish to estimate. Define also $\theta_{yk} = \Pr(Y = y|K = k)$, then

$$\theta_y = \sum_k \theta_{yk} w_k \tag{3.1}$$

where $w_k = \Pr(K = k)$. An estimator of θ_y can be formed from estimates of θ_{yk} and w_k following the form of (3.1). Following the notation of chapter two, define n as the number of units sampled from the domain, n_1 as the number of unit responders, n_0 as the number of unit NRs, n_{10} as the number of item NRs, and n_{11} as the number if item responders. Define further n_1^k as the number of unit responders of class k, n_{11}^k as the number of item responders of class k, n_{10}^k as the number of item NRs of class k, n_{11}^{ky} as the number of item responders of class k and $Y = y$, and n_{10}^{ky} as the number of item NRs of class k and $Y = y$.

Under the assumption of ignorability, then θ_{yk} can then be estimated without bias from the observed proportion amongst the item responders of class k, i.e.

$$\widehat{\theta}_{yk} = \frac{n_{11}^{ky}}{n_{11}^k}$$

However the $\{w_k\}$ must also be estimated in some way. The definition of the $\{k\}$ is determined in part by the enthusiasm variable which is not recorded for unit NRs. Hence the $\{w_k\}$ can only be estimated under an assumption about the distribution of K amongst the unit NRs. Following (2.7)

$$\begin{aligned} w_k \quad = \quad & \Pr(K = k|R = 1)\Pr((T,R) = (1,1)) \\ + \Pr(K = \; & k|R = 0)\Pr((T,R) = (1,0)) + \Pr(K = k|T = 0)\Pr(T = 0). \end{aligned} \tag{3.2}$$

Provided the set of domains, $\{d\}$, is defined in terms of Z variables then the MLEs of the terms of the form $\Pr((T,R) = (1,v))$ and $Pr(T = 0)$ are simply the observed proportions within the units sampled from the domain, i.e.

$$\widehat{\Pr}((T,R) = (1,v)) = \frac{n_{1v}}{n}, v = 0,1$$

$$\widehat{\Pr}(T = 0) = \frac{n_0}{n}$$

Under the assumption that the distribution of K among the unit NRs is the same as that

among the item NRs (Assumption 2 of chapter 2) then:

$$\Pr(K = k|T = 0) = \Pr(K = k|R = 0).$$

The MLE of the term on the right is the observed proportion

$$\widehat{\Pr}(K = k|R = 0) = \frac{n_{10}^k}{n_{10}}$$

and furthermore

$$\widehat{\Pr}(K = k|R = 1) = \frac{n_{11}^k}{n_{11}}$$

hence, substituting sample proportions for the probabilities in (3.2)

$$\widehat{w_k} = \frac{n_{11}^k}{n} + (\frac{n_{10} + n_0}{n}\frac{n_{10}^k}{n_{10}}).$$

Alternatively where there is more than one variable of interest then $\Pr(K = k|T = 0)$ may be estimated by considering the distribution of K amongst the item NRs to a variety of key questions. As discussed in section 2.7, one possibility might be to equate the distribution of K among the unit NRs to that observed among the item NRs to a single key question, or that among the item NRs to one or all of a group of key questions.

The set $\{\widehat{w}_k\}$ can be seen as a set of weights defined at the class level, whilst weighting is typically applied at the individual survey unit level. It is important to note that whilst $\{\widehat{w}_k\}$ is not variable of interest specific, the corresponding individual weights would need to be in order to generate the estimates (3.1) because of differential item non-response.

Let the subscript kj indicate the jth unit of weighting class k. Define $I_{kj}(g)$ to be an indicator function taking value 1 when the statement g is true for unit kj, and 0 when false. Then denoting the variable-specific individual weights by $\{\widehat{w}_{kY}\}$, (3.1) can be adapted to show that the estimator takes the form

$$\widehat{\theta}_y = \frac{1}{\sum_k \sum_{j=1}^{n_{11}^k} \widehat{w}_{kY}} \sum_k \sum_{j=1}^{n_{11}^k} I_{kj}(Y = y)\widehat{w}_{kY} \tag{3.3}$$

where $\widehat{w}_{kY} = \widehat{w}_k/n_{11}^k$ (in which case the first term in (3.3) is simply 1) or some constant multiple of this, and the summations are over the item responders only. The sum of the

weights over the item responders within domains of interest must be scaled so that estimates across domains can be simply computed. One possibility then is to ensure that the sum of the weights across the item responders in each domain equals the number of units sampled from each domain. Under this approach the weights can be considered to be the number of sampled units represented by the item responder. Hence the expression in (3.3) can be seen as an estimate of the proportion of units for which $Y = y$ among all those sampled.

Often under the standard weighting strategy for ignorable non-response the individual weights assigned are not variable specific and are based only on unit non-response. The effect of item non-response is ignored so item NRs contribute nothing to estimation, and the weights for item responders are unaffected by item non-response. However this is usually performed because item non-response is assumed to be approximately independent of Z, X, and Y. Then it may be assumed that the resulting bias is negligible, particularly where item non-response is very low. For the simplified analysis to deal with NINR, however, variable specific weighting should be performed for analysis. Since item non-response is not thought independent of Z and X, and the weights may well be highly variable under our approach, it may have a large effect.

3.2.4 Construction and Use of Multi-purpose Datasets

To analyse the data from surveys, assuming ignorable non-response, 'multi-purpose datasets' are often constructed. These consist of the unit responders' data together with a single set of individual weights. Equipped with such a dataset, users can analyse the data without understanding how the weights were derived. Furthermore the weighting approach leads to quick analyses, for which appropriate functions are available within standard statistical software. This will be of particular benefit where it is desired to involve several statisticians in the performance of many 'routine' analyses. Another possible application arises where a dataset is to be archived and yet accessible to all, as are many datasets including that of the National Survey of Sexual Attitudes and Lifestyles (NATSSAL) at the ESRC Data

Archive at the University of Essex.

The simplified method to adjust for suspected NINR which is proposed in this section may be applied using 'multi-purpose datasets' because it is a weighting based procedure. Furthermore 'multi-purpose datasets' created for the method retain all the benefits of these datasets when used for analysis when non-response is assumed ignorable. In particular, standard survey analysis software may be used to implement the method using only the multi-purpose dataset (see next subsection). One difference however is that variable specific weights must be used, which must be derived from the weights included in the dataset.

One possibility for the weight included in the multi-purpose dataset is simply $\{\widehat{w}_k\}$. Users can then calculate $\{n_{11}^k\}$ and use these to calculate $\{\widehat{w}_{kY}\}$ for the variables of interest. Whilst this entails knowledge of the weighting classes, it requires no knowledge of the derivation of the classes or of $\{\widehat{w}_k\}$, and in any case an algorithm could be supplied with which the user could calculate $\{\widehat{w}_{kY}\}$ for each variable of interest.

3.2.5 Confidence Interval Construction

In chapter two the use of the bootstrap for confidence interval construction under the propensity to respond score approach is advocated. The delta method provides an alternative method to calculate confidence intervals where the score is categorical, i.e. when weighting classes are used as in the simplified approach here. In section 2.4 the delta method was however found inferior, not least because resulting confidence intervals are symmetric, which on the basis of simulation results does not always reflect the distribution of the estimator. The delta method was also found to generate variance estimates that were appreciably larger than the 'true' figures in some of the scenarios. However the bootstrap is a computer intensive method, and the delta method estimate of the variance would require an additional program to be supplied to the user. By contrast, the standard weighting class based variance estimator of survey analysis software (e.g. the 'svyprop' function of STATA Release 5) is quickly computed and can be simply implemented by the end user of

the multi-purpose dataset. Considering one domain, a basic variance estimator for $\widehat{\theta}_y$ under SRS, based on treating $\{\widehat{w}_k\}$ as fixed, is given as

$$\begin{aligned} \widehat{Var}_1(\widehat{\theta}_y) &= Var[\sum_k \widehat{\theta_{yk}}\widehat{w}_k] \\ &= \sum_k \widehat{w}_k^2 Var[\widehat{\theta_{yk}}] \\ &= \sum_k \frac{\widehat{w}_k^2}{n_{11}^k}\widehat{\theta_{yk}}[1-\widehat{\theta_{yk}}] \end{aligned} \tag{3.4}$$

where

$$\widehat{\theta_{yk}} = \frac{1}{n_{11}^k}\sum_{j=1}^{n_{11}^k} I_{kj}(Y=y). \tag{3.5}$$

Clearly in this variance estimator, Y is taken to follow a multinomial distribution, with differing parameters over the weighting classes. In essence the weighting classes are considered to be strata, and would be labelled as such in survey analysis software in order to generate the variance estimator above.

Defining a variable as $\delta_{ykj} = I_{kj}(Y=y)$, then the estimator $\widehat{Var}_1(\widehat{\theta}_y)$ in (3.4) may be alternatively written as

$$\widehat{Var}_1(\widehat{\theta}_y) = \frac{\sum_k \sum_{j=1}^{n_{11}^k} \widehat{w}_{kY}^2(\delta_{ykj} - \widehat{\theta_{yk}})^2}{(\sum_k n_{11}^k \widehat{w}_{kY})^2}. \tag{3.6}$$

In addition to specifying the strata, the user would need to specify when using the software that the individual level weights to be used are $\{\widehat{w}_{kY}\}$. If the weighting classes are not specified as strata then the variance estimator used by the software takes the form

$$\widehat{Var}_2(\widehat{\theta}_y) = \frac{\sum_k \sum_{j=1}^{n_{11}^k} \widehat{w}_{kY}^2(\delta_{ykj} - \widehat{\theta}_y)^2}{(\sum_k n_{11}^k \widehat{w}_{kY})^2},$$

but this would seem less natural than $\widehat{Var}_1(\widehat{\theta}_y)$ since we expect θ_{yk} to vary across the classes $\{k\}$ if adjustment for non-response is to have an effect.

Confidence intervals (CIs) would be constructed in the usual way under an assumption that $\widehat{\theta}_y$ follows a Normal distribution. Resulting intervals will therefore be symmetrical. Where the sample size is small, the logit of a proportion is often thought to follow a distri-

bution closer to the Normal distribution than the proportion itself. By the delta method, asymptotically:

$$Var[logit(\widehat{\theta}_y)] = \frac{Var(\widehat{\theta}_y)}{[\widehat{\theta}_y(1-\widehat{\theta}_y)]^2} \tag{3.7}$$

A variance estimate for $\widehat{\theta}_y$ can be calculated taking the estimator given in (3.4) and then a confidence interval can be calculated for logit$(\widehat{\theta}_y)$ following the form of (3.7). By transforming the lower and upper bounds with the logit^{-1} function, a confidence interval for $\widehat{\theta}_y$ can be generated. This may better reflect the distribution of $\widehat{\theta}_y$, and will be a non-symmetric interval.

A variance estimate for $\widehat{\theta}_y$ can also be calculated using the bootstrap, and this will incorporate the variability due to the estimation of the class weights. This method is however computer intensive.

3.2.6 An Example

Consider the example of estimating the proportion of virgins in the UK, taking data from the NATSSAL, as in section 2.5. For the analysis presented in that section, the domains of interest were based on sex and age-group, information which was available for all units including unit NRs (i.e. sex and age are Z variables). Weighting classes must then be selected to create subsets within these domains, using additional variables related to item response, and including the enthusiasm to respond variable, which in this case is interviewee embarrassment. In the logistic regression model of the odds of item response, presented in Table 2.1 and described in section 2.3.1, embarrassment is seen to be the dominant term in the model. Furthermore the number of classes within each domain that would be created by using the levels of embarrassment to define classes is four, which may be considered adequate. Hence weighting classes are here defined by sex, age-group, and embarrassment.

Consider the proportion of virgins within the domain of men aged 45-59, for which $\widehat{\theta}$ (from both this simplified approach and under the full approach) is 4.2%. Assuming a

Method	Estimated 95% Confidence Interval
Bootstrap (full approach)	2.3 - 6.4%
Bootstrap	2.5 - 6.2%
Simplified estimation	2.6 - 5.8%
Logit estimation	2.9 - 6.1%
Delta method	1.8 - 6.6%

Table 3.1 A comparison of the estimated confidence intervals from different methods

sampling scheme of SRS, to calculate a simplified variance estimate for this domain (3.4), we need only sum over four weighting classes (the embarrassment levels). The resulting confidence intervals from the simplified method, from the logit transformed estimator (3.7), from the delta method, from the bootstrap, and from the bootstrap (following the full approach) are presented in Table 3.1.

For this example, no method is seen to produce an interval approximately equal to that from the bootstrap (following the full approach). The bootstrap (following the simplified approach) is seen to provide the closest match. As seen in the simulation study the delta method has produced a variance estimate that is too large, and the simplified and logit methods have produced variance estimates that are too small since the variability of the class weights is ignored.

3.2.7 Discussion

The main simplification proposed in this section is the use of a categorical response propensity score in order to form weighting classes, rather than defining the weights as the inverse of a continuous score. Whilst categorising in this way can be regarded as discarding some information, this simplified approach will still allow adjustment for suspected NINR, and may produce broadly similar estimates to the approach of chapter two. The standard error of the estimates from the simplified approach may however be smaller, particularly when item non-response is high.

The variance estimator from standard survey analysis software, under the simplified response propensity approach, is inferior to the bootstrap, since it does not reflect the full

complexity of the approach nor the asymmetric distribution of the estimates. However the ease with which this estimator can be computed may be considered to offset this disadvantage. That analyses allowing for NINR can be performed routinely by several statisticians unfamiliar with the details of the survey may be considered a strong advantage of the simplified approach.

3.3 Extension to Complex Sampling Schemes and to Regression Analysis

3.3.1 Introduction

Thus far the development of a response propensity method to deal with NINR has been limited to consideration of the direct estimation of simple population parameters and has assumed that the sampling scheme is SRS. Hence the method remains of somewhat limited practical use. In this section we develop an extension to the more realistic requirements of regression analysis and the analysis of complex surveys.

In most national population surveys the sampling scheme is more complex than SRS. The sampling designs include stratification, the selection of units in multiple stages (e.g. wards, then households), and differential unit selection probabilities. These schemes are described as complex sampling schemes. Strata may be incorporated into the sampling design to ensure that a sufficient number of units are recruited from each of a group of domains, and possibly to increase efficiency for the estimation of key parameters. A multi-stage design is used primarily because a single stage design is impossible, or to save costs. In face-to-face interviewing, sampling addresses in clusters (e.g. wards, post-code sectors) will clearly be a cheaper way of obtaining a given number of interviews than SRS. For example in the NATSSAL the primary sampling units were electoral wards which were selected with probability proportional to size from a list stratified by region. Furthermore since one person was interviewed at each address, the probabilities of selection were not equal across all eligible units. See Wadsworth *et al.* (1993) for further details.

Ignoring any of the aspects of complex sampling designs will lead either to bias or incorrect standard errors. If the multi-stage nature of the sampling is ignored in analysis then generally estimated standard errors will be too small. Equivalently note that there is generally less information in a dataset from a multi-stage sampling design than from one of equal size under SRS. If the stratification is ignored then the estimators of the standard

errors will generally be biased. If the unequal selection probabilities are ignored then biased estimates usually result. The articles within Skinner *et al.* (1989) provide an overview of the issues in analysis raised by complex sampling schemes.

In chapter two and the previous section we have presented techniques to estimate a proportion or equally a distribution of a categorical variable. Interest may also centre on regression analysis, which examines the associations between the variables of interest and various explanatory variables. Obvious possibilities include logistic and linear regression analysis.

3.3.2 Extension to Complex Sampling Schemes

The extension of the full response propensity method of chapter two to complex sampling schemes is conceptually straightforward, but interval estimation is made more complex. The extension of the simplified method described in the previous section to complex sampling schemes is straightforward, and can be routinely implemented provided specialist survey software is available to the end user (e.g. 'svyprop' in STATA release 5). This is so because the adjustment for NINR occurs through weighting, and indeed the $\{\widehat{w}_{kY}\}$ have an intuitive interpretation as the proportion (equivalently the number) of all units represented by that item responder. Under complex survey designs, units are often selected with differing probabilities. For example, often where households are sampled, only 1 person is selected per household for a survey focusing on individual views/behaviours. In this case the relative selection probability is the inverse of the number of eligible individuals in the household. Selection weights, here the number of eligible units in the household, can be considered to be the inverse of the relative selection probabilities. Weights to adjust for both differential unit selection and suspected NINR are developed in this section.

Define Q where $Q = 1$ represents selection, $Q = 0$ otherwise. Then, denoting unit and item response by T and R as in chapter two, note that

$$Pr(Q = 1, T = 1, R = 1) = \Pr(T = 1, R = 1|Q = 1)\Pr(Q = 1). \qquad (3.8)$$

The weights proposed to adjust for NINR in chapter two $\{\widehat{w}_{kY}\}$ are estimates of the inverses of the probability of unit and item response given selection (excluding a constant multiplier). In (3.8) we see the inverse of $Pr(Q=1,T=1,R=1)$ will be the product of the weight for NINR and the selection weight. Hence weights formed by multiplying together the weights for non-response and selection can be considered to be weights for the combined probability of selection and response. These weights can then be used in analysis.

To apply the simplified response propensity approach, using standard survey analysis software, the revised weights, and the primary sampling units (PSU) (e.g. wards) that arise through multi-stage designs would be specified in the usual way. The weighting classes could be specified as strata as in section 3.2.5. The resulting estimates and confidence intervals will then be adjusted for both the complex survey design and the NINR. Software such as 'svyprop' in STATA Release 5, for example, can be used for this purpose. Define m_k to be the number of PSUs in weighting class k, and index the PSUs by $\{g : g = 1, 2, ...\}$. Define the number of item responders of weighting class k and PSU g by n_{11}^{kg}. Define w_{kYj} to be the 'true' inverse selection/response probability of unit j in weighting class k (or a constant multiple of this probability), and $\widehat{w}_{kYj}$ to be an estimate of this calculated as suggested in the previous paragraph. Define Z_k, $\widehat{Z}_k$, Z_{kg} and $\widehat{Z}_{kg}$ by

$$Z_k = \sum_{g=1}^{m_k} \sum_{j=1}^{n_{11}^{kg}} w_{kYj}(\delta_{ykj} - \theta_{yk})$$

$$\widehat{Z}_k = \sum_{g=1}^{m_k} \sum_{j=1}^{n_{11}^{kg}} \widehat{w}_{kYj}(\delta_{ykj} - \widehat{\theta}_{yk})$$

$$Z_{kg} = m_k \sum_{j=1}^{n_{11}^{kg}} w_{kYj}(\delta_{ykj} - \theta_{yk})$$

$$\widehat{Z}_{kg} = m_k \sum_{j=1}^{n_{11}^{kg}} \widehat{w}_{kYj}(\delta_{ykj} - \widehat{\theta}_{yk})$$

Then the variance estimator used by such software, is

$$\widehat{Var}_3(\widehat{\theta}_y) = \frac{\sum_k \frac{1}{m_k(m_k-1)} \sum_{g=1}^{m_k} (\widehat{Z}_{kg} - \widehat{Z}_k)^2}{(\sum_k \sum_{g=1}^{m_k} \sum_{j=1}^{n_{11}^{kg}} \widehat{w}_{kYj})^2}$$

where the revised weights $\{\widehat{w}_{kYj}\}$ now vary even within the weighting classes.

To extend the full response propensity approach to complex sampling schemes, point estimation introduces no extra complexity, as the revised weights would simply be used. However interval estimation may require more programming and computer time. The bootstrap resampling would need to be performed separately within strata, and would also need to reflect the multi-stage nature of sampling, e.g. resample wards, and then units within each ward.

3.3.3 Extension to Regression

The extension to regression also proceeds in a straightforward manner for the simplified response propensity approach because the adjustment for NINR occurs through weighting, and because existing survey analysis software incorporates weighting. These techniques are based upon pseudolikelihood (e.g. 'svyreg', of STATA release 5), about which much more detail is provided in chapter 4, and some discussion of earlier work in section 1.2.2. Briefly, what could be described as a 'weighted' pseudolikelihood (Lawless *et al.*, 1998) is defined as the sum of the likelihood contributions which would arise from each unit under selection/response with equal probability, each multiplied by the corresponding weight. Note that the pseudolikelihood approach enables the use of a wide variety of models, including both linear and logistic regression models. Maximum pseudolikelihood estimates are those values of the parameters that maximise this pseudolikelihood. Variance estimators can be defined in terms of the first and second derivatives of the pseudolikelihood (see chapter 4 and also 'svyreg', STATA Release 5 User Guide, 1997).

The extension of the full response propensity approach to regression is again straightforward for point estimation, but adds complexity to the interval estimation. For point estimation, as with the simplified approach, standard survey analysis software can be used to perform a pseudolikelihood analysis using the weights calculated under the method. Bootstrap resampling would require maximum pseudolikelihood estimates to be computed

for each sampled dataset. This may add considerably to the computing time.

3.3.4 Discussion

The extensions of the simplified response propensity method proposed in section 3.2 to allow regression analysis and to allow for complex sampling schemes are seen to be straightforward, and indeed these extensions may be implemented in standard survey analysis software. The reason for this is seen to be that the adjustment for NINR occurs through weighting, and that interval estimation ignores the variability from weighting class formation and weight estimation. The extensions of the full response propensity approach are conceptually straightforward, again because the approach is based on weighting. However interval estimation by the bootstrap method may become very demanding of computing time. Hence when performing regression analysis of complex survey data, the differences between the full and simplified approach in terms of programming skills and computing time are accentuated relative to simple analyses under SRS. The full approach may nevertheless be recommended for situations where the number of variables of interest is small and adequate computing time is available.

Other approaches to adjust for NINR, unless based on weighting, would not lead to such natural extensions to complex sampling schemes and regression analysis. This may be considered a strong advantage of the response propensity approach.

3.4 Comparison with the Approach of Baker and Laird

3.4.1 Introduction

The response propensity method to adjust for NINR presented in chapter two, and its simplification presented in section 3.2, are applicable only to specific data structures. There must be suitable variables recorded for all unit responders such that the assumption of ignorable non-response, conditional on these factors, can be considered reasonable. These variables must therefore include at least one variable representing the enthusiasm to respond. Furthermore there must also be item non-response for the method to be applied. However a more generally applicable method, mentioned in section 1.2.1, has been proposed by Baker and Laird (1988), and extended by Chambers and Welsh (1993) and Park and Brown (1994). In this section we examine this more general approach, and compare it with the response propensity approach, focusing on the application to an example.

3.4.2 Likelihood and Goodness of Fit

To consider the approach of Baker and Laird, let Y be the variable of interest, and Z the vector of explanatory variables recorded for all units, which we shall assume for simplicity are all categorical. Let T^* indicate response, with the asterisk to denote the combination of unit and item response, though typically the distinction has not been made in the relevant literature.

The likelihood then takes the form

$$L = [\prod_z \prod_y Pr(Y = y, T^* = 1|Z = z)^{m_{zy1}}][\prod_z Pr(T^* = 0|Z = z)^{m_{z+0}}] \qquad (3.9)$$

where m_{zy1} denotes the number of responders for whom $Z = z$, and $Y = y$, and m_{z+0} denotes the number of non-responders (NRs) for whom $Z = z$. The models considered are of the form

$$Pr(Y, T^*|Z) = Pr(Y|Z)Pr(T^*|Y, Z). \qquad (3.10)$$

Log-linear models are fitted for each of the two terms in the structure, a response model, and a margin model for Y, fitted to the margin obtained by summing over responders and NRs. The fitting of the models typically uses an E-M or a Newton-Raphson algorithm, and assumptions that some terms (e.g. higher order interactions) are not required.

Only explanatory variables observed for all cases are included, so in our notation of chapter two, X variables are excluded from both of the models of (3.10). By contrast, in the response propensity approach, the propensity score is based in part on information from X variables. In both approaches the variable of interest is modelled as conditionally independent of response. In the general approach response is non-ignorable, i.e. response is conditionally dependent upon the variable of interest. In the response propensity approach, once a distribution of the propensity score has been assumed for the unit NRs, then response is modelled as conditionally independent of the variable of interest. In the general approach, the information with which to fit non-ignorable response models is derived from the way in which Y and T^* vary together across the levels of the Z variables, and is also dependent on the margin model.

Baker and Laird propose a deviance statistic, G^2 to test goodness of fit, of the form

$$G^2 = -2 \log \frac{l_{FIT}}{l_{FULL}}$$

where l_{FIT} is the likelihood under the model, and l_{FULL} the likelihood under a saturated margin model and the full ignorable response model (e.g. response dependent on the Z variables, including all interaction terms). However as is recognised by Forster and Smith (1998), such a statistic is of limited use. Immediately apparent is that the full ignorable response model and the saturated margin model provide a perfect fit, and this will occur whatever the degree of non-ignorability that would be apparent if the data for the non-responders were present. Forster and Smith suggest that the such goodness of fit statistics should be used informally only to reject those models that lead to a very poor fit of the observed data.

3.4.3 Application to an Example

To clarify the general approach and to perform a limited comparison with the response propensity model we fit some models to our example dataset, taking virginity, V, as the variable of interest. In this example our Z variables again consist of age (grouped for presentation), A, sex, S, and urbanicity, U.

We will take a saturated margin model for V, denoted by $VASU$, since we expect the proportion of virgins to vary across sex/age/urbanicity categories in a complex way. With a saturated margin model, the degrees of freedom available for the T^* model are the same as for the full ignorable T^* model i.e. the number of levels of the cross classification of A, S, and U minus 1, here equal to 15. By removing some of the terms from that full ignorable response model, denoted $ASUT^*$, we can introduce V, creating non-ignorable models.

The E-M algorithm is used to fit these models; for further details of the fitting process see Baker and Laird (1988). Briefly, model fitting starts by making an initial assignment of the NRs to virgin and non-virgin. Then the models of T^* and V are fitted and the NRs of each age/sex/urbanicity class reassigned in the same proportion as in the fitted model, and the process repeated to convergence. We have performed the process initially with two response models, which we feel are the most natural choices. In this context we feel the priority in the choice of response models is to make few assumptions (i.e. include interaction terms) rather than to select a parsimonious model for response. The first model includes main effects of virginity, sex, urbanicity, and age, and all interactions between age, sex, and virginity. We denote this model $ASVT^*/UT^*$. If it is thought that interactions between virginity and urbanicity are important, then the other model to consider is $AUVT^*/ST^*$. These response models both have 13 parameters.

Previous authors have described the convergence of the algorithm as slow, and we agree with that assessment. A criterion of convergence of the value of the likelihood function (3.9) at the MLEs is a natural choice. We aim to estimate the percentage of virgins in age/sex/urbanicity categories in the non-responders to within 0.05% of the MLEs. On the

Sex	Age	Observed % virgins	% virgins SVT^*/AT^*	% virgins resp. prop.
M	16-24	18.2	24.6	20.4
	25-34	3.14	5.16	4.09
	35-44	3.15	5.46	4.15
	45-59	2.49	4.79	4.22
F	16-24	16.3	25.8	18.2
	25-34	2.07	4.56	2.73
	35-44	1.36	3.10	1.63
	45-59	2.49	6.43	2.86

Table 3.2 A comparison of the fitted values from two models

basis of examining the changes in the estimated MLEs and log-likelihood together for some initial models, we decided to judge that convergence had occurred when consecutive values of the log-likelihood differed by less than 10^{-6}.

Both $ASVT^*/UT^*$ and $AUVT^*/ST^*$ converged at the boundary, in the sense that in one or more age/sex/urbanicity class either all or none of the NRs are assigned as virgins. Two models were also fitted without the variable urbanicity, AVT^*, and SVT^*/AT^*. Of these the first also converged on the boundary, but SVT^*/AT^* converged within the boundaries, and the fitted results are presented in Table 3.2, together with the observed proportion among the item responders, and the proportion estimated by the full response propensity approach of chapter two.

The estimates obtained from the SVT^*/AT^* model are higher than those from the response propensity model in all age/sex domains, particularly for women. Nevertheless the estimates would not be considered implausible.

3.4.4 Confidence Interval Construction

Baker and Laird (1988) suggest profile likelihood as a suitable method of confidence interval construction for the fitted values, and propose the use of the E-M algorithm to calculate the values of the likelihood. This is a slow process since for each value of the parameter of interest (here proportion of virgins in a demographic domain) the E-M algorithm must be used to maximise the likelihood holding the parameter of interest fixed, iterating till

convergence. Initial exploration of the profile likelihood suggests that confidence intervals for the estimated proportion of virgins from the SVT^*/AT^* model are much wider than those we present under the full response propensity approach. The approximate 95% confidence interval for the percentage of virgins amongst men aged 45-59 is 1.4 - 13.5%, which in width compares unfavourably with the bootstrap interval from the response propensity approach fitted in chapter two of 2.3 - 6.4%.

3.4.5 Discussion

The assumptions of both the response propensity approach and the more general approach are untestable, but an appropriate sensitivity analysis can be performed in either case. For the response propensity approach a method of performing sensitivity analyses is outlined in section 2.6. For the more general approach perhaps the most natural form of sensitivity analysis would consist of fitting a range different plausible response and margin models that provide an adequate fit to the observed data, and observing how the MLEs of the parameters of interest vary across these models.

However in some scenarios the general approach may be felt to be *a priori* less appropriate. For example one might feel that not only is a non-ignorable response model required, but that the effect of the variable of interest on response differs fundamentally between demographic classes. In this case interaction terms between the demographic variables and the variable of interest are required in the response model. For example, presented with data concerning homosexual experience (the variable of interest), response, and gender, one might well feel that the interaction term between gender and homosexual experience would be needed in the response model. However, with a saturated margin model, this saturated response model would be overparameterised in the approach of Baker and Laird (1988).

In the analysis of surveys with both unit and item non-response, and with suitable enthusiasm to respond variables, the response propensity model is likely to be selected in preference to that of Baker and Laird because it makes use of the specific data structure

and more of the available data. The simplification and extensions described in sections 3.2 and 3.3 are also available under the response propensity approach. Nevertheless, where the approach of Baker and Laird can be considered appropriate, estimates from applying suitable models under the approach may also be of interest.

Chapter 4

Incorporating Retrospective Data into an Analysis of Time to Illness

4.1 Introduction

To analyse the time from an initiating event to illness in an illness-death model (Kalbfleisch and Lawless, 1988) a prospective study design such as the prevalent cohort may be used. Techniques of analysis are well developed (Wang *et al.*, 1993). In this chapter we consider an alternative design, termed the augmented prevalent cohort (APC) study, where a random sample of those alive is recruited and subsequently followed. This can be regarded as a prevalent cohort augmented by 'retrospective' information from units who are ill but alive at recruitment. We aim to demonstrate a benefit from the incorporation of the retrospective information. The extent of this benefit will clearly depend on the efficiency of the techniques of analysis employed. We aim to develop techniques to analyse the time to illness, based on APC studies, that have high efficiency and yet are simple to implement. These techniques require information concerning the time to death.

Our work was initially motivated by the desire to analyse progression to AIDS amongst the women of the UK MRC Collaborative Study of HIV Infection in Women, which can be considered to be an APC study. Further details of the study can be found in The Study Group (1996, 1998). Another motivating example is the need that may arise to investigate the effect of a covariate, not as yet recorded, on progression to AIDS, for which living patients within an HIV seroconversion cohort could be studied.

To aid comparison of the APC and prevalent cohort designs we define prevalent cohort analysis, which when applied to data from an APC study would analyse only data from the nested prevalent cohort study. Regarding then the prevalent cohort study as a sub-study within an APC study, intuitively one might expect the analysis of an APC study to provide higher efficiency than the analysis of the corresponding prevalent cohort study, particularly if the amount of retrospective data is substantial. Nevertheless the analysis of a prevalent cohort study inevitably requires fewer assumptions than analysis of an APC study, and may therefore benefit from a greater degree of robustness.

We investigate the analysis of APC studies through both full likelihood and pseudolikelihood techniques. The full likelihood approach involves specifying a parametric form for the illness, subsequent survival, and death without illness processes, and also for the distribution of initiating event times in the time period of interest. The pseudolikelihood approach involves specifying a parametric form for survival after illness, and the distribution of the initiating event times. For the illness process, a parametric or semi-parametric form may be taken. The pseudolikelihood approach to fitting regression models has been suggested for a variety of studies where selection is outcome-dependent (Prentice, 1986, Kalbfleisch and Lawless, 1988, Wild, 1991, Schill *et al.*, 1993, Scott and Wild, 1997, Breslow and Holubkov, 1997, and Samuelsen, 1997). The approach has also been suggested for the related problem in survey analysis of non-ignorable non-response (Skinner, 1989, Skinner, 1996). A more general overview has been provided by Hu and Lawless (1997) and Lawless *et al.* (1999).

In section 4.6, a small simulation study illustrates the relative efficiency of these methods. In section 4.7, we apply the techniques to an example, a subset of data from the UK Register of HIV Seroconverters.

4.2 Notation and Key Assumption

Let the time of the initiating event be X, and denote time from X to illness, death without illness, or end of follow-up alive and without illness by T. For those units progressing to illness before recruitment or during follow-up, denote the time from $X+T$ to death or end of follow-up alive by D. Let δ_1 and δ_2 together indicate the type of event at $X+T$, $\delta_1 = 1$ representing illness, $\delta_1 = 0$ otherwise, and $\delta_2 = 1$ representing death without illness, $\delta_2 = 0$ otherwise. Hence the three possible scenarios are $(\delta_1, \delta_2) = (1,0)$ if the unit develops illness at time $X+T$, $(\delta_1, \delta_2) = (0,1)$ if the unit dies without illness, and $(\delta_1, \delta_2) = (0,0)$ if the unit is censored alive and illness-free. Let δ_D indicate the type of event at $X+T+D$, $\delta_D = 1$ indicates death, $\delta_D = 0$ indicates end of follow-up. Furthermore let Z represent a vector of relevant covariates. Let $R = 1$ represent inclusion in the study, $R = 0$ non-inclusion. Let the subscript i be used to denote the values of these variables for unit i. Let the fixed calendar time points of earliest initiating event of the units recruited, the latest event, and the start of recruitment be denoted ϕ_1, ϕ_2, and ϕ_E respectively. Figure 4.1. presents these quantities for three hypothetical study units. Unit 1 is dead before entry and so not recruited, unit 2 is recruited already ill, unit 3 is recruited without illness at entry.

To develop a likelihood (and later a pseudolikelihood) we need the following **key assumption**:

- Conditional on Z, the illness, death without illness, and death after illness processes are independent of calendar time. (Note that Z may include treatment variables).

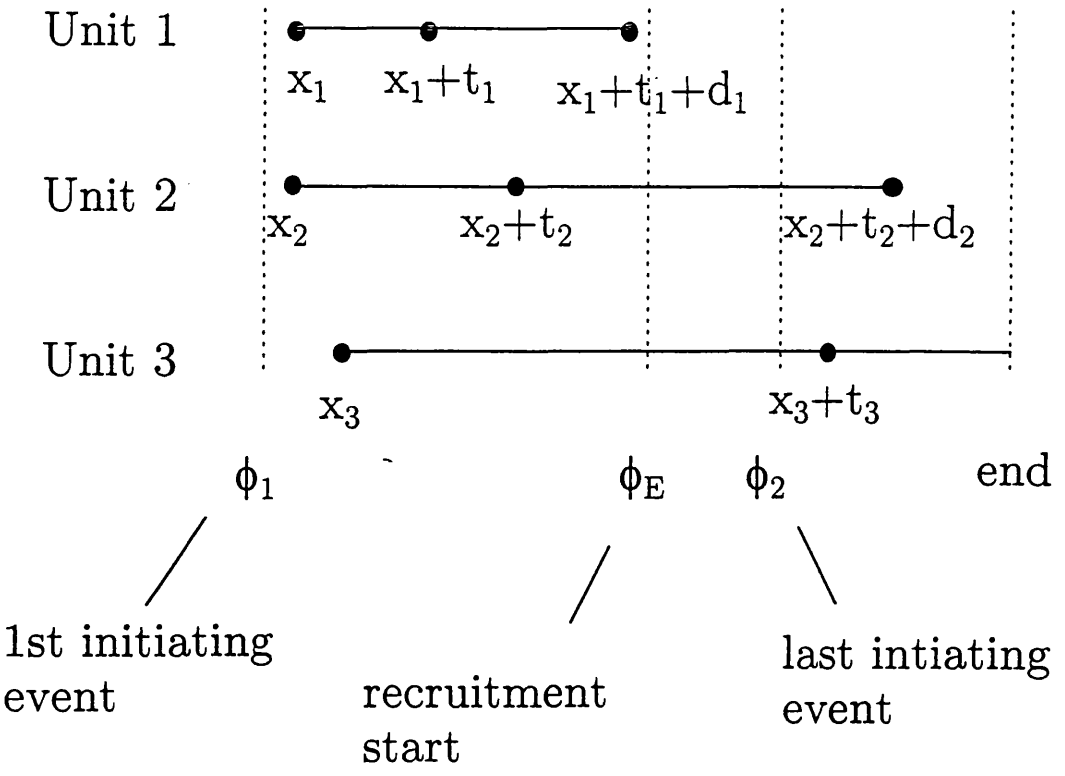


Figure 4.1 Diagrammatic representation of data from three hypothetical units

Models for the processes involved are most easily specified in terms of hazard functions. The three hazards involved are represented in Figure 4.2. Let $\lambda_1(t|Z = z, \varphi)$ denote the hazard of illness, $\lambda_2(t|Z = z, \theta)$ denote the hazard of death without illness. For units progressing to illness, let $\lambda_D(d|Z = z, \eta)$ represent the subsequent hazard of death. A natural choice for the form of the hazard functions is the proportional hazards form, where $\lambda(t|Z = z, \alpha, \beta) = \lambda_B(t|\alpha)\exp(z\beta)$, where $\lambda_B(t|\alpha)$ is the 'baseline' hazard function, and β is the regression parameter of interest. In this case interest will typically focus on a regression parameter β, representing the effect of the variables of interest Z, on the hazard of illness, $\lambda_1(t|Z = z, \varphi)$. This vector of parameters β is a subset of φ. For simplicity in the following sections we treat Z as a vector of time-independent covariates, but all the techniques can be simply adapted to deal with time-dependent covariates. For simplicity we also treat the illness process as independent of the death after illness process, although a parametric dependence of the hazard λ_D on the time to illness can be simply incorporated.

Let $f_X(x|Z = z, \theta)$ represent the distribution of the initiating event times.

4.3 Prevalent Cohort Analysis

In a prevalent cohort analysis, retrospective data are ignored, and the time to illness data are regarded as left truncated, at $T_r = max(\phi_E - X, 0)$. For example, of the units presented in Figure 4.1, units 1 and 2 would be excluded, and unit 3 would be considered left-truncated at time $\phi_E - x_3$. Times to death without illness are treated as censored values of the time to illness. The likelihood, L, takes the following form (Wang *et al.*, 1993):

$$L = \prod_i I(x_i + t_i > \phi_E)\lambda_1(t_i|Z = z_i, \varphi)^{\delta_{1i}} S_{PC}(t_i|Z = z_i, T_r = t_r, \varphi) \qquad (4.1)$$

where

$$S_{PC}(t|Z = z, T_r = t_r, \varphi) = \exp[-\int_{t_r}^{t} \lambda_1(u|Z = z_i, \varphi)du]$$

and $I(g)$ is an indicator function that takes value one when g is true and zero otherwise.

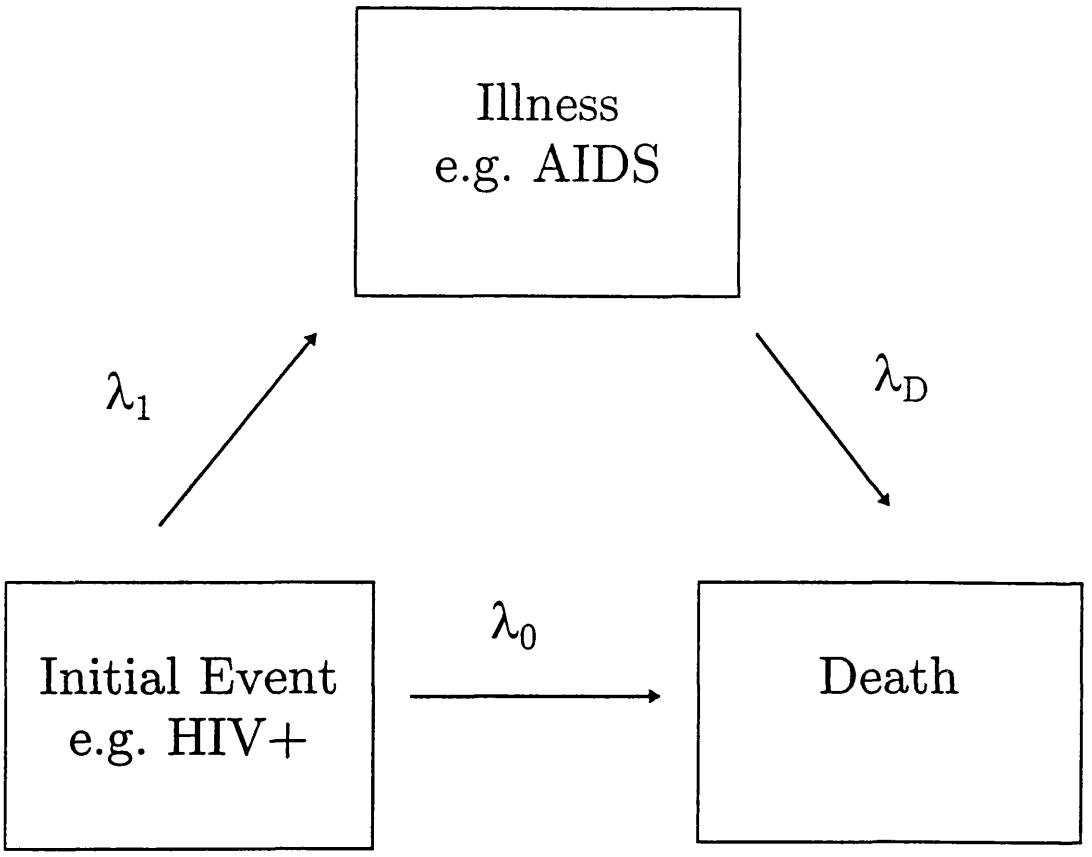


Figure 4.2 The HIV illness death model with corresponding hazard functions

A partial likelihood can alternatively be used in estimation, and this takes the following form:

$$L_{par} = \prod_i I(x_i + t_i > \phi_E) \frac{\delta_{1i} \exp[z_i\beta]}{\sum_j I(t_{rj} < t_i < t_j) \exp[z_j\beta]}$$

This function L_{par} is a function of β alone.

4.4 Full Likelihood Development

The full likelihood takes the following form:

$$L = \prod_i \frac{f_T(t_i|Z = z_i, \varphi, \theta) f_D(d_i|Z = z_i, \eta)^{\delta_{1i}}}{\Pr(R = 1|Z = z_i, \varphi, \theta, \eta)} \tag{4.2}$$

where f_T is the density of T. Similarly f_D is the density of D. The denominator provides the adjustment for the recruitment scheme. The form of f_T following the standard form for 'competing risks' (see Kalbfleisch and Prentice, 1980) is given as

$$f_T(t_i|Z = z_i, \varphi, \theta) = \lambda_1(t_i|Z = z_i, \varphi)^{\delta_{1i}} \lambda_2(t_i|Z = z_i, \theta)^{\delta_{2i}} S(t_i|Z = z_i, \varphi, \theta)$$

where $S(t)$ the survival function is defined as

$$S(t|X = x, Z = z, \varphi, \theta) = \exp[-\int_0^t \lambda_1(u|Z = z, \varphi) + \lambda_2(u|Z = z, \theta) du].$$

The form of f_D takes the standard form for a failure time distribution:

$$f_D(d_i|Z = z_i, \eta) = \lambda_D(d_i|Z = z_i, \eta)^{\delta_{Di}} \exp[-\int_0^{d_i} \lambda_D(u|Z = z_i, \eta) du].$$

The denominator of (4.2), $Pr(R = 1|Z = z)$, takes the form (where ϕ_1 and ϕ_2 are fixed calendar time points as described in section 4.2)

$$\int_{\phi_1}^{\phi_2} \Pr(R = 1|Z = z, X = v) f_X(v|Z = z, \theta) dv$$

where $\Pr(R = 1|Z = z, X = v)$ takes the form (ignoring the dependence on Z):

$$\int_0^{\infty} [\Pr(R = 1|T = u, X = v, \delta_1 = 1)\lambda_1(u) + \Pr(R = 1|T = u, X = v, \delta_2 = 1)\lambda_2(u)] S(u) du \tag{4.3}$$

The recruitment probabilities are defined as

$$\begin{aligned}
\Pr(R &= 1|T=t, X=x, Z=z, \delta_1=1) = \exp[-\int_0^{\max(\phi_E-x-t,0)} \lambda_D(u|Z=z,\eta)du] \\
\Pr(R &= 1|T=t, X=x, Z=z, \delta_2=1) = I(x+t>\phi_E) \qquad (4.4)
\end{aligned}$$

Where it is desired not to make any assumption about the form of $f_X(x|Z=z,\theta)$ then $\Pr(R=1|Z=z,X=x)$ may be used in place of $Pr(R=1|Z=z)$ as the denominator of (4.2), where x is the observed seroconversion time. Limited simulations suggest that analysis without assumptions about $f_X(x|Z=z,\theta)$ is less efficient, but that the loss in efficiency may be small.

L is a function of θ, φ, and η, and clearly takes a complex form, particularly due to the need to evaluate $Pr(R=1)$ for each woman (see (4.3)). With simple parametric forms for the three hazards involved, and perhaps using numerical integration, it would be possible to maximise L using a maximisation routine such as 'nlminb' in S-PLUS, to derive the MLE of φ, the vector of parameters of interest. Another level of complexity would arise however through the need to derive asymptotic variance estimates for the MLE of φ. These would be derived from the information matrix in the usual way. An alternative approach, particularly when one element of φ is of particular interest, might be to compute a profile likelihood based confidence interval.

4.5 Pseudolikelihood Development

An alternative approach to model fitting is through pseudolikelihood and this may be expected to be easier to implement but somewhat less efficient than full maximum likelihood. Our approach to formulating a pseudolikelihood has some similarities to that of Kalbfleisch and Lawless (1988) where they consider the case-cohort study. The main difference is that in an APC study the recruitment probabilities need to be estimated whereas in the case-cohort study the recruitment probabilities are known.

For the pseudolikelihood, a finite population must be defined from which the sample is

taken, and which would provide a suitable analysis if data were available from all members of the population. Here we consider the finite population to be all units with initiating event between times ϕ_1 and ϕ_2. Our approach requires estimation of the distribution of calendar initiating event times in the finite population, $f_X(x|Z = z)$, for which external information could be used if available.

With the notation established in sections 4.2 and 4.4, the log-likelihood contribution from unit i can be written as

$$l_i = \delta_{1i} \log[\lambda_1(t_i|z_i, \varphi)] + \delta_{2i} \log[\lambda_2(t_i|z_i, \theta)] + \log[S(t_i|z_i, \varphi, \theta)]$$

Defining the size of the finite population to be N, the log-likelihood, l_C, that would have been observed if data were available from all units can be written as

$$l_C = \sum_{i=1}^{N} l_i$$

A pseudolikelihood, l_p, estimates l_C through weighting the individual contributions from the recruited units by p_i^{-1} , where p_i is the recruitment probability for unit i, conditional on T, δ_1,and δ_2. The form of l_p is:

$$l_p = \sum_{i=1}^{N} \frac{R_i}{p_i} l_i \tag{4.5}$$

Here proportional hazards models will be used, but note that pseudolikelihood can equally well be used with other models. Consider a model where the intensity (hazard) λ_1 takes the form

$$\lambda_1(t|Z = z, \alpha, \beta) = \lambda_B(t|\alpha) exp(z\beta)$$

where $\varphi = (\alpha, \beta)$, and the form of $\lambda_2(t)$ is left unspecified, though assumed to be functionally independent of β. The form of the likelihood contribution from unit i is given by

$$l_i = \delta_{1i}\{\log[\lambda_B(t_i|z_i, \alpha)] + z_i\beta\} + \delta_{2i} \log[\lambda_2(t_i|z_i, \theta)] + \log[S(t_i|z_i, \varphi, \theta)].$$

where the last term takes the form

$$\log[S(t|z, \alpha, \beta, \theta)] = -\int_0^t \lambda_B(u|\alpha) \exp(z\beta) + \lambda_2(u|z, \theta) du.$$

Considered as a function of α and β, the pseudo-log-likelihood l_p takes the form $const + l_p^*$,

where l_p^*, the pseudo-log-likelihood corresponding to treating death without AIDS as a censoring event, is given by

$$l_p^* = \sum_{i=1}^{N} \frac{R_i}{p_i}[\delta_{1i}\{\log[\lambda_B(t_i|z_i,\alpha)] + z_i\beta\} - \int_0^{t_i} \lambda_B(u|\alpha)\exp(z_i\beta)du]. \tag{4.6}$$

Note that p_i^{-1} can be considered as the number of units from the finite population 'represented' by unit i. Note also that the recruitment probabilities, $\{p_i\}$, must be estimated. They are defined as the average of the recruitment probabilities given in (4.4), over the distribution of X between ϕ_1 and ϕ_2. These are given as

$$\begin{aligned} \Pr(R &= 1|Z=z, T=t, \delta_1=m_1, \delta_2=m_2) = \\ \int_{\phi_1}^{\phi_2} f_X(x|Z &= z)\Pr(R=1|X=x, Z=z, T=t, \delta_1=m_1, \delta_2=m_2)dx \end{aligned} \tag{4.7}$$

All the expressions are evaluated at the MLE of η. One extra probability, $\Pr(R=1|Z=z, T=t, \delta_1=0, \delta_2=0)$, is required. This is the probability of recruitment if the unit 'drops out' or is otherwise censored. In the example in section 4.7, almost all units are followed till study close, and so drop-out can be ignored in analysis. However, in general, it may be necessary to specify a model for the drop-out process, which may or may not depend on T.

The MLE of η is calculated by maximising the likelihood function obtained by treating those who develop illness as a prevalent cohort of times from illness to death, a function that takes a similar form to (4.1), and can be specified as

$$L_D = \prod_i \lambda_D(d_i|Z=z_i,\eta)^{\delta_{Di}} \exp[-\int_{\max(\phi_E - x_i - t_i, 0)}^{d_i} \lambda_D(u|Z=z_i,\eta)du]. \tag{4.8}$$

Information with which to estimate $f_X(x|Z=z)$ may be taken from an external source, or alternatively from the dataset itself. For example, note that for two cohorts defined by two values of X, the number of units from each cohort that would have been recruited from both cohorts can be calculated, based on the follow-up data, i.e. the number of units who would have been alive at study entry if the initiating event were at either value of X. The ratio of the number of such units at two values of X may be used to estimate the ratio of the values of the density of f_X at the two values.

The pseudo log-likelihood, l_p^*, can be differentiated with respect to φ to produce a pseudo-score function, which can be used to calculate maximum pseudo-likelihood estimates (MPLEs).

4.5.1 Parametric Pseudolikelihood

As with full likelihood, any parametric form for λ_1, λ_2 and λ_D can be specified, including in particular, parametric proportional hazards models. An estimator of the asymptotic variance of the MPLEs for this study design can be derived following the procedures proposed by Kalbfleisch and Lawless (1988) and Samuelsen (1997). The expressions in (4.7) provide consistent estimates of the recruitment probabilities, and furthermore they are independent of φ. In the following developments we treat the estimated recruitment probabilities as known, and for simplicity of notation they are denoted as $\{p_i\}$. The effect of treating the estimated recruitment probabilties as known on the variance estimates is considered in section 4.8.

Kalbfleisch and Lawless (1988) present the asymptotic variance estimator as

$$\hat{V}(\widehat{\varphi}) = \hat{A}(\widehat{\varphi})^{-1} + \hat{A}(\widehat{\varphi})^{-1}\hat{B}(\widehat{\varphi})\hat{A}(\widehat{\varphi})^{-1}$$

where

$$\begin{aligned}
\hat{A}(\widehat{\varphi}) &= -\frac{\partial^2}{\partial\varphi\partial\varphi^T} l_p(\varphi)|_{\varphi=\widehat{\varphi}} \\
\hat{B}(\widehat{\varphi}) &= \sum_{i=1}^{N} \frac{1-p_i}{p_i^2} R_i s_i(\widehat{\varphi}) s_i(\widehat{\varphi})' \\
s_i(\varphi) &= \frac{\partial}{\partial\varphi} l_i
\end{aligned}$$

and φ is the vector of parameters of interest.

4.5.2 Semi-Parametric Pseudo-likelihood

To develop a pseudo-likelihood which does not involve the baseline hazard function, we use an approach discussed in Johansen (1983), and undertaken for the case-cohort design by Kalbfleisch and Lawless (1988). First we maximise l_p^* given in (4.6) with respect to the

baseline hazard, $\lambda_B(t)$. This approach is based on the assumption that β is known and finding the (piecewise constant) $\lambda_B(t)$ that maximises l_p^*. Then this $\lambda_B(t)$ is substituted back into l_p^*, and the resulting maximised function, l_{pmax} is a function of β alone. For the APC study design, $l_{pmax}(\beta)$ takes the form

$$l_{p\max}(\beta) = \sum_{i=1}^{N} \frac{R_i \delta_{1i}}{p_i} [z_i\beta - \log \sum_{j=1}^{N} \frac{R_j Y_j(t_i) e^{z_j\beta}}{p_j}] \tag{4.9}$$

where $Y_j(t_i)$ indicates whether unit j is 'at risk' at time t_i (i.e. alive, without illness, and uncensored at t_i).

The complete cohort data log-partial likelihood for the estimation of β as proposed by Cox (1972) takes the form:

$$l_{Cox}(\beta) = \sum_{i=1}^{N} \delta_{1i} [z_i\beta - \log \sum_{j=1}^{N} Y_j(t_i) e^{z_j\beta}] \tag{4.10}$$

where again death without illness is treated as a censoring event.

We can see that $l_{pmax}(\beta)$ (4.9) is an intuitive estimate of $l_{Cox}(\beta)$ (4.10). Furthermore $l_{pmax}(\beta)$, defined for the APC study design, takes a similar form to the corresponding function proposed by Kalbfleisch and Lawless (1988) for the case-cohort design, and that proposed and used in estimation by Samuelsen (1997) for the nested case-control design. For the case-cohort design, Prentice (1986) proposes a pseudo-partial-likelihood of slightly different form.

Differentiation of (4.9) with respect to β produces a score function which is only asymptotically unbiased, although its bias can be expected to be small (Kalbfleisch and Lawless, 1988). Solving the corresponding score equations leads to maximum partial pseudolikelihood estimates (MPPLEs). We establish consistency of the MPPLE, and present an estimator for the asymptotic variance of the MPPLE, briefly outlining the assumptions required. This development has many similarities with the approach of Samuelsen (1997), but the form of the difference between the pseudo-score and the score, and the form of the variance estimator are different due to the different study designs considered. In the development, as in that of the parametric pseudolikelihood, the set of estimated recruitment probabilities,

$\{p_i\}$, is treated as known.

Firstly we note that the expectations of the pseudo-score and pseudo-information are not their complete cohort counterparts. With use of the notation of Samuelsen (1997), we further define

$$S^{(r)}(\beta,t) = \sum_{j=1}^{N} Y_j(t) Z_j(t)^{\otimes r} e^{Z_j(t)\beta},$$

and

$$\tilde{S}^{(r)}(\beta,t) = \sum_{j=1}^{N} \frac{R_j}{p_j} Y_j(t) Z_j(t)^{\otimes r} e^{Z_j(t)\beta}, \tag{4.11}$$

where $r = 0, 1, 2$; and the operator $\otimes$ is defined by $v^{\otimes 0} = 1, v^{\otimes 1} = v$, and $v^{\otimes 2} = v'v$. The pseudo and full cohort log-partial-likelihoods, score functions and information matrices can be represented partly in terms of these $\tilde{S}^{(r)}$ and $S^{(r)}$ functions.

Then $E[\tilde{S}^{(r)}(\beta,t)] = S^{(r)}(\beta,t)$. Under the assumption that

$$\sup_t \frac{1}{N} |\tilde{S}^{(r)}(\beta,t) - S^{(r)}(\beta,t)| \to 0$$

in probability, and of log-concavity of the Cox full cohort partial likelihood, then under standard convergence conditions for Cox's (1972) partial likelihood estimators, the MPPLE is consistent.

To examine the limiting distribution of the pseudo-score $\tilde{U}(\beta)$, first note that

$$\tilde{U}(\beta) - U(\beta) = \sum_{i=1}^{N} \frac{R_i \delta_{1i}}{p_i} [z_i - \frac{\tilde{S}^{(1)}(\beta,t_i)}{\tilde{S}^{(0)}(\beta,t_i)}] - \sum_{i=1}^{N} \delta_{1i} [z_i - \frac{S^{(1)}(\beta,t_i)}{S^{(0)}(\beta,t_i)}].$$

After some manipulation this difference can be shown to equal:

$$\sum_{i=1}^{N} \delta_{1i} z_i (\frac{R_i}{p_i} - 1) + \sum_{i=1}^{N} \frac{\delta_{1i}}{\tilde{S}^{(0)}(\beta,t_i)} \{\sum_{j=1}^{N} [(\frac{S^{(1)}(\beta,t_i)}{S^{(0)}(\beta,t_i)} - \frac{R_i z_j(t_i)}{p_i}) Y_j(t_i) e^{Z_j(t_i)\beta} \frac{R_j}{p_j}]\}.$$

By interchanging the order of summation in the second sum, and evaluating all terms at the true parameter value β_0, the second sum can be shown to equal $\sum_{j=1}^{N} \frac{-R_j W_j}{p_j}$ where:

$$W_j = \sum_{i=1}^{N} [\frac{R_i z_j(t_i)}{p_i} - \frac{S^{(1)}(\beta_0,t_i)}{S^{(0)}(\beta_0,t_i)}] Y_j(t_i) e^{z_j(t_i)\beta_0} \frac{\delta_{1i}}{\tilde{S}^{(0)}(\beta_0,t_i)}$$

and furthermore

$$\sum_{j=1}^{N} W_j = \sum_{i=1}^{N} \delta_{1i} \frac{S^{(1)}(\beta_0,t_i)}{\tilde{S}^{(0)}(\beta_0,t_i)} (\frac{R_i}{p_i} - 1).$$

Hence

$$\tilde{U}(\beta) - U(\beta) = \sum_{j=1}^{N}(1 - \frac{R_j}{p_j})[W_j - \delta_{1j}z_j + \frac{\delta_{1j}S^{(1)}(\beta_0, t_j)}{\tilde{S}^{(0)}(\beta_0, t_j)}].$$

If W_j^* is defined to be:

$$Wj^* = \int [z_j(t) - \frac{S^{(1)}(\beta_0, t)}{S^{(0)}(\beta_0, t)}]Y_j(t)e^{z_j(t)\beta_0}\lambda_B(t)dt,$$

then $W_j - W_j^* \rightarrow 0$, in probability.

Define V_j by

$$V_j = \frac{S^{(1)}(\beta_0, t_j)}{\tilde{S}^{(0)}(\beta_0, t_j)}.$$

Under standard convergence conditions, $N^{-1}S^{(0)}(\beta, t)$ converges to a function bounded away from zero. Then $V_j - V_j^* \rightarrow 0$, in probability, where V_j^* is independent of the sampling, and is defined as

$$V_j^* = \frac{S^{(1)}(\beta_0, t_i)}{S^{(0)}(\beta_0, t_i)}.$$

Under regularity conditions,

$$N^{-\frac{1}{2}}\tilde{U}(\beta_0)$$

and

$$N^{-\frac{1}{2}}U(\beta_0) + N^{-\frac{1}{2}}\sum_{j=1}^{N}(1 - \frac{R_j}{p_j})(W_j^* - \delta_{1j}z_j + \delta_{1j}V_j^*) \tag{4.12}$$

have the same limiting distribution. Furthermore it can be shown that the two terms in (4.12) are uncorrelated.

The covariance of the first term of (4.12), the score, is $\Sigma_N = E\{I(\beta_0)\}$, the information matrix. Consider the covariance matrix of the second term of (4.12):

$$\Delta_N = \text{cov}[\sum_{j=1}^{N}(1 - \frac{R_j}{p_j})(W_j^* - \delta_{1j}z_j + \delta_{1j}V_j^*)] = E\sum_{j=1}^{N}(W_j^* - \delta_{1j}z_j + \delta_{1j}V_j^*)^{\otimes 2}\frac{1 - p_j}{p_j}$$

Under the usual assumptions about convergence to positive semi-definite matrices, $N^{-1}\Sigma_N \rightarrow \Sigma$ and $N^{-1}\Delta_N \rightarrow \Delta$. If the usual Taylor expansion argument holds then $N^{\frac{1}{2}}(\tilde{\beta} - \beta)$ is approximately normal with expectation zero and covariance matrix $\Sigma^{-1} + \Sigma^{-1}\Delta\Sigma^{-1}$. Consistent estimators of Σ and Δ are respectively $N^{-1}\tilde{I}(\tilde{\beta})$ (where $\tilde{I}(\beta)$ is the pseudo-information

matrix) and:

$$N^{-1}\sum_{i=1}^{N} R_i[\tilde{W}_i(\tilde{\beta}) - \delta_{1i}z_i + \delta_{1i}\tilde{V}_i(\tilde{\beta})]^{\otimes 2}\frac{1-p_i}{p_i^2}$$

where

$$\tilde{W}_j(\beta) = \sum_{i=1}^{N}\frac{R_i}{p_i}[z_j(t_i) - \frac{\tilde{S}^{(1)}(\beta,t_i)}{\tilde{S}^{(0)}(\beta,t_i)}]Y_j(t_i)e^{z_j(t_i)\beta}\frac{\delta_{1i}}{\tilde{S}^{(0)}(\beta,t_i)}$$

and

$$\tilde{V}_j(\beta) = \frac{\tilde{S}^{(1)}(\beta,t_j)}{\tilde{S}^{(0)}(\beta,t_j)}.$$

4.6 Simulation Study

A small simulation study was performed to illustrate the relative efficiency of the methods proposed in sections 4.3-4.5, and their dependence on features of the dataset. The data structure is simplified so that death without illness is removed. Thus all units become ill and then die. In all simulations the distribution of X is uniform across the interval 0 to 80, and the start of study recruitment is at time 80. Time 80 is also the end of study recruitment time so that recruitment occurs at one time point only. There is no end of follow-up, all recruited units are followed up till death. The variable Z is a binary time-independent variable, and this defines two equally sized groups in the complete population. This may not be true in the recruited sample. The baseline hazard of illness is defined to be 0.03 over the time interval 0-30, and 0.04 after then. The baseline hazard of death after illness is defined to be either low, which is 0.06 on the interval 0-15 and 0.08 thereafter, or high which is 0.1 on the interval 0-15, and 0.13 thereafter. The simulation of the dataset is then further specified by the effect of Z on the illness process, and on the death process, measured by log-hazard ratios (HR). The different scenarios then have different ratios of the number of cases recruited over the number of cases available for prevalent cohort analysis, which may be expected to have a key impact on relative efficiencies. This ratio is presented in the last column of Table 4.1, where the results are presented from 1000 simulations of each type.

Table 4.1 The relative efficiency of potential methods: a simulation study

Death Haz.	log (HR) illness	log (HR) death	mean - full	mean - pseudo	mean - pseudo (semi)	mean - prev	s.d. - full	s.d. -pseudo (% R.E.)	s.d. -pseudo (semi) (% R.E.)	s.d. -prev (% R.E.)	mean est s.e. - pseudo (% of s.d.)	mean est s.e. - pseudo(semi) (% of s.d.)	ratio of units
low	0	-0.3	0.004	0.004	0.005	0.009	0.1083	0.1214 (89.2)	0.1223 (88.6)	0.1496 (72.4)	0.1163 (95.1)	0.1166 (95.3)	1.53
low	0	0	0.007	0.003	0.003	0.009	0.1101	0.1211 (90.9)	0.1216 (90.5)	0.1478 (74.5)	0.1194 (98.6)	0.1195 (98.3)	1.47
low	-0.4	-0.3	-0.402	-0.402	-0.402	-0.404	0.1056	0.1165 (90.6)	0.1177 (89.7)	0.1397 (75.6)	0.1152 (98.9)	0.1165 (99.0)	1.43
low	-0.4	0	-0.396	-0.395	-0.395	-0.396	0.1068	0.1216 (87.8)	0.1241 (86.1)	0.1368 (78.1)	0.1187 (97.6)	0.1202 (96.9)	1.38
low	-1	-0.3	-1.004	-1.004	-1.005	-1.006	0.1087	0.1202 (90.4)	0.1282 (84.8)	0.1325 (82.0)	0.1174 (97.7)	0.1251 (97.6)	1.30
low	-1	0	-1.001	-1.003	-1.003	-1.004	0.1117	0.1247 (89.6)	0.1330 (84.0)	0.1381 (80.9)	0.1212 (97.2)	0.1296 (97.4)	1.27
high	0	-0.3	-0.001	0.001	0.001	-0.004	0.1104	0.1290 (85.6)	0.1288 (85.7)	0.1517 (72.8)	0.1284 (99.5)	0.1283 (99.6)	1.35
high	0	0	-0.001	-0.001	-0.001	-0.004	0.1154	0.1352 (85.4)	0.1357 (85.0)	0.1568 (73.6)	0.1325 (98.0)	0.1322 (97.4)	1.31
high	-0.4	-0.3	-0.398	-0.395	-0.395	-0.404	0.1107	0.1310 (84.5)	0.1341 (82.6)	0.1418 (78.1)	0.1264 (96.5)	0.1277 (95.2)	1.28
high	-0.4	0	-0.395	-0.397	-0.397	-0.394	0.1088	0.1281 (84.9)	0.1304 (83.4)	0.1409 (77.2)	0.1306 (102.0)	0.1318 (101.1)	1.25
high	-1	-0.3	-1.000	-1.001	-1.000	-1.001	0.1178	0.1322 (89.1)	0.1405 (83.8)	0.1401 (84.1)	0.1274 (96.4)	0.1361 (96.9)	1.20
high	-1	0	-1.003	-1.002	-1.002	-1.005	0.1162	0.1338 (86.8)	0.1456 (79.8)	0.1404 (82.8)	0.1322 (98.8)	0.1418 (97.4)	1.18

Columns 4-7 give the mean estimate of the log(HR) of Z on the illness process. Note that there is no evidence of bias in any method. In columns 8-11 the standard deviations (s.d.) of the estimates of the log(HR) of Z on the illness process are presented, and in columns 9-11 the relative efficiencies (R.E.) of the methods compared to the full likelihood are given. There is some loss of efficiency in the use of the pseudolikelihood, which in these simulations has relative efficiency in the range 84.5-90.5%. The semi-parametric pseudolikelihood approach is seen to have efficiency only slightly less than the parametric pseudolikelihood. The relative efficiency of the prevalent cohort analysis is substantially lower than the pseudolikelihood methods when the proportion of retrospective data is high, since these data are ignored. However when the proportion of retrospective data is low the efficiency is comparable, indicating no benefit from recruitment of retrospective cases, unless full likelihood is used in analysis.

The performance of the variance estimators for the pseudolikelihood techniques is seen to be acceptable. Since the recruitment probabilities are treated as fixed, there is some underestimation of the variance, but this is small. It can be seen in columns 12 and 13 of Table 4.1 that the mean values of the estimated standard error range from 95.1-102.0% of the observed standard deviation.

4.7 Example

As an example we consider data from the UK Register of HIV Seroconverters, for which details can be found in UK Register of HIV Seroconverters Steering Committee (1996, 1998). This register contains data on estimated times of HIV seroconversion, times of first AIDS diagnoses, and times of death together with covariates such as age. Data are censored at the end of 1994. We define a hypothetical study to include only data from those people alive at the start of 1994, so that data from those people in the UK register dying before this point in time are missing. Data are also restricted to cases where the time between last

HIV negative test and first HIV positive test is 36 months or less. Illness is here defined to be AIDS, the initiating event HIV seroconversion, and interest centres on the effect of age on time to AIDS.

Whilst the complete dataset consists of 961 cases, our hypothetical study dataset consists of 878 cases, of which 79 developed AIDS before 1/1/94, and a further 58 developed AIDS during 1994. 46 cases also died during 1994, of which 6 had no AIDS diagnosis.

We have fitted simple models to examine the effect of age at seroconversion on time to AIDS, using semi-parametric pseudolikelihood, and semi-parametric prevalent cohort analysis. For illustrative purposes, the only covariate in our models of the hazard of AIDS, and of death after AIDS is age at seroconversion. The estimates and confidence intervals from the methods are also compared with the estimate and interval derived by the standard maximum partial likelihood approach from the complete register dataset. All likelihood maximisation was performed in S-PLUS using the 'nlminb' function. The following subsections describe the models selected.

4.7.1 Modelling Survival After AIDS

To model survival after first AIDS diagnosis, we use a proportional hazards model with the hazard following a piecewise constant form on the time intervals <400 days, and >400 days. As a covariate we select only age at seroconversion. For our hypothetical study dataset, using (4.8), we find a hazard ratio for an increase of one year in age at seroconversion to be 1.0031.

4.7.2 Modelling the Seroconversion Time Distribution

For the pseudolikelihood we need to model the distribution of seroconversion times, $f_X(x)$. We select a model that assumes that f_X is independent of Z, and also that it takes a piecewise uniform form, on two time intervals, before and after 1/1/1990. This dichotomy is chosen on the basis of the complete register data, and the rate before 1/1/1990 is taken

to be half that afterwards on the basis of the times of seroconversion in the register.

4.7.3 Modelling the Hazard of Illness

We use a proportional hazards model for the time from seroconversion to AIDS. We include a single time-independent covariate, age at seroconversion. The baseline hazard is unspecified, and we follow a semi-parametric approach.

4.7.4 Results

Using the complete dataset, the semi-parametric estimate of the hazard ratio for each year of age at seroconversion is 1.048, with 95% confidence interval 1.033 - 1.064. The estimated recruitment probabilities for the pseudolikelihood range from 0.29 to 1. The MPPLE of the hazard ratio is 1.048, with estimated 95% confidence interval 1.030 to 1.066. The estimated hazard ratio from the prevalent cohort analysis is 1.023 with 95% confidence interval 0.994 to 1.053.

The pseudolikelihood analysis has led to a confidence interval only slightly wider than that from the entire cohort, and substantially narrower than that from the prevalent cohort analysis. For this hypothetical example the benefit from recruiting retrospective cases is great. This is likely to be the case because whilst the proportion of retrospective cases is small (10%) the amount of person-years at risk of AIDS and the number of AIDS events are both much greater for the pseudolikelihood than for the prevalent cohort analysis. The validity of the estimated standard error for the MPPLE is addressed in the following section.

4.8 The Effect of Estimation of the Recruitment Probabilities on the Variance

As reported in section 4.6, the proposed variance estimators for the MPLEs and MPPLEs were seen to perform acceptably well in our simulation study. Nevertheless the proposed

Death Haz.	log(HR) illness	log(HR) death	s.d.	% of s.d. (est. pars)	s.d. -semi	% of s.d. (est. pars)
low	0	-0.3	0.1178	97.0	0.1186	97.0
low	0	0	0.1206	99.6	0.1215	99.9
low	-0.4	-0.3	0.1124	96.5	0.1137	96.6
low	-0.4	0	0.1179	97.0	0.1199	96.6
low	-1	-0.3	0.1185	98.6	0.1256	98.0
low	-1	0	0.1199	96.2	0.1269	95.4
high	0	-0.3	0.1251	97.0	0.1255	97.4
high	0	0	0.1310	96.9	0.1314	96.8
high	-0.4	-0.3	0.1276	97.4	0.1309	97.6
high	-0.4	0	0.1268	99.0	0.1299	99.6
high	-1	-0.3	0.1310	99.1	0.1399	99.6
high	-1	0	0.1289	96.3	0.1406	96.6

Table 4.2 The reduction in variation when the recruitment probabilities are known: a simulation study

variance estimators are slightly deflated due to treating the recruitment probabilities as known. To further investigate this issue, 500 simulations were run alongside those reported in section 4.6. In these simulations MPLEs and MPPLEs were calculated based on 'true' recruitment probabilities, calculated from the known parameters of the death after illness process. The standard deviations of the estimates are presented in Table 4.2. The fifth and seventh columns present the standard deviations as a percentage of those observed when the recruitment probabilities were estimated, as presented in Table 4.1. The variance of the estimates when the recruitment probabilities are known is seen to be only slightly less than the variance when these probabilities were estimated. The standard deviations of the MPLEs and MPPLEs range from 95.4% to 99.9% of those observed when the recruitment probabilities are estimated. This is consistent with the results of the main simulation study and suggests that over the range of scenarios considered, the increase in the variability in the MPPLEs and MPLEs due to the estimation of the recruitment probabilities is minimal.

As another example, and to explore the adequacy of the estimated standard error for the MPPLE of the effect of age on the hazard of AIDS in our hypothetical study in section 4.7, 500 further simulations were performed based on the example dataset. Values of the parameters of the death after AIDS process were simulated from a multivariate Normal distribution with the means taken as the MLEs and using the estimated covariance matrix.

The vector of MLEs and the estimated covariance matrix are

$$\begin{pmatrix} 0.003015 \\ -6.953 \\ -6.307 \end{pmatrix} \text{ and } \begin{pmatrix} 0.000308 & -0.009090 & -0.009971 \\ -0.009090 & 0.327405 & 0.294506 \\ -0.009971 & 0.294506 & 0.366412 \end{pmatrix},$$

where the first row and column refer to the log hazard ratio parameter, the second to the log baseline hazard for time < 400 days, and the third to time > 400 days. The function 'rmultnorm' described in the help for the function 'rnorm' was used to generate the random values of the parameters in S-PLUS. From each simulated set of parameters the recruitment probabilities of all the units in the hypothetical study were calculated and a semiparametric pseudolikelihood analysis performed. The standard deviation of the resulting estimates of the log hazard ratio from this variation in the recruitment probabilities was small but appreciable, and at 1.58×10^{-3} was 18% of the estimated standard error 8.76×10^{-3} used in section 4.7. This variance of the estimate as the recruitment probabilities vary and the time to illness data is fixed is asymptotically equal to the variance of the expectation of the estimate conditional on the recruitment probabilities. This latter term is the inflation factor required to obtain the unconditional variance of the estimate from the variance conditional on the recruitment probabilities.

An 18% inflation factor is clearly not of a magnitude that can be ignored, and represents a very different scenario to those of the main simulation study. The length of follow-up, at one year, is not enough to estimate the parameters of the survival after illness process with sufficient accuracy for this source of variability to be ignored. Nevertheless a clear benefit from the APC design arises, since the estimated standard error inflation factor (18%) is much smaller than the increase (69%) in the standard error from the use of the nested prevalent cohort. An alternative in such cases would be to use external information to estimate the recruitment probabilities. Typically such external information would be available, except where the survival after illness process depends on the covariates considered in the analysis of time to illness. In this hypothetical APC study, the effect of age at seroconversion on survival after AIDS is minimal, and data about survival after AIDS is abundant, so that a survival model could be formulated where the variance of the parameters could be considered

to be negligible. With such external information, the variance estimators proposed in section 4.5 would be considered appropriate. Where no such external information is available, then a jackknife or bootstrap procedure could be recommended for variance estimation.

4.9 Discussion

The augmented prevalent cohort (APC) study design in which a sample of living patients is recruited, whether ill or not, is seen to be of practical use when interest centres on the time to illness. When interest centres on both time to illness and time from illness to death, then the APC design may be particularly appropriate. The benefit from an APC design arises primarily where the initiating event is well-defined, such as birth. In other studies extra assumptions may be required around the timing of the initiating event, the knowledge that the initiating event had occurred (e.g. the need to have had an HIV test), and the loss to follow-up process. This extra complexity and need for assumptions may be considered unacceptable. For this reason we did not proceed with the analysis of the MRC Collaborative Study of HIV Infection in Women, which was our original motivating example. Since the times of HIV seroconversion were unknown in this study, as in many other HIV studies, the times from entry to illness and entry to death have been taken as the outcomes of interest in analysis by other authors (Study Group, 1998).

The analysis of APC studies based on pseudolikelihood techniques has been shown to be feasible and relatively simple to implement. The variance estimators proposed may be considered appropriate when there is a substantial amount of information concerning survival after illness within the study, or when appropriate external information is available. In other scenarios, the variance estimates should be compared with those from a jackknife or bootstrap procedure. The extra complexity of the full likelihood relative to the pseudolikelihood approaches is in most realistic scenarios unlikely to offset the rather modest efficiency gains. However when it is desired not to specify a model for the seroconversion times, a

conditional likelihood can be used. On the basis of limited simulations (not presented) the drop in efficiency, relative to the full likelihood where a model is specified for seroconversion times, may be small.

Our simulation study and our example are indicative of the sort of gain that may be available from an APC study design relative to the corresponding nested prevalent cohort study, i.e. from recruiting those alive with the illness in addition to those alive without the illness. Note, however, that we have not compared the APC design with the prevalent cohort design with the same number of recruits. This comparison is likely to be complex, and the APC design may have greater efficiency in some contexts, but not in general.

The analysis of time to illness from an APC study inevitably requires more assumptions than the analysis of a prevalent cohort study. Hence an APC study could be considered preferable to the 'nested' prevalent cohort study only when the expected efficiency gain is substantial. Furthermore due to the extra assumptions required in the analysis of an APC study, an investigation of sensitivity to the assumptions and/or confirmatory evidence from a prevalent cohort analysis is desirable.

Chapter 5

Discussion

There are conceptual similarities between non-ignorable non-response and outcome-based selection, and we have seen that techniques for the analysis of the two types of study have some overlap, notably in pseudolikelihood based techniques such as those developed in sections 4.5 and 3.3.3. There is, however, a natural difference in the emphasis of the analysis of the two study types. In studies with outcome-based selection, the selection mechanism is generally known or estimable, whereas in surveys non-ignorable non-response is something that may be suspected but without additional information or assumptions, its direction or magnitude cannot be estimated. Hence in the analysis of studies with outcome-dependent selection the goal would naturally be the direct estimation of the parameters of interest, with some attention paid to the impact of distributional assumptions. In the analysis of surveys where non-ignorable non-response is suspected, the focus of the analysis would naturally be an assessment of the sensitivity of the estimates to the assumptions made.

5.1 Non-ignorable Non-response

There are two main approaches to formulating a sensitivity analysis for the problem of non-ignorable non-response. The first approach consists of examining estimates from a range of models which provide adequate fits to the observed data (e.g. Baker and Laird, 1988). The second approach involves specifying appropriate models for the data structure such that the strength of non-ignorability is determined by one or more parameters. The range of estimates obtained as these parameters vary over a range that might be considered

plausible for the study in question can be considered. For the case of item non-response to a continuous variable, Copas and Li (1997) develop an approach in which the degree of non-ignorability is determined by a single parameter. They then consider primarily the issue of sensitivity where the plausible range for the non-ignorability parameter is centred on zero, representing ignorable non-response. For the analysis of categorical data subject to non-response, Forster and Smith (1998) develop an approach where the degree of non-ignorability is determined by a group of parameters about which assumptions could be made. They suggest that a Bayesian approach is suitable, in which a prior distribution can be specified for each parameter. For their approach they state that a standard sensitivity analysis is in general impractical due to the large number of parameters. The response propensity method of chapter two could be described most naturally as an example of this second approach in that two parameters are introduced which determine the strength of non-ignorability. The method is specific to the analysis of surveys where item and unit non-response arise, as would generally be the case. The method takes full advantage of this data structure and the available information in formulating the assumptions about the unit non-responders. For a sensitivity analysis as proposed in section 2.6, plausible ranges for the two parameters are required. These ranges would be centred not at the values that represent ignorable non-response, but on the values that assign the same distribution for the unit non-responders as that imputed for the item non-responders. In all these methods the selection of prior distributions or of plausible ranges for the parameters to determine non-ignorability is clearly entirely subjective, and additional data would be of great value.

The issue of local sensitivity, i.e. sensitivity to very small departures from ignorability, may be of key interest in some cases, perhaps for example when interest centres on the relationship between two variables. In other cases, for example when interest centres on the prevalence of sensitive behaviours, the local sensitivity of key parameters to non-ignorable non-response is obviously high, by definition. Another potential aspect of a sensitivity analysis, which is described by Copas and Li (1997), is to examine what degree of non-

ignorability would cause a significant test result to become non significant.

A key benefit of the approach developed in chapter two arises from the fact that estimation proceeds by weighting the item responders. As is clear in chapter three this leads to natural extensions to complex sampling schemes and to all forms of regression analysis, and also to simplifications that do not require specialist knowledge, advanced programming skills or great computing time. As is also clear from the comparison with the approach of Baker and Laird (1988) in section 3.4, a weighting based procedure avoids boundary solutions. An advantage relative to other procedures that weight the non-responders, e.g. the method of Bartholomew (1961), might be considered to be that all available information can be used to formulate the hypotheses about the non-responders. The main disadvantage of the response propensity approach may be considered to be that it may only be applied to surveys where there is both item and unit non-response, and where information concerning the enthusiasm to participate of unit responders is available.

Future work in the field of analysis where non-ignorable non-response is suspected may follow either of the two broad approaches described. New methods such as the response propensity approach may be developed for specific scenarios. Methods may alternatively attempt to provide a general framework, where the strength of non-ignorability is controlled by a small number of readily interpretable parameters. With regard to the response propensity approach, further work is needed into the optimum method of formulating the response propensity score when there are several variables of interest, and how this may then best be applied in estimation.

The problem of non-ignorable non-response in surveys is not something that will be solved entirely by improving survey designs. Where the relationship between the variables of interest and non-response is broadly understood or at least strong suspicions exist, then strategies can be developed to combat the problem. For example if non-ignorable unit non-response is believed to occur because people with little or no experience of the subject matter of the survey feel that their opinions or experiences are of little interest, then the

introduction to the survey might attempt to reassure such people of their value to the researchers. Equally item non-response may be thought to be non-ignorable because those people with larger values (e.g. larger number of sexual partners in the last year) cannot accurately recall the required information, and so refuse to answer. In this case units can be encouraged to estimate values or place them into ranges if necessary. Another area for further research into survey design is the development of suitable measures of the enthusiasm to participate, a key component of the response propensity approach.

5.2 Response-dependent Selection

In chapter 4 a study design which is termed the augmented prevalent cohort (APC) is considered. Whilst this design has been previously applied where interest centres on both the time to illness and the time from illness to death, the analysis of the time to illness has not involved those units who are ill before study entry. Where the proportion of such units is substantial, then a clear benefit from their incorporation into an analysis of time to illness is demonstrated. The approaches to the analysis of the time to illness from an APC study have some similarities with those developed for other studies where selection depends on the variables of interest. In such studies the full likelihood can in general be written as in section 4.4, where the numerator of each unit's contribution is the joint density of the variables, and the denominator is the probability of selection. Since the parameters of interest feature in both the numerator and denominator, the maximisation of such likelihoods is often complex. Furthermore since the form of all parts of the likelihood must be specified parametrically then mis-specification may be a serious problem. The advantage of full likelihood is maximal efficiency, and yet in many situations the disadvantages of the full likelihood have led to the development of alternative techniques. The aim of such techniques then is to be at least asymptotically unbiased, simpler to implement and yet without a great loss of efficiency relative to full likelihood. The issue of fewer assumptions may also be a motivation for

alternative methods.

For perhaps the most commonly applied study design featuring response-based selection, the case-control study of epidemiology, straightforward analysis is possible based on a logistic model for the variable of interest. Indeed this a major reason for the popularity of the design. The regression parameter representing the effect of covariates on the odds of being a case can be estimated by treating the sample as a prospective sample (i.e. selection based on covariates, not whether case or control), as described by Farewell (1979) and Prentice and Pyke (1979). For the prevalent cohort study, analysis may also proceed in a straightforward manner by truncating the time to illness at entry, as described in section 4.3, and in greater detail by Wang *et al.* (1993).

However in many cases, such simple approaches are not available, and this would seem to be the case with the APC study design, and also with a variety of studies based on two-stage sampling schemes, and other cohort study designs such as the case-cohort. Furthermore, even with the case-control design, once this is extended to incorporate stratification, and if models other than the logistic are fitted, then maximum likelihood based approaches become much more complicated (Scott and Wild, 1997). Pseudolikelihood based techniques have been developed for many response-based selection problems, and that developed in chapter 4 can be seen as a relatively natural extension of these methods to the APC study design. The pseudolikelihood developed in chapter 4 could be described as a weighted pseudolikelihood as could those developed by Kalbfleisch and Lawless (1988) and Samuelsen (1997), and this seems the most natural for the APC design. A key difference however between the APC study and other designs considered such as the nested case-control and case-cohort, is that in the APC design the recruitment probabilities need to be estimated. This leads to additional complications for the variance estimation under the pseudolikelihood approaches. The work in chapter 4 suggests that for the APC design, the variance estimators proposed which ignore the variability due to the estimation of the recruitment probabilities may often be acceptable. However where there is little information concerning survival after illness

within the study, and little appropriate external information available, then the jackknife could be recommended for variance estimation.

Work by Lawless *et al.* (1999) demonstrates that under some study designs various forms of pseudolikelihood may be available, and the variability in efficiency can be quite large. For the analysis of related designs to the APC, e.g. where marginal information is available, a different form of pseudolikelihood may be more appropriate. Furthermore the possibility of developing iterative procedures to maximise the full likelihood may be available, and Scott and Wild (1997) develop such a procedure based on a pseudolikelihood method. Further work into methods to maximise full likelihoods may increase the applicability of this approach, and in some cases remove the need for psuedolikelihood based methods.

Selection based on the variable of interest can provide dramatically greater efficiency than other study designs. Whilst the case-control and choice-based sampling designs provide a natural approach to cross-sectional studies, and designs for cohort studies are relatively well developed, many more designs may be developed in the future. Pseudolikelihood based techniques of analysis provide a very flexible approach to modelling data from studies with response-based selection. Much work remains to determine the relative efficiency of these methods, comparing different approaches to formulating pseudolikelihoods and comparing these methods with alternatives. For example, Samuelsen (1997) recently advocated the use of the weighted pseudolikelihood technique in the analysis of nested case-control studies, in preference to standard techniques. Further work in the field of pseudolikelihood based techniques might focus on developing the robustness of such techniques to model mis-specification. Where the recruitment probabilities must be estimated, as in the APC study design, further work may be directed to the development of variance estimators that are easy to implement and yet reflect all the sources of variability.

References

Anderson, J.A. (1972) Separate sample logistic discrimination. *Biometrika*, **59**, 19-35.

Baker, S.G. (1995) Marginal regression for repeated binary data with outcome subject to non-ignorable non-response. *Biometrics*, **51**, 1042-1052.

Baker, S. G. and Laird, N. M. (1988) Regression analysis for categorical variables with outcome subject to nonignorable nonresponse. *J. Am. Statist. Ass.*, **83**, 62-69.

Bartholomew, D.J. (1961) A method of allowing for 'not-at-home' bias in sample surveys. *Applied Stats*, **10**, 52-59.

Bloch, D.A. and Segal, M.R. (1989) Empirical comparison of approaches to forming strata. *J. Am. Statist. Ass.*, **84**, 897-905.

Borgan, O., Goldstein, L. and Langholz, B. (1995) Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann Stats*, **23**, 1749-1778.

Breslow, N.E. and Holubkov, R. (1997) Maximum likelihood estimation of logistic regression parameters under two-phase outcome-dependent sampling. *J. R. Statist. Soc. B*, **59**, 447-461.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995) *Measurement Error in Nonlinear Models*. London: Chapman & Hall,

Cassel, C.M., Sarndal, C.E. and Wretman, J.H. (1983) Some uses of statistical models in connection with the nonresponse problem. *Incomplete Data in Sample Surveys, Vol. 3.* (eds W.G.Madow and I.Olkin), pp 143-160. New York: Academic Press

Chambers, R. L. and Welsh, A. H. (1993) Log-linear models for survey data with non-ignorable non-response. *J. R. Statist. Soc. B*, **55**, 157-170.

Chapman, D.W. (1976) A survey of nonresponse imputation procedures. *Proc. Soc. Statist. Sect., Am. Statist. Ass.*, 245-251.

Conaway, M.R. (1994) Causal nonresponse models for repeated categorical measurements. *Biometrics*, **50**, 1102-1116.

Copas, A.J., Johnson, A.M. and Wadsworth, J. (1997) Assessing participation bias in a

sexual behaviour survey: implications for measuring HIV risk. *AIDS*, **11**, 783-790.

Copas, J.B. and Li, H.G. (1997) Inference for non-random samples (with discussion). *J. R. Statist. Soc. B*, **59**, 55-95.

Cox, D.R. (1972) Regression models and life tables (with discussion). *J. R. Statist. Soc. B*, **34**, 187-220.

Czajka, J.L., Hirabayashi, S.M., Little, R.J.A. and Rubin, D.B. (1992) Projecting from advance data using propensity modeling: an application to income and tax statistics. *J. Bus. Econ. Statist.*, **10**, 117-131.

Deville, J., Srandal, C. and Sautory, O. (1993) Generalized raking procedures in survey sampling. *J. Am. Statist. Ass.*, **88**, 1013-1020.

Diggle, P. and Kenward, M.G. (1994) Informative drop-Out in longitudinal data analysis. *Applied Statistics*, **43**, 49-93.

Farewell, V.T. (1979) Some results on the estimation of logistic models based on retrospective data. *Biometrika*, **66**, 27-32.

Fay, R.E. (1986) Causal models for patterns of nonresponse. *J. Am. Statist. Ass.*, **81**, 354-365.

Flanders, W.D. and Greenland, S. (1991) Analytic methods for two-stage case-control studies and other stratified designs. *Stats. in Med.*, **10**, 739-747.

Forster, J.J. and Smith, P.W.F. (1998) Model-based inference for categorical survey data subject to non-ignorable non-response. *J. R. Statist. Soc. B*, **60**, 57-70.

Goksel, H., Judkins, D.R. and Mosher, W.D. (1991) Nonresponse adjustments for a telephone follow-up to a national in-person survey. *Proc. Sect. Surv. Res. Meth., Am. Statist. Ass.*, 581-586.

Gong, G. and Samaniego, F.J. (1981) Pseudo maximum likelihood estimation: theory and applications. *Ann Stats*, **9**, 861-869.

Hausman, J.A. and Wise, D.A. (1981) Stratification on endogenous variables and estimation: the Gary income maintenance experiment. *Structural Analysis of Discrete Data*

with Econometric Applications (eds. Manski, C.F., McFadden, D.). Cambridge, Mass: MIT Press.

Heitjan, D.F. (1993) Ignorability and coarse data: some biomedical examples. *Biometrics*, **49**, 1099-1109.

Hoem, J.M. (1985) Weighting, misclassification, and other issues in the analysis of sample surveys of life histories. *Longitudinal analysis of labor market data*, edited by Heckmann, J.J, & Singer, B. Cambridge: Cambridge University Press.

Holt, D. and Smith, T.M.F. (1979) Post stratification. *J. R. Statist. Soc. A*, **142**, 33-46.

Hsieh, D.A., Manski, C.F. and McFadden, D. (1985) Estimation of response probabilities from augmented retrospective observations. *J. Am. Statist. Ass.*, **80**, 651-662.

Hu, X.J. and Lawless, J.F. (1997) Pseudolikelihood estimation in a class of problems with response-related missing covariates. *Can. J. Statist.* **25**, 125-142.

Iannacchione, V.G., Milne, J.G. and Folsom, R.E. (1991) Response probability weight adjustments using logistic regression. *Proc. Sect. Surv. Res. Meth. Am. Statist. Ass.*, 637-642.

Ibrahim, J.G. and Lipsitz, S.R. (1996) Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics*, **52**, 1071-1078.

Johansen, S. (1983) An extension of Cox's regression model. *Int. Statist. Rev.* **51**, 165-174.

Johnson, A. M., Wadsworth, J., Wellings, K. and Field, J. (1994) *Sexual attitudes and lifestyles.* Oxford: Blackwell.

Kalbfleisch, J.D. and Lawless, J.F. (1980) *The statistical analysis of failure time data.* New York: Wiley.

Kalbfleisch, J.D. and Lawless, J.F. (1988) Likelihood analysis of multi-state models for disease incidence and mortality. *Statist. Med* **7**, 149-160.

Kalton, G. (1983) *Compensating for missing survey data.* Ann Arbor: University of

Michigan.

Kalton, G. and Kasprzyk, D. (1986) The treatment of missing survey data. *Survey Meth*, **12**, 1-16.

Keiding, N. (1989) Retrospective estimation of diabetes incidence from information in a prevalent population and historical mortality. *Am. J. Epidem.* **130**, 588-600.

Langholz, B. and Goldstein, L. (1996) Risk set sampling in epidemiologic cohort studies. *Statist. Science*, **11**, 35-53.

Lawless, J.F., Wild, C.J. and Kalbfleisch, J.D. (1999) Semiparametric methods for response-selective and missing data problems in regression. *J. R. Statist. Soc. B*, **61**, 413-438.

Little, R.J.A. (1986) Missing-data adjustments in large surveys. *J. Bus. Econ. Statist.*, **6**, 287-296

Little, R.J.A. (1988) Survey nonresponse adjustments for estimates of means. *Int. Statist. Rev.*, **54**, 139-157

Little, R.J.A. (1993) Pattern-mixture models for multivariate incomplete data. *J. Am. Statist. Ass.*, **88**, 125-134.

Little, R.J.A. (1995) Modelling the drop-out mechanism in repeated-measures studies. *J. Am. Statist. Ass.*, **90**, 1112-1121.

Little, R.J.A. and Rubin, D.R. (1987) *Statistical analysis with missing data.* New York: Wiley.

Little, R.J.A. and Wang, Y. (1996) Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*, **52**, 98-111.

Manski, C.F. and McFadden, D. (1981) Alternative estimators and sample designs for discrete choice analysis. *Structural analysis of discrete data with econometric applications.* (eds. Manski CF, McFadden D) Cambridge, Mass: MIT Press.

Park, T. and Brown, M. B. (1994) Models for categorical data with nonignorable nonresponse. *J. Am. Statist. Ass.*, **89**, 44-52.

Paulino, C.D.M. and Pereeira, C.A.B. (1995) Bayesian methods for categorical data under informative general censoring. *Biometrika*, **82**, 439-446.

Politz, A. and Simmons, W. (1949) An attempt to get the 'not at homes' into the sample without callbacks, Parts I and II. *J. Am. Statist. Ass.*, **44**, 9-31.

Prentice, R.L. (1986) A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73**, 1-11.

Prentice, R.L. and Breslow, N.E. (1978) Retrospective studies and failure time models. *Biometrika*, **65**, 153-158.

Prentice, R.L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika*, **66**, 403-411.

Raab, G.M. (1997) Discussion of 'Inference for non-random samples' by Copas, J.B., and Li, H.G. *J. R. Statist. Soc. B*, **59**, 80-81.

Reilly, M. and Pepe, M.S. (1995) A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*, **82**, 299-314.

Rosenbaum, P.R. and Rubin, D.B. (1983) The central role of the propensity score on observational studies for causal effects. *Biometrika*, **70**, 41-55.

Rosenbaum, P.R. and Rubin, D.B. (1984) Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Statist. Ass.*, **79**, 516-24.

Rubin, D.B. (1987) *Multiplie imputation for nonresponse in surveys.* New York: Wiley.

Samuelsen, S.O. (1997) A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika,* **84**, 379-394.

Schafer, J.L. (1997) *Analysis of incomplete multivariate data.* London: Chapman and Hall.

Schill, W., Jockel, K.H., Drescher, K. and Timm, J. (1993) Logistic analysis in case-control studies under validation sampling. *Biometrika*, **80**, 339-352.

Scott, A.J. and Wild, C.J. (1991) Fitting logistic regression models in stratified case-control studies. *Biometrics*, **47**, 497-510.

Scott, A.J. and Wild, C.J. (1997) Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84**, 57-71.

Skinner, C.J. (1989) Domain means, regression and multivariate analysis. *Analysis of complex surveys*, edited by Skinner, C.J., Holt, D., & Smith, T.M.F., pp 59-87. Chichester: Wiley

Skinner, C.J. and Coker, O. (1996) Regression analysis for complex survey data with missing values of a covariate. *J. R. Statist. Soc. A* **159**, 265-274.

Skinner, C.J., Holt, D. and Smith, T.M.F. (Eds) (1989) *Analysis of complex surveys.* Chichester: Wiley

The study group for the MRC collaborative study of HIV infection in women. (1996) Ethnic differences in women with HIV infection in Britain and Ireland. *AIDS* **10**, 89-93.

The study group for the MRC collaborative study of HIV infection in women. (1998) Survival and progression of HIV disease in women attending GUM/HIV clinics in Britain and Ireland. submitted

UK Register of HIV Seroconverters Steering Committee (1996) The UK Register of HIV-seroconverters, methods and analytical issues. *Epidemiol Infect* **117**, 305-312

UK Register of HIV Seroconverters Steering Committee (1998) The AIDS incubation period in the UK estimated from a national register of HIV seroconverters. *AIDS* **12**, 659-667.

Wacholder, S. and Weinberg, C.R. (1994) Flexible maximum likelihood methods for assessing joint effects in case-control studies with complex sampling. *Biometrics*, **50**, 350-357.

Wadsworth, J., Field, J., Johnson, A.M., Bradshaw, S.A. and Wellings, K. (1993) Methodology of the National Survey of Sexual Attitudes and Lifestyles. *J. R. Statist. Soc. A*, **156**, 407-421.

Wadsworth, J., Johnson, A.M., Wellings, K. and Field, J. (1996) What's in a mean? - an examination of the inconsistency between men and women in reporting sexual partnerships.

J. R. Statist. Soc. A, **159**, 111-123.

Wang, M.C., Brookmeyer, R. and Jewell, N.P. (1993) Statistical models for prevalent cohort data. *Biometrics* **49**, 1-11.

Weinberg, C.R. and Wacholder, S. (1993) Prospective analysis of case-control data under general multiplicative-intercept risk models. *Biometrika*, **80**, 461-465.

Wild, C.J. (1991) Fitting prospective regression models to case-control data. *Biometrika* **78**, 705-717.