On the use of data augmentation to model biomedical data in the presence of selection bias.

Thesis submitted for the degree of

Doctor of Philosophy

At the University of Leicester

by

Jack Bowden B.Sc M.Sc

Department of Health Sciences

University of Leicester

July 2007
Abstract

Conducting biomedical experiments can be exceedingly expensive, therefore it is vital that maximum information is gleaned from existing research and that new experiments are designed as efficiently as possible. Unfortunately, a frequent consequence of efficient design is that trial data no longer constitute an ordinary ‘random’ sample. If not accounted for this can invalidate the analysis, or severely reduce the generalisability of any conclusions. Often statistical methods that model non-randomly selected data can be described theoretically, but involve integrals that are hard to evaluate because of their complexity or unknown analytical form.

Importance sampling is a general, Monte-Carlo method of evaluating integrals by simulation. Although widely used in Physics and Chemistry, it is relatively rarely employed within the biomedical sciences. In this thesis standard importance sampling techniques for likelihood inference are explored, and the properties of the recently proposed conditional likelihood method of Clayton (2003) are investigated and applied to models for biomedical data. This technique is first adapted for models of biased data in a meta-analysis, where its ability to easily cope with complex publication criteria is a clear asset. It is then shown how Clayton’s method can form the second stage of a novel, two-step approach for
analysing ascertainment biased data in genetic epidemiology. Although Clayton’s method works well, it is shown that small adaptations to his importance sampling algorithm can improve the performance further, especially when the number of parameters in the statistical model is large. Bias correction in these two fields is extremely challenging, but it is argued that importance sampling can obtain, with an attractive level of effort, approximate answers when no exact solution exists.
Acknowledgments

I would like to thank Professor John Thompson and Professor Paul Burton for the time and effort they have put in to supervising this work and me as an individual. They have taught me how to properly focus my thoughts and hone my critical faculties, whilst leaving my enthusiasm for research intact. I would also like to thank Professor Nigel Stallard for giving me access to some of his pre-published work and for patiently answering my often confused questions about the nature of bias. Finally, I would like to acknowledge the importance of David Clayton’s 2003 *Biometrika* paper, on influencing my work over the last 3 years. I hope that I have made some contribution towards disseminating its contents to the wider biostatistical audience it deserves.
# Table of Contents

1 Introduction .................................................. 1
   1.1 Overview of thesis ........................................ 7

2 Background .................................................. 10
   2.1 Classical statistical theory ............................ 10
      2.1.1 Models for selection biased data .............. 12
   2.2 Methods for integral approximation ............... 16
      2.2.1 Numerical integration .......................... 16
      2.2.2 Monte-Carlo integration ...................... 19
   2.3 Importance sampling .................................. 22
      2.3.1 Introduction ...................................... 22
      2.3.2 An example of MCML in practice .......... 25
      2.3.3 A conditional likelihood alternative ....... 27
   2.4 Summary .................................................. 32

3 A correction for publication bias in meta-analysis .... 33
   3.1 Introduction ............................................. 34
   3.2 Homocysteine and the MTHFR gene ................. 37
   3.3 Using pseudo-studies to correct for publication bias 40
   3.4 Simulation studies ..................................... 42
      3.4.1 Average performance ......................... 42
      3.4.2 The effect of updating $\theta'$ ............... 44
      3.4.3 Variation between different pseudo-data sets 46
      3.4.4 Misspecification of the selection criterion 47
   3.5 Factors influencing study selection ................. 48
      3.5.1 Extensions to the basic selection model .... 49
   3.6 A simple criterion ..................................... 51
   3.7 Complex criteria ....................................... 53
      3.7.1 A criterion incorporating effect size ....... 53
      3.7.2 Study-specific criteria ......................... 54
   3.8 Discussion ............................................... 59
# TABLE OF CONTENTS

4 Using pseudo-data to validate proposed selection criteria 62
   4.1 Complete selection model misspecification 63
   4.2 Misspecification of threshold criteria 65
      4.2.1 Conditional likelihood inference under selection model misspecification 68
   4.3 Using pseudo-data to rank \( C \) 72
      4.3.1 The Kolmogorov-Smirnov test 72
      4.3.2 A procedure for ranking \( C \) 73
   4.4 Application to the homocysteine data 76
   4.5 Discussion 81

5 Ascertainment bias in genetic epidemiology 86
   5.1 Introduction 87
   5.2 A model for ascertained family disease data 89
      5.2.1 Is Clayton's method appropriate? 90
   5.3 The effect of complete ascertainment for complex models 92
   5.4 A partial solution for GLMMs? 96

6 A two-stage approach to correcting for ascertainment bias in complex models 99
   6.1 Parameter estimation for the ascertained population 100
   6.2 Moving from \( \alpha^*, \theta^* \) to \( \alpha, \theta \) 101
      6.2.1 Starting values for \( \alpha', \theta' \) 103
   6.3 Simulation studies 104
      6.3.1 Simulation of the second stage 104
      6.3.2 Combined simulation of both stages 106
      6.3.3 Further improvement of the population estimates \( \hat{\alpha}, \hat{\theta} \) 108
   6.4 Application to complex models 110
   6.5 Discussion 116

7 Data augmentation 120
   7.1 Improving standard importance sampling 121
      7.1.1 Defensive importance sampling 122
   7.2 Defensive data augmentation 123
      7.2.1 Reverse Logistic Regression 125
   7.3 Combining RLR and conditional likelihood inference 127
   7.4 Simulation studies 129
      7.4.1 Truncated normal example 129
      7.4.2 Affected sibling pairs 131
   7.5 Discussion 135
# TABLE OF CONTENTS

8 Discussion and conclusions 137
  8.1 Meta-analysis ......................................................... 138
  8.2 Genetic epidemiology ................................................ 140
  8.3 Data augmentation .................................................. 142
  8.4 Conclusion ............................................................. 143

Appendices 144

A Additional material 145
  A.1 Computer programs .................................................. 145
    A.1.1 Program code for Chapter 3 ............................... 145
    A.1.2 Program code for Chapter 6 ............................... 149
    A.1.3 Program code for Chapter 7 ............................... 155
  A.2 Supplementary figures ............................................ 160
  A.3 The Li-Mantel estimator ......................................... 164

B Addenda 167
## List of Figures

2.1 *Left:* An illustration of the Trapezoid rule of integration for an unknown function $g(x)$. *Right:* A plot of the 5 most optimal points $s_1, ..., s_5$ with which to base the numerical integral for a standard normal probability density function over the range $(-\infty, \infty)$. .......................................................... 17

2.2 By simulating data $y_1, ..., y_5$ from the distribution $h(y|\theta)$ A Monte-Carlo estimate for $E[Y]$ can be obtained by calculating their mean value, $\bar{y}$. .................. 21

2.3 *(Left)* The accuracy, for varying $m$, of $\left(\frac{1}{m} \sum_{j=1}^{m} \frac{h(y_j|\theta)}{h(y_j|\theta')}\right)^{-1}$ as a Monte-Carlo estimate for $\frac{\frac{\partial}{\partial \theta}}{\frac{\partial}{\partial \theta'}}$ when $\theta' = 0.5$. *(Right)* The value of $\hat{\theta}_m$ that maximises (2.3.5), given $\theta' = 0.5$, $\theta$ is 0. Point estimates and 95% confidence intervals for $\hat{\theta}_m$ are shown, as well as the number of runs failing to converge. ........... 26

2.4 Data structure for Clayton’s algorithm. $y_{i0}$, $i = 1, ..., n$ are from $f(y|\theta)$ and pseudo-data $y_{ij}$, $j = 1, ..., m$ from $f(y|\theta')$. .......................................................... 28

2.5 Conditioning on a covariate $x_{i0}$. $y_{i0}$, $i = 1, ..., n$ from $f(y_{i0}|x_i, \theta)$ and pseudo-data $y_{ij}$, $j = 1, ..., m$ from $f(y|x_i, \theta')$. .......................................................... 31

3.1 The principle of Mendelian randomisation. Observable gene-phenotype and gene-disease relationships can be combined to derive an estimate for the phenotype-disease relationship that is free from confounding and reverse causation. .......................................................... 38

3.2 A funnel plot of the 33 trials reporting a mean difference in homocysteine, as analysed by Minelli et al. (2004). The y axis scale is the inverse each study’s reported standard error, which is also called the ‘precision’. Vertical lines indicate fixed and random effects estimates for the mean difference, unadjusted for publication bias. ................................. 38

3.3 Simulated average performance of the pseudo-data method in estimating $\theta$ when the number of pseudo-values $m = 1, 5, 10$ and the number of studies in the meta-analysis $n = 10, 25, 50$. The true value for $\theta$ is 3. Circles represent the average estimate from 200 simulations and vertical lines represent the average 95% confidence interval widths. ................................. 44
3.4 The 95 percent confidence interval widths for $\hat{\theta}$ in 4000 simulations for $\theta = 3$ when $\theta' = 3.4$ (simulations 1:2000) or $\theta' = 3$ (simulations 2001:4000) Horizontal lines indicate the means. .................................................. 45

3.5 Histograms of the distribution of $\hat{\theta}$ when one data set is analysed with each of 200 different pseudo-data sets. When $\theta' = 3.4$, the average estimate for $\theta$ was 3.00. Consequently when $\theta' = 3$, the average estimate was 3.01. Although this might appear to be less accurate, with respect to the true value of 3, it more accurately reflects the true answer for this particular data set. ...... 46

3.6 Top: A schematic diagram that represents the true relationship between the unknown distribution of the real data ($f_1$ to $f_4$) and the assumed observable study covariates ($p,d$). Bottom: A schematic diagram that represents a simplified, one-to-one relationship between study covariates and complex distributions, that can be dealt with using Clayton's conditional likelihood. ........................................ 57

3.7 Publication bias corrected estimates for the difference in homocysteine and their 95 % confidence intervals when analysed by Trim and Fill, (Duval and Tweedie, 2000), and using pseudo-data with different criteria as described in Sections 3.6, and 3.7. ..................................................... 60

4.1 A Venn diagram representation of two sets $C$ and $C_1$, subsets of all possible values of $y$. Set $C$ can be thought of as containing all data points $y$ that pass the true selection criterion. Set $C_1$ is the set of all data points that pass a proposed selection criterion. ............................................................ 64

4.2 A Venn diagram representation of three sets $C,C_1$ and $C_2$. Set $C$ contains all data points $y$ that pass the true selection criterion. Sets $C_1,C_2$ are the sets of all data points that pass proposed selection criteria. Set $C_1$ is a subset of $C$, which is itself a subset of $C_2$. .................................................. 69

4.3 Top: The effect of selection model misspecification on the resulting maximum conditional likelihood estimates for $\theta$, when criterion $C$ is assumed $\hat{\theta} \approx 3$. When criteria $C_1$ and $C_2$ are assumed, $\theta$ is estimated to be 2.4 and 3.5 respectively. .................................................. 70

4.4 Left: Empirical cdf plots for the real data set (black) and pseudo-data set (red), obtained by correctly guessing the selection criterion. Right: Empirical cdf plots for the real data set (black) and pseudo-data set (blue), obtained by incorrectly guessing the selection criterion. ........................................ 71

4.5 Top: Estimates for $\theta$ obtained by assuming range of possible selection criteria. Middle: The proportion of Kolmogorov-Smirnov two-sided tests rejected, for varying selection criteria. Bottom: The distribution of Kolmogorov-Smirnov D-statistics, for varying selection criteria. ........................................ 75
LIST OF FIGURES

4.6 Top: Publication biased corrected estimates \( \hat{\theta} \) for the mean difference in homocysteine, as a function of the publishable effect size \( T \) and probability of random publication \( \pi_2 \). Bottom: Kolmogorov-Smirnov D-statistics as a function of \( \hat{\theta}|C \) and \( \pi_2 \). ....................................................... 78

5.1 Distributions of linear predictors, \( \eta = \alpha + \xi \), for the 'general' and 'ascertained' population, under a simple GLMM model. ....................................................... 95

6.1 A schematic representation of the two-stage approach. If it is infeasible to take the 'direct' route to the general population parameters \( \alpha, \theta \), it may be possible to estimate the ascertained parameters \( \alpha^*, \theta^* \) (stage 1) and use these estimates along with pseudo-data to estimate \( \alpha, \theta \) (stage 2). ........................................................ 103

6.2 Top: the mean and average 95% confidence interval for \( \alpha, \theta \) obtained from pseudo-data cycles 1 (\( \alpha' = -3.5, \theta' = 5.85 \)), 2 (\( \alpha' = -5.15, \theta' = 4.72 \)) and 3 (\( \alpha' = -5.00, \theta' = 4.51 \)). Bottom: the distribution of 1000 estimates for \( \alpha, \theta \), obtained from the second cycle of pseudo-data generation. ........................................... 105

6.3 100 estimates for the general population parameter \( \alpha = -5 \) obtained with the stage 1 validation step (\( \hat{\alpha}_2 \)) and without (\( \hat{\alpha}_1 \)). The mean estimate is improved from -4.63 to -5.02. ....................................................... 109

6.4 General population parameter estimates for Burton (2003) simulation scenario E, using Clayton's conditional likelihood and assuming perfect information on the ascertained parameters. \( \hat{\alpha} = -6.12, \hat{\beta}_b = 0.67, \hat{\beta}_q = 0.39, \hat{\sigma}_2 = 2.01, \hat{\sigma}_2' = 2.14 \). ........................................... 113

6.5 Computer output showing the empirical variance-covariance matrices of the linear predictors \( \eta \) of the general and ascertained population, under Burton's complex model. ....................................................... 114

7.1 Augmented real data \( y_{i0} \) \( i = 1, \ldots, n \) and pseudo-data \( y_{ij} \) \( j = 1, \ldots, m \). \( \theta \) is an unknown parameter, \( \theta'_{1}, \ldots, \theta'_{m} \) are 'known' parameters that are used to generate the pseudo-data. Note, \( \theta'_{1}, \ldots, \theta'_{m} \) may, or may not be distinct. ........................................... 124

7.2 The bivariate distribution of an epidemiological response in the general (left) and ascertained (right) population of sibling pairs. ....................................................... 132

7.3 Augmented real and pseudo-data, matched through the sharing of a dependent variable, or covariate, \( x \). ........................................................ 132

A.1 The distribution of studies in the four publication categories described in Chapter 3, for varying \( T \). ....................................................... 160

A.2 Publication bias corrected estimates for the between study standard deviation \( \tau \) and their 95% confidence intervals, using pseudo data with different criteria as described in Chapter 3. ........................................... 161
A.3 *Top:* The effect of selection model misspecification on the resulting maximum likelihood estimates for $\theta$, when criterion $C$ is assumed $\hat{\theta} \approx 3$. When criteria $C_1$ and $C_2$ are assumed, $\theta$ is estimated to be 2.4 and 3.5 respectively.161

A.4 *Top:* Average bias corrected estimates $\hat{\theta}$ obtained using equation 4.5.4, assuming an identical simulation scenario for the real data to that of section 4.3.2 (real data simulated from a $N(\theta = 3, 1)$ distribution, conditional on being greater than 2.7). Each value is the result of 1000 independent simulations. The size of the real data set was allowed to vary from 10 to 200.

*Bottom:* The standard deviation of the average estimates for $\hat{\theta}$, as a function of study size. ................................................................. 162

A.5 100 estimates for the general population parameter $\theta = 4.5$ obtained with the stage 1 validation step ($\hat{\theta}_2$) and without ($\hat{\theta}_1$). The mean estimate is improved from 3.93 to 4.51. ................................................................. 163

A.6 The distribution of 300 ascertained population parameter estimates $\hat{\beta}_b, \hat{\beta}_q$ obtained directly from the linear predictors $\eta^*$. Their mean values were 0.73 and 0.42 respectively, a negligible change from their general population true values of 0.7 and 0.4 ................................................................. 163

A.7 The distribution of 300 ascertained linear predictors for parents and siblings under Burton's complex model, suggesting that two distinct parameters $\alpha_p^*, \alpha_s^*$ exist. ......................................................... 164
List of Tables

3.1 The effect that misspecifying the p-value (≤ 0.05) required for publication has on the resulting estimate for θ and its observed standard error (true criteria p ≤ 0.05, true θ = 3) .............................................................................. 47

3.2 Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation, when the publication criterion is assumed to depend on significance. .......................... 53

3.3 Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation, when publication is assumed to be dependent on significance and a minimum effect size T. ............... 54

3.4 Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation, when a study-specific publication criterion is assumed (see Section 3.7 for details) .................. 59

4.1 Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation obtained from the simple publication criterion detailed in Section 3.6, and the Kolmogorov-Smirnov D statistics that they imply. ................................. 79

4.2 Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation obtained from the complex publication criterion detailed in Section 3.7.1, and the Kolmogorov-Smirnov D statistics that they imply. ................................. 79

6.1 Over four scenarios, the average results (standard errors) obtained by analysing 1000 independent data sets, each containing 2000 families of size 5. Columns 3-4 show the average value of $\hat{a}^*$, $\hat{b}^*$ compared to their average estimate $\tilde{a}^*$, $\tilde{b}^*$, obtained by maximisation of the truncated binomial likelihood (6.1.1). Columns 5-6 show how the average estimate for α, θ differ when the exact and estimated values for $\hat{a}^*$, $\hat{b}^*$ are used. The results shown are obtained from a second cycle of pseudo-data generation, for cycle 1 $\alpha' =0.7\alpha$, $\theta' =1.3\theta$. 107
7.1 A comparison of the characteristics of the 3 competing data augmentation algorithms discussed. .................................. 129

7.2 A comparison of 3 competing data augmentation methods performance on the truncated normal example previously described. Column 3 shows the average estimate for $\theta$, column 4, shows the average $\sqrt{\text{mse}}$ for 5000 independent simulations. Coverage calculation assumes asymptotic normality for the distribution of $\theta$ estimates. ........................................ 130

7.3 Parameter estimates (average $\sqrt{\text{mse}}$), and coverages for the complex model using Clayton's method, and our combined approach. Coverages again relate to a nominally 95% confidence interval. ................................. 133
List of abbreviations

CI  Confidence interval.
EM  Expectation-maximisation.
GLM Generalised linear model.
GLMM Generalised linear mixed model.
KS  Kolmogorov-Smirnov.
MCMC Markov chain Monte-Carlo.
MCML Monte-Carlo maximum likelihood.
MLE Maximum likelihood estimate.
MTHFR Methylenetetrahydrofolate reductase.
PDF Probability density function.
RCT randomised controlled trial.
RMSE Root mean squared error.
RLR Reverse logistic regression.
Chapter 1

Introduction

For the majority of scientists, the process of answering a particular research question will involve the collection and analysis of data. By measuring the absolute and relative responses of clinical trial participants, biomedical scientists can obtain empirical evidence to support, or refute, a medical hypothesis. Because they are thought to produce the most reliable evidence, the first type of experimental design a clinician should consider is a randomised controlled trial (RCT). In its most basic form, patients are randomly assigned to either receive an ‘active’ treatment, or no treatment at all, and are then assessed by a common outcome measure. If the number of patients in the trial is large, randomisation ensures that there will be no systematic differences between the active and control groups, with respect to any human characteristic. As a consequence, if there are systematic differences between the two groups with respect to the outcome measure, this difference can be causally attributed to the active treatment, with much more certainty.

RCT’s on human participants are not always possible, for a host of reasons. For example, it may be unethical to conduct an RCT to measure the effects of exposure to a
toxic substance on the risk of disease because there are other known side effects. For this reason RCT's are generally limited to assessing treatments for disease, rather than their causes. However, even when this is the case, it may still be unethical to include a patient into a trial if it is feared that they are not capable of understanding the aim of the trial as well as the potential risks and benefits that inclusion could bring. This is termed 'informed consent'. The Declaration of Helsinki, see Shamoo and Resnik (2003), first proposed in 1964, is a set of ethical principles that those wishing to conduct human experimentation are now duty bound to follow. However, RCT's are often avoided for practical rather than ethical reasons. For example, if the disease in question has a late onset, such as some forms of cancer, it may prove too costly, in terms of time and money, to monitor patients for the length of time needed to observe sufficient cancer cases.

When the 'gold standard' RCT is not appropriate there are a number of alternative trial designs that can be used. One such alternative is a case-control study, in which subjects are initially classified by disease status, and then their exposure to a particular factor is estimated. If a subgroup of the population can be identified that naturally experiences raised levels of a certain exposure, for example factory workers and a particular toxic substance, then a prospective cohort study could be implemented. In this design, subjects with and without the raised exposure are followed forward until they can be classified by disease outcome. For a full review of study designs see Pocock (1983). As is the case with all non-RCT trials, when patients are recruited, or indeed recruit themselves, onto a specific treatment arm, rather than being randomly assigned to one, this can produce a sample of individuals who differ systematically from the hypothesised population, from which they were believed to have come. This can have an unforeseen and undesirable
affect on the results obtained from a subsequent statistical analysis and is often referred to as *selection Bias*.

In a case-control study investigating the association between exposure to a particular factor and a disease, the underlying assumption is that the disease cases and controls come from the same population, differing only by their disease status. It is usually straightforward to recruit disease cases, but deciding on a suitable control group is often more problematic, and can hugely effect the result. Moss et al. (1987) conducted a case-control study into the relationship between sexual promiscuity and AIDS. Alongside a group made up of diagnosed AIDS cases, a control group of non-AIDS sufferers were recruited, from sexually transmitted infection clinics. When the AIDS risk of individuals with less than five sexual partners was compared to those with more than 100, the odds ratio was small and indicative of no association. This was explained by the fact that the control group were chosen by a flawed selection process, not independent of sexual promiscuity. When a more sensible strategy for selecting the controls was implemented, the odds ratio increased dramatically. **Selection bias can stop researchers from detecting an association when it is truly present.**

Glesby and Hoover (1996) discuss a common problem in observational studies that assess the effectiveness of therapies for terminal illnesses, such as cancer. Study subjects who live longer have more opportunities to transfer to different treatment regimens, than those who die early on and, because of this, an artificial correlation between those treatments and improved survival rates can be observed. This is commonly referred to as *survivor selection bias*. Similarly, when treatment related deaths of patients, due to the
side effects of a certain drug, negate the possibility of those same patients experiencing a disease recurrence, this can unfairly give the impression that it is more effective, with respect to disease recurrence, than a less toxic alternative. This phenomenon has been termed competing risk bias, (Kim, 2007). Selection bias can produce spurious associations, when no true association is present.

There will always be ethical and financial factors that preclude scientists from using certain experimental designs. When randomising patients to treatments is not possible or appropriate, there will always be unknown confounding factors that could affect the results, as well as researchers who employ poorly conceived data collection strategies. This means the total elimination of selection bias is not an achievable goal. When bias is present in the data before the point of analysis, the words of R.A Fisher are particularly relevant

"It is a statistical commonplace that the interpretation of a body of data requires knowledge of how it was obtained" - (Fisher, 1934)

In other words, if data have passed a selection criterion, this should be accounted for within a statistical model. Unfortunately, there is substantial scope for researchers to produce a biased analysis of their results, even when there is no inherent bias in the data. When there are number of possible statistical methods that could be performed, and a number of subgroups of the original data that could be analysed in isolation, the selective reporting of only the statistically significant results is common. This is a huge problem in current genetics research, in which for example the association between thousands of mutations and a disease can now be tested simultaneously and instantaneously. The result of selective reporting is a literature containing many false-positive findings, caused by multiple
testing and not any true biological mechanism. Evidence to the extent of this problem is the number of published results that are never reproduced, as noted by Campbell and Rudan (2002), and echoed by Ioannidis (2005) in his paper "Why most published research findings are false".

Whilst it may be true that genuinely positive findings are more exciting than negative ones, when scientific journals selectively print only the positive results, they induce publication bias. The result is a skewed distribution of findings in the public domain that does not represent the true position in a research field, and there is a wealth of empirical evidence that supports its existence, (Stern and Simes, 1997; Sutton et al., 2000). The detrimental affect to science is two-fold; firstly health professionals who read this literature can be lead to make important far reaching decisions based on flawed evidence, secondly, it increases the chance that authors will selectively report their findings in order to be published, (Chan et al., 2004). Francis Bacon perhaps pinpoints the cause of this behaviour when he stated that human beings have an in built mechanism for

"The affirmative or active to effect [us] more than the negative or primitive" -

quoted by Scargle (2000)

As a society we are conditioned to expect biased information to come from 'users' of statistics, such as politicians and marketing agencies, but it is painfully evident that there is a practice of misuse in biomedical research too. Selection bias represents a potentially serious barrier to the advancement of science, and scientists themselves are, in part, to blame.
Efforts have been made by researchers to tackle the problem of selective reporting, with initiatives such as the CONSORT statement, (Begg et al., 1996; Campbell et al., 2004). This statement provides a list of guidelines for reporting results clearly and objectively, and has proved a useful resource for both the authors of clinical trial publications, as well as the editors of medical journals. It can not of course remove the problem of publication bias, but should at least guarantee that published studies are of higher quality. Has it worked? Devereaux et al. (2002), attempted to answer this question by comparing the characteristics of randomised controlled trials published in journals that either did or did not endorse the CONSORT statement. They found that whilst more results were reported in the CONSORT endorsing journals, a positive finding, standards of reporting were still on average far lower than the CONSORT checklists required. My view, as a statistician, is that I can not help to remove the causes of selection bias, by examining the facets of the human condition that lead to bad scientific practice, although this is clearly a worthy endeavor. However, I can help to develop methods for statistical inference that account for its inevitable presence. This is the fundamental aim of my thesis.

There are many other types of bias, which result from processes that are totally distinct from issues of selection, and as such will not be given attention in this thesis. For example recall bias is often thought to be present in case-control studies, because diseased patients are more likely to produce an accurate historical account of their exposure to particular factor than their non-diseased counterparts. A good example of this is lung cancer patients and smoking habits. When recall bias is present, the association between disease and exposure can be overestimated, (Neugebauer and Ng, 1990). Recall bias is a term restricted for use within case-control studies, but because it is essentially an artefact
1.1 Overview of thesis

of differences in measurement error (between cases and controls), it is closely related to the more general concept of regression dilution bias. For example, if a variable such as blood pressure, or cholesterol level, is used in a statistical model to predict the risk of disease for a cohort of subjects, unless the natural variation of this variable within and between each person is taken into account, the results from an analysis are likely to be biased, (Frost and Thompson, 2000). For a definitive list of all currently known biasing processes that can occur in biomedical experiments, as well as a description of their possible effects, see Delgado-Rodriguez and Llorca (2004).

1.1 Overview of thesis

In this thesis I propose methods to analyse two types of biomedical data, each affected by a different form of selection bias. The first relates to family disease data in genetic epidemiological studies, collected in order to determine the relative importance of genetic and lifestyle factors in influencing disease risk. If the genetic component of disease risk is sufficiently large, further research can be sanctioned to find the particular gene or genes responsible. When the disease of interest is rare, recruiting families at random from the general population will yield few sufferers and hence little useful information. A long established solution to this problem is to recruit families who have at least one diseased member; these form an ascertainment population. This is an example of a cost efficient design, but the results of the analysis are of limited use unless they are corrected, so as to be applicable to the general, and not just the ascertained, population.

The second example relates to the process of synthesising the results of multiple stud-
ies in a meta-analysis. By exploiting the joint power of information across lots of small studies, research findings have an increased level of certainty. Obtaining the maximum amount of useful information from existing research can negate the need to conduct further research in a similar vein, so that scarce resources can be better utilised elsewhere. However, because it is often only practical for published studies to be included in a meta-analysis, there is always a serious concern that the published studies over represent positive findings. Strong empirical evidence of precisely this practice has been found by Chan et al. (2004), which appears to make this concern justified.

There is a need to support both areas of research with robust statistical methodology. Methods for bias correction in genetic epidemiology can usually be described theoretically because the sampling mechanism is known, but in practice usually involve complex integrals that are intractable. Methods to correct for publication bias in meta-analysis are hampered by the fact that the selection mechanism generating the published studies is unknown. Importance sampling is a general, Monte-Carlo method for statistical inference. Its popularity stems from the fact that it is often easier to evaluate the expected value of a function, than to find its integral. In this thesis we develop methods that adapt standard importance sampling methods, in order to help model selection biased data in these two areas.

In Chapter 2 selection bias is introduced from a statistical viewpoint. Notation is developed that will be used throughout the thesis. Several methods of integral approximation suitable for correcting for selection bias are introduced. A recently proposed conditional likelihood approach suggested by Clayton (2003), that utilises importance sampling, is
1.1. Overview of thesis

described. In Chapter 3 the problem of publication bias in meta-analysis is reviewed, and it is demonstrated that Clayton's method can be modified for use in this setting. Chapter 4 discusses the idea of selection model misspecification, an important concept to consider in the context of bias correction in meta-analysis. A method to rank competing publication criteria is suggested. In Chapter 5 the problem of ascertainment bias in genetic epidemiology is discussed, and previously proposed methods of correction are reviewed. Chapter 6 describes a novel use of Clayton's conditional likelihood approach to correct for ascertainment bias. In Chapter 7 the theory of methods for utilising importance sampling inside full and conditional likelihoods is reviewed in detail, and new algorithms are suggested. The performance of the existing and newly proposed methods are investigated via simulation. The main findings from all the chapters are synthesised in chapter 8 and conclusions are drawn.
Chapter 2

Background

This chapter reviews the basic theory of likelihood inference and defines a notation for analysing selection biased data. Methods for integral approximation are discussed, in particular the theory of Monte-Carlo integration and importance sampling. Finally, the conditional likelihood approach of Clayton (2003) is explained in detail.

2.1 Classical statistical theory

Consider the problem of fitting the model \( p(y|x, \theta) \) to data \( y_1, \ldots, y_n \), where \( x \) is a covariate and \( \theta \) is an unknown parameter of interest. In practice it is usual to assume that the data \( y \) are selected randomly. The joint probability of \( y_1, \ldots, y_n \), given this statistical model is then equal to

\[
\prod_{i=1}^{n} p(y_i|x, \theta) = \prod_{i=1}^{n} L(\theta|y_i; x) = L(\theta|y; x)
\]

\( L(\theta|y; x) \) is termed the likelihood function given the available data \( x \) and \( y \). Maximum likelihood theory tells us to find the value of \( \theta, \hat{\theta} \) for which
\[ L(\hat{\theta}|y; x) \geq L(\theta^*|y; x) \quad \text{for all } \theta^* \]

\( \hat{\theta} \) is the value for the unknown parameter \( \theta \) that is best supported by the data and for that reason it is termed the \textit{maximum likelihood estimate} (MLE). For a rigorous treatment of likelihood theory see Edwards (1972) or Pawitan (2001). It is standard practice to work with the natural logarithm of the likelihood function \( l(\theta|y; x) \), since

\[
\log(L(\theta|y; x)) = \sum_{i=1}^{n} l(\theta|y_i; x)
\]

is much easier to calculate and does not alter the value of \( \hat{\theta} \). \( S(\theta|y; x) \) is called the \textit{score function} and is defined as

\[
S(\theta|y; x) = \frac{dl(\theta|y; x)}{d\theta}
\]

When the likelihood function is unimodal, the score function satisfies the condition that \( S(\theta|y; x) = 0 \), and can therefore be used to locate \( \hat{\theta} \). However, a more sophisticated scheme for calculating \( \hat{\theta} \) is now described. The \textit{Information} function for \( \theta \) is defined as

\[
\mathcal{I}(\theta) = -\frac{d^2l(\theta|y; x)}{d\theta^2}
\]

It is common practice to assume that the log-likelihood function \( l(\theta|y; x) \) is approximately quadratic around \( \hat{\theta} \). This means that the score function can be approximated by the first order Taylor series expansion

\[
S(\hat{\theta}|y; x) = S(\theta_0|y; x) + \mathcal{I}(\theta_0)(\hat{\theta} - \theta_0)
\]
about a point $\theta_0$. Treating $\theta_0$ as an initial guess, an iterative procedure for estimation of $\hat{\theta}$ is possible, via the formula

$$
\theta_{a+1} = \theta_a - I^{-1}(\theta_a)S(\theta_a|y; x)
$$

where $\hat{\theta} = \theta_q$ when $|\theta_{q+1} - \theta_q|$ is sufficiently small. This is known as the Newton-Raphson method, see Suli and Mayers (2003) for more details. When $\theta$ is a vector of $k$ parameters, the score function is a partial derivative with respect to each element of $\theta$, and the information function becomes a $k \times k$ matrix. The maximum likelihood framework, as well as the principle of the Newton-Raphson method, will be used for statistical inference throughout this thesis.

2.1.1 Models for selection biased data

If the assumption that $y$ are selected randomly does not hold and selection depends on the size of $y$, then failure to allow for the selection mechanism can result in severe bias.

To account for the selection, the likelihood must be based on

$$
p(y|x, \theta, Selected) = \frac{p(y|x, \theta)p(Selected|y; x, \theta)}{p(Selected|x, \theta)} \tag{2.1.1}
$$

Under random selection $p(Selected|y; x, \theta) = p(Selected|x, \theta)$ and (2.1.1) reduces to $p(y|x, \theta)$.

However, when selection is not random (given $x$) and $y$ is continuous, the likelihood will involve the integral

$$
p(Selected|x, \theta) = \int p(y|x, \theta)p(Selected|y; x, \theta)dy
$$
which, for many models is intractable. For discrete data the integral is replaced by a summation over all possible values of \(y\) and this too may be difficult to evaluate. The log likelihood for \(n\) independently selected data values \(y_1, \ldots, y_n\) becomes

\[
l(\theta; y; x, \text{Selected}) = \sum_{i=1}^{n} \left[ \log p(y_i|x_i, \theta) + \log \{p(\text{Selected}|y_i; x_i, \theta)\} - \log \{p(\text{Selected}|x_i, \theta)\} \right]
\]

When selection does not depend on \(\theta\), other than through the observed \(y_i\), the middle term can be dropped from the likelihood, which becomes

\[
l(\theta; y; x, \text{Selected}) = \sum_{i=1}^{n} \left[ \log p(y_i|x_i, \theta) - \log \{p(\text{Selected}|x_i, \theta)\} \right] \quad (2.1.2)
\]

It is sometimes appropriate to model the selection process by assuming that the probability of selection depends on whether \(y\) meets a certain criterion. For example, it could be that there exist probabilities \(\pi_1, \pi_2\) and a subset of all possible values of \(y\), \(C\), such that

\[
p(\text{Selected}|y; x, \theta) = \begin{cases} 
\pi_1 & \text{when } y \in C \\
\pi_2 & \text{when } y \notin C
\end{cases}
\]

so that the probability of selection becomes

\[
p(\text{Selected}|x, \theta) = \pi_1 \int_{y \in C} p(y|x, \theta) \, dy + \pi_2 \int_{y \notin C} p(y|x, \theta) \, dy \quad (2.1.4)
\]

Extensions to selection models with more subsets, or ranges that depend on \(x\), are possi-
ble. If the probability of selection was thought to be different for a large number, \( N \), of distinct subsets \( C_1, C_2, \ldots, C_N \), then \( p(Selected|x, \theta) \) would start to resemble a continuous, rather than a discrete step function.

In some areas of biostatistics, it is appropriate to consider a special case of selection model (2.1.3), where \( \pi_1 = 1 \) and \( \pi_2 = 0 \). For example, in genetic epidemiology, researchers often recruit related individuals into a study, so that many of the effects that genetic and lifestyle factors have on a particular condition can be estimated simultaneously. However it is often only possible to recruit families of individuals if (and only if) at least one member exhibits a medical condition that causes them to be 'ascertained' by a health professional. In this context \( y \) is a vector of responses for a family, \( y \in C \) if at least one member of the family is affected, and

\[
p(y|x, \theta, Selected) = \frac{p(y|x, \theta)}{\int_{y \in C} p(y|x, \theta) dy} \tag{2.1.5}
\]

In genetic studies this is called 'complete' ascertainment, and is discussed in more detail in Chapter 5.

In sequential clinical trials a similar dichotomous selection process is encountered when a trial is designed with the flexibility to be stopped at any number of time points \( T_i, i = 1, \ldots, n \), depending on the performance of a particular treatment. In order to correctly model the data at time point \( i \), equation (2.1.5) must be considered, where \( y \in C \) indicates, for example the set of all subjects \( y \) recruited up to the \( i \)'th inspection, (Whitehead, 1983). In Chapter 3 it is argued that equation (2.1.5) is also a sensible way to model the
distribution of study effect sizes \( y \) in a meta-analysis, where membership of \( C \) indicates that the effect size passed a publication criterion.

To simplify the notation the covariate \( x \) will be dropped where it is not essential and; the pdf of the full data \( p(y|\theta) \) will now be referred to as \( h(y|\theta) \), the selection probability \( p(Selected|\theta) \) will be written \( z(\theta) \) and the pdf of the selected data \( p(y|\theta, Selected) \) will be \( f(y|\theta) \), so that (2.1.5) becomes

\[
f(y|\theta) = \frac{h(y|\theta)}{z(\theta)} \tag{2.1.6}
\]

\( z(\theta) \) can be thought of as the normalising constant to the un-normalised density \( h \), and its accurate calculation is essential if we are to correct for selection bias. However, normalising constants are required in many other forms of statistical inference, such as the analysis of censored data, see Sun (2006) for a full review. For example, it is common for a data set of survival times \( y_1, \ldots, y_n \) to be made up of two distinct subgroups; those patients that experienced an event (such as death, or myocardial infarction) during a trial, and those who were event free up to a cutoff point \( T \), when the trial stopped. Letting \( y_1, \ldots, y_k \) index the \( k \) subjects who experienced a failure within the trial, and \( y_{k+1}, \ldots, y_n \) indicate the true but unobserved failure times of the other individuals, the full likelihood of the data can be expressed as
As science progresses, there is a need to fit increasingly complex models to statistical data. When models are complex, the integration required in the evaluation of the normalising constant \( z(\theta) \) can be intractable. One way to proceed with statistical inference in this case to produce an approximation to \( z(\theta) \). In Section 2.2 two basic methods of integral approximation are introduced.

### 2.2 Methods for integral approximation

#### 2.2.1 Numerical integration

Numerical integration is a term used for a large family of algorithms that provide a numerical approximation to any function’s integral. When this function has an unknown form, but has a known value at more than one point, the general Newton-Coates formula, (Whittaker and Robinson, 1967), can provide an approximation to the integral between those points. For example, when the value of a function \( g(x) \) is known at two locations, \( x_1, x_2 \), the Newton-Coates formula tells us to approximate the integral

\[
\int_{x_1}^{x_2} g(x) \, dx
\]

with a simple average value of \( (x_2 - x_1) \frac{g(x_2) - g(x_1)}{2} \), which is commonly known as the
Trapezoid rule, see Figure 2.1 (Left). If the value of $g$ is known at $n$ equally spaced points $x_1, ..., x_n$ then the Newton-Coates formula can be extended to infer

$$
\int_{x_1}^{x_n} g(x) \, dx \approx \sum_{i=1}^{n} w_i g(x_i)
$$

where the $w_i$ are weights that are calculated using Lagrange polynomials, (Abramowitz and Stegun, 1972). The Newton-Coates formula makes numerical integration possible for relatively undefined functions. For this reason, it has been used to help solve systems of differential equations in the engineering sciences, (Lambert, 1991). However, the stipulation that the points for which $g$ is known must be equally spaced is a consequence of the fact that the weights only depend on the location of the points, and not on the function $g$ itself. This 'one size fits all' characteristic can make the algorithm prone to failure in specific circumstances.

![Figure 2.1: Left: An illustration of the Trapezoid rule of integration for an unknown function $g(x)$. Right: A plot of the 5 most optimal points $s_1, ..., s_5$ with which to base the numerical integral for a standard normal probability density function over the range $(-\infty, \infty)$.](image)

In this thesis, we are faced with the relatively simple problem of integrating a known density function with an unknown parameter, such as $h(y|\theta)$. When the range for this
integral is well defined, a special form of numerical integration, called *Gaussian quadrature* can be used, (Barbeau, 1989). It is a more sophisticated method than the Newton-Coates formulas because, for a given \( n \), the range of the random variable \( Y \) is used to determine the most optimal points, or Abscissa \( s_1, \ldots, s_n \) and weights \( w(s_i) \) with which to achieve the integral approximation. For example, if \( Y \) is normally distributed, the appropriate range is \((-\infty, \infty)\) and the optimal weight function \( w(s_i) \) is known to equal \( e^{-s_i^2} \). Figure 2.1 (Right) shows a plot of 5 optimal Abscissa \( s_1, \ldots, s_5 \) below the density function of a univariate normal distribution. Letting \( h(y|\theta) \) be this standard normal distribution, they could be used to approximate the integration necessary for the evaluation of the expected value, or variance of \( Y \), using a weighted sum, as in (2.2.1). It is important to note that whilst their locations are not equally spaced, they are still highly ordered and symmetric about a central point. In Chapter 6 Gaussian quadrature is used to integrate a likelihood function for a population of families in a genetic epidemiological study, with respect to a genetic component of disease risk.

Is numerical integration an appropriate method with which to estimate the selection probability \( z(\theta) \)? Firstly, it may be difficult to find the best weight function \( w \) to numerically integrate \( h(y|\theta) \) over the set of values for \( y \) that are in the selected set \( C \). When the range of \( y \) is \([0, \infty)\) it is the *Laguerre* function \( e^{-s_i} \), when the range is \([-1,1]\) the *Chebyshev* function \((1 - s_i)^{1/2}\) is used, but none of these special cases could suffice. Secondly, as \( n \) increases numerical approximations will tend to become more accurate, but the price can be severe in terms of computational effort. For example, if \( n \) quadrature points are sufficient to approximate a one-dimensional integral of the function \( h(y|\theta) \) with respect to \( y \), a two-dimensional analogue of \( h(y|\theta) \) would require \( n^2 \) - a grid of \( n \times n \) points. In general for
an $m$-dimensional integration, $n^m$ quadrature points would need to be evaluated, which as
$m$ increases, gets very large. This is commonly referred to as the 'curse of dimensionality',
a phrase first coined by Bellman (1961).

Methods for avoiding extraneous computation using numerical methods have been pro­
posed. These include rules to subdivide an integral's area into more easily manageable
parts, as well as systems for sequentially increasing the number of integration points eval­
uated, until the result remains unchanged, given a prespecified tolerance, (McKeeman,
1962; Gander and Gautsch, 2000; Rabe-Hesketh et al., 2002). This area of research is
referred to as adaptive quadrature. However, whilst improving numerical integration’s
performance, they do not totally remove the shortcomings. This has lead to more effi­
cient algorithms for complex integral approximation being sought. One such method is
Monte-Carlo integration.

2.2.2 Monte-Carlo integration

Monte-Carlo methods are a set of extremely powerful tools that harness the power of
random numbers, as opposed to fixed but optimal abscissa, to perform statistical inference.
Metropolis and Ulam (1949) are credited with naming and formalising the approach,
which came to prominence due to its extensive use in the Manhattan project, a World
War Two development programme for nuclear weapons. In order to explain Monte-Carlo
integration simply, let $Y_1, ..., Y_n$ be a sequence of independent and identically distributed
random variables, with probability density function $h(y|\theta)$, where $\theta = (\mu, \sigma^2)$, $E(Y) = \mu$
and $Var(Y) = \sigma^2$. Consider the problem of estimating $\mu$, which is identical to evaluating
2.2. Methods for integral approximation

\[ E[Y] = \int y h(y|\theta) dy \]

If \( Y_1, \ldots, Y_n \) could actually be observed, \( \mu_n = \frac{1}{n} \sum_{i=1}^{n} Y_i \), could be evaluated. The simplicity of \( \mu_n \), the average value of \( n \) random variables, belies the fact that it is an extremely desirable quantity, because as \( n \) gets large, it approaches \( \mu \), as a result of the Law of Large Numbers, (Grimmett and Stirzaker, 1992). More formally, the expected squared error loss

\[
E(\mu_n - \mu)^2 = \frac{1}{n^2} \left\{ \text{Var} \left( \sum_{i=1}^{n} Y_i \right) + E \left( \sum_{i=1}^{n} Y_i \right)^2 \right\} - 2 \mu \frac{1}{n} E \left( \sum_{i=1}^{n} Y_i \right) + \mu^2 \\
= \frac{1}{n^2} \left( n\sigma^2 + n^2 \mu^2 \right) - \mu^2 \\
= \frac{\sigma^2}{n} \tag{2.2.2}
\]

So, as \( n \to \infty \quad E(\mu_n - \mu)^2 \to 0 \). The central limit theorem can be used to state that

\[
\frac{\sqrt{n}}{\sigma} (\mu_n - \mu)
\]

is approximately distributed as \( N(0,1) \) as \( n \to \infty \). Consequently, \( \mu_n \) will converge to \( \mu \) at rate \( \frac{1}{\sqrt{n}} \). This is the fundamental principle of Monte-Carlo integration. Although in this example the parameter of interest is one-dimensional, the real power of Monte-Carlo integration is that this convergence result holds, in principle, for a parameter space of any dimension. In other words Bellman's 'curse' does not apply.

\( \mu_n \) can not be calculated because it is a function of true unobservable random variables, but it can be approximated with \( \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} y_i \), by simulating a random sample of data.
y_1, ..., y_n from the density h(y|\theta). This can be achieved for a range of standard density functions in most statistical software packages. For an in depth discussion on the issue of simulating data for the purposes of Monte-Carlo integration, see Ripley (1987).

Figure 2.2: By simulating data y_1, ..., y_5 from the distribution h(y|\theta) A Monte-Carlo estimate for E[Y] can be obtained by calculating their mean value, \bar{y}.

Figure 2.2 shows the principle of Monte-Carlo integration, as applied to the same normal example in Figure 2.1 (Right). If five realisations from a N(\theta, 1) distribution can be simulated, their mean value could be used to approximate the integral \int yh(y|\theta)dy with \frac{1}{5} \sum_{i=1}^{5} y_i. Note that by being random draws from h, the data points are neither equally spaced, as in Figure 2.1 (Left), nor symmetrically spaced, as in Figure 2.1 (Right). Approximation of the selection probability z(\theta) = \int_{C} h(y|\theta)dy using Monte-Carlo integration, reduces to calculating the proportion of data points that pass the selection criteria since, for a sample of data points y_1, ..., y_n from h(y|\theta)

\[
z(\theta) = E[Y \in C] = \int_{C} I_C(y) \ h(y|\theta)dy
\approx \frac{1}{n} \sum_{i=1}^{n} I_C(y_i) \tag{2.2.3}
\]
where the indicator function $I_C(y_i) = 1$ if $y_i \in C$ or $0$ if $y_i \notin C$.

In most statistical applications $\theta$ will be unknown, so it is not possible to simulate data from $h(y|\theta)$. This means, although this basic method is extremely elegant and simple, it is, in general, of little practical use. In the next section a particular Monte-Carlo method called importance sampling, that can be more easily incorporated into maximum likelihood inference, is discussed. From now on, I refer to data that is assumed to come from $h(y|\theta)$ as ‘real’ data. Conversely I will refer to data that has been simulated for the purposes of Monte-Carlo integration as ‘pseudo’-data. All real data will be given the suffix 0, as in $y_0$, and $y_j, j = 1, ..., m$ will refer to simulated data.

## 2.3 Importance sampling

### 2.3.1 Introduction

Importance sampling is a common Monte-Carlo method that can be used for approximating integrals such as $z(\theta)$ in selection problems, (Srinivasan, 2002). Consider a real data point $y_0$, from the complex distribution $f(y_0|\theta)$. It may be hard to evaluate $z(\theta)$ directly but, providing the rules for membership of $C$ are known, comparatively easy to take an arbitrary value $\theta'$ and to simulate pseudo-data $y_j, j = 1, ..., M$ from $h(y_j|\theta')$. Ideally in this situation, importance sampling would work by noting that, for any arbitrary value $\theta^*$
2.3. Importance sampling

\[ z(\theta^*) = \int I_C(y) h(y|\theta^*) dy \]
\[ = \int I_C(y) \frac{h(y|\theta^*)}{h(y|\theta')} h(y|\theta') dy \]
\[ \approx \frac{1}{M} \sum_{j=1}^{M} \frac{I_C(y_j) h(y_j|\theta^*)}{h(y_j|\theta')} \]
\[ = \frac{z(\theta^*)}{m} \quad (2.3.1) \]

where \( m \) is the number of points out of \( M \) for which \( I_C(y_j) = 1 \). Because (2.3.1) holds for an arbitrary value \( \theta^* \) it is much easier to incorporate into the maximum likelihood framework, for example to evaluate

\[ L(\theta^*|y) \approx \frac{h(y|\theta^*)}{z(\theta^*)} \quad (2.3.2) \]

When \( \theta^* \) is equal to the true value \( \theta \), the asymptotic variance of (2.3.1) is equal to \( \frac{\sigma^2}{m} \) where

\[ \sigma^2 = \int \left( I_C(y) \frac{h(y|\theta)}{h(y|\theta')} - z(\theta) \right)^2 h(y|\theta') dy \]
\[ = \int I_C^2(y) h^2(y|\theta) \frac{dy}{h(y|\theta')} - z(\theta)^2 \quad (2.3.3) \]

Owen and Zhou (2000) point out that the most optimal importance sample could be obtained if one additionally chose \( \theta' = \theta \), because then clearly (2.3.3) would equal 0. This tells us that importance sampling will be most efficient when \( \theta' \) is close to \( \theta \), as this zero variance will not be achieved in practice. Although it is simply an estimate for
2.3. Importance sampling

the normalising constant $z(\theta)$ that is required, and an approximation for it can easily be obtained via (2.3.1), a possible downside to estimating $z(\theta)$ is that $m$ has the ability to vary. If this process is repeated several times, different simulated data sets of size $M$ would give rise to different values of $m$; so that the variance of $z(\theta)m$ will also have a variance. To remove this extra level of complication, $m$ can be fixed by simulating pseudo-data from $h(y|\theta')$, until $m$ obtain membership of $C$. Crucially, the $m$ data points can now be viewed as a random sample from $f(y|\theta')$. By using pseudo-data $y_j$, $j = 1, ..., m$, $z(\theta)$ can alternatively be estimated to within a constant of proportionality via importance sampling since

$$
\int \frac{h(y|\theta)}{h(y|\theta')} f(y|\theta') dy = \frac{z(\theta)}{z(\theta')} \int \frac{f(y|\theta)}{f(y|\theta')} f(y|\theta') dy \\
= \frac{z(\theta)}{z(\theta')} \\
\approx \frac{1}{m} \sum_{j=1}^{m} \frac{h(y_j|\theta)}{h(y_j|\theta')}
$$

(2.3.4)

There are many instances in statistics where ratios of normalising constants are needed. For example, in a frequentist analysis a likelihood ratio test might be performed to test a particular hypothesis $H_0 : \theta = \theta_1$ by calculating the ratio

$$W(\theta_1) = 2 \log \left\{ \frac{L(\theta_1|y)}{L(\theta|y)} \right\} = 2 \log \left\{ \frac{h(y|\theta_1)}{h(y|\theta)} \right\} - 2 \log \left\{ \frac{z(\theta_1)}{z(\theta)} \right\}
$$

A very similar problem in Bayesian inference is the calculation of Bayes factors, (McCulloch and Rossi, 1992; Han and Carlin, 2001), for which the ratio of two posterior densities, rather than two likelihoods, must be considered. Doucet et al. (2000) show how impor-
2.3. Importance sampling

tance sampling of this sort can be implemented in the Bayesian setting, but in this thesis I focus on its use within the maximum likelihood framework. Identity (2.3.4) allows us to approximate the log of likelihood (2.3.2), for a sample of real data \( y_{i0}, i = 1, \ldots, n \) and pseudo-data \( y_j, j = 1, \ldots, m \) by

\[
l(\theta|y, \theta') \approx \sum_{i=1}^{n} \left[ \log \left( \frac{h(y_{i0} | \theta)}{h(y_{i0} | \theta')} \right) - \log \left( \frac{1}{m} \sum_{j=1}^{m} \frac{h(y_j | \theta)}{h(y_j | \theta')} \right) \right]
\]

(2.3.5)

The \( \hat{\theta}_m \) that maximises equation (2.3.5) is the Monte-Carlo approximation of the MLE. The idea of substituting an importance sample into a likelihood in order to estimate \( \hat{\theta}_m \) is usually attributed to Penttinen (1984), whose purpose was to avoid the explicit calculation of integrals required to model spatial point pattern interaction. The properties of (2.3.5) are discussed in detail by Geyer and Thompson (1992). As \( m \to \infty \),

\[
\frac{1}{m} \sum_{j=1}^{m} \frac{h(y_j | \theta)}{h(y_j | \theta')} \to \frac{z(\theta)}{z(\theta')}
\]

and consequently \( \hat{\theta}_m \to \hat{\theta} \), almost surely. From now on I will refer to this method, as Monte-Carlo maximum likelihood - MCML for short.

2.3.2 An example of MCML in practice

The performance of likelihood (2.3.5) is clearly adversely affected when \( m \) is small, but also when the distance between \( \theta \) and \( \theta' \) is large. Since the variance of the Monte-Carlo estimate \( \frac{z(\theta)}{z(\theta')} \) will equal to

\[
\frac{1}{mz(\theta')} \int \frac{h^2(y | \theta)}{h(y | \theta')} dy - \left( \frac{z(\theta)}{z(\theta')} \right)^2
\]

if \( h(y | \theta') \) approaches 0 faster than \( h^2(y | \theta) \) as \( y \) moves away from \( \theta \), then this variance could be infinite.
To highlight the need for improvements to the basic method, a simple simulation study is shown. Real data \( y_i, i = 1, \ldots, 100 \) were simulated from a standard \( N(\theta = 0, 1) \) distribution, and selected with probability \( \pi_1 = 1 \), if \( y_i \geq 0 \), and probability \( \pi_2 = 0 \) otherwise. The mean value of this sample, \( \bar{y}_0 \), is approximately 0.79, which is clearly a biased estimate for \( \theta \). For correct estimation each \( h(y_i | \theta) \) must be conditioned on the probability of selection, \( z(\theta) \), which in this case can be equated to \( \Phi(\frac{\theta}{1}) \), the cumulative distribution function of a standard normal distribution. If \( z(\theta) \) can not be calculated, but a reasonable guess for \( \theta \) of 0.5 can be provided, then an importance sampling estimate of \( \frac{z(\theta)}{z(0.5)} \) could be obtained by simulating \( m \) pseudo-data points \( y_j, j = 1, \ldots, m \) from a \( N(\theta' = 0.5, 1) \) distribution, subject to being \( \geq 0 \), and calculating the mean value of the ratio \( \frac{h(y_j | \theta)}{h(y_j | \theta')} \) over \( j \).

![Figure 2.3: (Left) The accuracy, for varying \( m \), of \( \left( \frac{1}{m} \sum_{j=1}^{m} \frac{h(y_j | \theta)}{h(y_j | \theta')} \right)^{-1} \) as a Monte-Carlo estimate for \( \frac{z(\theta)}{z(\theta')} \) when \( \theta' = 0.5 \). (Right) The value of \( \hat{\theta}_m \) that maximises (2.3.5), given \( \theta' = 0.5, \theta \) is 0. Point estimates and 95% confidence intervals for \( \hat{\theta}_m \) are shown, as well as the number of runs failing to converge.](image)

It is important to look at the behaviour of the inverse of this ratio because it is the denomi-
nator in MCML. Figure 2.3 (Left) shows, for 100 independent simulations of pseudo-data, the average value of the importance sampling estimate for \( \frac{z(0.5)}{z(\theta)} \) as a function of \( \theta \) for \( m=1,2 \) and 5. When \( \theta = \theta' = 0.5 \) the ratio is trivially equal to 1, but when \( \theta \neq \theta' \) the required ratio is overestimated, crucially around the point \( \theta = 0 \), which is circled. Figure 2.3 (right) shows the impact that this overestimation has on the eventual estimate \( \hat{\theta}_m \) in MCML. The average value, over 1000 simulations of real and pseudo data, for \( \hat{\theta}_m \) is shown as a function of \( m \). When \( m \) is small, the overestimation of \( \frac{z(0.5)}{z(\theta)} \) causes each observation \( y_{i0} \) to be overcorrected for selection bias, meaning that \( \hat{\theta}_m \) is underestimated. When \( m \) is small, quasi-Newton maximisation routines can fail to find a maximum, or instead fail to assign a finite value to \( Var(\hat{\theta}_m) \). The percentage of runs ending in failure for each \( m \) is shown. These failures were excluded from the calculations and so do not affect the plots.

### 2.3.3 A conditional likelihood alternative

Clayton (2003) suggested an elegant approach to selection bias correction in genetics research, using the analogy of data from a matched case-control study, which is now described. His starting point was to consider the evaluation of the log-likelihood function for a series of observations \( y_1, ..., y_n \), from \( h(y|\theta) \) sampled under complete ascertainment rule (2.1.3). This took the form

\[
\ell(\theta|y) = \sum_{i=1}^{n} \log \left( \frac{h(y_i|\theta)}{\int_C h(y_i|\theta)dy} \right)
\]

Clayton noted that the score function was equal to

\[
\sum_{i=1}^{n} \log \left( S(\theta|y_i) - \frac{\int_C S(\theta|y_i)h(y_i|\theta)dy}{\int_C h(y_i|\theta)dy} \right) \tag{2.3.6}
\]
where \( S(\theta|y_i) = d\log(h(y_i|\theta)/d\theta \). Standard maximum likelihood theory tells us to find
the \( \hat{\theta} \) for which (2.3.6) equals zero, but when integration over the set \( C \) is intractable,
an alternative method must be used. Clayton postulated that the second term in (2.3.6)
could be approximated using the general method of data augmentation, (Tanner and Wong,
1987; Wei and Tanner, 1990). To illustrate Clayton’s method it is helpful to consider a
matrix of real data \( y_{i0} \) and pseudo-data \( y_{ij} \), augmented together, as in Figure 2.4.

\[
\begin{pmatrix}
\theta & \theta' & \theta' & \ldots & \theta' \\
y_{10} & y_{11} & y_{12} & \ldots & y_{1m} \\
y_{20} & y_{21} & y_{22} & \ldots & y_{2m} \\
\ldots & \ldots & \ldots & \ldots & \ldots \\
y_{n0} & y_{n1} & y_{n2} & \ldots & y_{nm}
\end{pmatrix}
\]  

(2.3.7)

Figure 2.4: Data structure for Clayton’s algorithm. \( y_{i0} \ i = 1, \ldots, n \) are from \( f(y|\theta) \) and
pseudo-data \( y_{ij}, j = 1, \ldots, m \) are from \( f(y|\theta') \).

As before, if it is assumed that all real data follows the complex distribution \( f(y|\theta) \), then
the pseudo-data is simulated to come from \( f(y|\theta') \). Considering row \( i \) as forming a set
\( \zeta_i = \{y_{i0}, y_{i1}, \ldots, y_{im}\} \), Clayton constructs a likelihood whose \( i \)'th term is the conditional
probability that, given there is only 1 real data point in \( \zeta_i \), that it is \( y_{i0} \). This takes the
form

\[
l_{\text{cond}}(\theta|\theta', y) = \sum_{i=1}^{n} \log \left( \frac{h(y_{i0}|	heta')}{z(\theta)} \prod_{k=1}^{m} \frac{h(y_{ik}|	heta')}{z(\theta')} + \sum_{j=1}^{m} \frac{h(y_{ij}|	heta)}{z(\theta)} \prod_{k=0, k \neq j}^{m} \frac{h(y_{ik}|	heta')}{z(\theta')} \right) 
\]  

(2.3.8)

dividing by \( \prod_{k=0}^{m} \frac{h(y_{ik}|	heta')}{z(\theta')} \) and canceling the ratio \( \frac{z(\theta')}{z(\theta)} \) that is common to all terms yields
the greatly simplified expression
Clayton showed that the score function for his conditional likelihood equalled

\[
S(\theta | y_i) - \frac{\sum_{y \in \mathcal{C}_i} S(\theta | y) \frac{h(y \mid \theta)}{h(y \mid \theta')} \frac{h(y) \mid \theta)}{h(y) \mid \theta')}{\sum_{y \in \mathcal{C}_i} h(y) \mid \theta')}
\]

This score function was now free of the troublesome integrals in (2.3.6), which were replaced by ratios of real and pseudo-data, evaluate under the density \(h\). The information function for his conditional likelihood naturally shared this property too, which meant that \(\hat{\theta}\) could, if so desired, be estimated by the Newton-Raphson algorithm described in Section 2.1. Careful inspection of the \(i^{th}\) term of (2.3.9) reveals that whilst ascertainment probabilities cancel out, what remains on the denominator is the Monte-Carlo estimate

\[
\sum_{j=1}^{m} \frac{h(y_{ij} \mid \theta)}{h(y_{ij} \mid \theta')} = m \frac{z(\theta)}{z(\theta')} 
\]

Thus Clayton's method manages to incorporate importance sampling into an exact conditional likelihood, as opposed to the approximate full likelihood approach of Geyer and Thompson (1992). This is totally justified because conditional likelihoods have the same asymptotic properties as standard likelihoods, (Andersen, 1970).

In order to understand the value of conditional likelihoods, and conditional arguments in general, it is helpful to think of every full likelihood as the product of a marginal likelihood and a conditional likelihood, (Qin and Zhang, 2005). This implies that the amount of information about \(\theta\), \(I(\theta)\), in a full log-likelihood \(l(\theta | y)\), is the sum of the information
in the conditional likelihood \( l_{\text{cond}} \) and the marginal likelihood \( l_{\text{marg}} \). Put formally

\[
\begin{align*}
    l(\theta|y) &= l_{\text{cond}}(\theta|y) + l_{\text{marg}}(\theta|y) \\
    I(\theta) &= I_{\text{cond}}(\theta) + I_{\text{marg}}(\theta)
\end{align*}
\]

(2.3.10)

At optimum efficiency, when \( \theta' = \theta \), Clayton shows that the expected value of the information function of conditional log-likelihood (2.3.9), \( I_{\text{cond}}(\theta) \), is of the order \( \frac{m}{m+1} I(\theta) \).

This means that the amount of information, \( I_{\text{marg}}(\theta) \), that is lost by canceling out the marginal selection probability \( z(\theta) \), is equal to \( \frac{I(\theta)}{m+1} \). Clearly, as \( m \to \infty \) this will become small. For a fully rigorous and detailed explanation of conditional likelihoods, and their role in statistical inference see Cox and Reid (1987) or Reid (1995).

In the context of modeling selection biased data, Clayton's idea of conditioning out the selection probability is very appealing because it is of no intrinsic interest, merely a nuisance quantity. When a single selection probability \( z(\theta) \) is required, MCML and Clayton's conditional likelihood can be implemented. But when \( y_{i0} \) is dependent on a continuous covariate \( x_{i0} \), each real observation from \( f(y|x, \theta) \) possesses a unique normalising constant

\[
z(\theta|x_i) = \int_{y_{i0} \in C} h(y_{i0}|x_i, \theta) dy_{i0}
\]

which must be calculated. In order to learn about \( z(\theta|x_i) \) via importance sampling, pseudo-data must be simulated conditional on \( x_i \), and then it can only be informative about \( z(\theta|x_i) \). However, the matched nature of real and pseudo-data in Clayton's conditional
2.3. Importance sampling

likelihood means that generalisation to the covariate case is immediate, and all that is required is for pseudo-data $y_{ij}$ to be simulated from $f(y_{ij}|x_i, \theta')$. MCML on the other hand can not deal so immediately with a covariate. Figure 2.3 shows that $m$ needs to be fairly large for a single normalising constant $z(\theta)$ to be estimated, the addition of a covariate would multiply the amount of pseudo-data required $n$ times, which would add considerable computational effort.

\[
\begin{pmatrix}
\theta & \theta' & \theta' & \ldots & \theta' \\
x_1 & y_{i0} & y_{i1} & y_{i2} & \cdots & y_{1m} \\
x_2 & y_{20} & y_{21} & y_{22} & \cdots & y_{2m} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x_n & y_{n0} & y_{n1} & y_{n2} & \cdots & y_{nm}
\end{pmatrix}
\]

Figure 2.5: Conditioning on a covariate $x_{i0}$. $y_{i0}$ $i = 1, \ldots, n$ from $f(y_{i0}|x_i, \theta)$ and pseudo-data $y_{ij}$, $j = 1, \ldots, m$ from $f(y|x_i, \theta')$.

Perhaps the only criticism of (2.3.9), is that a single pseudo-data distribution with a common parameter $\theta'$ must be assumed. It is not possible to weight each real observation $y_{i0}$ with pseudo observations from different pseudo-data distributions, as the required cancelations in equation (2.3.8) do not occur. When $\theta$ is multidimensional, it will be hard to propose a single $\theta'$ 'close' to $\theta$ and the further apart they are, the smaller the information $I(\theta)$ will be. Naturally this means that the variance of any estimate for $\theta$, which is the inverse function of $I(\theta)$, will be large. This point is discussed in more detail in Chapter 7.
2.4 Summary

This chapter has introduced the notion for selection bias and shown that the process of modeling selection biased data shares similarities with other areas of statistics, requiring the calculation of normalising constants. Monte-Carlo integration and in particular importance sampling is identified as a general method for calculating $z(\theta)$ in a fashion that avoids unnecessary computation. A method to incorporate importance sampling into maximum likelihood inference, namely Clayton's conditional likelihood, was introduced. Because of its many positive attributes, as well as its recent proposal, I believe it merits further investigation, both in terms of its theoretical properties as well as its potential for practical application to selection biased biomedical data. In Chapter 3 I investigate the use of Clayton's conditional likelihood to correct for publication bias in meta-analysis.
Chapter 3

A correction for publication bias in meta-analysis

Adjustment for publication bias in meta-analysis has important parallels with the correction for ascertainment bias in genetic studies. In this chapter Clayton's conditional likelihood, an importance sampling based method originally described as a means of correcting for ascertainment bias, is modified for use in meta-analysis when publication bias is suspected. The method involves simulating sets of pseudo-studies under the assumed meta-analytic model using guesses for the unknown parameters, as in Chapter 2. The pseudo-studies are subjected to the same selection criteria that are believed to have operated on the original studies and are both then used to estimate the adjusted values of the unknown parameters. This method is used to re-analyse a published meta-analysis of the effect of the MTHFR gene on homocysteine levels, a substance that is thought to regulate the risk of cardiovascular disease. Simulation studies show that the pseudo-data method is unbiased; they give an indication of the number of pseudo-data values required and suggest that a two-stage adjustment produces less variable estimates. This method
3.1. Introduction

Meta-analysis is the process of quantitatively combining the results of many studies that have looked at the same endpoint. It is usually undertaken as part of a broader systematic review of research on a particular topic and is especially useful when the individual studies are small and underpowered. A key assumption, implicit in any meta-analysis, is that the systematic review has identified a representative sample of studies; unfortunately this assumption may not hold in practice. Even if the review is thorough, it will rarely be possible to identify those studies that do not make it into the literature. Nor will it usually be possible to identify those numerous sub-analyses that were actually undertaken in analysing a particular study, but were not deemed 'useful enough' to report in the final published article. The problems of publication and reporting bias, sometimes together known as dissemination bias, have been a major concern in many meta-analyses. Although the potential impact of dissemination bias can be very great, it is very difficult to correct for because, by its very nature, we do not observe the unpublished or unreported data and cannot easily investigate the process that led to the bias. Despite the lack of evidence, there is a strong suspicion that publication is often linked to the size and/or statistical significance of the effect, with studies that exhibit a large effect or significant association.
being more likely to be published than those that draw less positive conclusions. Several methods have been proposed for detecting or correcting for publication bias, see Rothstein et al. (2005) for a full review. The starting point is usually a visual inspection of a ‘funnel plot’ of the data from the studies identified in the literature search; this plot shows each study’s estimated effect against a measure of the estimate’s precision, (Light and Pillemer, 1984). In the absence of dissemination bias the funnel plot should be symmetrical about the common average effect. Any departure from symmetry suggests, but does not prove, dissemination bias and should be investigated further. Most methods devised to detect publication bias test the assumption that there is no correlation between the effect size and the precision of the trials, (Begg and Mazumdar, 1994; Egger and Davey-Smith, 1997; Macaskill et al., 2001). If a significant level of correlation is found then this is taken as evidence for the presence of publication bias. ‘Trim and Fill’, (Duval and Tweedie, 2000), is a method that both detects and corrects for publication bias. It involves augmenting the meta-analysis with extra imaginary studies created with effect sizes and standard errors that make the funnel plot more symmetrical.

Historically, meta-analysis methodologists have commonly thought of two models being in operation when the data are generated, the ‘effect size model’ and the ‘selection model’. The effect size model generates the outcome measure for each study, for instance, a mean difference or a log-odds ratio, whilst the ‘selection model’ determines whether this measure is published. Iyengar and Greenhouse (1988) imagined a selection model that allocated each potential effect estimate a probability that, if that estimate were produced in a study and put forward for publication, it would then be published. They assumed that this weight was dependent on the p-value associated with the effect size. Copas (1999) added
3.1. Introduction

a further stage of complexity by assuming that the selection model was a function of two parameters, the estimated effect size and its standard error. Selection models have fallen into two distinct camps; those that assume the parameters of the selection model are unknown but estimable via maximum likelihood or Bayesian methods, (Givens and Smith, 1997), and those that assume the selection model’s parameters are known but then investigate the sensitivity to changes in the known values. Hedges (1984) used this latter approach in the context of estimating the true mean difference, $\theta$, of data coming from a certain truncated model, with observed difference $\hat{\theta}$. He investigated how changing the severity of the truncation in the selection model changed the ratio between the observed mean difference $\hat{\theta}$ and the true mean difference $\theta$.

Publication bias is a particular form of selection bias and as such has parallels in other areas of statistics. In particular there is an array of techniques for correcting for potential bias due to survey sampling in which some subjects have a greater probability of inclusion than others. Providing that those probabilities are known, it is generally possible to correct for this selection bias by weighting the data inversely to their probability of inclusion, (Rao, 1965).

In this chapter, I investigate the use of pseudo-data to correct for publication bias. In Section 3.2, I present a motivating example and analyse it without correction for publication bias and then re-analyse it using a ‘Trim and Fill’ correction. Section 3.3 describes Clayton’s pseudo-data method, as it would be applied within a meta-analysis. Section 3.4 uses simulation to investigate the performance of the adjustment for situations in which the correct selection criterion is both known and unknown. Sections 3.6 and 3.7 apply
the pseudo-data adjustment to the motivating example, with different assumptions about
that it assumes a known selection model and uses sensitivity analysis to investigate how
the publication bias correction changes with different selection criteria. At face value, the
method also appears to have similarities to the methods of Duval and Tweedie (2000)
and Givens and Smith (1997) in that it too involves data augmentation, however, there
is a fundamental difference. Whereas they augment meta-analyses with missing studies, I
augment them with whole new meta-analyses that have also been subject to an assumed
selection criteria.

3.2 Homocysteine and the MTHFR gene

Wald and Law (2002) published a meta-analysis of genetic association studies with the
aim of investigating whether homocysteine is causally linked to cardiovascular disease.
They considered a single common mutation in the methylenetetrahydrofolate reductase,
or MTHFR, gene that has been associated with a 20 % rise in the average level of homo­
cysteine and with a corresponding rise in cardiovascular disease. Wald and Law’s paper
concluded that the relationship is causal and that lowering a person’s homocysteine level,
perhaps through the use of folic acid, would reduce the risk of cardiovascular disease. For
a thorough review of genetic association studies see Cordell and Clayton (2005).

Minelli et al. (2004) re-analysed the homocysteine data to illustrate the use of Mendelian
randomisation to estimate the size of the causal association between homocysteine and
cardiovascular disease risk. Figure 3.1 is a schematic diagram that illustrates the idea of
Mendelian randomisation, which will be referred to in more detail in the next section. By
3.2. Homocysteine and the MTHFR gene

Figure 3.1: The principle of Mendelian randomisation. Observable gene-phenotype and gene-disease relationships can be combined to derive an estimate for the phenotype-disease relationship that is free from confounding and reverse causation.

using the genotype as an instrumental variable they obtained an estimate of the association between homocysteine and cardiovascular disease that was, assumed to be, free from confounding and reverse causation.

Figure 3.2: A funnel plot of the 33 trials reporting a mean difference in homocysteine, as analysed by Minelli et al. (2004). The y axis scale is the inverse each study's reported standard error, which is also called the 'precision'. Vertical lines indicate fixed and random effects estimates for the mean difference, unadjusted for publication bias.

Figure 3.2 shows a funnel plot of the data on the average difference in homocysteine levels in the 33 published studies used in Minelli et al. (2004). Each of the trials measured the
average difference in homocysteine levels, in μmol/l, between groups of people with two
copies of the mutation, TT, and people with no copies of the mutation, CC. The estimates
for the overall average difference in homocysteine are 1.95 and 2.68 μmol/l under fixed and
random effects models, respectively. These both seem likely to be overestimates because
the funnel plot is asymmetric and suggestive of publication bias. Furthermore, the Begg
test concludes that publication bias is highly likely (p < 0.001).

There is also a strong suggestion of 'heterogeneity' in the data, that is, evidence to support
the view that all studies are not measuring the same common mean. This can be assessed
using the Chi-Squared test of Cochran (1954), which shows that the proportion of varia-
tion in the effect sizes that can not be explained by chance alone, is 82%. In meta-analysis
this is referred to as the $I^2$ statistic, see Higgins and Thompson (2002).

A commonly used, nonparametric way to correct for publication bias is to use 'Trim and
Fill', (Duval and Tweedie, 2000). Assuming a fixed effects model, Trim and Fill estimates
that 12 studies are missing due to dissemination bias; imputing the missing studies lowers
the predicted average homocysteine difference to 1.43 μmol/l (95 % C.I 1.12, 1.72). A
random effects Trim and Fill analysis estimates that 11 studies are missing, and predicts
the average homocysteine difference is 1.65 μmol/l (95 % C.I 0.86, 2.44). The between-
study variance is estimated to be 4.8 μmol²/l².

As with publication bias, heterogeneity can also induce funnel plot asymmetry. Terrin
et al. (2003) noted that Trim and Fill will often wrongly correct for publication bias when
only heterogeneity is present. Similarly, regression based methods to detect publication
bias have been shown to exhibit an inflated type I error when applied to meta analyses of heterogenous studies, Peters et al. (2006). Without making strong assumptions about the publication mechanism in operation, attempting to disentangle these two phenomena will clearly be problematic.

### 3.3 Using pseudo-studies to correct for publication bias

Consider a meta-analysis of $n$ studies in which each study reports the mean level, $y_i$, of some measurement together with its standard error $s_i$. If a fixed effects model is used and we are willing to accept that the variance, and therefore standard errors are known, as in Hardy and Thompson (1996), then estimation of the common mean, $\theta$, reduces to calculating the weighted average of the $y_i$'s, using weights that are inversely proportional to $s_i^2$.

Suppose however, that we believe that the $y_i$'s are only observed if they pass some selection criterion, synonymous with being a member of a set $C$; such as, $y_i$ is greater than a certain threshold or the ratio $y_i/s_i$ is greater than some value. The method of Clayton (2003) can be modified to enable us to estimate the common mean of all studies, $\theta$, while taking full account of the impact of the selection criterion $C$.

Assume that before selection a real study effect, which will now be referred to as $y_{i0}$, was generated from a normal distribution with mean $\theta$ and known standard error $s_i$, that is $y_{i0} \sim N(\theta, s_i^2)$. The complex distribution of this effect size can be expressed, in a form
similar to (2.1.6), as

$$f(y_{i0}|s_i, \theta) = \frac{h(y_{i0}|s_i, \theta)}{z(\theta|s_i)}$$

where $z(\theta|s_i)$ represents the probability of publication, or membership of $C$. A guess, $\theta'$, is made for the common mean and a new value $y_{i1}$ is generated from the distribution $N(\theta', s_i^2)$. This value is then compared to the selection criterion $C$. If it passes then it is accepted, if not the process is repeated, potentially several times, until a value $y_{i1}$ is generated that is accepted. In this way the original real meta-data $y_{i0}$, and pseudo-data $y_{i1}$, are matched through a shared standard error. Further sets of pseudo-data $y_{i2}, y_{i3}, \ldots, y_{im}$ are then generated to improve the performance of the algorithm. Given this mix of real meta-data generated with an unknown common mean $\theta$ and matching pseudo-data generated with a known common mean $\theta'$, Clayton's conditional likelihood

$$L_{cond}(\theta|y; s, \theta') = \prod_{i=1}^{n} \frac{h(y_{i0}|s_i, \theta)}{h(y_{i0}|s_i, \theta')} \frac{h(y_{i1}|s_i, \theta)}{h(y_{i1}|s_i, \theta')} + \sum_{j=1}^{m} \frac{h(y_{ij}|s_i, \theta)}{h(y_{ij}|s_i, \theta')}$$

(3.3.1)

can be maximised in order to estimate $\theta$, as shown in Chapter 2.

In practice random effects models are usually applied to meta-data, which requires two parameters to be estimated; the common mean $\theta$ and the between study standard deviation $\tau$. The model assumption now made is that $y_i \sim N(\theta_i, s_i^2)$ and $\theta_i \sim N(\theta, \tau^2)$. The algorithm follows in a very similar way to that for the fixed effects model except that now guesses are made for $\theta'$ and $\tau'$. The pseudo-studies for real study $i$ are generated from a normal distribution with mean $\theta'$ and variance $s_i^2 + \tau'^2$ but they are subjected to the selection criterion $C$ as before. Whilst $s_i^2$ is clearly an estimate, it has become standard
practice to assume, as we do here, that it is known. In this case the conditional likelihood will become

$$L_{\text{cond}}(\theta, \tau | y; s, \theta', \tau') = \prod_{i=1}^{n} \frac{h(y_{i0} | s_i, \theta, \tau)}{h(y_{i0} | s_i, \theta', \tau')} + \sum_{j=1}^{m} \frac{h(y_{ij} | s_i, \theta, \tau)}{h(y_{ij} | s_i, \theta', \tau')}$$ (3.3.2)

As stated in Chapter 2, the efficiency of this method will depend in part on the quality of the guesses at the unknown parameters, and partly on the number, $m$, of pseudo-data sets. This is now investigated, in the context of meta-analysis, by simulation.

### 3.4 Simulation studies

A simulation study was used to investigate the performance of Clayton's pseudo-data method as a way of correcting for publication bias. The computer code used to generate the real and pseudo-data, and to fit Clayton's conditional likelihood can be found in appendix A.1.1. Clayton (2003) calculates the first and second derivatives of the conditional likelihood and uses the Newton-Raphson algorithm for the maximisation. It is slightly slower, but much more robust, and easier to program, if a general maximisation algorithm such as `optim()`, (R Development Core Team, 2004) or `ml()`, (StataCorp, 2003), is used. These results were calculated using `optim()`.

#### 3.4.1 Average performance

In order to reflect the situation found in the homocysteine example, meta-analyses of $n$ studies were simulated in which each study reported a mean difference between two groups. The two groups were simulated to have the same number of subjects, $n_i$, drawn from an exponential distribution with mean 100. Individuals from the first group were simulated...
from a normal distribution with mean 10 and standard deviation 4, individuals from the second group were simulated from a normal distribution with mean 7 and standard deviation 4. Summary statistics were calculated for each study in the form of the observed mean difference and its standard error, $s_i$. Assuming a fixed effects structure, the data were used to estimate the true difference between the two groups, $\theta$; the correct value being 3.

The simulated studies were subjected to a publication criterion that required either an estimated difference greater than 2.7 or a significant p-value ($<0.05$) in a Z-test for the null hypothesis $\theta_i = 0$.

$m$ pseudo-studies were generated for each 'real' study. The pseudo-data were drawn from a normal distribution with mean $\theta'$ and variance $s_i^2$. The assumed mean for the pseudo-data, $\theta'$, was taken to be the weighted average difference in the 'real' data without adjustment for publication bias, this was approximately 3.4. The pseudo-studies were subjected to the same publication criterion as the real trials. That is, $C$ was assumed to be known exactly. The conditional likelihood was maximized to create the publication-adjusted estimate of $\theta$.

Three different sizes of meta-analyses ($n=10, 25$ and $50$) were simulated for each of three different numbers of pseudo-studies ($m=1, 5$ and $10$). Two hundred simulations were performed for each of the nine scenarios. Figure 3.3 shows the mean, publication bias corrected, estimate for $\theta$ for each scenario, as well as an average confidence interval, that is, $\pm 1.96$ times the average standard error of the estimate. As the size of the meta-analysis $n$ increases the precision of the estimates increases sharply. There is also a gain in precision
3.4. Simulation studies

3.4.2 The effect of updating $\theta'$

To illustrate the impact of having a good initial guess $\theta'$ on the final estimates, I considered the same scenario as in Section 3.4.1 but fixed $m$ at 5. Two sets of pseudo-data were generated, one with $\theta'$ equal to the true adjusted value, 3, and one with $\theta'$ consistent with the biased data, approximately 3.4. Both simulations were repeated 2000 times. For both
3.4. Simulation studies

Figure 3.4: The 95 percent confidence interval widths for $\hat{\theta}$ in 4000 simulations for $\theta = 3$ when $\theta' = 3.4$ (simulations 1:2000) or $\theta' = 3$ (simulations 2001:4000) Horizontal lines indicate the means.

The average estimates of $\theta$ were very close to 3. Figure 3.4 shows for each $\theta'$ the widths of the 2000 confidence intervals for $\theta$. The 95% confidence intervals were noticeably narrower when the correct value was used for $\theta$. The coverages of the two sets of intervals were both close to their nominal values; 94.25% for $\theta' = 3$ and 95.25% for $\theta' = 3.4$. It appears that the closer your guess $\theta'$ is to the true value of $\theta$, the higher the precision of your estimate with no loss in coverage. This suggests a two-stage procedure; guess $\theta'$ and use that guess to find the adjusted estimate, $\hat{\theta}_0$, then repeat the analysis with $\theta' = \hat{\theta}_0$ in order to obtain the final estimate and its confidence interval.
3.4.3 Variation between different pseudo-data sets

Using the same basic scenario as in the previous section with $n=50$ and $m=5$, a single real data set was generated and then re-analysed 200 times using different sets of pseudo-data. $\theta'$ was based on the unadjusted analysis and turned out to be 3.43.

Figure 3.5: Histograms of the distribution of $\hat{\theta}$ when one data set is analysed with each of 200 different pseudo-data sets. When $\theta'=3.4$, the average estimate for $\theta$ was 3.00. Consequently when $\theta'=3$, the average estimate was 3.01. Although this might appear to be less accurate, with respect to the true value of 3, it more accurately reflects the true answer for this particular data set.

Figure 3.5 shows the distribution of $\hat{\theta}_0$ after one maximization of the conditional likelihood. As the meta-analysis is large it is not surprising that the average estimate for our particular real data set is close to 3. In an attempt to improve the precision, the two-stage process described in Section 3.4.2 was used. Figure 3.5 also shows the distribution of the estimates for $\hat{\theta}$, obtained from this second stage. The centre of the distribution is still close to the true value but the amount of variability is much less demonstrating how the two-stage estimates reduce the variability due to the randomness in the pseudo-data. Without the two-stage estimation, a much larger number, $m$, of pseudo-data sets would have been
required.

3.4.4 Misspecification of the selection criterion

To investigate the effect of misspecification, 200 meta-analyses were simulated. Each contained 50 studies with a mean difference of $\theta = 3$ that had passed a simple selection criterion requiring a significant effect (p-value $\leq 0.05$). $m=5$ pseudo studies were generated for each original study but assuming a different p-value-related publication criteria (publish if p-value $\leq 0.02, 0.04, 0.05, 0.06$ or $0.08$). Table 3.1 shows, for each of the five assumed criteria, the average estimate $\hat{\theta}$ and its standard error. When the p-value for publication is assumed too low, so is the average $\hat{\theta}$. In other words, an over-correction is made. When the p-value is assumed too high, so is $\hat{\theta}$, due to under-correction. Table 3.1 also shows that when the assumed criterion implies the lowest publication probability (p-value $\leq 0.02$) the standard deviation of the estimates for $\hat{\theta}$ is largest. This has the effect of improving the coverage of the confidence intervals for $\hat{\theta}$ and suggests that, at least in this example, assuming a publication criterion that is too severe is preferable to assuming one that is too lenient.

<table>
<thead>
<tr>
<th>Misspecified p-value</th>
<th>Mean difference $\hat{\theta}$</th>
<th>Average 95% confidence interval width</th>
<th>Proportion of 95% intervals that include $\theta = 3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.02</td>
<td>2.92</td>
<td>0.57</td>
<td>0.94</td>
</tr>
<tr>
<td>0.04</td>
<td>2.99</td>
<td>0.53</td>
<td>0.95</td>
</tr>
<tr>
<td>0.05</td>
<td>3.00</td>
<td>0.55</td>
<td>0.94</td>
</tr>
<tr>
<td>0.06</td>
<td>3.03</td>
<td>0.53</td>
<td>0.94</td>
</tr>
<tr>
<td>0.08</td>
<td>3.07</td>
<td>0.50</td>
<td>0.91</td>
</tr>
</tbody>
</table>

Table 3.1: The effect that misspecifying the p-value ($\leq 0.05$) required for publication has on the resulting estimate for $\theta$ and its observed standard error (true criteria $p \leq 0.05$, true $\theta = 3$).
3.5 Factors influencing study selection

It has been established that Clayton’s conditional likelihood approach can work if the selection criterion $C$ is known, while in fact the process that controlled the publication of the studies will always have to be guessed at. Typically each study calculates a measure of the size of the effect and an associated p-value, confidence interval or standard error. It is quite likely that the chance of publication depends on the size of the effect and its significance. Studies that find a small non-significant effect size are less likely to be published and so quite often the unpublished studies are small. A cynical but common opinion is that, when they produce a result that is contrary to an investigator’s beliefs, small trials are also easier to ‘bury’.

Basing $C$ on the size of the effect and its significance may capture the most important influences on publication but the reality is certainly more complex. Factors such as the country in which the study was performed, the research record of the main investigators and the importance of the topic might all influence publication. Ioannidis (1998) showed that after the publication of a novel strong effect there is a high chance that the next published study will be negative. Presumably the negative data were not considered sufficiently interesting until they could be claimed to contradict the new positive result. In a similar way, once an association has been established in the literature, studies that question that association will be harder to publish; indeed there may even be a degree of self-censorship.

Selective reporting also enters the criterion $C$. An incidental finding from a study may
be included in a paper if it is large or highly significant but omitted if it is less striking or if it does not fit with previous literature. Thus, considering the example of the MTHFR gene, one might imagine a study that finds a strong link with cardiovascular disease but in a small sub sample of subjects finds no association between MTHFR and homocysteine level. There would be a temptation not to report the negative finding for fear of bringing the positive result into question. We will never know what the true criteria governing dissemination (publication and reporting) or even whether the concept of a 'true publication criterion' is meaningful in some Platonic form. But the strength of the pseudo data approach is that the criteria can be varied in a sensitivity analysis and they can be made as complex as one wishes without adding any excessive computational demands.

3.5.1 Extensions to the basic selection model

By writing the publication probability as \( z(\theta, \tau|s) \) I am assuming that the publication model is 'complete'. That is studies are always published if and only if they meet the criterion \( C \). Whilst it may be reasonable to make such an assumption in genetic epidemiology, when the selection model is sometimes unambiguous and clear, more flexibility is certainly required when thinking about possible publication models in meta-analysis. In this context rewriting equation (2.1.4) from Chapter 2

\[
p(\text{Published}|s_i, \theta, \tau) = \pi_1 \int_{y_i \in C} h(y_i|s_i, \theta, \tau)dy_i + \pi_2 \int_{y_i \notin C} h(y_i|s_i, \theta, \tau)dy_i \quad (3.5.1)
\]

so when the publication model is complete \( \pi_1 = 1, \pi_2 = 0 \), the probability of publication simplifies to \( \int_{y_i \in C} h(y_i|s_i, \theta, \tau)dy_i = z(\theta, \tau|s_i) \), as in the denominator of equation (3.3.2).

Under a complete selection function, for pseudo-data that appears on the \( i \)'th denominator
of Clayton’s conditional likelihood

\[
\sum_{j=1}^{m} \frac{h(y_{ij}|s_i, \theta, \tau)}{h(y_{ij}|s_i, \theta', \tau')} \approx m \frac{z(\theta, \tau|s_i)}{z(\theta', \tau'|s_i)}
\]

which is the unknown selection probability \(p(\text{Published}|s_i, \theta, \tau)\) up to a constant (\(\frac{m}{z(\theta', \tau'|s_i)}\) being this constant). It would also be advantageous to be able to investigate publication models in which studies have a small chance of being published even if they do not meet the specific selection criterion \(C\). When this is the case, \(\pi_1 = 1, \pi_2 \neq 0\) and the distribution of the published studies becomes

\[
f(y_i|s_i, \theta, \tau) = \frac{h(y_i|s_i, \theta, \tau)p(\text{Published}|y_i; s_i, \theta, \tau)}{p(\text{Published}|s_i, \theta, \tau)} \tag{3.5.2}
\]

where \(p(\text{Published}|y_i; s_i, \theta, \tau)\) is either 1 or \(\pi_2\), depending on whether or not \(y \in C\), and

\[
p(\text{Published}|s_i, \theta, \tau) = \pi_1 \int_{y_i \in C} h(y_i|s_i, \theta, \tau)dy_i + \pi_2 \int_{y_i \notin C} h(y_i|s_i, \theta, \tau)dy_i
\]

\[
= \int_{y_i \in C} h(y_i|s_i, \theta, \tau)dy_i + \pi_2 \left\{ 1 - \int_{y_i \in C} h(y_i|s_i, \theta, \tau)dy_i \right\}
\]

\[
= z(\theta, \tau|s_i)(1 - \pi_2) + \pi_2 \tag{3.5.3}
\]

If this is truly the publication model, then as long as pseudo-data is simulated given the correct \(C\) and \(\pi_2\), then

\[
\sum_{j=1}^{m} \frac{h(y_{ij}|s_i, \theta, \tau)}{h(y_{ij}|s_i, \theta', \tau')} \approx m \frac{z(\theta, \tau|s_i)(1 - \pi_2) + \pi_2}{z(\theta', \tau'|s_i)(1 - \pi_2) + \pi_2}
\]

which also equals the unknown selection probability \(p(\text{Published}|s_i, \theta, \tau)\), up to a con-
3.6 A simple criterion

As long as the pseudo-data augmented with real study $i$ is subjected to the same selection process, publication probabilities, and additionally the constant $\pi_2$, cancel in a conditional likelihood as before.

Furthermore, one may feel it is also unrealistically restrictive to assume that every published study in a meta-analysis has met the same publication criterion. If study $i$ is subjected to a selection model making it a member of a unique but known set, $C_i$, and the complex distribution for study $i$, $f_i(y_i|s_i, \theta, \tau)$, can be expressed as

\[
\frac{h(y_i|s_i, \theta, \tau)}{z_i(\theta, \tau|s_i)}
\]

then it is immediately clear that all of the publication probabilities $z_1(\theta, \tau|s_1), ..., z_n(\theta, \tau|s_n)$ would cancel in Clayton’s conditional likelihood. This means there is a considerable amount of flexibility that can be utilised when tailoring the pseudo data approach to meet the needs of particular meta-analyses, depending on what publication process was thought to have operated.

3.6 A simple criterion applied to the homocysteine data

Minelli et al. (2004) estimated the average difference in the homocysteine levels between TT and CC subjects, $\theta$, to be 2.7 $\mu$mol/l (95% C.I 2.1, 3.4) and estimated the odds ratio for the effect of TT versus CC genotype on cardiovascular disease, $b$, to be 1.21. Combining these two estimates in a ‘Mendelian Randomisation’ analysis, with no specific
account being taken of publication bias, they calculated an estimate for the odds ratio of cardiovascular disease for a unit change in homocysteine to be

\[ b^{1/\theta} = 1.21^{1/2.7} \approx 1.07 \]

see Figure 3.1. Alternatively, a random effects model could be fitted to the data to estimate \( \theta \), accounting for the publication bias that appears to be present. Assuming that the true average difference in homocysteine for each study is normally distributed, 

\[ y_{i0} \sim N(\theta, s_i^2 + \tau^2), \quad i = 1, ..., 33, \]

where \( s_i^2 \) is the 'known' within-study variance and \( \tau^2 \) is the unknown between-study variance. I chose \( \theta' \) to be the pooled, weighted estimate of the average homocysteine difference, which is \( 1.954 \, \mu \text{mol/l} \) and \( \tau'^2 \) with the mean of the ten most precise average difference estimates, which was \( 1.05 \, \mu \text{mol}^2/\text{l}^2 \).

A plausible, but simple publication criterion \( C \) based solely on the p-value is suggested. Imagine that studies are always published whenever their p-value is \( \leq 0.05 \), if their p-value is \( > 0.05 \) then a study may still be published but with a decreased probability. This publication model is equivalent to equation (3.5.3). Including \( \tau_2 \), the publication probability for a non-significant study, in the criterion appears to be sensible. This is because 13 out of the 33 published trials in the MTHFR meta-analysis have a p-value greater than 0.05. In a sensitivity analysis \( \tau_2 \), is allowed to vary between 0.1 and 0.5.

Each study's mean difference, \( y_{i0} \), ten pseudo-data values \( y_{i1}, y_{i2}, ..., y_{i10} \) were drawn from a normal distribution with mean \( \theta' \) and variance \( s_i^2 + \tau^2 \). The pseudo-data were subjected to the criterion \( C \) and the data that failed the criterion were replaced. Maximisation of the conditional likelihood gives the results shown in Table 3.2. The estimates for the average
### 3.7. Complex criteria

<table>
<thead>
<tr>
<th>Probability of publication for non significant p-value ($\pi_2$)</th>
<th>Mean difference in Homocysteine ($\hat{\theta}$)</th>
<th>Between-study standard deviation ($\hat{\tau}$)</th>
<th>Estimate of phenotype-disease odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.32(0.24)</td>
<td>1.42(0.10)</td>
<td>1.16</td>
</tr>
<tr>
<td>0.2</td>
<td>1.60(0.17)</td>
<td>1.39(0.22)</td>
<td>1.13</td>
</tr>
<tr>
<td>0.3</td>
<td>1.80(0.17)</td>
<td>1.41(0.24)</td>
<td>1.11</td>
</tr>
<tr>
<td>0.4</td>
<td>1.97(0.14)</td>
<td>1.42(0.10)</td>
<td>1.10</td>
</tr>
<tr>
<td>0.5</td>
<td>2.10(0.14)</td>
<td>1.40(0.17)</td>
<td>1.10</td>
</tr>
</tbody>
</table>

Table 3.2: Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation, when the publication criterion is assumed to depend on significance.

difference in homocysteine between the TT and CC groups were between 1.32 and 2.10 $\mu$mol/l depending on the strength of the selection criterion reflected in the parameter $p$. The effect of $C$ can also be seen in the standard errors that become larger as the selection becomes more severe. In this example, the heterogeneity estimate is much less sensitive to changes in C. The final column of Table 3.2 shows the impact that publication bias has on the Mendelian randomisation estimates of Minelli et al. (2004). The odds ratio of a unit change in homocysteine on cardiovascular disease is almost certainly larger than they suggested with adjusted estimates ranging from 1.10 to 1.16.

### 3.7 Complex publication criteria

In this section, the selection criterion $C$ is allowed to depend on the measured effect size as well as the p-value. Next it is allowed to vary between studies.

#### 3.7.1 A criterion incorporating effect size

First, it is assumed that significant studies ($p \leq 0.05$) are always published and that non-significant studies are only published if they have an effect size that exceeds a threshold,
3.7. Complex criteria

$T$. Non-significant studies that have a small effect ($<T$) are also allowed to be published, with a constant probability $\pi_2$ of 0.05. In a sensitivity analysis the threshold effect size, $T$, at which a study becomes certain to be published was varied between 0 and 1 $\mu$mol/l in steps of 0.2. Thus at one extreme, any positive association is publishable while at the other, only strong effects are certain to be published when they are non-significant. Table 3.3 details the results of the sensitivity analyses. The corrected estimates for $\theta$ lie between 1.37 and 2.08 with greater uncertainty when the criterion $C$ is most severe. The estimates for the between-study variance lie between 2.99 and 3.65. This in turn causes the inferred phenotype-disease odds ratio to vary between 1.10 and 1.15.

<table>
<thead>
<tr>
<th>Assumed effect size publishable $T$</th>
<th>Mean difference in Homocysteine $\hat{\theta}$</th>
<th>Between-study standard deviation $\hat{\tau}$</th>
<th>Estimate of phenotype-disease odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>2.08(0.30)</td>
<td>1.73(0.37)</td>
<td>1.10</td>
</tr>
<tr>
<td>0.2</td>
<td>1.98(0.32)</td>
<td>1.82(0.45)</td>
<td>1.10</td>
</tr>
<tr>
<td>0.4</td>
<td>1.86(0.37)</td>
<td>1.80(0.33)</td>
<td>1.11</td>
</tr>
<tr>
<td>0.6</td>
<td>1.58(0.58)</td>
<td>1.91(0.57)</td>
<td>1.13</td>
</tr>
<tr>
<td>0.8</td>
<td>1.53(0.49)</td>
<td>1.42(0.44)</td>
<td>1.13</td>
</tr>
<tr>
<td>1.0</td>
<td>1.37(0.14)</td>
<td>1.40(0.54)</td>
<td>1.15</td>
</tr>
</tbody>
</table>

Table 3.3: Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation, when publication is assumed to be dependent on significance and a minimum effect size $T$.

3.7.2 Study-specific criteria

The assumed selection models in Sections 3.6 and 3.7.1 were not study-specific since they assumed that all studies were subject to the same selection criterion. One could argue that it is unreasonable to assume each study has to meet the same standards; since they are reported in a range of journals, by different authors and at different times. Rather than assuming that a general criterion exists for all studies, the publication criteria could be
allowed to vary depending on the character of the editor who reviewed them. I originally proposed to group all studies into one of four classes, according to study specific $p_i$ and $d_i$, where

- $p_i = 0$ if the p-value for trial $i$ is greater than 0.05, and 1 otherwise.
- $d_i = 0$ if the average difference in homocysteine for trial $i$ is less than some specified threshold $T$, and 1 otherwise.

Given the real homocysteine effect sizes, $y_{i0}$, their standard errors $s_i$ and a choice for $T$, study $i$ can be assigned ‘covariates’ $p_i$ and $d_i$ taking the values (0,0), (1,0), (0,1) or (1,1).

I then wanted to assume that the population of journal editors could be divided into 4 groups denoted by; $E_1$, $E_2$, $E_3$ and $E_4$, where

- $E_1 =$ All editors that will publish a paper independently of the values of $p$ and $d$
- $E_2 =$ All editors that will only publish a paper with a ‘significant’ p-value ($p = 1$)
- $E_3 =$ All editors that will only publish a paper with a ‘large’ effect ($d = 1$)
- $E_4 =$ All editors that will only publish a paper with a ‘significant’ p-value and ‘large’ effect ($p = 1, d = 1$)

As discussed in Section 3.5, it is a common belief that large studies are more likely to be published than small studies and I wanted this to be reflected in the study specific publication model. However, since large studies have an increased chance of reporting a significant p-value, they are more likely to have a $p$ covariate of 1. The implicit correlation between study size and statistical significance meant that I thought it unnecessary to
further characterise each study as 'small' or 'large' with an additional covariate.

The existence of multiple criteria will mean that real studies have the potential to be from one of four distributions

\[ f_c(y|\theta, \tau, s) = \frac{h(y|\theta, \tau, s)}{z_c(\theta, \tau, s)} \quad c = 1, \ldots, 4 \]  \hspace{1cm} (3.7.1)

Although all distributions share a common numerator, their denominators will be different, in each case

\[
\begin{align*}
  z_1(\theta, \tau|s) &= \int_{(0,0)\cup(1,0)\cup(0,1)\cup(1,1)} h(y|\theta, \tau, s)dy = 1 \\
  z_2(\theta, \tau|s) &= \int_{(1,0)\cup(1,1)} h(y|\theta, \tau, s)dy \\
  z_3(\theta, \tau|s) &= \int_{(0,1)\cup(1,1)} h(y|\theta, \tau, s)dy \\
  z_4(\theta, \tau|s) &= \int_{(1,1)} h(y|\theta, \tau, s)dy
\end{align*}
\]

where for example \( \int_{(0,1)} \) denotes the integral over all values of \( y \) which have covariate values \((0,1)\). Figure 3.6 (Top) shows the extent of what can be inferred about the possible distribution of a real study, given that its characteristics \( p, d \) can be observed.

Under this formulation, if a study has observable covariates \((0,0)\), it is possible to say for certain, that \( f_1 \) is the correct distribution that generated it. However, for any other covariate values, there is not a one-to-one correspondence between themselves and a complex distribution \( f_c \), and so it's distribution will be unknown. For example, if a study has covariates \((0,1)\) it could have been generated from either \( f_1 \) or \( f_2 \). Clayton's algorithm
Figure 3.6: Top: A schematic diagram that represents the true relationship between the unknown distribution of the real data \((f_1 \text{ to } f_4)\) and the assumed observable study covariates \((p, d)\). Bottom: A schematic diagram that represents a simplified, one-to-one relationship between study covariates and complex distributions, that can be dealt with using Clayton’s conditional likelihood.

will work under the assumption that the real studies have passed a known publication criterion, as this is then used to simulate pseudo studies under the same conditions. If this is not the case, the required cancelations will not occur and a conditional likelihood of the form shown in equation (3.3.2) can not be constructed. In an effort to make the study specific method tractable using Clayton’s conditional likelihood, I created a one-to-one correspondence between all covariate combinations and a publication criterion, as illustrated in Figure 3.6 (Bottom).

If the real homocysteine trial \(i\) has values \((0,0)\) then I assumed that because it does not have a ‘large’ effect size nor is it significant, it has not been subject to any selection process. This corresponds to being from \(f_1\), and so pseudo-data are also simulated under no selection mechanism. If the real homocysteine trial \(i\) has values \((1,0)\), I assumed that it was from \(f_2\) and corresponding pseudo-data is simulated conditional on having a sig-
nificant p-value. Conversely, if the real homocysteine trial $i$ has values (0,1), I assumed that the trial was published by editor $E_3$ due to its effect size. This corresponds to being from $f_3$, and so pseudo-data are accepted if they have a large enough effect, regardless of statistical significance. If the published study was both significant and had a large effect size, implying covariate values (1,1), I assumed that it was from $f_4$, and so corresponding pseudo-data must also have (1,1) covariate values. Finally, I added a random component to trial publication by giving each pseudo-study a small probability, $\pi_2=0.05$, of being published, regardless of its outcome.

$T$ is varied between 0 and 1 in steps of 0.2 as before. Throughout this range of $T$, the number of trials falling into each category only varies slightly; 1-3 trials are categorized as (0,0), 0 or 1 trials are categorized as (0,1), 11 or 12 are categorized as (1,0) and 19 or 20 trials are categorized as (1,1). A plot detailing the change in the frequency of group memberships as a function of $T$ can be found in Appendix A.1. However, whilst varying $T$ between 0 and 1 varies the publication category very little, it will alter the distribution of pseudo-trials considerably because the value of $T$ is used to generate the pseudo-data. So augmenting our real data with varying types of pseudo-data still gives us an array of estimates for the true mean difference $\theta$. Table 3.4 shows the results of 100 simulations under these six scenarios.

The estimate, $\hat{\theta}$, is 1.20 when $T$ is 0 and between 1.00 and 1.06 when $T$ is between 0.2 and 1. The estimates, $\hat{\theta}$, vary very little with no discernable trend. Adopting this study-specific publication criterion causes us to reduce markedly our estimate for the true mean homocysteine difference. It causes the phenotype-disease odds ratio to be estimated between 1.17 and 1.21.
### 3.8 Discussion

Publication bias is one of the major problems in meta-analysis and it is likely that all systematic reviews are affected to some extent by publication and reporting bias. Tests based on the funnel plot will usually have low power and so publication bias that might have a material impact on the final result can easily be missed. By the very nature of the problem we do not observe the studies that are not published and it would be very difficult to investigate the selection process directly. As a result, selection models always rely on a series of untestable assumptions and it is imperative that sensitivity analyses are used to assess the robustness of the final conclusions to changes in these assumptions.

---

<table>
<thead>
<tr>
<th>Assumed effect size publishable $T$</th>
<th>Mean difference in Homocysteine $\hat{\theta}$</th>
<th>Between-study standard deviation $\hat{\tau}$</th>
<th>Estimate of phenotype-disease odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>1.20(0.28)</td>
<td>1.71(0.17)</td>
<td>1.17</td>
</tr>
<tr>
<td>0.2</td>
<td>1.06(0.36)</td>
<td>1.78(0.22)</td>
<td>1.20</td>
</tr>
<tr>
<td>0.4</td>
<td>1.06(0.26)</td>
<td>1.74(0.17)</td>
<td>1.20</td>
</tr>
<tr>
<td>0.6</td>
<td>1.05(0.32)</td>
<td>1.75(0.17)</td>
<td>1.20</td>
</tr>
<tr>
<td>0.8</td>
<td>1.00(0.32)</td>
<td>1.76(0.17)</td>
<td>1.21</td>
</tr>
<tr>
<td>1.0</td>
<td>1.04(0.33)</td>
<td>1.71(0.17)</td>
<td>1.20</td>
</tr>
</tbody>
</table>

Table 3.4: *Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation, when a study-specific publication criterion is assumed (see Section 3.7 for details).*

Figure 3.7 shows the publication bias corrected estimates for $\theta$ obtained by all the publication models described. A similar plot for the between study standard deviation estimates can be found in appendix A.2.
Figure 3.7: Publication bias corrected estimates for the difference in homocysteine and their 95% confidence intervals when analysed by Trim and Fill, (Duval and Tweedie, 2000), and using pseudo-data with different criteria as described in Sections 3.6, and 3.7.

The use of pseudo-data offers a novel method of correcting for publication bias that is easy to implement even when the publication criterion is complex. Extensions of the method to the meta-analyses of log odds ratios and other statistics that have an approximate normal distribution are immediate. More complex distributions require minor changes to the conditional likelihood and may be more difficult to simulate but present no theoretical problems. For instance, in a meta-analysis of odds ratios in which some studies are small and contain zero cells, it would be better to use a binomial model, as in Shi and Copas (2002).

The analysis of the homocysteine data and the simulations suggest that stable results can be obtained by using five pseudo-observations matched to each real study together
with a two-stage analysis in which the initial guess used to generate the pseudo-data is updated after the first analysis. In most applications, publication criteria based on significance and effect size will be important but, if other more complex criteria are thought to have operated, then these can be investigated without any added practical problems.

For the homocysteine meta-analysis, our results suggest that, if one is willing to accept that its studies have passed some sort of publication criteria, then the true average difference in homocysteine levels between the TT and CC genotypes has probably been overestimated. An important consequence of this is that, in a Mendelian randomization analysis, because the genotype-phenotype relationship is overestimated, the derived phenotype-disease relationship will have been underestimated. All of the corrected phenotype-genotype odds ratio estimates are larger than the uncorrected estimate used by Minelli et al. (2004). Although adjusting for publication bias usually moves estimates towards the null, this example illustrates how adjustment can, in some circumstances, make a derived statistic larger.
Chapter 4

Using pseudo-data to validate proposed selection criteria

Simulations show that importance sampling methods, such as Clayton’s conditional likelihood, can produce asymptotically unbiased results when pseudo-data are generated under a known selection criterion. In reality, when analysing a real meta-data set, such as the homocysteine example in the previous chapter, we will have to make an educated guess at the true publication criterion, and there will always be some uncertainty about this guess. In order to begin to acknowledge this uncertainty, a sensitivity analysis was performed, by considering a range of possible publication criteria that could have operated on the studies. Is this all that can be done? In this chapter I explore, in a simple example, what occurs when selection biased data are analysed under a misspecified selection model. A simulation-based approach is proposed to detect this misspecification and the method is then applied to the homocysteine data. I presented the initial ideas in this chapter at the International Society for Clinical Biostatistics conference in 2006.
4.1 Complete selection model misspecification

Keeping the notation of Chapter 3, the most basic assumption was that the statistical model generating a study’s effect, \( y \), conditional on passing a certain publication criterion \( C \), could be modeled as

\[
f(y|s, \theta) = \frac{h(y|s, \theta)}{\int_{y \in C} h(y|s, \theta) dy}
\]

This formulation assumed that publication was a certainty for study \( y \), if it passed a selection criteria making it a member of \( C \), and had zero probability otherwise. If the publication mechanism was understood, that is the rules for membership of \( C \) were known, then an estimate for \( \theta \) could be derived by maximising the true log-likelihood

\[
l(\theta|y; s, C) = \sum_{i=1}^{n} \left\{ \log(h(y_i|s_i, \theta)) - \log \int_{y_i \in C} h(y_i|s_i, \theta) dy_i \right\}
\]

providing the integration over \( C \) is tractable. This maximisation is identical to finding the \( \hat{\theta} \) that solves the score equation

\[
S(\theta|y, s, C) = \sum_{i=1}^{n} \{ S(\theta|y_i; s_i) - b(\theta|s_i, C) \} = 0
\]

Where \( S(\theta|y_i; s_i) = d \log \{ h(y_i|s_i, \theta) \} / d\theta \) and

\[
b(\theta|s, C) = d \log \left\{ \int_{y_i \in C} h(y_i|s_i, \theta) dy_i \right\} / d\theta = \frac{\int_{y_i \in C} S(\theta|y_i; s_i) h(y_i|s_i, \theta) dy_i}{\int_{y_i \in C} h(y_i|s_i, \theta) dy_i}
\]

\( S(\theta|y; s) \) can be thought of as the 'crude' score function. If \( y \) is certain to be selected, then \( S(\theta|y; s, C) = \sum_{i=1}^{n} S(\theta|y_i; s_i) \) and \( b(\theta|s, C) \) would equal 0. This is identical to a
set \( C \) that encompasses all possible values of \( y \). When a non-trivial selection criterion \( C \) is in place, then \( \hat{\theta} \) is chosen so that \( \sum_{i=1}^{n} S(\hat{\theta}|y_i, s_i) = \sum_{i=1}^{n} b(\hat{\theta}|s_i, C) \). This tells us two things. Firstly, the selection mechanism uniquely determines the size of the bias in the uncorrected parameter estimate for \( \theta \). Secondly, in order to consistently correct for this bias the selection criterion must be correctly specified.

Let us imagine that it is incorrectly assumed selection is solely dependent on membership of a different set \( C_1 \), not \( C \). Figure 4.1 is a Venn diagram illustration of these two sets over the state space of \( y \) with respect to the density \( h \). Some \( y \)'s will be a member of both sets, some will be a member of one set and not the other. To understand how statistical inferences will alter when the selection criteria is misspecified in this way, it is noted that

\[
I(\theta|y; s, C_1) = \sum_{i=1}^{n} \left\{ \log(h(y_i|s_i, \theta)) - \log \int_{y_i \in C_1} h(y_i|s_i, \theta) dy_i \right\}
= \sum_{i=1}^{n} \left\{ \log(h(y_i|s_i, \theta)) - \log \int_{y_i \in C} h(y_i|s_i, \theta) dy_i - \log \frac{\int_{y_i \in C_1} h(y_i|s_i, \theta) dy_i}{\int_{y_i \in C} h(y_i|s_i, \theta) dy_i} \right\}
\]

(4.1.3)
4.2 Misspecification of threshold criteria

where

\[
\int_{y \in C} h(y|s, \theta) dy = \int_{y \in C} h(y|s, \theta) dy + \int_{y \in (C \cap C')} h(y|s, \theta) dy - \int_{y \in (C \cap C')} h(y|s, \theta) dy
\]

Where \( C^c \) means the complement of \( C \). The score function that will unwittingly be solved to obtain \( \hat{\theta} \) is now

\[
S(\theta|y, s, C_1) = \sum_{i=1}^{n} \{ S(\theta|y_i, s_i) - (b(\theta|s_i, C) + \delta(\theta|s_i, C, C_1)) \} = 0 \tag{4.1.4}
\]

where \( \delta(\theta|C, C_1, s) = b(\theta|C_1, s) - b(\theta|C, s) \). If the integral over \( C_1 \) is larger than the integral over the true region \( C \), then \( \delta(\theta|C, C_1, s) \) will be positive. If the converse is true, then \( \delta(\theta|C, C_1, s) \) will be negative. \( C_1 \) is a 'larger' set than \( C \), with respect to \( h \) when

\[
\int_{y \in (C \cap C')} h(y|s, \theta) dy > \int_{y \in (C \cap C')} h(y|s, \theta) dy.
\]

4.2 Misspecification of threshold criteria

In this section the effect of selection model misspecification on estimating a parameter \( \theta \) is illustrated, given data \( y_1, ..., y_n \sim N(\theta, \sigma^2) \), where \( \sigma^2 \) is known. \( y_1, ..., y_n \) are assumed to have passed the same complete selection criterion \( C \), which equates to a threshold condition \( y \geq T \). The log-likelihood for the selected data is therefore

\[
l(\theta|y, \sigma, C) \approx \sum_{i=1}^{n} \left\{ \log \frac{1}{\sigma} \phi \left( \frac{y_i - \theta}{\sigma} \right) - \log \Phi \left( \frac{\theta - T}{\sigma} \right) \right\} \tag{4.2.1}
\]

Copas and Jackson (2004) use this example to illustrate publication biased data in a meta-analysis. If all studies were subject to the same, known variance, a simple threshold
4.2. Misspecification of threshold criteria

criterion for $y$ would equate to only being able to observe studies that are statistically significant at some predefined level. For example, if $\theta = 0$, $y$ would have to be $\geq 1.96\sigma$ to achieve statistical significance with a type one error of 5%.

Stallard et al. (2007) encounter this type of selection based data when considering the performance of two competing treatments, in the setting of analysing a sequential clinical trial. Given two treatments $Q_1, Q_2$, assessed using outcome measures $x_1, x_2$, where

$$
\begin{align*}
    x_1 &\sim N(\theta_1, \frac{\sigma^2}{2}) \\
    x_2 &\sim N(\theta_2, \frac{\sigma^2}{2})
\end{align*}
$$

treatment $Q_1$ is taken forward for further tests whenever $x_1 > x_2$. Designs of this sort are seen as a cost-efficient method of fast tracking the most promising compounds, in early stage drug development. However, conditional on being greater than $x_2$, $x_1$ is not an unbiased estimate for $\theta_1$. By defining another variable $y = x_1 - x_2$, Stallard et al. showed that the problem could be equally thought of as using $y$ to estimate $\theta = \theta_1 - \theta_2$, given $y > 0$. For an arbitrary threshold $T$, in order to calculate the size of the bias in $y$, it is necessary to calculate

$$
E[Y|Y \geq T] = \frac{\int_T^\infty \frac{y}{\sigma} \phi \left( \frac{y-\theta}{\sigma} \right)}{\Phi \left( \frac{\theta-T}{\sigma} \right)} = \frac{\theta - \int_{-\infty}^T \frac{y}{\sigma} \phi \left( \frac{y-\theta}{\sigma} \right)}{\Phi \left( \frac{\theta-T}{\sigma} \right)}
$$
and by using a result from Todd et al. (1996)

\[
\int_{-\infty}^{T} \frac{y}{\sigma} \phi \left( \frac{y - \theta}{\sigma} \right) dy = -\sigma \phi \left( \frac{T - \theta}{\sigma} \right) + \theta \Phi \left( \frac{T - \theta}{\sigma} \right) \quad (4.2.2)
\]

Stallard et. al show that

\[
E[Y|Y > T] = \theta + \frac{\sigma \phi \left( \frac{\theta - T}{\sigma} \right)}{\Phi \left( \frac{\theta - T}{\sigma} \right)} \quad (4.2.3)
\]

The score function for likelihood (4.1.1), given the selection model is specified correctly is therefore

\[
S(\theta|y, \sigma, C) = \frac{1}{\sigma^2} (y - \theta) - \frac{\sigma}{\sigma} \log \Phi \left( \frac{\theta - T}{\sigma} \right) / d\theta
\]

\[
= \frac{1}{\sigma^2} (y - \theta) - \frac{1}{\sigma} \phi \left( \frac{\theta - T}{\sigma} \right) \Phi \left( \frac{\theta - T}{\sigma} \right)
\]

Substituting in \(E[Y|Y > T]\) from (4.2.3), the expected value of the score function at \(\theta\) is

\[
E[S(\theta|y, \sigma, C)] = \frac{1}{\sigma^2} (\theta - \theta) + \frac{1}{\sigma} \left( \phi \left( \frac{\theta - T}{\sigma} \right) - \phi \left( \frac{\theta - T}{\sigma} \right) \right)
\]

\[
= 0 \quad (4.2.4)
\]

So, under correct selection model specification, consistent estimation of \(\theta\) is possible. The simple nature of this single threshold selection criterion allows us to propose an incorrect threshold \(T_1\) such that the set \(C_1 \subset C\), or alternatively, a threshold \(T_2\) such that \(C \subset C_2\), as Figure 4.2 describes. The integral over sets \(C_1\) and \(C_2\) with respect to \(h\) can then be
4.2. Misspecification of threshold criteria

written as

\[
\int_{y \in C_1} h(y|s, \theta) \, dy = \int_{y \in C} h(y|s, \theta) \, dy - \int_{y \in (C \cap C_2^c)} h(y|s, \theta) \, dy
\]

\[
\int_{y \in C_2} h(y|s, \theta) \, dy = \int_{y \in C} h(y|s, \theta) \, dy + \int_{y \in (C_2 \cap C_2^c)} h(y|s, \theta) \, dy
\]

For an arbitrary, incorrect threshold \( T_w \), \( E[S(\theta|y, \sigma, C_w)] \) will only equal 0 for a \( \hat{\theta}|C_w \), where

\[
\left( \frac{\phi\left( \frac{\hat{\theta}|C_w - T_w}{\sigma} \right)}{\Phi\left( \frac{\hat{\theta}|C_w}{\sigma} \right)} + \frac{\theta|C_w}{\sigma} \right) = \left( \frac{\phi\left( \frac{\theta - T}{\sigma} \right)}{\Phi\left( \frac{\theta - T}{\sigma} \right)} + \frac{\theta}{\sigma} \right)
\] (4.2.5)

Although selection bias is a problem in sequential clinical trials, selection model misspecification is not an issue, because it is the researcher who decides on the criteria for comparing treatment performance. In that respect, it is equivalent to the problem that sometimes exists in genetic epidemiology. However, as I have discussed in Chapter 3, there will always be much more uncertainty surrounding the publication model for a group of studies in a meta-analysis.

4.2.1 Conditional likelihood inference under selection model misspecification

In this thesis I am interested in exploring the behaviour of importance sampling based solutions to correct for selection bias. The effect of model misspecification on the performance of Clayton's conditional likelihood is now explored. Real data \( y_1, \ldots, y_{n=200} \) are simulated from a \( N(\theta = 3, 1) \) distribution, subject to the condition of being greater than
4.2. Misspecification of threshold criteria

Figure 4.2: A Venn diagram representation of three sets $C, C_1$, and $C_2$. Set $C$ contains all data points $y$ that pass the true selection criterion. Sets $C_1, C_2$ are the sets of all data points that pass proposed selection criteria. Set $C_1$ is a subset of $C$, which is itself a subset of $C_2$.

$C = 2.7$. The mean value is 3.61, which is approximately $3 + \frac{0.3}{\Phi(0.3)}$. For each real data point, $m = 5$ pseudo data points $y_{ij}$ are simulated from a $N(0^2 = 2, 1)$ distribution, subject to being $\geq 2.7$. The augmented real and pseudo-data allow us to approximate log-likelihood (4.1.1) with

\[
l_{\text{cond}}(\theta|y; C, \theta^*) = \sum_{i=1}^{n} \left[ \log \frac{\phi(y_i - \theta)}{\phi(y_i - \theta^*)} - \log \left\{ \frac{\phi(y_i - \theta)}{\phi(y_i - \theta^*)} + \sum_{j=1}^{m} \frac{\phi(y_{ij} - \theta)}{\phi(y_{ij} - \theta^*)} \right\} \right] \tag{4.2.6}
\]

The black line in Figure 4.3 represents the value of the crude log likelihood $\log \frac{\phi(y - \theta)}{\Phi(y - \theta^*)}$, as a function of $\theta$, the solid blue line represents the value of the importance sampling derived estimate for $\log \Phi(\theta - 2.7)$, as a function of $\theta$. The solid red line shows the sum of these two quantities, $l_{\text{cond}}(\theta|y; \theta^*, C)$, and since $C$ is correctly specified, $l_{\text{cond}}(\theta|y; C)$ is maximised for a $\hat{\theta}$ close to 3.

The dotted blue lines plot the importance sampling derived estimates for the selection
4.2. Misspecification of threshold criteria

Figure 4.3: Top: The effect of selection model misspecification on the resulting maximum conditional likelihood estimates for $\theta$, when criterion $C$ is assumed $\theta \approx 3$. When criteria $C_1$ and $C_2$ are assumed, $\theta$ is estimated to be 2.4 and 3.5 respectively.

The dotted red lines show where the resulting conditional log-likelihood is maximised.

When the assumed selection criterion is too severe, as for the set $C_1$, Clayton’s likelihood overcorrects $\theta$, and when assumed selection criteria is to lenient ($C_2$) it undercorrects $\theta$.

Figure A.3 shows an identical plot for the full likelihood case, where Monte-Carlo estimates for the selection probability $\Phi(\theta - 2.7)$ are replaced with an accurate numerical approximation, using the pnorm() function in R. Each assumed complete selection set $C_i$ leads to a different estimate $\hat{\theta}|C_i$. What is required is a pseudo-data-based method that can test the assumption that $C_i = C$.

Let us suppose that $C_i = C$ and $E[\hat{\theta}|C_i] = \theta$ and that our pseudo-data method has been used to obtain $\hat{\theta}$. If further data $y_1^*, \ldots, y_n^*$ were simulated given $\hat{\theta}|C_i$ and subjected
to the same true selection criteria, due to the unbiasedness of $\hat{\theta}|C_1$, this data should on average come from the same distribution as the original real data. One way to assess the similarity of two data sets is to look at their empirical cumulative distribution functions. Figure 4.4 (left) plots the empirical distribution function for 2 data sets generated from a $N(\theta = 3, 1)$ distribution subject to being greater than 2.7 (criterion $C$). We can think of one data set being 'real' - generated from $f(y|\theta)$, and one being pseudo-data, generated from $f(y|\hat{\theta}|C)$.

![Empirical cdf plots](image)

Figure 4.4: **Left**: Empirical cdf plots for the real data set (black) and pseudo-data set (red), obtained by correctly guessing the selection criterion. **Right**: Empirical cdf plots for the real data set (black) and pseudo-data set (blue), obtained by incorrectly guessing the selection criterion.

Figure 4.4 (right) plots the same empirical cdf for the previous real data set, against a cdf of data simulated from a $N(\theta = 2.4, 1)$ distribution, passing the criterion of being greater than 3 (criterion $C_1$). This illustrates the case of pseudo-data generated under $f(y|\hat{\theta}|C_1)$ (see Figure 4.3).
4.3 Using pseudo-data to rank $C$

Clearly the two empirical plots differ the most when $\theta$ and $C$ are misspecified. In this section I describe a formal statistical test to assess the similarity of two data sets, based on their empirical distribution functions and describe how this might be used to decide, from a range of possible selection criterion, which is the most likely.

4.3.1 The Kolmogorov-Smirnov test

The Kolmoogov-Smirnov test is non parametric statistical method that can be used to test the assumption that two samples of data are from the same distribution. The 'distribution free' property is perfect for assessing data that has passed a selection criteria, since the selection process will in general cause any observed data to follow a non-standard distribution. Suppose one wishes to test the assumption that two sets of observations $x_1, \ldots, x_m$ and $y_1, \ldots, y_n$ are from the same continuous distribution $f$. This can be formulated as the null hypothesis

$$H_0 : P(X \leq a) = P(Y \leq a) \quad \forall a$$

(4.3.1)

In order to test $H_0$, versus the two-sided alternative $H_1 : P(X \leq a) \neq P(Y \leq a)$ for at least one $a$, a single list of all $x$ and $y$ observations is made, by ordering them in magnitude from smallest to largest. The $x$ and $y$ observations are then re-labeled as $z_{(1)}, z_{(2)}, \ldots, z_{(N)}$ where $N = n + m$. Now it is necessary to define the indicator variables $I_i, i = 1, \ldots, N$ where $I_i$ is 1 if $z_{(i)}$ is an $x$ observation and 0 if $z_{(i)}$ is a $y$ observation. A further variable $\omega_j$ is created using the formula
4.3. Using pseudo-data to rank $C$

\[
\omega_j = \left[ \frac{j m}{N} - I_1 - \ldots - I_j \right], j = 1, \ldots, N. \tag{4.3.2}
\]

Finally, letting $d$ be the greatest common divisor of $m$ and $n$, the statistic

\[
D = \frac{N}{d} \max \{|\omega_1|, \ldots, |\omega_N|\} \tag{4.3.3}
\]

can be calculated and $H_0$ can be rejected with type one error $\alpha$ if $D \geq D(\alpha, m, n)$. For a full explanation of this test, in all its forms, as well as relevant statistical tables for $D(\alpha, m, n)$ see Hollander and Wolfe (1973). The Kolmogorov-Smirnov test can be implemented using the `ks.test()` function in R.

4.3.2 A procedure for ranking $C$

For the truncated normal example, the selection probabilities could clearly be approximated analytically. In general this will not always be the case, and it might be appropriate to correct for bias via pseudo-data simulation and conditional likelihood maximisation, as in Chapter 3. Let it be assumed that real data $y_1, \ldots, y_n$ has been observed from the complex distribution $f(y|\theta)$ previously described. A general simulation based procedure, to firstly correct for selection bias, and secondly to check the plausibility of the correction could be to:

1. Simulate pseudo-data given $\theta'$ and assumed selection criterion $C_1$ and maximise Clayton’s conditional likelihood to obtain $\hat{\theta}|C_1$

2. Re-simulate pseudo-data given $\hat{\theta}|C_1$ and assumed selection criterion $C_1$ a large num-
4.3. Using pseudo-data to rank $C$

number of times, each time performing a Kolmogorov-Smirnov test for the null hypothesis

$$H_{0,1} : \hat{\theta}|C_1 = \theta|C$$

3. Repeat steps 1 and 2 for a range for assumed criteria $C_1, \ldots, C_k$ and choose $\hat{\theta}|C_r$, $1 \leq r \leq k$, such that $P(H_{0,r} = FALSE) \leq P(H_{0,i} = FALSE)$, for all $i$ in $1, \ldots, k$.

Figure 4.5 shows the results of a simulation study to assess the ability of this procedure to find the correct selection criteria given a range of possible guesses. Real data $y_1, \ldots, y_{100}$ were simulated from a $N(\theta = 3, 1)$ distribution, subject to being greater than 2.7. Thirty possible selection criteria, corresponding to selection thresholds between 1.8 and 3.2 were assumed, and each used, along with a common $\theta'$ of 2, to simulate pseudo-data $y_{ij}, i = 1, \ldots, 100, j = 1, \ldots, 5$. The real and pseudo-data were then augmented and Clayton's conditional likelihood was used to obtain estimates $\hat{\theta}|C_k$, $k = 1, \ldots, 30$. Figure 4.5 (Top) shows the average estimates for $\theta$ for 100 independent simulations, and the average 95\% confidence interval. There is a clear linear trend in the corrected estimates. When the assumed cutoff is too lenient, $\theta$ is overestimated, and when the assumed cutoff is too severe, $\theta$ is underestimated. When the correct cutoff of 2.7 is used $\theta$ is approximately unbiased.

Figure 4.5 (Middle) shows the results of the second pseudo-data stage of this procedure. For each of the 100 independent simulations and 30 different scenarios, pseudo-data were re-simulated given the implied estimate for $\theta$, $\hat{\theta}|C_k$ and the proportion of times that the p-value associated with the D-statistic was less than 0.01 and 0.05 was recorded. As the proposed selection criteria nears the true selection criteria, the proportion of $H_0$ hypothe-
Figure 4.5: Top: Estimates for $\theta$ obtained by assuming range of possible selection criteria. Middle: The proportion of Kolmogorov-Smirnov two-sided tests rejected, for varying selection criteria. Bottom: The distribution of Kolmogorov-Smirnov $D$-statistics, for varying selection criteria.
4.4. Application to the homocysteine data

In the previous chapter I showed how pseudo studies, simulated from an assumed publication criteria, could be used to derive a bias-corrected estimate for the mean homocysteine difference, reported by 33 published genetic association studies, as well as an estimate for the between study variance. Estimating these two quantities required a random effects model, so that real study \( y_i \) in the original meta-analysis was assumed to come from a
N(\theta, s_i^2 + \tau^2) distribution. In keeping with the approach taken in this chapter, I firstly chose a selection model that is simply a function of the effect size \( y \), in an attempt to find the most likely effect size for triggering publication. As before it was assumed that \( y \in C_k \) if \( y \geq T_k \).

Figure 4.6 (top) plots the results of the estimated mean homocysteine difference \( \theta \), when the selection mechanism for \( y \) is greater than the threshold \( T \), between -1 and 2, and additionally \( \pi_2 \) is varied between 0 and 0.08. The results were obtained by simulating \( m=10 \) pseudo studies from a \( N(\theta', s_i^2 + \tau^2) \), for each real study \( i \) and maximising Clayton’s conditional likelihood. The values of the pseudo-data parameters, and the rationale for choosing them, is explained in Chapter 3. The estimate for the bias in \( \theta \) becomes more positive as the effect size assumed to be associated with successful publication increases. But, at the same time, the estimated bias in \( \theta \) goes down as the probability of random publication, \( \pi_2 \), increases.

Figure 4.6 (bottom) shows the average Kolmogorov-Smirnov statistic \( D \) obtained by further pseudo-data simulation, on the assumption that \( \theta = \hat{\theta}|C_k, \tau = \hat{\tau}|C_k \). For all values of \( \pi_2 \) there is good agreement that the effect size \( T \), that minimises the Kolmogorov-Smirnov statistic, is approximately 0.5 \( \mu \text{mol/l} \).

Tables 4.1 and 4.2 show the Kolmogorov-Smirnov \( D \) statistics obtained from assuming the selection models described in Sections 3.6 and 3.7.1. These selection models assumed a global publication criterion for the homocysteine data, based on the p-value, and then the p-value plus effect size respectively. I decided not to implement the Kolmogorov-Smirnov test on the results from the previous chapter which assumed a study-specific criterion, due
4.4. Application to the homocysteine data

Figure 4.6: Top: Publication biased corrected estimates $\hat{\theta}$ for the mean difference in homocysteine, as a function of the publishable effect size $T$ and probability of random publication $\pi_2$. Bottom: Kolmogorov-Smirnov $D$-statistics as a function of $\hat{\theta}|C$ and $\pi_2$. 
to concerns about possible bias. In each case the $D$ statistics shown are the average of 500 independent simulations. The last column contains a p-value, derived via simulation rather than the large sample approximation used by the Kolmogorov-Smirnov test. This was obtained by simulating 999 'real' and 'pseudo' data sets, in the ratio 1:10 as before and given the assumed parameter estimates and selection model, and calculating the rank of the 1 observed $D$ statistic, out of the 999 simulated $D$ statistics.

<table>
<thead>
<tr>
<th>Probability of publication for non-significant p-value ($\pi_2$)</th>
<th>Mean difference in homocysteine (mean (sd))</th>
<th>Between-study standard deviation ($\hat{\tau}$) (mean (sd))</th>
<th>Average K-S D statistic given $\hat{\theta},\hat{\tau}$ (mean (sd))</th>
<th>P-value for $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>1.32(0.24)</td>
<td>1.42(0.10)</td>
<td>0.149(0.015)</td>
<td>0.523</td>
</tr>
<tr>
<td>0.2</td>
<td>1.60(0.17)</td>
<td>1.39(0.22)</td>
<td>0.142(0.016)</td>
<td>0.588</td>
</tr>
<tr>
<td>0.3</td>
<td>1.80(0.17)</td>
<td>1.41(0.24)</td>
<td>0.145(0.017)</td>
<td>0.559</td>
</tr>
<tr>
<td>0.4</td>
<td>1.97(0.14)</td>
<td>1.42(0.10)</td>
<td>0.145(0.015)</td>
<td>0.563</td>
</tr>
<tr>
<td>0.5</td>
<td>2.10(0.14)</td>
<td>1.40(0.17)</td>
<td>0.146(0.016)</td>
<td>0.551</td>
</tr>
</tbody>
</table>

Table 4.1: Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation obtained from the simple publication criterion detailed in Section 3.6, and the Kolmogorov-Smirnov D statistics that they imply.

<table>
<thead>
<tr>
<th>Assumed effect size publishable $T$</th>
<th>Mean difference in homocysteine (mean (sd))</th>
<th>Between-study standard deviation ($\hat{\tau}$) (mean (sd))</th>
<th>Average K-S D statistic given $\hat{\theta},\hat{\tau}$ (mean (sd))</th>
<th>Simulated p-value for $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0</td>
<td>2.08(0.30)</td>
<td>1.73(0.37)</td>
<td>0.131(0.010)</td>
<td>0.685</td>
</tr>
<tr>
<td>0.2</td>
<td>1.98(0.32)</td>
<td>1.82(0.45)</td>
<td>0.129(0.012)</td>
<td>0.697</td>
</tr>
<tr>
<td>0.4</td>
<td>1.86(0.37)</td>
<td>1.80(0.33)</td>
<td>0.132(0.011)</td>
<td>0.671</td>
</tr>
<tr>
<td>0.6</td>
<td>1.58(0.58)</td>
<td>1.91(0.57)</td>
<td>0.133(0.014)</td>
<td>0.660</td>
</tr>
<tr>
<td>0.8</td>
<td>1.53(0.49)</td>
<td>1.42(0.44)</td>
<td>0.170(0.016)</td>
<td>0.370</td>
</tr>
<tr>
<td>1.0</td>
<td>1.37(0.14)</td>
<td>1.40(0.54)</td>
<td>0.179(0.020)</td>
<td>0.306</td>
</tr>
</tbody>
</table>

Table 4.2: Parameter estimates (standard errors) for the mean homocysteine difference and the between-study standard deviation obtained from the complex publication criterion detailed in Section 3.7.1, and the Kolmogorov-Smirnov D statistics that they imply.

When publication is assumed to be a function of a fixed significant p-value of less than 0.05 as well as an unknown probability of publication $\pi_2$, the most likely value of $\pi_2$ given
4.4. Application to the homocysteine data

the available choices, is 0.2. This, along with an estimated mean homocysteine difference of 1.60 and between study standard deviation 1.39, produces the smallest average $D$ statistic, of 0.142. Assuming normality, an approximate 95% confidence interval for this statistic is (0.110, 0.173). The $p$-value for this $D$ statistic, derived by simulation, is 0.588. This indicates that there is no evidence to reject a null hypothesis $H_0$ that the homocysteine data, and pseudo-data simulated given these ‘optimal’ parameter estimates and publication criterion, are from different sources. However, there is clearly no evidence to say that this particular criterion and parameter estimate combination is significantly different from the other 4 options.

When publication is assumed to be a function of a fixed significant $p$-value, a fixed value of $\pi_2$ and an unknown effect size $T$, as in Section 3.7.1, the most likely value of $T$, given the available choices, is 0.2. This, along with an estimated mean homocysteine difference of 1.98 and between study standard deviation 1.82, produces the smallest average $D$ statistic, of 0.129 (95% confidence interval (0.105, 0.153)). Over the two publication classes, this particular criterion is in some sense the most likely, given that it is associated with the smallest value for $D$. But again, there is no power to distinguish it from other criteria either.

In my opinion, the small size and homogeneous nature of the homocysteine data present an insurmountable barrier to learning about the most likely possible publication criterion it was subjected to. The $p$-values calculated for all scenarios indicate that the null hypothesis can not be rejected, but it is clearly not possible for all alternative hypotheses to be true. This is a case of ‘absence of evidence’ rather than ‘evidence of absence’.
4.5 Discussion

When the distribution of a random variable $Y$ is $N(\theta, \sigma^2)$, before an unknown selection process is allowed to distort the distribution of observable realisations $y_1, \ldots, y_n$, Copas and Jackson (2004) have shown that the expected bias in the parameter $\theta$ satisfies the following inequality,

$$bias(\hat{\theta}) \leq \frac{\sigma}{p} \Phi^{-1}(p)$$

(4.5.1)

where $p$ denotes the probability of selection. The only assumption Copas and Jackson make is that the selection probability function does not increase if $\sigma$ increases. Furthermore, when the selection model is the single step function which equals 1 when $y \geq \theta - \sigma \Phi^{-1}(p)$, and 0 otherwise, (4.5.1) becomes an equality. I could have used this formula to calculate the bias in our parameter $\theta$ when it was assumed that $y \in C$ if $y \geq T$, with $\pi_1 = 1, \pi_2 = 0$ (complete selection) since

$$y \geq \theta - \sigma \Phi^{-1}( \frac{\theta - T}{\sigma} )$$

$$y \geq \theta - (\theta - T)$$

$$y \geq T$$

and therefore
bias(\hat{\theta}) = \frac{\sigma}{\Phi\left(\frac{\theta - T}{\sigma}\right)} \phi \left[ \Phi^{-1}\Phi\left(\frac{\theta - T}{\sigma}\right) \right] \\
= \frac{\sigma \phi\left(\frac{\theta - T}{\sigma}\right)}{\Phi\left(\frac{\theta - T}{\sigma}\right)}

Copas and Jackson's result shows that, for a given selection probability \( p \), complete selection models induce the maximum bias. It is not surprising therefore that when I firstly assumed a simple threshold selection function for the homocysteine data in Section 4.4, the bias correction was largest for \( \pi_2 \) equal to 0. Assuming a complete selection model for meta-data will in general be a prudent choice since it will often return a statistic closer to the null position. However, as shown in Chapter 3, the homocysteine example was an exception, because the inverse of the corrected mean effect was then used to derive another quantity of interest.

Another result from Todd et al. (1996) is

\[
\int_{-\infty}^{T} \frac{y^2}{\sigma^2} \phi\left(\frac{y - \theta}{\sigma}\right) dy = (\theta^2 + \sigma^2) \Phi\left(\frac{T - \theta}{\sigma}\right) - \sigma(T + \theta) \phi\left(\frac{T - \theta}{\sigma}\right)
\]

which allows us to calculate \( E[Y^2|Y > T] \) to be

\[
\int_{-\infty}^{\infty} \frac{y^2}{\sigma^2} \phi\left(\frac{y - \theta}{\sigma}\right) dy = \frac{\theta^2 + \sigma^2 - \int_{-\infty}^{T} \frac{y^2}{\sigma^2} \phi\left(\frac{y - \theta}{\sigma}\right) dy}{\Phi\left(\frac{\theta - T}{\sigma}\right)} = \frac{\theta^2 + \sigma^2 - (\theta^2 + \sigma^2)(1 - \Phi\left(\frac{\theta - T}{\sigma}\right)) - \sigma(T + \theta) \phi\left(\frac{\theta - T}{\sigma}\right)}{\Phi\left(\frac{\theta - T}{\sigma}\right)} = \theta^2 + \sigma^2 + \sigma(T + \theta) \phi\left(\frac{\theta - T}{\sigma}\right) \Phi\left(\frac{\theta - T}{\sigma}\right)
\]
expression (4.2.3) and (4.5.3) can then be combined to derive an expression for the variance of $Y | Y > T$

$$\text{Var}[Y | Y > T] = E[Y^2 | Y > T] - E[Y | Y > T]^2$$

$$= \theta^2 + \sigma^2 + \sigma(T + \theta) \frac{\phi \left( \frac{\theta - T}{\sigma} \right)}{\Phi \left( \frac{\theta - T}{\sigma} \right)} - E[Y | Y > T]^2$$

$$= \sigma^2 + E[Y | Y > T](T - E[Y | Y > T]) + \theta(E[Y | Y > T] - T)$$

Rearranging for $\theta$, and substituting theoretical quantities for statistics derived from the observed data $y$ gives

$$\theta = \frac{\text{Var}[Y | Y > T] - \sigma^2 - E[Y | Y > T](T - E[Y | Y > T])}{E[Y | Y > T] - T}$$

$$\approx \frac{\text{Var}(y) - \sigma^2 - \bar{y}(T - \bar{y})}{\bar{y} - T} \quad (4.5.4)$$

So, for large samples of $y$, pseudo-data, and even maximum likelihood approaches are not strictly necessary to derive a bias corrected estimate for $\theta$, this simple formula should suffice. The results of a small simulation study to investigate the accuracy of equation (4.5.4) are detailed in Figure A.4.

When meta-data is thought to have been subject to publication bias, and the selection model is thought to be complex, importance sampling approaches that provide Monte-Carlo approximations to complex integrals, become an attractive option with which to achieve a correction. The adapted Kolmogorov-Smirnov method to test the validity of
any assumed publication model could complement this approach. It is commonly used in medical statistics to assess differences in survival rates between two populations interest, (Fleming et al., 1980), and also in meteorology to monitor changes in weather patterns over time, (Freedman, 1979). I do not believe it has been applied in this context before. I could have attempted to use the Chi-Square test in its place to assess the similarity of the real and pseudo-data. However in order to implement this test I would have had to split the data into a number of discrete regions to generate the observed and expected frequency counts. This would have added a further level of complication and subjectivity, because the value of the Chi-Square statistic would obviously vary depending on the number and size of regions chosen.

The fundamental assumption that must be satisfied to make the Kolmogorov-Smirnov test applicable in this setting, is that $E[\hat{\theta}|C] = \theta$. In other words, the maximum likelihood estimate for $\theta$ given correct model specification, that is approximated via importance sampling methods, should be unbiased. In our truncated normal example this was theoretically true, and it was demonstrated that when the sample size was sufficiently large, then this property was preserved. However, Stallard et al. (2007) show that when the sample size is small, the maximum likelihood estimate for $\theta$ can be infinitely biased. Although their proof is somewhat longer, this can be understood at a simple heuristic level by imagining how the estimate obtained for $\theta$ in (4.5.4) alters as $\bar{y}$ tends to the threshold $T$, as it might do by chance if $\bar{y}$ is based on a handful of observations.

In other settings, the MLE for a parameter of interest may be biased even if it is based on a large sample size, for example if the total number of parameters in a statistical model
is also large. When $E[\hat{\theta}|C] \neq \theta$, the Kolmogorov-Smirnov D statistic can not be expected to be minimised for a $T_k$ that is the true threshold for membership of $C$. Furthermore, the size of meta-data sets as well as their heterogeneous nature will mean there is little or no power to learn about the publication model that was in operation, or distinguish one proposal from another. This was certainly the case for the homocysteine example.

The Kolmogorov-Smirnov test can only be applied to continuous data, and I made this assumption when assuming that the reported mean homocysteine differences followed a normal distribution. However, due to the fact that the studies only reported this difference to one decimal place, several of these values were identical. To overcome this problem, and enable $R$ to calculate a Kolmogorov-Smirnov statistic, the value of a handful of studies reported mean differences were altered, up or down, by 0.001. Whilst this enabled the software to work superficially, the small size of the data set, as well as its pseudo-continuous nature made me wary of trusting the p-values attributed by `ks.test()` to the $D$ statistics. That is why I preferred to derive my own p-values via simulation. In situations where the meta-data is non-continuous, for example frequency counts, then the Chi-Square test would still be a valid option whereas the Kolmogorov-Smirnov test would not.
Chapter 5

Ascertainment bias in genetic epidemiology

Ascertainment bias in epidemiological studies is an unavoidable consequence of sampling strategies that enable the aetiology of rare diseases to be explored. Historically, the modest aim of such studies was often to investigate the prevalence of a disease amongst the offspring of 'at-risk' parental unions. This was known as the segregation-ratio, (Sham, 1998), and its estimation provided important, though crude, information about the mechanism of genetic inheritance that may be operating. In recent years methodological advances have enabled scientists to fit complex statistical models to account for the influence of genetic and environmental factors on raising, or lowering, an individual's disease risk. However, whilst these models offer a more realistic approximation to biological reality, they have also made the process of correctly accounting for the effects of an ascertainment process much more challenging. After considering whether Clayton's conditional likelihood approach could be used to correct for ascertainment bias in this setting, the partial solution of Burton et al. (2001) is introduced, that is currently the topic of considerable debate.
5.1 Introduction

The standard way to estimate parameters that reflect the distribution of disease in a population is to collect data on a random sample of people from that population in an epidemiological study. If the disease risk is thought to be influenced by measurable covariates, for example by age and gender, then the population prevalence and the effect of the covariates can be simultaneously modeled using a generalized linear model (GLM) (McCullagh and Nelder, 1989), such as logistic regression. Let $y_i$ be the binary, $(0,1)$, disease status of individual $i$ recruited into a study. The underlying disease risk for individual $i$ is denoted by $\mu_i$, so that $y_i \sim \text{bern}(\mu_i)$ and

$$\log \left\{ \frac{\mu_i}{1 - \mu_i} \right\} = \eta_i = \alpha + \beta x_i$$

(5.1.1)

where $x$ indicates an observable covariate with a corresponding coefficient $\beta$. Often the pattern of disease prevalence shows more variation than would be permitted by the mean-variance relationship of a GLM for binomial data. Typically such 'over-dispersion' (Dey et al., 1997), is due to unobserved covariates and one way to accommodate the extra variability is by including normally distributed random effects within the linear predictor; so producing a generalized linear mixed model (GLMM) (Schall, 1991; Breslow and Clayton, 1993). When this over-dispersion can be linked to a tendency for disease cases to cluster within families, indicating that shared genetic or environmental random effects could be responsible, a genetic epidemiological study becomes an attractive option. Now, sets of related rather than unrelated individuals must be recruited.

Let $y_{ij}$ be the binary, $(0,1)$, disease status of the $j^{th}$ member of the $i^{th}$ family recruited...
onto a study and let $y_i$ denote the vector of responses for the full family. The underlying probability that an individual has the disease is denoted by $\mu_{ij}$, so that $y_{ij} \sim \text{bern}(\mu_{ij})$ and

$$\log \left\{ \frac{\mu_{ij}}{1 - \mu_{ij}} \right\} = \eta_{ij} = \alpha + \beta x_{ij} + \xi_{ij}$$  (5.1.2)

Here $x_{ij}$ is an observed covariate and the random effect $\xi_{ij}$ represents the impact of unobserved covariates. Across families it is assumed that the random effects have a zero mean and a covariance matrix $\Sigma_i$ that depends on a vector of parameters $\theta$. The covariance structure of the random effects within each family can be selected to reflect our beliefs about how the genetic and environmental factors are shared, as in Burton et al. (1999). If it can be shown in a variance component analysis that the genetic component of variation in disease risk is sufficiently large, then a strong case is made for further, more sophisticated techniques such as linkage to be employed, with the hope of identifying the gene (or most probably genes) responsible. For a thorough description of how variance component analysis complements genetics research in general see Burton et al. (2005, 2007).

Under this model the likelihood for a random sample of families from the general population is obtained by integrating over the random effects to give

$$L(\alpha, \beta, \theta|y) \propto \prod_{i=1}^{n} \int_{R_i} \left\{ \prod_{j=1}^{m_i} \frac{\exp y_{ij}(\alpha + \beta x_{ij} + \xi_{ij})}{1 + \exp (\alpha + \beta x_{ij} + \xi_{ij})} \right\} \phi(\xi_i|0, \Sigma_i(\theta)) d\xi_i$$  (5.1.3)

Here $\phi$ represents a multivariate normal density and $R_i$ represents the full range of the random effect for the $i^{th}$ family. Estimates of the parameters may be obtained by maximising this likelihood over $\alpha, \beta, \theta$. 

5.1. Introduction
5.2 A model for ascertained family disease data

The GLMM framework would provide a good model for randomly selected individuals or perhaps families, but if the disease of interest is rare, recruiting at random will yield very few sufferers, and hence little useful information. For example, if the disease of interest affects 1 family in a thousand, on average 10,000 families would have to be recruited in order to observe just 10 disease cases. Not only would a data set of this sort be extremely sparse, providing little or no power for parameter estimation, it would be extremely costly to create. To circumvent these problems one can choose to oversample ‘information-rich’ families. For example, one could select all families in the population that have at least one member affected by the disease of interest. This is called complete ascertainment, (Elston and Sobel, 1979; Ewens and Shute, 1986; Kraft and Thomas, 2000). If the probability of ascertainment is proportional to the number of affected individuals in a family, rather than being a constant value, then the sampling strategy is referred to as ‘single’ ascertainment, (Cordell and Olson, 2000).

Suppose that family \( i \) is only recruited if their responses, \( y_i \), satisfy the condition \( y_i \in C_i \).

For instance \( C_i \) might be the set of responses such that \( \sum_{j=1}^{m_i} y_{ij} \geq 1 \). So that each selected family must have at least one affected member. Under this selection criterion the probability that family \( i \) is recruited into the study is

\[
p(y_i \in C_i | \alpha, \beta, \theta) \propto \sum_{y_i \in C_i} \int_{R_i} \left\{ \prod_{j=1}^{m_i} \frac{\exp y_{ij}(\alpha + \beta x_{ij} + \xi_{ij})}{1 + \exp(\alpha \beta + x_{ij} + \xi_{ij})} \right\} \phi(\xi_i | 0, \Sigma_i(\theta)) d\xi_i \quad (5.2.1)
\]

The correctly adjusted likelihood, allowing for ascertainment becomes
where I change the symbol in the summation for the ascertainment probability to \( u \) in order to distinguish it from the observed data \( y_{ij} \). If \( y \) was a continuous variable this sum would naturally be replaced by an integral. Evaluation of this likelihood is complicated by the fact that it involves several integrals, that are possibly multidimensional, and unless they are tractable the calculation is likely to be prohibitively slow via numerical methods, a view supported by Noh et al. (2005).

5.2.1 Is Clayton's method appropriate?

In Chapter 2 the relative merits of Monte-Carlo integration over that of numerical integration for complex problems were considered. Clayton's conditional likelihood, a method that utilised a special type of Monte-Carlo integration, was introduced, and in Chapter 3 it was used to correct for publication bias in meta-analysis. Can his idea be applied to make direct inferences about parameters when the data takes the form of binary, correlated observations sampled under an ascertainment mechanism, as modeled by likelihood (5.2.2)?

Equation (2.3.9) showed that Clayton's algorithm conditions out the probability of ascertainment, denoted by \( z(\theta) \), on the denominator of each term \( h(y|\theta) \). This leaves a conditional likelihood whose \( i \)th term requires the evaluation of \( (m + 1) \) real and pseudo-data points only under the density \( h \). For Clayton's method to be appropriate it must therefore be easier to evaluate \( h \) \( (m + 1) \) times, than to evaluate \( h \) once and \( z \) once.
5.2. A model for ascertained family disease data

In (5.2.2)

- \( h(y|\theta) \) is analogous to \( \int_{R_i} \left\{ \prod_{j=1}^{m_i} \frac{\exp[y_{ij}(\alpha + \beta z_{ij} + \xi_{ij})]}{1 + \exp(\alpha + \beta z_{ij} + \xi_{ij})} \right\} \phi(\xi_i|0, \Sigma_i(\theta)) d\xi_i \)

- \( z(\theta) \) is analogous to \( \sum_{y_i \in C_i} \int_{R_i} \left\{ \prod_{j=1}^{m_i} \frac{\exp[y_{ij}(\alpha + \beta z_{ij} + \xi_{ij})]}{1 + \exp(\alpha + \beta z_{ij} + \xi_{ij})} \right\} \phi(\xi_i|0, \Sigma_i(\theta)) d\xi_i \)

Neither \( h \) or \( z \) can begin to be calculated until the distribution of random effects in the model has been integrated out. Clayton’s method saves us from performing the integrations for \( z \), but it requires \( m \) extra integrations over \( h \). Because the dimension of this integral is linked to the number of members in the family, data augmentation will add considerably more computation than it will take away.

In general, the appropriateness of Clayton’s method will depend on the structure of the data as well as the ‘complexity’ of the model. Imagine that \( y_1, ..., y_n \) are i.i.d observations \( \sim N(\theta, \sigma^2) = h(y|\theta, \sigma^2) \) subject to a selection process. Because the mean and variance are orthogonal quantities that are parameterised separately, \( h(y|\theta, \sigma^2) \) can be instantly evaluated. This is an example of normal data from a ‘simple’ model, as in the distribution of homocysteine levels in Chapter 3. Imagine now that \( y_1, ..., y_m \) are a vector of correlated observations \( \sim MVN_m(\theta, \Sigma) = h(y|\theta, \Sigma) \). Although the model is more complex, Clayton’s method would still be appropriate as \( h \) is normally distributed and easy to evaluate. This scenario could be thought of as normal data from a ‘complex’ model. Imagine that \( y_1, ..., y_n \) follow a \( bin(n, \theta) = h(y|n, \theta) \) but have been subject to a selection process. Because there is a single \( \theta \) for all observations, \( h(y|n, \theta) \) can be evaluated easily without prior integration over any random effects and so Clayton’s method can still be used. This could be thought of as non-normal data, but from a simple model. Unfortunately, likeli-
5.3. The effect of complete ascertainment for complex models

The effect of complete ascertainment for complex models does not fit into any of these categories, because it models non-normal data, coming from a highly complex model. θ's mean value is replaced by multiple fixed effects α and β, and its variance is modeled by Σθ.

5.3 The effect of complete ascertainment for complex models

Under this strategy the ascertained population being sampled from differs from the general population in two distinct ways. Most obviously, if no account is taken for the fact that all 'at risk' families with no affected members are missing from the sample, a naive estimate of the prevalence of disease in the ascertained population will be positively biased; potentially seriously so. This is referred to by Burton et al. (2001) as 'classical' ascertainment bias and has a long history, (Fisher, 1934; Haldane, 1938). If random effects are present, the ascertainment process imposes a much more subtle change. Because the probability of ascertaining a family is related to the underlying disease risks, as reflected in the magnitude of the random effects, the pattern of random effects in the ascertained population and the general population will differ systematically. This phenomenon is unrelated to classical ascertainment bias and is certainly harder to intuitively understand, because the disease risks modeled via the linear predictor η are never observed. To demonstrate both the separate and combined effects of these two processes I consider a simple example.

Consider a simple GLM model based on logistic regression for family i, with shared linear predictor \( \eta_i = \alpha \), where \( \alpha = -5 \). This illustrates the case where there is no heterogeneity of disease risk between families in the general population. Every individual has a disease
risk $\mu$ of approximately 0.7%, which is identical in this case to the expected population prevalence. One thousand families, each made up of $m_t = m = 5$ siblings, are simulated under complete ascertainment, so that they have at least 1 affected member. Of this, 986 have 1 affected member, 14 have 2 affected members and there are no families with 3, 4 or 5 affected members. A naive estimate of the population prevalence of disease, ignoring the ascertainment criterion, could be obtained by dividing the total number of affected individuals by the total number of observed individuals

$$\frac{\exp(\alpha)}{1 + \exp(\alpha)} = \mu = \frac{\sum_{k=1}^{5} k a_k}{5 \sum_{k=1}^{5} a_k}$$

(5.3.1)

where $a_k$ symbolises the number of families with $k$ affected members. Crucially, $a_0$ is missing. Using our simulated data the population prevalence $\mu$ was grossly overestimated to be 30%, implying and $\alpha$ of $\log\left(\frac{0.93}{1-0.93}\right) = -0.84$. In order to take into account the effect of ascertainment it must be acknowledged that the distribution of the number of affected siblings in the ascertained population follows a truncated binomial distribution, (Finney, 1949; Wilkinson, 1961; Yao and Tai, 2000), so that the probability of observing $k$ affected individuals, given $\mu$ is

$$P_5(k, \mu) = \binom{5}{k} \mu^k (1-\mu)^{(5-k)} \frac{\mu}{1 - (1-\mu)^5}$$

Thomas and Gart (1971) show that when the family size is fixed the maximum likelihood estimator $\hat{\mu}$ may be found by solving the equation.
5.3. The effect of complete ascertainment for complex models

\[ \sum_{k=1}^{5} i a_k = \frac{5\hat{\mu} \sum_{k=1}^{5} a_k}{1 - (1 - \hat{\mu})^5} \]  

But when the family size is greater than 3, as in our example, (5.3.2) has no closed form and an iterative procedure is needed to find \( \hat{\mu} \). Li and Mantel (1968) proposed an elegant solution to deriving an estimate for \( \mu \), no matter what the family size, via the formula

\[ \frac{\exp(\alpha)}{1 + \exp(\alpha)} = \mu \approx \frac{(\sum_{k=1}^{5} k a_k) - a_1}{5(\sum_{k=1}^{5} a_k) - a_1} \]  

See Appendix A.3 for a detailed explanation. Using our simulated data, equation (5.3.3) can be used to estimate \( \hat{\mu} \).

\( \hat{\mu} = \frac{(1(986) + 2(14) + 3(0) + 4(0) + 5(0)) - 986}{5(1000) - 986} = 0.00697 \)

Meaning \( \hat{\alpha} = \log\left(\frac{6.97 \times 10^{-3}}{1 - 6.97 \times 10^{-3}}\right) = -4.96 \). The Li-Mantel estimator works because when there is no heterogeneity in the disease risk (no random effect), the underlying risk is the same in the general and the ascertained population.

Now consider a GLMM based on logistic regression, identical to the previous example, but with the addition of a single random effect that is common to all members of the same family. So \( \xi_{ij} = \xi_i \sim N(0, \theta) \) for all \( j \), and \( \theta \) is equal to 4.5. Figure 5.1 compares the distributions of linear predictors \( \eta_i = \alpha + \xi_i \) for 1000 sibships, from the general population with 1000 linear predictors sampled under the ascertainment criterion that selected families must have at least one affected member. The mean of these ascertained linear...
The variance of the ascertained linear predictors was also shrunk, from 4.47 in the general population, to 2.44 in the ascertained population.

![Histograms of linear predictors, \( \eta = \alpha + \xi \), for the 'general' and 'ascertained' population, under a simple GLMM model.](image)

Figure 5.1: Distributions of linear predictors, \( \eta = \alpha + \xi \), for the 'general' and 'ascertained' population, under a simple GLMM model.

The linear predictors from the ascertained population went on to produce; 692 families with 1 affected member, 185 with 2 affected members, 66 with 3 affected members, 41 with 4 affected members and 16 families with 5 affected members respectively. Although the mean disease risk under the GLMM model is the same as the overall disease risk in the GLM case, the heterogeneity in disease risk causes a totally different pattern of affected families in the ascertained population. Employing the Li-Mantel estimator to correct for classical ascertainment bias with the GLMM data yields an estimate for \( \alpha \) of \( \log(\frac{0.188}{1-0.188}) = -1.46 \). Because the underlying disease risk is different in the ascertained
population, the assumption of a simple truncated binomial distribution for the disease data is no longer valid, and consequently the Li-Mantel estimator does not work. A proper simultaneous correction for 'classical' bias, as well as accounting for the effects of heterogeneity in disease risk, requires the evaluation of likelihood (5.2.2), and has been the subject of recent debate, (Burton et al., 2001; Glidden and Liang, 2002; Epstein et al., 2002; Burton, 2003; Noh et al., 2005).

5.4 A partial solution for GLMMs?

Burton et al. (2001) investigated the effect of ascertainment bias on disease data \( y_{ij} \sim \text{bern}(\mu_{ij}) \) from the logistic model

\[
\log \left\{ \frac{\mu_{ij}}{1 - \mu_{ij}} \right\} = \eta_{ij} = \alpha + \beta b_{ij} + \xi_i \tag{5.4.1}
\]

where \( b_{ij} \) was a zero-centered binary covariate, \( \xi_i \sim N(0, \theta) \). In their example each family was made up of 3 siblings (no parents) and were sampled subject to the criterion that they contained at least 1 affected member. Estimation of the general population parameters \((\alpha, \beta, \theta)\) would have involved a complex likelihood analogous to (5.2.2). Furthermore, since there was covariate information on each individual, these integrations would have to be evaluated separately for each family. Consequently full evaluation was prohibitively slow whether the likelihood is maximised or it is combined with priors in a Bayesian analysis.

Burton et al. (2001) proposed a novel method for the analysis of the disease data generated from model (5.4.1) that did not ignore the ascertainment process, but was nevertheless computationally easier than the correct maximisation of likelihood (5.2.2). They proposed
5.4. A partial solution for GLMMs?

an adapted Bayesian routine for estimation, based on Markov chain Monte Carlo (MCMC) methods implemented in WinBUGS, (Spiegelhalter et al., 2005). At simulation \( r \) in the MCMC loop, Burton conditioned likelihood (5.2.2) on the current value of the random effects \( \xi^r \) so that the posterior density

\[
p(\alpha^r, \beta^r, \theta^r | y) \propto \prod_{i=1}^{n} \frac{\left\{ \prod_{j=1}^{n_i} \frac{\exp(y_{ij}(\alpha^r + \beta^r b_{ij} + \xi^r_j))}{1+\exp(\alpha^r + \beta^r b_{ij} + \xi^r_j)} \right\}}{\prod_{j=1}^{n_i} \frac{1}{1+\exp(\alpha^r + \beta^r b_{ij} + \xi^r_j)}} g(\alpha^r, \beta^r, \theta^r) \tag{5.4.2}
\]

could be immediately evaluated. \( g \) here represents the prior distribution of the general population parameters. Over many iterations, the MCMC algorithm integrates this function over the random effects, which for non-informative priors, is virtually the same as integrating a likelihood with this form. Burton et al. (2001) acknowledged that (5.4.2) does not solve the ascertainment problem, in the sense that it can return parameter estimates that pertain to the general population. In order to achieve a complete correction it is necessary to marginalise - integrate over the distribution of random effects, and then condition the standard likelihood by the probability of ascertainment. Burton et al. (2001) effectively reverse this order, by conditioning on the probability of ascertainment before marginalising, a fact noted by Glidden and Liang (2002); Epstein et al. (2002); Noh et al. (2005). Burton et al. (2001) offers a convincing argument, backed up with empirical evidence, as to how parameter estimates obtained from his procedure pertain to the ascertained population. In a simulation study Burton et al. (2001) simulated 1000 sibships with \( \alpha \) set to -5, \( \theta \) set to 4.5 and the covariate \( \beta \) set to be random draws from a zero-centered Bernoulli distribution, making it very similar to the previous example. The mean and variance of the linear predictors in the ascertained sample were estimated directly to be -2.23 and 2.42 respectively. By specifying uninformative priors Burton used posterior likelihood (5.4.2)
to estimate the mean and variance to be -2.39 and 2.45 respectively.

Burton et. al’s method appears to possess the ability to estimate parameters in a GLMM model that pertain to the ascertained population. This was not accepted by Glidden and Liang (2002) or Epstein et al. (2002), but has more recently been supported by Nie (2003). However, what is not in dispute is the fact that the ascertained parameters are of strictly limited utility. So why bother estimating them at all? In Chapter 6 I take up this point and argue that, even though they possess little value in their own right, they have the potential to form the first stage of a two-stage solution to general population parameter estimation.
Chapter 6

A two-stage approach to correcting for ascertainment bias in complex models

Following on from Chapter 5, it is hopefully clear that correction for ascertainment bias is a vital part of the analysis of genetic epidemiology studies, whenever subjects are recruited under a non-random sampling strategy. It was shown that, if the model includes both fixed and random effects, correct adjustment for ascertainment often requires extensive integration, which can be computationally infeasible. In this chapter I propose a two-stage method for ascertainment bias correction. In the first stage parameters are estimated that pertain to the ascertained population, via a truncated likelihood taking its inspiration from Burton et al. (2001). In the second stage these are converted into general population parameter estimates using Clayton’s conditional likelihood. The method is illustrated with simulations based on a simple model and it is then described how the method can be used with complex models. The two-stage approach avoids some of the integration required in
6.1 Parameter estimation for the ascertained population

Returning to the simple GLMM example in Chapter 5, in which disease data for a population of families of size 5 are simulated using a linear predictor of the form $\eta_i = \alpha + \xi_i$, I denote the parameters in the general population to be $\alpha, \theta$ and now denote the parameters of the ascertained population, that is all families with at least one affected member, to be $\alpha^*, \theta^*$. They reflect the mean and variance of the distribution of ascertained linear predictors, shown in Figure 5.1. If the linear predictors of an ascertained sample, denoted by $\eta^*_i = \alpha^* + \xi^*_i$, were directly accessible, it would be a simple matter to obtain their empirical mean and variance in order to estimate $\alpha^*$ and $\theta^*$. Instead of course it is the disease outcomes $y$ that are observed.

Under a model in which each member of a family has the same probability of disease, in the general population the number of affected members in families of a given size will follow a binomial distribution. As shown in Chapter 5, the impact of sampling under an ascertainment criterion is a complete absence of 'at risk' families, which by chance contain no affected members. The distribution of the number of affected members per family in the ascertained population will be a zero-truncated binomial. For families of size 5 the likelihood would be
6.2. Moving from $\alpha^*, \theta^*$ to $\alpha, \theta$

$$L(\alpha^*, \theta^* | y) \propto \prod_{i=1}^{n} \frac{\prod_{j=1}^{m_i} \exp \left( \frac{\exp(y_{ij} (\alpha^* + \xi^*_i))}{1 + \exp(\alpha^* + \xi^*_i)} \right)}{1 + \exp(\frac{1}{1 + \exp(\alpha^* + \xi^*_i)})} \phi(\xi^*_i | 0, \theta^*) d\xi^*_i$$  \hspace{1cm} (6.1.1)

Under models with covariates in the linear predictor, or a complex correlation structure between the random effects, the distribution of the number of affected members of a family will no longer be binomial. None the less, the equivalent probabilities can easily be evaluated and the likelihood, equivalent to (6.1.1) still only requires one integration. Burton's posterior likelihood (5.4.2) appears to be a generalised Bayesian analogy to (6.1.1). Maximisation leads directly to parameter estimates that pertain to the ascertained population. It is clearly computationally easier to get estimates for the ascertained population than it is for the general population. However, as discussed in Chapter 5, $\alpha^*, \theta^*$ are of limited interest. In the next section a procedure is described that enables $\alpha^*, \theta^*$ to be used to form a bridge to the general population parameters $\alpha, \theta$.

### 6.2 General population parameter estimation via $\alpha^*, \theta^*$

Let $h()$ denote a general normal density and $f()$ a complex and unknown density, obtained under complete ascertainment. Burton et al. (2001), Glidden and Liang (2002) observed, and Figure 5.1 illustrates, that when the random effects are normally distributed in the general population they are also approximately normally distributed in the ascertained population. So, with respect to the linear predictor $\eta$,

$$f(\eta | \alpha, \theta) = \frac{h(\eta | \alpha, \theta)}{z(\alpha, \theta)}$$

$$\approx h(\eta | \alpha^*, \theta^*)$$  \hspace{1cm} (6.2.1)
6.2. Moving from $\alpha^*, \theta^*$ to $\alpha, \theta$

where $z(\alpha, \theta)$ represents an unknown ascertainment probability. The distribution of linear predictors generated from $\alpha, \theta$ and then subjected to the ascertainment criterion $C$ is approximately the same as the distribution of linear predictors with parameters $\alpha^*, \theta^*$. To generate a sample of ascertained linear predictors one could either return to the general population, generate linear predictors given $\alpha, \theta$ and then subject them to the selection criterion, or one could choose to generate them directly from $\alpha^*, \theta^*$. This implies the following scheme:

- Re-simulate a sample of linear predictors $\eta^*$ from $h(\eta^*|\alpha^*, \theta^*)$
- Assume that the $\eta^*$ have been drawn from the complex distribution $f(\eta^*|\alpha, \theta) = \frac{h(\eta^*|\alpha, \theta)}{z(\alpha, \theta)}$.
- Maximise $L(\alpha, \theta|\eta^*) = \prod h(\eta^*|\alpha, \theta) \cdot \frac{1}{z(\alpha, \theta)}$ to obtain estimates for $\alpha, \theta$.

Since $z(\alpha, \theta)$ is unknown $L(\alpha, \theta|\eta^*)$ can not be maximised directly. However, the density $h()$ and the ascertainment criterion $C$ are known so it is possible to use Clayton’s method for the maximization. In this current context, the parameter estimates for the ascertained population can be used to simulate linear predictors $\eta_i^*$ for family $i$, which can be augmented with pseudo linear predictors $\eta_{ik}^*, k = 1, \ldots, D$, generated from guesses at the general population parameters and then subjected to the selection criterion. In this case Clayton’s conditional likelihood becomes

$$L_{\text{cond}}(\alpha, \theta|\alpha^*, \theta^*, \eta^*, \eta^*) = \prod_{i=1}^{n} \left\{ \frac{h(\eta_i^*|\alpha, \theta)}{h(\eta_i^*|\alpha^*, \theta^*)} \cdot \frac{h(\eta_{ik}^*|\alpha, \theta)}{h(\eta_{ik}^*|\alpha^*, \theta^*)} \right\}$$

(6.2.2)

It is important to remember that (6.2.2) is made up solely of simulated data. Once the real data has been used to estimate $\alpha^*, \theta^*$ it is discarded and $\alpha^*, \theta^*$ are then used
6.2. Moving from $\alpha^*, \theta^*$ to $\alpha, \theta$

to simulate $\eta^*$. Maximisation of this conditional likelihood produces estimates for $\alpha, \theta$. As discussed in Chapter 3, increasing $D$ will reduce the Monte-Carlo error component of the parameters' variances. However there is obviously a trade-off with computational efficiency. The optimal choice of $D$ will vary depending on the model, computational facilities and accuracy requirements.

Stage 2

Figure 6.1: A schematic representation of the two-stage approach. If it is infeasible to take the 'direct' route to the general population parameters $\alpha, \theta$, it may be possible to estimate the ascertained parameters $\alpha^*, \theta^*$ (stage 1) and use these estimates along with pseudo-data to estimate $\alpha, \theta$ (stage 2).

Figure 6.1 is a graphical illustration of our two-stage procedure. Given observed disease data $y$, maximisation of the 'correct' likelihood, analogous to (5.2.2), would lead directly to estimates for the general population parameters $\alpha, \theta$. However, if this is not computationally feasible, maximisation of likelihood (6.1.1) in stage 1 leads to ascertained population parameter estimates $\alpha^*, \theta^*$. Stage 2 can then be implemented by using these parameters to simulate linear predictors $\eta^*$, combining them with pseudo linear predictors $\eta'$ and constructing a conditional likelihood to estimate $\alpha, \theta$.

6.2.1 Starting values for $\alpha', \theta'$

It is clear from Figure 5.1 that the ascertainment process forces $\alpha^*$ to be greater than $\alpha$ and $\theta^*$ to be smaller than $\theta$, so the ascertainment parameter estimates from stage 1 can be
informative for deciding appropriate pseudo-data starting values in stage 2. However, as illustrated in Chapter 3, poor starting values do not invalidate this method. Once estimates for the general population parameters $\alpha, \theta$ have been obtained, then stage 2 can be repeated by simulating pseudo-data with $\alpha' = \hat{\alpha}, \theta' = \hat{\theta}$. Stage 2 should be repeated in this manner until the general population parameter estimates have satisfactorily converged.

6.3 Simulation studies

6.3.1 Simulation of the second stage

One thousand data sets, each of 2000 families were simulated under the scenario described in Section 6.1. The chosen parameters for the general population were $\alpha = -5, \theta = 4.5$ and all families with at least one affected member were retained. As the data were simulated, the disease outcomes, $y$, as well as their corresponding linear predictors were available. For each data set it was therefore possible to estimate $\alpha^*, \theta^*$ directly by calculating their mean and variance. From now on I refer to ascertained parameter estimates derived directly from the linear predictors to be $\tilde{\alpha}^*, \tilde{\theta}^*$. A preliminary analysis was undertaken aimed solely at demonstrating that Clayton's method can be used in stage 2 to translate good estimates at the level of the ascertained sample to estimates in the general population.

For the purposes of this analysis, 500 families linear predictors were re-simulated from a $N(\tilde{\alpha}^*, \tilde{\theta}^*)$ distribution, which represented the 'real' data. These re-simulated linear predictors were augmented with $D = 5$ sets of 500 pseudo linear predictors based on a guess at the general population parameters, $\alpha' = -3.5, \theta' = 5.85$, chosen to be $0.7\alpha, 1.3\theta$. That is family data were drawn from the hypothesised general population $(N(\alpha', \theta'))$, and then
subjected to the ascertainment rule. I used 500 re-simulated families' linear predictors, a quarter of the original number, because it was sufficient for maximisation purposes and increased the computational speed of our simulations considerably. Using the 500 re-simulated linear predictors and the matched sets of pseudo linear predictors, conditional likelihood (6.2.2) was maximised to obtain estimates of the general population parameters. This maximisation was carried out using the optimisation function `optim()`, (R Development Core Team, 2004).

![Graphs showing simulation results](image-url)

**Figure 6.2:** Top: the mean and average 95% confidence interval for $\alpha, \theta$ obtained from pseudo-data cycles 1 ($\alpha' = -3.5, \theta' = 5.85$), 2 ($\alpha' = -5.15, \theta' = 4.72$) and 3 ($\alpha' = -5.00, \theta' = 4.51$). Bottom: the distribution of 1000 estimates for $\alpha, \theta$, obtained from the second cycle of pseudo-data generation.

Since the performance of Clayton's method is affected by the quality of guesses $\alpha', \theta'$, for each data set I augmented the re-simulated linear predictors with 3 'cycles' of pseudo-data, where cycles 2 and 3 benefited from updating $\alpha', \theta'$ to be the current estimate for $\alpha, \theta$ derived from maximisation of (6.2.2). The average estimates for $\alpha, \theta$ over the 1000

---

6.3. Simulation studies

---
6.3. Simulation studies

data sets after 1 cycle of pseudo-data were; -5.15 and 4.72, after 2 cycles; -5.00 and 4.51; and after 3 cycles -5.01 and 4.52.

The histograms in Figure 6.2 (Bottom) show the distribution of the estimates for $\alpha, \theta$ after two cycles of conditional likelihood maximisation. Figure 6.2 (Top) show the mean and average 95% confidence intervals for $\hat{\alpha}, \hat{\theta}$ for cycles 1, 2 and 3. Preliminary simulations clearly demonstrate that if one can obtain good estimates of the ascertained parameters, such as $\tilde{\alpha}^*, \tilde{\theta}^*$, then it is possible to transform them into good estimates for the general population parameters $\alpha, \theta$. In this example two cycles are sufficient to give very accurate estimates, but in general the cycles should be repeated until the estimates are unchanged.

6.3.2 Combined simulation of both stages

In reality, the linear predictors, and hence $\tilde{\alpha}^*, \tilde{\theta}^*$, are unobservable and so $\alpha^*, \theta^*$ need to be estimated from the disease data $y$, which can be done by maximising likelihood (6.1.1). This likelihood involves integrating over the random effects, $\xi$, but because there is only one random effect in our example it is simple and efficient to do this integration numerically, using a quadrature routine (see Chapter 2 for details). Twenty-one point Gauss-Hermite quadrature was used for the integration and the Nelder-Mead simplex algorithm (also implemented in \texttt{optim()} ) was used for the maximisation. This method does not require numerical derivatives and was found to perform better than the standard quasi Newton-Raphson approaches used by \texttt{optim()}.  

Table 6.1 demonstrates the performance of our two-stage procedure for a range of general population values, chosen so that the proportion of families ascertained varied from approximately 12% to 50%, again starting values for $\alpha', \theta'$ were chosen to be $0.7\alpha$ and
### Table 6.1

Over four scenarios, the average results (standard errors) obtained by analysing 1000 independent data sets, each containing 2000 families of size 5. Columns 3-4 show the average value of $\alpha^*, \theta^*$ compared to their average estimate $\hat{\alpha}^*, \hat{\theta}^*$, obtained by maximisation of the truncated binomial likelihood (6.1.1). Columns 5-6 show how the average estimate for $\alpha, \theta$ differ when the exact and estimated values for $\alpha^*, \theta^*$ are used. The results shown are obtained from a second cycle of pseudo-data generation, for cycle $1 \alpha = 0.7\alpha, \theta = 1.3\theta$.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>General population parameter values</th>
<th>Estimated ascertained parameter values $\alpha^<em>, \theta^</em>$</th>
<th>General population estimates given $\alpha^<em>, \theta^</em>$</th>
<th>General population estimates given $\alpha^<em>, \theta^</em>$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$\alpha = -5$</td>
<td>$\hat{\alpha}^* = -2.19(0.04)$</td>
<td>$\hat{\alpha}^* = -2.16(0.08)$</td>
<td>$-5.01(0.37)$</td>
</tr>
<tr>
<td></td>
<td>$\theta = 4.5$</td>
<td>$\hat{\theta}^* = 2.50(0.09)$</td>
<td>$\hat{\theta}^* = 2.29(0.24)$</td>
<td>$4.52(0.59)$</td>
</tr>
<tr>
<td>(12.3%)</td>
<td></td>
<td></td>
<td></td>
<td>$3.97(0.79)$</td>
</tr>
<tr>
<td>2</td>
<td>$\alpha = -4$</td>
<td>$\hat{\alpha}^* = -3.16(0.02)$</td>
<td>$\hat{\alpha}^* = -3.14(0.13)$</td>
<td>$-4.00(0.08)$</td>
</tr>
<tr>
<td></td>
<td>$\theta = 1$</td>
<td>$\hat{\theta}^* = 0.89(0.03)$</td>
<td>$\hat{\theta}^* = 0.82(0.26)$</td>
<td>$1.00(0.09)$</td>
</tr>
<tr>
<td>(12.5%)</td>
<td></td>
<td></td>
<td></td>
<td>$0.95(0.31)$</td>
</tr>
<tr>
<td>3</td>
<td>$\alpha = -3$</td>
<td>$\hat{\alpha}^* = -1.84(0.03)$</td>
<td>$\hat{\alpha}^* = -1.84(0.07)$</td>
<td>$-3.00(0.14)$</td>
</tr>
<tr>
<td></td>
<td>$\theta = 2$</td>
<td>$\hat{\theta}^* = 1.38(0.04)$</td>
<td>$\hat{\theta}^* = 1.35(0.17)$</td>
<td>$1.99(0.21)$</td>
</tr>
<tr>
<td>(30.2%)</td>
<td></td>
<td></td>
<td></td>
<td>$1.94(0.36)$</td>
</tr>
<tr>
<td>4</td>
<td>$\alpha = -2$</td>
<td>$\hat{\alpha}^* = -0.90(0.03)$</td>
<td>$\hat{\alpha}^* = -0.92(0.05)$</td>
<td>$-2.00(0.18)$</td>
</tr>
<tr>
<td></td>
<td>$\theta = 3$</td>
<td>$\hat{\theta}^* = 1.86(0.07)$</td>
<td>$\hat{\theta}^* = 1.88(0.17)$</td>
<td>$3.00(0.38)$</td>
</tr>
<tr>
<td>(50.2%)</td>
<td></td>
<td></td>
<td></td>
<td>$3.09(0.50)$</td>
</tr>
</tbody>
</table>
1.3θ respectively. The results shown are from a second cycle of pseudo-data generation using the same procedure described previously. Column 5 of Table 6.1 provides further evidence, across all scenarios, that if $\alpha^*, \theta^*$ are known, then $\alpha, \theta$ can be estimated with negligible bias. Column 4 shows the average estimates for the ascertained parameters, $\hat{\alpha}^*, \hat{\theta}^*$, obtained via maximisation of (6.1.1). $\alpha^*$ appears to be estimated with little bias. $\theta^*$ is generally less well estimated, the bias being greatest when $\theta$ is large and $\alpha$ is large and negative. Column 6 shows the average general population estimates $\hat{\alpha}, \hat{\theta}$, that can be obtained in stage 2 when the estimates $\hat{\alpha}^*, \hat{\theta}^*$ are carried forward from stage 1. A selection of the computer programs written to perform this simulation study can be found in appendix A.1.2

6.3.3 Further improvement of the population estimates $\hat{\alpha}, \hat{\theta}$

The accuracy of this two-stage approach to general population parameter estimation relies on accurate estimation of the ascertained parameters $\alpha^*$ and $\theta^*$. In scenarios 2, 3 and 4 of Table 6.1 the estimates derived from the observed data, $\hat{\alpha}^*$ and $\hat{\theta}^*$ show no evidence of bias and consequently they produce good population estimates in the second stage. However in scenario 1, on average, the variance $\hat{\theta}^*$ needed to be increased by 0.21 to be consistent with $\hat{\theta}^*$. Likewise, although clearly a much smaller problem, the estimate $\hat{\alpha}^*$ needed to be increased on average by a 0.03 in order to be consistent with $\hat{\alpha}^*$. Because of this, subsequent estimation of $\hat{\alpha}$ and $\hat{\theta}$ was affected. I now describe a method to guess the likely magnitude of the bias from stage 1, and use this to improve the general population estimates in stage 2.

Let $\hat{\alpha}^*_1, \hat{\theta}^*_1$ be the estimates of the ascertained parameters obtained from analysing a partic-
ular data set $y$, using likelihood (6.1.1) and let $\hat{\alpha}_1, \hat{\theta}_1$ be the general population parameter estimates derived from the conditional likelihood based on pseudo-data with cycling. The accuracy of the estimates can be validated and if necessary further adjusted as follows,

- Simulate new disease data $y_{new}$ using the distribution $p(y_{new}|\hat{\alpha}_1, \hat{\theta}_1)$ and the ascertainment rule. Keep the simulated linear predictors.

- Calculate $\hat{\alpha}_{new}^*, \hat{\theta}_{new}^*$ directly from the simulated linear predictors and $\hat{\alpha}_{new}^*, \hat{\theta}_{new}^*$ from the disease data $y_{new}$ using likelihood (6.1.1).

- Re-simulate linear predictors $\eta^*$ from a $N(\hat{\alpha}_{1}^* + (\hat{\alpha}_{new}^* - \hat{\alpha}_{new}^*), \hat{\theta}_{1}^* + (\hat{\theta}_{new}^* - \hat{\theta}_{new}^*))$ and use as before to derive updated general population parameter estimates $\hat{\alpha}_2, \hat{\theta}_2$.

Figure 6.3: 100 estimates for the general population parameter $\alpha = -5$ obtained with the stage 1 validation step ($\hat{\alpha}_2$) and without ($\hat{\alpha}_1$). The mean estimate is improved from -4.63 to -5.02.

Figure 6.3 illustrates how this extra adjustment impacts on the estimates obtained for $\alpha$ under scenario 1 for 100 data sets of 2000 families. The average estimate for $\hat{\alpha}$ obtained
with 2 cycles of pseudo-data, but without the extra adjustment is -4.63. This is in line with the results from Table 6.1. When the extra adjustment is applied, the average estimate for \( \hat{\alpha} \) is -5.02, while for \( \hat{\theta} \) (not shown) the estimate is improved from 3.94 to 4.51. The equivalent results for \( \theta \) are shown in appendix A.2.

So, even though, in some circumstances, maximisation of likelihood (6.1.1) leads to inaccurate estimation of the ascertained parameters, (Epstein et al., 2002), it is still possible to return good quality estimates for the general population parameters using the two-stage method.

### 6.4 Application to complex models

Building on previous work in Burton et al. (2001), Burton (2003) investigated the effect of ascertainment bias on disease data \( y_{ij} \sim \text{bern}(\mu_{ij}) \) from the logistic model

\[
\log \left\{ \frac{\mu_{ij}}{1 - \mu_{ij}} \right\} = \eta_{ij} = \alpha + \beta_h b_{ij} + \beta_q q_{ij} + \xi_{ij} \tag{6.4.1}
\]

where \( b_{ij}, q_{ij} \) were binary and continuous covariates. In that study each family had 5 members made up of 2 parents \((j=1,2)\) and 3 children \((j=3,4,5)\) and they shared genetic and environmental random effects so that each family’s random effects \( \xi_i \) followed a multivariate normal distribution with zero mean and covariance matrix
6.4. Application to complex models

\[ \Sigma_\theta = \begin{pmatrix} V & U_1 & U_2 & U_2 & U_2 \\ U_1 & V & U_2 & U_2 & U_2 \\ U_2 & U_2 & V & U_2 & U_2 \\ U_2 & U_2 & U_2 & V & U_2 \\ U_2 & U_2 & U_2 & U_2 & V \end{pmatrix} \]

\[ V = \sigma_c^2 + \sigma_a^2, \quad U_1 = \sigma_c^2, \quad U_2 = \sigma_c^2 + \frac{1}{2}\sigma_a^2 \]  \hspace{1cm} (6.4.2)

In this model \( \sigma_c^2 \) represents the variance of an environmental random effect and \( \sigma_a^2 \) represents the variance of a genetic random effect. Each parent is assumed to be genetically independent but to share, on average, half of their genetic random effect with their offspring. Similarly, each sibling is assumed to share, on average, half their genetic variance with every other sibling. \( U_1 \) is therefore the covariance between parents, and \( U_2 \) is the covariance between any parent-child or child-child combination. Burton (2003) imposed the complete ascertainment criteria that all families must have at least 1 affected sibling, that is \( \sum_{j=3}^5 y_{ij} \geq 1 \).

Estimation of the general population parameters \( (\alpha, \beta_b, \beta_q, \theta) \) in the traditional way would have involved a complex likelihood analogous to (5.2.2). The complex correlation structure would make the integral in the denominator of (5.2.2) three-dimensional and the integral on its numerator five-dimensional. Using the same method proposed in Burton et al. (2001), the posterior distribution (shown at iteration \( r \))

\[ p(\alpha^r, \beta_b^r, \beta_q^r, \theta^r | y) \propto \prod_{i=1}^n \left\{ \frac{\prod_{j=1}^{m_i} \exp y_{ij}(\alpha^r + \beta_b^r b_{ij} + \beta_q^r q_{ij} + \theta_{ij})}{1 + \exp (\alpha^r + \beta_b^r b_{ij} + \beta_q^r q_{ij} + \theta_{ij})} \right\} ^{m_i} \frac{g(\alpha^r, \beta_b^r, \beta_q^r, \theta^r)}{1 + \exp (\alpha^r + \beta_b^r b_{ij} + \beta_q^r q_{ij} + \theta_{ij})} \]  \hspace{1cm} (6.4.3)

was instead used to model the disease data, to return the parameter estimates pertain-
6.4. Application to complex models

ing to the ascertained population. On the whole there was reasonable agreement with
the empirical mean and variance estimates from the ascertained linear predictors and the
parameters estimated via (6.4.3). Burton showed that the linear predictors for each fam-
ily were approximately multivariate normally distributed, by comparing the Mahalanobis
distances to a $\chi^2$ distribution, (Mahalanobis, 1936). Because the assumption of normality
appears to be satisfied, if likelihood (6.4.3), or another method could be relied upon to
estimate the parameters that describe this distribution, Clayton’s method could be used
to estimate the general population parameters as before.

To demonstrate this I generated data from scenario E in Burton (2003), by simulating an
ascertained population of 2000 families, from which the general population parameters and
covariances took the values $\alpha = -6$, $\beta_b = 0.7$, $\beta_q = 0.4$, $U_1 = 2$ and $U_2 = 3$. The covariance
values for $U_1, U_2$ imply that $\sigma_c^2 = \sigma_a^2 = 2$. In this first stage, the fixed effects were obtained
directly from the linear predictor $\eta^*$, from which the empirical variance/covariance matrix
was calculated, and used to estimate $\Sigma_{\theta^*}$, without attempting to explicitly estimate any
of the variance components individually.

Figure 6.4 shows the estimates obtained for the general population parameters over 300
independent simulations of scenario E, using the two-stage approach. For the second stage,
linear predictors $\eta^*$ were re-simulated from a $\mathcal{MVN}_5$ distribution using the empirical vari-
ance/covariance matrix, and were then augmented with pseudo linear predictors gener-
ated from general population parameters estimates of $\alpha' = -5, \beta_b' = 0.72, \beta_q' = 0.42, \sigma_a'^2 = 
2.5, \sigma_c'^2 = 1.5$. I allowed myself to make near perfect guesses for the regression coefficients
$\beta_b, \beta_q$ because they are the easiest parameters to estimate. This is because their values
are essentially unchanged by the ascertainment process, as Figure A.6 illustrates. The
results show that after one cycle of pseudo-data, approximately unbiased estimates for the general population parameters can be obtained. Small errors in the estimates could be removed by updating our guesses for the pseudo-data parameters, as in Section 6.3.

An important question is, can the two-stage approach be performed in its entirety for this complex model? Only, of course, if the ascertained parameters can be explicitly estimated. After some investigation, the task of explicit ascertained parameter estimation was found to be much more problematic than first appreciated. Under Burton’s com-
plex model and ascertainment scheme, more parameters were required to characterise the ascertained population than the general population. This has certainly been overlooked until now.

To demonstrate this, disease data for 2000 families were again simulated under Burton's complex model (6.4.1) (scenario E) with and without ascertainment. Figure 6.5 (left) shows the mean value of the linear predictors for the parents ($P_1, P_2$) and siblings ($S_1, S_2, S_3$) in the general population as well as their empirical variance-covariance matrix. The mean and variance structure of families in the general population are totally consistent with values for $\alpha, \sigma_c^2, \sigma_s^2$ of -6.2 and 2 respectively. Figure 6.5 (right) highlights the pattern of linear predictors in the ascertained population. Firstly, the mean value of the linear predictors for the parents and siblings are now distinct. This is also shown by Figure A.7. Two separate parameters $\alpha_p^*, \alpha_s^*$ must now be estimated.

<table>
<thead>
<tr>
<th># General population</th>
<th># Ascertained population</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>$\alpha$</td>
</tr>
<tr>
<td>$P$ $S$</td>
<td>$P$ $S$</td>
</tr>
<tr>
<td>-6.05 -6.08</td>
<td>3.63 -3.34</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$\Sigma$</th>
<th>$\Sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$  $P_2$ $S_1$ $S_2$ $S_3$</td>
<td>$P_1$  $P_2$ $S_1$ $S_2$ $S_3$</td>
</tr>
<tr>
<td>3.175 3.119 3.171 4.121 3.107</td>
<td>2.144 2.173 2.032 3.204 1.960</td>
</tr>
</tbody>
</table>

Figure 6.5: Computer output showing the empirical variance-covariance matrices of the linear predictors $\eta$ of the general and ascertained population, under Burton's complex model.

Secondly, and most importantly, the structure of the empirical variance-covariance matrix
is now more complex. Over multiple simulations it has become clear that the linear
predictors for family $i$ in the ascertained population follow a $MVN(\alpha_i^*, \Sigma^*_\theta)$ where

$$\alpha^* = \begin{pmatrix}
\alpha_p^* + \beta_p^* b_{11} + \beta_q^* q_{11} \\
\alpha_p^* + \beta_p^* b_{12} + \beta_q^* q_{12} \\
\alpha_p^* + \beta_p^* b_{13} + \beta_q^* q_{13} \\
\alpha_p^* + \beta_p^* b_{14} + \beta_q^* q_{14}
\end{pmatrix}
\quad
\Sigma^*_\theta = \begin{pmatrix}
V_0^* & U_1^* & U_2^* & U_3^* & U_4^* \\
U_1^* & V_0^* & U_2^* & U_3^* & U_4^* \\
U_2^* & U_1^* & V_1^* & U_3^* & U_4^* \\
U_3^* & U_2^* & U_1^* & V_1^* & U_4^* \\
U_4^* & U_3^* & U_2^* & U_1^* & V_1^* 
\end{pmatrix}
$$

Three distinct covariances now also appear. This distortion is because families are made
up of parents and siblings, but that the ascertainment criterion only applied to the sib­
lings. It was clear that if the MCMC approach of Burton (2003) was going to be used to
perform the first stage of the two-stage procedure effectively, a new algorithm would be
needed to account for this newly identified complexity.

Over many simulations of different general population parameter values it was observed
that the inequalities $U_1^* \leq U_3^* \leq U_2^*$ and $V_0^* \leq V_1^*$ always held. Disregarding the genetic
and environmental interpretations that the general population variance components have,
this suggested that; $U_2^*$ could be represented by a parameter $\sigma^{*2}$, $U_1^*$ could be represented
by $\sigma^{*2} - \sigma_p^{*2}$ and $U_3^*$ could be represented by $\sigma^{*2} - \sigma_e^{*2}$. In order to estimate the param­
ters to describe the covariance structure of (6.4.4) inside an MCMC loop, linear predictors
would need to be simulated for each family with the same pattern of covariances. Defining
the variables $x_1, ..., x_8$ to be

$$x_1 \sim N(0, \sigma^{*2})$$
$$x_2, x_6, x_7, x_8 \sim N(0, \sigma_p^{*2})$$
$$x_3, x_4, x_5 \sim N(0, \sigma_e^{*2})$$
At iteration \( r \), if realisations of \( x_1^r, \ldots, x_5^r \) could be obtained from the above distributions, new variables \( \delta_{P1}, \ldots, \delta_{S3} \), could be derived for each family, where

\[
\begin{align*}
\delta_{P1} &= x_1^r + x_2^r \\
\delta_{P2} &= x_1^r - x_2^r \\
\delta_{S1} &= x_1^r + x_3^r + x_4^r + x_5^r \\
\delta_{S2} &= x_1^r - x_3^r + x_5^r + x_6^r \\
\delta_{S3} &= x_1^r - x_4^r - x_5^r + x_6^r
\end{align*}
\]

\( \delta_{P1}, \ldots, \delta_{S3} \) could then replace the \( \xi^r \) terms in posterior likelihood (6.4.3). \( \delta_{P1}, \delta_{P2} \) imply that the variance of the parents \( V_0^* = \sigma^2 + \sigma_p^2 \) and \( \delta_{S1}, \delta_{S2}, \delta_{S3} \) imply that variance term of the siblings \( V_1^* = \sigma^2 + 2\sigma_s^2 + \sigma_p^2 \). The program code for fitting this model in Winbugs is shown in Appendix A.1.2. Initial simulations have not proved to be successful due to poor mixing and lack of convergence. Normal distributions with zero means and large variances were used for the fixed effects parameters. Various prior distributions for the random effects variances were tested, including gamma distributions and also uniform distributions (for the square-root of the variance parameters).

6.5 Discussion

The two-stage approach offers a practical method of parameter estimation for GLMM's in the presence of ascertainment bias. Providing that it is possible to estimate the parameters of the ascertained population then it is relatively simple to convert these into estimates for the parameters of the general population. Dividing the estimation into two stages greatly simplifies the first stage and enables us to convert the second stage onto the scale of the linear predictors making a much simpler calculation.
There are many methods that could be used for each stage of this procedure. In simple problems, such as the one on which the simulations were based, it is easy to use maximum likelihood and numerical integration for the first stage. For more complex models the MCMC approach of Burton et al. (2001) appears to be an excellent candidate. Noh et al. (2005) suggests a hierarchical likelihood approach that is analogous to the method of Burton and therefore provides another way to estimate ascertained parameters. Similarly, there are alternatives to Clayton's conditional likelihood approach in the second stage. For instance, pseudo-data could be used as an importance sample to construct an approximation to the full likelihood, as in Penttinen (1984); Geyer and Thompson (1992).

However, as discussed in Chapter 2, problems that involve covariates are more easily dealt with inside conditional likelihoods.

I demonstrated this two-stage method in its entirety on a standard example with no covariates and a single random effect, previously used by Glidden and Liang (2002); Noh et al. (2005); Burton et al. (2001), so that it could be easily understood. However, generally the models for real genetic epidemiological studies will be much more complex than this. Typically they will include measured covariates and several correlated random effects. The families are likely to be of varying size and to contain more than one generation. In fact, precisely the sort of model as that of Burton (2003). Generalisation of the two-stage approach to complex models is immediate in theory, because Burton's adapted Bayesian approach (with uninformative priors) is a direct analogy to the truncated likelihood (6.1.1). In the complex case, if (6.4.3) could estimate parameters that accurately described the distribution of linear predictors in the ascertained population, then Figure 6.4 shows that
they can be immediately used in the second stage of the two-stage estimation procedure. Furthermore the second stage is hardly any more computationally intensive, when the distribution of the linear predictors is multivariate, rather than univariate normal, and a conditional likelihood quickly produces estimates for the general population parameters. However, estimating the ascertained population’s variance parameters in Burton’s complex model proved impossible, regardless of the prior distribution chosen. I do not believe that all complex models, with multiple covariates and random effects, will suffer from this problem. The complex correlation structure of the ascertained population in Burton (2003), that leads to identifiability issues, is mostly a result of the fact that, whilst each family is made up of parents and children, the ascertainment criteria is applied to the children only. As further work it would be interesting to find sufficiently complex model and ascertainment mechanism for which the ascertained variance components were identifiable. Once this has been achieved it would be necessary to conduct a thorough sensitivity analysis on how the choice of prior distributions for the ascertained parameters affected the resulting posterior estimates. An general issue with choosing prior distributions for variance parameters is deciding on the appropriate scale. Priors for the actual variance, its standard deviation or its precision are all valid options, but can hugely affect the result, as Lambert et al. (2005) point out.

In this chapter I have not considered the case in which the distribution of genetic and environmental random effects is non-normal. As Glidden and Liang (2002) point out, and Noh et al. (2005) have since thoroughly investigated, if this were the case then the distribution of linear predictors in the ascertained population would not be normally distributed, and our two-stage procedure is not so straightforward to implement. However
in most genetic settings I believe the assumption of normality is likely to be reasonable.

The two-stage method has the potential to be adapted for use when analysing any type of ascertained data under a GLMM model. In essence, a problem involving non-normal data and requiring prohibitively slow multiple integrations, is turned into a problem involving normal data. The separate parameterisations of the mean and variance terms replaces some of the integration and enables parameters to be estimated efficiently and accurately.
Chapter 7

Understanding and improving

Clayton’s conditional likelihood

Monte-Carlo Maximum Likelihood (MCML), (Geyer and Thompson, 1992), incorporates standard importance sampling into likelihood inference. However, as shown in Chapter 2, standard importance sampling can fail dramatically, even for relatively simple problems and so the performance of MCML can be poor. In this chapter the ‘defensive’ strategy of Hesterberg (1995), for improving standard importance sampling, is discussed and it is shown that Clayton’s conditional likelihood naturally exhibits a weak version of this property. A method for improving the performance of MCML called ‘Reverse Logistic Regression’ (RLR), (Geyer, 1994), is then introduced as it too makes use of defensive sampling. The link between Clayton’s and Geyer’s methods is described, and a new method is proposed that combines their most attractive properties. Crucially, this modification allows three desirable characteristics to be synthesised. Firstly, covariates can be included easily in a model. Secondly, a mixture distribution of pseudo observations can be sampled without the parameters of the mixture distribution being explicitly estimated. Thirdly,
Inference can be based on maximisation of a single likelihood function. I presented the main ideas in this chapter at the *International Biometric Society* conference in 2006.

### 7.1 Improving standard importance sampling

Monte-Carlo integration has become an extremely widely used technique for solving problems that are analytically intractable or computationally infeasible (with standard numerical methods), due to their high dimension. As shown in Chapter 2, importance sampling is in essence a technique that improves the performance and scope of Monte-Carlo integration, but there is a continual drive to improve its efficiency still further so that new and increasingly complex systems can be investigated. A common way to measure the success of an importance sampling procedure is the Monte-Carlo variance it produces for a particular parameter estimate. In statistics, methods that produce parameter estimates with a smaller variance are generally preferred because they naturally provide more power for hypothesis testing. Additionally however, if the Monte-Carlo variance of an estimate is too large, importance sampling based algorithms can, in certain cases, fail to find a solution altogether. This makes the topic of 'variance reduction' doubly important when Monte-Carlo methods are in use. Consider the general problem of obtaining a value for the integral

\[ E[I(Y)] = \int I(y) h(y|\theta) dy = z(\theta) \]

for which
\[ V ar[I(Y)] = \int I^2(y)h(y|\theta)dy - E[I(Y)]^2 = \sigma^2 \]

Integrals of this sort are encountered in many areas of science. For example in physics \( I(y) \) could be a model for quantum particles in high dimensional space, (Creutz, 1980). In telecommunications \( I(y) \) is used to estimate congestion probabilities for mobile telephone networks, (Paschalidis and Vassilaras, 2004). In chemistry, \( I(y) \) is used to predict the fluidity of gases at various temperatures, (Toschi et al., 2003). In the case of selection bias \( I(y) \) could be an indicator function that denotes membership of a set \( C \), thereby making \( z(\theta) \) the true probability of selection.

### 7.1.1 Defensive importance sampling

The standard Monte-Carlo method tells us to simulate data \( y_1, \ldots, y_n \) from \( h(y|\theta) \) and to estimate \( z(\theta) \) by

\[ \hat{z}(\theta) = \frac{1}{n} \sum_{i=1}^{n} I(y_i) \]

In a general sense, when the functions \( I(y) \) and \( h(y|\theta) \) are very ‘different’, simulating from \( h(y|\theta) \) might be inefficient. For whatever reason, pseudo-data \( y_1, \ldots, y_n \) can be simulated from a distribution \( h(y|\theta'_1) \) and inferences about \( z(\theta) \) can be based on the importance sample \( \frac{1}{n} \sum_{i=1}^{n} I(y_i)h(y|\theta)/h(y|\theta'_1) \) instead. While \( h(y|\theta'_1) \) may generally be a good density to work with, there may be particular values of \( y \) for which \( h(y|\theta'_1) \) approaches 0 but \( h(y|\theta) \) does not, thereby inflating the variance considerably. Hesterberg (1995) proposed a method to guard against this, by showing that, if data could be sampled from the mixture distribution

\[ h_{mix}(y) = \alpha_0 h(y|\theta) + \sum_{j=1}^{m} \alpha_j h(y|\theta'_j) \]  

where \( \sum_{j=0}^{m} \alpha_j = 1 \), then the variance of \( E[I(Y)] \) would equal
(7.1.2) tells us that if the original variance $\sigma^2$ is finite, then the variance of the Monte-Carlo estimate will also be finite. Hesterberg termed this property 'defensive' for obvious reasons.

Obtaining importance sampling estimates for $z(\theta)$ is not my primary focus in this thesis. Importance sampling has been used inside a conditional liklelihood function, so that statistical inferences can be made about an unknown parameter $\theta$. Because $\theta$ is unknown, it would therefore seem impossible to exploit Hesterberg's defensive principle within a data augmentation routine, since pseudo-data cannot be simulated given $\theta$. However, on the contrary, it is key to understanding both why Clayton's method performs so well, and also how it can be improved.

### 7.2 Defensive data augmentation

Let matrix (7.2.1) contain real data $y_{i0}$ from $f(y|\theta)$, and pseudo-data $y_{ij}$ simulated from $f(y|\theta_j)$, $j = 1, ..., m$, the $\theta_j$'s being a range of guesses for $\theta$.

Assuming (briefly) that the pseudo-data $y_{ij}$ in matrix (7.2.1) has come from a single
7.2. Defensive data augmentation

\[
\begin{pmatrix}
\theta & \theta_1' & \theta_2' & \ldots & \theta_m' \\
y_{10} & y_{11} & y_{12} & \ldots & y_{1m} \\
y_{20} & y_{21} & y_{22} & \ldots & y_{2m} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
y_{n0} & y_{n1} & y_{n2} & \ldots & y_{nm}
\end{pmatrix}
\]  

(7.2.1)

Figure 7.1: Augmented real data \( y_{i0} \ i = 1, \ldots, n \) and pseudo-data \( y_{ij} \ j = 1, \ldots, m \). \( \theta \) is an unknown parameter, \( \theta_1', \ldots, \theta_m' \) are 'known' parameters that are used to generate the pseudo-data. Note, \( \theta_1', \ldots, \theta_m' \) may, or may not be distinct.

The mixture distribution \( f(y|\theta') \), meaning that \( \theta_1' = \theta_2' = \ldots = \theta_m' \), Clayton’s argument can be exploited to calculate the conditional probability that, given \( y_{i0} \) is in the set \( \zeta_i = \{ y_{i0}, y_{i1}, \ldots, y_{im} \} \), \( y_{i0} \) is 'real'. This probability greatly simplifies to

\[
\frac{h(y_{i0}|\theta)}{h(y_{i0}|\theta')} + \sum_{j=1}^{m} \frac{h(y_{ij}|\theta)}{h(y_{ij}|\theta')}
\]

(7.2.2)

and forms the \( i \)'th term of his conditional likelihood (2.3.9). Because it is made up of real data, given \( \theta \), and pseudo-data, given \( \theta' \), the form of Clayton’s denominator means that it can claim the defensive property of Hesterberg. Clayton’s mixture distribution is, in effect

\[
\frac{1}{m+1} f(y|\theta) + \frac{m}{m+1} f(y|\theta')
\]

(7.2.3)

In Section 2.3.2 it was shown that even for a relatively simple model of selection biased data, the performance of MCML, (Geyer and Thompson, 1992), was poor. In Chapters 3 to 6 Clayton’s method was chosen, above MCML, to implement bias corrections in the areas of meta-analysis and genetic epidemiology, a task which it performed well. I believe that its defensive qualities, which are totally lacking in MCML, can partly explain this
success. Can $m$ different, and therefore strongly defensive, pseudo-data distributions ever be used in a data augmentation routine for likelihood inference?

### 7.2.1 Reverse Logistic Regression

Geyer (1994) suggested a two-step procedure called ‘Reverse Logistic Regression’ (RLR), to reduce the variance of MCML estimates over a larger range of $\theta$, possibly in answer the specific criticisms by Green (1992) and Ogata (1992). In order to explain RLR in a way that facilitates an easy comparison with Clayton’s conditional likelihood, let matrix (7.2.1) contain real data $y_{i0}$ from $f(y|\theta)$, and pseudo-data $y_{ij}$ simulated from $f(y|\theta^*_j)$, as before. This time $\theta^*_1,...,\theta^*_m$ are distinct. Although there are $m$ separate pseudo-data distributions, Geyer suggests that one could assume that all pseudo-data $y_{ij}$ $i = 1,...,n$ $j = 1,...,m$ have come from a single mixture distribution

$$f_{mix} = \sum_{j=1}^{m} \frac{1}{m} f(y|\theta^*_j) = \sum_{j=1}^{m} \frac{1}{m} \frac{h(y|\theta^*_j)}{z(\theta^*_j)} \quad (7.2.4)$$

Because there are $n$ data points from each of the $m$ pseudo-data distributions, each one is given an equal weight of $\frac{1}{m}$. The probability that, given $y$ is from the mixture distribution $f_{mix}$, it is actually generated from $\theta^*_j$ can be shown to equal

$$p_j(y; z(\theta^*_j)) = \frac{h(y|\theta^*_j)}{\sum_{k=1}^{m} \frac{h(y|\theta^*_k)}{z(\theta^*_k)}}$$

The fact is, we know which $\theta^*_j$ generated each pseudo-data point, but Geyer cleverly noted this conditional probability statement offers a convenient vehicle for estimation of $\phi_j = \frac{z(\theta^*_j)}{z(\theta^*_m)}, j = 1,...,m - 1$ by maximisation of
7.2. Defensive data augmentation

\[ l(\phi_1, ..., \phi_{m-1}|y) = \sum_{j=1}^{m} \sum_{i=1}^{n} \log \{ p_j(y_{ij}, \phi_j) \} \]  

(7.2.5)

It is necessary for \( \phi_m \) to be fixed at a constant value in order to act as a pivot for estimating \( \phi_1, ..., \phi_{m-1} \). These first stage estimates, derived solely using pseudo-data can then be inserted into an approximate profile likelihood,

\[ l(\theta|\hat{\phi}_1, ..., \hat{\phi}_{m-1}, y) = \sum_{i=1}^{n} \left[ \log \left\{ \frac{h(y_{i0}|\theta)}{\sum_{j=1}^{m} h(y_{i0}|\theta) \phi_j} \right\} - \log \left\{ \frac{1}{mn} \sum_{j=1}^{m} \sum_{i=1}^{n} \frac{h(y_{ij}|\theta)}{\phi_k} \right\} \right] \]  

(7.2.6)

to be maximised with respect to \( \theta \).

It was shown in Section 2.3.2 that replacing the exact selection probability \( \frac{z(\theta)}{z(\theta')} \) with a Monte-Carlo estimate made MCML unstable. However, simulating importance weights from a range of distributions and replacing exact selection probabilities \( \phi_1...\phi_m \) with \( \hat{\phi}_1...\hat{\phi}_m \), will mean that they will be informative over a wider range of \( \theta \), and should greatly reduce the variance of \( \hat{\theta} \), as shown by Hesterberg (1995) and Owen and Zhou (2000).

When all observations \( y_{i0}, i = 1, ..., n \) have a common probability of selection that requires the same normalising constant \( z(\theta) \), but there is considerable uncertainty as to the true value of \( \theta \), RLR should work well. Every pseudo-data point \( y_{ij} \) is used in the estimation of \( \phi_j \) via maximisation of likelihood (7.2.5), which makes them highly accurate. However, since a normalising constant \( \phi \) must be calculated for every distribution, the
7.3. Combining RLR and conditional likelihood inference

variance reduction in $\hat{\theta}$ comes at the cost of computational speed. Although this two step procedure is a nice solution, the error in the estimation of the first stage $\phi$'s is not properly carried forward, and subsequently the coverage of the confidence intervals for $\theta$ will be sub-optimal. The performance of Geyer's RLR method is investigated in Section 7.4.

7.3 Combining RLR and conditional likelihood inference

I now consider how Clayton's conditional likelihood is linked to RLR. If Geyer's mixture distribution reasoning is applied only to the real data (column 0 of matrix (7.2.1), under the condition that $\theta_1 = \theta_2 = ... = \theta_m$, then the log-likelihood for $\theta$ can be simplified to

$$l(\theta|y; \theta') = \sum_{i=1}^{n} \sum_{j=0}^{0} \log \left( \frac{h(y_{i0}|\theta)}{\frac{h(y_{i0}|\theta)}{z(\theta)} + m \frac{h(y_{i0}|\theta')}{z(\theta')}} \right)$$

Replacing by an $m$-term Monte-Carlo estimate $\frac{1}{m} \sum_{j=1}^{m} \frac{h(y_{ij}|\theta)}{h(y_{ij}|\theta')}$ sees (7.3.1) reduce to Clayton's original expression (7.2.2). Applying Geyer's argument to the real data in matrix (7.2.1) without the restriction that the $\theta_j$'s are equal, and replacing the ratio $\frac{z(\theta)}{z(\theta_j)}$ by a 1-term Monte-Carlo estimate $\frac{h(y_{ij}|\theta)}{h(y_{ij}|\theta_j)}$, the following new likelihood can be derived.
7.3. Combining RLR and conditional likelihood inference

\[
l(\theta|y; \theta_1', \ldots, \theta_m') = \sum_{i=1}^n \log \left( \frac{h(y_{i0}|\theta)}{z(\theta)} + \sum_{j=1}^m \frac{h(y_{ij}|\theta_j')}{z(\theta_j')} \right)
\]

Equation (7.3.2) maintains all the attractive properties of Clayton's original conditional method. It is a single expression in which \( \theta \), and only \( \theta \) is estimated. Additionally, the pseudo-data used to weight each real observation \( y_{i0} \) is allowed to come from a range of distributions, so it can be considered maximally defensive, in the true fashion that Hesterberg intended.

When a single selection probability \( z(\theta) \) is required, all the algorithms previously discussed can be implemented. But, as discussed in Chapter 2, if \( y_{i0} \) is dependent on a continuous covariate \( x_{i0} \), then each real observation requires a unique normalising constant \( z(\theta|x_i) \).

The computational demands of Geyer's method are now extreme. Since there are \( n \) unique normalising constants, \( nm - 1 \) quantities \( \phi_{ij} = \frac{z(\theta|x_i)}{z(\theta_j|x_i)} \) must be estimated in order to implement his first stage. Furthermore, as only pseudo-data simulated conditional on \( x_i \) can be used to approximate \( z(\theta|x_i) \), much more pseudo-data is required. By contrast, Clayton's method, and the combined approach, can still be implemented with no extra difficulty. This is because the normalising constant ratios \( \phi \) are either conditioned out or replaced with Monte-Carlo estimates, and pseudo-data \( y_{ij}, j = 1, \ldots, m \) is only ever used to weight real data point \( y_{i0} \).
Table 7.1 summarises the strengths and weaknesses of the data augmentation algorithms discussed in this chapter.

<table>
<thead>
<tr>
<th>Method (equation)</th>
<th>Nuisance parameters estimated</th>
<th>Likelihoods to maximise</th>
<th>Mixture distribution for pseudo-data?</th>
<th>Covariates easily dealt with?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Geyer’s RLR ( (7.2.5) ) and ( (7.2.6) )</td>
<td>m-1</td>
<td>2</td>
<td>YES</td>
<td>NO</td>
</tr>
<tr>
<td>Clayton’s Conditional likelihood ( (2.3.9) )</td>
<td>0</td>
<td>1</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Combined method ( (7.3.2) )</td>
<td>0</td>
<td>1</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

Table 7.1: A comparison of the characteristics of the 3 competing data augmentation algorithms discussed.

In the next section I conduct two simulation studies to highlight exactly how the characteristics of each algorithm impacts on performance. The first example uses data from a simple model, similar to that in Section 2.3.2. The second example involves data from a more complex, genetic epidemiology setting.

### 7.4 Simulation studies

#### 7.4.1 Truncated normal example

The ability of the existing and newly proposed methods, listed in Table 7.1, are investigated for the truncated normal example described in Section 2.3.2. Table 7.2 shows the results of 5000 independent simulations. In each case 100 real data points \( y_{10}, \ldots, y_{1100} \) were sampled.
from a $N(\theta = 0, 1)$ distribution, subject to being greater or equal to 0. This equated to a selection probability of $\Phi(0) = 50\%$. For all methods that permitted a mixture distribution of pseudo-data, matrix (7.2.1) was constructed by setting $m$ equal to 5 and $\theta_1', \ldots, \theta_5'$ equal to $(-0.5, -0.25, 0, 0.25, 0.5)$. To demonstrate Clayton's method, $m = 5$ lots of pseudo-data points were sampled (from a single truncated distribution) for each real point, with $\theta' = 0.3$. This meant that, in both cases, the average distance between $\theta$ and $\theta'$ was the same. A second, near identical simulation was performed for all methods, the only change being that all real and pseudo-data was subject to a more severe truncation of being greater or equal to 1. This equated to a selection probability of $\Phi(-1) \approx 16\%$. Calculations were performed using the statistical package R and the optimisation function `optim()`. A selection of the program code written to implement these methods is shown in Appendix A.1.3.

<table>
<thead>
<tr>
<th>Method</th>
<th>Selection criterion passed</th>
<th>Average $\theta$ estimate</th>
<th>Average RMSE</th>
<th>Coverage of 95% C.I</th>
<th>average CPU time (secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$RLR$</td>
<td>$\geq 0$</td>
<td>-0.01</td>
<td>0.18</td>
<td>93%</td>
<td>2.34</td>
</tr>
<tr>
<td></td>
<td>$\geq 1$</td>
<td>-0.01</td>
<td>0.25</td>
<td>93%</td>
<td>2.62</td>
</tr>
<tr>
<td>Conditional</td>
<td>$\geq 0$</td>
<td>-0.01</td>
<td>0.18</td>
<td>96%</td>
<td>0.30</td>
</tr>
<tr>
<td>likelihood</td>
<td>$\geq 1$</td>
<td>-0.01</td>
<td>0.25</td>
<td>95%</td>
<td>0.33</td>
</tr>
<tr>
<td>Combined method</td>
<td>$\geq 0$</td>
<td>0.00</td>
<td>0.19</td>
<td>95%</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td>$\geq 1$</td>
<td>0.00</td>
<td>0.26</td>
<td>95%</td>
<td>0.36</td>
</tr>
</tbody>
</table>

Table 7.2: A comparison of 3 competing data augmentation methods performance on the truncated normal example previously described. Column 3 shows the average estimate for $\theta$, column 4, shows the average $\sqrt{\text{MSE}}$ for 5000 independent simulations. Coverage calculation assumes asymptotic normality for the distribution of $\theta$ estimates.

Column 2 indicates which selection criterion was in operation and column 3 shows the average estimate obtained for $\theta$ (its true value being 0), over the 5000 simulations. Column 4 shows the average root mean squared error (RMSE) for these estimates. The results
show that all methods can return approximately unbiased estimates for $\theta$. Unsurprisingly, all methods perform better, in terms of their mean squared error and speed of convergence, when the selection probability equals 50% rather than 16%. However, there are differences in their performance when this selection probability is held constant. Geyer's RLR method is the slowest. This is unsurprising because it had to estimate 4 nuisance parameters, as well as $\theta$. The coverage of the 95% confidence intervals for $\theta$ produced by Geyer's RLR method are sub-optimal, indicating that their variances are underestimated. Clayton's original method appears to be the best performing algorithm, by marginally outperforming our combined approach, in terms of computational speed and mean squared error.

### 7.4.2 Affected sibling pairs

An example of ascertainment biased data, from a fictional genetic epidemiological study is now introduced. Let

$$ y_{ik} = \alpha + \beta x_{ik} + \epsilon_{ik}, \quad i = 1, \ldots, n \quad k = 1, 2 \quad (7.4.1) $$

be the model that predicts the continuous response $y_{ik}$ of sibling $k$ in family $i$. To account properly for the within family and between family variation in the general population

$$ \epsilon_i \sim MVN(0, \Sigma) \quad \Sigma = \begin{pmatrix} \sigma^2 + \tau^2 & \tau^2 \\ \tau^2 & \sigma^2 + \tau^2 \end{pmatrix} $$

Figure 7.2 shows the distribution of a simulated continuous epidemiological endpoint in a general and an ascertained population of sibling pairs. The ascertained sibling pairs all pass the condition of having at least one 'affected' member, by nature of their endpoint being greater than a particular threshold.
7.4. Simulation studies

Figure 7.2: The bivariate distribution of an epidemiological response in the general (left) and ascertained (right) population of sibling pairs.

The distribution of the response $y$ in the general and ascertained populations in Figure 7.2 was generated by true parameter values $\alpha = 5$, $\beta = 3$ and $\sigma^2 = \tau^2 = 2$. The continuous covariates $x$ were independently drawn from a $U(-1,1)$ distribution. The data were subject to 'complete ascertainment' as described in equation (2.1.3) and in Chapter 5, the criterion being $y_i \in C$ if $y_{i1} \cup y_{i2} \geq 8$. This equated to an average selection probability for a sibling pair of approximately 23%. Let $\theta = (\alpha, \beta, \Sigma)$ and $h(y_i | \theta, x_i)$ represent a multivariate normal distribution with mean $\alpha + \beta x_i$ and variance $\Sigma$.

$$
\begin{pmatrix}
\text{Cov} & \theta & \theta' & \ldots & \theta'_m \\
x_{1.0} & y_{1.0} & y_{1.1} & y_{1.2} & \ldots & y_{1.m} \\
x_{2.0} & y_{2.0} & y_{2.1} & y_{2.2} & \ldots & y_{2.m} \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\
x_{n.0} & y_{n.0} & y_{n.1} & y_{n.2} & \ldots & y_{n.m}
\end{pmatrix}
$$

(7.4.2)

Figure 7.3: Augmented real and pseudo-data, matched through the sharing of a dependent variable, or covariate, $x$.

Row $i$ of data array (7.4.2) contains; a real sib pair response $y_{i.0}$, their covariate informa-
### 7.4. Simulation studies

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter estimates</th>
<th>CPU time (secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha = 5$</td>
<td>$\beta = 3$</td>
</tr>
<tr>
<td><strong>Conditional likelihood</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source 1</td>
<td>4.53(2.16)</td>
<td>2.98(0.29)</td>
</tr>
<tr>
<td>Coverage</td>
<td>95%</td>
<td>95%</td>
</tr>
<tr>
<td>Source 2</td>
<td>4.93(0.53)</td>
<td>3.01(0.18)</td>
</tr>
<tr>
<td>Coverage</td>
<td>95%</td>
<td>94%</td>
</tr>
<tr>
<td>Source 3</td>
<td>4.15(2.60)</td>
<td>3.40(1.37)</td>
</tr>
<tr>
<td>Coverage</td>
<td>91%</td>
<td>94%</td>
</tr>
<tr>
<td><strong>Combined method</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Source 4</td>
<td>4.75(0.76)</td>
<td>3.07(0.25)</td>
</tr>
<tr>
<td>Coverage</td>
<td>95%</td>
<td>92%</td>
</tr>
</tbody>
</table>

Table 7.3: Parameter estimates (average $\sqrt{mSE}$), and coverages for the complex model using Clayton’s method, and our combined approach. Coverages again relate to a nominally 95% confidence interval.

The presence of a continuous covariate makes them extremely computationally demanding, so I did not attempt to apply Geyer’s RLR method on this example. As before, if $\theta'_1 = \theta'_2 = ... = \theta'_m = \theta'$ then Clayton’s conditional likelihood method can be implemented by approximating the selection probability ratio $\frac{z(\theta' | x_i)}{z(\theta'_j | x_i)}$ with the $m$-term Monte-Carlo estimate $\frac{1}{m} \sum_{j=1}^{m} \frac{h(y_{i,j} | \theta, x_i)}{h(y_{i,j} | \theta', x_i)}$.

If instead pseudo-data are sampled from $m$ different distributions then our combined approach can be implemented by approximating the selection probability ratio $\frac{z(\theta' | x_i)}{z(\theta'_j | x_i)}$ with a 1-term Monte-Carlo estimate $\frac{h(y_{i,j} | \theta, x_i)}{h(y_{i,j} | \theta', x_i)}$.

Table 7.3 shows the results of 500 independent simulations. In each case a real data set of 200 affected sibling pairs were simulated from model (7.4.1), using the parameter values previously described. $m = 5$ lots of pseudo sibling pairs were simulated for each real sib pair and parameter estimates were obtained by maximising the appropriate likelihoods.
7.4. Simulation studies

For Clayton’s method they were generated from one of three different sources of pseudo-data

- **Source 1**: $\alpha' = 3, \beta' = 5, \tau^2 = 0.5, \sigma'^2 = 4$

- **Source 2**: $\alpha' = 5, \beta' = 3, \tau^2 = 2, \sigma'^2 = 2$

- **Source 3**: $\alpha' = 7, \beta' = 1, \tau^2 = 4, \sigma'^2 = 0.5$

Source 2, the case where $\theta = \theta'$ was chosen to illustrate Clayton’s method at its optimum. Sources 1 and 3 represent the realistic situation where $\theta \neq \theta'$. For our combined method a single pseudo sibling pair was simulated from each of 5 different distributions, for each real sib pair, so that

- **Source 4**: 

  $(\alpha', \beta', \tau^2, \sigma'^2) = (3,5,0.5,4), (4,4,1,3), (5,3,2,2), (6,2,3,1), (7,1,4,0.5)$

The parameter values chosen for source 4 cover the range of pseudo-data simulated in sources 1 to 3.

Clayton’s method is the best performing algorithm when $\theta = \theta'$ (source 3). Computationally, it is the fastest, it has the smallest average mean squared error, and its coverage is the most optimal. But, in reality there might be a great deal of uncertainty as to the value of $\theta$, especially when it is multi-dimensional. When this uncertainty is included into the pseudo-data generation process, via pseudo-data sources 1,3 and 4, it no longer appears to be the best option. The best performing method in terms of mean squared error and coverage is now the combined approach, although it is marginally slower than Clayton’s algorithm.
7.5 Discussion

Correcting data coming from a truncated normal distribution for selection bias is clearly a trivial task that does not necessitate a technique such as importance sampling. As seen in Chapter 4 results from Todd et al. (1996), or Copas and Jackson (2004), or even equation (4.5.4), could have been used to obtain an analytic estimate for the bias. In the absence of exact theoretical solutions to selection bias, accurate numerical approximations to selection probabilities should be considered. For example, in the case of a univariate normal distribution, the pnorm() function in R, (Wichura, 1988). Indeed in Section 7.4.2 the importance sampling selection probability estimates could have been replaced with a numerical approximation to the normalising constant of a multivariate normal distribution, using pmvnorm() in R, (Genz et al., 2006). Such numerical approximations should be utilised wherever possible because they have zero variance.

Many genetic epidemiological study designs are even more complex than the sibling pair example. They often involve variable family sizes from multiple generations. This increases the complexity of the within family correlation structure and hence the dimension of the integration required to calculate the appropriate normalising constant. This makes numerical approximations less accurate and more computationally intensive. In genetic epidemiology perhaps the most challenging models to successfully correct for selection bias occur when the data are highly correlated and non-normal, as was discussed in Chapter 5. Neither standard importance sampling nor numerical integration provide an adequate framework for statistical inference in this case, but in Chapter 6 I showed that a two-stage approach can provide an approximate but feasible solution. In complex problems, the
combined likelihood (7.3.2) is perfect for an initial analysis when relatively little is known about $\theta$. Pseudo-data can be simulated from a range of distributions to produce initial estimates, which can then inform successive pseudo data generation cycles, as illustrated in Chapters 3 and 6. Once $\theta$ has been estimated accurately a final simulation phase should be implemented using pseudo data sampled from this singular $\theta$ and Clayton’s method should be used. This produces the most reliable variance estimates, by nature of being an exact conditional likelihood.

When the outcome variable $y$ is known to depend on a real covariate $x$, as is often the case in biomedical research, it is unwise to use RLR to derive intermediate parameter estimates $\hat{\phi}$, for the purpose of improving the performance of MCML. However, Geyer’s conditional argument for RLR is, in my opinion, extremely sound in its own right. In thinking of it merely as an addition to MCML, its value has been overlooked.
Chapter 8

Discussion and conclusions

Selection bias, if not properly corrected for, can have a potentially disastrous effect on biomedical research, by throwing into doubt the scientific validity of a result, or by severely reducing its applicability in a wider context. It is therefore vital that statistical methodology is developed to make bias correction possible. By viewing selection bias as a problem of mathematical integration, importance sampling was identified as an efficient and underused method, that could offer a solution. Following investigation, Clayton's conditional likelihood, that utilised importance sampling via data augmentation, was identified as a perfect vehicle for performing statistical inference in the presence of selection bias. Such was its attractiveness, practical applications for this approach immediately presented themselves. I now review the work of this thesis, in an attempt to highlight both its achievements as well as its limitations, and point to future work that could be conducted in this area.
8.1 Meta-analysis

In Chapter 3 I explored the possibility of using Clayton’s conditional likelihood to model publication biased data in the field of meta-analysis. By simulating ‘pseudo studies’ under a known publication criteria, approximately unbiased estimates for the mean study effect and the between study standard deviation were obtained. Whilst data augmentation was not new to this area per se, see for example the Trim and Fill method of Duval and Tweedie (2000), the notion of augmenting real studies with pseudo studies that had passed a publication criteria (and where hence observable), as apposed to missing and therefore unobservable studies, was new. It has the potential to be a useful tool. Its biggest strength is that publication models of any complexity can be implemented. I believe also that it is fairly intuitive and easy to implement. This is especially important in the context of meta-analysis, since the majority of practitioners will come from a non-statistical background.

Rather than a mean difference, it is much more common for meta-analyses to be based on log odds ratios, the natural statistic calculated after a case-control study. Generalising Clayton’s conditional likelihood, from handling mean differences to log odds ratios, is immediate if they are assumed to be normally distributed, but what happens when this assumption breaks down, for example if the number of exposed cases or controls is 0? A common fix is to add 0.5 to each cell and to proceed with the normal assumption as before, although there is little formal justification for doing so. Shi and Copas (2002) suggested a new method to for meta-analyses of $2 \times 2$ tables, that made no assumption of normality. Their conditional argument required the evaluation of a complex integral in
8.1. Meta-analysis

order for inferences to be made about the log odds ratios mean and variance parameters, for which they used a specialised MCMC algorithm, (McLachlan and Krishnan, 1997). Although Clayton’s approach does not appear to be applicable in this setting, importance sampling could be used to integrate their complex function instead of MCMC, and may be faster to converge.

Clayton’s method can handle selection models of any complexity and, if used in a meta-analysis, each study can be assumed to have passed a unique publication criterion. But, as demonstrated in Chapter 3, it can not be used when studies are only known to have passed, at best, a number of possible criteria. However, by imagining the existence of a ‘missing variable’, \( w \) say, that indicated which criterion had truly been in operation, it is possible that the EM algorithm, (Dempster et al., 1977), could be used to correct for publication bias, in a way that allowed for this extra uncertainty.

In reality, the set of rules that govern publication are unknown. Due to its inevitability, I investigated the consequences that selection model misspecification would have on correct estimation, and proposed a pseudo-data based Kolmogorov-Smirnov test to rank competing publication criteria, in order of likelihood (Chapter 4). The rationale behind this test was that the closer the true selection model to the proposed selection model, the closer one’s corrected estimate would be to the true answer. This idea was attractive because it demonstrated how pseudo-data could be simultaneously used for model validation and bias correction. It is certainly a novel application of the Kolmogorov-Smirnov test. However, the typical size and heterogeneous nature of meta data meant that there will very often be little or no power to distinguish one publication criterion from another. This
was certainly true for the homocysteine data. In that respect, the argument of Copas and Jackson (2004), who advocate the bounding of bias as the only practical step possible, appears strong.

This leaves the question, should we attempt to correct for bias, using the pseudo-data method or any other method, at all? I believe that methods to correct for bias do have a place, particularly when there are strong, well founded prior beliefs that a particular publication criterion has been in operation. But this will rarely be the case. Inevitably one's prior beliefs about the selection model will influence the result, and this should always be acknowledged. In summary, we should correct for publication bias, but at the same time, not expect our correction to be unbiased.

8.2 Genetic epidemiology

In Chapter 5 the subject of ascertainment bias was introduced, from its original conception and discussion by Fisher (1934) and Li and Mantel (1968), when estimating disease prevalence, to the more contemporary situation whereby genetic and environmental random effects require estimation within the GLMM framework. Although Clayton's method was proposed to deal with bias in genetic studies, it became obvious that the non-normal, correlated nature of the disease data prohibited a straight-forward application of data augmentation. The Bayesian approach of Burton et al. (2001) appeared to offer a partial solution the problem of ascertainment bias, by claiming the ability to return parameter estimates for the ascertained population, but this was the subject of some debate.
By showing, in Chapter 6, that it was a Bayesian analogy to integrating a truncated binomial likelihood, I believe I gave a clearer explanation of Burton’s MCMC approach. This offered a more concrete justification as to why ascertained population parameter estimates could be returned. It was immediately apparent that, whilst the ascertained parameters were of little interest in their own right, they could be used to re-simulate the risk profiles of the ascertained population, thus enabling Clayton’s conditional likelihood to produce fully corrected estimates. Because it has significantly clarified Burton’s work, and made an important contribution to a current and lively academic debate, I rank this as the most important achievement of the thesis.

The two step approach was demonstrated on a fairly simple example with a single random effect, previously used in the literature by Burton et al. (2001) and Glidden and Liang (2002), so as to be easily understood. The random effect was chosen to be normally distributed, since this made Clayton’s data augmentation approach very straightforward. The scientific justification for assuming a normal random effects distribution in this setting is that it models the situation where a large number of unknown genes each exert a small, additive effect on the increasing the risk of disease. However, this will certainly not be universally appropriate. It would be useful, as further work, to extend the two step approach to cover situations where the variation in disease risk is more skewed. For this reason gamma and t-distributions have been proposed by Noh et al. (2005). Lee and Thompson (2007) have also recently explored the topic of incorporating non-normal random effects into statistical models, but in the context of meta analysis. These papers would undoubtedly both prove useful resources when carrying out this further work.
At present, the combined effect of a model such as Burton (2003) that incorporates; multiple fixed and random effects, members from different generations and a complex ascertainment criterion, is beyond the two-stage approach. This is because the ascertained parameters obtained from the first stage can not be accurately estimated. However, it should be pointed out that no practical solution currently exists for models of this complexity. Further work is required to clarify what sort of models can be fitted. If a slight modification to Burton's model could be shown to greatly improve the ability to estimate the ascertained population's parameters, given that the two-stage approach is a practical solution, this could potentially influence the design of future genetic epidemiological studies.

8.3 Data augmentation

Clayton's conditional likelihood was initially chosen over the MCML approach of Geyer and Thompson (1992) because of its superior performance in simulation studies. In Chapter 7 my focus shifted towards probing the reasons for this discrepancy. Clayton's conditional likelihood was shown to naturally exhibit the 'defensive' property of Hesterberg (1995), whereas MCML did not. Although this was an interesting finding in its own right, I felt that, in some circumstances, Clayton's approach could be improved. By combining Clayton's algorithm with Geyer's RLR approach, I proposed a new approximate likelihood that encapsulated the most advantageous properties of both.

In a simple example this combined algorithm was shown to perform marginally worse that Clayton's conditional likelihood. However in a complex genetic epidemiological sce-
nario, for which pseudo-data generation was not so straightforward, it was shown to perform the best. This illustrates the fact that defensive importance sampling is not always necessary, and can in fact reduce the effectiveness of importance sampling routines, as noted by Owen and Zhou (2000).

For further work it will be interesting to investigate if there are other importance sampling routines, different to the form used in this thesis, that could be employed to model selection biased data. An excellent source for these new ideas is the work of Meng and Wong (1996), who proposed a general identity for all possible varieties of importance sampling, of which the form utilised in Clayton's method is a special case.

8.4 Conclusion

In recent years Bayesian statisticians have taken full advantage of the scope and flexibility afforded by Monte-Carlo integration, with advancements to MCMC methodology. To complement this, user-friendly software has also been developed, (Plummer, 2004; Spiegelhalter et al., 2005), which often makes the Bayesian approach the most practical and attractive option for biomedical researchers to perform their data analysis. This has certainly contributed to a steep rise in its popularity, to the extent that one could be forgiven for thinking that Monte-Carlo methods were the sole preserve of the Bayesian community.

It is my view that, in challenging situations where statistical inference is hindered by complex integrals, such as the analysis of data in the presence of selection bias, Bayesian
methods should not necessarily be the first port of call. I believe that by incorporat-
ing Monte-Carlo methods into the maximum likelihood framework, via techniques such as data augmentation, the most meritable properties of both the Bayesian and frequen-
tist philosophies can be preserved. Firstly, prior information about the likely value of a parameter, $\theta$, can be utilised by specifying parameter values, $\theta'$, for ‘pseudo’ data gener-
ation. Monte-Carlo integration then enables intractable likelihoods to be approximated and maximised with relative ease. Secondly, because Monte-Carlo integration is a partly random process, no matter how close the true value of $\theta$ is to one’s prior beliefs about its value (reflected in $\theta'$) a distribution of estimates, rather than a single value is obtained for $\hat{\theta}$. This is true even if the parameter $\theta$ represents a variance term, and in this way, uncertainty is allowed to ‘propagate’ throughout the model. However, whilst these are undeniably Bayesian characteristics, they can be achieved without having to specify a prior distribution for $\theta$, and therefore inferences are ‘real’ data driven, and not based on a posterior distribution. However, a valid Bayesian counter-argument might be that it is better to explicitly quantify the effect of ‘pseudo’ data with a prior distribution.

In the future I hope to contribute to making inferential methods based on data aug-
mentation, that exploit the principles of Monte-Carlo integration, more accessible and amenable to the wider biostatistical community.
Appendix A

Additional material

A.1 Computer programs

All programs were written using the statistical package R.

A.1.1 Program code for Chapter 3

```r
realgen = function(n =100,mu1=10,mcu2=5,sdeffect=5,
siglevel=1.96,asclevel=1){

# simulate trial sizes from
# truncated exponential distribution
# no trial less than 5 subjects

mu.p = 0.5*(mu1-mu2)
number = round(rexp(10*n,0.01),0)
number = number[number > 5]
```
number = number[1:n]

# set up data vectors and
# selection indicator variables

diff = NULL
vardiff = NULL
p = NULL
std.error = NULL
is.pvalue = NULL
is.effect = NULL
score = NULL

for(i in 1:n){
    is.pvalue[i] = 0
    score[i] = is.pvalue[i]

    # subject data to
    # publication criteria

    while(score[i] < asclevel){
        a = rnorm(number[i],mu1,sdeffect)
        b = rnorm(number[i],mu2,sdeffect)
        mdiff[i] = mean(a) - mean(b)
        vardiff[i] = var(a)+var(b)
        std.error[i] = sqrt((var(a)+var(b))/(number[i]))
        p[i] = mdiff[i]/std.error[i]

        if(p[i] > siglevel){is.pvalue[i]=1}
        if(p[i] < siglevel){is.pvalue[i]=0}

        score[i] = is.pvalue[i]
    }
}

return(list(mdiff=mdiff, p=p, std.error=std.error, sdeffect=sdeffect,
            asclevel=asclevel, siglevel=siglevel,
            is.pvalue=is.pvalue, vardiff=vardiff,
            number=number, n=n, mu1=mu1, mu2=mu2))
pseudogen = function(Real=real,m=10,thetadash=theta.p){
  # source characteristics
  n = Real$n; number = Real$number; std.error = Real$std.error
  asclevel = Real$asclevel; mu.p = Real$mu.p
  siglevel = Real$siglevel; sigdiff = Real$sigdiff

  # set up data vectors and # selection indicator variables
  mdiff = matrix(nrow=n,ncol=m)
  p = matrix(nrow=n,ncol=m)
  is.pvalue = matrix(nrow=n,ncol=m)
  is.effect = matrix(nrow=n,ncol=m)
  score = NULL

  for(i in 1:n){
    for(j in 1:m){
      is.pvalue[i,j] = 0
      is.effect[i,j] = 0
      score = is.pvalue[i,j] + is.effect[i,j]

      # subject pseudo data to # same criterion as real data
      while(score < asclevel){
        mdiff[i,j] = rnorm(1,thetadash,std.error[i])
        p[i,j] = mdiff[i,j]/std.error[i]

        if(mdiff[i,j] > sigdiff){is.effect[i,j]=1}
        if(mdiff[i,j] < sigdiff){is.effect[i,j]=0}

        if(p[i,j] > siglevel){is.pvalue[i,j]=1}
        if(p[i,j] < siglevel){is.pvalue[i,j]=0}

        score = is.effect[i,j] + is.pvalue[i,j]
      }
    }
  }
}
```r
cond.lik <- function(z,Pseudo=pseudo$mdiff,Real=real$mdiff,
             MU=theta.p,Number = real$number,
             Std.error=real$std.error) {

  # set up data matrices and
  # parameter vector

  MUreal   = z[1]
  dset     = cbind(Real,Pseudo)
  n        = (dim(dset)[1])
  M        = (dim(dset)[2])
  W        = matrix(nrow = n,ncol=M)
  L        = NULL

  for(i in 1:n){

    # calculate likelihood for the
    # ith set of m+1 observations

    for(j in 1:M){

      A     = max(dnorm((dset[i,j]),MUreal,Std.error[i]),0.00000000000000000001)
      B     = max(dnorm((dset[i,j]),MU,Std.error[i]),0.00000000000000000001)
      W[i,j] = A/B
    }
  }
}
```

A.1. Computer programs

L[i] = W[i,1]/sum(W[i,])
#
# sum likelihood over n

logL = log(L)
l = - sum(logL)
#

A.1.2 Program code for Chapter 6

Simulate 'real' family data

familysim = function(ALPHA=-5,family.no=2000,sigmasq=4.5,nsibs=5){

  famsize = nsibs
  ETA = matrix(nrow=family.no,ncol=famsize)
  subject = rep(seq(1,famsize),family.no)
  family = sort(rep(seq(1,family.no),famsize))
  disease = NULL
  SCORE = NULL
  count = 0

  #################################################
  # create linear predictor for binomial response #
  #################################################

  for(i in 1:family.no) {

    lb = famsize*(i-1)+1
    ub = famsize*i
    disease[lb:ub] = rep(0,famsize)
    score = sum(disease[lb:ub])

    #########################
    # ascertain if at least 1 affected #
    #########################

    while(score < 1){

        ...
A.1. Computer programs

count = count + 1

ETA[i,] = rep(rnorm(1, ALPHA,sqrt(sigmasq)),famsize)
k = as.vector(ETA[i,])
MU = exp(k)/(1+exp(k))
disease[lb:ub] = rbinom(length(MU),1,MU)
score = sum(disease[lb:ub])
SCORE[i] = score

ETA = as.vector(t(ETA))
MU = exp(ETA)/(1+exp(ETA))
data = data.frame(subject,family,ETA,MU,disease)
sibdata = data$disease

return(list(data=data,n=family.no,ALPHA=ALPHA, sigmasq=sigmasq,SCORE=SCORE,
   count=count,famsize=famsize,sibdata=sibdata))

Pseudo family data was simulated with a highly similar program.

Stage 1 maximisation

step1.lik = function(z,A=A21,W=W21,y=Y,N=5,points=p){
    alpha = z[1]
sigmasq = z[2]
L = 0
c = log(sqrt(pi))

for(i in 1:N){
    I = 0
    for(k in 1:points){
        zeta = A[k]*sqrt(2)*sqrt(sigmasq)
\[ p = \frac{\exp(\alpha+zeta)}{1+\exp(\alpha+zeta)} \]
\[ \text{lik1} = y_{2,i} \times \log(p) + (N-y_{2,i}) \times \log(1-p) - \log(1-(1-p)^N) \]
\[ \text{constant} = y_{2,i} \times \log(0.03) + (N-y_{2,i}) \times \log(0.97) - \log(1-(0.97)^N) \]
\[ LL = \text{lik1-constant} \]
\[ I = I + W[k] \times \exp(\text{LL}) \]

\[
L = L + y[1,i] \times (\log(I)-c)
\]

\[
1 = -L
\]

The points \( A \) and weights \( W \) used for the numerical integration were

<table>
<thead>
<tr>
<th>A</th>
<th>W</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-5.5503519 3.720365e-14</td>
</tr>
<tr>
<td>2</td>
<td>-4.7739923 8.818611e-11</td>
</tr>
<tr>
<td>3</td>
<td>-4.1219955 2.571230e-08</td>
</tr>
<tr>
<td>4</td>
<td>-3.5319729 2.171886e-06</td>
</tr>
<tr>
<td>5</td>
<td>-2.9799912 7.478399e-05</td>
</tr>
<tr>
<td>6</td>
<td>-2.4535521 1.254982e-03</td>
</tr>
<tr>
<td>7</td>
<td>-1.9449629 1.141407e-02</td>
</tr>
<tr>
<td>8</td>
<td>-1.4489343 6.017965e-02</td>
</tr>
<tr>
<td>9</td>
<td>-0.9614996 1.921203e-01</td>
</tr>
<tr>
<td>10</td>
<td>-0.4794507 3.816691e-01</td>
</tr>
<tr>
<td>11</td>
<td>0.0000000 4.790237e-01</td>
</tr>
<tr>
<td>12</td>
<td>0.4794507 3.816691e-01</td>
</tr>
<tr>
<td>13</td>
<td>0.9614996 1.921203e-01</td>
</tr>
<tr>
<td>14</td>
<td>1.4489343 6.017965e-02</td>
</tr>
<tr>
<td>15</td>
<td>1.9449629 1.141407e-02</td>
</tr>
<tr>
<td>16</td>
<td>2.4535521 1.254982e-03</td>
</tr>
<tr>
<td>17</td>
<td>2.9799912 7.478399e-05</td>
</tr>
<tr>
<td>18</td>
<td>3.5319729 2.171886e-06</td>
</tr>
<tr>
<td>19</td>
<td>4.1219955 2.571230e-08</td>
</tr>
<tr>
<td>20</td>
<td>4.7739923 8.818611e-11</td>
</tr>
<tr>
<td>21</td>
<td>5.5503519 3.720365e-14</td>
</tr>
</tbody>
</table>
The numerical integration phase performed by \texttt{step1.lik} was followed by 2nd stage conditional likelihood phase, performed by \texttt{step2.lik}.

\{
\texttt{step2.lik} = function(z)

\texttt{real.eta} = \texttt{real$data$ETA[real$data$subject==1]}
\texttt{pseudo.eta} = \texttt{pseudo$data$ETA[real$data$subject==1]}
\texttt{n} = \texttt{real$n}
\texttt{pseudo.eta} = \texttt{matrix(pseudo.eta,nrow=n*m,ncol=1,byrow=T)}
\texttt{real.eta} = \texttt{matrix(real.eta,nrow=n,ncol=1,byrow=T)}
\texttt{mudash} = \texttt{pseudo$ALPHA}
\texttt{sigmasqdash} = \texttt{pseudo$sigmasq}
\texttt{sddash} = \texttt{sqrt(sigmasqdash)}

#############################
# set up maximisation vector #
#############################

\texttt{mu} = z\[1\]
\texttt{sigmasq} = z\[2\]
\texttt{sd} = \texttt{sqrt(sigmasq)}
\texttt{L} = \texttt{NULL}

for(i in 1:n) {

#############################
# calculate weights #
#############################

\texttt{realy} = \texttt{real.eta[i,1]}
\texttt{A} = \texttt{max(dnorm(realy,mu,sd),0.000000000000000000000001)}
\texttt{Adash} = \texttt{max(dnorm(realy,mudash,sddash),0.000000000000000000000001)}
\texttt{B} = \texttt{NULL}
\texttt{Bdash} = \texttt{NULL}

for(j in 1:m)

\texttt{k} = m*(i-1)+j
\texttt{pseudoy} = \texttt{pseudo.eta[k,1]}
\texttt{B[j]} = \texttt{max(dnorm(pseudoy,mu,sd),0.000000000000000000000001)}
\texttt{Bdash[j]} = \texttt{max(dnorm(pseudoy,mudash,sddash),0.000000000000000000000001)}
Winbugs code to implement an MCMC algorithm for ascertained population parameter estimation.

```plaintext
model{
  for(i in 1(NUMFAMS)){
    ones[i] <- 1;
    x1[i] ~ dnorm(0,tau);
    x2[i] ~ dnorm(0,tauP);
    x3[i] ~ dnorm(0,tauS);
    x4[i] ~ dnorm(0,tauS);
    x5[i] ~ dnorm(0,tauS);
    x6[i] ~ dnorm(0,tauP);
    x7[i] ~ dnorm(0,tauP);
    x8[i] ~ dnorm(0,tauP);
    k1 <- 5*(i-1)+1
    k2 <- 5*(i-1)+2
    k3 <- 5*(i-1)+3
    k4 <- 5*(i-1)+4
    k5 <- 5*(i-1)+5
    y1[i] <- x1[i] + x2[i];
  }
}
```
y2[i] <- x1[i] - x2[i] ;
y3[i] <- x1[i] + x3[i] + x4[i] + x6[i] ;
y4[i] <- x1[i] - x3[i] + x5[i] + x7[i] ;
y5[i] <- x1[i] - x4[i] - x5[i] + x8[i] ;

logit(pmean[k1])  <- alphaP + bb*b[k1] + bq*q[k1] + y1[i];
logit(pmean[k2])  <- alphaP + bb*b[k2] + bq*q[k2] + y2[i];
logit(pmean[k3])  <- alphaS + bb*b[k3] + bq*q[k3] + y3[i];
logit(pmean[k4])  <- alphaS + bb*b[k4] + bq*q[k4] + y4[i];
logit(pmean[k5])  <- alphaS + bb*b[k5] + bq*q[k5] + y5[i];

prob.temp1[i] <- 1-(1-pmean[k3])*(1-pmean[k4])*(1-pmean[k5])
prob.temp2[i] <- min(prob.temp1[i],0.9999)

prob.ascertained[i] <- max(prob.temp2[i],0.0001)

dist[i] <- pow(pmean[k1],pheno[k1])*pow((1-pmean[k1]),(1-pheno[k1]))* 
pow(pmean[k2],pheno[k2])*pow((1-pmean[k2]),(1-pheno[k2]))* 
pow(pmean[k3],pheno[k3])*pow((1-pmean[k3]),(1-pheno[k3]))* 
pow(pmean[k4],pheno[k4])*pow((1-pmean[k4]),(1-pheno[k4]))* 
pow(pmean[k5],pheno[k5])*pow((1-pmean[k5]),(1-pheno[k5]))* 
(1/prob.ascertained[i])

p[i] <- dist[i]/1.0E4

ones[i] ~ dbern(p[i])

bb ~ dnorm(0,0.000001)
bq ~ dnorm(0,0.000001)
alphaP ~ dnorm(0,0.000001)
alphaS ~ dnorm(0,0.000001)
sigma ~ dgamma(0.000001,0.000001)
sigmaP ~ dgamma(0.000001,0.000001)
sigmaS ~ dgamma(0.000001,0.000001)

tau <- 1/(sigma*sigma)
tauP <- 1/(sigmaP*sigmaP)
tauS <- 1/(sigmaS*sigmaS)

# output stats

main[1] <- alphaP
main[2] <- alphaS
main[3] <- bb
main[4] <- bq
A.1.3 Program code for Chapter 7

New pseudo data generation program, that simulates from multiple distributions.

```r
pseudogen = function(real, m=100, thetadash=c(2.5, 3, 3.5)){
  n = length(thetadash)
  sd = real$sdeffect
  mdiff = matrix(nrow=m, ncol=n)
  score = NULL
  sigdiff = real$sigdiff

  for(i in 1:m){
    for(j in 1:n){
      score = -100
      while(score < sigdiff){
        mdiff[i,j] = rnorm(1, thetadash[j], sd)
        score = mdiff[i,j]
      }
    }
  }

  return(list(mdiff=mdiff, thetadash=thetadash, sd=sd))
}
```
Programs for Geyer’s 2 step RLR

# Stage 1: estimate normalising constants#

\[
gl.lik <- \text{function(a)}\{
\]

\[
Pseudo = \text{pseudo$mdiff}
\]

\[
\text{thetadash} = \text{pseudo$thetadash}
\]

\[
lb = 0.000000000000000001
\]

\[
sd = \text{pseudo$sd}
\]

\[
n = \text{length(thetadash)}
\]

# set up parameters

\[
z = a[1:(n-1)]
\]

\[
dset = \text{Pseudo}
\]

\[
M = (\text{dim(dset)}[1])
\]

\[
n = (\text{dim(dset)}[2])
\]

\[
phi = -\log(c(z,1))
\]

\[
constant = \log(1/n)
\]

\[
eta = \exp(\phi+constant)
\]

\[
\text{numerator} = \text{matrix(nrow }=M,\text{ncol}=n)
\]

\[
\text{denominator} = \text{matrix(nrow }=M,\text{ncol}=n)
\]

\[
L = \text{matrix(nrow }=M,\text{ncol}=n)
\]

\[
\text{for}(i \text{ in } 1:M)\{
\]

\[
\text{for}(j \text{ in } 1:n)\{
\]

\[
p = \text{max(dnorm((dset[i,j]),thetadash[j],sd),lb)}
\]

\[
\text{num} = \text{dnorm((dset[i,j]),thetadash,sd)}
\]

\[
\text{numerator}[i,j] = p*\text{eta}[j]
\]

\[
\text{denominator}[i,j] = \text{num}*\text{eta}
\]

\[
L[i,j] = \text{numerator}[i,j]/\text{denominator}[i,j]
\]

\}

\}

\[
l = -\text{sum(log(L)}
\]
A.1. Computer programs

# Stage 2: plug in normalising
# constant estimators to approximate full likelihood and estimate theta

\[
g2.1ik \leftarrow \text{function}(z)\{ \\
\text{Pseudo} = \text{pseudo}$\text{mdiff} \\
\text{thetadash} = \text{pseudo}$\text{thetadash} \\
\text{lb} = 0.00000000000000000001 \\
\text{sd} = \text{pseudo}$\text{sd} \\
\text{n} = \text{length(thetadash)} \\
\text{Real} = \text{real}$\text{mdiff} \\
\phi = c(\text{results}$\text{estimate}[1:(n-1)],1) \\
\text{nreal} = \text{real}$n \\
\}
\]

# set up parameters

\[
\text{theta} = z[1] \\
\text{dset} = \text{Pseudo} \\
\text{M} = (\text{dim}(\text{dset})[1]) \\
\text{n} = (\text{dim}(\text{dset})[2]) \\
\text{L1} = \text{NULL} \\
\text{L2} = \text{matrix(nrow=M,ncol=n)} \\
\]

# calculate crude real data likelihood

\[
\text{for}(f \text{ in } 1:\text{nreal})\{ \\
\text{hytheta} = \text{max(}dnorm((\text{Real}[f]),\text{theta},\text{sd}),\text{lb}) \\
\text{hymix} = \text{dnorm((Real[f]),thetadash,sd)/phi} \\
\text{L1}[f] = \text{hytheta/sum(hymix)} \\
\}
\]

\[
\text{for}(i \text{ in } 1:M)\{ \\
\text{for}(j \text{ in } 1:n)\{ \\
\text{hztheta} = \text{max(}dnorm((\text{dset}[i,j]),\text{theta},\text{sd}),\text{lb}) \\
\text{hzmix} = \text{dnorm((dset[i,j]),thetadash,sd)/phi} \\
\text{L2}[i,j] = \text{hztheta/sum(hzmix)} \\
\}
\}
\]

\[
\text{l} = -(\text{sum(log(L1))} - \text{nreal*(log(mean(L2)))})
\]
Combined approach

```r
combined.lik <- function(a){

    Real = real$mdiff
    Pseudo = pseudo$mdiff
    lb = 0.00000000000000000001
    sd = pseudo$sd
    thetadash = pseudo$thetadash

    # set up parameters
    theta = a[1]
    dset = cbind(Real,Pseudo)
    N = (dim(dset)[1])
    M = length(thetadash)
    L = NULL

    for(i in 1:N){
        p1 = max(dnorm(dset[i,1], theta, sd), lb)
        p5 = 0
        for(j in 1:M){
            p2 = max(dnorm(dset[i,1], thetadash[j], sd), lb)
            p3 = max(dnorm(dset[i,(j+1)], theta, sd), lb)
            p4 = max(dnorm(dset[i,(j+1)], thetadash[j], sd), lb)
            p5 = p5 + (p2/p1)*(p3/p4)
        }
        L[i] = (1/(1 + p5))
    }
}
```
\[ l = -\text{sum}(\log(L)) \]
Figure A.1: *The distribution of studies in the four publication categories described in Chapter 3, for varying T.*
Figure A.2: *Publication bias corrected estimates for the between study standard deviation* $\tau$ *and their 95% confidence intervals, using pseudo data with different criteria as described in Chapter 3.*

Figure A.3: *Top: The effect of selection model misspecification on the resulting maximum likelihood estimates for* $\theta$, *when criterion* $C$ *is assumed* $\theta \approx 3$. *When criteria* $C_1$ *and* $C_2$ *are assumed, $\theta$ is estimated to be 2.4 and 3.5 respectively.*
Figure A.4: Top: Average bias corrected estimates $\hat{\theta}$ obtained using equation 4.5.4, assuming an identical simulation scenario for the real data to that of section 4.3.2 (real data simulated from a $N(\theta = 3, 1)$ distribution, conditional on being greater than 2.7). Each value is the result of 1000 independent simulations. The size of the real data set was allowed to vary from 10 to 200. Bottom: The standard deviation of the average estimates for $\hat{\theta}$, as a function of study size.
Figure A.5: 100 estimates for the general population parameter $\theta = 4.5$ obtained with the stage 1 validation step ($\hat{\theta}_2$) and without ($\hat{\theta}_1$). The mean estimate is improved from 3.93 to 4.51.

Figure A.6: The distribution of 300 ascertained population parameter estimates $\hat{\beta}_b, \hat{\beta}_q$ obtained directly from the linear predictors $\eta^*$. Their mean values were 0.73 and 0.42 respectively, a negligible change from their general population true values of 0.7 and 0.4.
A.3. The Li-Mantel estimator

The following proof has been adapted and generalised from a special case illustrated in Burton et al. (2001). Let $\mu$ be the shared probability of disease across a hypothetical general population, so that it is identical to the expected population prevalence. Our hypothetical population is made up of families with $k$ members. Let $n_0, n_1, \ldots, n_k$ be the number of families with 0, 1, 2, ..., $k$ affected members. Under a complete ascertainment criteria, all families with at least 1 affected member are recruited onto a study, so that $n_A$, the number of families in the ascertained population $= \sum_{i=1}^{k} n_i$, and $n_G$ the number of families in the general population $= n_0 + n_A$. Let $m_A$ be the number of individuals in the ascertained population $= kn_A$, and likewise $m_G = kn_G$. Finally, let $a_A$ and $a_G$ symbolise the number of affected individuals in each population. However, because complete ascertainment is assumed, $a_A = a_G$ and so $a_G$ redundant. Clearly, the number of families in the general population $n_G$ is linked to $n_A$ via

$$n_G = \frac{n_A}{1 - (1 - \mu)^k}$$  \hspace{1cm} (A.3.1)
where \( 1 - (1 - \mu)^k \) is the probability of ascertainment. Of course \( n_G \) is unknown because \( n_0 \) is not observed. Nevertheless, the expected number of affected individuals is equal to \( kn_G \mu \), and the expected number of families with 1 affected member is equal to \( kn_G \mu(1 - \mu)^{k-1} \). By rearranging A.3.1 in terms of \( n_A \), the expected number of individuals in the ascertained population is equal to \( kn_G (1 - (1 - \mu)^k) \).

Li and Mantel (1968) consider the expression

\[
\frac{E[a_A - n_1]}{E[m_A - n_1]} = \frac{kn_G \mu(1 - (1 - \mu)^{k-1})}{kn_G (1 - (1 - \mu)^k - \mu(1 - \mu)^{k-1})} \tag{A.3.2}
\]

\( kn_G \) is common to the numerator and denominator and can be removed, this is convenient because \( n_G \) is unknown. Although not obvious the denominator of A.3.2 can be simplified by noting that

\[
(1 - \mu)^k = (1 - \mu)^{k-1}(1 - \mu) = (1 - \mu)^{k-1} - \mu(1 - \mu)^{k-1}
\]

so that A.3.2 becomes

\[
\frac{kn_G \mu(1 - (1 - \mu)^{k-1})}{kn_G (1 - (1 - \mu)^k + (1 - \mu)^k - (1 - \mu)^{k-1})} = \frac{\mu(1 - (1 - \mu)^{k-1})}{(1 - (1 - \mu)^{k-1})} = \mu
\]

Li and Mantel (1968) replaces the expected values with observed values in A.3.2 to estimate
the prevalence $\hat{\mu}_L$. When $k = 2$ the Li-Mantel estimator for $\mu$ is equal to the maximum likelihood estimator $\hat{\mu}$. Thomas and Gart (1971) show that the asymptotic bias in the Li-Mantel estimator for $\mu$ is

$$
\frac{(k - 1)\mu^2(1 - \mu)^{k-1}(1 - (1 - \mu^k))}{k(\sum_{i=1}^k a_i)(1 - (1 - \mu)^{k-1})^2} + O((\sum_{i=1}^k a_i)^{-2})
$$

and point out that in many cases, the bias in the Li-Mantel estimator is smaller than that of the maximum likelihood estimator. This, coupled with its ease of calculation compared to $\hat{\mu}$, make it a very attractive option for prevalence estimation.
Appendix B

Addenda

Published works to which I contributed during my PhD.


Copies of the published works that relate solely to this thesis are attached.
Bibliography


Chan, A., Hrobjartsson, A., Haahr, M., Gotzsche, P., and Altman, D. (2004). Empir-


StataCorp (2003). *STATA Statistical Software Release 8.0*. Stata Corporation, College Station, TX.


