HARVARD UNIVERSITY

THE GRADUATE SCHOOL OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Division

Department of Statistics

Committee

have examined a thesis entitled

Controlling for an Ecological Parameter in
Meta-Analyses and Hierarchical Models

presented by Martin W. McIntosh

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature

Typed name Carl N. Morris

Signature Frederick C. Mosteller

Typed name Frederick C. Mosteller

Signature

Typed name

Date May 17, 1996

Controlling for an Ecological Parameter in Meta-analyses and Hierarchical Models

A thesis presented

by

Martin McIntosh

to

The Department of Statistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of

Statistics

Harvard University
Cambridge, Massachusetts
May, 1996

UMI Number: 9631547

UMI
300 North Zeeb Road
Ann Arbor, MI 48103

Abstract

Meta-analysis, the synthesis of quantitative results from many clinical studies, is an important and controversial method in medical research. This is especially true if only some studies suggest the treatment to be beneficial, leaving medical practitioners in a state of confusion. A meta-analysis may force consensus by using differences in treatment, patient, and study design characteristics to determine which study conditions can be expected to produce positive results. Rarely are individual patient data made available, so one is forced to use aggregate measures of patient characteristics instead (for example, average age of patients in each study). This thesis proposes using aggregate measures as explanatory variables in meta-analyses. Most importantly, we propose using a covariate measuring the aggregate health of the treated population, possibly constructed from the outcome of the control group (the "population risk"). Studies measuring population attributes are called ecological covariates.

Using ecological covariates as explanatory variables is not straight forward. Strong observed associations may be due to measurement error attenuation rather than meaningful differences in the studies. For example, a plot of trials treatment effect estimates versus the mortality rate in the control group may reveal a strong association, but this does not imply that the sicker populations benefit from a treatment differently than healthier ones. That conclusion is not excluded, however, and so his thesis is concerned with estimating any true association that may exist. For demonstration, the proposed method is shown to help resolve a recent controversy over magnesium therapy, a treatment for acute myocardial infarction. We argue that magnesium therapy is beneficial despite the results of a recent large clinical trial.

Previous attempts to control for ecological covariates have been either incorrect or inadequate. In this thesis we represent clinical trials in a hierarchical model that can be viewed as a general measurement error model and does not have any of the restrictions of previous methods. When measurement errors are normal we find that we can quantify the bias with convenient expressions, and from this we derive rules that let us assess when we can ignore the measurement error. With normal measurement error we derive method of moments, maximum likelihood, and Bayes estimates for the hierarchical model.

The most general model we derive allows treatment effect and population risk estimates to be functions of natural exponential families. We compute Bayes estimates with a Metropolised Markov Chain Monte Carlo method. We also give technical attention to finding accurate posterior approximations for our MCMC algorithm.

Contents

List of Tables

List of Figures

Acknowledgments

Each of the faculty have contributed to my education, and I thank each of them. In my first years here I learned the beauty and practicality of foundational statistical theory from Art Dempster. From Don Rubin I learned a tremendous amount working hard to figure out how he would view an applied problem, and that has always lead to a deeper understanding of all statistics. Herman Chernoff's resources appear to be without limit, and his pleasant demeanor made all department functions a joy. I am not the first to say that Alan Zaslavsky knows everything! He has been a tremendous resource for all of my work. This department has been blessed with many talented junior faculty, including Hal Stern, Jun Liu, and recently John Barnard. I hope when I become a faculty member I am as helpful as they have been.

Most importantly I thank Carl Morris and Fred Mosteller for their help, guidance, and support. Working with Fred has been a blessing. Fred is wise and his advice will stay with me forever. I was working with Fred when I discovered the topic of my first paper and this thesis. Without his support, kind advice, and patience, neither would have been completed.

I am terribly indebted to Carl Morris. If not it were not for Carl I would not be here. Carl is a true teacher. His love for statistics is infecting, and I thank him for sharing that with me. Carl's great strength is always keeping his "eye on the ball" when working on applied problems. I hope I can do the same. Anyone who knows Carl can find his influence in the text of this thesis. No matter how busy he is, Carl always finds time to talk. Over the years he has been a teacher, a role model, and a friend. I cannot imagine a better advisor.

This thesis would be incomplete without my collaborators. From the Technology Assessment Group (TAG) at the Harvard School of Public Health, I thank Fred Mosteller, Cathy Berkey, and Elliott Antman, and from the New England Medical Center (NEMC), I thank Joe Lau, Chris Schmid, Joe Cappelleri, and John Ioannidis. The inspiration for this thesis began at the TAG, and it continued at the NEMC. I cannot imagine a better group of collaborators, and I hope we continue this work together in the future.

Many students and other faculty have been great resources and made my time here an wonderful intellectual experience, especially L.J Wei and Cindy Christiansen, and students Mark Glickman, Igor Perisic, Yignian Wu, Sally Thurston, and my classmate Mike Larsen. This graduate program is strong because students "hang out" and learn from each other. Dale Rinkel, Shelley Weiner, and especially Betsey Cogswell deserve credit for making the department a nice place to be.

Several friends have made the years at Harvard a joy, including Jeanne-Anne, Jeff, Jean, Edyta, Giorgio, and Paola, and many former and current students, especially Steve, Sarah and a future star of statistics, Jennifer. I will miss all of them.

Most importantly, I thank my parents and grandparents, who have been supportive and encouraging throughout my entire life.

The work contained in this thesis was supported generously by Agency for Health Care Policy Research grants No.1R01HS07118-01, No.1R01HS07306-01, No.1R01-HS08532-01, and PORT-HS06341.

Preface

I was first introduced to this topic while working with Frederick Mosteller's Technology Assessment Group (TAG) at the Harvard School of Public Health. Our subgroup included Elliott Antman, Frederick Mosteller, Catherine Berkey, and me. I started working with TAG by arrangements made by my advisor, Carl Morris, who was supporting me financially and otherwise during this time. We were all working together as part of the larger "PORT" project (Patient Outcome Review Team) in Barbara McNeil's department of health care policy at the Harvard Medical School.

At one of our meetings Elliott presented summaries of nine clinical trials that evaluated magnesium therapy for treating acute myocardial infarction. One of most recent trials (ISIS-4), by far the largest of them, seemed to show that the treatment did not offer any benefit. Because of the size of the trial, without any further explanation of its results, magnesium therapy was likely on its way out of medical use. Elliott had the idea of plotting the treatment effect estimates against the control group risk. He has scientific theories suggesting that the highest risk patients should benefit from magnesium, and that for many reasons ISIS-4 excluded those patients. We will show in Chapter 1 a strong observed pattern between the estimates of magnesiums effect and the control group mortality rates (see Figure 1.2 on page 9). My contributions to this effort were first to recognize that the observed association may have purely statistical sources, and second to propose a crude method to account for that portion of the association. That work can be found in McIntosh (1996).

During this time a group of physicians and statisticians at the New England Medical Center, consisting of Joseph Lau, Christopher Schmid, Joseph Cappelleri, John Ioannidis, and Thomas Chalmers, was investigating similar patterns in several hundreds of meta-analyses that they have compiled. Much of the work after McIntosh (1996) was done in collaboration with this group. As far as I know, they were the first to suggest that using the control group mortality rate as a covariate, a technique they call "control rate meta-regression", should be a general procedure for all meta-analyses. They have a large body of ongoing work on this topic, much of which uses the methods contained in this manuscript (with Schmid *et al.*, 1995; Lau *et al.*, 1995, as two examples).

Through my connections with the New England Medical Center, I have also been able to work with researchers from the Cochrane Collaboration. In particular, I have collaborated with a group, lead by Les Irwig, that investigates methods for performing meta-analyses of diagnostic tests. This group showed me that the methods contained in this thesis may be applied to correct measurement error biases in their methods as well.

Associations similar to those found by Elliott and the NEMC group are commonly referred to both formally and informally in the statistical and medical literature, and many are unaware that the associations have statistical sources. Although I have pointed out this problem, I was not the first (see Senn, 1991), but I am unaware of any other work besides McIntosh (1996) that suggests methods to correct the bias.

The original work on this topic was crude in that it treated meta-analysis without covariates, and assumed outcomes had normal distributions. In practice the most common outcomes are functions of binomial distributions, and so that work is limited. This manuscript extends the types of outcomes to allow many outcome distributions, where the binomial distribution is a special case. If the biases are small enough to be ignored then simple regression methods do lead to valid inferences, and so it is advantageous to know when that case holds. So in addition to more sophisticated methods for correcting biases, Chapter 3 of this manuscript gives rules to determine when we may be ignore the biases.

Much of the work in this manuscript has analogies to measurement error models and problems of ecological inference, and so these methods have application in those areas as well. Readers who are interested in ecological regression methods may find Chapter 3 useful, and readers interested in measurement error models may find Chapter 4 and Chapter 5 useful. Readers interested primarily in meta-analyses and not the technical details may wish to read Chapter 1, Chapter 2 and the examples sections of Chapter 4 and Chapter 5.

Overall this manuscript may be the first systematic study of using the control group outcome as a covariate. Although the manuscript begins by treating the normal model, it always works toward deriving procedures that may be used to evaluate data, like the magnesium data, that contains trials too small for the normal distribution assumption to be valid. While progressing through the text the reader will be helped by keeping that development program in mind.

Chapter 1

Introduction

The term 'meta-analysis' refers to the systematic method of summarizing, or synthesizing, the quantitative evidence of several independent studies. Each of k studies provides a point estimate $\hat{\theta}_{yi}$ of θ_{yi}, perhaps an experimental treatment effect, and an estimated standard error $\hat{\sigma}_{yi}$, where $i = 1 \cdots k$. The appropriate synthesis method to use depends on the nature of the data. If the studies are homogeneous so that all $\theta_{yi} = \theta_0$, then the differences among $\hat{\theta}_{yi}$ are due to experimental error only, and a weighted average of $\hat{\theta}_{yi}$, with weights $1/\hat{\sigma}^2_{yi}$, gives a minimum variance estimate of θ_0. We call this the "fixed effects" estimate.

Individual experiments that do not achieve statistical significance on their own may achieve it when their data are pooled, and so conducting a meta-analysis of several small studies is an attractive alternative to conducting a single large study. For an alternative view, Peto *et al.* (1988) argue that even meta-analyses require large trials, because if a drug is determined to have benefit, it will end up being prescribed to large populations, and only large clinical trials can mimic that use.

The conclusions of studies often differ from one another by amounts greater than can be accounted for by experimental error. Thus quantification of the disagreement is another contribution of meta-analysis. When studies disagree we call them "heterogeneous". Heterogeneous studies are commonly synthesized with a random effects model (a random effects meta-analysis), a method generally recommended in a National Academy of Science report (Graver *et al.*, 1992). A random effects meta-analysis allows θ_{yi} to differ, treating the θ_{yi} as if they are sampled from a population with mean θ_0 and standard deviation τ_y. Thus random effects meta-analysis views $\hat{\theta}_{yi}$ as having two sources of variability; within study experimental error, σ^2_{yi}, and between study heterogeneity, τ^2_y.

For example, Table 1.1 records a sample of size 10 from 33 placebo controlled randomized clinical trials (RCT's) that evaluated streptokinase, a treatment of acute myocardial infarction (AMI) (the complete data are given on page 48). Each study contributes a treatment effect estimate, $\hat{\theta}_{yi}$, which we choose to be the log of the relative risk of mortality. The sum of both experimental error and heterogeneity results in the estimates ranging from .94 (harmful) to -2.565 (beneficial). In Chapter 2 we find that a fixed effects procedure estimates the mean treatment effect as $\hat{\theta}_0 = -0.231$, and a random effects procedure gives $\hat{\theta}_0 = -0.235$ and $\hat{\tau}_y = 0.132$. Although both models give similar estimates for θ_0, they do not have the same interpretation. If the fixed effects model holds then $\theta_0 < 0$ implies that each $\theta_{yi} < 0$, but if the fixed effects model holds, because the θ_{yi} vary, even if $\theta_0 < 0$, some current or future study may still have $\theta_{yi} > 0$. In Chapter 2 we find that a future RCT for streptokinase has probability near 0.06 of not being beneficial. Thus we can clearly interpret θ_0 in the fixed effects, but its interpretation is unclear in the random effects model.

Although difficult to interpret, Mosteller and Colditz (1994) view heterogeneity as informative, and an opportunity for researchers to "produce answers to new questions that cannot be addressed easily in individual studies." For example, there may be differences in the administered 'dose' or the 'sex' of the patients or differences in controlled factors (i.e., did the studies control for sex). Including these factors as covariates in a random effects model may capture information. As an example, Berkey *et al.* (1994) relate heterogeneity of the effect of a tuberculosis vaccine to the distance from each trial to the equator.

Unfortunately, meaningful covariates are often unavailable for every study included in

Table 1.1: A sample of ten clinical trials that evaluated streptokinase for treating acute myocardial infarction (sorted by magnitude of treatment effect). The columns are: trial name; treatment and control group mortality rates, $\hat{p}_t$ number $\hat{p}_c$; treatment and control group sizes, n_t and n_c; treatment effect estimate in log relative risk and its estimated standard error, $\log(\frac{\hat{p}_t}{\hat{p}_c})$ and $\hat{\sigma}_y$; log-odds of mortality in the control group, $\log(\frac{\hat{p}_c}{1-\hat{p}_c})$. The unweighted means given in the final row are computed from the complete list of 33 trials (see data appendix).

Trial	$\hat{p}_t$	$\hat{p}_c$	n_t	n_c	$\log(\frac{\hat{p}_t}{\hat{p}_c})$	$\hat{\sigma}_y$	$\log(\frac{\hat{p}_c}{1-\hat{p}_c})$
1	0.286	0.111	14	9	0.944	1.033	-2.079
3	0.241	0.179	83	84	0.300	0.304	-1.526
5	0.253	0.218	249	234	0.149	0.165	-1.278
11	0.063	0.071	859	882	-0.128	0.179	-2.565
12	0.107	0.130	5860	5852	-0.190	0.051	-1.905
15	0.092	0.120	8592	8595	-0.263	0.045	-1.995
17	0.203	0.290	123	107	-0.354	0.234	-0.897
23	0.190	0.333	21	21	-0.560	0.546	-0.693
28	0.053	0.158	19	19	-1.099	1.108	-1.674
33	0.000	0.200	29	30	^{a}-2.565	1.468	-1.327
Mean	0.114	0.166	561	558	-0.451	0.577	-1.725

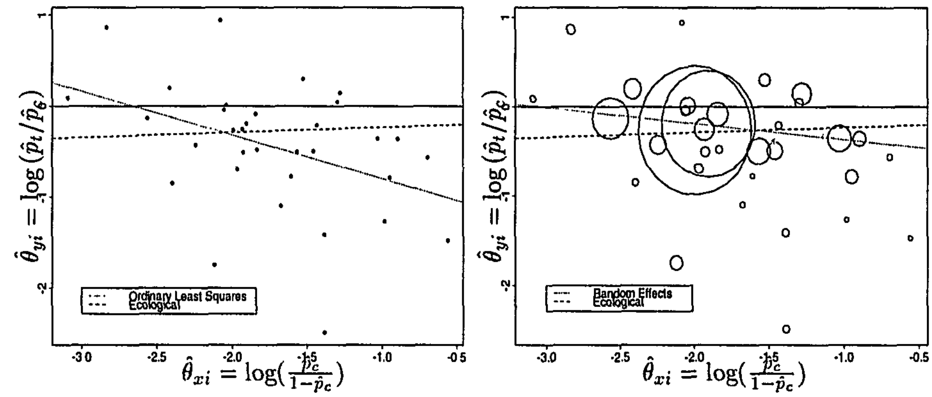
^{a}: computed assuming one death out of $n_t + 1 = 30$ patients.

a synthesis, or covariates that are available are insufficient to capture a substantial portion of the heterogeneity. The final column in Table 1.1 contains the log-odds of control group mortality for the streptokinase data, which is one possible summary of the aggregate 'health' of the population treated in each study. We refer to values that summarize the aggregate health of a population as the *population risk*. Scientific and design factors that influence treatment efficacy may also influence the population risk, and so the population risk may be a useful covariate. For example, less healthy populations may be older, have more serious medical histories, or are perhaps treated by less experienced medical staff.

Figure 1.1 plots the estimated treatment effects versus the population risk for the data in Table 1.1. Figure 1.1(a) plots the raw data and Figure 1.1(b) plots the data with point size reflecting the size of the trials, with the large trials having larger points. The solid horizontal line in each plot represents a null treatment effect; studies falling above the solid line suggest streptokinase is harmful, and points falling below this line suggest streptokinase is beneficial. Because the points above the line also tend to have low population risk, this plot seems to suggest that streptokinase does not benefit (or may even harm) healthy populations, but will benefit less healthy populations. The line in Figure 1.1(a) with label "Ordinary Least Squares" is estimated by a least squares regression with the population risk included as a covariate, and the line in Figure 1.1(b) with label "Random Effects" is estimated by a random effects regression (see Section 2.2). Both regression slopes are statistically significant, and thus appear to provide statistical evidence supporting the hypothesis that differences in population risk explains the heterogeneity. If this were true, ethical issues arise, because there may exist identifiable populations that do not benefit from streptokinase. However, in this manuscript we argue that the slightly upward sloping line with label "Ecological" better estimates the true association between treatment efficacy and population risk. Despite the apparent association, the effect of streptokinase is nearly constant across these studies.

The observed associations in Figure 1.1 can be completely explained by the attenuation of experimental or measurement error, and not necessarily due to any scientifically meaningful differences between the studies. McIntosh (1996) demonstrated this affect for general definitions of treatment effect and proposed a method to deattenuate the error and estimate any underlying association. The deattenuation method has many limitations, the most severe being the requirement of normally distributed measurement error. This assumption makes the method valid only for meta-analyses of large studies. This thesis extends the results of McIntosh (1996) to correct for this and other deficiencies.

An outline of this thesis is as follows. Chapter 2 discusses random effects meta-analysis in more detail, and introduces notation used throughout the text. We represent the usual random effects model in a 2-stage hierarchical model, with the first stage representing the experimental error (the distribution of $\hat{\theta}_{yi}$) and the second stage, the 'structural model',



(a) Streptokinase unweighted plot

(b) Streptokinase plot with points size proportional to $\sqrt{n_{ti} + n_{ci}}$

Figure 1.1: Plot of log risk ratio, $\log\left(\frac{\hat{p}_t}{\hat{p}_c}\right)$, versus the control group log odds, $\log(\frac{\hat{p}_c}{1-\hat{p}_c})$, for the streptokinase data. The solid horizontal line is the line indicating no effect of treatment across all levels of population risk. The slightly increasing line with label "Ecological" is estimated by the method recommended in this thesis. Plot labeled (a) gives a line with label "Ordinary Least Squares" that it estimated with ordinary least squares regression. Plot with label (b) gives a line with label "Random Effects" that is estimated by random effects regression.

representing the distribution of the true treatment effects (the distribution of the θ_{yi}). We briefly review estimation methods that are popular in current practice.

In Chapter 3 we extend the random effects model of Chapter 2 to incorporate the population risk, and the hierarchical representation of the previous chapter becomes bivariate, with first stage representing the experimental error of the treatment effect and population risk estimates, and the second stage representing the structural association relating their true values. We assume the experimental error has a normal distribution, and so this model is valid only for meta-analyses of large studies, but this assumption makes it convenient to demonstrate and discuss the effect of experimental error attenuation. This chapter extends

McIntosh (1996) because it allows covariates in the structural model, quantifies the measurement error attenuation to a much greater degree, and derives a means to determine if the biases are small enough to be ignored. Chapter 4 derives and demonstrates maximum likelihood, Bayesian, and method of moment procedures to estimate the model Chapter 3 constructs.

The model presented in Chapter 3 has many limitations that restrict its general use. Chapter 5 reviews the requirements of a more general procedure and proposes a model that meets those requirements. In particular, we extend the experimental error, the first stage of the hierarchical model, to include distributions that are functions of natural exponential families. The most useful of these distributions are the binomial, Poisson, and normal. We extend the structural model of Chapter 3, the second stage, to allow more complex associations between the true treatment effect and population risk than the model of Chapter 3 allows; for example, polynomial functions. We derive Bayesian estimation procedures based on a "Metropolised" data augmentation (Tanner and Wong, 1987; Gelman *et al.*, 1995), a Markov-Chain Monte-Carlo (MCMC) algorithm, and a natural extension of the estimation procedures given in Chapter 3. Thus Chapter 2, Chapter 3, Chapter 4 and Chapter 5 present a progression of models for meta-analysis, with less restrictive assumptions and more general estimation procedures.

At the end of Chapter 4 and Chapter 5 we investigate the statistical properties of the estimating procedures each proposes. In particular we evaluate the frequency properties of interval estimates of structural model parameters. We demonstrate that the procedures in both Chapter 3 and Chapter 5 adequately account for the experimental error attenuation when studies are large, and that when studies are small, the methods of Chapter 5 adjust correctly.

Before proceeding, we introduce another data set, and demonstrate that although associations like those in Figure 1.1 can be explained by measurement error, estimating a true underlying association can be valuable to help answer important scientific questions, and also to affect policy.

Example: The magnesium trials

Table 1.2 records the results of nine clinical trials evaluating intravenous magnesium for treating AMI. The treatment effect estimates (in log relative risk of mortality) range from 0.210 (harmful) to -2.249 (beneficial). Most importantly, the single largest clinical trial, ISIS 4 with over 50 thousand patients (94% of the total), suggests possible harm of magnesium therapy ($p = 0.050$). This is contrary to what was expected. Not only have most previous studies suggested that magnesium is beneficial, but there is other scientific evidence to support this hypothesis (Antman, 1995a,b, summarizes biological theory, animal experiments, and epidemiological studies). How we way the evidence from ISIS 4 is controversial. For many, the magnitude of the ISIS 4 trial overturns all previous evidence (Gupta, 1996). Once a popular treatment for AMI fewer than ten percent of clinicians now use magnesium therapy (Antman, 1995a).

In Chapter 2 we show that, because of the size of ISIS 4, if we synthesize the magnesium trials with a fixed effects model, the conclusion that magnesium has no benefit ($\hat{\theta}_0 = 0.021$ with standard error 0.030). If these results are synthesized with a random effects model a different conclusion follows ($\hat{\theta}_0 = -0.469$ with standard error 0.226, $\tau_y = 0.447$), and on average magnesium can be expected to be beneficial. However, even if the random effect model is preferred, because of the size of τ_y, perhaps 19% of treated populations may not derive any benefit from, and perhaps may be harmed by, magnesium therapy. Explanation of the heterogeneity could be very valuable.

Figure 1.2 plots the treatment effect (log relative risk) against population risk estimates (log-odds of control group mortality). As with the streptokinase data, we observe a negative association. Any negative association can be explained in part by attenuation of experimental error, but here the deattenuated line (with label "Ecological") also has a negative

Table 1.2: Data from nine clinical trials evaluating intravenous magnesium for treatment of AMI (sorted by magnitude of treatment effect). The columns are: trial name; treatment and control group mortality rates, $\hat{p}_t$ and $\hat{p}_c$; treatment and control group sizes, n_t and n_c; treatment effect estimate in log relative risk and its standard error, $\log(\frac{\hat{p}_t}{\hat{p}_c})$ and $\hat{\sigma}$; log odds of mortality in the control group, $\log(\frac{\hat{p}_c}{1-\hat{p}_c})$. The final row gives the unweighted means of the columns.

Trial	$\hat{p}_t$	$\hat{p}_c$	n_t	n_c	$\log(\frac{\hat{p}_t}{\hat{p}_c})$	$\hat{\sigma}$	$\log(\frac{\hat{p}_c}{1-\hat{p}_c})$
Feldsted	0.067	0.054	150	148	0.210	0.460	-2.862
ISIS 4	0.076	0.072	29011	29039	0.051	0.031	-2.556
Abraham	0.021	0.022	48	46	-0.043	1.399	-3.807
LIMIT 2	0.078	0.103	1150	1150	-0.271	0.134	-2.169
Morton	0.025	0.056	40	36	-0.799	1.203	-2.833
Rasmussen	0.067	0.170	135	135	-0.938	0.374	-1.583
Ceremuzynski	0.040	0.130	25	23	-1.182	1.118	-1.897
Schecter '95	0.042	0.173	96	98	-1.384	0.536	-1.561
Schechter	0.017	0.161	59	56	-2.249	1.037	-1.653
Means	0.048	0.104	3235	3233	-0.734	0.699	-2.324

slope that is statistically significant (p = 0.023). The ISIS 4 trial falls close to this line suggesting that its result can be explained by its healthier population, and that less healthy populations are likely to find beneficial from magnesium therapy.

The evidence summarized above is purely statistical, but is consistent with other evidence supporting magnesium as a beneficial treatment for some populations. Antman (1995a) discusses several differences in some of the studies for which information is available. Based on scientific considerations, the most important difference may be the duration from the onset of chest pain to administration of magnesium. Biological evidence suggests that for magnesium to be effective, it should be administered shortly a after onset of chest pain. The Rasmussen trial treated all patients in less than three hours after onset. Trials other than Rasmussen to not give information on time until treatment, but some give information on time until their patients were randomized. Because patients cannot be treated before randomization, time until randomization provides a lower bound for time until treat-

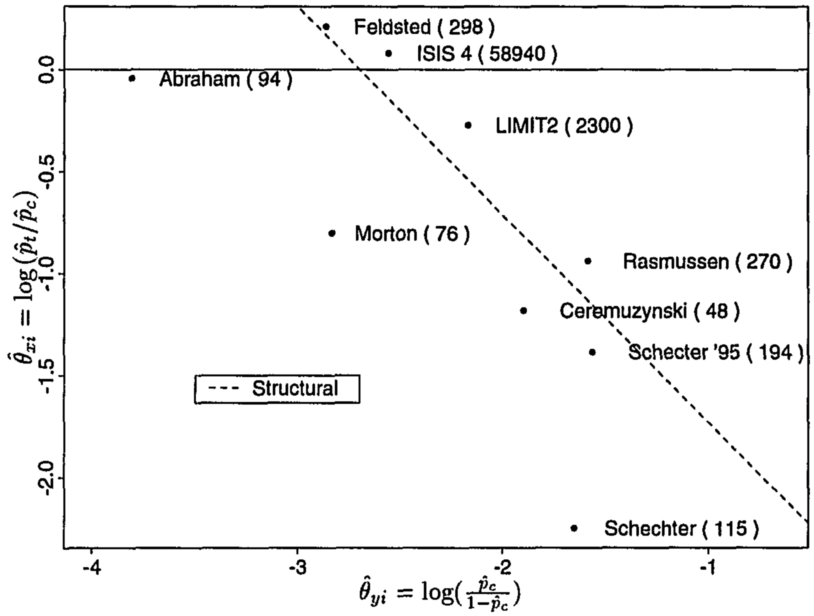


Figure 1.2: Plot of log risk ratio, $\log\left(\frac{\hat p_t}{\hat p_c}\right)$, versus the control group log odds, $\log(\frac{\hat p_c}{1-\hat p_c})$, for the magnesium data. The line with label "Ecological" is estimated by the method recommended in this thesis. The solid horizontal line is the line indicating no effect of treatment across all levels of population risk. The values in parentheses give the total number of patients in the treatment and control groups.

ment. The LIMIT 2 study randomized patients at a median of three hours after onset of chest pain, and ISIS 4 randomized patients a median of eight hours after onset (for certain high risk subgroups the median was over twelve hours!). Notice that the treatment effect estimates fall in the order suggested by these values.

If indicators of timing were available for all trials we could include them as covariates in a random effects meta-analysis. But its effect can be measured indirectly though the population risk. Because sicker patients are less likely to survive until randomization, trials with long durations until randomization will be associated with lower control group mortality rates (and thus lower population risk). Additionally, patients with "conditions associated with high risk" were excluded from ISIS 4 (ISIS-4 Collaborative Group, 1995), and this too can be measured by the population risk. Because the structural model in Figure 1.1 measures both timing and selection criteria indirectly, it may control for these differences in all trials, not just those that provide that information.

The meta-analytic method proposed in this thesis adds statistical evidence to the scientific evidence supporting a new clinical trial to evaluate magnesium therapy[1]. Others question the ethics of a new trial (for example see Gupta, 1996), calling any new trial "unethical". Despite the scientific arguments on both sides, the method can add valuable statistical evidence into important scientific and policy debates.

[1] A clinical trial to further evaluate magnesium, organized by Dr. Elliott Antman, has recently received scientific approval from the National Institute of Health

Chapter 2

Preliminaries: Notation and Random Effects Meta-analysis

Many popular methods for combining information use the normal random effects model (for example see Cochran, 1954; Gilbert *et al.*, 1988; Olkin, 1995; Morris and Normand, 1992; DerSimonian and Laird, 1986). In this chapter we review the use of random effects models for combining information, and introduce its representation in a hierarchical model. This chapter has four parts: First, we introduce notation describing all observed and estimated quantities from a clinical trial. Second, we use the random effects model to represent the process that generates the observed data. Conducting a meta-analysis then involves estimating the parameters of the random effects model, and the third part of this chapter discusses a few of the most popular estimation methods. Finally, we present examples and demonstrate the claims made in Chapter 1, that when explanation of heterogeneity is insufficient, a research synthesis does not lead to a consensus opinion.

2.1 Observed and Estimated Quantities

Commonly, a treatment effect estimate summarizes the conclusion of a clinical trial, but the treatment effect estimate is actually computed from two estimates. For example, the data in Table 1.1 and Table 1.2 give the mortality rates in the treatment group, $\hat{p}_{ti}$, and control group, $\hat{p}_{ci}$, and we compute the treatment effect by taking the natural log of their ratio; $\hat{\theta}_{yi} = \log\left(\frac{\hat{p}_{ti}}{\hat{p}_{ci}}\right)$. The mortality rates measure the true underlying treatment and control group mortality rates, p_{ti} and p_{ci}. We call p_{ti} and p_{ci} the treatment and control group estimands, and call $\hat{p}_{ci}$ and $\hat{p}_{ci}$ the treatment and control group estimates. The treatment effect estimate $\hat{\theta}_{yi}$ measures the treatment effect estimand, $\theta_{yi} = \log\left(\frac{p_{ti}}{p_{ci}}\right)$. We generalize the process of data collection, treatment effect definition and estimation as follows.

Treatment and Control Outcomes

Clinical trials estimate the mean population outcome under a treatment, μ_{ti}, and a mean population outcome under a control μ_{ci}. For example, μ_{ti} and μ_{ci} could be mortality or survival rates, mean survival time, or the mean reduction in some patient outcome such as cholesterol level. In practice we do not observe μ_{ti} and μ_{ci} but they are instead measured with error from the treatment and control estimates $\hat{\mu}_{ti}$ and $\hat{\mu}_{ci}$, respectively. If we assume individual subjects in the treatment and control groups have outcomes distributed with means μ_{ti} and μ_{ci}, and known variances $V(\mu_{ti})$ and $V(\mu_{ci})$, then $\hat{\mu}_{ti}$ and $\hat{\mu}_{ci}$ have sampling distributions given by

$$\hat{\mu}_{ti} \mid \mu_{ti} \overset{\text{ind}}{\sim} \left[\mu_{ti}, \frac{V(\mu_{ti})}{n_{ti}}\right] \qquad i = 1 \cdots k \tag{2.1}$$

$$\hat{\mu}_{ci} \mid \mu_{ci} \overset{\text{ind}}{\sim} \left[\mu_{ci}, \frac{V(\mu_{ci})}{n_{ci}}\right] \qquad i = 1 \cdots k \tag{2.2}$$

With the square bracket notation above, the first argument represents the mean and second argument represents the variance of the random quantity. Here $V(\mu)$ represents a known variance function that depends on the mean parameter μ. For example, a Bernoulli observation has mean p and variance $V(p) = p(1-p)$. We will use $\mu_i = (\mu_{ti}, \mu_{ci})'$, and $\hat{\mu}_i = (\hat{\mu}_{ti}, \hat{\mu}_{ci})'$ to represent the treatment and control estimands and estimates together. We assume that within each study that $\hat{\mu}_{ti}$ and $\hat{\mu}_{ci}$ are independent, and we also assume independence between trials. That is, we assume independence of $\hat{\mu}_i$ and $\hat{\mu}_j$, $i \neq j$.

The assumptions we make here, that of known variance function $V(\cdot)$ and independence $\hat{\mu}_{ti}$ and $\hat{\mu}_{ci}$, typically hold true when the studies are clinical trials.

Treatment Effect

Functions of μ_i define the treatment effect, and we denote it by $\theta_{yi} = \theta_y(\mu_i)$. Because we observe μ_{ti} indirectly by $\hat{\mu}_{ti}$, we also observed the treatment effect indirectly by a function of $\hat{\mu}_i$. Customarily that function is $\hat{\theta}_{yi} = \theta_y(\hat{\mu}_i)$, but that is not necessary. Although $\hat{\mu}_i$ has simple well known distribution, the distribution of $\hat{\theta}_{yi}$ may be complicated, depending on $\theta_y(\cdot)$. Customarily its mean and variance is estimated by the delta method by

$$\hat{\theta}_{yi} \mid \theta_{yi} \sim \left[\theta_{yi}, \sigma^2_{yi}\right] \qquad i = 1 \cdots k \tag{2.3}$$

where $\sigma^2_{yi} \doteq \nabla\theta'_{yi}\mathbf{V}(\mu_i, n_i)\nabla\theta_{yi}$, and $\mathbf{V}(\mu_i, n_i) = \mathrm{diag}(V(\mu_{ti})/n_{ti}, V(\mu_{ci})/n_{ci})$ denotes the diagonal variance covariance matrix of $\hat{\mu}_i$, and $\nabla\theta_{yi} = (\frac{\partial\theta_y}{\partial\mu_{ti}}, \frac{\partial\theta_y}{\partial\mu_{ci}})'$. We typically evaluate (2.3) at $\hat{\mu}_{ti}$ to form the estimated standard error $\hat{\sigma}_{yi}$.

If n_{ti} and n_{ci} are large, and $\nabla\theta_{yi}(\cdot)$ is continuous at μ_i, then (2.3) can be assumed to have an approximate normal distribution. Otherwise (2.3) leads to a poor approximation.

Population Risk and Covariates

For completeness, we define the population risk estimate and estimand here, although we will not address its use until Chapter 3. Like θ_{yi}, we also define the population risk estimand as a function of μ_i, denoted by $\theta_{xi} = \theta_x(\mu_i)$, and denote its estimate by $\hat{\theta}_{xi} = \theta_x(\hat{\mu}_i)$. Together we denote the treatment effect and population risk estimates as $\theta_i = (\theta_{yi}, \theta_{xi})'$, and its estimand as $\hat{\theta}_i = (\hat{\theta}_{yi}, \hat{\theta}_{xi})'$. We restrict $\theta(\mu_i)$ to be a bijective (one-to-one) function of μ_i so that there exists an inverse function mapping θ_i to μ_i, and denoted by $\mu_i = \mu(\theta_i)$.

Each study may also contribute a vector of study level covariates, Z_i. For example, Z_i may contain the treatment 'dose'. Table 2.1 summarizes all the observed quantities from a clinical trial and represents the data available for a typical meta-analysis (compare to Table 1.1 and Table 1.2). Table 2.2 summarizes all the notation just introduced.

Table 2.1: Representation of observed quantities from a meta-analysis

					Treatment	Population	
Trial	Outcomes		Size		Effect	Risk	Covariates
i	$\hat{\mu}_{ti}$	$\hat{\mu}_{ci}$	n_{ti}	n_{ci}	$\hat{\theta}_{yi}$	$\hat{\theta}_{xi}$	Z_i
1	$\hat{\mu}_{t1}$	$\hat{\mu}_{c1}$	n_{t1}	n_{c1}	$\hat{\theta}_{y1}$	$\hat{\theta}_{x1}$	Z_1
2	$\hat{\mu}_{t2}$	$\hat{\mu}_{c2}$	n_{t2}	n_{c2}	$\hat{\theta}_{y2}$	$\hat{\theta}_{x2}$	Z_2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	$\hat{\mu}_{tk}$	$\hat{\mu}_{ck}$	n_{tk}	n_{ck}	$\hat{\theta}_{yk}$	$\hat{\theta}_{xk}$	Z_k

Table 2.2: Summary of notation for observed and estimated quantities.

Symbol/Definition	Comment
$\mu_{ti}, \hat{\mu}_{ti}$	estimand and estimate of the treatment outcome.
$\mu_{ci}, \hat{\mu}_{ci}$	estimand and estimate of the control outcome.
$\mu_i = (\mu_{ti}, \mu_{ci})'$	implicit function of θ_i: $\mu_i = \mu(\theta_i)$
$\hat{\mu}_i = (\hat{\mu}_{ti}, \hat{\mu}_{ci})'$	implicit function of $\hat{\theta}_i$: $\hat{\mu}_i = \mu(\hat{\theta}_i)$
$\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_k)'$	also written as $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\theta})$
$\hat{\boldsymbol{\mu}} = (\hat{\mu}_1, \hat{\mu}_2, \cdots, \hat{\mu}_k)'$	also written as $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}(\hat{\boldsymbol{\theta}})$
$\theta_{yi}, \hat{\theta}_{yi}$	treatment effect estimand, estimate for trial i
$\theta_{xi}, \hat{\theta}_{xi}$	population risk estimand, estimate for trial i
$\theta_i = (\theta_{yi}, \theta_{xi})'$	implicit function of μ_i: $\theta_i = \theta(\mu_i)$
$\hat{\theta}_i = (\hat{\theta}_{yi}, \hat{\theta}_{xi})'$	implicit function of $\hat{\mu}_i$: $\hat{\theta}_i = \theta(\hat{\mu}_i)$
$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)'$	also written as $\boldsymbol{\theta}(\boldsymbol{\mu})$
$\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k)'$	also written as $\boldsymbol{\theta}(\hat{\boldsymbol{\mu}})$.
$Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{ip})'$	p-dimensional vector of study level covariates
$\mathbf{Z} = (Z_1, Z_2, \cdots, Z_k)'$	
$\eta_i = Z_i'\beta$	expected mean of θ_{yi} for trial i
$\hat{\eta}_i = Z_i'\hat{\beta}$	estimated mean estimate of θ_{yi} for trial i
$\theta_0, \hat{\theta}_0$	prior mean and estimate when no covariates are used

2.1.1 Example:

We now use the streptokinase data to demonstrate these definitions and notation.

Treatment and control outcomes

The streptokinase data summarized in Table 1.1 has control group mortality rates as estimands, and so $\mu_i \equiv p_i = (p_{ti}, p_{ci})'$, where p_{ti} represents the true treatment group death rates, and p_{ci} represents the true control group death rate. We estimate them by $\hat{p}_i = (\hat{p}_{ti}, \hat{p}_{ci})'$, which are the average mortality of the treatment and control patients. If an individual patient in a trial has a Bernoulli outcome (for example, life=0 and death=1), then $V(p) = p(1-p)$ and expressions (2.1) and (2.2) give the distribution of $\hat{p}_i$ as

$$\begin{aligned} \hat{p}_{ti} \mid p_{ti} &\overset{\text{ind}}{\sim} \left[p_{ti}, \frac{p_{ti}(1-p_{ti})}{n_{ti}}\right] & i = 1 \cdots k \\ \hat{p}_{ci} \mid p_{ci} &\overset{\text{ind}}{\sim} \left[p_{ci}, \frac{p_{ci}(1-p_{ci})}{n_{ci}}\right] & i = 1 \cdots k \end{aligned}$$

Treatment effect and population risk

We may choose one of many commonly uses measures of treatment effect, including; log relative risk, $\theta_y(p_i) = \log(p_{ti}/p_{ci})$; risk difference, $\theta_y(p_i) = p_{ti} - p_{ci}$; log-odds, $\theta_y(p_i) = \log\left(\frac{p_{ti}}{1-p_{ti}}\right) - \log\left(\frac{p_{ci}}{1-p_{ci}}\right)$. Some choices for population risk include; log-odds of mortality in the control group, $\theta_x(p_i) = \log(\frac{p_{ci}}{1-p_{ci}})$; sum of the log odds in treatment and control group; $\theta_x(p_i) = \log(\frac{p_{ci}}{1-p_{ci}}) + \log(\frac{p_{ti}}{1-p_{ti}})$; absolute risk $\theta_x(p_{ci}) = p_{ci}$.

Figure 1.1 uses log relative risk treatment effect and $\nabla\theta_y(\mu_i) = (1/p_{ti}, -1/p_{ci})'$, and (2.3) yields the following approximate sampling variance of $\hat{\theta}_{yi}$:

$$\hat{\sigma}^2_{yi} = \widehat{\text{Var}}(\hat{\theta}_{yi} \mid \theta_{yi}) \doteq \frac{1}{n_{ti}}\left(\frac{1-\hat{p}_{ti}}{\hat{p}_{ti}}\right) + \frac{1}{n_{ci}}\left(\frac{1-\hat{p}_{ci}}{\hat{p}_{ci}}\right) \tag{2.4}$$

Customarily analysis proceeds by treating $\hat{\theta}_{yi} \sim N(\theta_i, \hat{\sigma}^2_{yi})$.

2.2 Random Effects Meta-analysis: The normal model

We assume each of k independent clinical studies contributes a treatment effect estimate, $\hat{\theta}_{yi}$, its approximate standard error, $\hat{\sigma}^2_{yi}$ (perhaps estimated by (2.3)), and a vector of covariates, Z_i, $i = 1 \cdots k$. The previous section described how the observed random quantities $\hat{\theta}_{yi}$ relate to their estimands θ_{yi} *within* a clinical study, but meta-analysis concerns studying the distribution of θ_{yi} *between* the studies. We consider two competing models for describing the behavior of effects between studies: the fixed effects model and the random effects model. The fixed effects model assumes each experiment estimates the same quantity, $\theta_{y1} = \theta_{y2} = \cdots = \theta_{yk} = \theta_0$, and experimental error explains all differences among the $\hat{\theta}_{yi}$. In notation we write the fixed effects model as $\hat{\theta}_{yi} = \theta_0 + e_i$, where e_i are independent normally distributed errors with mean 0 and variance σ^2_{yi}.

The random effects model allows the true treatment effects to differ, so that both experimental error and heterogeneity account for the variability between the $\hat{\theta}_{yi}$. In notation we write $\hat{\theta}_{yi} = \theta_{yi} + e_i$, and $\theta_{yi} = \theta_0 + d_i$, where e_i are defined above and d_i are independent and identically distributed normal error terms with mean 0 and variance τ^2_y. Note that the random effects model reduces to the fixed effects model if $\tau^2_y = 0$.

These models change only slightly if we include study level covariates. The fixed effects model then has true effects θ_{yi} falling exactly on a regression line $\eta_i = Z_i'\beta$, and the random effects model has the true treatment effects varying around the regression line with variance τ^2_y. In notation, we write $E(\theta_{ti} \mid \beta, \tau^2_y) = \eta_i = Z_i'\beta$, and $\mathrm{E}((\theta_{yi} - \eta_i)^2 \mid \beta, \tau^2_y) = \tau^2_y$.

We may express the random effects model hierarchically by

$$p(\hat{\theta}_{yi} \mid \theta_{yi}, \phi) = N\left(\theta_{yi}, \sigma^2_{yi}\right) \tag{2.5}$$

$$p(\theta_{yi} \mid \phi) = N\left(\eta_i, \tau^2_y\right) \tag{2.6}$$

$$\eta_i = Z_i'\beta \tag{2.7}$$

where $\phi = (\beta', \tau^2_y)'$. Expression (2.5) describes the distribution of the treatment effect

Table 2.3: Hierarchical representation of the random effects model

Experimental Model	Between Study Model
$p(\hat{\theta}_{yi} \mid \theta_{yi}, \phi) = N\left(\theta_{yi}, \sigma^2_{yi}\right)$ $i = 1 \cdots k$ σ^2_{yi} known	$p(\hat{\theta}_{yi} \mid \phi) = N\left(\eta_i, \sigma^2_{yi} + \tau^2_y\right)$ $i = 1 \cdots k$
Structural Model	Posterior Model
$p(\theta_{yi} \mid \phi) = N\left(\eta_i, \tau^2_y\right)$ $i = 1 \cdots k$ $\eta_i = Z_i'\beta$ $\phi = (\beta', \tau_y)'$ ϕ unknown	$p(\theta_{yi} \mid \hat{\theta}_{yi}, \phi) = N\left(\eta^*_i, \tau^2_y B_i\right)$ $i = 1 \cdots k$ $\eta^*_i = \eta_i B_i + \hat{\theta}_{yi}(1 - B_i)$ $B_i = \frac{\sigma^2_{yi}}{\sigma^2_{yi} + \tau^2_y}$

estimate within each study, and (2.6), referred to as the structural model, describes the distribution of the true effects between each study.

Notice that distributions (2.5) and (2.6) together specify a joint distribution of $\hat{\theta}_{yi}$ and θ_{yi}: $p(\hat{\theta}_{yi}, \theta_{yi}) = p(\hat{\theta}_{yi} \mid \theta_{yi}, \phi)p(\theta_{yi} \mid \phi)$. We will return to this point often throughout this manuscript. The joint distribution has an equivalent specification with the marginal distribution $p(\hat{\theta}_{yi} \mid \phi)$ and the conditional distribution $p(\theta_{yi} \mid \hat{\theta}_{yi}, \phi)$ by: $p(\hat{\theta}_{yi}, \theta_{yi} | \phi) = p(\theta_{yi} | \hat{\theta}_{yi}) p(\hat{\theta}_{yi} | \phi)$. Table 2.3 summarizes these representations. Distributions (2.5) and (2.6) are given in the upper and lower cells of the left column, respectively. The upper right corner describes the marginal distribution $p(\hat{\theta}_{yi} \mid \phi)$, the between study distribution. The lower right corner describes the posterior distribution, via Bayes rule, $p(\theta_{yi} \mid \hat{\theta}_{yi}, \phi)$. Note that the variance in the upper right corner contains both the experimental and systematic components of error.

2.3 Estimation

Given the random effects model (2.5) and (2.6), a meta-analysis involves making inferences about the unknown parameters ϕ. The between study model summarized in Table 2.3 is a linear model with residual variance $\sigma_{yi}^2 + \tau_y^2$. If we treat the experimental error σ_{yi}^2 as known, then if τ_y^2 was also known, we would estimate β by a weighted linear regression with weights $w_i(\tau_y) = 1/(\sigma_{yi}^2 + \tau_y^2)$, by

$$\hat{\beta} = (\mathbf{Z}'\mathbf{D}(\tau_y)\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{D}(\tau_y)\hat{\theta}_y \tag{2.8}$$

where $w_i(\tau_y) = 1/(\sigma_{yi}^2 + \tau_y^2)$, $\mathbf{D}(\tau_y) = \mathrm{diag}(w_1(\tau_y), w_2(\tau_y), \cdots, w_k(\tau_y))$, and $\mathrm{B}_i(\tau_y) = \sigma_{yi}^2/(\sigma_{yi}^2 + \tau_y^2)$. If we have no covariates we can re express (2.8) more familiarly as

$$\hat{\theta}_0(\tau_y) = \frac{\sum_{i=1}^k w_i(\tau_y)\hat{\theta}_{yi}}{\sum_{i=1}^k w_i(\tau_y)} \tag{2.9}$$

The random effects model above has two extreme cases; $\tau_y^2 = 0$ and $\tau_y^2 = \infty$. When $\tau_y^2 = 0$ then the weights become $w_i(\tau_y) = 1/\sigma^2$. Because we assume σ_{yi}^2 are known, inferences for β then involves a standard weighted regression. When $\tau_y^2 = \infty$ then the weights become $w_i(\tau_y) = 1$, and thus inferences for β involves a simple linear regression. Between these two extremes, when $0 < \tau_y^2 < \infty$, estimation of β involves a regression with unknown weights $w_i(\tau_y)$. It should be not surprising to find much of the effort in performing a random effects meta-analysis involves estimating the heterogeneity component τ_y^2. We next summarize popular methods for estimating ϕ.

2.3.1 Cochran/Dersimonian and Laird

Perhaps the most common citation for random effects meta-analysis is DerSimonian and Laird (1986), but Cochran (1954) contains their method almost completely, so we will refer to it as the CDL (Cochran, Dersimonian and Laird) method. Their method estimates a mean treatment effect θ_0 and τ_y only, and does not allow covariates, and so they base estimation on (2.9). They first test whether the fixed effect or random effect model holds by computing the test statistic

$$Q_w = \sum_{i=1}^{k} w_i(0)(\hat{\theta}_{yi} - \hat{\theta}_0(0))^2$$

which has an approximate χ^2_{k-1} distribution when $\tau_y^2 = 0$. Thus Q_w can be used to test $H_0 : \tau_y^2 = 0$ versus $H_A : \tau_y^2 > 0$. If they do not reject the test at a given significance level, then they treat $\tau_y^2 = 0$, otherwise DerSimonian and Laird (1986) estimate τ_y^2 by

$$\hat{\tau}_y^2 = \max\left[0, \frac{Q_w - (k-1)}{\sum_{i=1}^k w_i(0) - \frac{\sum_{i=1}^k w_i^2(0)}{\sum_{i=1}^k w_i(0)}}\right] \tag{2.10}$$

Expression (2.10) is derived by setting Q_w equal to its expectation. They then estimate θ_0 by $\hat{\theta}_0(\hat{\tau}_y)$ according to (2.9).

2.3.2 Morris (1983)

Most approaches that view the random effects model estimation hierarchically are based at least conceptually on the empirical bayes procedure in Morris (1983b) (for an early example of Empirical Bayes procedures applied to meta-analysis see Gilbert *et al.*, 1988). This procedure also allows study level covariates Z_i, a p dimensional vector, in the structural model, a feature omitted in Cochran (1954) and DerSimonian and Laird (1986).

Morris estimates β and τ_y^2 with the following iterative procedure. Start with an initial guess $\hat{\tau}_y^2$ for τ_y^2. Then with weights $w_i(\hat{\tau}_y)$ iterate between estimating β by (2.8) and then update the estimate of τ_y^2 by

$$\hat{\tau}_y^2 = \frac{\sum_{i=1}^k w_i(\hat{\tau}_y)\left\{(k/(k-p))(\hat{\theta}_{yi} - Z_i'\hat{\beta})^2 - \sigma_{yi}^2\right\}}{\sum_{i=1}^k w_i(\hat{\tau}_y)} \tag{2.11}$$

These steps are repeated until the estimates converge.

2.3.3 Maximum Likelihood and Bayesian inference

We may also evaluate the random effects model by maximum likelihood or Bayesianly. If we assume (2.5) and (2.6) have normal distribution, then the between trial distribution (upper right corner of Table 2.3) is also normally distributed, and inferences for ϕ can be made from the likelihood

$$L(\phi \mid \hat{\theta}) = \prod_{i=1}^{k} \frac{1}{\sqrt{\sigma_{yi}^2 + \tau_y^2}} \exp\left\{ -\frac{1}{2} \frac{(\hat{\theta}_{yi} - Z_i'\beta)^2}{\sigma_{yi}^2 + \tau_y^2} \right\} \tag{2.12}$$

We may compute maximum likelihood estimation numerically or by the EM algorithm (Dempster *et al.*, 1977). For Bayes estimation (2.12) multiplied by a prior distribution $p(\phi)$ forms a posterior distribution. See DuMouchel and Waternaux (1992) for a discussion of issues for specifying $p(\phi)$. We may perform fully Bayes estimation with a Markov chain Monte Carlo method (MCMC) (for example Gelman *et al.*, 1995, pages 78-90). DuMouchel (1990) and Morris and Normand (1992), give approximate Bayes procedures.

2.4 Examples

Here we describe how we use the random effects model to combine information from k independent studies when the $\hat{\theta}_{yi}$ have approximate normal distribution. For outcomes having binomial distribution the normal approximation to $\hat{\theta}_{yi}$ holds only when $n_{ti}p_{ti} > 5$ and $n_{ci}p_{ci} > 5$ and also $n_{ti}(1 - p_{ti}) > 5$ and $n_{ci}(1 - p_{ci}) > 5$. Referring to the magnesium data and the complete streptokinase data (see 3.4 on page 48) we find that most of the magnesium trials and nearly half of the streptokinase data fails to meet this criteria. In the description that follows we should only consider this model valid for that subset of the trials that meets the criteria. We present the normal model here because its description conveys all the concepts we require, and because in practice it is common to find the normal model used even when this criteria fails. We will treat the nonnormal aspect of the trials as this

Table 2.4: Parameter estimates from random effects meta-analysis for streptokinase and magnesium data. Values in parentheses are the standard error estimates of $\hat{\theta}_0$, except for the Bayes estimate which is a posterior standard deviation. The columns are: 'Fixed', estimate assuming the fixed effects model; 'Morris', estimate using Morris (1983a); 'CDL', estimate from Cochran/Dersimonian and Laird method; 'Bayes', Bayes procedure estimated by MCMC simulation using priors uniform on θ_0 and τ_y.

		Model/Method			
Data	Estimate	Fixed	Morris	CDL	Bayes
Streptokinase	$\hat{\theta}_0$	-0.231 (0.046)	-0.235 (0.056)	-0.231 (0.046)	-0.241 (0.064)
	$\hat{\tau}_y$	0[a]	0.132	0.087[b]	0.162
Magnesium	$\hat{\theta}_0$	0.021 (0.030)	-0.469 (0.226)	-0.412 (0.193)	-0.523 (0.338)
	$\hat{\tau}_y$	0[a]	0.447	0.126[c]	0.675

[a]: by definition; [b]: p-value 0.190;[c]: p-value 0.001

manuscript proceeds.

Following common practice we apply these procedures to the data in Table 1.1 and Table 1.2, assuming normally the normal error approximation holds. Table 2.4 summarizes those results. For the streptokinase data, the CDL method cannot reject $H_0 : \tau_y^2 = 0$, and so that estimate equals the fixed effects estimate (in the table we computed $\hat{\theta}_0$ with $\tau_y^2 = 0$, but we computed $\hat{\tau}_y$ according to (2.10)). The random effects models give only slightly different estimates, but every estimate is statistically significant, so we may be confident that on average clinical trials for streptokinase show benefit.

Even though the random effects and fixed effects procedures give similar estimates for θ_0, we interpret the random effects models differently because they allow heterogeneity. Using $\hat{\theta}_0$ and $\hat{\tau}_y^2$, we can estimate the probability that a future trial for streptokinase, θ_y^+, will be harmful. We estimate the mean of a future trial as $\theta_i^+ = \hat{\theta}_0$, and its standard error by $\mathrm{se}(\theta_i^+) = \hat{\sigma}^+ = \sqrt{\widehat{\mathrm{se}}^2(\hat{\theta}_0) + \hat{\tau}_y^2}$. Treating $(\theta_y^+ - \hat{\theta}_0)/\hat{\sigma}^+$ as having an approximate t_{k-1}

distribution yields

$$\begin{aligned} P\left(\theta_y^+ > 0\right) &\approx P\left(\frac{\theta_y^+ - \hat{\theta}_0}{\sigma^+} > \frac{\hat{\theta}_0}{\sigma^+}\right) \\ &= P\left(t_{k-1} > \frac{-\hat{\theta}_0}{\sigma^+}\right) \end{aligned} \tag{2.13}$$

Using the Morris estimates from Table 2.4, expression (2.13) yields

$$P\left(t_{32} > \frac{0.231}{\sqrt{0.056^2 + 0.131^2}}\right) = 0.058$$

Thus, although we may be confident that $\theta_0 < 0$, because of the large heterogeneity, general use of streptokinase has significant risk.

For the magnesium data, the fixed and random effects models yield substantially different estimates for θ_0. The fixed effects estimate gives a statistically significant and positive estimate (harmful), but the random effects models show a statistically significant negative effect (benefit) of magnesium. The Cochran/Dersimonian and Laird method rejects $H_0 : \tau_y^2 = 0$ ($p < 0.001$), so the random effects model likely holds. However, even if the random effects model holds, we cannot be confident that every $\theta_{yi} < 0$. We estimate the probability that a future trial of magnesium shows harm, $P(\theta_y^+ > 0)$, according to (2.13) as $P(t_8 > \frac{-.469}{\sqrt{0.226^2+0.447^2}}) = 0.188$, a substantial risk!

2.5 Discussion

When the fixed effects model holds the standard error of $\hat{\beta}$ is smallest, and so it may be tempting to use the fixed effects model whenever possible. Perhaps this observation leads DerSimonian and Laird (1986) to suggest testing for $H_0 : \tau_y^2 = 0$ before proceeding. However, Cochran (1954) points out that testing before determining which procedure to use (fixed effects or random effects) underestimates the variance of $\hat{\theta}_0$ (and thus $\hat{\beta}$), and the conservative approach is to always use the random effects procedure. The procedure of Morris (1983b) has good small sample properties when making inferences for β (Morris, 1983b; Laird and Louis, 1989), and also has the advantage of allowing covariates.

Work by Morris (Morris, 1983b, 1995) and Everson (1995) document that with small k maximum likelihood procedures with hierarchical models tend to under estimate the variance component τ_y^2, resulting in underestimates of the standard errors of $\hat{\beta}$. With small k we prefer Bayes or approximate Bayes methods (e.g., Morris, 1983b; Morris and Normand, 1992; DuMouchel, 1990) with small k.

Meta-analyses of the streptokinase and magnesium data suggests that, although $\theta_0 < 0$ (beneficial), their general use involves significant risk. We can reduce the risk by determining which conditions are likely to produce beneficial results. One solution is to include includes important scientific and design factors as covariates, but rarely are covariates available for all studies. Plots like Figure 1.1 and Figure 1.2 suggest that the population risk estimate $\hat{\theta}_{xi}$ may offer a partial solution. The next chapter investigates $\hat{\theta}_{xi}$ as a covariate.

Chapter 3

Controlling for an Ecological Covariate in Normal Hierarchical Models

Chapter 2 presented a random effects model commonly used to synthesize the quantitative results of clinical studies, and demonstrated that their conclusions may be unsatisfactory, because even if we find a treatment is beneficial on average, with large heterogeneity the treatment may still harm some populations. A satisfactory synthesis must determine which populations or treatment conditions are associated with beneficial effects of treatment. Figure 1.1 and Figure 1.2 suggest that the population risk may be helpful. This chapter investigates using the population risk as an explanatory variable in a meta-analysis. We intend to investigate the ecological model[1] that relates θ_{yi} to θ_{xi} and Z_i. Recall that we observe θ_{yi} and θ_{xi} only indirectly by $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$. The primary result of this chapter demonstrate that the association that relates $\hat{\theta}_{yi}$ to $\hat{\theta}_{xi}$ and Z_i does not represent the association that relates θ_{yi} to $\hat{\theta}_{xi}$ and Z_i, and quantifies the bias. That result can be found in equations (3.14) and (3.15).

[1] We use *ecological* following its use in sociology and epidemiology to mean pertaining to groups or environment. For a brief history see Steward (1990)

3.1 The Model

We extend the hierarchical model from Chapter 2 to include the population risk by decomposing the joint distribution of $(\hat{\theta}_i', \theta_i')'$ into within and between components as

$$\begin{aligned} p(\hat{\theta}_i, \theta_y \mid \phi) &= p(\hat{\theta}_i \mid \theta_i, \phi) p(\theta_i \mid \phi) \\ &= \underbrace{p(\hat{\theta}_i \mid \theta_i, \phi)}_{measurement} \underbrace{\underbrace{p(\theta_{yi} \mid \theta_{xi}, \phi)}_{ecological} p(\theta_{xi} \mid \phi)}_{structural} \end{aligned}$$

For reasons we soon make apparent, the first factor on the right hand side above, the within trial distribution of $\hat{\theta}_i$, is named the measurement model. The second component, the between trial distribution of θ_i, represents the structural model, and contains within it the ecological model that we are interested in.

3.1.1 Measurement Model

We assume each of k clinical studies contributes a treatment and control outcome $\hat{\mu}_i$, and we estimate the treatment effect and population risk from it by $\hat{\theta}_i = \theta(\hat{\mu}_i)$. Analogous to Chapter 2, we derive the within group distribution of $\hat{\theta}_i$ from the distribution of $\hat{\mu}_i$. Expressions (2.1) and (2.2) show that the $\hat{\mu}_i$'s have a well know distribution: $\hat{\mu}_i$ is unbiased for μ_i and has diagonal covariance matrix $\text{Var}(\hat{\mu}_i) = \mathbf{V}(\mu_i, n_i) = \text{diagonal}(V(\mu_{ti})/n_{ti}, V(\mu_{ci})/n_{ci})$. Although the components of $\hat{\mu}_i$ are independent, the components of $\hat{\theta}_i$ are dependent because they are both functions of $\hat{\mu}_i$. Their joint distribution depends on the definitions of θ_{yi} and θ_{xi} but we estimate it with a Taylor series by

Measurement error model:

$$p(\hat{\theta}_i \mid \theta_i, \phi) = N_2\left(\begin{pmatrix} \theta_{yi} \\ \theta_{xi} \end{pmatrix}, \Sigma_i = \begin{pmatrix} \sigma^2_{yi} & \beta_{wi}\sigma^2_{xi} \\ \beta_{wi}\sigma^2_{xi} & \sigma^2_{xi} \end{pmatrix}\right) \tag{3.1}$$

with variance

$$\Sigma_i = \mathbf{J}'(\mu_i)\mathbf{V}(\mu_i)\mathbf{J}(\mu_i) \tag{3.2}$$

and where $\mathbf{J}(\cdot)$ represents the Jacobian of the transformation $\theta(\cdot)$ defined by

$$\mathbf{J}(\mu) = \frac{\partial\theta}{\partial\mu} = \begin{pmatrix} \frac{\partial\theta_y(\mu)}{\partial\mu_t} & \frac{\partial\theta_y(\mu)}{\partial\mu_c} \\ \frac{\partial\theta_x(\mu)}{\partial\mu_t} & \frac{\partial\theta_x(\mu)}{\partial\mu_c} \end{pmatrix} \tag{3.3}$$

Note the parameterization of the off diagonal elements of Σ_i; $\text{cov}\left(\hat{\theta}_{yi}, \hat{\theta}_{xi}\right) = \beta_{wi}\sigma^2_{xi}$. The parameter β_{wi} is named the within trial regression slope because within each trial, if θ_i were observed, we could predict $\hat{\theta}_{yi}$ from $\hat{\theta}_{xi}$ by

$$E(\hat{\theta}_{yi} \mid \hat{\theta}_{xi}, \theta_i) = \theta_{yi} + \beta_{wi}(\hat{\theta}_{xi} - \theta_{xi}) \tag{3.4}$$

Even though we do not observe θ_i, because we assume Σ_i are known, the β_{wi} are known.

With large n_{ti} and n_{ci} and with $\mathbf{J}(\cdot)$ continuous at μ_i, then (3.1) has approximate normal distribution. We typically evaluate Σ_i at $\mu_i = \hat{\mu}_i$. With small n_{ti} or n_{ci}, the approximation is poor, but to make the exposition clearer, throughout this chapter we assume approximation (3.1) holds.

3.1.2 Structural and Ecological Models

The structural model describes the between trial distribution of θ_{yi}. We construct it here in two pieces. We represent the association relating θ_{yi} to θ_{xi} and Z_i as

Ecological Model:

$$p(\theta_{yi} \mid \theta_{xi}, \phi) = N_1\left(\beta_0 + Z_i'\beta_z + \beta_\theta(\theta_{xi} - \gamma_0), \tau^2_{y|x}\right) \tag{3.5}$$

The ecological model given in expression (3.5) relates the population treatment effect to its population risk and covariates Z_i. The coefficient of θ_{xi}, β_θ, is named the ecological regression slope. Without loss of generality we assume Z_i has mean zero and does not include the error term. We do this so the intercepts β_0 represents the unconditional mean of θ_{yi} and θ_{xi}.

We also define a model that predicts the population risk from Z_i, by

Population Risk Model:

$$p(\theta_{xi} \mid \phi) = N_1\left(\gamma_0 + Z_i'\gamma_z, \tau^2_x\right) \tag{3.6}$$

We use $\gamma = (\gamma_0, \gamma_z)'$. Expression (3.6) relates the population risk to the covariates Z_i. We define it as a necessity of the estimation procedures in Chapter 4, but it may also be found practically useful for investigating which values of Z_i are associated with high population risk.

We may express the ecological and population risk models (3.5) and (3.6) as a bivariate distribution by

Structural Model:

$$\begin{aligned} p(\theta_i \mid \phi) &= N_2\left(\omega_0 + \omega_z Z_i, \Lambda\right) \\ &= N_2\left(\begin{pmatrix} \beta_0 + Z_i'\tilde{\beta}_z \\ \gamma_0 + Z_i'\gamma_z \end{pmatrix}, \Lambda = \begin{pmatrix} \tau^2_y & \beta_\theta\tau^2_x \\ \beta_\theta\tau^2_x & \tau^2_x \end{pmatrix}\right) \end{aligned} \tag{3.7}$$

where $\tilde{\beta}_z = \beta_z + \beta_\theta\gamma_z$ and $\tau^2_y = \tau^2_{y|x} + \beta^2_\theta\tau^2_x$. Specification (3.7) shows that the θ_i follow a multivariate multiple regression with unconditional mean $\omega_0 = (\beta_0, \gamma_0)'$. We also point out that, without controlling for θ_{xi}, θ_{yi} has a normal distribution with mean $\beta_0 + Z_i'\tilde{\beta}_z$ and variance $\tau_y^2 = \tau_{y|x}^2 + \beta_0^2\tau_x^2$. We collect all unknown parameters and denote them by $\phi = (\beta', \gamma', \tau_{y|x}, \tau_x)'$.

3.1.3 Aggregate Model

Models (3.1) and (3.7) describe the within trial distribution of $\hat{\theta}_i$, and the between trial distribution of θ_i, respectively. The between trial distribution of $\hat{\theta}_i$ is given by the marginal distribution $p(\hat{\theta}_i \mid \phi)$ and is expressed by

$$\begin{aligned} p(\hat{\theta}_i \mid \phi) &= N_2\left(\omega_0 + \omega_z Z_i, \Sigma_i + \Lambda\right)) \\ &= N_2\left(\begin{pmatrix} \beta_0 + Z_i'\tilde{\beta}_z \\ \gamma_0 + Z_i'\gamma_z \end{pmatrix}, \begin{pmatrix} \sigma_{yi}^2 + \tau_y^2 & \beta_{wi}\sigma_{xi}^2 + \beta_0\tau_x^2 \\ \beta_{wi}\sigma_{xi}^2 + \beta_0\tau_x^2 & \sigma_{xi}^2 + \tau_x^2 \end{pmatrix}\right) \end{aligned} \tag{3.8}$$

Expression (3.8) follows from the rules of conditional expectation and variance:

$$E(\hat{\theta}_i \mid \phi) = EE(\hat{\theta}_i \mid \theta_i, \phi) = \omega_0 + \omega_z Z_i$$

$$\text{Var}(\hat{\theta}_i \mid \phi) = E(\text{Var}(\hat{\theta}_i \mid \theta_i, \phi)) + \text{Var}(E(\hat{\theta}_i \mid \theta_i,)) = \Sigma_i + \Lambda$$

Table 3.1 summarizes the measurement, structural, and aggregate models, but expresses them with the parameters of the ecological model (3.5). Notice that the aggregate model has two sources of variability; the within trial measurement variance Σ_i, and the structural variance, Λ. The lower right corner of Table 3.1 gives the posterior distribution $p(\theta_i \mid \hat{\theta}_i, \phi)$, which we will find useful in Chapter 4 when we treat estimation.

Notice that marginally both $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$ follow the random effects model summarized in Chapter 2. This means that we may estimate the parameters β_0, $\tilde{\beta}_z$ and τ_y^2, and γ_0, γ_z by the methods discussed there (for example the method of Morris, 1983b). We will return to this point again in Section 3.3 and Section 3.4.

Table 3.1: Hierarchical representation of the bivariate normal random effects model.

Measurement Model (Observed)	Ecological Model (Observed)
$p(\hat{\theta}_i \mid \theta_i) = N_2(\theta_i, \Sigma_i)$ $\hat{\theta}_i = (\hat{\theta}_{yi}, \hat{\theta}_{xi})'$ $\theta_i = (\theta_{yi}, \theta_{xi})'$ $\Sigma_i = \begin{pmatrix} \sigma^2_{yi} & \beta_{wi}\sigma^2_{xi} \\ \beta_{wi}\sigma^2_{xi} & \sigma^2_{xi} \end{pmatrix}$ Σ_i known	$p(\hat{\theta}_i \mid \boldsymbol{\phi}) = N_2(\boldsymbol{\eta}_i, \Sigma_i + \Lambda)$ $\Sigma_i + \Lambda =$ $\begin{pmatrix} \sigma^2_{yi} + \tau^2_{y\mid x} + \beta^2_\theta\tau^2_x & \beta_{wi}\sigma^2_{xi} + \beta_\theta\tau^2_x \\ \beta_{wi}\sigma^2_{xi} + \beta_\theta\tau^2_x & \sigma^2_{xi} + \tau^2_x \end{pmatrix}$
Structural Model	**Posterior**
$p(\theta_i \mid \boldsymbol{\phi}) = N_2(\boldsymbol{\eta}_i, \Lambda)$ $\boldsymbol{\eta}_i = \begin{pmatrix} \beta_0 + Z_i'\tilde{\boldsymbol{\beta}}_z \\ \gamma_0 + Z_i'\gamma_z \end{pmatrix}$ $\tilde{\boldsymbol{\beta}}_z = \boldsymbol{\beta}_z + \beta_\theta\gamma_z$ $\Lambda =$ $\begin{pmatrix} \tau^2_{y\mid x} + \beta^2_\theta\tau^2_x & \tau^2_x\beta_\theta \\ \beta_\theta\tau^2_x & \tau^2_x \end{pmatrix}$ $\boldsymbol{\phi} = (\boldsymbol{\beta}, \gamma, \tau^2_y, \tau^2_x)'$ $\boldsymbol{\phi}$ unknown	$p(\theta_i \mid \hat{\theta}_i, \boldsymbol{\phi}) = N_2(\boldsymbol{\eta}^*_i, \Lambda^*_i)$ $\boldsymbol{\eta}_i{}^* = \mathbf{B}_i\eta_i + (I - \boldsymbol{B}_i)\hat{\theta}_i$ $\mathbf{B}_i = (\Sigma_i + \Lambda)^{-1}\Sigma_i$ $\Lambda^*_i = \Lambda\mathbf{B}_i$

3.2 Bias

The previous section decomposed the variability of $\hat{\theta}_i$ into its within and between components. We also defined the ecological model relating θ_{yi} to θ_{xi} and Z_i. Because we only observe θ_i through $\hat{\theta}_i$, it may be tempting to estimate the parameters of the structural model by regressing $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$ and Z_i. This section demonstrates that this leads to inconsistent estimates of the structural model.

To make this demonstration we first derive the expectation of an individual $\hat{\theta}_{yi}$ given $\hat{\theta}_{xi}$ and Z_i and show that the coefficient of Z_i is not $\boldsymbol{\beta}_z$ and the coefficient of $\hat{\theta}_{xi}$ is not β_θ. We then demonstrate that when we use weighted or ordinary least squares methods and regress $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$ and Z_i, their coefficients do not represent $\boldsymbol{\beta}_z$ and β_θ either. The result of this demonstration lead to methods that allow us to evaluate the potential for bias prior to performing any regression.

Single observation bias

Because we assume the aggregate model has a bivariate normal distribution, the expectation of $\hat{\theta}_{yi}$ is linear in $\hat{\theta}_{xi}$ and Z_i, and we express this expectation by

$$\mathrm{E}(\hat{\theta}_{yi} \mid \hat{\theta}_{xi}, \phi) = \beta_0 + Z_i'\boldsymbol{\beta}^{lm}_{z,i} + \beta^{lm}_{\theta,i}(\hat{\theta}_{xi} - \gamma_0) \tag{3.9}$$

$$\mathrm{Var}(\hat{\theta}_{yi} \mid \hat{\theta}_{xi}, \phi) = \sigma^2_{yi} + \tau^2_{y|x} + \tau^2_x(\beta^2_\theta - (\beta^{lm}_{\theta,i})^2) - (\beta^{lm})^2\sigma^2_{xi} \tag{3.10}$$

where we make the definitions

$$\beta^{lm}_{\theta,i} \equiv \beta_\theta(1 - \mathrm{B}_{xi}) + \beta_{wi}\mathrm{B}_{xi}$$

$$\boldsymbol{\beta}^{lm}_{z,i} \equiv \boldsymbol{\beta}_z + \mathrm{B}_{xi}(\beta_{wi} - \beta_\theta)\gamma_z$$

$$\mathrm{B}_{xi} \equiv \frac{\sigma^2_{xi}}{\tau^2_x + \sigma^2_{xi}}$$

We rewind the reader that in the expressions above σ^2_{xi}, σ^2_{yi} and β_{wi} are known. The superscript ^{lm} means linear model. We derive (3.9) and (3.10) below, but first make a few observations.

Expression (3.9) demonstrates that trials with similar $\hat{\theta}_{xi}$ do not, on average, have their $\hat{\theta}_{yi}$ fall on the ecological regression line. Because B_{xi} and β_{wi} are different for each observation, then $\hat{\theta}_{yi}$ follows a regression on $\hat{\theta}_{xi}$ with each observation having a different slope and residual variance. How far from the regression line each falls is governed by B_{xi}, the population risk "shrinkage factor", and β_{wi}, the within trial regression slope. Notice that $0 \leq \mathrm{B}_{xi} \leq 1$. The magnitude of B_{xi} indicates how poorly the population risk has been measured. With B_{xi} near zero then θ_{xi} is well measured, and we can expect $\hat{\theta}_{yi}$ to fall nearer the ecological regression line than if B_{xi} were large. We next demonstrate expressions (3.9) and (3.10) by applying the sweep operator on the marginal covariance of $(\hat{\theta}_{yi}, \hat{\theta}_{xi}, Z_i')$.

Marginal representation

In this chapter and the next we will find different representations of the aggregate model (3.8) useful. The representation summarized in Table 3.1 conditions on the covariate Z_i, but we also find the representation that does not condition on Z_i useful.

Collect all observed quantities together and denote them by $W_i = (\hat{\theta}_{yi}, \hat{\theta}_{xi}, Z_i')'$. We let Λ_{zz} represent the variance of Z_i, and $\boldsymbol{\mu}_z$ represents the mean of Z_i. Then W_i has distribution

$$W_i|\phi \sim \left[\mu_w = \begin{pmatrix} \beta_0 \\ \gamma_0 \\ \boldsymbol{\mu}_z \end{pmatrix}, \Psi_i = \begin{pmatrix} \Sigma_i + \Lambda + \boldsymbol{\omega}_z' \Lambda_{zz} \boldsymbol{\omega}_z & \boldsymbol{\omega}_z \Lambda_{zz} \\ \Lambda_{zz} \boldsymbol{\omega}_z' & \Lambda_{zz} \end{pmatrix} \right] \qquad (3.11)$$

We can derive expressions (3.9) and (3.10) by applying successive sweep operations to Ψ_i. By the associative property of the sweep operator,

$$\begin{aligned}
&\text{Sweep}[\hat{\theta}_x, Z](\Psi_i) \qquad (3.12)\\
&= \text{Sweep}[\hat{\theta}_x]\,(\text{Sweep}[Z]\,(\Psi_i))\\
&= \text{Sweep}[\hat{\theta}_x]\left(\begin{pmatrix} \Sigma_i + \Lambda & \omega_z\Lambda_{zz} \\ \Lambda_{zz}\omega_z & \Lambda_{zz} \end{pmatrix}\right)\\
&= \text{Sweep}[\hat{\theta}_x]\left(\begin{pmatrix} \sigma^2_{yi} + \tau^2_y & \beta_{wi}\sigma^2_{xi} + \beta_\theta\tau^2_x & \tilde{\beta}'_z \\ \beta_{wi}\sigma^2_{xi} + \beta_\theta\tau^2_x & \sigma^2_{xi} + \tau^2_x & \gamma'_z \\ -\tilde{\beta}_z & -\gamma_z & \Lambda_{zz} \end{pmatrix}\right)\\
&= \begin{pmatrix} \sigma^2_{yi} + \tau^2_y - \frac{(\beta_{wi}\sigma^2_{xi}+\beta_\theta\tau^2_x)^2}{\sigma^2_{xi}+\tau^2_x} & \frac{\beta_{wi}\sigma^2_{xi}+\beta_\theta\tau^2_x}{\sigma^2_{xi}+\tau^2_x} & \tilde{\beta}'_z - \frac{\beta_{wi}\sigma^2_{xi}+\beta_\theta\tau^2_x}{\sigma^2_{xi}+\tau^2_x}\gamma'_z \\ \frac{\beta_{wi}\sigma^2_{xi}+\beta_\theta\tau^2_x}{\sigma^2_{xi}+\tau^2_x} & \frac{1}{\sigma^2_{xi}+\tau^2_x} & \frac{\gamma'_z}{\sigma^2_{xi}+\tau^2_x} \\ \tilde{\beta}'_z - \frac{\beta_{wi}\sigma^2_{xi}+\beta_\theta\tau^2_x}{\sigma^2_{xi}+\tau^2_x}\gamma'_z & \frac{-\tilde{\beta}_z}{\sigma^2_{xi}+\tau^2_x} & \Lambda_{zz} + \frac{\gamma_z\gamma'_z}{\sigma^2_{xi}+\tau^2_x} \end{pmatrix} \qquad (3.13)
\end{aligned}$$

The notation Sweep$[\hat{\theta}_x, Z]$ means we first sweep the matrix on the pivots corresponding to components of Z, and then on $\hat{\theta}_x$. By the properties of the Sweep operator, the first column, from top to bottom, of (3.13) contains $\text{Var}(\hat{\theta}_{yi} \mid \hat{\theta}_{xi}, Z_i, \phi)$ and the coefficients of $(\hat{\theta}_{xi}, Z'_i)$ (for example see Dempster, 1969, page 62). Recognizing that $\beta^{ls}_{\theta,i} = \frac{\beta_{wi}\sigma^2_{xi}+\beta_\theta\tau^2_x}{\sigma^2_{xi}+\tau^2_x}$, the single observation bias results follows.

3.2.1 Bias from a regression estimate

Expressions (3.9) and (3.10) show that each $\hat{\theta}_{yi}$ follows a linear model in $\hat{\theta}_{xi}$ and Z_i but where the coefficients are different for each observation. It is not clear then what a regression of $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$ and Z_i estimates. We give those that result in expressions (3.14) and (3.15), as a consequence of Theorem 1 which we derive next.

Theorem 1: *Let $W_i = (\hat{\theta}_{yi}, \hat{\theta}_{xi}, Z'_i)$, $i = 1 \cdots k$, have distribution given by (3.11). Let d_{ik} be a sequence of weights with $0 \leq d_{ik} < 1$ and $\sum_{i=1}^{k} d_{ik} = 1$.*

If

(1) $max(d_{ik}) \to 0$ *as* $k \to \infty$.

(2) $\lim_{k\to\infty} \sum_{i=1}^{k} d_{ik}\Psi_i = \bar{\Psi}$.

(3) The elements of W_i *have fourth moments bounded by* $C \leq \infty$.

(4) Ψ *is positive definite.*

then a weighted least squares regression of $\hat{\theta}_{yi}$ *on* $\hat{\theta}_{xi}$ *and* Z_i *with weights* d_{ik} *consistently estimates a coefficient of* θ_{xi} *which is*

$$\beta_\theta^{ls} \equiv \beta_\theta(1 - \overline{B}_x^{ls}) + \bar{\beta}_w^{ls}\overline{B}_x^{ls} \tag{3.14}$$

and consistently estimates a coefficient of Z_i *which is*

$$\beta_z^{ls} \equiv \beta_z + \overline{B}_x^{ls}(\bar{\beta}_w^{ls} - \beta_\theta)\gamma_z \tag{3.15}$$

and consistently estimates residual variance which is

$$V^{ls} = \bar{\sigma}_y^2 + \tau_{y|x}^2 + \tau_x^2(\beta_\theta^2 - (\beta_\theta^{ls})^2) - (\beta^{ls})^2\bar{\sigma}_x^2 \tag{3.16}$$

where we have defined

$$\overline{B}_x^{ls} \equiv \frac{\bar{\sigma}_x^2}{\bar{\sigma}_x^2 + \tau_x^2} \tag{3.17}$$

$$\bar{\sigma}_x^2 \equiv \sum_{i=1}^{k} \sigma_{xi}^2 d_{ik} \tag{3.18}$$

$$\bar{\beta}_w^{ls} \equiv \frac{\sum_{i=1}^{k} \beta_{wi}\sigma_{xi}^2 d_{ik}}{\sum_{i=1}^{k} \sigma_{xi}^2 d_{ik}} \tag{3.19}$$

Proof:

Define the sample mean and covariance of W_i by

$$\begin{aligned}\overline{W}_k &= \sum_{i=1}^{k} d_{ik} W_i \\ S_{W_k} &= \sum_{i=1}^{k} d_{ik}(W_i - \overline{W}_k)(W_i - \overline{W}_k)' \\ &= \sum_{i=1}^{k} d_{ik}(W_i - \mu_w)(W_i - \mu_w)' && (3.20) \\ &- (\overline{W}_k - \mu_w)(\overline{W}_k - \mu_w)' && (3.21)\end{aligned}$$

From (3.11) each term in the summation (3.20) has expectation $d_{ik}\Psi_i$, and so by assumption (2), the expectation of this sum converges to $\bar{\Psi}$. We now show that (3.21) converges in probability to zero, and (3.20) converges in probability to $\bar{\Psi}$.

For any vector $A = (a_1, a_2, \cdots, a_k)'$ with $|a_i| < a^* < \infty$, then $A'W_i$ as variance $A'\Psi_i A$ and by assumption (3) $A'\Psi_i A \leq C^* < \infty$ for some $C^* < \infty$. Then by Chebychev's inequality and assumption (1)

$$\begin{aligned}P(|A'(\bar{W}_k - \mu_w)| > \epsilon) &\leq \frac{1}{\epsilon^2}\mathrm{Var}(A'(\bar{W}_k - \mu_w)) \\ &= \frac{1}{\epsilon^2}\sum_{i=1}^{k} d_{ik}^2 \mathrm{Var}(A'(W_i - \mu_w)) \\ &= \frac{C^* \max(d_{ik})}{\epsilon^2}\sum_{i=1}^{k} d_{ik} \\ &= C^* \max(d_{ik}) \to 0\end{aligned}$$

so (3.21) converges in probability to 0. Now let the lm element of S_{W_k} be denoted by s_k^{lm}, and denote the l-th element of W_i and μ_w by W_{il} and μ_{wl}. Then by (3) and (1)

$$\begin{aligned}
\mathrm{Var}(S_k^{lm}) &= \mathrm{Var}\left(\sum_{i=1}^{k} d_{ik}(W_{il}-\mu_{wl})(W_{im}-\mu_{wm})'\right) \\
&= \sum_{i=1}^{k} d_{ik}^2 \mathrm{Var}\left((W_{il}-\mu_{wl})(W_{im}-\mu_{wm})\right)' \\
&= C\max(d_{ik})\sum_{i=1}^{k} d_{ik} \\
&= C\max(d_{ik}) \to 0
\end{aligned}$$

We have just established that $S_{W_k} \to \bar{\Psi}$ in probability. Now by assumption (4)

$$\begin{aligned}
\lim_{k\to\infty} \mathrm{Sweep}[\hat{\theta}_{xi}, Z](S_{W_k}) &= \mathrm{Sweep}[\hat{\theta}_{xi}, Z](\lim_{k\to\infty} S_{W_k}) \\
&= \mathrm{Sweep}[\hat{\theta}_{xi}, Z](\bar{\Psi})
\end{aligned}$$

The results (3.14) and (3.15) follows by replacing Ψ_i by $\bar{\Psi}$ in (3.12).

□

In words Theorem 1 states that a linear regression of $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$ and Z_i estimates the of $\hat{\theta}_{yi}$ as if it were

$$\mathrm{E}^{ls}(\hat{\theta}_{yi} \mid \hat{\theta}_{xi}, \phi) = \beta_0 + Z_i'\beta_z^{ls} + (\hat{\theta}_{xi} - \gamma_0)\beta_\theta^{ls} \tag{3.22}$$

When $d_{ik} = 1/k$, Theorem 1 quantifies the bias that results when we use ordinary least squares regression (ols) to estimate the ecological model, otherwise Theorem 1 quantifies the bias from weighted least squares regression. We will use superscripts ^{ols} and ^{wls} to denote ordinary and weighted least squares, respectively, whenever we find it useful to distinguish these cases, and ^{ls} when no distinction needs to be made.

3.2.2 Discussion of Bias

Expression (3.22) shows the least squares regression of $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$ and Z_i yields inconsistent estimates of the ecological model parameters. We refer to $\overline{B}_x^{ls}$ as the average shrinkage, but it is actually the shrinkage evaluated at $\bar{\sigma}_x^2$, the weighted average population risk variance. We also refer to $\bar{\beta}_w$ as the average measurement error slope.

Because $0 \leq \overline{\mathrm{B}}_x^{ls} \leq 1$, the coefficient of $\hat{\theta}_{xi}$ estimates a quantity between the ecological slope, β_0, and the weighted average measurement error slope, $\bar{\beta}_w^{ls}$. The total bias is $\beta_0^{ls} - \beta_0 = \overline{\mathrm{B}}_x^{ls}(\bar{\beta}_w^{ls} - \beta_0)$, or $100 \times \overline{\mathrm{B}}_x^{ls}\%$ of the difference between the average measurement slope and the ecological slope. To have a small proportion of bias $\overline{\mathrm{B}}_x^{ls}$ must be small, and this means that $\bar{\sigma}_x^2$ must be small compared to τ_x^2. A bias of $100 \times p\%$ or less requires $\bar{\sigma}_x^2/\tau_x^2 < \frac{p}{1-p}$. Because $\overline{\mathrm{B}}_x^{ls}$ depends on the sample sizes n_{ci}, Theorem 1 shows that a least squares procedure peculiarly estimates a quantity that depends on the sample size. Figure 3.1 gives a representation of these effects for the equal variance case (i.e., with equal Σ_i).

The proportion of bias from an ols is given by $\overline{\mathrm{B}}_x^{ols}$. A weighted least squares can result in more or less bias than the bias from an unweighted procedure. If observations with large σ_{xi}^2 have small d_i, then $\overline{\mathrm{B}}_x^{wls} < \overline{\mathrm{B}}_x^{ls}$, and a weighted least squares procedure has less bias. In practice this will often be the case, because both σ_{xi}^2 and σ_{yi}^2 depend on sample size, and we commonly choose d_i related to $1/\sigma_{yi}^2$.

Including $\hat{\theta}_{xi}$ in a regression also effects the coefficients of Z_i. Whether the regression underestimates or overestimates the components of $\boldsymbol{\beta}_z$ depends on the sign of the components of $\boldsymbol{\gamma}_z$, and its magnitude depends on $\overline{\mathrm{B}}_x^{ls}$.

A few special cases are:

(1) $\underline{\beta_0 \equiv 0}$: When no ecological association exists a least squares regression estimates coefficient estimates are

$$\begin{aligned} \beta_0^{ls} &= \bar{\beta}_w \overline{\mathrm{B}}_x^{ls} \\ \boldsymbol{\beta}_z^{ls} &= \boldsymbol{\beta}_z + \overline{\mathrm{B}}_x^{ls} \bar{\beta}_w \boldsymbol{\gamma}_z \end{aligned} \tag{3.23}$$

Expression (3.23) shows that an association between $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$ may be observed even

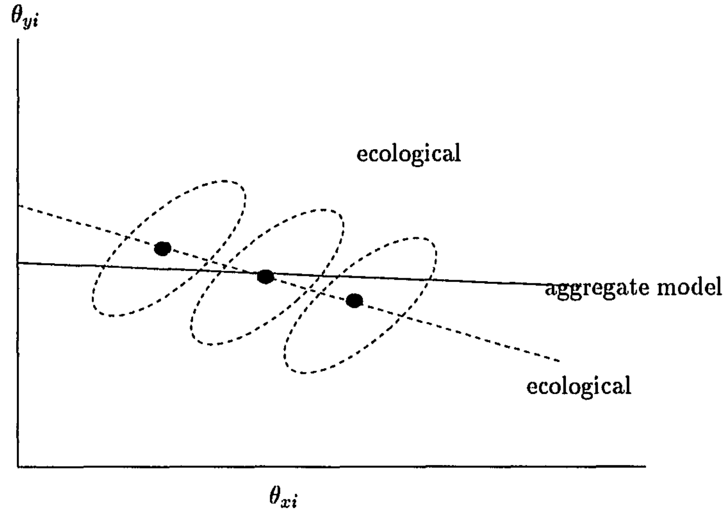


Figure 3.1: Example of bias for equal variance case: The solid points represents θ_i, and the downward sloping dotted line represents the ecological model relating θ_{yi} to θ_{xi}. Because every point falls exactly on the ecological regression line, the ecological model has $\tau_{y|x} = 0$. The ellipses surrounding the solid points represent the sampling distribution of $\hat{\theta}_i$ around their means θ_i, and the upward sloping lightly dotted lines represent the within trial regression lines. Because the ellipses have the same shape and size, this plot represents the case where all Σ_i are equal. The solid line represents the aggregate regression line. Note that its slope is between the structural slope and the measurement slope.

when no association between θ_{yi} and θ_{xi} exists. The sign of the slope depends on the sign of $\bar{\beta}_w$. This result is consistent with the claim made in Chapter 1, that despite the observed association in Figure 1.1, there may be no ecological association.

(2) $\underline{\beta_{wi} = 0}$: When $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$ have independent measurement error, then $\bar{\beta}_w = 0$, and

$$\beta_0^{ls} = \beta_0(1 - \overline{\mathrm{B}}_x^{ls}) \tag{3.24}$$

$$\beta_z^{ls} = \beta_z - \overline{\mathrm{B}}_x^{ls}\beta_0\gamma_z \tag{3.25}$$

Because $0 \leq \overline{\mathrm{B}}_x^{ls} \leq 1$, expression (3.24) shows that $| \beta_0 | > | \beta_0^{ls} |$, so we consistently underestimate structural slopes when $\bar{\beta}_w = 0$. We recognize expression (3.24) as giving the usual measurement error bias result when error is in the explanatory variable only (see for example, Miller, 1986; Fuller, 1987; Davies and Hutton, 1975). Thus the model summarized in Table 3.1 may be viewed as a general measurement error model.

(3) $\underline{\overline{\mathrm{B}}_x^{ls}(\bar{\beta}_w - \beta_0) = 0}$: Eliminating the effects of measurement error requires either $\overline{\mathrm{B}}_x^{ls} = 0$ or $\bar{\beta}_w^{ls} = \beta_0$. The former condition occurs when $\sigma_{xi}^2 = 0$, or when we have no population risk measurement error. Thus we may reduce the bias by simply acquiring more data to estimate the population risk. The latter condition, which implies the equality of the ecological slope and measurement slope, occurs only under very restrictive assumptions (see Langbein and Lichtman, 1978, for a review).

Least squares will estimate (3.16) as the residual variance. That expression can be either too large or too small, depending on many factors. We can shown that $\tau_{y|x}^2$ is underestimated whenever

$$\left(\frac{\beta_0^{ls}}{\beta_0}\right)^2 > 1 - \overline{\mathrm{B}}_x^{ls} \tag{3.26}$$

Because the right side of (3.26) is always less than 1, we estimate the residual variance too small whenever $\beta_0^{ls} > \beta_0$.

3.3 Example: Streptokinase

This section uses the streptokinase data to demonstrates the results of the previous sections using. We first derive the measurement error model with the large sample delta method approximations. We then use Theorem 1 to evaluate the bias we can expect to find with an ordinary least squares regression of $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$.

3.3.1 Estimating the measurement error model

To make the exposition clearer, we drop the index i in the derivation below. Deriving the structural model requires computing Σ, the measurement error variance. Recall that the streptokinase trials measure their outcomes as mortality rates, $\hat{p} = (\hat{p}_t, \hat{p}_c)$, and so $\hat{p}_i$ has variance matrix

$$\mathbf{V}(p,n) = \begin{pmatrix} \frac{p_t(1-p_t)}{n_t} & 0 \\ 0 & \frac{p_c(1-p_c)}{n_c} \end{pmatrix}$$

We choose log-relative risk treatment effects and log-odds of the mortality rate as the population risk. In notation we have $\theta = \theta(p) = \left(\log\left(\frac{p_t}{p_c}\right), \log\cdot\left(\frac{p_c}{1-p_c}\right)\right)'$. The measurement error variance Σ is formed by first computing the Jacobian of the transformation $\theta(\mu)$ by using (3.3), which gives

$$\mathbf{J}(p) = \begin{pmatrix} \frac{1}{p_t} & -\frac{1}{p_c} \\ 0 & \frac{1}{p_c(1-p_c)} \end{pmatrix}$$

We then estimate the measurement error variance by $\Sigma = \mathbf{J}'(\mu)\mathbf{V}(\mu,n)\mathbf{J}(\mu)$ which yields

$$\Sigma = \begin{pmatrix} \left(\frac{1-p_t}{p_t}\right)\frac{1}{nt} + \left(\frac{1-p_c}{p_c}\right)\frac{1}{n_c} & -\frac{1}{p_c n_c} \\ -\frac{1}{p_c n_c} & \frac{1}{p_c(1-p_c)}\frac{1}{n_c} \end{pmatrix} \tag{3.27}$$

Note that Σ has diagonal elements that equal the usual large sample variance approximations for log relative risk and log odds. We can now determine the within trial regression slope β_{wi}

$$\beta_w = \frac{\operatorname{cov}(\theta_y, \theta_x)}{\operatorname{var}(\theta_x)} = p_c - 1 \tag{3.28}$$

We typically estimate Σ_i and $\hat{\theta}_i$ by evaluating them at $p = \hat{p}$, but for the definition used here, we cannot compute Σ or $\hat{\theta}$ for any trial having $\hat{p}_t$ or $\hat{p}_c$ equaling zero or one. Following common practice, for every such $\hat{p}_t$, we replace $\hat{p}_t$ by $(n_t\hat{p}_t + 1/2)/(n_t + 1)$ and replace n_t

Table 3.2: Measurement error model estimates for sample of streptokinase data.

Trial	$\hat{\theta}_{yi}$	$\hat{\theta}_{xi}$	σ_{yi}	σ_{xi}	β_{wi}	B_{xi}
1	0.944	-2.079	1.033	1.061	-0.889	0.656
3	0.300	-1.526	0.304	0.285	-0.821	0.304
5	0.149	-1.278	0.165	0.158	-0.782	0.258
11	-0.128	-2.565	0.179	0.131	-0.929	0.427
12	-0.190	-1.905	0.051	0.039	-0.870	0.815
15	-0.263	-1.995	0.045	0.033	-0.880	0.199
17	-0.354	-0.897	0.234	0.213	-0.710	0.121
23	-0.560	-0.693	0.546	0.463	-0.667	0.502
28	-1.099	-1.674	1.108	0.629	-0.842	0.334
33	-2.565	-1.327	1.468	0.441	-0.790	0.560
Means	-0.451	-1.725	0.577	0.399	[a]-0.850	[b]0.55

[a] weighted average with weights proportional to $\sigma^2_{x_i}$.

[b] average bias evaluated at $\bar{\sigma}^2_x = 0.225$.

by $n_t + 1$. This is equivalent to adding a single patient to the treatment group and also adding half an observation to the number of observed deaths (for other recommendations see Mosteller and Tukey, 1977). We proceed similarly for $\hat{p}_c$.

The measurement error standard deviations and within trial regression slopes are summarized by Table 3.2. The final row of the table gives numerical averages of the columns except for the column containing β_{wi} and B_{xi} which gives $\bar{\beta}_w^{ols}$ and $\overline{\mathrm{B}}_x^{ls}$ computed by (3.18) and (3.17) with $d_i = 1$. The complete data version of this table can be found on page 48.

We have just completed estimating the parameters needed to represent the measurement error model summarized in Table 3.1. We next demonstrate using these values to assess the bias we may expect when performing a linear regression of $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$.

3.3.2 Evaluating the Bias and Estimation

We now use all $k = 33$ measurement error estimates to evaluate the bias of the regression line show in Figure 1.1. Recall that Chapter 1 claimed that measurement error can completely explain the negative association. We we demonstrate that claim quantitatively.

Theorem 1 states that the observed regression line is pulled away from β_θ toward $\bar{\beta}_w^{ols}$. Table 3.2 gives $\bar{\beta}_w^{ols} = -0.850$, and the discussion following Theorem 1 states that if $\beta_\theta = 0$, we can expect to observe $-0.850 < \hat{\beta}_\theta^{ls} < 0$. Where between these two values the expectation lies depends on $\overline{\mathrm{B}}_x^{ls}$, which we estimate next.

To estimate $\overline{\mathrm{B}}_x^{ls}$ we first estimate τ_x^2. Recall that marginally $\hat{\theta}_{xi}$ follows the random effects model summarized in Section 2.2, and so the parameters γ_z and τ_x^2 can be estimated by the methods summarized there. Here we prefer the method given by Morris (1983b) and summarized by Section 2.3.2. Morris gives a iterative procedure to estimate τ_x^2. We denote its estimate by $\hat{\tau}_x^2$. However, Morris points out that, because shrinkage factors (like $\overline{\mathrm{B}}_x^{ls}$) are convex functions of τ_x^2, using $\hat{\tau}_x^2$ to estimate $\overline{\mathrm{B}}_x^{ls}$ will lead to estimates of $\overline{\mathrm{B}}_x^{ls}$ that are too small. To correct this bias we follow his recommendation and estimate $\overline{\mathrm{B}}_x^{ls}$

$$\widehat{\overline{\mathrm{B}}}_x^{ols} = \left(\frac{k-p-1}{k-p-3}\right)\frac{\bar{\sigma}_x^2}{\bar{\sigma}_x^2 + \hat{\tau}_x^2} \tag{3.29}$$

In the expression above p represents the number of regressors Z_i (recall that here that Z_t does not include the constant term).

For the streptokinase data $\hat{\tau}_x^2 = 0.179$, $\bar{\sigma}_x^2 = 0.225$, $p = 0$, and so $\widehat{\overline{\mathrm{B}}}_x^{ols} = 0.550$, and so a least squares regression estimates the coefficient of $\hat{\theta}_{xi}$ 55% of the way from β_θ toward $\bar{\beta}_w^{ols}$. Table 3.4 gives $\bar{\beta}_w^{ols} = -0.850$, and so if $\beta_\theta = 0$ then we can expect to observe a slope $\beta^{ols} = \bar{\beta}_w^{ols}\overline{\mathrm{B}}_x^{ols} \approx (-0.850)(0.555) = -0.453$. This is very close to the observed least squares regression slope, $\hat{\beta}_\theta^{ols} = -0.482$. We have just demonstrated the claim made in Chapter 1, that the observed slope can be completely explained by measurement error.

Notice that β_θ is the only quantity in (3.16) that we have not estimated, and so we may solve for it and derive an method of moments estimator. Solving this equation for β_θ leads to a

$$\hat{\beta}_0^{mom} = \frac{\hat{\beta}_0^{ls} - \widehat{\bar{\mathrm{B}}}_x^{ols} \bar{\beta}_w}{1 - \widehat{\bar{\mathrm{B}}}_x^{ols}} \tag{3.30}$$

For the streptokinase data $\hat{\beta}^{ols} = -0.468$, and (3.30) yields $\hat{\beta}_0{}^{mom} = -0.053$. The method of moments estimator we treat in the next chapter gives (3.30) as a special case.

3.4 Discussion

The effect of using aggregates as covariates in meta-analyses in particular and hierarchical models in general is apparently not well known. Informally, Sinclair and Bracken (1994) point out that it is commonly known that the effects treatments often depend on the control group mortality rate. But recently, Lau *et al.* (1995) and Schmid *et al.* (1995) find that, although this association is very common in over four hundred meta-analyses, the association can be most often attributed to measurement error of the kind described here, and not to anything ecological, and so that informal observation is perhaps true less often than is believed. More formally, articles in statistical journals often use aggregate values as covariates without accounting for their effects (for example see Moses *et al.*, 1993; Brand and Kragt, 1992; Bryk and Raudenbush, 1992; Gelfand *et al.*, 1990). A few points are discussed below.

3.4.1 Meta-analyses

Brand and Kragt (1992) found a strong association between the treatment effect and population risk estimates in a collection of clinical trials that evaluated β-mimetics for treating pre-term labor. Commenting on their result, Senn (1991) showed that for a particular choice of $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$, some of that association could be explained by measurement error, but did not suggest a method to correct for it, stating only that, because measurement error can account for some association, "no further explanation is necessary." In response, Brand and Kragt (1991) suggest correcting the bias as follows.

First define $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$ as

$$\hat{\theta}_{yi} = \log\left(\frac{\hat{p}_{ti}}{1-\hat{p}_{ti}}\right) - \log\left(\frac{\hat{p}_{ci}}{1-\hat{p}_{ci}}\right) \tag{3.31}$$

$$\hat{\theta}_{xi} = \log\left(\frac{\hat{p}_{ti}}{1-\hat{p}_{ti}}\right) + \log\left(\frac{\hat{p}_{ci}}{1-\hat{p}_{ci}}\right) \tag{3.32}$$

then estimate the association

$$\hat{\theta}_{yi} = \beta_0 + \beta_1\hat{\theta}_{xi} \tag{3.33}$$

with a least squares regression. They argue that the bias is eliminated because specification (3.33) eliminates the correlation. Their suggestion has two difficulties: First, even if $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$ are uncorrelated, this simply implies that $\beta_{wi} = 0$, and Theorem 1 assures that estimates of β_1 are biased toward zero; Second, contrary to intuition, the specification above does not remove the correlation. The estimates $\hat{\theta}_{xi}$ and $\hat{\theta}_{yi}$ have covariance $\sigma_{xyi} = \mathrm{Var}\left(\log\left(\frac{\hat{p}_{ti}}{1-\hat{p}_{ti}}\right)\right) - \mathrm{Var}\left(\log\left(\frac{\hat{p}_{ti}}{1-\hat{p}_{ti}}\right)\right)$. Notice that $\sigma_{xyi} = 0$ only when $p_{ti} = p_{ci}$ or $p_{ti} = 1 - p_{ci}$.

Moses *et al.* (1993) use model (3.33) for meta-analysis of diagnostic tests, an important and recent application of meta-analyses (Mosteller and Colditz, 1994). In their application $\hat{p}_{ti}$ and $\hat{p}_{ci}$ represent the true-positive and false-positive rates for diagnostic tests, respectively, and they use model (3.33) to estimate a Receiver Operating Characteristic (ROC) curve. Diagnostic tests often have large differences between p_{ti} and p_{ci}, and so σ_{xyi} (and thus β_{wi}) may often be extreme for this application, leading to substantial bias (for an example diagnostic test meta-analysis with this method see Fahey *et al.*, 1995).

3.4.2 Hierarchical Models

Theorem 1 applies whenever hierarchical models use group aggregates as covariates. For example, suppose hypothetically we wish to control for the average age of the patients in the streptokinase trials. If risk factors include a patients age, then the population average age may be an attractive covariate : perhaps older patients benefit from the treatment? A hierarchical model may use age in the following model

$$p(\hat{\theta}_{yi} \mid \theta_{yi}, \phi) = N(\theta_{yi}, \sigma_{yi}^2)$$
$$p(\theta_{yi} \mid \phi) = N(\eta_i, \tau_y^2)$$

where the linear predictor is

$$\eta_i = \beta_0 + \beta^{ls}(\text{Age}_i - \overline{Age}) \tag{3.34}$$

However, if a patient's age is a risk factor then $\hat{\theta}_{yi}$ and Age_i may be correlated within each trial. Denote the within trial regression slope with β_{age} (for simplicity we will assume equal effects of age within all trials, so we do not need β_{age} to depend on i). Theorem 1 states

$$\hat{\eta}_i \approx \beta_0 + (\beta_\theta(1 - \overline{B}_{age}^{wls}) + \beta_{age}\overline{B}_{age}^{wls}) \times (\text{Age}_i - \overline{Age}) \tag{3.35}$$

Expression (3.35) shows that hierarchical models that use aggregate values as covariates lead to biased estimates of the linear predictor η_i. This may have particularly important implications for estimating the individual θ_{yi}, an important application of hierarchical models, because estimates of θ_{yi} are typically chosen to be between $\hat{\theta}_{yi}$ and $\hat{\eta}_i$.

It is commonly known that coefficients of group averages do not estimate individual effects (belief in the contrary is known as the ecological fallacy, see for example Robinson, 1995), but expression (3.35) shows that coefficients of group averages do not even estimate group effects (for an example of general hierarchical model that use aggregates see Bryk and Raudenbush, 1992, who estiamte the association between the average school math score and the schools average student economic status). To estimate group effects with aggregates the bias must be corrected or else determined to be insignificant. However, to correct the bias we must either know or have information about β_{wi}. For meta-analyses, we estimate β_{wi} from the distribution of $\hat{\mu}_i$, which we can do because clinical trials give us a known distribution of $\hat{\mu}_i$. But for other aggregates, such as average age, information about β_{wi} may be unavailable. Thus we must assess the bias.

We may assess the bias without β_{wi} if we can estimate $\overline{\mathrm{B}}_x^{ls}$. To estimate $\overline{\mathrm{B}}_x^{ls}$ we must have τ_x^2 and weights d_i. Even if we do not know β_{wi}, σ_{xi}^2 may be known, and we can estimate τ_x^2 by Morris (1983b), and use the adjustment given by (3.29). Deciding which weights to use poses the most difficult problem.

The bias results of Theorem 1 suggest that we compute $\overline{\mathrm{B}}_x^{wls}$ with $d_i \propto 1/(\sigma_{yi}^2 + \tau_{y|x}^2)$. However, recall that we cannot estimate $\hat{\tau}_{y|x}^2$ if we do not know β_{wi}. We may however, compute a lower bound for $\overline{\mathrm{B}}_x^{wls}$. Because $\tau_{y|x}^2 < \tau_y^2$, estimating $\overline{\mathrm{B}}_x^{wls}$ with $d_i \propto 1/(\sigma_{yi}^2 + \hat{\tau}_y^2)$ yields a lower bound for assessing the bias. We can estimate τ_y^2 because, as we pointed out at the end of Section 3.1, $\hat{\theta}_{yi}$ given Z_i follows the random effects model summarized in Chapter 2. If we do this and estimate $\overline{\mathrm{B}}_x^{wls}$ as small, then if we can also trust that the β_{wi} are not too large, then we may proceed with using the aggregate as a covariate. Otherwise, it should not be used.

3.4.3 Summary

This chapter constructed a hierarchical model that incorporates the treatment effect, the population risk, and covariates Z_i, and can be viewed as an extension of the random effects model summarized in Table 2.3. We demonstrated that the coefficients from a least squares regression that includes the population risk as a covariate inconsistently estimates the coefficients of the ecological model. Quantification of the bias leads to a simple method of moments correction for the bias (3.30), and also allows us to evaluate the bias even when we cannot correct for it. The next chapter considers additional methods to estimate the parameters of the structural model.

Table 3.3: Data from nine clinical trials evaluating intravenous magnesium for treatment of AMI. The columns are: the treatment and control group mortality rates, $\hat{p}_t$ and $\hat{p}_c$; the treament and control group sample sizes, n_t and n_c; the treatment effect estimate in loog relative risk, $\hat{\theta}_y = \log(\hat{p}_t/\hat{p}_c)$; a measure of the risk of mortality, $\hat{\theta}_x = \text{logit}\ (\hat{p}_c)$; the standard errors of the treatment effect and control group risk, σ_t and σ_r; the within trial regression coefficient for estimating $\hat{\theta}_y$ from $\hat{\theta}_x$, β_m. The data are sorted by $\hat{\theta}_y$. The final row gives the unweighted means of the columns.

Source	$\hat{p}_t$	$\hat{p}_c$	n_t	n_c	$\hat{\theta}_y$	$\hat{\theta}_x$	σ_y	σ_x	β_m
Feldsted	0.067	0.054	150	148	0.210	-2.862	0.460	0.364	-0.946
ISIS 4	0.078	0.072	29901	29039	0.080	-2.556	0.029	0.024	-0.931
Abraham	0.021	0.022	48	46	-0.043	-3.807	1.399	1.011	-0.978
LIMIT 2	0.078	0.103	1150	1150	-0.271	-2.169	0.134	0.097	-0.897
Morton	0.025	0.056	40	36	-0.799	-2.833	1.203	0.728	-0.944
Rasmussen	0.067	0.170	135	135	-0.938	-1.583	0.374	0.229	-0.830
eremuzynski	0.040	0.130	25	23	-1.182	-1.897	1.118	0.619	-0.870
Schecter'95	0.043	0.173	96	98	-1.384	-1.561	0.527	0.267	-0.827
Schechter	0.017	0.161	59	56	-2.249	-1.653	1.037	0.364	-0.839
Means	0.048	0.104	3511	3414	-0.730	-2.324	0.697	0.411	-0.895

Table 3.4: Complete streptokinase data with columns: treatment and control group mortality rates, $\hat{p}_t$ and $\hat{p}_c$, and sample sizes, n_t and n_c; treatment effect, $\hat{\theta}_y = \log(\hat{p}_t/\hat{p}_c)$; population risk, $\hat{\theta}_x$ = logit $(\hat{p}_c)$; standard errors of $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$, σ_t and σ_x; within trial regression coefficient, β_w. The final row gives the unweighted means of the columns.

Trial	$\hat{p}_t$	$\hat{p}_c$	n_t	n_c	$\hat{\theta}_y$	$\hat{\theta}_x$	σ_y	σ_x	β_w
^{a}1	0.286	0.111	14	9	0.944	-2.079	1.033	1.061	-0.889
2	0.132	0.056	53	54	0.866	-2.833	0.662	0.594	-0.944
3	0.241	0.179	83	84	0.300	-1.526	0.304	0.285	-0.821
4	0.100	0.082	219	207	0.201	-2.414	0.308	0.253	-0.918
5	0.253	0.218	249	234	0.149	-1.278	0.165	0.158	-0.782
^{a}6	0.048	0.043	21	23	0.091	-3.091	1.382	1.022	-0.957
7	0.224	0.214	49	42	0.047	-1.299	0.397	0.376	-0.786
8	0.116	0.115	164	157	0.010	-2.044	0.309	0.250	-0.885
9	0.109	0.113	55	53	-0.037	-2.058	0.544	0.434	-0.887
10	0.126	0.137	302	293	-0.082	-1.845	0.211	0.170	-0.863
11	0.063	0.071	859	882	-0.128	-2.565	0.179	0.131	-0.929
12	0.107	0.130	5860	5852	-0.190	-1.905	0.051	0.039	-0.870
^{a}13	0.156	0.192	32	26	-0.208	-1.435	0.575	0.498	-0.808
14	0.098	0.126	264	253	-0.250	-1.932	0.249	0.189	-0.874
15	0.092	0.120	8592	8595	-0.263	-1.995	0.045	0.033	-0.880
16	0.185	0.263	373	357	-0.353	-1.029	0.140	0.120	-0.737
17	0.203	0.290	123	107	-0.354	-0.897	0.234	0.213	-0.710
18	0.063	0.096	191	177	-0.424	-2.242	0.362	0.255	-0.904
19	0.086	0.138	35	29	-0.476	-1.833	0.721	0.539	-0.862
20	0.115	0.189	156	159	-0.492	-1.459	0.276	0.203	-0.811
21	0.105	0.173	352	376	-0.498	-1.565	0.192	0.136	-0.827
22	0.077	0.127	52	55	-0.504	-1.925	0.596	0.405	-0.873
^{a}23	0.190	0.333	21	21	-0.560	-0.693	0.546	0.463	-0.667
24	0.061	0.122	49	49	-0.693	-1.969	0.678	0.436	-0.878
^{a}25	0.077	0.167	13	12	-0.773	-1.609	1.157	0.775	-0.833
26	0.127	0.279	102	104	-0.783	-0.950	0.303	0.219	-0.721
^{a}27	0.036	0.083	28	24	-0.847	-2.398	1.193	0.739	-0.917
^{a}28	0.053	0.158	19	19	-1.099	-1.674	1.108	0.629	-0.842
^{a}29	0.077	0.273	13	11	-1.266	-0.981	1.080	0.677	-0.727
30	0.049	0.200	41	25	-1.411	-1.386	0.797	0.500	-0.800
^{a}31	0.083	0.364	12	11	-1.473	-0.560	1.037	0.627	-0.636
32	0.019	0.107	107	112	-1.746	-2.120	0.752	0.306	-0.893
^{b}33	0.000	0.200	29	30	-2.565	-1.327	1.468	0.441	-0.790
Mean	0.114	0.166	561	558	-0.451	-1.725	0.577	0.399	-0.834

^{a}: smallest nine trials, ^{b}: calculated as if $\frac{1}{2}/30$

Chapter 4

Normal Model Estimation

Chapter 3 represented the treatment effect and population risk in a normal hierarchical model, and demonstrated the difficulty when using the population risk to explain treatment heterogeneity. To use the population risk as a covariate correctly, we must account for the measurement error attenuation. This chapter proposes three methods to do that by estimating the structural parameters of the normal hierarchical model. First, we derive a simple method of moments estimate that generalizes the simple estimate given at the end of Chapter 3. In the next section we treat likelihood based inference and derive maximum likelihood and Bayesian estimation procedures. The method of moments and likelihood inference sections can be read independently. A separate section points to several issues that need to be considered when using these procedures. We then demonstrate each of the methods with the streptokinase data.

Recall that the normal approximation to the measurement error does not hold for nearly half of the streptokinase trials, and so we should not have complete confidence that the procedures derived here work well for those data. We demonstrate the effect of the small trials with a simulation study. The simulation study evaluates the performance of these procedures when the data have true normal distribution and when the data do not. The results show that when errors are normally distributed these procedures perform well, and give nearly unbiased estimates of β_θ, and truthful 90% and 95% confidence intervals. When errors are functions of binomial distributions the normal model admits significant bias, and may even increase the bias compared to using least squares methods that ignore measurement error. Thus for adequate analysis of the streptokinase and magnesium trials, we must move beyond the normal model.

4.1 Method of Moments Estimates

Expressions (3.14) and (3.15) of Theorem 1 lead immediately to method of moment estimators given by

$$\hat{\beta}_\theta^{mom} = \frac{\hat{\beta}^{ls} - \widehat{\overline{\mathrm{B}}}_x \bar{\beta}_w^{ls}}{1 - \widehat{\overline{\mathrm{B}}}_x^{ls}} \tag{4.1}$$

$$\hat{\boldsymbol{\beta}}_z^{mom} = \hat{\beta}_z - \widehat{\overline{\mathrm{B}}}_x(\bar{\beta}_w^{ls} - \hat{\beta}_\theta^{mom})\hat{\gamma}_z \tag{4.2}$$

Recall that: $\hat{\beta}^{ls}$ represents the estimated coefficient of $\hat{\theta}_{xi}$, and $\hat{\boldsymbol{\beta}}_z$ represents the estimated coefficient of Z_i that results when we regress $\hat{\theta}_{yi}$ on $\hat{\theta}_{xi}$ and Z_i; $\hat{\gamma}_z$ represents the coefficient of Z_i when we regress $\hat{\theta}_{xi}$ on Z_i; and $\widehat{\overline{\mathrm{B}}}_x$ is (3.17) evaluated at $\hat{\tau}_x^2$.

These estimates follow immediately from the statement of Theorem 1, which decomposed the bias due to a measurement error component into its effect on coefficients measured without error, Z_i, separate from its effect on those measured with error, $\hat{\theta}_{xi}$. Although this decomposition conveniently represents the propensity for bias in a single value, $\overline{\mathrm{B}}_x^{ls}$, we do not find it convenient for estimating purposes, nor is it clear how to assess its standard error. Here we re-express the results of Theorem 1 as a single expression.

Re-expression of Theorem 1 bias results

As with Chapter 3, we find an alternative representation of the aggregate model more useful for deriving the method of moments estimate. We proceed in similar manner to the derivation of (3.13), and treat the marginal distribution of $\hat{\theta}_{yi}$, $\hat{\theta}_{xi}$, and Z_i.

We collect the predictors $\hat{\theta}_{xi}$ and Z_i together as $X_i = (\hat{\theta}_{xi}, Z_i')'$ and represent their true values by $x_i = (\theta_{xi}, Z_i)$, and denote $\mathbf{X} = (X_1, X_2, \cdots, X_k)'$. We will denote the regressors and predictand together by $W_i = (\hat{\theta}_{yi}, X_i')'$ and their true values by $w_i = (\theta_{yi}, x_i')'$. We use p to represent the dimension of Z_i, but $p' = p + 1$ to represent the dimension of w_i. We think of the observed quantity W_i as w_i measured with error. Because Z_i, a component of both X_i and x_i, does have measurement error, it may be convenient to think of Z_i as having measurement error with zero variance. We can re express the measurement error distribution (3.1) by

$$W_i \mid w_i, \phi \sim \left[w_i = \begin{pmatrix} \hat{\theta}_{yi} \\ x_i \end{pmatrix}, \Sigma_i = \begin{pmatrix} \sigma^2_{yi} & \boldsymbol{\beta}'_{wi}\Sigma_{xxi} \\ \Sigma_{xxi}\boldsymbol{\beta}_{wi} & \Sigma_{xxi} \end{pmatrix} \right] \tag{4.3}$$

The measurement variance σ^2_{yi} has the same definition used in Chapter 3. The sub-matrix Σ_{xxi} represents the error variance for the predictors X_i, and the $\boldsymbol{\beta}_{wi}$ represent multivariate versions of the within trial regression slope. To make the present specification equal to (3.1) we set to zero all components of Σ_{xxi} and $\boldsymbol{\beta}_{wi}$ that correspond to elements of Z_i.

We re-express the structural model (3.7) by

$$w_i \mid \phi = \left[\mu_w = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \Lambda_w = \begin{pmatrix} \lambda^2_y & \boldsymbol{\beta}'_x\Lambda_{xx} \\ \Lambda_{xx}\boldsymbol{\beta}_x & \Lambda_{xx} \end{pmatrix} \right] \tag{4.4}$$

where the parameter $\boldsymbol{\beta}_x = (\boldsymbol{\beta}'_z, \beta_\theta)'$ represents the coefficients of the ecological model. Then W_i has marginal distribution given by

$$W_i \mid \phi = \left[\mu_w = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix}, \Sigma_i + \Lambda_w = \begin{pmatrix} \sigma^2_{yi} + \lambda^2_y & \boldsymbol{\beta}'_{wi}\Sigma_{xx} + \boldsymbol{\beta}'_x\Lambda_{xx} \\ \Sigma_{xxi}\boldsymbol{\beta}_{wi} + \Lambda_{xx}\boldsymbol{\beta}_x & \Sigma_{xxi} + \Lambda_{xx} \end{pmatrix} \right] \tag{4.5}$$

Keeping with the notation in Chapter 3, we define $\mathbf{B}_x = (\Sigma_{xxi} + \Lambda_{xx})^{-1}\Sigma_{xxi}$, a multivariate shrinkage factor.

The marginal model (4.5) allows us to re-express the single observation bias (3.9) by

$$\mathrm{E}(\hat{\theta}_{yi} \mid X_i, \phi) = \mu_y + (X_i - \mu_x)'(\Sigma_{xxi} + \Lambda_{xx})^{-1}(\Sigma_{xxi}\boldsymbol{\beta}_{wi} + \Lambda_{xx}\boldsymbol{\beta}_x) \tag{4.6}$$

$$= \mu_y + (X_i - \mu_x)'(\mathbf{B}_{xi}\boldsymbol{\beta}_{wi} + (I - \mathbf{B}_{xi})\boldsymbol{\beta}_x) \tag{4.7}$$

As with Theorem 1 on page 33, we let d_i, $i = 1 \cdots k$, represent a set of weights summing to 1, and denote the weighted sample variance of W_i by S_W, and $\bar{\Sigma}$ represents the weighted mean of the Σ_i. Theorem 1 states that S_W consistently estimates $\overline{\Sigma}+\Lambda_w$, and a least squares regression will with weights d_i estimates regression coefficients represented by Sweep$[X](\bar{\Sigma}+\Lambda)$. Thus we may now re-express (3.14) and (3.15) simultaneously by

$$\hat{\mathrm{E}}(\hat{\theta}_{yi} \mid X, \phi) = \mu_y + (X_i - \mu_x)'(\overline{\mathbf{B}}_x\bar{\boldsymbol{\beta}}_w^{ls} + (I - \overline{\mathbf{B}}_x)\boldsymbol{\beta}_x) \tag{4.8}$$

Representing a submatrix of $\overline{\Sigma}$ by $\overline{\Sigma}_{xx}$, then $\overline{\mathbf{B}}_x = (\overline{\Sigma}_{xx} + \Lambda_{xx})^{-1}\overline{\Sigma}_{xx}$ represents a multivariate version of the average $\overline{\mathbf{B}}_x^{ls}$, and we define $\bar{\boldsymbol{\beta}}_w$ so that $\overline{\Sigma}_{xx}\bar{\boldsymbol{\beta}}_w = \frac{1}{k}\sum_{i=1}^{k}\Sigma_{xxi}\boldsymbol{\beta}_{wi}$. Again, we accomplish this by setting to zero components of $\bar{\boldsymbol{\beta}}_w$ that multiply Z_i.

Expression (4.8) restates Theorem 1, except that here we do not separate the bias into components associated with θ_{xi} and those associated with Z_i.

Theorem 1 states that $\hat{\Lambda}_w = S_W - \overline{\Sigma}$ consistently estimates Λ_w. We use notation $\hat{\overline{\mathbf{B}}}_x$ to represent $\overline{\mathbf{B}}_x$ evaluated at $\hat{\Lambda}_{xx}$, a submatrix of $\hat{\Lambda}_w$. Letting us denote the coefficient estimate of X_i by $\hat{\boldsymbol{\beta}}_x^{ls}$, then equating it to the coefficient of X_i in (4.8) and solving for β_0, we get the following method of moments estimate

Method of Moments Estimate

$$\hat{\boldsymbol{\beta}}_x^{mom} = (I - \hat{\overline{\mathbf{B}}}_x)^{-1}(\hat{\boldsymbol{\beta}}_x^{ls} - \hat{\overline{\mathbf{B}}}_x\bar{\boldsymbol{\beta}}_w) \tag{4.9}$$

We may express the standard error of (4.9) in terms of the standard error of $\hat{\beta}_x^{ls}$ by

$$\widehat{V}_{mom} = (I - \hat{\overline{\mathbf{B}}}_x)^{-1}\mathrm{Var}(\hat{\beta}_x^{ls})(I - \hat{\overline{\mathbf{B}}}_x)^{-1} \tag{4.10}$$

where $\text{Var}(\hat{\beta}_x^{ls})$ gives the variance of the least squares estimator. However, we may not use the standard error estimate of $\hat{\beta}_x^{ls}$ given to use by a regression package. Least squares methods estimate variance by $\widehat{\text{Var}}(\hat{\beta}_{ls}) = \hat{\sigma}^2(X'DX)^{-1}$, where $\hat{\sigma}^2$ represents the mean squared error of the residuals $Y_i - X_i'\hat{\beta}^{ls}$. But by (3.35), $X_i'\hat{\beta}^{ls}$ is not the expectation of $\hat{\theta}_{yi}$. This causes the mean squared residuals to be too large, and the least squares estimate of variance to be too large also. We may derive the standard error of $\hat{\beta}_{ls}$ using "sandwich estimators" Liang and Zeger (1986), as follows.

We first recognize that the right hand side of (4.7) estimates the mean of $\hat{\theta}_{yi}$. That is, letting us denote the right side of (4.7) by η_{yi}, then $R_i = \hat{\theta}_{yi} - \eta_{yi}$ gives its error term. If we use $\hat{\eta}_i$ to denote η_i evaluated at $\hat{\Lambda}_w$ and X_i, then $\hat{R}_i = \hat{\theta}_{yi} - \hat{\eta}_{yi}$ is a consistent estimate of the error term $\hat{R}_i$. We use $\hat{\mathbf{R}} = \text{diagonal}(\hat{R}_i, \hat{R}_2, \cdots, \hat{R}_k)$ to denote a diagonal matrix with diagonal elements $\hat{R}_i$. follows.

The residual variance for $\hat{\theta}_{yi}$ is $\text{Var}(\hat{\theta}_{yi} \mid \hat{\theta}_{xi}, Z_i, \phi)$, which is given in (3.10) on page 30. Denote its value as V_i, and let $\mathbf{V}$ be a diagonal matrix with diagonal elements V_i. Then by definition the variance of $\hat{\boldsymbol{\beta}}_x^{ls}$, is expressed as

$$\begin{aligned}\text{Var}\left(\hat{\boldsymbol{\beta}}_x^{ls}\right) &= \text{Var}\left((\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\hat{\boldsymbol{\theta}}_y\right) \\ &= (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{V}\mathbf{D}'\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\end{aligned}$$

Notice that the residual variance V_i is different for every observation, and depends on unknown parameters. We may plug in point estimates to compute the V_i, but we find it more convenient to express the standard error using sandwich estimates of variance (Liang and Zeger, 1986, see). Because $\hat{R}_i$ consistently estimates R_i, we may replace $\mathbf{V}$ by $\mathbf{R}\mathbf{R}'$ and so

$$\text{Var}(\hat{\beta}_x^{ls}) = \frac{k}{k-p'}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\hat{\mathbf{R}}\hat{\mathbf{R}}'\mathbf{D}'\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1} \tag{4.11}$$

is consistent for V_{ls}. Substituting (4.11) into (4.10) gives us our standard error estimate.

The estimator (4.9) is mathematically equivalent to the estimator given by Fuller (1987)[1]. For other estimators of variance that treat more restricted conditions, such as the case with uncorrelated or equal measurement error variances (see Seber, 1977; Davies and Hutton, 1975).

4.2 Likelihood Inference

The measurement and structural models, and so the aggregate and posterior models, summarized in Table 3.1 have normal distribution. Direct likelihood inference for ϕ is made through the *observed likelihood* that results from k observations from the aggregate model

$$
\begin{aligned}
L(\phi \mid \hat{\boldsymbol{\theta}}) &\propto p(\hat{\boldsymbol{\theta}} \mid \phi) \\
&\propto \prod p(\hat{\theta}_i \mid \phi) \\
&\propto \prod_{i=1}^{k} \mid \Sigma_i + \Lambda \mid^{-1/2} \exp\left\{(\hat{\theta}_i - \eta_i)'(\Sigma_i + \Lambda)^{-1}(\hat{\theta}_i - \eta_i)\right\} \qquad (4.12)
\end{aligned}
$$

To compute maximum likelihood estimates we must find the mode of (4.12), and to compute standard errors we must evaluate the second derivative of its log at the mode. Bayes procedures make inferences from (4.12) multiplied by a prior distribution $p(\phi)$. Because of the manner that Λ and β_0 enters the likelihood, direct application of Bayes rule or maximization will be difficult. Here we make use of missing data methods, the EM (Dempster *et al.*, 1977) and data augmentation (Tanner and Wong, 1987) algorithms to make inferences for ϕ by treating θ_i as missing data (for good tutorials on these methods see Gelman *et al.* (1995), and for an example of their use on a simpler version of this model see McIntosh (1996)).

Preliminaries

[1] Fuller (1987), page 183, provides a complicated variance estimate, but there is an error (probably a typo) in his statement of the estimate. The error is apparent because the dimensions of matrix products and the end of the theorem do not conform. I have not been able to determine what the estimator should be.

Although we find it difficult to make inferences from (4.12) directly, we find that indirect methods are simple. We notice that if $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)'$ were observed, then we would straight forwardly perform both maximum likelihood and Bayes inferences procedures. Although we do not observe θ_i, if we knew ϕ we could predict them. Recall from Table 3.1 that,

$$p(\theta_i \mid \theta_i, \phi) = N(\theta_i^*, \Lambda_i^*) \tag{4.13}$$

(see Table 3.1 for definitions of θ_i^* and Λ^*). Both the EM and data augmentation algorithms make use of this structure.

4.2.1 Maximum Likelihood Estimation: EM-Algorithm

We follow the notation from Section 4.1, and we let $W_i = (\hat{\theta}_{yi}, \hat{\theta}_{xi}, Z_i')$, $i = 1 \cdots k$ represent the observed data, and $w_i = (\theta_{yi}, \theta_{xi}, Z_i')$, $i = 1 \cdots k$ will be called the missing data. Together the values $(W_i, w_i), i = 1 \cdots k$ are called the complete data.

Because the structural model has a linear expectation with normal error, if we observed the complete data we would estimate ϕ from with the sufficient statistics $\overline{w} = \frac{1}{k}\sum_{i=1}^{k} w_i$ and $\overline{ww'} = \frac{1}{k}\sum_{i=1}^{k} w_i w_i'$. Because the sufficient statistics depend on unobserved data we call them the *complete data sufficient statistics* (we borrow these terms from Dempster *et al.*, 1977). Note too that if ϕ were known, we could predict the complete data sufficient statistics conditional on the observed data $\mathbf{W}$ and ϕ using (4.13).

The EM algorithm maximizes the observed likelihood (4.12) by alternating between computing the expected complete data sufficient statistics (E-step) and computing maximum likelihood estimates from them (M-Step) as follows. Let $\phi^{(n)}$ represent an estimate of the structural parameter ϕ ('n' means the nth iteration). The EM algorithm proceeds by first computing the E-Step as if $\phi = \phi^{(n)}$.

E-Step:

$$\bar{w}^{(n)} = E(\overline{w} \mid \hat{\boldsymbol{\theta}}, \phi = \phi^{(n)}) = \frac{1}{k}\sum_{i=1}^{k} \begin{pmatrix} \theta_i^* \\ Z_i \end{pmatrix} = \begin{pmatrix} \bar{\theta}^* \\ \bar{Z} \end{pmatrix} \tag{4.14}$$

$$\overline{ww'}^{(n)} = E(\overline{ww'}^{(n)} \mid \hat{\boldsymbol{\theta}}, \phi = \phi^{(n)}) = \frac{1}{k}\sum_{i}^{k} \begin{pmatrix} \theta_i^*\theta_i^{*\prime} + \Lambda^* & \theta_i^* Z_i' \\ Z_i\theta_i^{*\prime} & Z_i Z_i' \end{pmatrix} \tag{4.15}$$

The 'starred' terms above are defined in Table 3.1 (lower right corner), and we evaluate them at $\phi = \phi^{(n)}$. With these expected sufficient statistics, the M-Step updates the parameter $\phi^{(n+1)}$ by computing the maximum likelihood estimate of ϕ. Letting $S_w^{(n)} = \overline{ww'}^{(n)} - \overline{w}^{(n)}\overline{w}'^{(n)}$, we complete the M-Step by setting $\phi^{(n+1)}$ to

M-Step:

$$\omega_0^{(n+1)} = \bar{\theta}^* \tag{4.16}$$

$$\text{Sweep}[Z](S_w^{(n)}) = \begin{pmatrix} \Lambda^{(n+1)} & \hat{\omega}_z^{(n+1)} \\ \hat{\omega}_z^{(n+1)} & \mathbf{Z}'\mathbf{Z}^{-1} \end{pmatrix} \tag{4.17}$$

Iterating between the E-Step and M-Step gives a sequence of parameter estimates $\phi^{(n)}, \phi^{(n+1)}, \cdots$ that converges to a value that maximizes (4.12). We assess convergence by computing the log of the observed likelihood (4.12) after each iteration, and monitoring its increases. Dempster *et al.* show that the observed likelihood increases after each step, and so when the successive increases become small, we conclude that the algorithm has converged. Standard errors can be computed by the SEM algorithm (Meng and Rubin, 1991), but in this manuscript we prefer to estimate standard errors by computing numerical second derivatives of the log of the complete data likelihood (4.12) around its mode (for example Gelman *et al.*, 1995, page 273). Once the EM algorithm finds the mode, computing numerical derivatives on a computer is quick and efficient.

4.2.2 Bayes Estimates

For Bayes estimation we wish to investigate the posterior distribution proportional to the observed likelihood (4.12) multiplied by some prior, $p(\phi)$. Because this does not have a convenient closed form, we estimate it by simulation, using the data augmentation algorithm (Tanner and Wong, 1987). The data augmentation algorithm is a Markov Chain Monte Carlo (MCMC) algorithm and a special form of a Gibbs Sampler (see Gelfand and Smith, 1990, for an overview of MCMC methods). The steps of the data augmentation algorithm have analogies to the EM algorithm, where instead of computing expectations of missing data (E-Step) and computing maximum likelihood (M-Step), we simulate the missing data (Augmentation Step), and simulate parameters (Parameter Step).

Notice that if ϕ were known, then we could simulate θ_i from (4.13) by

Augmentation Step:

$$\theta_i \sim p(\theta_i \mid \hat{\theta}_i, \phi) = .N(\theta_i^*, \Lambda_i^*)$$

Conversely, if $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)'$ were observed, then the posterior distribution $p(\phi \mid \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}, \phi)$ has a familiar form. We represent a normal density with mean $\boldsymbol{\mu}$ and variance Σ as $G(\cdot \mid \boldsymbol{\mu}, \Sigma)$. To improve readability we let its dimension be implied by the dimension of $\boldsymbol{\mu}$. Define the linear predictors $\eta_{yi} = X_i'\beta$, $\eta_{xi} = Z_i'\gamma$. Then we compute the parameter step by

Parameter Step:

$$\begin{aligned} p(\phi \mid \boldsymbol{\theta}, \hat{\boldsymbol{\theta}}) &\propto p(\phi) p(\hat{\boldsymbol{\theta}} \mid \boldsymbol{\theta}, \phi) p(\boldsymbol{\theta} \mid \phi) \\ &\propto p(\phi) p(\boldsymbol{\theta}_y \mid \boldsymbol{\theta}_z, \phi) p(\boldsymbol{\theta}_z \mid \phi) \\ &\propto p(\phi) \prod_{i=1}^{k} \{G(\theta_{yi} \mid \eta_{yi}, \tau_y^2) G(\theta_{xi} \mid \eta_{xi}, \tau_x^2)\} \end{aligned} \tag{4.18}$$

The term involving $\hat{\boldsymbol{\theta}}$ drops from the derivation above because it does not depend on unknown parameters. Because the structural model is defined by the product of two linear regressions, the parameter step derivation involves two successive posterior computations of a normal linear model, and can be found in many standard texts (see for example Box and Tiao, 1992, p114). We derive the exact form of $p(\phi \mid \boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$ in Section 4.2.2.

The data augmentation algorithm alternates between the augmentation and parameter steps. First, with a current parameter $\phi^{(n)}$ ('n' meaning the nth iteration), we impute the missing data by $\theta_i^{(n)} \sim p(\theta_i \mid \hat{\theta}_i, \phi = \phi^{(n)})$, $i = 1 \cdots k$, as described in the augmentation step. We denote the values imputed with parameter $\phi^{(n)}$ by $\boldsymbol{\theta}^{(n)}$. Second, we update the parameter $\phi^{(n+1)}$ by simulating $\phi^{(n+1)} \sim p(\phi \mid \boldsymbol{\theta} = \boldsymbol{\theta}^{(n)}, \hat{\boldsymbol{\theta}})$ (shown below). The sequence $\phi^{(1)}$, $\phi^{(2)}$, $\phi^{(3)}$,..., converges to the posterior distribution $p(\phi \mid \hat{\boldsymbol{\theta}})$.

The choice of a starting value and assessing convergence for MCMC methods is very important. Here we use the method prescribed by Gelman and Rubin (1992) (see the companion piece by Geyer, 1992, for a different view). We outline that recommendation in Section 4.2.3.

Derivation of the parameter step

Before we derive the parameter step, we first define some notation. We find it most convenient to express the posterior distribution including the constant term in the predictors, so use $\tilde{Z}_i = (1, Z_i')$ and denote the predictors θ_{xi} and Z_i together as $\tilde{x}_i = (\tilde{Z}_i, \theta_{xi})$. We also use $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \tilde{Z}_2, \cdots, \tilde{Z}_k)'$ and $\tilde{\mathbf{x}} = (\tilde{x}_1, \tilde{x}_2, \cdots, \tilde{x}_k)'$, and represent the dimension of Z_i as p and the dimension of X_i as $p' = p + 1$. We use the following recognizable definitions from linear models: $\hat{\boldsymbol{\beta}}$, the estimate of $\boldsymbol{\beta}$, $\widehat{H}_{\hat{\beta}} = (\tilde{\mathbf{x}}'\tilde{\mathbf{x}})^{-1}$, and $s^2_{y|x}$ estimate of residual variance from a regression of θ_{yi} on $\tilde{\mathbf{x}}_i$; $\hat{\gamma}$, the estimate of γ, $\widehat{H}_{\hat{\gamma}} = (\tilde{\mathbf{Z}}'\tilde{\mathbf{Z}})^{-1}$, and s^2_x the estimate of residual variance, from a regression of θ_{xi} on Z_i. Finally, we use $IG(\tau^2 \mid \delta/2, q/2)$ to represent the inverse gamma density given by

$$IG(\tau^2 \mid \delta/2, q/2) \propto \left(\frac{1}{\tau^2}\right)^{\frac{q}{2}-1} \exp\left(-\frac{\delta}{2}\frac{1}{\tau^2}\right) d\left(\frac{1}{\tau^2}\right) \tag{4.19}$$

The likelihood portion of (4.18) can factor into components containing the sufficient statistics given by (see Box and Tiao, 1992)

$$
\begin{aligned}
L(\boldsymbol{\phi} \mid \boldsymbol{\theta}, \mathbf{Z}) &\propto \prod_{i=1}^{k} \{G(\theta_{yi} \mid \eta_{yi}, \tau_y^2) G(\theta_{xi} \mid \eta_{xi}, \tau_x^2)\} \\
&\propto G(\boldsymbol{\beta} \mid \hat{\boldsymbol{\beta}}, \tau_{y|x}^2 \widehat{H}_{\hat{\beta}}) \times IG(\tau_{y|x}^2 \mid \frac{(k-p)s_{y|x}^2}{2}, \frac{k-p+2}{2}) \qquad (4.20) \\
&\times G(\gamma \mid \hat{\gamma}, \tau_x^2 \widehat{H}_{\hat{\gamma}}) \times IG(\tau_{y|x}^2 \mid \frac{(k-p')s_x^2}{2}, \frac{k-p'+2}{2}) \qquad (4.21)
\end{aligned}
$$

Expression (4.20) factors the ecological model, and (4.21) factors the population risk model. We often find it convenient to choosing our priors from the class of conjugate priors because they lead to simple expressions for the posterior distributions. That is, if we choose priors given by

$$
p(\boldsymbol{\phi}) = G(\boldsymbol{\beta} \mid \boldsymbol{\beta}^0, \Sigma_\beta) \times IG(\tau_{y|x}^2 \mid \delta_{y|x}/2, q_{y|x}/2) \times G(\gamma \mid \gamma^0, \Sigma_\gamma) \times IG(\tau_x^2 \mid \delta_x/2, q_x/2)
$$

then the prior combines with the likelihood to produce the posterior distribution described by

$$
\begin{aligned}
\tau_x^2 &\sim IG(\tau_x^2 \mid \frac{\delta_x + (k-p)s_x^2}{2}, \frac{q_x + k - p}{2}) \qquad (4.22) \\
\gamma \mid \tau_x^2 &\sim G(\gamma \mid \gamma^*, \Sigma_\gamma^*) \qquad (4.23) \\
\tau_{y|x}^2 &\sim IG(\tau_{y|x}^2 \mid \frac{\delta_{y|x} + (k-p)s_{y|x}^2}{2}, \frac{q_{y|x} + k - p'}{2}) \qquad (4.24) \\
\gamma \mid \tau_y^2 &\sim G(\gamma \mid \gamma^*, \Sigma_\gamma^*) \qquad (4.25)
\end{aligned}
$$

Parameters $\boldsymbol{\beta}^0$ and Σ_β represent the mean and covariance of $\boldsymbol{\beta}$, and γ^0 and Σ_γ represent the prior mean and covariance for γ. With uniform priors for $\boldsymbol{\beta}$ and γ, then $\boldsymbol{\beta}^* = \hat{\boldsymbol{\beta}}$ and $\gamma^* = \hat{\gamma}$. Otherwise,

$$
\begin{aligned}
\boldsymbol{\beta}^* &= \mathbf{B}_\beta \boldsymbol{\beta}^0 + (I - \mathbf{B}_\beta)\hat{\boldsymbol{\beta}} \qquad (4.26) \\
\Sigma_\beta^* &= \tau_{y|x}^2 \widehat{H}_\beta \mathbf{B}_\beta \qquad (4.27)
\end{aligned}
$$

and

$$\gamma^* = \mathbf{B}_\gamma \gamma^0 + (I - \mathbf{B}_\gamma)\hat{\gamma} \tag{4.28}$$

$$\Sigma_\gamma^* = \tau_x^2 \widehat{H}_\gamma \mathbf{B}_\gamma \tag{4.29}$$

where

$$\mathbf{B}_\beta = (\widehat{H}_\beta \tau_x^2 + \Sigma_\beta)^{-1} \widehat{H}_\beta \tau_x^2 \tag{4.30}$$

$$\mathbf{B}_\gamma = (\widehat{H}_\gamma \tau_{y|x}^2 + \Sigma_\gamma)^{-1} \widehat{H}_\gamma \tau_{y|x}^2 \tag{4.31}$$

4.2.3 Evaluating Convergence

For the present discussion we use θ to denote a generic parameter. When assessing convergence of an MCMC algorithm we use the method prescribed by Gelman and Rubin (1992) (for a good tutorial see Gelman *et al.*, 1995, p331). Instead of using a single MCMC sequence, they use J independent sequences, each with a different starting value (that is, each has a different $\theta^{(1)}$). We use $\theta_j^{(i)}$ to denote the i-th iteration from chain j. Thus we have J independent sequences $\theta_j^{(1)}, \theta_j^{(2)}, \cdots, \theta_j^{(n)}$, for $j = 1 \cdots J$, and each converging to the target posterior distribution. But because they were started at different values, all samples together are more variable than the target distribution because they contain within and between sources of variability. When the chain converges, the between sequence variability should be undetectable (that is, the distribution should be independent of its starting point if it has converged). Gelman and Rubin use this to assess the chains convergence as follows

After each chain completes $2n$ iterations we examine the final half, for last n draws from each chain. For each sequence we compute its mean, which we denote by $\bar{\theta}_j$, and its variance, which we denote s_j^2, and also the grand mean of all J sequences, which we denote

$\bar{\theta}$. From these we compute the within and between sequence variance components by

$$B_V = \frac{n}{J-1}\sum_{j=1}^{J}(\bar{\theta}_j - \bar{\theta})^2$$

$$W_V = \frac{n}{J}\sum_{j=1}^{J} s_j^2$$

The quantity B_V measures the total variability within each sequence around the grand mean. The quantity W_V measures the total variance of each sequence around the individual means. Now define the quantity

$$\widehat{\mathrm{Var}}(\theta) = \frac{n-1}{n}W_V + \frac{1}{n}B$$

Gelman and Rubin assess convergence by computing

$$\sqrt{\hat{R}} = \sqrt{\frac{\widehat{\mathrm{Var}}(\theta)}{W_V}} \tag{4.32}$$

and stopping the MCMC algorithm when $\sqrt{\hat{R}}$ becomes small. They recommend stopping when $\sqrt{\hat{R}} < 1.2$, but with our analyses we use $\sqrt{\hat{R}} < 1.1$.

Intuitively, this rule stops the sequence when the within sequence variance dominates the between sequence variance. The expression (4.32) gets small as $n \to \infty$, and as the means from the J sequences come to agree about the posterior mean of θ.

4.3 Discussion

The behavior of the estimating procedures can be understood by examining the simple method of moments estimator (3.30) on page 42. That expression can be rewritten as

$$\hat{\beta}_0^{mom} = \frac{\hat{\beta}_0^{ls} - \widehat{\overline{\mathrm{B}}}_x^{ols}\bar{\beta}_w}{1 - \widehat{\overline{\mathrm{B}}}_x^{ols}} \tag{4.33}$$

$$= \left(\hat{\beta}_0^{ls} - \widehat{\overline{\mathrm{B}}}_x^{ols}\bar{\beta}_w\right)\left(1 + \frac{\bar{\sigma}_x^2}{\tau_x^2}\right) \tag{4.34}$$

The estimate of $1/\tau_x^2$ determines how far we estimate β_0 from the observed slope, $\hat{\beta}_0^{ls}$. This is true empirically, and we can also see this in the simple methods of moments estimate (4.34). If $1/\tau_x^2$ has small value, then the second factor on the right side has value near one, but as τ_x^2 gets small, the second factor approaches ∞.

Also, if $1/\tau_x^2$ has large value then the uncertainty of inferences for β_0 increases. Expression (4.34) demonstrates this because as $1/\tau_x^2$ increases, the second factor on the right hand side, a multiplicative factor for $\hat{\beta}_0^{ls}$, increases as well. There is also intuition for this result. If β_0 is the coefficient of a simple linear regression model then its uncertainty is determined by the ratio of the residual variance and the variance of the covariate; here $\tau_{y|x}^2/\tau_x^2$. Because both of these effects (bias adjustment and variance) depend on $1/\tau_x^2$, the inference for this parameter is very important when estimating β_0. In particular, we find inferences for β_0 sensitive to small values of τ_x^2.

The maximum likelihood and method of moments procedures may estimate τ_x^2 as zero (or negative), especially with small k. When this happens the estimates are invalid, because this implies that the ecological regression has a vertical line. But even when maximum likelihood or moment estimates have $\tau_x^2 \leq 0$, the likelihood supports larger values of τ_x^2 and the Bayes estimates are valid. When this occurs, we may either conclude that heterogeneity in the control group does not exist, or we must use the Bayes estimates. With small k we prefer the Bayes estimates.

With large k all reasonably vague priors result in similar inferences, but with small k inferences for β_0 are sensitive to the choice of δ_x and q_x. With $\delta = 0$, (4.19) provides two candidates that may be considered as vague prior distributions[2]; prior 1, uniform on the standard deviation ($q = -1$), and prior 2, uniform on the variance ($q = -2$). Prior 2 places weight on larger values of τ^2 than does prior 1. If we choose prior 1 or 2 for τ_x^2, the

[2] Jeffreys' prior, which is uniform on $\log(\tau^2)$, and corresponds to (4.19) with $q = 0$ and $\delta = 0$ does not lead to a proper posterior. See DuMouchel and Waternaux (1992) for a discussion.

Bayes procedure typically estimate τ_x^2 larger, and results in smaller bias adjustment, than the maximum likelihood procedure. Prior 2, because it places weight on larger values of τ^2, typically results in less bias adjustment than prior 1. Previous work (McIntosh, 1996; Schmid *et al.*, 1995) suggests using $\delta_{y|x} = \delta_x = 0$, $\delta_{y|x} = -1$ (uniform on $\tau_{y|x}$) and $\delta_x = -2$ (uniform on τ_x^2) as reference priors.

We can also choose parameters δ and q to represent informative opinion. If we express opinion in the form of a mean and variance, or two prior quantiles (for example the 1st and 99th), δ and q are chosen so that (4.19) matches this opinion (see Gelman *et al.* (1995), pages 139-140, for a discussion).

The model constructed in Chapter 3 has a bivariate normal hierarchical representation, and there exist many currently available packages to estimate such models (see for example Bryk and Raudenbush, 1992; Everson, 1995). Many of these may not be useful for the model we treat because our estimand is the ratio of two variance components. To use those methods for ecological inference they must estimate the components of the structural covariance matrix, Λ, but also provide the means necessary to estimate the uncertainty of their ratios (a ratio of variance components defines β_0). Also, many methods are primarily concerned with making inferences for the mean parameters ω, and their procedures are tuned so that these estimates have good frequency properties. In particular, Bayes estimates will typically use prior distributions that are in some way uninformative for Λ, but that may lead to informative priors for β_0, and so may not have good frequency properties when estimating β_0.

4.4 Data Analyses

Here we demonstrate the normal model estimating procedures on the complete streptokinase data. Recall that the normal model should be assumed valid only for clinical trials that have outcomes $\hat{p}_{ti}$, $\hat{p}_{ci}$ and samples n_{ti} and n_{ci} large enough for the normal approximation to hold. Here we analyze the entire data set, and include even those that do not fit this criteria, because we will follow this section with another that uses these data to evaluates the procedures. We wish to make it clear that in practice these procedures should not be applied to similar data without considerable evaluation of the procedures performance for that particular data set. The simulation results of the following section show that these procedures applied to data with nonormal measurement errors can introduce bias that exceeds that caused by error attenuation.

4.4.1 Streptokinase Data

We apply each estimating procedure to the streptokinase data. For Bayes estimates we use the convergence assessment method described in Section 4.2.3 with $J = 5$. We find for these data that each sequence running 2,000 iterations is more than enough to meet our convergence criteria (i.e., produces $\sqrt{\hat{R}} < 1.1$). So the Bayes estimates given below represent the summaries of 5,000 simulated values (from the second half of each sequence). We use priors uniform on β, γ, and as recommended Section 4.3, uniform on $\tau_{y|x}$ and τ_x^2.

Table 4.3 on page 75 summarizes the Bayes, maximum likelihood, and method of moments estimates of the streptokinase structural model. For comparison, that table also provides the ecological model estimates that result when ignoring the measurement error and including the population risk as a covariate in a random effects model. Figure 4.1 plots the ecological regression lines that Table 4.3 summarizes.

The random effects model estimates an ecological slope of -0.178, and is statistically significant (one sided p-value=0.036). Thus ignoring measurement error leads us to conclude that an association between the effect of treatment and the population risk exists. The structural procedures give a different conclusion. Each structural model estimating proce-

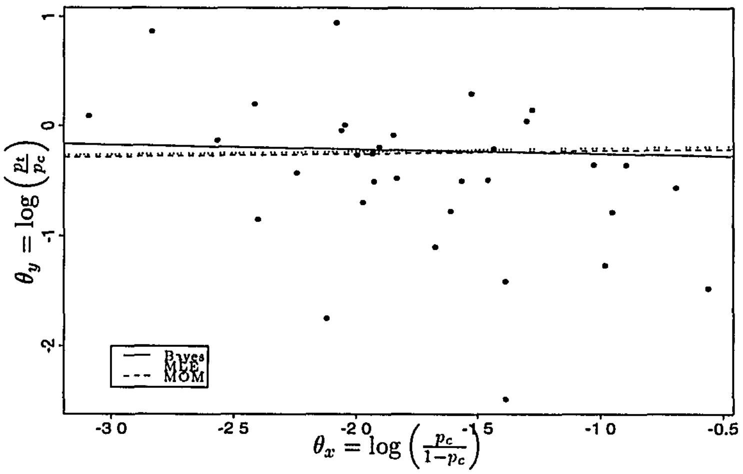


Figure 4.1: Estimated ecological models for the streptokinase data.

dure agrees about the structural slopes magnitude, giving point estimates very close to zero and interval estimates that contain $\beta_\theta = 0$ (see caption of Table 4.3 for a description of the interval estimates). Figure 4.1 displays each of these estimates, showing nearly horizontal ecological iegression lines. Notice that the maximum likelihood procedure gives narrower confidence intervals than does the Bayes procedure. We find that this typical for all data analyses in this manuscript and elsewhere (see McIntosh, 1996). Figure 4.2(a) shows the ecological slope posterior distribution to have normal shape, and saddles $\beta_\theta = 0$. We conclude from these results that there does not exist an ecological association for streptokinase.

To evaluate the overall effect of streptokinase we examine the estimates of β_0, the mean treatment effect. Each procedure gives interval estimates that exclude $\beta_0 = 0$, and so we have confidence that $\theta_{yi} < 0$ on average. We may also evaluate how likely we are to find some future trial, θ^+_{yi}, having $\theta^+_{yi} > 0$. If no structural association exists then θ^+_{yi} has normal distribution with mean β_0 and standard deviation $\tau_{y|x}$. Only the Bayes procedure estimates $\tau_{y|x} > 0$ and suggests that heterogeneity still exists after controlling for population risk. The data augmentation procedure provides us with a convenient method for answering this question.

We express the risk of finding a trial harmful by $P(\theta^+_{yi} > 0 \mid \hat{\theta})$, but if ϕ were known then we evaluate the risk by $P(\theta^+_{yi} > 0 \mid \phi)$, which we may calculate from a normal table. Without knowing ϕ we express that risk as $P(\theta^+_{yi} > 0 \mid \hat{\theta}) = E(P(\theta^+_{yi} > 0 \mid \phi))$ where we take the expectation with respect to the posterior distribution $p(\phi \mid \hat{\theta})$. Because we have 5,000 samples from this distribution, which we denote by $\phi^{(i)}$, $i = 1 \cdots 5000$, we may estimate the risk with its posterior mean by computing $\widehat{P}(\theta^+_{yi} > 0 \mid \hat{\theta}) = \frac{1}{5000}\sum_{1=1}^{5000} P(\theta_{yi} > 0 \mid \phi = \phi^{(i)})$. For the streptokinase data we have $\widehat{P}(\theta^+_{yi} \mid \hat{\theta}) = 0.073$. This exceeds the values 0.058 which we found with the simple random effects model in Chapter 2, and using expression (2.13). We therefore conclude that on average streptokinase has benefit, but its heterogeneity may be large enough to cause worry that some populations may be harmed.

The data analyses of Chapter 5 demonstrate uses of the augmentation procedure to estimate other useful parameters.

4.5 Evaluations

Here we use the streptokinase data to evaluate the performance of the normal model estimating procedures. Many of the streptokinase trials do not contain a sufficient number of events for the normal approximation to hold. This section intends to seek evidence for two points: First, to determine the operating characteristics when data have true normal error; Second, to determine the performance of the procedures when measurement errors cannot be assumed normally distributed. That is, to determine the performance of the procedures when the approximations fail.

As we have done throughout this manuscript, we focus our attention on the ecological slope and evaluate the frequency properties of point estimates and confidence intervals for

β_θ. This is true for the Bayes estimates, which for a fixed choice of prior distribution, we view as a frequency procedure. We conduct all simulations with $\beta_\theta = 0$. We do this not only because we have focused considerable energy throughout this manuscript on that case, but because the frequency characteristics when $\beta_\theta = 0$ has important policy and clinical implications.

The procedures frequency properties depend on many things, including the true measurement error, the size and number of the trials, k and n_i respectively, and also on the true value of the parameter, ϕ. Additionally, the performance of the Bayes procedures depend on the choice of priors, and so the Bayes procedures are actually a family of procedures. To limit the scope of the simulations we do the following.

For parameter values other than β_θ we choose to mimic the streptokinase, and we use their sample sizes as well. We use $\phi_s \equiv (\beta_0 = -0.239, \beta_\theta = 0, \gamma_0 = -1.771, \tau_{y|x} = 0.110, \tau_x = 0.462)$. We choose these values because they are the posterior mean estimates that result from the estimation procedures given in Chapter 5, which we believe to provide valid inferences with binomial error.

We choose prior distributions $p(\phi)$ to be uniform on β, γ, $\tau_{y|x}$, and τ_x^2. The reasons for choosing these priors come from our experience analyzing several data sets, and also from the discussion given in Section 4.3 and in McIntosh (1996). Those discussions show these priors for $\tau_{y|x}$ and τ_x lead to smaller adjustments of observed line, and smaller uncertainties in β_θ than other uniform priors. Thus if the Bayes methods perform well with these choices, coverage properties will not likely be worse with the other choices.

Finally, because the models derived in this chapter assume normal measurement error, and the real data have measurement error that are functions of binomial, we perform each set of simulations twice, once with normal measurement error and the other with binomial measurement error. That data has $k = 33$ trials, and so we use $\phi \equiv \phi_s$ to generate 33 pairs $\theta_i = (\theta_{yi}, \theta_{xi})$. For the binomial simulations we then calculate 33 pairs $p_i = \theta^{-1}(\theta_i)$, and then simulate 33 pairs $\hat{p}_i$ from binomial distributions with mean p_i and sample sizes $n_i = (n_{ti}, n_{ci})'$. For the normal simulations we use the same θ_i and then generate 33 pairs $\hat{\theta}_i \sim N(\theta_i, \Sigma_i)$, where Σ_i represents the large sample estimates of the measurement error variances given in Table 3.4 on page 48.

For the data analysis portion of the simulation we proceed as described in this chapter and calculate method of moments, maximum likelihood and Bayes estimates, noting that if we find any $\hat{p}_{ti} = 0$ or $\hat{p}_{ci} = 0$ we add a half observation to the number of events, and a full observation to the number of subjects in that clinical trial. For comparison we also investigate the performance of least squares estimates. In particular we evaluate ordinary least squares regression (OLS) and random effects meta-analysis (RE).

One technical point our simulations must consider is that both the maximum likelihood and method of moment procedures (from here on MLE and MOM), may give $\tau_x \leq 0$, in which case β_0 is not defined. When this happens we do not count it in our procedure evaluations, but we do make a note of the occurrence.

Each table that follows summarizes a simulation study. For each procedure, we give the average of its point estimates of β_0, and the empirical coverage of its 90% and 95% confidence intervals. Beneath each estimate we give the sample standard error of the estimates, which we use to evaluate whether we can trust that the procedures estimate the true values.

4.5.1 Complete Streptokinase data

Table 4.1 presents the results of the streptokinase simulations, giving the three structural model estimates on the top and the least squares estimates on the bottom. First notice the least squares estimates. In each instance their point estimates fall substantially far from their null values (all estimates are several standard errors from $\beta_0 = 0$), and the confidence intervals fall far short of their nominal values. In particular notice that the OLS

Table 4.1: Simulation results from all 33 streptokinase trials. Simulation results from the smallest nine streptokinase trials. Each estimate gives the empirical mean of its estimates for β_θ and the 90% and 95% confidence intervals. The values in parentheses give the standard error of the estimates.

	Gaussian Variation			Binomial Variation		
Method/Model	$\widehat{E}(\hat{\beta}_\theta)$	90%	95%	$\widehat{E}(\hat{\beta}_\theta)$	90%	95%
Structural Methods						
MOM	0.012	0.87	0.932	0.043	0.829	0.890
	(0.005)	(0.021)	(0.013)	(0.018)	(0.024)	(0.019)
MLE	-0.002	0.890	0.951	-0.001	0.893	0.950
	(0.011)	(0.024)	(0.016)	(0.011)	(0.023)	(0.016)
BAYES	-0.010	0.921	0.955	0.005	0.899	0.958
	(0.009)	(0.019)	(0.015)	(0.008)	(0.019)	(0.013)
Least Squares						
OLS	-0.414	0.580	0.724	-0.480	0.536	0.612
	(0.027)	(0.031)	(0.028)	(0.021)	(0.031)	(0.031)
RE	-0.155	0.596	0.784	-0.114	0.716	0.852
	(0.007)	(0.031)	(0.026)	(-0.006)	(0.0280	(0.022)

estimates for β_θ with both normal and binomial errors have mean greater than 0.4, and not significantly different from 0.453, the value we predicted in using the analytic methods in Section 3.3. Thus here we have evidence demonstrating the validity of Theorem 1. We find this result holds for all simulation we present here and in the next chapter. The RE procedure has smaller bias that the OLS procedure, and that too empirically verifies the claim made in Section 3.4, that RE method admits less bias than OLS estimates because the treatment weights are correlated with the population risk weights.

We also draw attention to the fact that the bias we find when we simulate with normal errors are very similar to what we find when we simulate with binomial errors. We find this to be true for all simulations we do here and in the next chapter, and so we have reason to believe that the analytic bias results given in (3.15) and (3.15) are not too far off when we have binomial error.

The top half Table 4.1 summarizes the estimates from the structural procedures. First notice that the likelihood methods perform well with both normal and binomial measurement error, each giving approximately unbiased estimates and truthful confidence intervals. The MOM estimates may slightly over adjust the observed regression line with normal error, and a bit more with binomial error, leading to a positive estimate for β_θ, and confidence intervals falling just shy of their nominal values. Otherwise, the MOM method provides a substantial reduction in bias over from least squares methods.

It may be surprising to find the likelihood procedures performing well with binomial measurement error. Examining the streptokinase data can understand why (the complete data can be found on page 48). Although many of the streptokinase trials have small size (n_i) with few events, many have substantial size. In particular 55% of the trials have over 50 subjects in both the treatment and control groups, and 45% of the trials have over 100 subjects in each group. Nearly every one of these large trials contain five or more events, and so certainly the normal measurement error approximation holds for the large trials. The smaller trials will not have normal measurement error but the random effects model weights the smaller trials less than the larger trials, thus diminishing their influence. We hypothesize that the presence of the large trials gives these procedures their good operating characteristics. The next section investigates this hypothesis.

4.5.2 Subsample of streptokinase trials

We take the smallest nine observations from the streptokinase trials for a simulation study (see the trials marked with ^{a} in Table 3.4 on page 3.4). We do this not only to determine if the performance we find with the complete streptokinase data is attributable to the several large trials, but also because in practice meta-analyses are commonly conducted with few small trials. The magnesium data we introduced in Chapter 1 provides a good example of this. For any method to be practically useful it must perform well for data with

Table 4.2: Simulation results from the smallest nine streptokinase trials. Each estimate gives the empirical mean of its estimates for β_θ and the 90% and 95% confidence intervals. The values in parentheses give the standard error of the estimates.

	Gaussian Variation			Binomial Variation		
Method/Model	$\hat{E}(\hat{\beta}_\theta)$	90%	95%	$\hat{E}(\hat{\beta}_\theta)$	90%	95%
[a]MOM	0.040	0.82	0.910	0.364	0.844	0.932
	(0.037)	(0.024)	(0.018)	(0.043)	(0.023)	(0.016)
[b]MLE	-0.203	0.936	0.976	-0.407	0.979	1
	(0.055)	(0.014)	(0.006)	(0.043)	(0.008)	NA
Bayes	-0.0400	0.92	0.98	-0.631	0.987	1
	(0.072)	(0.017)	(0.009)	(0.067)	(0.007)	NA
Least Squares						
OLS	-0.639	0.50	0.058	-0.612	0.372	0.532
	0.020	0.031	0.030	0.018	0.030	0.031
RE	-0.613	0.616	0.808	-0.565	0.447	0.928
	0.022	0.031	0.025	0.019	0.025	0.016

[a]: 18% percent estimated $\tau_x \leq 0$.

[b]: 8% percent estimated $\tau_x = 0$.

these characteristics.

Table 4.2 summarizes the simulations, and shows that the least squares methods perform worse that with the full data set. The results of Chapter 3 suggest that this should be the case, because removing the largest trials, which have smallest σ^2_{xi}, increases $\overline{\mathrm{B}}_x^{ls}$, and so introduces more bias. For these data we have $\overline{\mathrm{B}}_x^{ols} = 0.718$, $\bar{\beta}_w^{ols} = -0.942$, and so Theorem 1 from Chapter 2 predicts that we should observe a regression line with slope equal to $(0.718)(-0.942) = -0.677$. The empirical value, -0.639, falls within two standard errors of this value.

With normal measurement error, only the Bayes estimates perform well, appearing approximately unbiased with truthful confidence intervals. The maximum likelihood estimates over-adjust the regression line, and also estimates $\tau_x = 0$ approximately 8% of the time, giving undefined estimates of β_θ. This can be expected, because with small samples maximum likelihood procedures tend to underestimate variance components (for example see Morris, 1995; Everson, 1995). The method of moments estimates appear approximately unbiased with normal measurement error but its intervals undercover.

Every structural model procedure fails when errors are binomially distributed. In fact, comparing these results with the linear model estimates we find that the Bayes procedures may actually **increase** the bias compared with the RE model. Clearly the normal model has limited use for data sets that have measurement error dominated by binomial distributions.

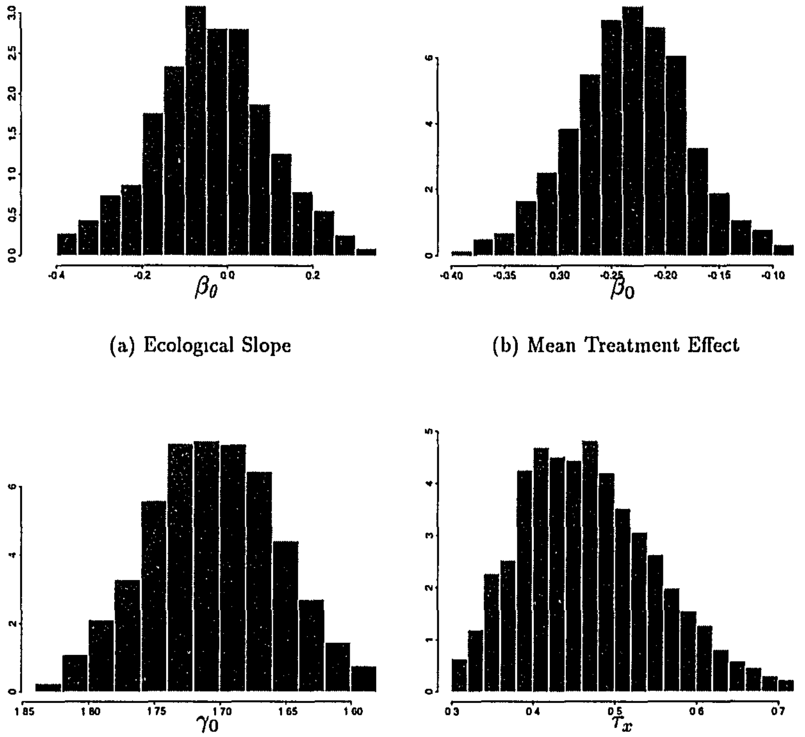
4.6 Summary and Conclusions

Chapter 3 uses a bivariate normal hierarchical model to quantify the biasing affects when population risks are used as covariates to explain clinical trial heterogeneity. This present chapter proposed three procedures to estimate the parameters of that model. The simulation evaluations in the previous section suggest that when measurement errors have true normal distribution, then the Bayes procedures perform well, giving approximately unbiased point estimates and valid coverage of its 90% and 95% confidence intervals. With a large number of trials, the maximum likelihood procedure perform similarly. The method of moments procedure eliminates a substantial amount of bias, but its coverage properties do not perform as well as the likelihood based procedures.

Most commonly clinical researcher use mortality rates to measure treatment and control outcomes, and so they have binomial distribution. To use the model and estimating procedures given in Chapter 3 and Chapter 4, we must approximate the joint distribution of treatment effects and population risk with a normal distribution. This is only an approximation, and the simulation results show that when the approximation fails for a substantial proportion of the trials, the procedures of this chapter perform poorly and may even increase the bias. The simulation results did show that that the quantity $\overline{\mathrm{B}}_x^{ls}$, which we use to quantify the bias, appears approximately valid even when errors are nonormal.

With small $\tau_{y|x}$ the likelihood procedures give large weight to the large trials. Because the large trials are more likely to have normally distributed measurement error then with small $\tau_{y|x}$ the normal procedures can tolerate some trials with nonormal error yet maintain good operating characteristics. However, similar data having larger $\tau_{y|x}$ cannot be adequately analyzed with these procedures. In general we cannot consider these procedures as valid when some of the clinical trials in a meta-analysis have few events.

We now recall the magnesium data introduced in Chapter 1 (see data on page 8). Concluding that $\beta_0 < 0$ for those data leads to important policy decisions. But that data has only 9 clinical trials and 4 of the treatment groups have only 1 death. Clearly the normal model procedure cannot be treated as valid for data of this type. The next chapter derives estimating procedures that we may use for analyzing these data.



(a) Ecological Slope

(b) Mean Treatment Effect

(c) Mean Population Risk

(d) Population Risk Standard Deviation

Figure 4.2: Histogram of marginal posterior distributions for streptokinase data, created from 5,000 draws from the data augmentation algorithm, with priors uniform on β_θ, β_0, γ_0, $\tau_{y|x}$ and τ_x^2.

Table 4.3: Results from streptokinase data analysis, with columns: β_θ, the structural model parameter; β_0, the mean treatment effect; $\tau_{y|x}$, the heterogeneity remaining after controlling for population risk; γ_0, the mean population risk; τ_x, the heterogeneity of the population risk; p-val/prob, the 1-sided p-value testing $H_0 : \beta_\theta = 0$ versus $H_a : \beta_\theta < 0$ (for MLE, MOM, or Random effects estimates), or the posterior probability $P(\beta_\theta > 0 \mid \hat{\theta})$ (for Bayes estimates). Each method provides point estimates and standard errors, and underneath them, 95% interval estimates. All intervals except the Bayes give 95% intervals based on a normal approximation. The Bayes procedure gives intervals computed from the upper and lower 2.5% quantiles of the posterior distribution. The posterior distribution is computed from 5,000 draws from the data augmentation algorithm using priors uniform on β_θ, β_0, γ_0, $\tau_{y|x}$, and τ_x^2.

	Ecological Model			Covariate Model		
Model/Method	β_θ	β_0	$\tau_{y\|x}$	γ_0	τ_x	p-val/prob
Bayes						
estimate(s.e.)	-0.040(0.147)	-.223(0.059)	0.144	-1.705(0.053)	0.470	0.384
interval	(-0.346,0.252)	(-0.356,-0.113)		(-1.806,-1.600)		
MLE						
estimate(s.e.)	0.034(0.103)	-0.222(0.034)	0	-1.701(0.090)	0.412	0.692
interval	(-0.167,0.236)	(-0.288,-0.155)		(-1.8774,-1.534)		
MOM						
estimate(s.e.)	0.028(0.320)	-0.244(0.036)	0	-1.726(0.098)	0.359	0.534
interval	(-0.599,0.655)	(-0.315,-0.173)		(-1.918,-1.533)		
Random Effects						
estimate(s.e.)	-0.178(0.095)	-0.235(0.050)	0	NA	NA	0.036
interval	(-0.364,0.008)	(-333,-.137)				

Chapter 5

Small Sample Inference

The simulation results of Chapter 4 show the large sample model performs poorly with small within study samples, due to the poorly approximated measurement error distribution. The model and estimation procedures we develop here intend to work well in that instance, but we also extend the ecological model. This chapter first lists requirements that a model and estimating procedure must meet for it to be applicable to a wide range of clinical research, then proposes a model that intends to meet those requirements. The model we derive naturally extends the hierarchical model derived in Chapter 3, and the estimating procedures naturally extend the data augmentation procedure of Chapter 4. We then demonstrate the features of the model on the magnesium data, providing convincing evidence for the existence of a structural slope. Finally, we investigate the frequency properties of this small sample procedure and provide evidence that it has valid frequency properties in instances that the large sample model fails.

5.1 Requirements of a General Model

For a model and inference procedure to be widely applicable, it should have the following features

- Allow a variety of measurement error models: The streptokinase and magnesium data have binomial outcomes $\hat{\mu}_i$, but other measurement distributions are possible, including Poisson (for failure rate data), normal (when treatment effects have continuous outcomes), and perhaps gamma (for combining estimates of variance).

- Allow general definitions of treatment effect and population risk: Different medical specialties prefer different measures of treatment efficacy (for example see Sinclair and Bracken, 1994), including relative risks, risk difference, log-odds ratios and other functions of them. A health professional should be allowed to determine the scale of the treatment effect and population risk.

- Allow complex ecological models: The structural model defined by Chapter 3 requires a treatment effect to be linear in the population risk. More complex associations should be allowed.

- Valid frequency inferences for or large within study and between study samples: It is typical for meta-analyses to contain few trials (k) with small samples sizes (n_i) in each trial. Inference procedures must have good frequency properties in these circumstances.

The model we propose next accommodates each of these requirements. The next two sections construct a hierarchical model and derives a Bayesian inference procedure. Much of the work has analogies to the derivations given in Chapter 3 and Chapter 4.

5.2 The Model

As with Chapter 4, we represent the observed and unobserved quantities in a hierarchal model, with first stage representing the measurement error model, and second stage representing the structural model. Here, instead of defining the measurement error distribution on $\hat{\theta}_i = (\hat{\theta}_{yi}, \hat{\theta}_{xi})'$, which we can only approximate, we treat the exact measurement error on the raw outcomes, $\hat{\mu}_i = (\hat{\mu}_{ti}, \hat{\mu}_{ci})'$. We also change the structural model so that a richer variety of ecological models can be defined.

The measurement error model

The large sample hierarchical model of Chapter 3 can approximate the distribution of $\hat{\theta}_i = \theta(\hat{\mu}_i)$ by a delta method because $\hat{\mu}_i$ has a known distribution, but otherwise does not use the parametric form of its distribution. Here we assume $\hat{\mu}_i$ follow a Natural Exponential Family with Quadratic Variance Function (NEF-QVF) (see Morris, 1982, 1983a, for extensive results concerning the properties of these distributions). That is, we now assume (2.1) and (2.2) on page 12 equals

$$p(\hat{\mu}_{ti} \mid \mu_{ti}) = \text{NEF-QVF}\left[\mu_{ti}, \frac{V(\mu_{ti})}{n_{ti}}\right] \qquad i = 1 \cdots k \tag{5.1}$$

$$p(\hat{\mu}_{ci} \mid \mu_{ci}) = \text{NEF-QVF}\left[\mu_{ci}, \frac{V(\mu_{ci})}{n_{ci}}\right] \qquad i = 1 \cdots k \tag{5.2}$$

where we assume $V(\mu) = v_2\mu^2 + v_2\mu + v_0$. We denote the joint distribution of $\hat{\mu}_{ti}$ and $\hat{\mu}_{ci}$ by $p(\hat{\mu}_i \mid \mu_i)$, and because $\hat{\mu}_{ti}$ and $\hat{\mu}_{ci}$ are independent then $p(\hat{\mu}_i \mid \mu_i) = p(\hat{\mu}_{ti} \mid \mu_{ti})p(\hat{\mu}_{ti} \mid \mu_i)$.

Many distributions commonly used in clinical research follow an NEF-QVF, perhaps the most useful being the normal ($V(\mu) = 1$), binomial ($V(\mu) = \mu(1-\mu)$), and the Poisson ($V(\mu) = \mu$), but the gamma ($V(\mu) = \mu^2$), negative binomial ($\mu = \mu^2$), and a sixth distribution Morris calls the NEF-GHS (GHS for generalized hyperbolic secant) are also NEF-QVF. Table 5.1 gives a selection of some NEF-QVF family members (see Morris, 1988, for a more complete list).

The NEF-QVF distributions have convenient mathematical properties that allow us to

Table 5.1: Some NEF-QVF distributions, with their means μ, their variance $V(\mu)/n$, natural parameter ϑ, cumulants generating function, $\Psi(\vartheta)$, and conjugate distribution.

$p(\hat{\mu} \mid \mu)$	μ	$V(\mu)/n$	ϑ	$\Psi(\vartheta)$	Conjugate=PD[$\mu_0, r; V(\mu)$]
$\frac{1}{n}$Bin(n,p)	p	$\frac{p(1-p)}{n}$	$\log\left(\frac{p}{1-p}\right)$	$\log(1+e^{\vartheta})$	Beta:$p^{r\mu_o-1}(1-p)^{r(1-\mu_0)-1}$
$\frac{1}{n}$Poiss(λ)	λ	λ/n	$\log(\lambda)$	e^{ϑ}	Gamma : $\lambda^{r\mu_0}e^{-r\mu_0}$
$\frac{\sigma}{n}N(\mu,1)$	μ	σ^2/n	μ	$\frac{\mu^2}{2}$	Normal : $\sqrt{r}e^{-\frac{r}{2}(\mu-\mu_0)^2}$

treat all the NEF-QVF family members without any more difficulty than if we treat only one of them.

The ecological and structural model

As with Chapter 4, we define the structural model as a product of the ecological model and the population risk model, but defined here as

$$p(\theta_{yi} \mid \theta_{xi}, \phi) = N\left(Z_i'\beta_z + \chi_i'\beta_0, \tau_{y|x}^2\right) \tag{5.3}$$

$$p(\theta_{xi} \mid \phi) = N\left(Z_i'\gamma_z, \tau_x^2\right) \tag{5.4}$$

where $\phi = (\beta_z', \beta_0', \gamma_z', \tau_{y|x}, \tau_x)$.

We use $\chi_i = \chi(\theta_{xi}, Z_i)$ to denote a known one-to-one function of θ_{xi} and Z_i. We allow χ_i to be vector valued so the ecological slope β_0 may be vector valued as well. We give examples of its use below, but first notice that we may make (5.3) equal to the structural model of Chapter 3 by choosing $\chi_i(\theta_{xi}, Z_i) = \theta_{xi}$. The population risk model (5.4) remains unchanged from Chapter 3.

We define the structural model as the product of the ecological and population risk model by

$$p(\theta_{yi}, \theta_{xi} \mid \phi)d\theta_i = p(\theta_{yi} \mid \theta_{xi}, \phi)p(\theta_{xi} \mid \phi)d\theta_i \tag{5.5}$$

We make the measure element $d\theta_i$ explicit to remind us that (5.5) defines a density on θ_i. At times we find it useful to view (5.5) as a density on μ_i, and the measure element will help us distinguish these two cases.

The ecological model we defined has many features that accommodate the requirements listed in Section 5.1. We give a few examples to illustrate them.

(1) Schmid *et al.* (1995), Lau *et al.* (1995), and Antman (1995b) wish to estimate the association $\log\left(\frac{p_{ti}}{p_{ci}}\right) = \beta_0 + \beta_\theta p_{ci}$. To use this ecological specification the model of Chapter 3 requires us to treat p_{ci} as having a normal distribution. This creates difficulty because $0 < p_{ci} < 1$, but the normal distribution does not enforce this, and it is possible for the data augmentation algorithm to simulate values of p_{ci} out of this range. A better specification treats $\theta_{xi} = \log\left(\frac{p_{ci}}{1-p_{ci}}\right)$ as having a normal distribution. The present model can accomplish this without changing the structural model by choosing $\theta_{xi} = \log\left(\frac{p_{ci}}{1-p_{ci}}\right)$ and $\chi_i = \frac{e^{\theta_{xi}}}{1+e^{\theta_{xi}}}$.

(2) If we let Z_{ij} represent the jth component of Z_i, then we may include an interaction term in the ecological model by choosing $\chi_i = (\theta_{xi}, Z_{ij}\theta_{xi})'$. In this instance β_θ has dimension 2.

(3) We may include a quadratic function by choosing $\chi_i = (\theta_{xi}, \theta_{xi}^2)$.

Section 5.4.1 uses specifications (1) and (3) when analyzing the magnesium data.

Recall that Chapter 3 represented the structural model as a bivariate normal distribution. Here we may represent (5.5) by a bivariate normal distribution only when $\chi_i = \theta_{xi}$. Otherwise, the structural model has no simpler form other than (5.5).

Joint distribution

Chapter 4 found the joint distribution $p(\hat{\theta}_i, \theta_i \mid \phi)$ useful when deriving estimation procedures. Here we proceed in a similar manner, except that we specify the joint distribution $p(\hat{\mu}_i, \mu_i \mid \phi)$.

The structural model (5.5) defines a density on θ_i, but observations from the measurement model provides a likelihood in terms of μ_i. So that they conform we must either parameterize the measurement model in terms of θ_i, or transform the structural model to be a density with respect to $d\mu_i$. We choose to do the latter for reasons that become apparent when we treat estimation.

To transform (5.5) to a distribution on μ_i, we substitute $\theta(\mu_i)$ for θ_i and change the measure element from $d\theta_i$ to $\frac{d\theta(\mu_i)}{d\mu_i} =\mid J(\mu_i) \mid d\mu_i$. Here $J(\mu_i)$ defines the Jacobian of the transformation of μ_i, and equals the Jacobian we use for the large sample variance approximation in Chapter 3 (see expression (3.3) on page 26). Thus we write the joint distribution of the observed $\hat{\mu}_i$ and unobserved μ_i as

$$p(\hat{\mu}_i, \mu_i \mid \phi) = \underbrace{p(\hat{\mu}_i \mid \mu_i)}_{(a)} \underbrace{p(\theta_y(\mu_i) \mid \theta_x(\mu_i), \phi) p(\theta_x(\mu_i) \mid \phi)}_{(b)} \underbrace{\mid J(\mu_i) \mid}_{(c)} \tag{5.6}$$

Expression (5.6) shows that $\hat{\mu}_i$ and μ_i have joint distribution defined as a product of two NEF-QVF distributions (expression with label (a)), multiplied by the product of two regression models (expression with label (b)), and then multiplied by the measure element (expression with label (c)).

5.3 Inference

Chapter 3 uses $p(\hat{\theta}_i, \theta_i|\phi)$ to derive the marginal distribution $p(\hat{\theta}_i \mid \phi)$ and the conditional distribution $p(\theta_i|\hat{\theta}_i, \phi)$. Here we express the analogous marginal distribution as $p(\hat{\mu}_i \mid \phi) = \int p(\hat{\mu}_i, \mu_i \mid \phi) d\mu_i$, which has no simple expression if we allow a general choice of NEF-QVF and $\theta(\mu_i)$, and the conditional distribution $p(\mu_i|\hat{\mu}_i, \phi)$ does not have the form of a known density. Because of this the method of moment and maximum likelihood procedures given in Chapter 3, which require closed form expressions for marginal and conditional distributions, cannot be applied. We do find that we may extend the Bayes methods to the present circumstance.

As with the Bayes procedure of Section 4.2.2 we investigate the posterior distribution of $\boldsymbol{\phi}$, which we denote by $p(\boldsymbol{\phi} \mid \hat{\boldsymbol{\mu}})$, indirectly through the joint posterior distribution $p(\boldsymbol{\phi}, \boldsymbol{\mu} \mid \hat{\boldsymbol{\mu}})$, by treating $\boldsymbol{\mu}$ as missing data. Recall from Chapter 3 that we use the notation $\boldsymbol{\mu}_t = (\mu_{t1}, \mu_{t2}, \cdots, \mu_{tk})$, $\boldsymbol{\mu}_c = (\mu_{c1}, \mu_{c2}, \cdots, \mu_{ck})$, and $\boldsymbol{\mu} = (\boldsymbol{\mu}_t, \boldsymbol{\mu}_c)$.

The joint posterior distribution of $\boldsymbol{\phi}$ and the unobserved means $\boldsymbol{\mu}$ resulting from k observations from (5.6) and a prior $p(\boldsymbol{\phi})$ is

$$
\begin{aligned}
& p(\boldsymbol{\mu}, \boldsymbol{\phi} \mid \hat{\boldsymbol{\mu}}) && (5.7)\\
\propto \;& p(\boldsymbol{\phi}) \underbrace{\prod_{i=1}^{k} p(\hat{\mu}_{ti} \mid \mu_{ti}) p(\hat{\mu}_{ci} \mid \mu_{ci})}_{\text{(a)}} \underbrace{\prod_{i=1}^{k} p(\theta_y(\mu_i) \mid \theta_x(\mu_i), \boldsymbol{\phi}) p(\theta_x(\mu_i) \mid \boldsymbol{\phi})}_{\text{(b)}} \underbrace{\prod_{i=1}^{k} J(\mu_i)}_{\text{(c)}} && (5.8)\\
\propto \;& p(\boldsymbol{\phi}) \underbrace{p(\hat{\boldsymbol{\mu}} \mid \boldsymbol{\mu})}_{\text{(a)}} \underbrace{p(\boldsymbol{\theta}_y(\boldsymbol{\mu}) \mid \boldsymbol{\theta}_x(\boldsymbol{\mu}), \boldsymbol{\phi}) p(\boldsymbol{\theta}_x(\boldsymbol{\mu}), \boldsymbol{\phi})}_{\text{(c)}} \underbrace{\mathbf{J}(\boldsymbol{\mu})}_{\text{(c)}} && (5.9)
\end{aligned}
$$

5.3.1 The data augmentation procedure

The data augmentation procedure given by Chapter 4 makes inferences for $\boldsymbol{\phi}$ by alternating between an augmentation step and a parameter step. In that procedure the parameter step simulates values of $\boldsymbol{\phi}$ from $p(\boldsymbol{\phi}|\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ as if $\boldsymbol{\theta}$ were known, then the augmentation step simulates $\boldsymbol{\theta}$ from $p(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}, \boldsymbol{\phi})$ as if $\boldsymbol{\phi}$ were known. Both steps are simple because the hierarchical model summarized in Table 3.1 on page 3.1 gives a normally distributed augmentation step. We use data augmentation to estimate the present model, but here the augmentation step simulates $\boldsymbol{\mu}$. However, because of the difficult form (5.9) we find that the augmentation steps do not follow a known distribution, and so direct simulation is difficult. We first describe the data augmentation procedure as it would proceed if we could perform the augmentation steps directly, then describe a method to perform it indirectly.

Letting $\boldsymbol{\mu} = \boldsymbol{\mu}^{(n-1)} = (\boldsymbol{\mu}_t^{(n-1)}, \boldsymbol{\mu}_c^{(n-1)})'$ represent the current values of the unobserved means $\boldsymbol{\mu}$, in brief notation we perform the data augmentation algorithm by alternating through the following steps.

$$\boldsymbol{\theta}^{(n)} \sim p(\boldsymbol{\theta}|\boldsymbol{\mu} = \boldsymbol{\mu}^{(n-1)}, \hat{\boldsymbol{\mu}}) \tag{5.10}$$

$$\boldsymbol{\mu}_t^{(n)} \sim p(\boldsymbol{\mu}_t|\boldsymbol{\mu}_c = \boldsymbol{\mu}_c^{(n-1)}, \boldsymbol{\phi} = \boldsymbol{\phi}^{(n)}) \tag{5.11}$$

$$\boldsymbol{\mu}_c^{(n)} \sim p(\boldsymbol{\mu}_c|\boldsymbol{\mu}_t = \boldsymbol{\mu}_t^{(n)}, \boldsymbol{\phi} = \boldsymbol{\phi}^{(n)}) \tag{5.12}$$

Expression (5.10) represents the parameter step and (5.11) and (5.12) define the augmentation steps. We divide the augmentation step into two steps so that they simulate from univariate distributions, making the algorithm we present in the next section simpler. In detail the parameter and augmentation steps are:

Parameter Step:

With current mean values $\boldsymbol{\mu}^{(n-1)}$, update the structural model parameters by

$$\begin{aligned} \boldsymbol{\phi}^{(n)} &\sim p(\boldsymbol{\phi} \mid \boldsymbol{\mu} = \boldsymbol{\mu}^{(n-1)}, \hat{\boldsymbol{\mu}}) && (5.13)\\ &= p(\boldsymbol{\phi} \mid \boldsymbol{\theta}(\boldsymbol{\mu}) = \boldsymbol{\theta}(\boldsymbol{\mu}^{(n-1)}), \hat{\boldsymbol{\mu}}) && (5.14)\\ &\propto p(\boldsymbol{\phi})p(\boldsymbol{\theta}_y^{(n-1)} \mid \boldsymbol{\theta}_x^{(n-1)}, \boldsymbol{\phi})p(\boldsymbol{\theta}_x^{(n-1)} \mid \boldsymbol{\phi}) && (5.15) \end{aligned}$$

Expression (5.14) follows because we restrict θ_i and μ_i to be one to one functions. Comparing the parameter step (5.15) to the parameter step (4.18) on page 57 we find them to be equal and so we perform the parameter step by the methods derived in Section 4.2.2. □

Augmentation Step

With current parameter estimate $\boldsymbol{\phi}^{(n)}$, simulate the missing means in two steps by

(1) For $i = 1 \cdots k$, impute the missing treatment μ_{ti} by

$$\begin{aligned} \mu_{ti}^{(n)} &\sim p(\mu_{ti} \mid \hat{\mu}_i, \mu_{ci} = \mu_{ci}^{(n-1)}, \boldsymbol{\phi} = \boldsymbol{\phi}^{(n)}) && (5.16)\\ &\propto p(\hat{\mu}_{ti} \mid \mu_{ti})p(\theta_y(\mu_{ti}, \mu_{ci}^{(n-1)}) \mid \theta_x(\mu_{ti}, \mu_{ci}^{(n-1)}), \boldsymbol{\phi}^{(n)}) \mid J(\mu_{ti}, \mu_{ci}^{(n-1)}) \mid && (5.17) \end{aligned}$$

(2) For $i = 1 \cdots k$ impute the missing control means μ_{ci} by

$$\mu_{ci}^{(n)} \sim p(\mu_{ci} \mid \hat{\mu}_i, \mu_{ti} = \mu_{ti}^{(n)}, \phi = \phi^{(n)}) \tag{5.18}$$

$$\propto p(\hat{\mu}_{ci} \mid \mu_{ci}) p(\theta_y(\mu_{ti}^{(n)}, \mu_{ci}) \mid \theta_x(\mu_{ti}^{(n)}, \mu_{ci}), \phi^{(n)}) \mid J(\mu_{ti}^{(n)}, \mu_{ci}) \mid \tag{5.19}$$

□

If we could simulate from densities proportional to (5.17) and (5.19) then alternating between the parameter and augmentation steps generates a sequence $\phi^{(n)}, \phi^{(n)} \cdots \phi^{(n)}$ that converges to the posterior distribution $p(\phi \mid \hat{\mu})$. Because those densities do not represent any named distribution, direct simulation is not possible without special effort.

One computationally intensive solution approximates the augmentation distributions by a discrete distribution, or grid, and treats simulations from that approximation as if they were exact. This procedure has the colorful name "the Griddy Gibbs" algorithm (Ritter and Tanner, 1992). We will find that we need several thousand iterations for our data augmentation procedure to converge, and because we must perform the augmentation steps $2k$ times at each iteration, any method that is too computationally intensive can be impractical. Because we do not use the data augmentation algorithm to make inferences for μ, but rather as a device to make inferences for ϕ, we may wish to sacrifice accuracy of augmentation step for computational efficiency.

5.3.2 Metropolised data augmentation procedure

Although we find the augmentation steps difficult to perform exactly, we find that we can form good approximations. The Metropolis-Hastings algorithm (Hastings, 1970) is an MCMC algorithm and general form of the Gibbs sampler that allows the steps to be performed approximately (for a tuorial on these methods see Gelman *et al.*, 1995, , Chapter 11, page 320). Here we will use their algorithm in the augmentation step. We outline the algorithm here, and discuss making the approximation in the section that follows.

Before we describe the Metropolis algorithm, we give a few preliminaries. If we let $p(\cdot)$ represent a true distribution, then we use $\hat{p}(\cdot)$ to denote a distribution that approximates it. We also define the *importance weights* by

$$I_{p,\hat{p}}(\mu_1 \mid \mu_2, \phi) = \frac{p(\mu_1 \mid \mu_2, \phi)}{\hat{p}(\mu_1 \mid \mu_2, \phi)} \tag{5.20}$$

Intuitively, the importance weights give the relative density of μ_1 under the true density $p(\cdot)$ compared to the approximate density $\hat{p}(\cdot)$ when both are conditioned on μ_2 and ϕ. For example, if $I_{p,\hat{p}}(\mu_1 = 4.2 \mid \mu_2, \phi) = 3$, then μ_1 is 3 times more likely to take on the value 4.2 with the true density than with the approximate density.

In words, the Metropolised augmentation step uses the importance weights as follows. We treat updating μ_{ti} for now. At each step we approximate the augmentation distribution by a known distribution and simulate a candidate $\tilde{\mu}_{ti}$. We then compare the importance weight of the candidate with the importance weight of the current observation $\hat{\mu}_{ti}^{(n-1)}$. If their ratio exceeds 1, then we set $\mu_{ti}^{(n)} = \tilde{\mu}_{ti}$. Otherwise, we either choose $\mu_{ti}^{(n)} = \tilde{\mu}_{ti}$ or $\mu_{ti}^{(n)} = \mu_{ti}^{(n-1)}$, with probability that depends on the ratio of the importance weights. In notation the steps are

Metropolised Augmentation Step:

(1) Impute $\mu_{ti}^{(n)}$ for $i = 1 \cdots k$ by

First draw a candidate value for $\mu_{ti}^{(n)}$ by

$$\tilde{\mu}_{ti} \sim \hat{p}(\mu_{ti} \mid \mu_{ci} = \mu_{ci}^{(n-1)}, \phi = \phi^{(n)}) \tag{5.21}$$

then update $\mu_{ti}^{(n)}$ by

$$\mu_{ti}^{(n)} = \begin{cases} \tilde{\mu}_{ti} & \text{with probability } \min\left(\frac{I_{p,\hat{p}}(\tilde{\mu}_{ti} | \mu_{ci}^{(n-1)}, \phi^{(n)})}{I_{p,\hat{p}}(\mu_{ti}^{(n-1)} | \mu_{ci}^{(n-1)}, \phi^{(n)})}, 1 \right) \\ \mu^{(n-1)} & \text{otherwise} \end{cases} \tag{5.22}$$

(2) Impute $\mu_{ci}^{(n)}$ for $i = 1 \cdots k$.

$$\tilde{\mu}_{ci} \sim \hat{p}(\mu_{ci} \mid \mu_{ti} = \mu_{ti}^{(n+1)}, \phi = \phi^{(n)}) \tag{5.23}$$

$$\mu_{ci}^{(n)} = \begin{cases} \tilde{\mu}_{ci} & \text{with probability } \min\left(\frac{I_{p,\hat{p}}(\tilde{\mu}_{ci} | \mu_{ti}^{(n)}, \phi^{(n)})}{I_{p,\hat{p}}(\mu_{ci}^{(n-1)} | \mu_{ti}^{(n)}, \phi^{(n)})}, 1 \right) \\ \mu_{ci}^{(n-1)} & \text{otherwise} \end{cases} \tag{5.24}$$

□

Notice that if $\hat{p}(\cdot) = p(\cdot)$, so that the approximating density equals the exact density, then $I_{p,\hat{p}}(\cdot) = 1$, and the Metropolis algorithm always accepts the candidate. In this instance the Metropolised algorithm equals the exact augmentation algorithm, and so we see that the Metropolis algorithm generalizes the Gibbs sampler.

The Metropolis algorithm has an advantage if accurate approximating densities can be formed quickly. With poor approximations the ratio of importance weights will often be small, and the Metropolis algorithm converges slowly. Efficient implementation requires accurate approximations so that the importance ratios are as close to 1 as possible.

5.3.3 Density Approximations Based on Pearson Densities

So far we have described how we perform an MCMC estimation procedure if we have approximations to (5.17) and (5.19). Here we draw attention to creating approximations based on fitting the modes of (5.17) and (5.19) to known distributions. Because we have univariate augmentation steps either Newton's method or computational searching can find modes quickly and simply. We first discuss a few preliminaries.

We need to approximate densities that have form similar to the right hand side of (5.17) and (5.19). For clarity, in the discussions that follow we drop the indices and use $p_n(\mu)d\mu$ to denote the right hand sides of (5.17) or (5.19) at the n^{th} step of the data augmentation procedure. We also represent the log density by $\ell_n(\mu) = \log(p_n(\mu))$.

The most common way to approximate densities matches the first and second derivatives of $\ell_n(\mu)$ to a normal distribution. If $\ell'_n(\mu_0) = 0$, so that $\ell_n(\mu)$ has mode μ_0, and letting $\sigma^2 = -1/\ell''(\mu_0)$, then $\hat{p}_n(\mu_t) = N(\mu_0, \sigma^2)$ approximates $p_n(\mu)$. For our application this approximation performs poorly. Recall that we define $p_n(\mu)$ as the product of an NEF-QVF likelihood (expression with label (a) in (5.6)), multiplied by a transformation of two normal densities (expression with labels (b) and (c) in (5.6)). These distributions may have skew, a feature the normal distribution cannot accommodate.

Morris (1988) describes a procedure that generalizes the normal approximation by using general Pearson families to approximate univariate distributions. The Pearson family contains a large number of named distributions, including the beta, gamma, normal, F, and others, and many have skew. Before we describe the approximating procedure, we first give some preliminary facts about NEF-QVF distributions and Pearson families. For the results we outline next see Morris (1982, 1983a, 1988) for detailed references.

NEF-QVF Distributions

If a random quantity $\hat{\mu}$ has NEF-QVF$[\mu, V(\mu)/n]$ distribution, its density may be written as

$$p(\hat{\mu} \mid \mu) = \exp\left(n\hat{\mu}\vartheta - n\Psi(\vartheta)\right) h_n(\hat{\mu}) \tag{5.25}$$

where the term $h_n(\hat{\mu})$ does not depend on μ. When $\hat{\mu}$ are discrete, then (5.25) defines a probability mass function, otherwise it defines a probability density function. The parameter ϑ, a one to one function of μ, is called the *natural parameter*. The function $\Psi(\vartheta)$ generates cumulants and has first derivative $\Psi'(\vartheta) = \mu$, the mean of $\hat{\mu}$, and second derivative $\Psi''(\vartheta) = V(\mu)$, so that $\text{Var}(\hat{\mu}) = \Psi''(\vartheta)/n$. With large n, (5.25) has an approximate normal shape.

For example, if $n\hat{p}$ has binomial(n, p) distribution then $\hat{p}$ is a NEF-QVF$[p, p(1-p)/n]$, with $\vartheta = \log\left(\frac{p}{1-p}\right)$ and $\Psi(\vartheta) = \log(1 + e^{\vartheta})$. Notice that $\Psi'(\vartheta) = e^{\vartheta}/(1 + e^{\vartheta}) = p$. Table 5.1 gives the definitions of $\Psi(\cdot)$ and ϑ for some NEF-QVF distributions.

Pearson Distributions

A distribution that mimics (5.25) except that we treat it as a density on the parameter ϑ_* has the form

$$p(\vartheta_* \mid r, \mu_0) = K_{r,\mu_0} \exp(r\mu_0\vartheta_* - r\Psi_*(\vartheta_*)) \tag{5.26}$$

The constant K_{r,μ_0} normalizes the density to have unit integral. Random quantities with densities having the form (5.26) are related to random variables following Pearson distributions if $\Psi''_*(\vartheta_*)$ evaluated at the mode of (5.26) has a quadratic form in μ_0: $V_*(\mu_0) = v_2\mu_0 + v_1\mu_0 + v_0 > 0$. Although ϑ_* does not follow a Pearson distribution, the transformation $\mu_* = \mu_*(\vartheta_*) = \Psi'(\vartheta_*)$ does, and has mean μ_0 and variance $V_*(\mu_0)/(r - v_2)$. We also use $\vartheta_* = \vartheta_*(\mu) = \Psi'^{-1}_*(\mu)$ to denote the inverse of that transformation. The Pearson densities are characterized by the variance function V_* and we represent a quantity having a Pearson distribution with parameters r and μ_0 by $\mu_* \sim \mathrm{PD}(r, \mu_0; V_*)$.

The log of (5.26) has first and second derivatives expressed as

$$\begin{aligned} \ell'(\vartheta_*) &= r\mu_0 - r\Psi'_*(\vartheta_*) \\ &= r(\mu_0 - \mu_*) \end{aligned} \tag{5.27}$$

and

$$\begin{aligned} \ell''(\vartheta_*) &= -r\Psi''(\vartheta_*) \\ &= -rV_*(\mu_*) \end{aligned} \tag{5.28}$$

For convenience we express the derivatives above once in terms of ϑ_*, and then again in terms of μ_*. The first derivative shows that a Pearson density has mode at $\mu_* = \mu_0$ or equivalently, a mode at $\vartheta_* = \vartheta_*(\mu_0)$. The curvature at the mode (the second derivative) gets more negative as r increases, and importantly, as r gets large, then $\sqrt{r}(\mu_* - \mu_0)$ and $\sqrt{r}(\vartheta_* - \vartheta_0)$ converge in distribution to the normal distribution. The Pearson family is the conjugate to the NEF-QVF if $\Psi_*(\cdot)$ in (5.26) equals $\Psi(\cdot)$ in (5.25).

For example, if μ_* has a beta distribution, the conjugate to the binomial, having mean p_0 and variance $\frac{p_0(1-p_0)}{r-1}$, then $\vartheta_* = \Psi'(p) = \log\left(\frac{p}{1-p}\right)$ has distribution written as $\vartheta \sim k_{r,p_0}\exp(rp_0\vartheta_* - r\log(1+e^{\vartheta_*}))$. Table 5.1 gives the conjugate Pearson densities for some NEF-QVF members.

Computing an approximating density

Morris advocates using Pearson densities to approximate the mean of posterior distributions, μ, but here we will adapt his method and approximate the density of the natural parameter ϑ. The primary reason for doing this is that, because μ may have restricted range, a normal approximation for μ (like those described at the beginning of this section) does not give a good fit. For example, if $n\hat{\mu}$ has a binomial distribution then $0 < \mu < 1$. Natural parameters, however, do not have restricted range, and so normal approximations to ϑ work better than normal approximations to μ. For example, for binomial we have $-\infty \leq \vartheta = \log(\mu/(1-\mu)) \leq \infty$. In practice it is most common to approximate ϑ by a normal distribution, and so comparing the Pearson and normal approximation methods is fairest if we compare how they perform on ϑ. This does not cause difficulty for our augmentation steps because simulating a candidate ϑ, denoted $\tilde{\vartheta}$, can be used to form the candidate $\tilde{\mu}$ by $\tilde{\mu} = \mu(\tilde{\vartheta})$.

Thus we will transform the augmentation distributions (5.17) and (5.19) to define densities ϑ, the natural parameter of the NEF-QVF measurement error distribution. We denote that density by $p_n(\vartheta)d\vartheta$, and we derive it by transforming $p_n(\mu)d\mu$ to a density on $\vartheta = \vartheta(\mu)$, leading to

$$p_n(\vartheta)d\vartheta = p_n(\mu)d(\mu(\vartheta)) \tag{5.29}$$

$$= p_n(\mu)d(\Psi'(\vartheta)) \tag{5.30}$$

$$= p_n(\mu)V(\mu)d\vartheta \tag{5.31}$$

Recall that μ is an implicit function of ϑ, and so the above equation means that the probability density for ϑ equals the density $p_n(\mu)$ evaluated at $\mu(\vartheta)$, then multiplied by the measure element $V(\mu(\vartheta))$. We use $\ell_n(\vartheta)$ to denote $\log(p_n(\vartheta))$

We are now ready to approximate the candidate draw of the augmentation step. Morris recommends choosing an approximating Pearson density with the restriction that the range of ϑ_* be the same as the range of ϑ, then select r and μ_0 to equate the first and second derivatives of $\ell_n(\vartheta)$ to (5.27) and 5.28). If $\ell'(\vartheta_0) = 0$ so that $\ell_n(\vartheta)$ has mode ϑ_0 then equating the first and second derivatives (5.27) and (5.28) selects r and μ_0 as

$$\mu_0 = \mu_*(\vartheta_0) \quad \text{and} \quad r = -1/\ell''(\vartheta_0)V_*(\mu_*(\vartheta_0)) \tag{5.32}$$

Note that if we choose to approximate $p_n(\vartheta)$ with a normal distribution then $V_*(\mu) = 1$ and (5.32) reduces to the usual normal approximation for ϑ. In this sense Morris generalizes the usual approximation method.

Performing the candidate draw

To simulate the candidate $\tilde{\vartheta}$ we recognize that $\mu_* = \Psi'(\vartheta)$ has distribution $PD(r, \mu_0; V_*)$. Thus we simulate a candidate by $\tilde{\mu}_* \sim PD(r, \mu_0; V_*)$ and then compute $\tilde{\vartheta} = \vartheta_*(\tilde{\mu}_*)$. Recall however that our initial aim was to simulate the candidate $\tilde{\mu}$, the mean parameter. We complete the augmentation step by computing $\tilde{\mu} = \mu(\tilde{\vartheta})$.

To summarize, at each augmentation step we choose an approximating Pearson distribution characterized by $V_*(\cdot)$, then choose parameters r and μ_0 by (5.32). Now simulate $\tilde{\mu}_* \sim PD(r, \mu_0; V_*)$ and complete the candidate draw by

$$\tilde{\mu} = \mu(\vartheta_*(\tilde{\mu}_*)) \tag{5.33}$$

Expression (5.33) simplifies for particular choices of $V_*(\cdot)$. For example, if we choose the Pearson density to be the conjugate for the NEF-QVF then $V_* = V$ and the candidate draw (5.33) becomes $\tilde{\mu} = \tilde{\mu}_*$. If however we use the usual normal approximation then $V_* = 1$ and (5.33) becomes $\tilde{\mu} = \mu(\tilde{\mu}_*)$.

Although this may seem complicated it is very efficient to implement by computer and as we find next, gives very efficient approximations. Before we give examples and discuss choosing V_* we must first point out the correct approximate distribution with which to compute the importance ratio (5.20). The importance ratio is expressed with densities on μ, but (5.26) is expressed in terms of ϑ. To express this in terms of μ, we must multiply it by the change of measure $d\vartheta/d\mu = 1/\Psi''(\vartheta)$. Thus in the denominator of (5.20) we use

$$\hat{p}(\tilde{\mu}_{ti} \mid \mu_{ci} = \mu_{ci}^{(n-1)}, \phi = \phi^{(n)}) = \exp(r\mu_0\tilde{\vartheta} - r\Psi_*(\tilde{\vartheta}))/\Psi''(\tilde{\vartheta}) \qquad (5.34)$$

5.3.4 Examples

There may be many possible Pearson families to choose from when forming the approximating density. To help make that decision Morris gives two diagnostics that require evaluating $\ell_n'''(\vartheta)$ at the mode, and so requires additional computation. We find that although his diagnostics do lead to the best choice of V_*, we find empirically that the choice $V_* = V$ often works best, and when it is not best, it still performs well and so in the interest of computational efficiency, we do not compute the diagnostics.

(1) We demonstrate the approximation method analytically with an example that yields simple expressions for r and μ_0. Although simple, the model we define may be used to correctly implement the meta-analysis methods proposed by Moses *et al.* (1993) and Van Houwelingen *et al.* (1993).

Suppose we wish to evaluate the association between $\theta_y = \log\left(\frac{p_t}{1-p_t}\right)$ and $\theta_x = \log\left(\frac{p_c}{1-p_c}\right)$ with the linear model $E(\theta_y \mid \theta_x) = \beta_0 + \beta_0\theta_x$. Because binomial observations have natural parameter $\vartheta = \log\left(\frac{p}{1-p}\right)$, we can express this model more generally by $\theta_y = \vartheta_t$ and $\theta_x = \vartheta_c$. We define $\eta = \beta_0 + \beta_0\theta_x$, and so the augmentation distribution (5.17) requires that we approximate a density for ϑ_t which has log looking like

$$\ell_n(\vartheta_t) = \underbrace{n_t\hat{\mu}_t - n_t\Psi(\vartheta_t)}_{(a)} - \underbrace{\frac{1}{2\tau^2_{y|x}}(\vartheta_t - \eta)^2}_{(b)} \tag{5.35}$$

and has first and second derivatives given by

$$\ell'(\vartheta_t) = n_t\hat{\mu}_t\vartheta_t - n_t\Psi(\vartheta_t) - \frac{1}{\tau^2_{y|x}}(\vartheta_t - \eta) \tag{5.36}$$

$$\ell''(\vartheta_t) = -n_t\Psi''(\vartheta_t) - \frac{1}{\tau^2_{y|x}} \tag{5.37}$$

$$= -n_tV(\mu) - \frac{1}{\tau^2_{y|x}} \tag{5.38}$$

We first make a few observations regarding the shape of (5.35) with different amounts of information. The sample size n_t indicates a level of information about ϑ contained in (a), and the magnitude of $1/\tau_{y|x}$ indicates the information for ϑ contained in (b). With small $1/\tau^2_{y|x}$ (b) contains little information and term (a) influences the shape of the distribution. Because term (a) represents an NEF-QVF likelihood choosing $V_* = V$ leads to the best approximation. With large n (a) has nearly normal shape, or with large $1/\tau^2_{y|x}$ then (a) and (b) together have normal shape regardless of n, and the best approximating distribution has $V_* = 1$, the normal distribution. An MCMC algorithm will take on many values of $\tau_{y|x}$ so it is likely that for some iterations $V_* = V$ works best and for other choosing $V_* = 1$ works best.

If $\ell'(\vartheta^0_t) = 0$, $\ell_n(\vartheta_t)$ has mode ϑ^0_t then (5.32) leads us to choose

$$\mu_0 = \mu_*(\vartheta^0_t) \quad \text{and} \quad r = n_t\frac{V(\mu_0)}{V_*(\mu_0)} + \frac{1}{V_*(\mu_0)}\frac{1}{\tau^2_{y|x}} \tag{5.39}$$

Now notice that when $1/\tau^2_{y|x}$ and/or n_t have large value, then r is large also (this follows from (5.39) in particular and (5.32) in general). Because the Pearson family converges to the normal distribution with large r, then even when $V_* = 1$ is best, the choice of $V_* = V$ will not perform poorly. The next example demonstrates this graphically.

(2) Throughout the previous chapters we have chosen $\theta_y = \log\left(\frac{p_t}{p_c}\right)$ and $\theta_x = \log\left(\frac{p_c}{1-p_c}\right)$. This choice does not lead to convenient expressions for r and μ_0, but we may evaluate the approximations graphically. Recall that this choice of θ_{yi} and θ_{xi} has Jacobian

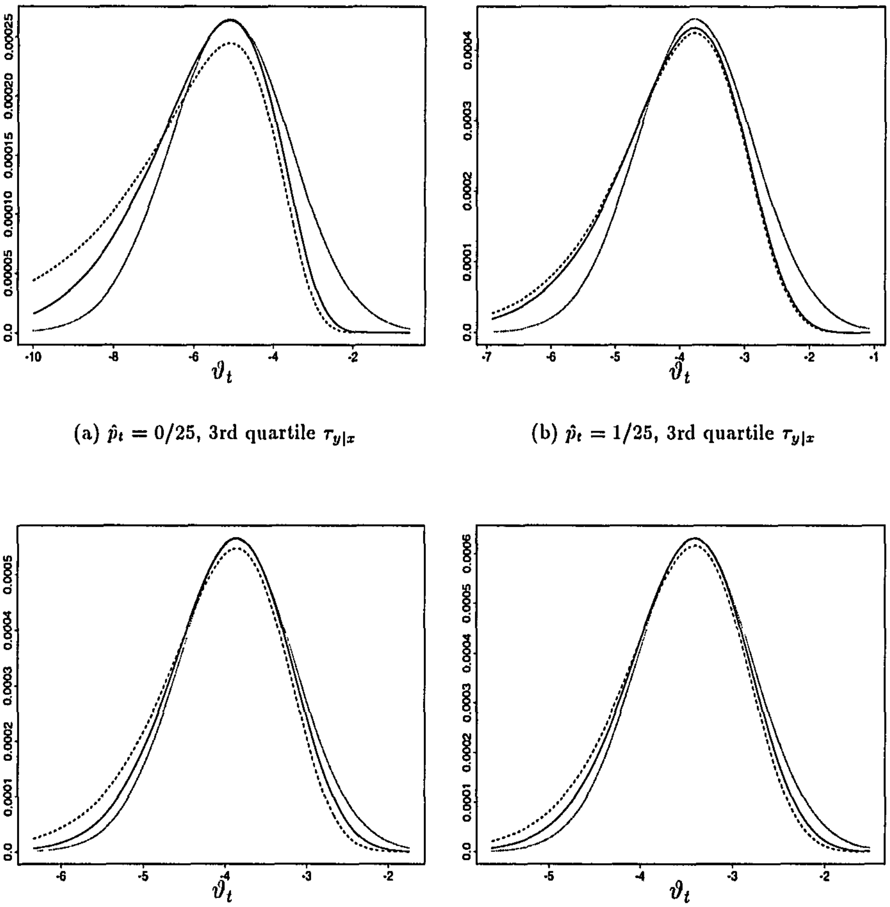
$$J(p) = \begin{pmatrix} \frac{1}{p_t} & -\frac{1}{p_c} \\ 0 & \frac{1}{p_c(1-p_c)} \end{pmatrix}$$

and so $\mid J(p) \mid = \frac{1}{p_t p_c(1-p_c)}$. This leads to augmentation steps for p_{ti} with the following form

$$\begin{aligned} p_n(\mu(\vartheta))V(\mu(\theta))d\vartheta_t = n_t\hat{\mu}_t \log\left(\frac{p_t}{1-p_t}\right) - n_t \log(1-p_t) && (5.40)\\ - \frac{1}{2\tau^2_{y|x}}\left(\log\left(\frac{p_t}{p_c}\right) - \beta_0 - \beta_\theta \log\left(\frac{p_c}{1-p_c}\right)\right)^2 && (5.41)\\ - \log\left(\frac{1}{p_t}\right) && (5.42)\\ - \log\left(V(p_t)\right) && (5.43) \end{aligned}$$

Term (5.42) represents $\log\left(\frac{d}{p_t}\log\left(\frac{p_t}{p_c}\right)\right)$, the component of $\log(\mid J(p_t,p_c)\mid)$ that involves p_t, and (5.43) represents log of the change in measure $\frac{dp}{d\vartheta}$. Although it has difficult form, especially if expressed in terms of $\vartheta_t = \log\left(\frac{p_t}{1-p_t}\right)$, its major mode and second derivative can be computed quickly on a computer.

At any particular iteration of the MCMC algorithm the shape of the true augmentation density depends on $\phi^{(n)} = (\beta_0^{(n)}, \beta_\theta^{(n)}, \tau_{y|x}^{(n)}, \tau_x^{(n)})'$ and $p_c^{(n-1)}$, and also on the observed mortality rate $\hat{p}_t$ and treatment group sample size n_t. Figure 5.1 shows examples of the true and approximate distributions when parameters are chosen to be typical values for the magnesium data, as given to us by the data analysis of Section 5.4.1 (see Table 5.3 on page 98 for the actual values). We use the posterior means of β_0 and β_θ and select $n_t = 25$ to correspond to the smallest magnesium trial. To show the shape of the augmentation density with different sources of information we have $\hat{p}_t$ take values 0/25 and 1/25, and we assign $\tau_{y|x}$ to be the median and 3rd quartile of its posterior distribution.



(a) $\hat{p}_t = 0/25$, 3rd quartile $r_{y|x}$

(b) $\hat{p}_t = 1/25$, 3rd quartile $r_{y|x}$

(c) $\hat{p}_t = 0/25$, median $r_{y|x}$

(d) $\hat{p}_t = 1/25$, median $r_{y|x}$

Figure 5.1: Plot of Normalized Density and Approximations: The curves are exact posterior(—-), Pearson approximation with $V_* = 1$ (normal) (.......), Pearson approximation with $V_* = V$ (logit of a beta) (- - - - -).

Figure 5.1(a) and Figure 5.1(b) show the true distribution (heavy solid line) when the binomial likelihood heavily influences its shape (with large $\tau_{y|x}$ and small n_t). Figure 5.1(c) and Figure 5.1(d) show the true distribution with small $\tau_{y|x}$ so the structural model greatly influences its shape. The heavily dotted line gives the density approximation when $V_* = V$ (the conjugate choice), and the lightly dotted line represents $V_* = 1$ (the normal approximation).

The two top figures show that with little information for ϑ_t the augmentation distribution has heavy skew, and the Pearson distribution with $V_* = V$ follows it nicely. In Figure 5.1(a) the Pearson approximation seems to over estimate the left tail a bit, but this does not cause alarm. With the Metropolis algorithm, it is better to over estimate a tail than underestimate it, because if underestimated too greatly the Metropolis algorithm will have difficulty obtaining samples in that region, and convergence slows.

Figure 5.1(c) and Figure 5.1(d) use smaller heterogeneity components than do the figures on the top, and look considerably more normal. The Pearson conjugate approximation appears to fit better than the normal approximation in the left tail, but worse in the right tail. But in every case, the Pearson approximation fits well. The next section provides further evidence that the Pearson approximations perform well.

5.4 Data analysis

We now apply the small sample procedure to the magnesium and streptokinase data. To demonstrate the capabilities of the extended structural model we use several ecological models. Because of the important policy implications of the magnesium data, we concentrate on those data analyses.

For each of the analyses that follow we run $J = 5$ independent MCMC sequences and monitor β_0 to assess convergence (see Section 4.2.3 for description of the convergence

Table 5.2: Data from nine clinical trials evaluating intravenous magnesium for treatment of AMI (sorted by magnitude of treatment effect). The columns are: trial name; treatment and control group mortality rates, $\hat{p}_t$ and $\hat{p}_c$; treatment and control group sizes, n_t and n_c; treatment effect estimate in log relative risk and its standard error, $\log(\frac{\hat{p}_t}{\hat{p}_c})$ and $\hat{\sigma}$; log odds of mortality in the control group, $\log(\frac{\hat{p}_c}{1-\hat{p}_c})$. The final row gives the unweighted means of the columns.

Trial	$\hat{p}_t$	$\hat{p}_c$	n_t	n_c	$\log(\frac{\hat{p}_t}{\hat{p}_c})$	$\hat{\sigma}$	$\log(\frac{\hat{p}_c}{1-\hat{p}_c})$
Feldsted	0.067	0.054	150	148	0.210	0.460	-2.862
ISIS 4	0.076	0.072	29011	29039	0.051	0.031	-2.556
Abraham	0.021	0.022	48	46	-0.043	1.399	-3.807
LIMIT 2	0.078	0.103	1150	1150	-0.271	0.134	-2.169
Morton	0.025	0.056	40	36	-0.799	1.203	-2.833
Rasmussen	0.067	0.170	135	135	-0.938	0.374	-1.583
Ceremuzynski	0.040	0.130	25	23	-1.182	1.118	-1.897
Schecter '95	0.042	0.173	96	98	-1.384	0.536	-1.561
Schechter	0.017	0.161	59	56	-2.249	1.037	-1.653
Means	0.048	0.104	3235	3233	-0.734	0.699	-2.324

method). The Metropolised data augmentation algorithm typically requires over 4,000 iterations to reach convergence, which is over twice the number of iterations the large sample procedure to requires. As Section 4.3 recommends we choose prior distributions uniform on β, γ, $\tau_{y|x}$ and τ_x^2.

5.4.1 Data Analyses of the Magnesium Trials

Chapter 1 introduced the magnesium data, but for convenience we reproduce it here.

We analyze the magnesium data with three different ecological models. We first use an ecological model that has been treated throughout this manuscript, a specification we call the "standard" model. We also estimate the association favored by Lau *et al.* (1995) and Antman (1995b) that relates the log relative risk linearly to the control group mortality rate. Because this is the specification preferred in many ongoing projects at the New England Medical Center, we call this the "NEMC" model. Finally, we use the standard model but add a squared term to the ecological component. Although these models are quite different, we find next that they give strikingly similar conclusions.

Standard model

We use treatment effect and population risk definitions $\theta_{yi} = \log\left(\frac{p_{ti}}{p_{ci}}\right)$ and $\theta_{xi} = \log\left(\frac{p_{ci}}{1-p_{ci}}\right)$ to analyze the magnesium data. The top half of Table 5.3 summarizes the posterior distribution of the structural model parameters. The bottom half of the table summarizes the posterior distributions of some derived quantities we introduce below. Figure 5.2 shows the marginal posterior distribution of selected structural parameters.

The small sample procedure gives the ecological slope posterior mean as $\hat{\beta}_\theta = -1.024$, with only 2.3% of the samples falling above zero, and a 95% interval estimate for β_θ from -2.323 to -0.020. Figure 5.2(a) shows the posterior distribution of β_θ has slightly right skew. We conclude from these results that the magnesium data has a non zero ecological slope.

We may also wish to estimate the overall mean treatment effect of magnesium when we do not control for the population risk[1]. We compute the posterior distribution of the mean treatment effect, which we we denote by μ_y, from the MCMC sequence as follows. If we let $\beta_0^{(i)}$ and $\beta_\theta^{(i)}$ represent the ecological coefficients at iteration i and $\gamma_0^{(i)}$ represent the mean population risk at iteration i then $\mu_y^{(i)} = \beta_0^{(i)} + \beta_\theta^{(i)}\gamma_0^{(i)}$ represents the mean treatment effect estimate at that iteration. The histogram of $\mu_y^{(i)}$ estimates the marginal posterior distribution of μ_y, and Figure 5.2(b) plots it. Table 5.3 summarizes the posterior distribution and gives a posterior mean $\hat{\mu}_y = -.582$ and a 95% interval from -1.579 to 0.210, with 7% of the mass falling above $\mu_y = 0$. These estimates differ from when we apply the large sample procedure (not shown) which gives $\hat{\mu}_y = -0.470$ but with a 95% interval that excludes $\mu_y = 0$. We conclude that on average trials of magnesium show benefit but there

[1] We remind the reader that the intercept β_0 represents the expected value of θ_{yi} at $\theta_{xi} = 0$. This differs from the specification given in Chapter 4 where, because of the deviations in mean parameterization of the ecological model, the intercept term represented μ_y.

Table 5.3: Results from magnesium data analyses, with columns: Mean, posterior mean estimate; Std.Dev., posterior standard deviation; P-val, the fraction of samples greater than zero, representing a Bayesian 1-sided p-value; Quantiles, the quantiles of the posterior distribution.

	Posterior Estimates			Quantiles				
	Mean	Std.Dev.	P-val	0.025	0.05	0.50	0.95	0.975
β_θ	-1.024	0.496	0.023	-2.323	-1.977	-1.023	-0.271	-0.02
β_0	-2.795	1.137	0.012	-5.958	-5.179	-2.759	-1.123	-0.625
$\tau_{y\|x}$	0.348	0.295	1	0.005	0.012	0.275	0.909	1.394
γ_0	-2.245	0.294	0	-3.028	-2.868	-2.238	-1.788	-1.664
τ_x	0.717	0.356	1	0.240	0.283	0.645	1.356	1.098
μ_y	-0.516	0.392	0.07	-1.618	-1.327	-0.49	0.071	0.214
ζ	-2.788	0.540	0.001	-5.035	-4.25	-2.65	-2.423	-2.347

is substantial risk that some will cause harm.

We may evaluate how likely it is for a future trial of magnesium to be harmful as follows. Suppose a future trial has population risk c standard deviations from the mean: $\theta^+_{xi} = \gamma_0 + c\tau_x$. With ϕ known, a future trial θ^+_{yi} has normal distribution with mean $\beta_0 + \beta_\theta \theta^*_{xi}$ and variance $\tau^2_{y|x}$, and so $P(\theta^+_{yi} > 0|\phi)$ can be evaluated from a normal table. Averaging these values over the samples $\phi^{(i)}$ gives a posterior mean estimate of risk. Choosing $c = 0$, so we estimate the risk of a trial with average population risk, we get $\widehat{P}(\theta_{yi} > 0|\hat{\theta}) = 0.14$, a significant risk of harm. If we choose $c = 1$, so the population risk is one standard deviation from its mean, then we get $\widehat{P}(\theta_{yi} > 0|\hat{\theta}) = 0.001$. Thus treating sicker populations gives substantially smaller risk.

By setting to zero the expected treatment effect, $0 = \theta_y = \beta_0 + \beta_\theta \theta_x$, and solving for θ_x we estimate the population risk that can expect to have $\theta_{yi} = 0$. We represent that point by $\zeta = -\beta_0/\beta_\theta$, and approximate its posterior distribution from the histogram of $\zeta^{(i)} = -\beta_0^{(i)}/\beta_\theta^{(i)}$. We will find that when $\beta_\theta^{(i)}$ is near zero then $\zeta^{(i)}$ is extremely large or small, depending on the sign of $\beta_0^{(i)}$. Thus Figure 5.2(c) plots the posterior distribution of ζ trimming 0.1% of its tails, and Table 5.3 summarizes the trimmed distribution. The parameter ζ has posterior mean $\hat{\zeta} = -2.738$, with a 95% quantile at -2.423. This means that with probability 0.95 trials with $\theta_{xi} < -2.423$ can expect to find $\theta_{yi} > 0$. The ISIS 4 trial has population risk $\theta_{xi} = -2.602$, lying just near the mode of the posterior distribution for ζ (its mode is -2.650). This is not likely a coincidence. Because ISIS 4 falls almost exactly on the line $\theta_{yi} = 0$, and because of its large size, it contributes a considerable amount of leverage on the posterior distribution of ζ.

Here we briefly evaluate the performance of our data augmentation density approximations by examining the acceptance probabilities. Table 5.4 summarizes the jumping probabilities we found for each augmented parameter, giving their 1%, 10% and median values. We find that the worst performing augmentation step comes from imputing the Abraham control group mortality rate, and that has a median acceptance rate of 96.9%. We consider this substantially close to the optimal value of 1. Notice the acceptance rate for ISIS 4. Because of its large size its imputation step certainly is almost exactly normal, and here we see the Pearson approximation, which fits a beta distribution to the augmentation step, finds a median acceptance rate of 0.987 when augmenting p_{ci} and .992 when augmenting p_{ti}. We conclude that our Pearson approximations perform extremely well.

NEMC Model

We now define $\theta_{yi} = \beta_0 + \beta_1 p_{ci}$, where p_{ci} represents the control group mortality rate. So that we may continue to treat $\theta_{yi} = \log\left(\frac{p_{ci}}{1-p_{ci}}\right)$ as having normal distribution we choose $\chi_i(\theta_{xi}, Z_i) = \frac{e^{\theta_{xi}}}{1+e^{\theta_{xi}}}$. Table 5.5 summarizes the posterior distributions. Notice that the population risk parameters γ_0 and τ_x in Table 5.5 have posterior distribution nearly identical to those found when using the standard model. This is because the population risk model remains unchanged so the posterior distributions must also remain unchanged. Any difference between the two estimates must be due to the variability of the MCMC algorithm.

We find the ecological slope highly significant, with posterior mean $\hat{\beta}_0 = -12.472$, and only 1% of its mass falling above zero. The population mortality rate that gives zero

Table 5.4: Assessment of augmentation step approximations. Left columns gives the augmentation performance for the control group and the left gives it for the right. Column give quantiles of the jumping probabilities.

	Control			Treatment		
Trial	1%	10%	50%	1%	10%	50%
Feldsted	0.214	0.730	0.969	0.665	0.925	0.994
ISIS 4	0.785	0.909	0.987	0.811	0.931	0.992
Abraham	0.012	0.678	0.969	0.685	0.913	0.996
LIMIT 2	0.331	0.746	0.973	0.705	0.914	0.995
Morton	0.437	0.722	0.977	0.666	0.932	0.996
Rasmussen	0.538	0.775	0.970	0.653	0.893	0.993
Ceremuzynski	0.625	0.837	0.978	0.605	0.874	0.993
Schecter'95	0.942	0.976	0.997	0.951	0.976	0.997
Schechter	0.030	0.631	0.967	0.730	0.944	0.996

treatment effect, ζ, has posterior expectation $\hat{\zeta} = 0.063$.

Comparing different models

Although the NEMC and standard specifications suggests an ecological association exists, their specifications are linear on different scales. We view our procedure as being primarily data analytic, and so we do not think that of there being a "true" model, but we do have concern if two different specifications lead to different conclusions (we will comment on this point again in the conclusion to this manuscript). Here we compare the conclusions of NEMC and standard models.

Figure 5.3 plots the NEMC structural model on the scale of the standard model. That is, the NEMC curve represents $\theta_y = \beta_0 + \beta_1 \left(\frac{e^{\theta_x}}{1+e^{\theta_x}}\right)$. We find the standard and NEMC models agree in the region containing most of the observations. We will comment on this further when we estimate the quadratic model. We may also compare ζ by transforming the NEMC estimate to the scale of the standard model. We get logit $(0.063) = -2.700$, which is close to the standard models estimate of -2.788. Furthermore, McIntosh (1996) used the large sample model with $\theta_{yi} =$ logit $(\theta_{yi}) -$ logit (θ_{xi}) and $\theta_{xi} = p_{ci}$ to estimate

Table 5.5: Magnesium data posterior summary with NEMC model.

	Posterior Estimates			Quantiles				
	Mean	Std.Dev.	P-val	0.025	0.05	0.50	0.95	0.097
β_θ	-12.472	4.137	0.01	-23.411	-20.085	-12.199	-6.556	-3.976
β_0	0.816	0.468	0.945	-0.5927	-0.292	0.843	1.496	1.662
$\tau_{y\|x}$	0.267	0.256	1	0.008	0.012	0.202	0.734	0.899
γ_0	-2.243	0.299	0	-3.128	-2.974	-2.244	-1.788	-1.678
τ_x	0.701	0.339	1	0.297	0.301	0.701	1.401	1.22
ζ	0.063	0.025	0.965	-0.036	-0.005	0.071	0.084	0.089

the ecological association for magnesium and found $\zeta = 0.062$, which is strikingly close to the value estimated by the NEMC model. Thus we find that changing the definitions of both θ_{yi} and θ_{xi} do not change our conclusions about which populations can benefit from magnesium therapy.

5.4.2 Quadratic Model

We investigate more complex ecological associations by introducing a quadratic term into the standard model and estimate $\theta_{yi} = \beta_0 + \beta_{\theta,1}\theta_{xi} + \beta_{\theta,2}\theta_{xi}^2$. We accomplish this by choosing $\chi_i(\theta_{xi}, Z_i) = (\theta_{xi}, \theta_{xi}^2)'$. Table 5.6 with the standard model.

The 95% posterior interval for the quadratic term is between -11.86 and 3.99, with over 10% of its mass falling above zero. We have little evidence suggesting a quadratic ecological model, and so in the interest of parsimony, we prefer the standard model.

For completeness Figure 5.3 includes the quadratic ecological model constructed from the posterior means given by Table 5.6. Notice that quadratic model agrees with the other models in the region where the majority of the data are found, with the Abraham study being one possible exception.

If the quadratic model were to hold, would we expect a future trial similar to Abraham

Table 5.6: Quadratic model results: Parameter $\beta_{\theta,2}$ is the coefficient of θ_{xi}^2, $\beta_{\theta,1}$ is the coefficient of θ_{xi} and $\beta_{\theta,0}$ is the intercept when estimating the regression $\theta_{yi} = \beta_0 + \beta_{\theta,1}\theta_{xi} + \beta_{\theta,2}\theta_{xi}^2$.

	Posterior Estimates			Quantiles				
	Mean	Std.Dev.	P-val	0.025	0.05	0.50	0.95	0.975
$\beta_{\theta,2}$	-1.713	4.092	0.113	-11.868	-8.392	-1.415	1.054	3.991
$\beta_{\theta,1}$	-9.008	19.375	0.094	-56.679	-41.08	-7.388	3.616	16.123
β_0	-11.886	22.752	0.067	-68.699	-49.932	-9.828	2.03	16.088
$\tau_{y\|x}$	0.285	0.342	1	0.003	0.006	0.172	0.963	1.254
γ_0	-2.283	0.291	0	-3.064	-2.893	-2.275	-1.831	-1.722
τ_x	0.716	0.357	1	0.207	0.248	0.644	1.368	1.657

to find $\theta_{yi} > 0$? We must be cautious when making such interpretations. The curves in Figure 5.3 represent the associations relating θ_{yi} and θ_{xi}, but the data points represent $\hat{\theta}_{yi}$ and $\hat{\theta}_{xi}$. To answer this question we must first estimate Abraham's true θ_{xi}. Although this manuscript did not treat estimating θ_{xi}, we briefly give two ways to estimate θ_{xi}.

One way uses the augmentation steps of the data augmentation algorithm to estimate the posterior distribution of Abraham's θ_{xi}. The mean values of $\theta_{xi}^{(i)}$ may be used as an estimate of θ_{xi}. For Abraham we find that θ_{xi} have average value -2.922, which places Abraham in the region of the structural models where they agree. Another method to estimate the individual θ_{xi} uses Bayes or empirical Bayes methods and estimates them without regard to $\hat{\theta}_{yi}$. As Chapter 3 and Chapter 4 pointed out, the population risks marginally follow a univariate hierarchical model and so the population risk model may be analyzed separately from the ecological model. For example, if trials are large enough so that we may assume normally distributed within trial measurements, then we may use Morris (1983b) to estimate θ_{xi}. If the data have a Poisson distribution, or have binomial distribution with small event probabilities (so that the Poisson approximation holds) then we may use the PRIMM software program (reference here) to estimate the individual θ_{xi}. For completeness, if we apply Morris (1983b) to estimate θ_{yi} for Abraham we find $\hat{\theta}_{yi} = -2.862$.

We may use this when making predictions with the ecological models. Clearly evaluating and developing methods of forecasting and prediction presents an opportunity for future research.

As with all data analyses we must be careful and interpret our model only in the range for which we have information. We have most information at $\theta_{xi} = \gamma_0$, the mean population risk, and as θ_{xi} moves away from γ_0 the information diminishes. Although we have no quantitative results, we recommend treating the structural models as valid only in the region given by $\hat{\gamma}_0 \pm 1.5\hat{\tau}_x$. The vertical lines in Figure 5.3 represent this region for the magnesium data. The argument for choosing values less than 1.96 (a 95% interval) is in consideration of the uncertainty of estimating tails of any distribution. We choose 1.5 for idiosyncratic reasons.

Because we find a significant ecological slope with a variety of treatment effect, population risk, and ecological models, we conclude their exists overwhelming evidence for the existence of an ecological association for magnesium. As we began in Chapter 1, we find that differences in the population risk can account for substantial between trial heterogeneity.

5.4.3 Streptokinase data

Table 5.7 summarizes the posterior distribution for the streptokinase data when using the standard model, and Table 5.8 summarizes the posterior distribution when we add a squared term. Figure 5.4 plots the two structural models.

The standard model gives the posterior mean for β_0 as $\hat{\beta}_0 = 0.042$, with almost 60% of its mass falling above zero. We conclude from this that the streptokinase does not have an ecological association linear in this population risk. Overall streptokinase has mean treatment effect $\hat{\mu}_y = -0.241$, with a 95% posterior interval from -0.43 to -0.144, and so we can be confident that on average trials of streptokinase are beneficial. Although

Table 5.7: Streptokinase results for standard specification.

	Posterior Estimates			Quantiles				
	Mean	Std.Dev.	P-val	0.025	0.05	0.50	0.95	0.975
β_{0}	0.042	0.157	0.598	-0.382	-0.245	0.048	0.258	0.312
β_0	-0.175	0.283	0.320	-0.984	-0.737	-0.169	0.247	0.329
$\tau_{y\|x}$	0.110	0.098	1	0.003	0.005	0.083	0.293	0.348
γ_0	-1.771	0.097	0	-1.989	-1.963	-1.774	-1.605	-1.566
τ_x	0.462	0.081	1	0.317	0.328	0.453	0.607	0.647
μ_y	-0.241	0.060	0	-0.43	-0.389	-0.244	-0.165	-0.144

Table 5.8: Streptokinase quadratic model specifications.

	Posterior Estimates			Quantiles				
	Mean	Std.Dev.	P-val	0.025	0.05	0.50	0.95	0.975
β_2	0.157	0.788	0.609	-1.829	-1.463	0.19	1.348	1.584
β_1	0.428	2.746	0.582	-6.376	-4.995	0.494	4.68	5.475
β_0	-0.171	2.297	0.473	-5.979	-4.383	-0.129	3.382	4.233
$\tau_{y\|x}$	0.585	0.147	1	0.313	0.344	0.57	0.854	0.931
γ_0	-1.769	0.107	0	-2.005	-1.98	-1.757	-1.576	-1.553
τ_x	0.482	0.1	1	0.305	0.322	0.471	0.662	0.718

their is no ecological association, the probability of a future trial causing harm is only $\widehat{P}(\theta_{yi}^{+} > 0|\hat{\theta}) = 0.042$, nearly half that found by the large sample procedure. Thus the small sample procedure suggests that heterogeneity may be low enough so that general use of streptokinase has minimal risk.

Table 5.8 does not find the quadratic term significant, and so in the interest of parsimony we prefer the standard model. Figure 5.4 plots the two structural models together, and shows that both give close agreement over the range of the observed data.

5.5 Summary of simulations

The small sample procedure requires a substantial amount of time (in hours) to reach convergence. For example, the magnesium trials typically achieved convergence shortly after one hour of computing time. This makes comprehensive analysis with simulation prohibitive because typically several hundred runs are required for each study. Here we will use simulation and concentrate on evaluating the performance of the small sample procedure for only the situation where the large sample procedure fails most completely. We also conduct a few smaller simulation studies to provide evidence that the small sample procedure does not do worse than the large sample procedure in other circumstances. That is, we wish to provide evidence that the small sample procedure will not do worse than, and at times will do better than, the large sample procedure. We restrict our attention to binomial measurement.

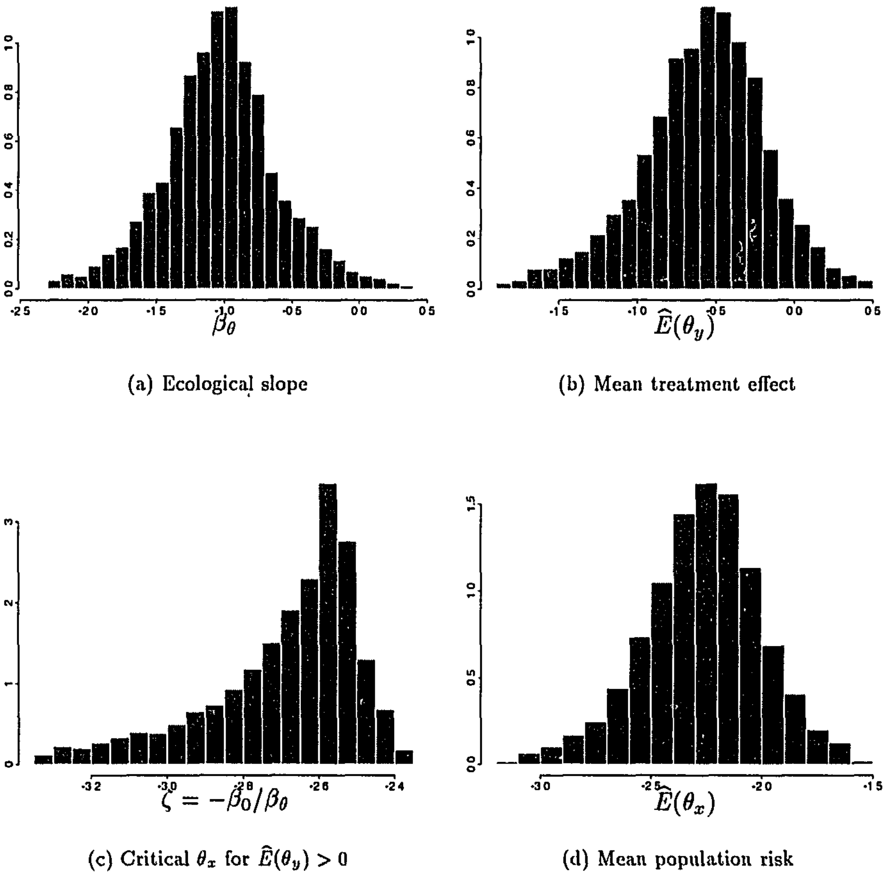
Recall that the three sets of simulations given in Section 4.5 used parameters ϕ and sample sizes k and $n_{i.}$ to represent streptokinase trials. (see page 4.5 for a detailed description of the simulation procedure). We found that the large sample Bayes procedure appears nearly unbiased for both, but the maximum likelihood procedure gives adequate correction only when the measurement errors have normal distribution. The top two rows of Table 5.9 summarizes the performance of the small sample procedure with these configurations. The row with label "Streptokinase" summarizes the performance when simulations match the streptokinase data. We only perform these simulations to test the hypothesis that the small sample procedure does not do worse, and so we use only 100 iterations. Because each estimate falls within two standard errors of their true values, these simulation results are consistent with the small sample procedure being unbiased and having honest coverage

A third set of simulations from Section 4.5 used the smallest nine streptokinase trials for a simulation study. Recall that the least squares methods all estimated regression slopes

Table 5.9: Summary of small sample simulations. The top row gives the simulation results when the small sample procedure is applies to all 33 streptokinase trials when ϕ matches the posterior mean estimates in Table 5.7. The row with label "Small Streptokinase" gives the results when only the smallest nine streptokinase trials are used for simulation. Compare these results to those in Section 4.5.

		Coverage	
Method	$\tilde{E}(\hat{\beta}_0)$	90%	95%
Structural Methods:			
Streptokinase:	0.005	0.87	0.93
(n=100)	(0.014)	(0.033)	(0.026)
Small Streptokinase	-0.098	0.94	0.97
(n=250)	(0.051)	(0.015)	(0.011)

near 0.5, and the structural methods found the quite disturbing result that the Bayes and maximum likelihood procedures *increased* the bias compared to the least squares methods. Recall that the large sample procedure Bayes estimates of β_0 averaged -0.631, nearly ten standard errors from $\beta_0 = 0$. We focus our attention on this set of simulations and so we use a greater number of repetitions ($n = 250$). The row in Table 5.9 with label "Small Streptokinase" summarizes this simulation study. The small sample procedure estimates of β_0 averaged $\beta_0 = -0.098$, and is -0.098/0.051=-1.92 standard errors from $\beta_0 = 0$. This suggests that there may be mild under adjustment of the slope, but otherwise corrects a substantial portion of the bias. The interval estimates appear to be slightly greater than nominal coverage, and so we may consider the small sample procedure as giving slightly conservative interval estimates.



(a) Ecological slope

(b) Mean treatment effect

(c) Critical θ_x for $\widehat{E}(\theta_y) > 0$

(d) Mean population risk

Figure 5.2: Posterior distributions for selected magnesium data parameters. Computed from the second half of 5 parallel runs of the Metropolised data augmentation algorithm, for a total of 10,000 iterations.

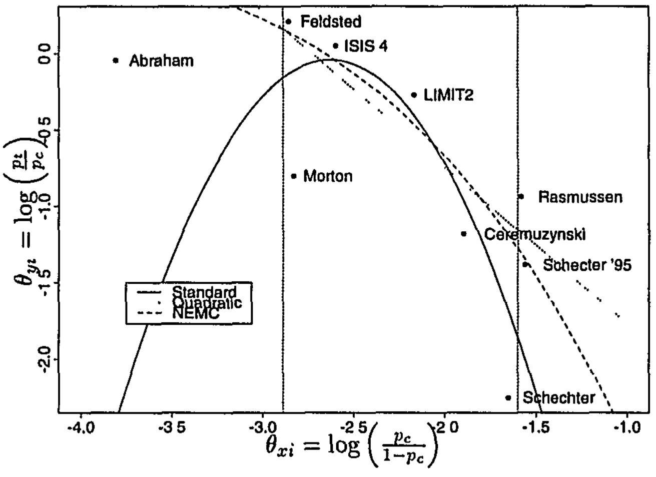


Figure 5.3: Plot of magnesium data ecological models.

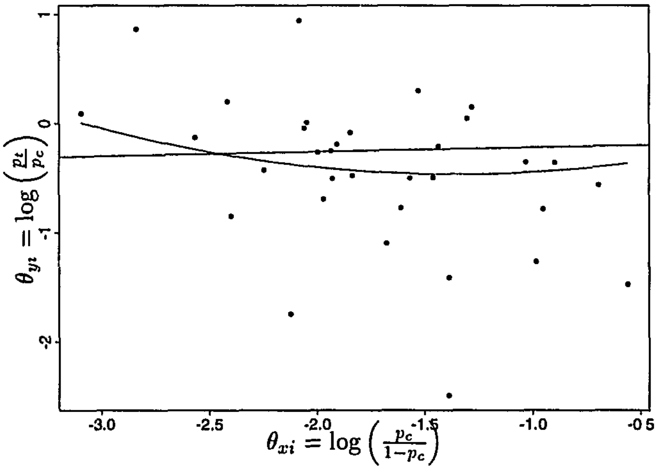


Figure 5.4: Plot of streptokinase ecological models.

Chapter 6

Conclusions and Summary

This thesis proposes a procedure to perform meta-analysis using ecological covariates. Ecological covariates are aggregate values that measure attributes, or the ecology, of the treated populations. Perhaps the most useful ecological covariate is constructed from the control group outcomes, a value we call the "population risk." It is most useful because it is always available and the information we need to correct for biases can be always be estimated from clinical summaries.

The population risk also has practical use because it has an intuitive meaning that physicians find appealing, and because its value is affected by many factors that physicians would like to control for but cannot because of the lack of covariates. This is dramatically demonstrated by the magnesium data, where "time until treatment", an important factor that is available for only a few studies, correlates highly with the population risk.

This meta-analysis method has already been proven useful for making health policy decisions. The conclusion of the magnesium data analyses has been used as evidence supporting a larger body of theories proposed by Dr. Elliott Antman, and also as part of a justification for a new clinical trial. The work for this manuscript began because Dr. Antman wished to determine if higher risk populations may benefit from magnesium treatment.

Researchers Christopher Schmid, Joseph Lau, Joseph Cappelleri, and John Ionnidis, of the New England Medical Center, are also conducting a large body of related work using this technology. They demonstrate the value of using the control group risk as a covariate to help answer many important medical questions. For example, Lau *et al.* (1995) shows that controlling for population risk has some value for explaining the discrepancies between large and small clinical trials. They also find that in meta-analyses where some individual data are available the association of patients individual risks with treatment efficacy has magnitude similar to the effect estimate from the methods described in Chapter 4. That is, when concluding that sicker populations benefit from treatment, it also seems to follow that sicker patients within the trials benefit more than the patients who are less sick.

At first it appears that controlling for the population risk should be as simple as including its value as a covariate in a linear model. However, we have shown that doing this may lead to coefficient estimates that are meaningless. Correcting this bias requires considerable time and effort, and also requires that we know the within trial regression slopes, which may not be available. If the magnitude of measurement error is small then simply using the ecological covariates in a linear model *does* lead to meaningful inferences. It is valuable to know when we can ignore the measurement error so that we can use ecological covariates when the within trial regression slopes are not known. Toward this end we also provide a method to quantify the measurement error bias so that we may determine when the measurement error may be ignored. We discuss this at the end of Chapter 3.

We develop two models and two corresponding estimating procedures to estimate the ecological model. One method, given in Chapter 3 and Chapter 4, assumes that the population risk and treatment effects have normal distributions. Simulation results show that the likelihood based inferences given by Chapter 4 have good frequency properties under many conditions, but fails when within study samples are not large enough for the normal measurement error approximation to hold. The model of Chapter 5 correct these deficiencies. Limited simulation results suggest that has been accomplished.

Because the normal model of Chapter 4 is simpler and quicker to use, it would be useful to know it leads to valid inferences, because the procedure in Chapter 5 is more complicated. The simulation results at the end of Chapter 4 suggest that the non-normal measurement error can be ignored when "some" of the trials have large samples and heterogeneity is not "too" great. No concise rule is available. Deriving such a rule is an opportunity for further research.

Possibly one definition of treatment effect and population risk will find significant ecological association, but another choice will not. In fact, if an ecological association does not exist with one specification, and ecological association will necessarily exist on another. For example, if the log-relative risk is constant as the control group mortality changes, then the log-odds ratio will not be constant. Perhaps the best discussion of this can be found in Sinclair and Bracken (1994). The data analyses in Chapter 5 demonstrated that many different ecological models lead to strikingly similar conclusions for the streptokinase and magnesium data, and our experience with other data sets shows that this kind of multiplicity often holds true. We have no assurance that this will hold true for all data sets. Clearly, developing methods for model checking or evaluation is an important topic for future research.

References

Antman, E. M. (1995a). Magnesium in acute MI. *Circulation* **92**, 2367–2372.

Antman, E. M. (1995b). Randomized clinical trials of magnesium in acute myocardial infarction: Big numbers do not tell the whole story. *The American Journal of Cardiology* **75**, 391–393.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., and Colditz, G. A. (1994). A random-effect regression model for meta-analysis. *Statistics in Medicine* **14**, 395–411.

Box, G. E. P. and Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis.* John Wiley & Sons.

Brand, R. and Kragt, H. (1991). Letters to the editor. *Statistics in Medicine* **13**, 297–298.

Brand, R. and Kragt, H. (1992). Importance of trends in the interpretation of an overall odds ratio in the meta-analysis of clinical trials. *Statistics in Medicine* **11**, 2077–2082.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Sage Publications.

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics* **1**, 101–129.

Davies, R. B. and Hutton, B. (1975). The effect of errors in the independent variables in linear regression (corr: V64 p655). *Biometrika* **62**, 383–392.

Dempster, A. P. (1969). *Elements of Continuous Multivariate Analysis.* Addison-Wesley.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the Em algorithm (c/r: P22-37). *Journal of the Royal Statistical Society, Series B, Methodological* **39**, 1–22.

DerSimonian, R. and Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

DuMouchel, W. (1990). Bayesian metaanalysis. In *Statistical Methodology in the Pharmaceutical Sciences*, 509- 529.

DuMouchel, W. and Waternaux, C. (1992). Discussion of: Hierarchical models for combining information and for meta-analyses. In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, 338- 339.

Everson, P. (1995). *Inference in Hierarchical Normal Models.* Ph.D. thesis, Department of Statistics, Harvard University.

Fahey, M. T., Irwig, L., and Macaskill, P. (1995). Meta-analysis of pap test accuracy. *American Journal of Epidemiology* **141**, 680–687.

Fuller, W. A. (1987). *Measurement Error Models.* John Wiley & Sons.

Gelfand, A. E., Hills, S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association* **85**, 972–985.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398–409.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (1995). *Bayesian Data Analysis.* Chapman & Hall.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences (disc: P483-501, 503-511). *Statistical Science* **7**, 457–472.

Geyer, C. J. (1992). Practical Markov chain Monte Carlo (disc: P483-503). *Statistical Science* **7**, 473–483.

Gilbert, J. P., McPeek, M., and Mosteller, F. (1988). Progress in surgery and anesthesia: benefits and risks of innovative therapy. In *Costs, Risks, and Benefits of Surgery*, 124–169.

Graver, D., Draper, D., Greenhouse, J., Hedges, L., Morris, M., and Waternaux, C. (1992). *Combining Information. Statistical Issues and Opportunities for Research.* Washington DC: National Academy Press.

Gupta, V. k. (1996). Letters to the editor. *Cardiovascular Drugs and Therapy (to appear)* .

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109.

ISIS-4 Collaborative Group (1995). Isis-4: A randomized factorial trial assessing early oral captopril, oral mononitrate, and intravenous magnesium sulphat in 58050 patients with suspected acute myocardial infarction. *The Lancet* **345**, 669–684.

Laird, N. M. and Louis, T. A. (1989). Empirical Bayes confidence intervals for a series of related experiments. *Biometrics* **45**, 481–495.

Langbein, L. I. and Lichtman, A. J. (1978). *Ecological Inference.* Sage Publications.

Lau, J., Schmid, H. C., Ioannidis, J. P., McIntosh, M. W., Cappelleri, J. C., Lau, J., and Chalmers, T. C. (1995). The importance of the baseline risk in explainaing disrepancies between clinical trials. a control rate meta-regression approach. *(submitted to) Controlled Clinical trials* .

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73**, 13–22.

McIntosh, M. W. (1996). The control group risk as an explanantory variable in research syntheses of clinical trials. *Statistics in Medicine(to appear)* .

Meng, X.-L. and Rubin, D. B. (1991). Using Em to obtain asymptotic variance-covariance matrices: The Sem algorithm. *Journal of the American Statistical Association* **86**, 899–909.

Miller, Rupert G., J. (1986). *Beyond ANOVA, Basics of Applied Statistics.* John Wiley & Sons.

Morris, C. N. (1982). Natural exponential families with quadratic variance functions. *The Annals of Statistics* **10**, 65–80.

Morris, C. N. (1983a). Natural exponential families with quadratic variance functions: Statistical theory. *The Annals of Statistics* **11**, 515–529.

Morris, C. N. (1983b). Parametric empirical Bayes inference: Theory and applications (c/r: P55-65). *Journal of the American Statistical Association* **78**, 47–55.

Morris, C. N. (1988). Approximating posterior distributions and posterior moments. In *Bayesian Statistics 3*, 327– 344.

Morris, C. N. (1995). Hierarchical models for educational data: An overview. *Journal of Educational and Behavioral Statistics* **20**, 190–200.

Morris, C. N. and Normand, S. L. (1992). Hierarchical models for combining information and for meta-analyses (disc: P335-344). In *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, 321– 335.

Moses, L. E., Shapiro, D., and Littenberg, B. (1993). Combining independent studies of a diagnostic test into a summary Roc curve: Data-analytic approaches and some additional considerations. *Statistics in Medicine* **12**, 1293–1316.

Mosteller, F. and Colditz, G. A. (1994). Understanding research synthesis (meta-analysis). *Annual Review of Public Health* **17**, 1–23.

Mosteller, F. and Tukey, J. W. (1977). *Data Analysis and Regression: A Secondary Course in Statistics.* Addison-Wesley.

Olkin, I. (1995). Meta-analysis: Reconciling the results of independent studies. *Statistics in Medicine* **14**, 457–472.

Peto, R., Collins, R., and Gray, R. (1988). Large scale randomized evidence: Large, simple trials and overviews of trials. In *Annals of the New York Academy of Sciences*, 314– 339.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-gibbs sampler. *Journal of the American Statistical Association* **87**, 861–868.

Robinson, W. S. (1995). Ecological correlations and the behavior of individuals. *American Sociological Review* **15**, 351–357.

Schmid, H. C., McIntosh, M. W., Cappelleri, J. C., Lau, J., and Chalmers, T. C. (1995). Measuring the impact of the control rate in meta-analysis of clinical trials. *(abstract in) Controlled Clinical trials* **16**, 665.

Seber, G. A. F. (1977). *Linear Regression Analysis.* John Wiley & Sons.

Senn, S. (1991). Letters to the editor. *Statistics in Medicine* **13**, 293–296.

Sinclair, J. C. and Bracken, M. B. (1994). Clinically useful measures of effect in binary analyses of randomized trials. *Journal of Clinical Epidemiology* **47**, 881–889.

Steward, J. H. (1990). Ecology. In *Encyclopedia of the Social Sciences*, vol. 4.

Tanner, M. A. and Wong, W. H. (1987). The calculation of posterior distributions by data augmentation (c/r: P541-550). *Journal of the American Statistical Association* **82**, 528–540.

Van Houwelingen, H. C., Zwinderman, K. H., and Stijnen, T. (1993). A bivariate approach to meta-analysis. *Statistics in Medicine* **12**, 2273–2284.